THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

November 2008 Vol. 9 No. 1 (ISSN 1727-5997)

Profile

Intelligent Systems in Nanjing University. . . . . *Yang Gao, Lin Shang & Yubin Yang* 1

Conference Report

ECAI 2008 Workshops on Configuration and Recommender Systems: Two Converging Research Fields. . . . . *Markus Zanker & Juha Ttiihonen* 3

Feature Articles

Cross-domain Text Classification using Wikipedia. . . . . *Pu Wang, Carlotta Domeniconi, & Jian Hu* 5

An Ensemble of Classifiers with Genetic Algorithm Based Feature Selection . . . . . *Zili Zhang & Pengyi Yang* 18

A Reliable Basis for Approximate Association Rules. . . . . *Yue Xu,Yuefeng Li & Gavin Shaw* 25

Parimputation: From Imputation and Null-Imputation to Partially Imputation . . . . . *Shichao Zhang* 32

Book Review

Computational Methods of Feature Selection . . . . . *Longbing Cao & David Taniar* 39

Announcements

Related Conferences, Call For Papers/Participants 41

On-line version: http://www.comp.hkbu.edu.hk/~iib (ISSN 1727-6004)

The Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society deals with tools and systems using biologically and linguistically motivated computational paradigms such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality. If you are a member of the IEEE Computer Society, you may join the TCII without cost. Just fill out the form at http://computer.org/tcsignup/.

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published once a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCII Executive Committee

2) Feature Articles

3) R&D Profiles (R&D organizations, interview profile on individuals, and projects etc.)

4) Book Reviews

5) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. William K. Cheung; Email:william@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997(printed)1727-6004(on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing—Google(www.google.com), The ResearchIndex(citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and **DBLP** Computer Science Bibliography (www.informatik.uni-trier.de/»ley/db/index.html).*

Intelligent Systems in Nanjing University

Intelligent systems is a major research theme in Nanjing University, with the support from the State Key Laboratory for Novel Software Technology of China, one of the top laboratories in the information technology field in the whole country. Currently, the research carried out by the intelligent systems group at Nanjing University mainly focuses on the following topics:

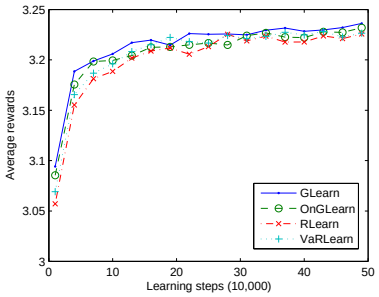
- Fundamental methods of intelligent computing, particularly reinforcement learning, incremental learning, granular computing and rough sets.
- Intelligent agents and multi-agent systems.
- Content-based multimedia (images and 3D models) retrieval.

The intelligent systems group aims to design intelligent algorithms and systems to deal with real problems efficiently. The group has developed software tools and applications covering wide areas, including medical treatment, public security, and virtual games.

I. RESEARCH

- Reinforcement learning

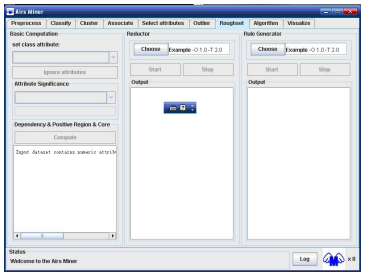
Reinforcement learning is an online, incremental learning technology, by which intelligent agents interact with the surrounding world by trial-and-error, and learn the optimal policy of decision sequences according to reinforcement signals. Our group has studied various algorithms for reinforcement learning problems, including average reward reinforcement learning, multi-agent reinforcement learning, relational reinforcement learning, function approximation in reinforcement learning and option discovery, etc. For example, our G-learning algorithm addresses the average reward domain, and is more stable than the classical R-learning and Q-learning. The following figure shows four curves of the average rewards computed in every 10,000 steps in the access-control queuing task. As shown in this figure, G-learning outperforms R-learning and its variation in the learning speed.



Besides the theoretical algorithm research, the group also studies the application of reinforcement learning. The proposed algorithms have been applied to video game tasks such as Tetris. The group attended the RL 2008 competition and won the 6th place in the game of Tetris. Research in learning classifier systems is related to reinforcement learning as well. Our group has successfully applied the learning classifier systems technology to different problem domains including data mining, medical data analysis and image steganography detection.

- Rough set

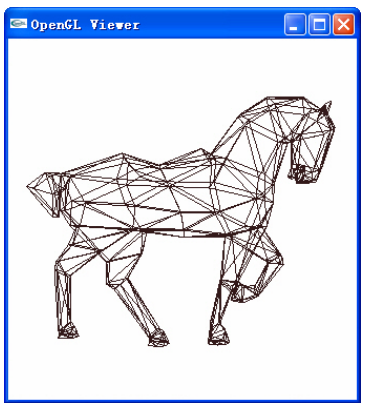
Rough set theory is a sound mathematical tool to deal with imprecise, uncertain, and vague information. Most of the group's work in this area can be characterized by rough-set-based hybrid approaches in classification and features selection. For incomplete information systems, rule generation by the GDT(General Distribution Table) approach and a default rule extracting method were proposed. Incremental algorithms are important, as data sources are increasingly in quantity. The group has investigated the incremental updating algorithms for core computing in the dominance-based rough set model. For face recognition, our group applies multiple applications of a reduction process based on approximation quality, and achieves fewer attributes. For natural language processing, we explore a RoughTree classifier with rough set and semi-naive Bayesian hybrid in decision tree representation, and conduct a dependency parsing task on Chinese corpus embedded with Nivre's deterministic parsing algorithm. The results show that RoughTree has better performance on a dependency parsing task. Combining rough set with other data mining algorithms, the group has developed a software tool, AirsMiner1.1, for data analysis.

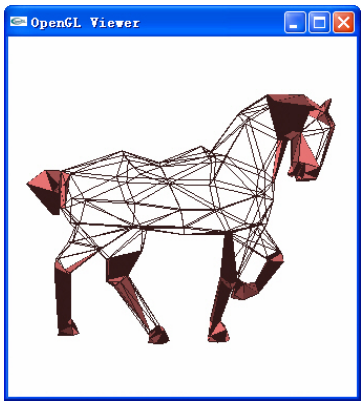


- Content-based 3D models retrieval

With the development of computer graphics and the progress in multimedia hardware technologies, 3D models are increasingly demanded in many application fields. The content-based 3D model retrieval researchers in this group has been exploring this challenging research field since 2004. The group proposes a novel retrieval performance metric, GSSS(Get Score from Similarity Sequence), which has been proved better than the widely used Precision-Recall(PR) curve benchmark on most shape-based 3D model retrieval cases.

One of the key issues in Content-Based 3D Model Retrieval is to define appropriate feature descriptors. To address this issue, we have developed a novel 3D feature extraction algorithm combining global and local characteristics, by which a 3D model is first segmented into several meaningful parts, and then representative 3D feature descriptors are extracted from those parts.

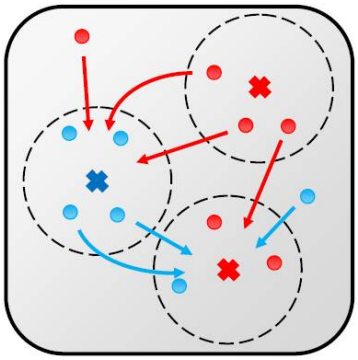




The group has also carried out research in semantic-based retrieval and made significant progress. A small-scale semantic network has been constructed and used to retrieve text-based annotations of 3-D models. The experimental results prove that these annotations are more understandable for human beings than those generated solely based on visual features.

- Intelligent agent and multi-agent systems

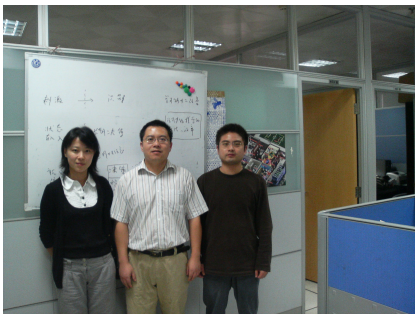
Agent technology is key for designing and developing distributed systems and adaptive software. The research conducted by the group includes developing multi-agent belief revision modeling, agent negotiation and multi-agent learning. We propose a multi-sentence batch belief revision model, a computational method, and a persuasive multi-agent multi-issue negotiation model. In multi-agent reinforcement learning, our group designed a meta-Q learning algorithm which overcomes the shortcomings of Nash Q-learning algorithm, and discovers the Pareto efficiency solutions as well.



The group applies agent technology to first-person shooting games. For example, we developed a RL-DOT (Reinforcement Learning-based DOmination Team) algorithm to learn the winning policy in UT's domination mode.

II. Education and Activities

The intelligent systems group has three tenured faculty members, Dr. Yang Gao(the leader of the group), Dr. Lin Shang and Dr. Yubin Yang. In the Department of Computer Science and Technology of Nanjing University, the group offers courses in Artificial Intelligence, Image Processing, Multimedia and Intelligent Agents. Currently, more than twenty graduate research students join the group and actively participate in our research projects. Within the past five years, eighteen MSc degrees and two PhD students have graduated from the group. The research of the intelligent systems group has led to a number of publications, most of which are published at top conferences and journals. Moreover, the group has won two Science and Technology Advancement Awards of Jiangsu Province, China in the past five years.



The group have also organized and hosted many research conferences held in Nanjing, China, such as:

- The 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), May, 2007.
- The 2nd Chinese Conference on Agent Theory and Applications(Agent'08), Apr. 2008.
- The 3rd, 4th, 5th and 6th Chinese Workshop on Machine Learning and Applications (MLA'05,MLA'06,MLA'07 and MLA'08).

In the department of computer science of Nanjing University, the group offers courses in artificial intelligence, image processing, multimedia and intelligent agents. Currently, more than twenty graduate students in the group actively participate in our research projects. Within the past five years, eighteen M.S. degrees and two Ph.D. degrees have been granted in the group.

III. Concluding Remarks

The research work of Intelligent Systems group has been funded by National Grand Fundamental Research 973 Program of China(No.2002CB312002, 2009CB32702), National Science Fund for Distinguished Young Scholars(No.60325207), National Natural Science Foundation of China(No.60775046, 60721002, 60503012, 60505008 and 60875011), Natural Science Foundation of Jiangsu Province, China(No.BK2007520) and Jiangsu Province Science and Technology Project(No. BE2006011). We have explored various interests in intelligent systems such as machine learning, multi-agent systems, content-based multimedia (images and 3D models) retrieval. To contact us, please use the following information.

Contact Information:
Department of Computer Science and Technology, Nanjing University,
No.22 Hankou Road, Nanjing
210093, China
gaoy@nju.edu.cn
shanglin@nju.edu.cn
yangyubin@nju.edu.cn
Website: ai.nju.edu.cn

Configuration and Recommender Systems: Two Converging Research Fields

BY MARKUS ZANKER[1,2] AND JUHA TIIHONEN[1]

Configuration (CS) and recommender systems (RS), two successful applications of AI techniques, have enjoyed wide-spread implementation for more than a decade. In addition to supporting sales-related functions, both have become immensely popular and active research fields in the context of Web-based commerce.

The history of knowledge-based configuration systems dates back to the first rule-based systems for ensuring the technical correctness of customer-defined orders for computer systems [2]. With the application of the Mass Customization paradigm to industries like automobiles, machinery, computers or furniture, configurators have been in widespread use ever since. Mittal and Frayman [4] defined configuration as a special type of design activity, with the key feature that the artifact being designed is assembled from a set of pre-defined components. Consequently, a configurator computes valid configurations, i.e. product instances, that conform to a given generic product structure and comply with a set of restrictions ensuring for instance compatibility, connectivity and customer requirements.

The field has continued to evolve since the late 1980s, exploring various higher-level knowledge representation mechanisms in order to enable shorter system development cycles, provide higher maintainability and more flexible reasoning. With the advent of Web-based commerce, configuration systems have had to satisfy new requirements such as online availability, ease-of-use or personalized interaction modes.

Tailoring the configuration process to the assumed informational needs and technical capabilities of the user or personalizing selection options based on past interaction logs has recently been identified as an avenue for further research in the field. These challenges connect configuration research with the field of recommender systems. RS help users to identify those items that will most probably interest them out of large sets of choices. One of the first application domains for a recommendation technique termed *collaborative filtering* was a personalized online news platform [5]. However, generating personalized recommendations for a user based on the opinions of peers with similar preferences quickly became popular in other online commerce domains such as movies, music or books. For instance, the online superstore *amazon.com* very successfully converts visitors into buyers by supporting their decision making with personalized recommendations. Since then a variety of additional recommendation approaches such as content-based, knowledge-based, utility-based and hybrid variants thereof have been explored [3]. More recent application domains with complex product items such as financial services or travel packages require the merging of knowledge based approaches with statistical learning methods like collaborative filtering. Thus, further interaction between both research fields seems inevitable in the near future.

This conference report will in the following outline the additional contents and discussions of both workshops.

I. ECAI 2008 - WORKSHOP ON CONFIGURATION SYSTEMS

The Workshop on Configuration Systems was the 11th in the series started at the AAAI'96 Fall Symposium (Cambridge, MA) and has continued in association with IJCAI, AAAI, and ECAI conferences since 1999. In addition to researchers from a variety of different fields, the events have attracted a significant number of industrial participants from major configurator vendors like SAP, Oracle, ILOG, and Tacton, as well as from end-users like Siemens, HP, or DaimlerChrysler.

Fig. 1. Rio-Antirrio bridge, a cable-stayed bridge crossing the Gulf of Corinth near Patras

The 2008 Patras workshop was a 1.5 day event that took place from July 21st to 22nd [6]. Its program consisted of nine technical papers and three invited talks. The invited talk of *Markus Stumptner* from University of South Australia titled "Reconfiguration from First Principles - with a fair bit of pragmatism in the mix" addressed the challenging topic of reconfiguration. Reconfiguration is used to modify an existing configuration to satisfy new requirements, usually with the goal of implementing minimal changes to existing individual products. Existing work was reviewed and new ideas were presented, providing practical solutions to existing and new application areas such as web service composition and responding to time-bounded changes in sales quotation processes. Research must continue in the future if languages capable of modeling general reconfiguration problems in commercial environments and inference systems based on general purpose reasoning mechanisms are to become a reality. The industrial invited talk of *Andreas Falkner* from Siemens on "Two Decades' Experience

[1]Organizer and Chair of the ECAI 2008 Workshop on Configuration Systems
[2]Organizer and Chair of the ECAI 2008 Workshop on Recommender Systems

in Developing Product Configurators" discussed the breadth of configuration problems in a large manufacturing company. For a long time Configurators have played a vital role in Siemens' business processes. Adopting new solutions and satisfying user requirements has in the meantime lead to the introduction of 6th generation of in-house developed configuration systems. A demonstration of the new S'UPREME configurator comprehensively illustrated the current state-of-the-art. Albert Haag from SAP gave the invited talk "What Makes Product Configuration Viable in a Business?". He reflected on the commercial and technical promises of The PLAKON system envisioned in 1985 and concluded that the original commercial expectations have not yet been met. Total cost of ownership, return on investment in configuration systems, technology gaps as well as integration hassles are still impediments for the pervasive deployment of configuration technology. Finally, exemplary business scenarios, commercial obstacles and an outlook on emerging trends were presented.

In addition, the workshop on configuration systems had three sessions, discussing technical papers on the following topics:

- Fundamentals: modeling and constraint based systems (4 papers)
- Personalization and Interactivity (3)
- Process Integration and Long-term management (2)

To summarize, presentations and discussions exhibited emerging trends and continuously active topics. It appears that application domains are being extended beyond traditional products to service industries and software configuration. Personalized interaction modes, long-term management of configurators and their knowledge bases, as well as reconfiguration and integration with other systems is becoming increasingly important.

II. ECAI 2008 - Workshop on Recommender Systems

The workshop continued the series of successful Workshops on Recommender Systems over the past decade, following in the footsteps of a similar one at ECAI 2006 or the Joint Workshop on Intelligent Techniques for Web Personalization and Recommender Systems at AAAI 2007 and 2008 to name only a few.

All submissions underwent a double-blind peer review process by at least three members of the international programme committee. The workshop was also held in conjunction with the 18th European Conference on Artificial Intelligence on July 22nd in Patras, Greece [7].

The workshop started with an invited talk titled "Revealing the Magic of Product Recommendation" by *Thomas Roth-Berghofer* who is Senior Researcher at the German Research Center for Artificial Intelligence (DFKI). He introduced the audience to the case-based recommendation paradigm that on the one hand exploits domain specific knowledge encoded as similarity functions on product items or cases, and on the other hand utilizes community knowledge by evaluating past system interactions. Due to their knowledge on item similarities, such systems are capable of explaining why a specific product was proposed to the user. Despite the fact that explanations for recommendations rarely appear in practical applications, they may help to stimulate users' trust in a system and make its function more transparent.

In addition the workshop consisted of three technical sessions dealing with the following topics:

- Social and Interactivity Aspects of Recommender Systems (4 papers)
- Algorithms and Security (4)
- Recommender Systems and Knowledge Management (3)

Accordingly, the workshop received a wide spectrum of technical contributions ranging from different recommendation scenarios like Web 2.0 or help-desk agents to algorithm improvements or attack strategies. The Netflix competition strongly stimulates research on algorithm improvements for collaborative filtering and subsequently the workshop included two technical papers presenting new strategies in the field. However, discussions at the workshop questioned if the current state-of-practice that evaluates algorithms' accuracy based on few historic datasets really captures their true performance and their effect on the user. As a result, a special issue on "Measuring the impact of personalization and recommendation on user behaviour" by the International Journal on Human-Computer Studies is now calling for contributions addressing this research question this autumn [1]. As online superstores diversify their product portfolios the potential of cross domain recommendations is an additional item for future research.

References

[1] CFP: Measuring the impact of personalization and recommendation on user behaviour. *International Journal on Human-Computer Studies*, 66(10):756, 2008.

[2] Virginia E. Barker, Dennis E. O'Connor, Judith Bachant, and Elliot Soloway. Expert systems for configuration at digital: Xcon and beyond. *Communications of the ACM*, 32(3):298–318, 1989.

[3] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[4] Sanjay Mittal and Felix Frayman. Toward a generic model of configuration tasks. In 11^{th} *International Joint Conferences on Artificial Intelligence*, pages 1395–1401, Menlo Park, California, 1989.

[5] P. Resnick, N. Iacovou, N. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Computer Supported Collaborative Work (CSCW)*, Chapel Hill, NC, 1994.

[6] Juha Tiihonen, Alexander Felfernig, Markus Zanker, and Tomi Maennistoe, editors. *Proceedings of the ECAI 2008 Workshop on Configuration Systems*. ECAI 2008 - University of Patras (ISBN 978-960-6843-01-3), Greece, 2008.

[7] Markus Zanker, Alexander Felfernig, and Robin Burke, editors. *Proceedings of the ECAI 2008 Workshop on Recommender Systems*. ECAI 2008 - University of Patras (ISBN 978-960-6843-06-8), Greece, 2008.

Markus Zanker is an assistant professor at the Department for Applied Informatics at the University of Klagenfurt and cofounder and director of ConfigWorks GmbH, a provider of interactive selling solutions. He received his MS and Ph.D. degree in Computer Science and MBA in business administration from Klagenfurt University. His research interests focus on knowledge-based systems, in particular in the fields of interactive sales applications such as product configuration and recommendation.

Juha Tiihonen is a researcher and project manager at the Department of Computer Science and Engineering of Helsinki University of Technology (HUT). He received his MS and Lic.Sc. degrees in Computer Science from HUT, his Ph.D. is being finalized. His main interest is product and service configuration in its various forms, including modeling and operations management aspects of business processes and design for configuration.

Cross-domain Text Classification using Wikipedia

Pu Wang, Carlotta Domeniconi, and Jian Hu

***Abstract*—Traditional approaches to document classification requires labeled data in order to construct reliable and accurate classifiers. Unfortunately, labeled data are seldom available, and often too expensive to obtain, especially for large domains and fast evolving scenarios. Given a learning task for which training data are not available, abundant labeled data may exist for a different but related domain. One would like to use the related labeled data as auxiliary information to accomplish the classification task in the target domain. Recently, the paradigm of transfer learning has been introduced to enable effective learning strategies when auxiliary data obey a different probability distribution.**

A co-clustering based classification algorithm has been previously proposed to tackle cross-domain text classification. In this work, we extend the idea underlying this approach by making the latent semantic relationship between the two domains explicit. This goal is achieved with the use of Wikipedia. As a result, the pathway that allows to propagate labels between the two domains not only captures common words, but also semantic concepts based on the content of documents. We empirically demonstrate the efficacy of our semantic-based approach to cross-domain classification using a variety of real data.

***Index Terms*—Text Classification, Wikipedia, Kernel methods, Transfer learning.**

I. Introduction

Document classification is a key task for many text mining applications. For example, the Internet is a vast repository of disparate information growing at an exponential rate. Efficient and effective document retrieval and classication systems are required to turn the massive amount of data into useful information, and eventually into knowledge. Unfortunately, traditional approaches to classification requires labeled data in order to construct reliable and accurate classifiers. Labeled data are seldom available, and often too expensive to obtain, especially for large domains and fast evolving scenarios. On the other hand, given a learning task for which training data are not available, abundant labeled data may exist for a different but related domain. One would like to use the related labeled data as auxiliary information to accomplish the classification task in the target domain. Traditional machine learning approaches cannot be applied directly, as they assume that training and testing data are drawn from the same underlying distribution. Recently, the paradigm of transfer learning has been introduced to enable effective learning strategies when auxiliary data obey a different probability distribution.

A co-clustering based classification algorithm has been proposed to tackle cross-domain text classification [17]. Let D_i be the collection of labeled auxiliary documents, called *in-domain* documents, and D_o be the set of (*out-of-domain*) documents to be classified (for which no labels are available). D_i and D_o may be drawn from different distributions. Nevertheless, since the two domains are related, e.g., baseball vs. hockey, effectively the conditional probability of a class label given a word is similar in the two domains. The method leverages the shared dictionary across the in-domain and the out-of-domain documents to propagate the label information from D_i to D_o. This is achieved by means of a two-step co-clustering procedure [17]. Specifically, it is assumed that class labels for D_i and D_o are drawn from the same set of class labels (for example, one class label may be "sport"; the documents in D_i are about baseball, and those in D_o are about hockey). Two co-clustering steps are carried out: one finds groups of documents and words for the out-of domain documents, and the other discovers groups of labels and words. In both cases, the set of words considered is the union of the terms appearing in D_i and D_o.

Thus, the words shared across the two domains allow the propagation of the class structure from the in-domain to the out-of-domain. Intuitively, if a word cluster $\hat{w}$ usually appears in class c in D_i, then, if a document $d \in D_o$ contains the same word clusters $\hat{w}$, it is likely that d belongs to class c as well.

The co-clustering approach in [17] (called CoCC) leverages the common words of D_i and D_o to bridge the gap between the two domains. The method is based on the "Bag of Words" (BOW) representation of documents, where each document is modeled as a vector with a dimension for each term of the dictionary containing all the words that appear in the corpus. In this work, we extend the idea underlying the CoCC algorithm by making the latent semantic relationship between the two domains explicit. This goal is achieved with the use of Wikipedia. By embedding background knowledge constructed from Wikipedia, we generate an enriched representation of documents, which is capable of keeping multi-word concepts unbroken, capturing the semantic closeness of synonyms, and performing word sense disambiguation for polysemous terms. By combining such enriched representation with the CoCC algorithm, we can perform cross-domain classification based on a *semantic bridge* between the two related domains. That is, the resulting pathway that allows to propagate labels from D_i to D_o not only captures common words, but also semantic concepts based on the content of documents. As a consequence, even if the two corpora share few words (e.g., synonyms are used to express similar concepts), our technique is able to bridge the gap by embedding semantic information in the extended representation of documents. As such, improved classification accuracy is expected, as also demonstrated in our experimental results.

In our previous work [31], a thesaurus was derived from Wikipedia, which explicitly defines synonymy, hyponymy and associative relations between concepts. Using the thesaurus

Pu Wang and Carlotta Domeniconi are with the Department of Computer Science, George Mason University, Fairfax, VA, 22030 USA e-mail: pwang7@gmu.edu, carlotta@cs.gmu.edu.

Jian Hu is with Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, P.R. China email: jianh@microsoft.com

constructed from Wikipedia, semantic information was embedded within the document representation, and the authors proved via experimentation that improved classification accuracy can be achieved [30]. In this work, we leverage these techniques to develop a semantic-based cross-domain classification approach.

The rest of the paper is organized as follows. In Section II, we discuss related work. In Section III, the background on co-clustering and the CoCC algorithm is covered. Section IV describes the structure of Wikipedia, and how we build a thesaurus from Wikipedia [31]. Section V presents the methodology to embed semantics into document representation, and Section VI describes our overall approach to cross-domain classification. In Section VII experiments are presented, and Section VIII provides conclusions and ideas for future work.

II. Related Work

In this section, we review background work in the areas of transfer learning, and text classification using encyclopedic knowledge.

A. Transfer learning

Cross-domain classification is related to transfer learning, where the knowledge acquired to accomplish a given task is used to tackle another learning task. In [28], the authors built a term covariance matrix using the auxiliary problem, to measure the co-occurrence between terms. The resulting term covariance is then applied to the target learning task. For instance, if the covariance between terms "moon" and "rocket" is high, and "moon" usually appears in documents of a certain category, it is inferred that "rocket" also supports the same category, even without observing this directly in the training data. The authors call their method Informative Priors.

In [21], the authors model the text classification problem using a linear function which takes the document vector representation as input, and provides in output the predicted label. Under this setting, different text classifiers differ only on the parameters of the linear function. A meta-learning method is introduced to learn how to tune the parameters. The technique uses data from a variety of related classification tasks to obtain a good classifier (i.e., a good parameter function) for new tasks, replacing hours of hand-tweaking.

In [19], Dai et al. modified the Naive Bayes classifier to handle a cross-domain classification task. The technique first estimates the model based on the distribution of the training data. Then, an EM algorithm is designed under the distribution of the test data. KL-divergence measures are used to represent the distribution distance between the training and test data. An empirical fitting function based on KL-divergence is used to estimate the trade-off parameters of the EM algorithm.

In [18], Dai et al. altered the Boosting algorithm to address cross-domain classification problems. Their basic idea is to select useful instances from auxiliary data with a different distribution, and use them as additional training data for predicting the labels of test data. However, in order to identify the most helpful additional training instances, the approach relies on the existence of some labeled testing data, which in practice may not be available.

B. Text classification using encyclopedic knowledge

Research has been done to exploit ontologies for content-based categorization of large corpora of documents. In particular, WordNet has been widely used. Siolas et al. [13] build a semantic kernel based on WordNet. Their approach can be viewed as an extension of the ordinary Euclidean metric. Jing et al. [10] define a term similarity matrix using WordNet to improve text clustering. Their approach only uses synonyms and hyponyms. It fails to handle polysemy, and breaks multi-word concepts into single terms. Hotho et al. [9] integrate WordNet knowledge into text clustering, and investigate word sense disambiguation strategies and feature weighting schema by considering the hyponymy relations derived from WordNet. Their experimental evaluation shows some improvement compared with the best baseline results. However, considering the restricted coverage of WordNet, the effect of word sense disambiguation is quite limited. The authors in [5], [14] successfully integrate the WordNet resource for document classification. They show improved classification results with respect to the Rocchio and Widrow-Hoff algorithms. Their approach, though, does not utilize hypernyms and associate terms (as we do with Wikipedia). Although [4] utilized WordNet synsets as features for document representation and subsequent clustering, the authors did not perform word sense disambiguation, and found that WordNet synsets actually decreased clustering performance.

Gabrilovich et al. [7], [8] propose a method to integrate text classification with Wikipedia. They first build an auxiliary text classifier that can match documents with the most relevant articles of Wikipedia, and then augment the BOW representation with new features which are the concepts (mainly the titles) represented by the relevant Wikipedia articles. They perform feature generation using a multi-resolution approach: features are generated for each document at the level of individual words, sentences, paragraphs, and finally the entire document. This feature generation procedure acts similarly to a retrieval process: it receives a text fragment (such as words, a sentence, a paragraph, or the whole document) as input, and then maps it to the most relevant Wikipedia articles. This method, however, only leverages text similarity between text fragments and Wikipedia articles, ignoring the abundant structural information within Wikipedia, e.g. internal links. The titles of the retrieved Wikipedia articles are treated as new features to enrich the representation of documents [7], [8]. The authors claim that their feature generation method implicitly performs words sense disambiguation: polysemous words within the context of a text fragment are mapped to the concepts which correspond to the sense shared by other context words. However, the processing effort is very high, since each document needs to be scanned many times. Furthermore, the feature generation procedure inevitably brings a lot of noise, because a specific text fragment contained in an article may not be relevant for its discrimination. Furthermore, implicit word sense disambiguation processing is not as effective as explicit disambiguation, as we perform in our approach.

In [16], Banerjee et al. tackled the daily classification

task (DCT) [22] by importing Wikipedia knowledge into documents. The method is quite straightforward: using Lucene (http://lucene.apache.org) to index all Wikipedia articles, each document is used as a query to retrieve the top 100 matching Wikipedia articles. The corresponding titles become new features. This technique is prone to bring a lot noise into documents. Similarly to [22], documents are further enriched by combining the results of the previous n daily classifiers with new testing data. By doing so, the authors claim that the combined classifier is at least no worse than the previous n classifiers. However, this method is based on the assumption that a category may be comprised of a union of (potentially undiscovered) subclasses or themes, and the class distribution of these subclasses may shift over time.

Milne et al. [25] build a professional, domain-specific thesaurus of agriculture from Wikipedia. Such thesaurus takes little advantage of the rich relations within Wikipedia articles. On the contrary, our approach relies on a general thesaurus, which supports the processing of documents concerning a variety of topics. We investigate a methodology that makes use of such thesaurus, to enable the integration of the rich semantic information of Wikipedia into a kernel.

III. Co-clustering

Clustering aims at organizing data in groups so that objects similar to each other are placed in the same group, or cluster. Co-clustering exploits the duality between objects and features, and simultaneously performs clustering along both dimensions. For example, for text mining applications, co-clustering discovers groups of documents and groups of words, thus leveraging the interplay between documents and words when defining similar documents.

The authors in [20] model the data contingency table as a joint probability distribution between two discrete random variables, and define an information-theoretic co-clustering algorithm that maps rows and columns to row-clusters and column-clusters, respectively. Optimality is defined in terms of mutual information between the clustered random variables. Formally, let X and Y be two discrete random variables that take values in the sets $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_n\}$, respectively, and let $p(X, Y)$ be their joint probability distribution. The goal is to simultaneously cluster X into k disjoint clusters, and Y into l disjoint clusters. Let $\{\hat{x}_1, \hat{x}_2, ..., \hat{x}_k\}$ be the k clusters of X, and $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_l\}$ the l clusters of Y. Then, the objective becomes finding mappings C_X and C_Y such that $C_X : \{x_1, ..., x_m\} \rightarrow \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_k\}$, $C_Y : \{y_1, ..., y_n\} \rightarrow \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_l\}$. The tuple (C_X, C_Y) represents a co-clustering.

We can measure the amount of information a random variable X can reveal about a random variable Y (and vice versa), by using the mutual information $I(X;Y)$, defined as follows:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

The quality of a co-clustering is measured by the loss in mutual information $I(X;Y) - I(\hat{X};\hat{Y})$ (subject to the constraints on the number of clusters k and l) [20]. The smaller the loss, the higher the quality of the co-clustering.

A. Co-clustering based Classification Algorithm (CoCC)

The authors in [17] use co-clustering to perform cross-domain text classification. Since our approach is based on their technique, we summarize here the CoCC algorithm [17].

Let D_i and D_o be the set of in-domain and out-of-domain data, respectively. Data in D_i are labeled, and $\mathcal{C}$ represents the set of class labels. The labels of D_o (unknown) are also drawn from $\mathcal{C}$. Let $\mathcal{W}$ be the dictionary of all the words in D_i and D_o. The goal of co-clustering D_o is to simultaneously cluster the documents D_o into $|\mathcal{C}|$ clusters, and the words $\mathcal{W}$ into k clusters. Let $\hat{\mathcal{D}}_o = \{\hat{d}_1, \hat{d}_2, ..., \hat{d}_{|\mathcal{C}|}\}$ be the $|\mathcal{C}|$ clusters of D_o, and $\hat{\mathcal{W}} = \{\hat{w}_1, \hat{w}_2, ..., \hat{w}_k\}$ the k clusters of $\mathcal{W}$. Following the notation in [20], the objective of co-clustering D_o is to find mappings C_{D_o} and $C_{\mathcal{W}}$ such that

$$C_{D_o} : \{d_1, ..., d_m\} \rightarrow \{\hat{d}_1, \hat{d}_2, ..., \hat{d}_{|\mathcal{C}|}\}$$

$$C_{\mathcal{W}} : \{w_1, ..., w_n\} \rightarrow \{\hat{w}_1, \hat{w}_2, ..., \hat{w}_k\}$$

where $|D_o| = m$ and $|\mathcal{W}| = n$. The tuple $(C_{D_o}, C_{\mathcal{W}})$, or $(\hat{\mathcal{D}}_o, \hat{\mathcal{W}})$, represents a co-clustering of D_o.

To compute $(\hat{\mathcal{D}}_o, \hat{\mathcal{W}})$, a two step procedure is introduced in [17], as illustrated in Figure 1 (the initialization step is discussed later). Step 1 clusters the out-of-domain documents into $|\mathcal{C}|$ document clusters according to the word clusters $\hat{\mathcal{W}}$. Step 2 groups the words into k clusters, according to class labels and out-of-domain document clusters simultaneously. The second step allows the propagation of class information from D_i to D_o, by leveraging word clusters. Word clusters, in fact, carry class information, namely the probability of a class given a word cluster. This process allows to fulfill the classification of out-of-domain documents.

As in [20], the quality of the co-clustering $(\hat{\mathcal{D}}_o, \hat{\mathcal{W}})$ is measured by the loss in mutual information

$$I(\mathcal{D}_o; \mathcal{W}) - I(\hat{\mathcal{D}}_o; \hat{\mathcal{W}}) \quad (2)$$

Thus, co-clustering aims at minimizing the loss in mutual information between documents and words, before and after the clustering process. Similarly, the quality of word clustering is measured by

$$I(\mathcal{C}; \mathcal{W}) - I(\mathcal{C}; \hat{\mathcal{W}}) \quad (3)$$

where the goal is to minimize the loss in mutual information between class labels $\mathcal{C}$ and words $\mathcal{W}$, before and after the clustering process.

By combining (2) and (3), the objective of co-clustering based classification becomes:

$$\min_{\hat{\mathcal{D}}_o, \hat{\mathcal{W}}} \{I(\mathcal{D}_o; \mathcal{W}) - I(\hat{\mathcal{D}}_o; \hat{\mathcal{W}}) + \lambda(I(\mathcal{C}; \mathcal{W}) - I(\mathcal{C}; \hat{\mathcal{W}}))\} \quad (4)$$

where λ is a trade-off parameter that balances the effect of the two clustering procedures. Equation (4) enables the classification of out-of-domain documents via co-clustering, where word clusters provide a walkway for labels to migrate from the in-domain to the out-of-domain documents.

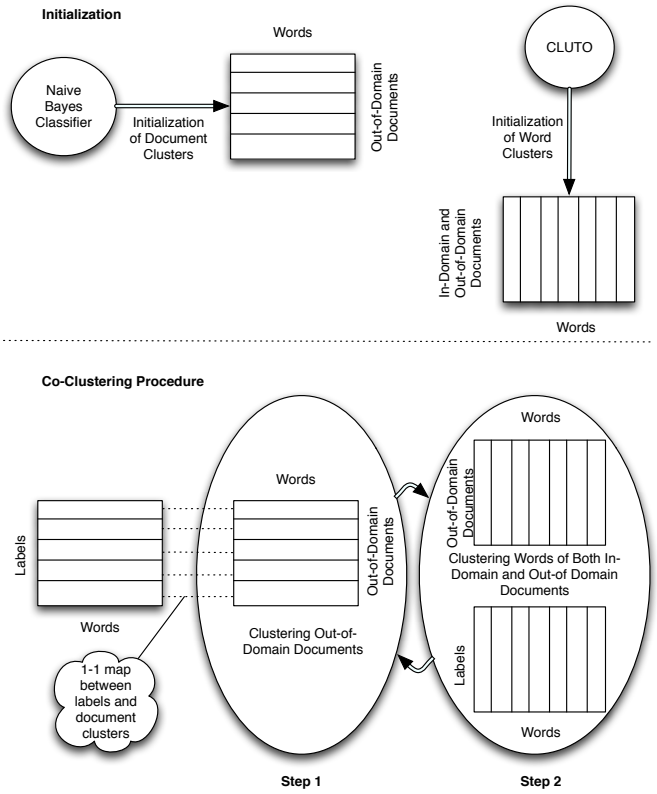


Fig. 1. Co-Clustering for Cross-domain Text Classification

To solve the optimization problem (4), the authors in [17] introduce an iterative procedure aimed at minimizing the divergence between distributions before and after clustering. To see this, lets first consider some definitions. $f(\mathcal{D}_o;\mathcal{W})$ represents the joint probability distribution of $\mathcal{D}_o$ and $\mathcal{W}$. $\hat{f}(\mathcal{D}_o;\mathcal{W})$ represents the joint probability distribution of $\mathcal{D}_o$ and $\mathcal{W}$ under co-clustering $(\hat{\mathcal{D}}_o;\hat{\mathcal{W}})$. Similarly, $g(\mathcal{C},\mathcal{W})$ denotes the joint probability distribution of $\mathcal{C}$ and $\mathcal{W}$, and $\hat{g}(\mathcal{C},\mathcal{W})$ denotes the joint probability distribution of $\mathcal{C}$ and $\mathcal{W}$ under the word clustering $\hat{\mathcal{W}}$. The marginal and conditional probability distributions can also be defined. In particular:

$$\hat{f}(d|\hat{w}) = \hat{f}(d|\hat{d})\hat{f}(\hat{d}|\hat{w}) = p(d|\hat{d})p(\hat{d}|\hat{w}) \quad (5)$$

$$\hat{f}(w|\hat{d}) = \hat{f}(w|\hat{w})\hat{f}(\hat{w}|\hat{d}) = p(w|\hat{w})p(\hat{w}|\hat{d}) \quad (6)$$

In [17], the following results are proven.

Lemma 1: For a fixed co-clustering $(\hat{\mathcal{D}}_o;\hat{\mathcal{W}})$, we can write the loss in mutual information as:

$$\begin{aligned} &I(\mathcal{D}_o;\mathcal{W}) - I(\hat{\mathcal{D}}_o;\hat{\mathcal{W}}) + \lambda I(\mathcal{C};\mathcal{W}) - I(\mathcal{C};\hat{\mathcal{W}}) \quad (7)\\ &= D(f(\mathcal{D}_o;\mathcal{W})||\hat{f}(\mathcal{D}_o;\mathcal{W})) + \lambda D(g(\mathcal{C},\mathcal{W})||\hat{g}(\mathcal{C},\mathcal{W})) \end{aligned}$$

where $D(\cdot||\cdot)$ is the KL-divergence defined as

$$D(p(x)||q(x)) = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

Lemma 2:

$$\begin{aligned} &D(f(\mathcal{D}_o,\mathcal{W})||\hat{f}(\mathcal{D}_o,\mathcal{W})) \\ &= \sum_{\hat{d}\in\hat{\mathcal{D}}_o}\sum_{d\in\hat{d}} f(d)D(f(\mathcal{W}|d)||\hat{f}(\mathcal{W}|\hat{d})) \quad (8)\\ &D(f(\mathcal{D}_o,\mathcal{W})||\hat{f}(\mathcal{D}_o,\mathcal{W})) \\ &= \sum_{\hat{w}\in\hat{\mathcal{W}}}\sum_{w\in\hat{w}} f(w)D(f(\mathcal{D}_o|w)||\hat{f}(\mathcal{D}_o|\hat{w})) \quad (9) \end{aligned}$$

Lemma 3:

$$\begin{aligned} &D(g(\mathcal{C},\mathcal{W})||\hat{g}(\mathcal{C},\mathcal{W})) \\ &= \sum_{\hat{w}\in\hat{\mathcal{W}}}\sum_{w\in\hat{w}} g(w)D(g((\mathcal{C}|w)||\hat{g}(\mathcal{C}|\hat{w}))) \quad (10) \end{aligned}$$

Lemma 1 states that to solve the optimization problem (4), we can minimize the KL-divergence between f and $\hat{f}$, and the KL-divergence between g and $\hat{g}$. Lemma 2 tells us that the minimization of $D(f(\mathcal{W}|d)||\hat{f}(\mathcal{W}|\hat{d}))$ for a single document d can reduce the value of the objective function of Equation (8). The same conclusion can be derived for the minimization of $D(f(\mathcal{D}_o|w)||\hat{f}(\mathcal{D}_o|\hat{w}))$ for a single word w. Similar conclusions can be derived from Lemma 3. Based on Lemmas 2 and 3, the approach described in Algorithm 1 computes a co-clustering $(\mathcal{C}_{\mathcal{D}_o},\mathcal{C}_{\mathcal{W}})$ that corresponds to a local minimum of the objective function given in Lemma 1 [17] .

Algorithm 1 The Co-clustering based Classification Algorithm (CoCC) [17]

1: **Input**: in-domain data D_i (labeled); out-of-domain data D_o (unlabeled); a set $\mathcal{C}$ of all class labels; a set $\mathcal{W}$ of all the word features; initial co-clustering $(\mathcal{C}^{(0)}_{\mathcal{D}_o},\mathcal{C}^{(0)}_{\mathcal{W}})$; the number of iterations T.
2: Initialize the joint distributions f, $\hat{f}$, g and $\hat{g}$
3: **for** $t \leftarrow 1,3,5,...,2T+1$ **do**
4: Compute the document clusters:

$$\mathcal{C}^{(t)}_{\mathcal{D}_o}(d) = \operatorname{argmin}_{\hat{d}} D(f(\mathcal{W}|d)||\hat{f}^{(t-1)}(\mathcal{W}|\hat{d})) \quad (11)$$

5: Update the probability distribution $\hat{f}^{(t)}$ based on $\mathcal{C}^{(t)}_{\mathcal{D}_o}$, $\mathcal{C}^{(t-1)}_{\mathcal{W}}$. $\mathcal{C}^{(t)}_{\mathcal{W}} = \mathcal{C}^{(t-1)}_{\mathcal{W}}$ and $\hat{g}^{(t)} = \hat{g}^{(t-1)}$.
6: Compute the word clusters:

$$\begin{aligned} \mathcal{C}^{(t+1)}_{\mathcal{W}}(d) = \operatorname{argmin}_{\hat{w}}\ & f(w)D(f(\mathcal{D}_o|w)||\hat{f}(\mathcal{D}_o|\hat{w})) \\ &+\lambda g(w)D(g((\mathcal{C}|w)||\hat{g}(\mathcal{C}|\hat{w}))) \quad (12) \end{aligned}$$

7: Update the probability distribution $\hat{g}^{(t+1)}$ based on $\mathcal{C}^{(t+1)}_{\mathcal{W}}$. $\mathcal{C}^{(t+1)}_{\mathcal{D}_o} = \mathcal{C}^{(t)}_{\mathcal{D}_o}$ and $\hat{f}^{(t+1)} = \hat{f}^{(t)}$.
8: **end for**
9: **Output**: The partition functions $\mathcal{C}^{(T)}_{\mathcal{D}_o}$ and $\mathcal{C}^{(T)}_{\mathcal{W}}$

The CoCC algorithm requires an initial co-clustering $(\mathcal{C}^{(0)}_{\mathcal{D}_o},\mathcal{C}^{(0)}_{\mathcal{W}})$ in input. As depicted in Figure 1, in [17] a Naive Bayes classifier is used to initialize the out-of-domain documents into clusters. The initial word clusters are generated using the CLUTO software [23] with default parameters. Once the co-clustering $(\mathcal{C}_{\mathcal{D}_o},\mathcal{C}_{\mathcal{W}})$ is computed by Algorithm 1,

the class of each document $d \in D_o$ is identified using the following [17]:

$$c = \arg\min_{c \in \mathcal{C}} D(\hat{g}(\mathcal{W}|c)||\hat{f}(\mathcal{W}|\hat{d}))$$

IV. Wikipedia as a Thesaurus

In the following sections, we present the methodology based on Wikipedia to embed semantics into document representation, and our overall approach to cross-domain classification. We start with a description of the fundamental features of the thesaurus built from Wikipedia [31].

Wikipedia (started in 2001) is today the largest encyclopedia in the world. Each article in Wikipedia describes a topic (or concept), and it has a short title, which is a well-formed phrase like a term in a conventional thesaurus [25]. Each article belongs to at least one category, and hyperlinks between articles capture their semantic relations, as defined in the international standard for thesauri [9]. Specifically, the represented semantic relations are: equivalence (*synonymy*), hierarchical (*hyponymy*), and associative.

Wikipedia contains only one article for any given concept (called *preferred term*). *Redirect* hyperlinks exist to group equivalent concepts with the preferred one. Figure 2 shows an example of a redirect link between the synonyms "puma" and "cougar". Besides synomyms, redirect links handle capitalizations, spelling variations, abbreviations, colloquialisms, and scientific terms. For example, "United States" is an entry with a large number of redirect pages: acronyms (U.S.A., U.S., USA, US); Spanish translations (Los Estados, Unidos, Estados Unidos); common mispellings (Untied States); and synonyms (Yankee land) [2].

Disambiguation pages are provided for an ambiguous (or polysemous) concept. A disambiguation page lists all possible meanings associated with the corresponding concept, where each meaning is discussed in an article. For example, the disambiguation page of the term "puma" lists 22 associated concepts, including animals, cars, and a sportswear brand.

Each article (or concept) in Wikipedia belongs to at least one category, and categories are nested in a hierarchical organization. Figure 2 shows a fragment of such structure. The resulting hierarchy is a directed acyclic graph, where multiple categorization schemes co-exist [25].

Associative hyperlinks exist between articles. Some are one-way links, others are two-way. They capture different degrees of relatedness. For example, a two-way link exists between the concepts "puma" and "cougar", and a one-way link connects "cougar" to "South America". While the first link captures a close relationship between the terms, the second one represents a much weaker relation. (Note that one-way links establishing strong connections also exist, e.g., from "Data Mining" to "Machine Learning".) Thus, meaningful measures need to be considered to properly rank associative links between articles. Three such measures have been introduced in [15]: *Content-based*, *Out-link category-based*, and *Distance-based*. We briefly describe them here. In Section V-B we use them to define the proximity between associative concepts.

The content-based measure is based on the bag-of-words representation of Wikipedia articles. Each article is modeled as a *tf-idf* vector; the associative relation between two articles is then measured by computing the cosine similarity between the corresponding vectors. Clearly, this measure (denoted as S_{BOW}) has the same limitations of the BOW approach.

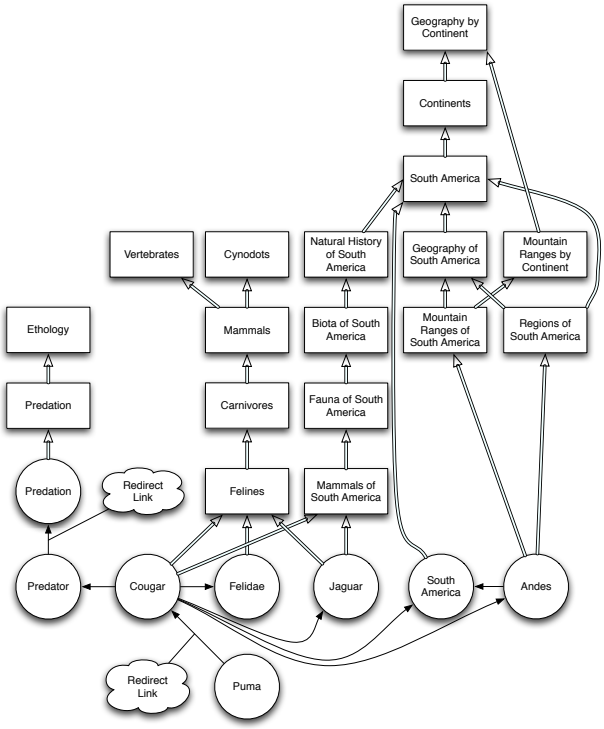


Fig. 2. A fragment of Wikipedia's taxonomy

The out-link category-based measure compares the out-link categories of two associative articles. The out-link categories of a given article are the categories to which out-link articles from the original one belong. Figure 3 shows (a fraction of) the out-link categories of the associative concepts "Data Mining", "Machine Learning", and "Computer Network". The concepts "Data Mining" and "Machine Learning" share 22 out-link categories; "Data Mining" and "Computer Network" share 10; "Machine Learning" and "Computer Network" share again the same 10 categories. The larger the number of shared categories, the stronger the associative relation between the articles. To capture this notion of similarity, articles are represented as vectors of out-link categories, where each component corresponds to a category, and the value of the i-th component is the number of out-link articles which belong to the i-th category. The cosine similarity is then computed between the resulting vectors, and denoted as S_{OLC}. The computation of S_{OLC} for the concepts illustrated in Figure 3 gives the following values, which indeed reflect the actual semantic of the corresponding terms: $S_{OLC}(\textit{Data Mining}, \textit{Machine Learning}) = 0.656$, $S_{OLC}(\textit{Data Mining}, \textit{Computer Network}) = 0.213$, $S_{OLC}(\textit{Machine Learning}, \textit{Computer Network}) = 0.157$.

The third measure is a distance measure (rather then a similarity measure like the first two). The distance between two articles is measured as the length of the shortest path connecting the two categories they belong to, in the acyclic graph of the category taxonomy. The distance measure is

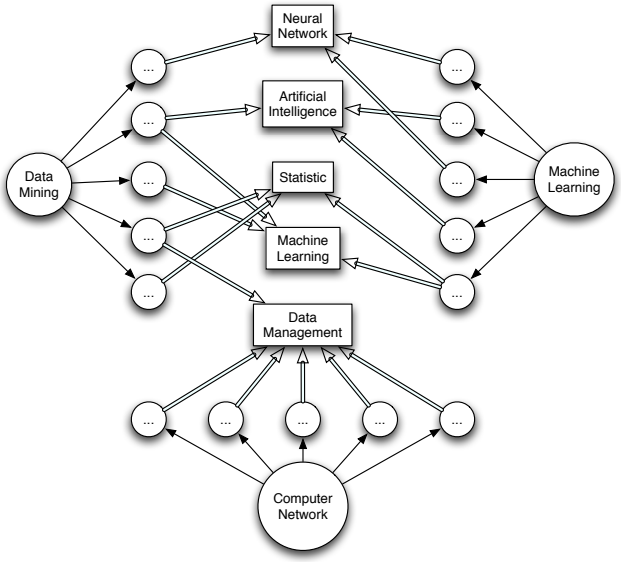


Fig. 3. Out-link categories of the concepts "Machine Learning", "Data Mining", and "Computer Network"

normalized by taking into account the depth of the taxonomy. It is denoted as D_{cat}.

A linear combination of the three measures allows to quantify the overall strength of an associative relation between concepts:

$$S_{overall} = \lambda_1 S_{BOW} + \lambda_2 S_{OLC} + (1-\lambda_1-\lambda_2)(1-D_{cat}) \quad (13)$$

where $\lambda_1, \lambda_2 \in (0,1)$ are parameters to weigh the individual measures. Equation (13) allows to rank all the associative articles linked to any given concept.

V. Concept-based Kernels

As mentioned before, the "Bag of Words" (BOW) approach breaks multi-word expressions, maps synonymous words into different components, and treats polysemous as one single component. Here, we overcome the shortages of the BOW approach by embedding background knowledge into a semantic kernel, which is then used to enrich the representation of documents.

In the following, we first describe how to enrich text documents with semantic kernels, and then illustrate our technique for building semantic kernels using background knowledge constructed from Wikipedia.

A. Kernel Methods for Text

The BOW model (also called Vector Space Model, or VSM) [29] of a document d is defined as follows:

$$\phi : d \mapsto \phi(d) = (tf(t_1,d), tf(t_2,d), \ldots, tf(t_D,d)) \in \mathcal{R}^D$$

where $tf(t_i,d)$ is the frequency of term t_i in document d, and D is the size of the dictionary.

The basic idea of kernel methods is to embed the data in a suitable feature space, such that solving the problem (e.g., classification or clustering) in the new space is easier (e.g., linear). A kernel represents the similarity between two objects (e.g., documents or terms), defined as dot-product in this new vector space. The kernel trick [12] allows to keep the mapping implicit. In other words, it is only required to know the inner products between the images of the data items in the original space. Therefore, defining a suitable kernel means finding a good representation of the data objects.

TABLE I
Example of document term vectors

	Puma	Cougar	Feline	...
d_1	2	0	0	...
d_2	0	1	0	...

In text classification, semantically similar documents should be mapped to nearby positions in feature space. In order to address the omission of semantic content of the words in VSM, a transformation of the document vector of the type $\tilde{\phi}(d) = \phi(d)S$ is required, where S is a semantic matrix. Different choices of the matrix S lead to different variants of VSM. Using this transformation, the corresponding vector space kernel takes the form

$$\begin{aligned} \tilde{k}(d_1,d_2) &= \phi(d_1)SS^\top\phi(d_2)^\top \\ &= \tilde{\phi}(d_1)\tilde{\phi}(d_2)^\top \end{aligned} \quad (14)$$

Thus, the inner product between two documents d_1 and d_2 in feature space can be computed efficiently directly from the original data items using a kernel function.

The semantic matrix S can be created as a composition of embeddings, which add refinements to the semantics of the representation. Therefore, S can be defined as:

$$S = RP \quad (15)$$

where R is a diagonal matrix containing the term weightings or relevance, and P is a *proximity matrix* defining the semantic similarities between the different terms of the corpus. One simple way of defining the term weighting matrix R is to use the inverse document frequency (*idf*).

P has non-zero off diagonal entries, $P_{ij} > 0$, when the term i is semantically related to the term j. Embedding P in the vector space kernel corresponds to representing a document as a less sparse vector, $\phi(d)P$, which has non-zero entries for all terms that are semantically similar to those present in document d. There are different methods for obtaining P [32], [1]. Here, we leverage the external knowledge provided by Wikipedia.

Given the thesaurus built from Wikipedia, it is straightforward to build a proximity (or similarity) matrix P. Here is a simple example. Suppose the corpus contains one document d_1 that talks about pumas (the animal). A second document d_2 discusses the life of cougars. d_1 contains instances of the word "puma", but no occurrences of "cougar". Vice versa, d_2 containts the word "cougar", but "puma" does not appear in d_2. Fragments of the BOW representations of d_1 and d_2 are given in Table I, where the feature values are term frequencies. The two vectors may not share any features (e.g., neither document contains the word "feline"). Table II shows a fragment of

TABLE II
EXAMPLE OF A PROXIMITY MATRIX

...	Puma	Cougar	Feline	...
Puma	1	1	0.4	...
Cougar	1	1	0.4	...
Feline	0.4	0.4	1	...
...				...

TABLE III
EXAMPLE OF "ENRICHED" TERM VECTORS

	Puma	Cougar	Feline	...
d_1'	2	2	0.8	...
d_2'	1	1	0.4	...

a proximity matrix computed from the thesaurus based on Wikipedia. The similarity between "puma" and "cougar" is one since the two terms are synonyms. The similarity between "puma" and "feline" (or "cougar" and "feline") is 0.4, as computed according to equation (13). Table III illustrates the updated term vectors of documents d_1 and d_2, obtained by multipling the original term vectors (Table I) with the proximity matrix of Table II. The new vectors are less sparse, with non-zero entries not only for terms included in the original document, but also for terms semantically related to those present in the document. This enriched representation brings documents which are semantically related closer to each other, and therefore it facilitates the categorization of documents based on their content. We now discuss the enrichment steps in detail.

B. Semantic Kernels derived from Wikipedia

The thesaurus derived from Wikipedia provides a list of concepts. For each document in a given corpus, we search for the Wikipedia concepts mentioned in the document. Such concepts are called *candidate concepts* for the corresponding document. When searching for candidate concepts, we adopt an exact matching strategy, by which only the concepts that explicitly appear in a document become the candidate concepts. (If an m-gram concept is contained in an n-gram concept (with $n > m$), only the last one becomes a candidate concept.) We then construct a vector representation of a document, which contains two parts: terms and candidate concepts. For example, consider the text fragment "Machine Learning, Statistical Learning, and Data Mining are related subjects". Table IV shows the traditional BOW term vector for this text fragment (after stemming), where feature values correspond to term frequencies. Table V shows the new vector representation, where boldface entries are candidate concepts, and non-boldface entries correspond to terms.

We observe that, for each document, if a word only appears in candidate concepts, it won't be chosen as a term feature any longer. For example, in the text fragment given above, the word "learning" only appears in the candidate concepts "Machine Learning" and "Statistical Learning". Therefore, it doesn't appear as a term in Table V. On the other hand, according to the traditional BOW approach, after stemming, the term "learn" becomes an entry of the term vector (Table IV).

TABLE IV
TRADITIONAL BOW TERM VECTOR

Entry	*tf*
machine	1
learn	2
statistic	1
data	1
mine	1
relate	1
subject	1

TABLE V
VECTOR OF CANDIDATE CONCEPTS AND TERMS

Entry	*tf*
machine learning	1
statistical learning	1
data mining	1
relate	1
subject	1

Furthermore, as illustrated in Table V, we keep each candidate concept as it is, without performing stemming or splitting multi-word expressions, since multi-word candidate concepts carry meanings that cannot be captured by the individual terms.

When generating the concept-based vector representation of documents, special care needs to be given to polysemous concepts, i.e., concepts that have multiple meanings. It is necessary to perform word sense disambiguation to find the specific meaning of ambiguos concepts within the corresponding document. For instance, the concept "puma" is an ambiguous one. If "puma" is mentioned in a document, its actual meaning in the document should be identified, i.e., whether it refers to a kind of animal, or to a sportswear brand, or to something else. In Section V-C we explain how we address this issue.

Once the candidate concepts have been identified, we use the Wikipedia thesaurus to select synonyms, hyponyms, and associative concepts of the candidate ones. The vector associated to a document d is then enriched to include such related concepts: $\phi(d) = (<$*terms*$>,<$*candidate concepts*$>,<$*related concepts*$>)$. The value of each component corresponds to a *tf-idf* value. The feature value associated to a related concept (which does not appear explicitely in any document of the corpus) is the *tf-idf* value of the corresponding candidate concept in the document. Note that this definition of $\phi(d)$ already embeds the matrix R as defined in equation (15).

We can now define a proximity matrix P for each pair of concepts (candidate and related). The matrix P is represented in Table VI. For mathematical convenience, we also include the terms in P. P is a symmetrical matrix whose elements are defined as follows. For any two terms t_i and t_j, $P_{ij} = 0$ if $i \neq j$; $P_{ij} = 1$ if $i = j$. For any term t_i and any concept c_j, $P_{ij} = 0$. For any two concepts c_i and c_j:

$$P_{ij} = \begin{cases} 1 & \text{if } c_i \text{ and } c_j \text{ are synonyms;} \\ \mu^{-depth} & \text{if } c_i \text{ and } c_j \text{ are hyponyms;} \\ S_{overall} & \text{if } c_i \text{ and } c_j \text{ are associative concepts;} \\ 0 & \text{otherwise.} \end{cases}$$

TABLE VI
PROXIMITY MATRIX

	Terms				Concepts			
Terms	1	0	$\cdots$	0	0	0	$\cdots$	0
	0	1	$\cdots$	0	0	0	$\cdots$	0
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	0	0	$\cdots$	1	0	0	$\cdots$	0
Concepts	0	0	$\cdots$	0	1	a	$\cdots$	b
	0	0	$\cdots$	0	a	1	$\cdots$	c
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	0	0	$\cdots$	0	b	c	$\cdots$	1

TABLE VII
COSINE SIMILARITY BETWEEN THE REUTERS DOCUMENT #9 AND THE WIKIPEDIA'S ARTICLES CORRESPONDING TO THE DIFFERENT MEANINGS OF THE TERM "STOCK"

Meanings of "Stock"	*Similarity with Reuters #9*
Stock (finance)	**0.2037**
Stock (food)	0.1977
Stock (cards)	0.1531
Stocks (plants)	0.1382
Stock (firearm)	0.0686
Livestock	0.0411
Inventory	0.0343

$S_{overall}$ is computed according to equation (13). *depth* represents the distance between the corresponding categories of two hyponym concepts in the category structure of Wikipedia. For example, suppose c_i belongs to category A and c_j to category B. If A is a direct subcategory of B, then $depth = 1$. If A is a direct subcategory of C, and C is a direct subcategory of B, then $depth = 2$. μ is a back-off factor, which regulates how fast the proximity between two concepts decreases as their category distance increases. (In our experiments, we set $\mu = 2$.)

By composing the vector $\phi(d)$ with the proximity matrix P, we obtain our extended vector space model for document d: $\tilde{\phi}(d) = \phi(d)P$. $\tilde{\phi}(d)$ is a less sparse vector with non-zero entries for all concepts that are semantically similar to those present in d. The strength of the value associated with a related concept depends on the number and frequency of occurrence of candidate concepts with a close meaning. An example of this effect can be observed in Table III. Let us assume that the concept "feline" is a related concept (i.e., did not appear originally in any of the given documents). "feline" appears in document d'_1 with strength 0.8, since the original document d_1 contains two occurrences of the synonym concept "puma" (see Table I), while it appears in d'_2 with a smaller strength (0.4), since the original document d_2 contains only one occurrence of the synonym concept "cougar" (see Table I). The overall process, from building the thesaurus from Wikipedia, to constructing the proximity matrix and enriching documents with concepts, is depicted in Figure 4.

C. Disambiguation of Concept Senses

If a candidate concept is polysemous, i.e. it has multiple meanings, it is necessary to perform word sense disambiguation to find its most proper meaning in the context where

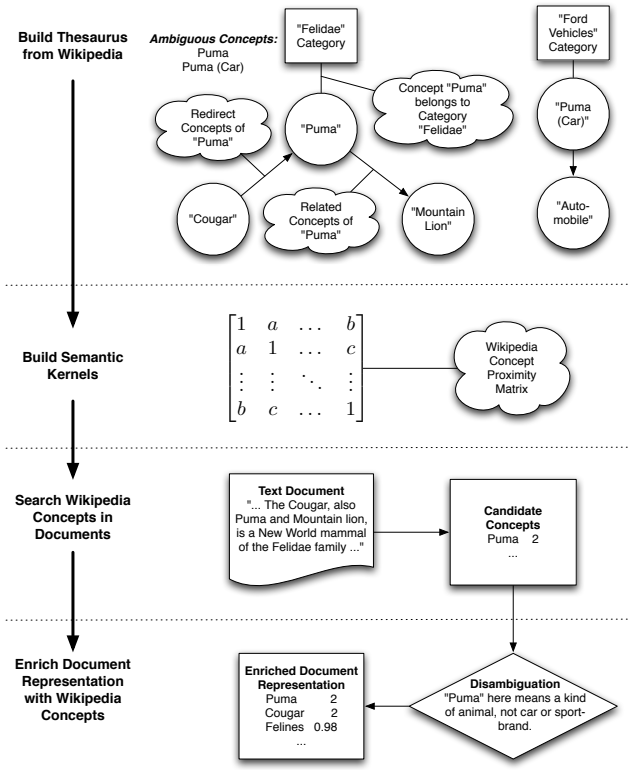


Fig. 4. The process that derives semantic kernels from Wikipedia

it appears, prior to calculating its proximity to other related concepts. We utilize text similarity to do explicit word sense disambiguation. This method computes document similarity by measuring the overlapping of terms. For instance, the Reuters-21578 document #9 [3] talks about stock splits, and the concept "stock" in Wikipedia refers to several different meanings, as listed in Table VII. The correct meaning of a polysemous concept is determined by comparing the cosine similarities between the *tf-idf* term vector of the text document (where the concept appears), and each of Wikipedia's articles (corresponding *tf-idf* vectors) describing the different meanings of the polysemous concept. The larger the cosine similarity between two *tf-idf* term vectors is, the higher the similarity between the two corresponding text documents. Thus, the meaning described by the article with the largest cosine similarity is considered to be the most appropriate one. From Table VII, the Wikipedia article describing "stock" (finance) has the largest similarity with the Reuters document #9, and this is indeed confirmed to be the case by manual examination of the document (document #9 belongs to the Reuters category "earn").

As mentioned above, document #9 discusses the stock split of a company, and belongs to the Reuters category "earn". The document contains several candidate concepts, such as "stock", "shareholder", and "board of directors". Table VIII gives an example of the corresponding related concepts identified by our method, and added to the vector representation of document #9 of the Reuters data set [30].

TABLE VIII
THE HYPONYM, ASSOCIATIVE, AND SYNONYM CONCEPTS INTRODUCED IN REUTERS DOCUMENT #9

Candidate Concepts	Hyponyms	Associative Concepts	Synonyms
Stock	Stock market Equity securities Corporate finance	House stock Bucket shop Treasury stock Stock exchange Market capitalization	Stock (finance)
Shareholder	Stock market	Board of directors Business organizations Corporation Fiduciary Stock	Shareholders
Board of directors	Business law Corporate governance Corporations law Management	Chief executive officer Shareholder Fiduciary Corporate governance Corporation	Boards of directors

VI. SEMANTIC-BASED CROSS-DOMAIN TEXT CLASSIFICATION

We apply the enriching procedure described in Section IV to all in-domain documents D_i and all out-of-domain documents D_o to perform cross-domain text classification. As a result, the representation of two related documents d_1 and d_2, such that $d_1 \in D_i$ and $d_2 \in D_o$, corresponds to two close vectors $\tilde{\phi}(d_1)$ and $\tilde{\phi}(d_2)$ in the extended vector space model. In other words, the extended vector space model applied to D_i and D_o has the effect of enriching the shared dictionary with concepts that encapsulate the content of documents. As such, related domains will have a shared pool of terms/concepts of increased size that has the effect of making explicit their semantic relationships.

We thus perform co-clustering based cross-domain classification by providing the CoCC algorithm (Algorithm 1) the extended vector space model of in-domain and out-of-domain documents. The set $\mathcal{W}$ now comprises the new dictionary, which includes terms and concepts (both candidate and related). We emphasize that concepts constitute individual features, without undergoing stemming, or splitting of multi-word expressions.

VII. EMPIRICAL EVALUATION

To evaluate the performance of our approach, we conducted several experiments using real data sets. We test scenarios for both binary and multiple category classification.

A. Processing Wikipedia XML data

The evaluation was performed using the Wikipedia XML Corpus [6]. The Wikipedia XML Corpus contains processed Wikipedia data parsed into an XML format. Each XML file corresponds to an article in Wikipedia, and maintains the original ID, title and content of the corresponding Wikipedia article. Furthermore, each XML file keeps track of the linked article ID, for every redirect link and hyperlink contained in the original Wikipedia article.

We do not include all concepts of Wikipedia in the thesaurus. Some concepts, such as "List of ISO standards", "1960s", and so on, do not contribute to the achievement of improved discrimination among documents. Thus, before building the thesaurus from Wikipedia, we remove concepts deemed not useful. To this end, we implement a few heuristics as explained below.

TABLE IX
NUMBER OF TERMS, CONCEPTS, AND LINKS AFTER FILTERING

Terms in Wikipedia XML corpus	**659,388**
Concept After Filtering	495,214
Redirected Concepts	413
Categories	113,484
Relations in Wikipedia XML corpus	**15,206,174**
Category to Subcategory	145,468
Category to Concept	1,447,347
Concept to Concept	13,613,359

First, all concepts of Wikipedia which belong to categories related to chronology, such as "Years", "Decades", and "Centuries", are removed. Second, we analyze the titles of Wikipedia articles to decide whether they correspond to useful concepts. In particular, we implement the following rules:

1) If the title of an article is a multi-word title, we check the capitalization of all the words other than prepositions, determiners, conjunctions, and negations. If all the words are capitalized, we keep the article.
2) If the title is one word title, and it occurs in the article more than three times [2], we keep the article.
3) Otherwise, the article is discarded.

After filtering Wikipedia concepts using these rules, we obtained about 500,000 concepts to be included in the thesaurus. Table IX provides a break down of the resulting number of elements (terms, concepts, and links) used to build the thesaurus, and therefore our semantic kernels. In particular, we note the limited number of redirected concepts (413). This is due to the fact that redirect links in Wikipedia often refers to the plural version of a concept, or to misspellings of a concept, and they are filtered out in the XML Corpus. Such variations of a concept, in fact, should not be added to the documents, as they would contribute only noise. For example, in Table VIII, the synonyms associated to the candidate concepts "Shareholder" and "Board of visitors" correspond to their plural versions. Thus, in practice they are not added to the documents.

B. Data Sets

We evaluated our approach using the 20 Newsgroups [11], and the SRAA [24] data sets. We split the original data in two corpora, corresponding to in-domain and out-of-domain documents. Different but related categories are selected for the two domains. Data sets across different classes are balanced.

20 Newsgroups. The 20 Newsgroups [11] data set is a popular collection of approximately 20,000 newsgroup documents, partitioned nearly evenly across 20 different newsgroups (about 1,000 per class).

We generated ten different data sets comprised of different combinations of categories. Each data set contains several top categories, which also define the class labels. Data are split into two domains based on their sub-categories. For example, one top category (i.e., class label) considered is "recreation"; the in-domain documents of this class talk about "autos" and "motorcycles", while the out-of-domain documents of the same class are concerned with "baseball" and "hockey" (they belong to the sub-category "sport"). This setting assures that documents in D_i and in D_o belong to different but related domains. Table X shows how categories were distributed for each data set generated from the 20 Newsgroups corpus. The setting of the six data sets for binary classification is the same as in [17].

SRAA. The SRAA [24] data set contains 73,218 articles from four discussion groups on simulated auto racing, simulated aviation, real autos, and real aviation. It is often used for binary classification, where the task can be defined as the separation of documents on "real" versus "simulated" topics, or as the separation of documents on "auto" vs. documents on "aviation". We generated two binary classification problems accordingly, as specified in Table XI.

TABLE X
SPLITTING OF 20 NEWSGROUPS CATEGORIES FOR CROSS-DOMAIN CLASSIFICATION

	Data Set	D_i	D_o
2 Categories	comp vs sci	comp.graphics comp.os.ms-windows.misc sci.crypt sci.electronics	comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x sci.med sci.space
	rec vs talk	rec.autos rec.motorcycles talk.politics.guns talk.politics.misc	rec.sport.baseball rec.sport.hockey talk.politics.mideast talk.religion.misc
	rec vs sci	rec.autos rec.sport.baseball sci.med sci.space	rec.motorcycles rec.sport.hockey sci.crypt sci.electronics
	sci vs talk	sci.electronics sci.med talk.politics.misc talk.religion.misc	sci.crypt sci.space talk.politics.guns talk.politics.mideast
	comp vs rec	rec.autos rec.sport.baseball comp.graphics comp.sys.ibm.pc.hardware comp.sys.mac.hardware	rec.motorcycles rec.sport.hockey comp.os.ms-windows.misc comp.windows.x
	comp vs talk	talk.politics.guns talk.politics.misc comp.graphics comp.sys.mac.hardware comp.windows.x	talk.politics.mideast talk.religion.misc comp.os.ms-windows.misc comp.sys.ibm.pc.hardware
3 Categories	rec vs sci vs comp	rec.motorcycles rec.sport.hockey sci.med sci.space comp.graphics comp.sys.ibm.pc.hardware comp.sys.mac.hardware	rec.autos rec.sport.baseball sci.crypt sci.electronics comp.os.ms-windows.misc comp.windows.x
	rec vs talk vs sci	rec.autos rec.motorcycles talk.politics.guns talk.politics.misc sci.med sci.space	rec.sport.baseball rec.sport.hockey talk.politics.mideast talk.religion.misc sci.crypt sci.electronics
	sci vs talk vs comp	sci.crypt sci.electronics talk.politics.mideast talk.religion.misc comp.graphics comp.sys.mac.hardware comp.windows.x	sci.space sci.med talk.politics.misc talk.politics.guns comp.os.ms-windows.misc comp.sys.ibm.pc.hardware
4 Categories	sci vs rec vs talk vs comp	sci.crypt sci.electronics rec.autos rec.motorcycles talk.politics.mideast talk.religion.misc comp.graphics comp.os.ms-windows.misc	sci.space sci.med rec.sport.baseball rec.sport.hockey talk.politics.misc talk.politics.guns comp.sys.mac.hardware comp.sys.ibm.pc.hardware comp.windows.x

TABLE XI
SPLITTING OF SRAA CATEGORIES FOR CROSS-DOMAIN CLASSIFICATION

Data Set	D_i	D_o
auto vs aviation	sim-auto & sim-aviation	real-auto & real-aviation
real vs simulated	real-aviation & sim-aviation	real-auto & sim-auto

C. Methods

In our experiments, we compare the classification results of the CoCC approach based on the BOW representation of documents, and of the CoCC approach based on the extended vector space model. We denote the first technique as CoCC *without enrichment*, and the second one as CoCC *with enrichment*.

The CoCC algorithm uses a Naive Bayes classifier to initialize the out-of-domain documents into clusters. Thus, we also report the results of the Naive Bayes classifiers, with and without enrichment, respectively.

D. Implementation Details

Standard pre-processing was performed on the raw data. Specifically, all letters in the text were converted to lower case, stop words were eliminated, and stemming was performed using the Porter algorithm [27] (candidate and related concepts, though, are identified prior to stemming, and kept unstemmed). Words that appeared in less than three documents were eliminated from consideration. Term Frequency (TF) was used for feature weighting when training the Naive Bayes classifier, and for the co-clustering based classification (CoCC) algorithm.

To compute the enriched representation of documents, we need to set the parameters λ_1 and λ_2 in Equation (13). These parameters were tuned according to the methodology suggested in [31]. As a result, the values $\lambda_1 = 0.4$ and $\lambda_2 = 0.5$ were used in our experiments.

The co-clustering based classification algorithm requires the initialization of document clusters and word clusters. As mentioned earlier, here we follow the methodology adopted in [17], and compute the initial document clusters using a Naive Bayes classifier, and the initial word clusters using the CLUTO software [23] with default parameters. The Naive

Bayes classifier is trained using D_i. The trained classifier is then used to predict the labels of documents in D_o. In our implementation, we keep track of class labels associated to clusters by the Naive Bayes classifier, to compute the final labels of documents in D_o.

The following implementation issue is worth a mention here. We observe that some words may only appear in D_i (or D_o). For such a word, and for a document $d \in D_o$ ($d \in D_i$, respectively), the estimation of $p(w|d)$ is zero. Furthermore, if all words w in a word cluster $\hat{w}$ only appear in D_i, since the CoCC algorithm only clusters documents $d \in D_o$, the estimation of $p(\hat{w}|\hat{d})$ becomes zero as well.

According to Equation (6), if $p(\hat{w}|\hat{d}) = 0$, then $\hat{f}(w|\hat{d})$ will also be zero. As a consequence, $D(f(W|d)||\hat{f}(W|\hat{d})) = \sum_{w \in W} f(w|d) \log \frac{f(w|d)}{\hat{f}(w|\hat{d})}$ becomes unbounded. In order to avoid this, in Equation (11), when $\hat{f}(w|\hat{d}) = 0$, Laplacian smoothing [26] is applied to estimate the probabilities. We proceed similarly for the computation of $D(f(D_o|w)||\hat{f}(D_o|\hat{w}))$ and $D(g(C|w)||\hat{g}(C|\hat{w}))$ in Equation (12).

E. Results

Table XII presents the precision rates obtained with Naive Bayes and the CoCC algorithm, both with and without enrichment, for all data sets considered. The results of the CoCC algorithm corresponds to $\lambda = 0.25$, and 128 word clusters. The precision values are those obtained after the fifth iteration. In the following, we study the sensitivity of our approach with respect to the number of iterations, the value of λ, and the number of clusters.

From Table XII, we can see that the CoCC algorithm with enrichment provides the best precision values for all data sets. For each data set, the improvement offered by CoCC with enrichment with respect to the Naive Bayes classifier (with enrichment), and with respect to CoCC without enrichment is quite significant. These results clearly demonstrate the efficacy of a semantic-based approach to cross-domain classification.

As shown in Table XII, the most difficult problem appears to be the one with four categories, derived from the 20 Newsgroups data set: rec vs talk vs sci vs comp. A closer look to the precision rates obtained for each category reveals that almost all documents of classes "recreation" and "talk" in D_o are correctly classified. The misclassification error is mostly due to the fact that the top categories "science" and "computers" are closely related to each other (in particular, the sub-category "electronics" of "science" may share many words with the category "computers"). As a consequence, several "science" documents are classified as "computers" documents. Nevertheless, CoCC with enrichment achieves 71.3% accuracy, offering a 8.9% improvement with respect to CoCC without enrichment, and a 17.5% improvement with respect to Naive Bayes. It is interesting to observe that in all cases the Naive Bayes classifier itself largely benefits from the enrichment process.

The authors in [17] have proven the convergence of the CoCC algorithm. Here, we show the precision achieved by CoCC with enrichment as a function of the number of iterations for the four multi-category problems considered in our

TABLE XII
CROSS-DOMAIN CLASSIFICATION PRECISION RATES

Data Set	w/o enrichment		w/ enrichment	
	NB	CoCC	NB	CoCC
rec vs talk	0.824	0.921	0.853	0.998
rec vs sci	0.809	0.954	0.828	0.984
comp vs talk	0.927	0.978	0.934	0.995
comp vs sci	0.552	0.898	0.673	0.987
comp vs rec	0.817	0.915	0.825	0.993
sci vs talk	0.804	0.947	0.877	0.988
rec vs sci vs comp	0.584	0.822	0.635	0.904
rec vs talk vs sci	0.687	0.881	0.739	0.979
sci vs talk vs comp	0.695	0.836	0.775	0.912
rec vs talk vs sci vs comp	0.487	0.624	0.538	0.713
real vs simulation	0.753	0.851	0.826	0.977
auto vs aviation	0.824	0.959	0.933	0.992

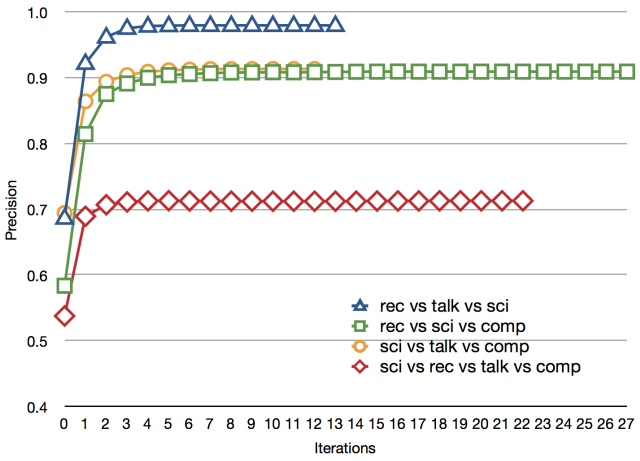


Fig. 5. CoCC with enrichment: Precision as a function of the number of iterations

experiments (see Figure 5). In each case, the algorithm reached convergence after a reasonable number of iterations (at most 27 iterations for the four data sets considered in Figure 5). The improvement in precision with respect to the initial clustering solution are confined within the first few iterations. During the subsequent iterations, the precision remains stable. We obtained a consistent result across all data sets. For this reason, in Table XII we provide the precision results obtained after the fifth iteration.

We also tested the sensitivity of CoCC with enrichment with respect to the λ parameter of Equation (4), and with respect to the number of clusters. We report the results obtained on the three category problem derived from the 20 Newsgroups data set: sci vs talk vs comp. Following the settings in [17], we used λ values in the range (0.03125, 8), with three different numbers of word clusters: 16, 64 and 128. Figure 6 shows the results. Overall, the precision values are quite stable. A reasonable range of values for λ is [0.25, 0.5].

The precision values as a function of different number of clusters are given in Figure 7. We tested different numbers of clusters between 2 and 512 for three different values of λ: 0.125, 0.25, and 1.0. As Figure 7 shows, the same trend was obtained for the three λ values. Precision increases significantly until a reasonable number of word clusters is achieved (too few word clusters do not allow discrimination

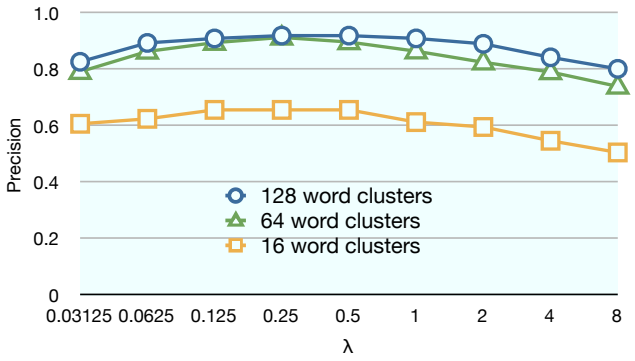


Fig. 6. CoCC with enrichment: Precision as a function of λ

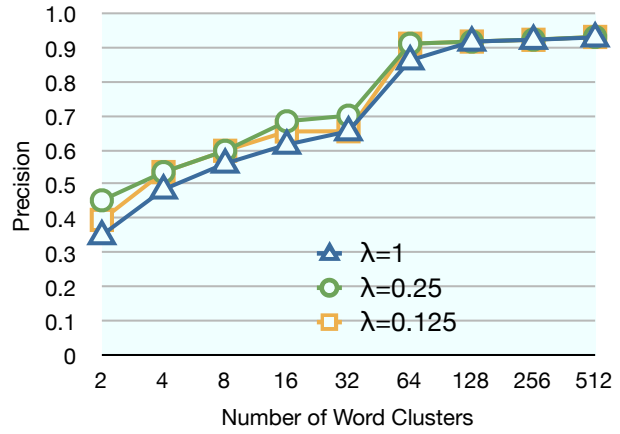


Fig. 7. CoCC with enrichment: Precision as a function of the number of word clusters

across classes). A value of 128 provided good results for all problems considered here (this finding is consistent with the analysis conducted in [17]).

VIII. Conclusions and Future Work

We extended the co-clustering approach to perform cross-domain classification by embedding background knowledge constructed from Wikipedia. In particular, we combine the CoCC algorithm with an enriched representation of documents, which allows to build a semantic bridge between related domains, and thus achieve high accuracy in cross-domain classification. The experimental results presented demonstrate the efficacy of a semantic-based approach to cross-domain classification.

The words shared between related domains play a key role to enable the migration of label information, and thus fulfill classification in the target domain. In our future work, we plan to explore alternate methodologies to leverage and organize the common language substrate of the given domains. We also plan to extend our approach to perform cross-language text classification, an interesting problem with difficult challenges.

Acknowledgment

This work was in part supported by NSF CAREER Award IIS-0447814.

References

[1] L. AlSumait and C. Domeniconi. Local Semantic Kernels for Text Document Clustering. In *Workshop on Text Mining, SIAM International Conference on Data Mining*, Minneapolis, MN, 2007. SIAM.

[2] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006.

[3] Carnegie Group, Inc. and Reuters, Ltd. *Reuters-21578 text categorization test collection*, 1997.

[4] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *International World Wide Web Conference*, Budapest, Hungary, 2003.

[5] M. de Buenega Rodriguez, J. M. Gomez-Hidalgo, and B. Diaz-Agudo. Using wordnet to complement training information in text categorization. In *International Conference on Recent Advances in Natural Language Processing*, 1997.

[6] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[7] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, 2005.

[8] E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In *National Conference on Artificial Intelligence (AAAI)*, Boston, Massachusetts, 2006.

[9] A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *Semantic Web Workshop, SIGIR Conference*, Toronto, Canada, 2003. ACM.

[10] L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang. Ontology-based distance measure for text clustering. In *Workshop on Text Mining, SIAM International Conference on Data Mining*, Bethesda, MD, 2006. SIAM.

[11] K. Lang. Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning*, Tahoe City, California, 1995. Morgan Kaufmann.

[12] J. Shawe-Taylor and N. Cristianini. *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[13] G. Siolas and F. d'Alché Buc. Support vector machines based on a semantic kernel for text categorization. In *International Joint Conference on Neural Networks (IJCNN'00)*, pages 205–209, Como, Italy, 2000. IEEE.

[14] L. A. Urena-Lopez, M. Buenaga, and J. M. Gomez. Integrating linguistic resources in TC through WSD. *Computers and the Humanities*, 35:215–230, 2001.

[15] P. Wang, J. Hu, H.-J. Zeng, L. Chen, and Z. Chen. Improving text classification by using encyclopedia knowledge. In *International Conference on Data Mining*, pages 332–341, Omaha, NE, 2007. IEEE.

[16] S. Banerjee. Boosting inductive transfer for text classification using wikipedia. In *International Conference on Machine Learning and Applications (ICMLA-2007)*, 2007.

[17] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Co-clustering based classification for out-of-domain documents. In *International Conference on Knowledge Discovery and Data Mining (KDD-2007)*, 2007.

[18] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *International Conference on Machine Learning (ICML-2007)*, 2007.

[19] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Transferring naive bayes classifiers for text classification. In *AAAI Conference on Artificial Intelligence (AAAI-2007)*, 2007.

[20] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, 2003.

[21] C. Do and A. Y. Ng. Transfer learning for text classification. In *Annual Conference on Neural Information Processing Systems (NIPS-2005)*, 2005.

[22] G. Forman. Tackling concept drift by temporal inductive transfer. In *Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-2006)*, 2006.

[23] G. Karypis. Cluto software for clustering high-dimensional datasets.

[24] A. K. McCallum Simulated/real/aviation/auto usenet data.

[25] D. Milne, O. Medelyan, and I. H. Witten. Mining domain-specific thesauri from wikipedia: A case study. In *International Conference on Web Intelligence*, 2006.

[26] T. M. Mitchell. Machine learning. In *McGraw Hill*, 1997.
[27] M. F. Porter An algorithm for suffix stripping *Program*, 14(3): 130–137, 1980.
[28] R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *International Conference on Machine Learning (ICML-2006)*, 2006.
[29] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
[30] P. Wang and C. Domeniconi. Building semantic kernels for text classification using wikipedia. In *Submit to International Conference on Knowledge Discovery and Data Mining (KDD-2008)*, 2008.
[31] P. Wang, J. Hu, H.-J. Zeng, L. Chen, and Z. Chen. Improving text classification by using encyclopedia knowledge. In *International Conference on Data Mining (ICDM-2007)*, 2007.
[32] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector space model in information retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 1985.

Pu Wang received a B.E. degree from Beihang University, Beijing, China, in 2004, and an M.S. degree from Peking University, Beijing, China, in 2007. From 2005 to 2007, he was an intern in the Machine Learning Group at Microsoft Research Asia, Beijing, China. He is currently a Ph.D. student in the Department of Computer Science at George Mason University, USA. His research interests focus on machine learning and data mining.

Carlotta Domeniconi received the Laurea degree in computer science from the University of Milano, Milan, Italy, in 1992, the M.S. degree in information and communication technologies from the International Institute for Advanced Scientic Studies, Salerno, Italy, in 1997, and the Ph.D. degree in computer science from the University of California, Riverside, in 2002. She is currently an Associate Professor in the Department of Computer Science, George Mason University, Fairfax, VA. Her research interests include machine learning, pattern recognition, data mining, and feature relevance estimation, with applications in text mining and bioinformatics. Dr. Domeniconi is a recipient of a 2004 Ralph E. Powe Junior Faculty Enhancement Award, and an NSF CAREER Award.

Jian Hu is currently an Assistant Researcher at Microsoft Research Asia, Beijing, China. He received Master's and Bachelor degrees from the Department of Computer Science and Technology at Shanghai Jiao Tong University, in 2006 and 2003 respectively. His current research interests include information retrieval, natural language processing, and Web usage data mining.

An Ensemble of Classifiers with Genetic Algorithm Based Feature Selection

Zili Zhang and Pengyi Yang

***Abstract*—Different data classification algorithms have been developed and applied in various areas to analyze and extract valuable information and patterns from large datasets with noise and missing values. However, none of them could consistently perform well over all datasets. To this end, ensemble methods have been suggested as the promising measures. This paper proposes a novel hybrid algorithm, which is the combination of a multi-objective Genetic Algorithm (GA) and an ensemble classifier. While the ensemble classifier, which consists of a decision tree classifier, an Artificial Neural Network (ANN) classifier, and a Support Vector Machine (SVM) classifier, is used as the classification committee, the multi-objective Genetic Algorithm is employed as the feature selector to facilitate the ensemble classifier to improve the overall sample classification accuracy while also identifying the most important features in the dataset of interest. The proposed GA-Ensemble method is tested on three benchmark datasets, and compared with each individual classifier as well as the methods based on mutual information theory, bagging and boosting. The results suggest that this GA-Ensemble method outperform other algorithms in comparison, and be a useful method for classification and feature selection problems.**

***Index Terms*—Ensemble Classifiers, Multi-objective Genetic Algorithms, Decision Tree, Artificial Neural Networks, Support Vector Machines.**

I. Introduction

MACHINE learning algorithms have been widely used in various fields to analyze and extract valuable information and patterns from large datasets with noise and missing values [1], [2], [3], [4]. One fundamental task of those learning algorithms is sample classification which is heavily relied on feature selection or extraction.

Learning algorithms are usually divided into two different categories: supervised learning, unsupervised learning [5]. In this work, we will focus on supervised learning.

The learning algorithms used in the classification process are refereed as classifiers, and several types of classifiers have been developed including decision trees, various types of artificial neural networks (ANN), support vector machines (SVM), and so on. Each of these classifiers uses different learning strategies. A common method used in supervised learning to improve classification accuracy and decrease computation complexity is feature selection [6]. In many applications, feature selection is essential as it can also help to identify important and meaningful traits [7], [8], [9]. Many feature selection approaches are available [6], [8], [10], [11], [12], [13], which can be categorized as deterministic or stochastic feature selection. However, deterministic feature selection results in high-dimensional datasets are often local optimal, while stochastic feature selection results are usually unstable [11].

A growing body of studies indicates that every single learning strategy has its own shortcomings and none of them could consistently perform well over all datasets. To overcome the shortcomings of individual methods, ensemble methods have been suggested as the promising measures [1], [14], [15], [16], [17]. For instance, the empirical study of ensemble system for data classification by Chandra and Yao [18] suggest that the ensemble systems tend to achieve higher accuracy and generalize better than single method. An ensemble of classifiers is a collection of classifiers that individual decisions are combined typically by means of weighed or un-weighted voting [15]. Some applications of different ensemble methods in real world datasets have demonstrated their power [20], [21], [22], [23].

The necessary and sufficient condition for an ensemble classifier to outperform its individual members is that the combined classifiers are accurate and diverse [16], [24]. In addition, previous studies have illustrated that the key requirements to successful ensemble methods are:

- the individual classifiers used to form the ensemble must have error rates less than 0.5 when classifying data, and
- the errors of those are uncorrelated at least in some extent [15].

In our previous work, we explored different hybrid algorithms–the combination of GA with decision tree (GADT), the combination of GA with artificial neural network (GANN), and the combination of GA with support vector machine (GASVM). They are used to analyze microarray data and SNP genotype data [7], [8]. All three algorithms have been proved powerful in sample classification and trait related feature selection.

In this study, a novel hybrid algorithm is proposed, which is the combination of a multi-objective Genetic Algorithm and an ensemble classifier. While the ensemble classifier, which consists of a decision tree classifier, an Artificial Neural Network (ANN) classifier, and a Support Vector Machine (SVM) classifier, is used as the classification committee, the multi-objective Genetic Algorithm is employed as the Random Subspacing (RS) method and the feature selector to facilitate data classification. This GA-Ensemble algorithm is essentially the combination of our previous algorithms. Nevertheless, the

Zili Zhang is with School of Engineering and Information Technology, Deakin University, Geelong, Victoria, Australia, 3217.
E-mail: zzhang@deakin.edu.au
Pengyi Yang is with Intelligent Software and Software Engineering Laboratory, Southwest University, Chongqing, China, 400715.
E-mail: mikesilver7@msn.com

objective of the use of ensemble classifiers and the combination with multi-objective GA feature selector is to further improve the overall data classification accuracy and feature selection reproducibility.

Since each of the three classifiers uses its own yet different learning strategies to classify the data, a diversely aggregated ensemble classifier can be obtained given the effective integration method was employed. This ensemble method differ itself from bagging and boosting strategies [25] because the diversity between classifiers is inherent in the inductive algorithms themselves other than manipulating the training dataset. The classification results over three benchmark datasets [28] are compared to see if this GA-Ensemble algorithm outperforms the individual ones. Furthermore, the results are also compared with those obtained by methods based on mutual information theory [13] and those obtained with bagging and boosting of decision tree [26].

The rest of the paper is structured as follows: Section II outlines the GA-Ensemble algorithm. The GA feature selector and ensemble classifiers in the GA-Ensemble algorithm are detailed in Sections III and IV, respectively. Evaluation is presented in Section V. Section VI concludes the paper.

II. Outline of the GA-Ensemble Algorithm

The GA-Ensemble algorithm proposed in this study is the combination of a multi-objective GA and an ensemble classifier consisting of a decision tree classifier, a standard multiple layer proceptron back propagation ANN classifier, and a support vector machine (SVM) classifier. A multi-objective evolutionary algorithm (which is similar to multi-objective GA) has been firstly employed in ensemble classifier construction by Chandra and Yao [27]. However, different from their application which optimize the diversity and the accuracy of the base classifiers explicitly, we incorporate these two optimization goals implicitly. Figure 1. illustrates the structure of the proposed system.

The learning steps of this algorithm can be described as following:

1) Initially, the global multi-objective GA randomly creates a set of chromosomes representing various feature sets.
2) Using all chromosomes in the set as the inputs of the classifiers. After classifiers evaluate certain feature set, they return the evaluation accuracies of this set to GA. GA then calculates the mean score and the consensus of this feature set.
3) After the whole population has been evaluated, GA selects favorite chromosomes with high fitness scores.
4) The crossover and mutation operations are then conducted on selected chromosomes with a predefined p_c (probability of crossover) and p_m (probability of mutation), respectively; and the next generation begins.
5) Repeat steps 2-4 until terminating generation is reached and the final chromosomes are printed out as the near optimal set of features for classification.

III. The Multi-Objective GA Feature Selector

The proposed ensemble approach utilized three classifiers, each will assess data and features with their own learning strategies. Thus, a multi-objective GA is employed to balance their assessments and facilitate their diversity. The fitness function of this multi-objective GA is defined as follows:

$$fitness_1(s) = \frac{\sum_{j=1}^{n} accuracy_j(s)}{n} \quad (1)$$

$$fitness_2(s) = consensus(s) \quad (2)$$

$$fitness(s) = \frac{fitness_1(s) + fitness_2(s)}{2} \quad (3)$$

Where $accuracy_j(s)$ specify the classification accuracy of the jth classifier upon the sth feature subset, while $consensus(s)$ specify the classification accuracy using consensus upon the sth feature subset.

The first part of the fitness function tries to optimize the target feature set into a subset which has superior power on accurate sample classification with not only one specific classifier but the whole classification committee. This part of the function improves the generalization ability of the resulting feature set [22]. As to the second part of the fitness function, it tries to optimize the target features set into a superior set in producing high consensus classification. This part of the function promotes the selected features in creating diverse classifiers implicitly, which in turn leads to the high sample classification accuracy [19].

The use of GA in this algorithm is two-fold. On the one hand, GA works as a RS method and a feature selector to select and rank different features based on their importance. This is extremely useful when the features in the given dataset are large and redundant, while the number of the samples are small. With the help of GA, informative features can be selected and uninformative ones will be removed. Otherwise, the uninformative features will increase the complexity of computation and introduce noisy and redundant data to the process [29]. By doing so, over-fitting can also be avoided in some extent. Moreover, the selected features can be further studied to find their special association with data.

On the other hand, when analyzing large datasets, feature selection is critical in improving the classification accuracy of the classifiers. It is widely acknowledged that the classification accuracy of ANN and SVM is affected by the size of the datasets. This is especially phenomenal when the number of the features is large. Moreover, it is both hard and unnecessary to use all data features as the inputs [7], [8] because it not only adds more computational expenses but also decreases the classification power of classifiers. By using GA, one can scale down the number of the inputs while also maintain or improve the classification accuracy of ANN and SVM. The need of combining GA with decision tree lies in that the decision tree algorithm is deterministic and it always uses the highest ranked feature – the feature with highest gain value – to split the dataset every time. This results in only one tree being created and it may be a locally optimal classifier, while an alternative one with a different splitting point can perform better [30]. This shortcoming is also more severe when the number of

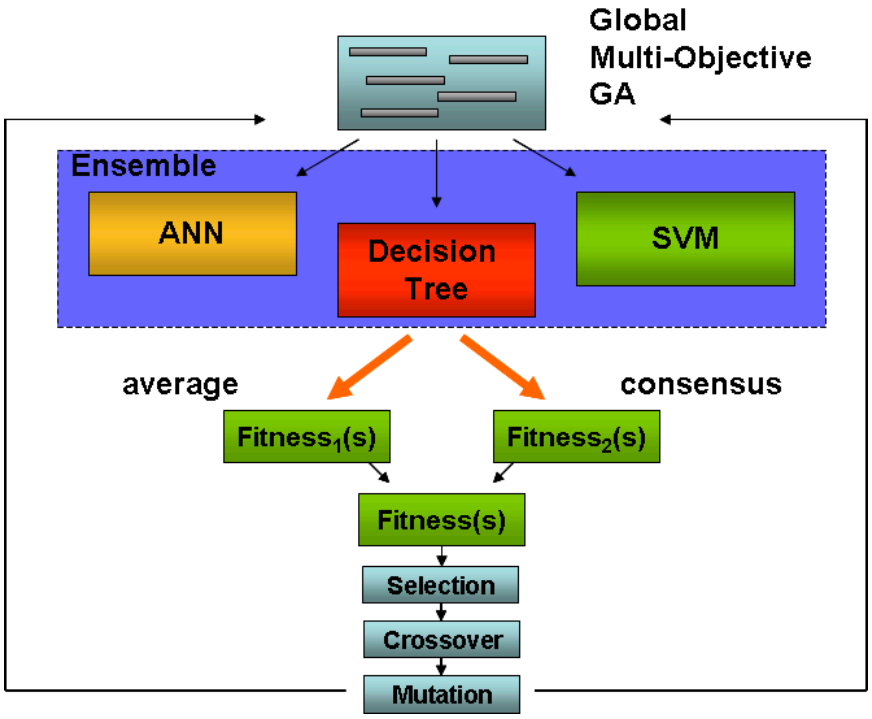


Fig. 1. The Architecture of the GA-Ensemble Algorithm

features been considered is large. When using GA to create different subsets, different decision trees can be produced and the favorite ones will be selected by GA for later iterations. This can help the decision tree to overcome the pitfall of the local optimal classification as well as identify important features.

IV. The Ensemble Classifier

Majority voting is one of the simplest strategies in implementing combination classifiers. Yet, the power of this strategy is comparable to other complex methods [31]. In n classifiers majority voting, consensus is made by k classifiers where

$$k = \begin{cases} n/2+1 & if\ n\ is\ even \\ (n+1)/2 & if\ n\ is\ odd \end{cases} \quad (4)$$

The three classifiers namely decision tree, ANN and SVM are integrated as a consensus committee. In the feature selection phase, each candidate feature combination produced by multi-objective GA will be fed into the ensemble classifier. After a feature combination has been input into the ensemble classifier, the three classifiers contained in this ensemble classifier will use the input features to learn and classify data sample, separately. When each classifier returns its classification accuracy by using certain feature combination, the multi-objective GA will calculate the consensus, using majority voting. Then the consensus score with the average score of three classifiers will be used as the fitness score of this feature combination.

In the evaluation phase, the best feature combination selected by GA-Ensemble is used to make sample classification with the test set. When a querying sample is input, three different classifiers will give their own prediction of which class this sample belongs to, and the majority voting is conducted to decide the final class it should be.

V. Evaluation

This section presents the experimental results we conducted.

A. Datasets

Three benchmark datasets, all obtained from UCI Repository [28], have been used to evaluate the proposed method. The first dataset is called Sonar dataset. The task is to discriminate between sonar signals bounced off a metal cylinder at different angles under various conditions and those bounced off a roughly cylindrical rock. The first class contains 111 samples and the second class contains 97 samples obtained from rocks under similar conditions. Each sample has 60 features representing the energy within a particular frequency band, integrated over a certain period of time [12]. The second dataset, named Ionosphere, contains 351 samples collected from radar signals, and 225 samples from it belong to class "good" while other 126 samples belong to class "bad". Each sample has 34 features. The last dataset is called Soybean (large) which contains 307 samples and 35 features. This dataset is different from the first two datasets in that the feature of the dataset is characterized as "categorical" instead of real numbers, and the number of the class is 19 instead of 2. Table 1 is the summary of the datasets used in evaluation.

All three datasets have long been utilized as evaluation datasets in many classification and feature selection studies [12], [13], [32], [33], [34] because they contain random noise, redundant features and the samples are linear inseparable.

TABLE I
DATASETS DESCRIPTION.

Dataset	Num. of Feature	Num. of Sample	Class Num.
Sonar [28]	60	208	2
Ionosphere [28]	34	351	2
Soybean [28]	35	683	19

B. GA-Ensemble Algorithm Implementation

The ANN adopted in this study is a three layers fully connected neural network. The number of the neuron in the first layer corresponds to the number of the input data features, and the number of the hidden neuron is the ceiling of the half of the input ones. Only one neuron is used in the output layer. The learning strategy of this ANN is consistent with that proposed by Brierley and Batty [35]. The learning rate from input layer to hidden layer is set to 0.4 and the learning rate 0.03 is set for the hidden layer to output layer. 1000 training epochs are used to train the ANN.

An easy to use yet very powerful SVM classifier package, SVMTorch II [36], is employed to construct SVM for the ensemble classifier. The default parameters are used when performing learning and classification. The kernel of the SVM is set as polynomial kernel with exponent of 2.

As for the decision tree, one of the most popular decision tree algorithm package C5.0, an improvement of C4.5 [37], is used to build decision tree and carry out simple classification.

Starting population size of GA is set to 100. The probability of crossover p_c and the probability of mutation p_m are 0.7 and 0.03, respectively. The single point mutation and crossover are used in genetic operation parts, and the binary tournament selection method [38] is adopted to select favorite gene combinations. The termination condition is that GA reaches the 50th generation, and the program terminates after the selected genes of the 50th generation are printed out.

C. Cross Validation

5-fold cross validation is conducted to evaluate the overall accuracy of all utilized methods. Both Sonar and Ionosphere datasets are randomly divided into five separate subsets, and while the four folds are used to train the algorithms, the remaining one fold is used to evaluate the classification accuracy of each method. The validation process repeats five times until all data in the sets are tested and the average classification accuracy is then calculated.

D. Z-score Calculation

In order to evaluate stability and reproducibility of each method, an independent re-run of every method is conducted. The feature combinations in last 10 generations of GA (after 40th generation) are extracted and the top 30 feature combinations with highest classification accuracy from the last 10 generations are then selected to compare in two independent runs to evaluate the stability and reproducibility.

A z-score or standard score is a measure of how many standard deviation units an individual raw score away from the mean of the distribution [39]. We employ this statistic method to calculate selection frequency of each feature. A high z-score of a feature indicates it's frequently selected. All evaluation results are z-score transformed as following [11]:

$$Z = [F_i - E(F_i)]/\sigma \tag{5}$$

where the $E(F_i)$ and σ are then calculated as following:

$$E(F_i) = P(feature_i) \cdot A \tag{6}$$

$$\sigma = \sqrt{P(feature_i) \cdot [1 - P(feature_i)] \cdot A} \tag{7}$$

In the above formulas, $A = 30$, which is the number of the top 30 feature combinations. $P(feature_i) = d/T$, where d is the feature combination length and T is the total feature number. F_i is the number of times $feature_i$ is selected.

E. Results

Previous studies show that the mean errors are relatively low and the classification accuracies are likely to be high when the combination size of the feature is small [12], [13]. Thus, in this study, feature sets with size of 6 and 12 for Sonar dataset and size of 5 and 10 for Ionosphere dataset as well as Soybean dataset are used to test our method. Firstly, each individual methods are tested separately, then the proposed GA-Ensemble approach is tested to compare with individual methods.

Tables 2 to 4 provide detail information of the results obtained with Sonar dataset, Ionosphere dataset and Soybean dataset for GADT, GANN, GASVM, and GA-Ensemble algorithms, respectively. All classification accuracies are calculated by averaging the 5-fold cross validation results with the best combination from each algorithm five times.

As shown in Tables 2, 3 and 4, GA-Ensemble method achieved the best classification accuracies, with 80.02% and 83.95% using 6-feature and 12-feature combinations in Sonar dataset, with 92.22% and 93.54% using 5-feature and 10-feature combinations in Ionosphere dataset, and with 94.97% and 95.37% using 5-feature and 10-feature combinations in Soybean dataset.

Table 5 provides the classification results obtained by using bagging and boosting of C4.5 algorithms. As can be seen, the results obtained by GA-Ensemble are comparable or better. The results of the first two datasets are also compared with those obtained with the method based on mutual information theory reported in [13]. In [13], several kinds of classification and feature selection methods are studied, which are all based on mutual information theory. The highest classification accuracies of Sonar dataset with 6 features and 12 features were obtained by 'TMFS with MIFS-U' and 'MIFS-U', with

TABLE II
CLASSIFICATION ACCURACY WITH SONAR DATASET

Methods	6 selected feature	12 selected feature
GADT	(F_{11} F_{25} F_{36} F_{38} F_{39} F_{45}) 79.72%	(F_{11} F_{14} F_{15} F_{21} F_{23} F_{33} F_{37} F_{42} F_{45} F_{47} F_{52} F_{60}) 80.63%
GANN	(F_{11} F_{17} F_{20} F_{27} F_{36} F_{46}) 78.61%	(F_1 F_4 F_7 F_{11} F_{12} F_{17} F_{20} F_{22} F_{49} F_{54} F_{57} F_{58}) 82.01%
GASVM	(F_9 F_{11} F_{12} F_{28} F_{37} F_{46}) 79.22%	(F_2 F_6 F_{11} F_{15} F_{20} F_{22} F_{31} F_{35} F_{36} F_{45} F_{46} F_{48}) 79.82%
GA-Ensemble	(F_4 F_9 F_{12} F_{36} F_{46} F_{48}) 80.02%	(F_3 F_{11} F_{16} F_{18} F_{19} F_{23} F_{32} F_{35} F_{37} F_{39} F_{45} F_{47}) 83.95%

TABLE III
CLASSIFICATION ACCURACY WITH IONOSPHERE DATASET

Methods	5 selected feature	10 selected feature
GADT	(F_3 F_4 F_5 F_{15} F_{27}) 91.47%	(F_2 F_3 F_4 F_5 F_6 F_8 F_{13} F_{23} F_{27} F_{32}) 90.90%
GANN	(F_1 F_5 F_{11} F_{25} F_{27}) 87.85%	(F_1 F_2 F_5 F_8 F_{18} F_{22} F_{24} F_{25} F_{27} F_{32}) 88.22%
GASVM	(F_1 F_5 F_8 F_{27} F_{29}) 91.45%	(F_2 F_3 F_4 F_5 F_6 F_7 F_{12} F_{16} F_{24} F_{27}) 93.16%
GA-Ensemble	(F_1 F_5 F_7 F_8 F_{27}) 92.22%	(F_1 F_5 F_8 F_9 F_{10} F_{20} F_{24} F_{26} F_{27} F_{32}) 93.54%

TABLE IV
CLASSIFICATION ACCURACY WITH SOYBEAN DATASET

Methods	5 selected feature	10 selected feature
GADT	(F_1 F_9 F_{15} F_{17} F_{35}) 92.18%	(F_1 F_{10} F_{12} F_{17} F_{18} F_{23} F_{29} F_{31} F_{32} F_{35}) 93.41%
GANN	(F_1 F_3 F_{15} F_{17} F_{32}) 94.65%	(F_3 F_5 F_{12} F_{15} F_{18} F_{19} F_{21} F_{22} F_{32} F_{35}) 94.85%
GASVM	(F_3 F_{15} F_{17} F_{26} F_{29}) 94.68%	(F_1 F_3 F_6 F_{14} F_{15} F_{22} F_{26} F_{29} F_{31} F_{35}) 94.29%
GA-Ensemble	(F_1 F_3 F_{17} F_{32} F_{35}) 94.97%	(F_1 F_3 F_{15} F_{17} F_{18} F_{22} F_{29} F_{31} F_{32} F_{35}) 95.37%

TABLE V
CLASSIFICATION ACCURACY WITH OTHER ENSEMBLE METHODS

Methods	Sonar Data	Ionosphere Data	Soybean
Bagging C4.5	75.48%	92.02%	92.83%
AdaBoosting C4.5	80.29%	91.74%	93.27%

79.31% and 81.51%, respectively. Those results are 0.5%-2% lower than those achieved by GA-Ensemble method. For Ionosphere datasets, the best classification accuracy is achieved by 'MIFS-U' ($\beta = 1.0$), which is 91.18% for 5 features and 92.02% for 10 features, respectively. For the proposed method, the classification accuracies are again 1%-2.5% better off.

Figure 2 illustrates the z-score of the features selected from Sonar dataset and Ionoshpere dataset, respectively. Two independent runs of each method are drawn on the same sub-graph to show the reproducibility and the stability. As can be seen from the diagram, the two independent runs using GA-Ensemble method are better overlapped compared with those using single classifier with single objective GA, in every case. These results demonstrate that GA-Ensemble method is comparatively more stable in feature selection. In addition, frequently selected features are calculated with high z-score. For Sonar dataset, F_{11}, F_{36}, F_{45} and F_{46} are the most frequently selected features. For Ionoshpere dataset, features F_1, F_3, F_5, F_7 and F_{27} are the favorite ones in the selected results. As for Soybean dataset, the favorite features are F_1, F_3, F_{15}, F_{17}, F_{32} and F_{35}. It is worth noting that the selection of the smaller feature sets are generally more stable than bigger ones, with GA selector.

VI. CONCLUSION

In this study, we demonstrated that the proposed GA-Ensemble method outperforms the GADT, GANN and GASVM algorithms in both classification accuracy and stability of feature selection with three benchmark datasets. The classification accuracy with GA-Ensemble method is also generally higher than that obtained by the methods based on mutual information theory, bagging and boosting of C4.5. The GA-Ensemble employs different classifiers to select features and use majority voting to make sample classification. The idea is that different classifiers will use their own learning

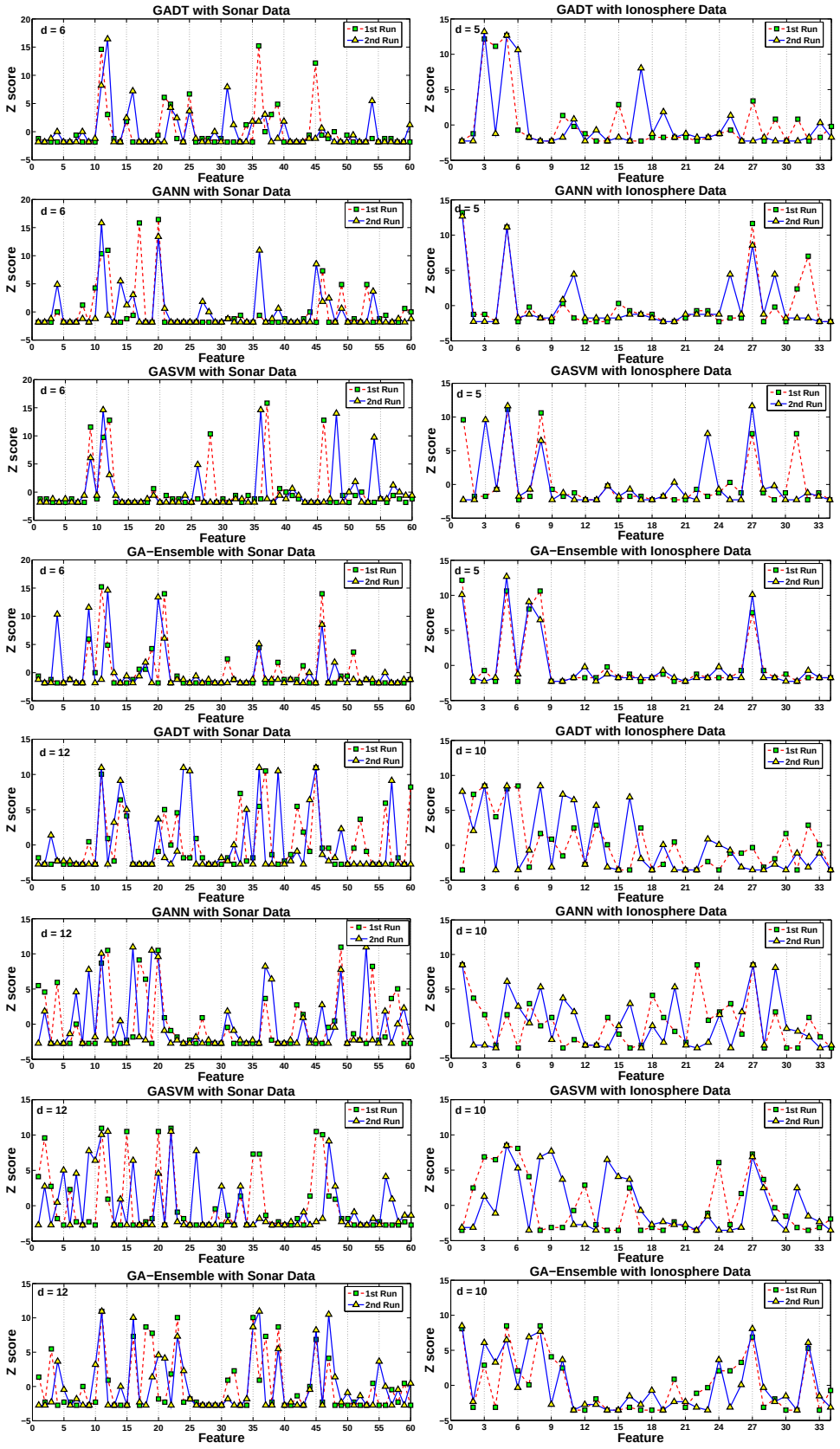


Fig. 2. Feature selection with Sonar dataset and Ionosphere dataset. Features selected by GADT, GANN, GASVM and GA-Ensemble methods. Z-score test is conducted to indicate the selected features. Note d is the feature combination length.

strategies to generate data classification hypothesis, and taking more hypotheses into consideration can improve data classi-

fication accuracy as well as generalization ability. The results suggest that the GA-Ensemble algorithm be a promising feature selection and sample classification algorithm.

REFERENCES

[1] L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their applications to handwritten numerals," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, pp. 418-435, 1992.

[2] C. Suen, C. Nadal, T. Mai, R. Legault, and L. Lam, "Computer recognition of unconstrained handwritten numerals," *Proc. IEEE*, vol. 80, pp. 1162-1180, 1992.

[3] A. Tan and D. Gilbert, "An empirical comparison of supervised machine learning techniques in bioinformatics," *Proceedings of the First Asia Pacific Bioinformatics Conference*, vol. 19, pp. 219-222, 2003.

[4] S. Cho and H. Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification," *Proceedings of the First Asia Pacific Bioinformatics Conference*, vol. 19, pp. 189-198, 2003.

[5] S. Theodoridis and K. Koutroumbas, *Pattern Recognition (Third Edition)*. Elsevier Press, 2006.

[6] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, pp. 1226-1238, 2005.

[7] P. Yang and Z. Zhang, "A Hybrid Approach to Selecting Susceptible Single Nucleotide Polymorphisms for Complex Disease Analysis," submitted to BMEI08 Conference.

[8] P. Yang and Z. Zhang, "Hybrid mehtods to select informative gene sets in microarray data classification," *Proceedings of AI 2007, LNAI 4830*, Springer, pp. 811-815, 2007.

[9] C. Ding and H. Peng, "Minimum Redundancy Feature Selection From Microarray Gene Expression Data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185-205, 2005.

[10] K. Srinivasa, K. Venugopal, and L. Patnaik, "Feature Extraction using Fuzzy C-Means Clustering for Data Mining Systems," *International Journal of Computer Science and Network Security*, vol. 6, no.3A, 2006.

[11] L. Li, C. Weinberg, T. Darden, and L. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, pp. 1131-1142, 2001.

[12] M. Skurichina and R. Duin, "Combining Feature Subsets in Feature Selection," *MCS 2005 LNCS 3541*, Springer, pp. 165-175, 2005.

[13] N. Kwak and C.-H. Choi, "Input Feautre Selection by Mutual Information Based on Parzen Windows," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667-1671, Dec. 2002.

[14] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search for ensemble feature selection," *Information Fusion*, vol. 6, pp. 83-98, 2005.

[15] T. Dietterich, "Ensemble Methods in Machine Learning," *Proceedings of the First International Workshop on MCS. LNCS*, Springer, vol. 1857, pp. 1-15, 2000.

[16] G. Giacinto and F. Roli, "An approach to the automatic design of multiple classifier systems," *Pattern Recognition Letters*, vol. 22, no. 1, pp. 25-33, 2001.

[17] D. Miller and L. Yan, "Critic-driven Ensemble Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 10, pp. 2833-2844, 1999.

[18] A. Chandra, and X. Yao, "Evolving hybrid ensembles of learning machines for better generalisation," *Neurocomputing*, vol. 69, pp. 686-700, 2006.

[19] D. Ruta and B. Gabrys, "Application of the Evolutionary algorithms for Classifier Selection in Multiple Classifier Systems with Majority Voting," *Proceedings of MCS 2001, LNCS 2096*, Springer, pp. 399-408, 2001.

[20] X. Li, S. Rao, Y. Wang, and B. Gong, "Gene Mining: A Novel and Powerful Ensemble Decision Approach to Hunting for Disease Genes Using Microarray Expression Profiling," *Nucleic Acids Research*, vol. 32, no. 9, pp. 2685-2694, 2004.

[21] S. Cho and P. Chanho, "Speciated GA for Optimal Ensemble Classifiers in DNA Microarray Classification," *Evolutionary Computation, 2004. CEC2004. Congress on.* vol. 1, pp. 590- 597, 2004.

[22] G. Bontempi, "A Blocking Strategy to Improve Gene Selection for Classification of Gene Expression Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 2, pp. 293-300, 2007.

[23] *Á*. Blanco, M. Martín-Merino, and J. de las Rivas, "Ensemble of Dissimilarity Based Classifiers for Cancerous Samples Classification," *PRIB 2007, LNBI 4774*, Springer, pp. 178-188, 2007.

[24] L. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Trans. Pattern Analysis and Machine Intell.*, vol. 12, pp. 993-1001, 1990.

[25] E. Bauer, R. Kohavi, "An empirical comparsion of voting classification algorithms: bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1-2, pp. 105-139, 1999.

[26] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139-157, 2000.

[27] A. Chandra, and X. Yao, "Ensemble learning using multi-objective evolutionary algorithm," *Journal of Mathematical Modelling and Algorithms*, vol. 5, no. 4, pp. 417-445, 2006.

[28] C. Blake and C. Merz, UCI repository of machine learning databases, http://www.ics.uci.edu/MLRepository.html, 2007.

[29] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and N. Yakhini, "Tissue Classification with Gene Expression Profiles," *Journal of Computational Biology*, vo. 7, pp. 559-584, 2000.

[30] E. Keedwell and A. Narayanan, *Intelligent Bioinformatics*. John Wiley & Sons, Ltd, pp. 159, 2005.

[31] L. Lam and Y. Suen, "Application of Majority Voting to Pattern Recognition: An Analysis of its Behaviour and Performance," *IEEE Transactions on Systems, Man, and Cybernetics* vol. 27, no. 5, pp. 553-568, 1997.

[32] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. on Neural Networks*, vol. 5, no. 4, pp. 537-550, July 1994.

[33] M. Tesmer and P. Estevez, "AMIFS: Adaptive Feature Selection by Using Mutual Information," *Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN2004, Budapest, Hungary*, July 26-29, pp. 303-308, 2004.

[34] M. Tan, and L. Eshelman, "Using weighted networks to represent classification knowledge in nosiy domains," *Proceedings of the Fifth International Conference on Machine learning*, pp. 121-134, 1988.

[35] P. Brierley and B. Batty, *Data Mining with Neural Networks - an Applied Example in Understanding Electricity Consumption Patterns.* Knowledge Discovery and Data Mining (ed Max Bramer). pp. 240-303, IEE, 1999.

[36] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.

[37] J. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.

[38] D. Goldberg and K. Deb, "A Comparative Analysis of Selection Schemes used in Genetic Algorithms," *Foundations of Genetic Algorithms*. San Mateo, CA: Morgan Kaufmann pp. 69-93, 1991.

[39] S.L. Jackson, *Research Methods and Statistics: A Critical Thinking Approach (2 edition)*. Wadsworth Publishing, 2005.

A Reliable Basis for Approximate Association Rules

Yue Xu, Yuefeng Li, Gavin Shaw

***Abstract*—For most of the work done in developing association rule mining, the primary focus has been on the efficiency of the approach and to a lesser extent the quality of the derived rules has been emphasized. Often for a dataset, a huge number of rules can be derived, but many of them can be redundant to other rules and thus are useless in practice. The extremely large number of rules makes it difficult for the end users to comprehend and therefore effectively use the discovered rules and thus significantly reduces the effectiveness of rule mining algorithms. If the extracted knowledge can't be effectively used in solving real world problems, the effort of extracting the knowledge is worth little. This is a serious problem but not yet solved satisfactorily. In this paper, we propose a concise representation called Reliable Approximate basis for representing non-redundant approximate association rules. We prove that the redundancy elimination based on the proposed basis does not reduce the belief to the extracted rules. We also prove that all approximate association rules can be deduced from the Reliable Approximate basis. Therefore the basis is a lossless representation of approximate association rules.**

***Index Terms*—Non-redundant association rule mining, approximate association rules, closed itemsets, certainty factor.**

I. Introduction

One big problem in association mining is the huge amount of the extracted rules which severely hinders the effective use of the discovered knowledge. Moreover, many of the extracted rules produce no value to the user or can be replaced by other rules thus considered redundant. Many efforts have been made on reducing the size of the extracted rule set. The approaches can be roughly divided to two categories, subjective approach and objective approach. In the subjective approach category, one technique is to define various interestingness measures and only the rules which are considered interesting based on the interesting measurements are generated[2], [3]. Another technique in this category is to apply constraints or templates to generate only those rules that satisfy the constraints or templates [1], [8], [11], [15]. In the objective approach category, the main technique is to construct concise representative bases of association rules without using user-dependent constraints. A concise representative basis contains much smaller number of rules and is considered lossless since all association rules can be derived from the basis. A number of concise representations of frequent patterns have been proposed, one of them, namely the closed itemsets, is of particular interest as they can be applied for generating non-redundant rules [9], [12], [19]. The notion of closed frequent itemset has its origins in the mathematical theory of Formal Concept Analysis introduced in the early 80s'[5], [16]. The use of frequent closed itemsets can greatly reduce the number of extracted rules and also provides a concise representation of association rules [13], [20]. Even though the number of extracted rules can be reduced by only using frequent closed itemsets, however, a considerable amount of redundancy still remains.

Yue Xu, Yuefeng Li and Gavin Shaw are with the Faculty of Information Technology, Queensland University of Technology, Brisbane, QLD4001, Australia (e-mail: {yue.xu,y2.li,gavin.shaw}@qut.edu.au).

Rules with confidence less than 1 are called Approximate rules and rules with confidence equal to 1 are called Exact rules. We have proposed a method to extract non-redundant exact rules [17]. In this paper, we present a concise representation basis called Reliable Approximate basis to extract non-redundant approximate rules. Most importantly, in this paper, we show that the redundancy elimination based on the proposed basis will not reduce the inference capacity of the extracted non-redundant rules. The certainty factor (CF) is an important and popular used measure of belief to inference rules [14]. We prove that the redundant rules eliminated by our approach have less or equal CF belief values than that of their corresponding non-redundant rules, and thus the elimination of such rules will not reduce the belief to the extracted rules. Moreover, we prove that all approximate association rules can be deduced from the reliable approximate basis, thus the reliable approximate basis is a lossless representation of approximate association rules.

The paper is organized as follows. Section II discusses redundancy in association rules and the elimination of the redundancy. Section III firstly introduces the proposed reliable basis for extracting non-redundant approximate rules, then presents a method to derive all approximate rules from the reliable basis. Experimental results are given in Section IV. Section V briefly discusses some related work. Finally, Section VI concludes the paper.

II. Redundancy and Redundancy Elimination

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct items, t be a transaction that contains a set of items such that $t \subseteq I$, T be a database containing different identifiable transactions. An association rule is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset I$ are sets of items called itemsets, and $X \bigcap Y = \emptyset$. Association rule mining is to find out association rules that satisfy the predefined minimum support (denoted as minsupp)and confidence (denoted as mincof) from a given database. The problem is usually decomposed into two subproblems: to find frequent itemsets and to generate association rules from those frequent itemsets. For the popular

TABLE I
ASSOCIATION RULES (MUSHROOM DATASET, MINSUPP=0.8, MINCONF=0.8)

	Rules (supp, conf)
1	gill-attachment-f ⇒ veil-type-p (0.97415,1.0)
2	veil-color-w ⇒ veil-type-p (0.97538 ,1.0)
3	gill-attachment-f,veil-color-w ⇒ veil-type-p (0.97317,1.0)
4	gill-attachment-f,ring-number-o ⇒ veil-type-p (0.89808,1.0)
5	gill-spacing-c,veil-color-w ⇒ veil-type-p (0.81487,1.0)
6	gill-attachment-f,gill-spacing-c ⇒ veil-type-p,veil-color-w (0.81265,1.0)
7	gill-attachment-f,gill-spacing-c⇒ veil-type-p (0.81265,1.0)
8	gill-attachment-f,gill-spacing-c,veil-type-p ⇒ veil-color-w (0.81265 ,1.0)
9	gill-attachment-f ⇒ veil-type-p,veil-color-w (0.97317,0.99899)
10	gill-attachment-f ⇒ veil-type-p,ring-number-o (0.89808,0.92191)
11	veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81487,0.83544)
12	veil-color-w ⇒ gill-attachment-f,gill-spacing-c,veil-type-p (0.81265,0.83317)
13	gill-attachment-f,veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81265,0.83506)
14	gill-attachment-f,veil-color-w ⇒ veil-type-p,ring-number-o (0.8971,0.92183)
15	gill-attachment-f,ring-number-o ⇒ veil-type-p,veil-color-w (0.8971,0.9989)
16	gill-spacing-c,veil-color-w ⇒ gill-attachment-f,veil-type-p (0.81265,0.99728)
17	gill-attachment-f ⇒ veil-color-w (0.97317,0.99899)
18	gill-attachment-f ⇒ ring-number-o (0.89808,0.92191)
19	gill-attachment-f,veil-color-w ⇒ gill-spacing-c (0.81265,0.83506)
20	gill-attachment-f,ring-number-o ⇒ veil-color-w (0.8971,0.9989)

TABLE II
CLOSED ITEMSETS AND MINIMAL GENERATORS (MUSHROOM DATASET, MINSUPP=0.8)

Closed itemsets	Minimal Generators	Support
{ veil-type-p }		1.0
{gill-attachment-f,veil-type-p}	{gill-attachment-f}	0.97415
{gill-spacing-c,veil-type-p}	{gill-spacing-c}	0.8385
{veil-type-p,veil-color-w}	{veil-color-w}	0.97538
{veil-type-p,ring-number-o}	{ring-number-o}	0.9217
{gill-attachment-f, veil-type-p,veil-color-w}	{gill-attachment-f, veil-color-w}	0.97317
{gill-attachment-f,veil-type-p, ring-number-o}	{gill-attachment-f, ring-number-o}	0.8981
{gill-spacing-c,veil-type-p, veil-color-w}	{gill-spacing-c, veil-color-w}	0.81487
{gill-attachment-f,gill-spacing-c, veil-type-p,veil-color-w}	{gill-attachment-f, gill-spacing-c}	0.81265
{gill-attachment-f,veil-type-p, veil-color-w,ring-number-o}	{veil-color-w, ring-number-o}	0.8971

used Mushroom dataset (http://kdd.ics.uci.edu/), with minimal support 0.8 and minimal confidence 0.8, we can generate 88 association rules, 20 of them are displayed in Table I.

The definition of closed itemsets comes from the closure operation of the Galois connection [5]. Let I denote the set of items and T denote the set of transactions, 2^I and 2^T are the power set of I and T, respectively. $\forall i \in I$ and $\forall t \in T$, if item i appears in transaction t, then i and t has a binary relation δ denoted as $i\delta t$. The Galois connection of the binary relation is defined by the following mappings where $X \subseteq I$, $Y \subseteq T$:

$$\tau : 2^I \rightarrow 2^T, \tau(X) = \{t \in T | \forall i \in X, i\delta t\} \quad (1)$$

$$\gamma : 2^T \rightarrow 2^I, \gamma(Y) = \{i \in I | \forall t \in Y, i\delta t\} \quad (2)$$

$\tau(X)$ is called the transaction mapping of X. $\gamma(Y)$ is called the item mapping of Y. $\gamma \circ \tau(X)$, called the closure of X, gives the common items among the transactions each of which contains X.

Definition 1: (Closed Itemsets) Let X be a subset of I. X is a closed itemset iff $\gamma \circ \tau(X) = X$.

Definition 2: (Generators) An itemset $g \in 2^I$ is a generator of a closed itemset $c \in 2^I$ iff $c = \gamma \circ \tau(g)$ and $g \subset \gamma \circ \tau(g)$. g is said a minimal generator of the closed itemset set c if $\nexists g' \subset g$ such that $\gamma \circ \tau(g') = c$.

For the Mushroom dataset, the closed itemsets and their minimal generators (minsupp=0.8) are given in Table II.

A challenge to association mining is the huge amount of the extracted rules. Recent studies have shown that using closed itemsets and generators to extract association rules can greatly reduce the number of extracted rules [13], [19]. However, considerable amount of redundancy still exists in the extracted association rules extracted based on closed itemsets. In this section, firstly some examples are given to show the existence of redundancy in the extracted rules, then we define the redundancy to be removed, and at the end of this section we prove that the elimination of the defined redundancy won't reduce the belief to the extracted non-redundant rules. In Section 3, we describe a concise representation of the defined non-redundant association rules, from which all approximate association rules can be derived.

A. Redundancy Definition

The rules in Table I are considered useful based on the predefined minimum support and confidence. However, some of the rules actually do not contribute new information. The consequent concluded by some rules can be obtained from some other rules without requiring more conditions but with higher or the same confidences. For example, in order to be fired the rules 5, 8, 13, and 20 in Table I require more conditions than that of rules 2, 6, 11, and 9, respectively, but conclude the same or less results which can be produced by rules 2, 6, 11, and 9. That means, without rules 5, 8, 13, and 20, we still can achieve the same result using other rules. Therefore, rules 5, 8, 13, and 20 are considered redundant to rules 2, 6, 11, and 9, respectively. Comparing to rules 2, 6, 11, and 9, the redundant rules 5, 8, 13, and 20 have a longer or the same antecedent and a shorter or the same consequent, respectively, and the confidence of the redundant rules is not larger than that of their corresponding non-redundant rules. The following definition defines such kind of redundant rules.

Definition 3: (Redundant rules) Let $X \Rightarrow Y$ and $X' \Rightarrow Y'$ be two association rules with confidence cf and cf', respectively. $X \Rightarrow Y$ is said a redundant rule to $X' \Rightarrow Y'$ if $X' \subseteq X$, $Y \subseteq Y'$, and $cf \leq cf'$.

Based on Definition 3, for an association rule $X \Rightarrow Y$, if there does not exist any other rule $X' \Rightarrow Y'$ such that the confidence of $X' \Rightarrow Y'$ is the same as or larger than the confidence of $X \Rightarrow Y$, $X' \subseteq X$ or $Y \subseteq Y'$, then $X \Rightarrow Y$ is non-redundant. Definition 3 is similar to Pasquier's definition of min-max association rules [13]. However, Pasquier's definition requires that a redundant rule and its corresponding non-redundant rule must have identical confidence and identical support, while Definition 3 here only requires that the confidence of the redundant rule is not larger than that of its corresponding non-redundant rule. In the following subsection, we prove that the requirement relaxation to redundancy will not reduce the belief to the extracted non-redundant rules.

B. Redundancy Elimination

The certainty factor theory were first introduced in MYCIN [14] to express how accurate and truthful a rule is and how

reliable the antecedent of the rule is. The certainty factor theory is based on two functions: measure of belief $MB(X,Y)$ and measure of disbelief $MD(X,Y)$ for a rule $X \Rightarrow Y$, as given below.

$$MB(X,Y) = \begin{cases} 1 & P(Y) = 1 \\ 0 & P(Y/X) \leq P(Y) \\ \frac{P(Y/X)-P(Y)}{1-P(Y)} & otherwise \end{cases} \quad (3)$$

$$MD(X,Y) = \begin{cases} 1 & P(Y) = 0 \\ 0 & P(Y/X) \geq P(Y) \\ \frac{P(Y)-P(Y/X)}{P(Y)} & otherwise \end{cases} \quad (4)$$

where, in the context of association rules, $P(Y/X)$ and $P(Y)$ are the confidence of the rule and the support of the consequent, respectively. The values of $MB(X,Y)$ and $MD(X,Y)$ range between 0 and 1 measuring the strength of belief or disbelief in consequent Y given antecedent X. $MB(X,Y)$ weighs how much the antecedent X increases the possibility of Y occurring. If the antecedent completely support the consequent, then $P(Y/X)$ will equal to 1 thus $MB(X,Y)$ will be 1. $MD(X,Y)$=1 indicates that the antecedent completely denies the consequent thus the disbelief in the rule reaches its highest value. The total strength of belief or disbelief in the association captured by the rule is measured by the certainty factor which is defined as follows:

$$CF(X,Y) = MB(X,Y) - MD(X,Y) \quad (5)$$

The value of a certainty factor is between 1 and -1. Negative values represent cases where the antecedent is against the consequent; positive values represent that the antecedent supports the consequent; while CF=0 means that the antecedent does not influence the belief to Y. Obviously, association rules with high CF values are more useful since they represent strong positive associations between antecedents and consequents. Indeed, the aim of association rule mining is to discover strong positive associations from large amount of data. Therefore, we propose that the certainty factors can be used to measure the strength of discovered association rules.

Theorem 1 below states that the CF value of a redundant rule defined by Definition 3 will never be larger than the CF value of its corresponding non-redundant rules. It means that, the association between the antecedent and consequent of the non-redundant rule is stronger than that of the redundant rule.

Theorem 1: Let $X \Rightarrow Y$ and $X' \Rightarrow Y'$ be two association rules. If $Y' \subseteq Y$, and $P(Y/X) \geq P(Y'/X')$, then $CF(X,Y) \geq CF(X',Y')$.

Proof: From Equation (5) we have
$CF(X,Y)$ - $CF(X',Y')$
$= MB(X,Y)$-$MB(X',Y')$+$MD(X',Y')$-$MD(X,Y)$

1) Assuming that $P(Y'/X') \geq P(Y')$. From condition $Y' \subseteq Y$, we have $P(Y) \leq P(Y')$. Because $P(Y/X) \geq P(Y'/X')$, so we have $P(Y/X) \geq P(Y)$. In this case, $MD(X',Y')$-$MD(X,Y)$ = 0. To prove the theorem, we need to prove that $MB(X,Y)$ - $MB(X',Y') \geq 0$. From Equation (3), we have:
$MB(X,Y)$ - $MB(X',Y')$ $= \frac{P(Y/X)-P(Y)}{1-P(Y)} - \frac{P(Y'/X')-P(Y')}{1-P(Y')}$
$= \frac{(P(Y/X)-P(Y))(1-P(Y'))-(P(Y'/X')-P(Y'))(1-P(Y))}{(1-P(Y))(1-P(Y'))}$
$= \frac{P(Y/X)-P(Y'/X')+P(Y'/X')P(Y)-P(Y/X)P(Y')-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$

Let $\alpha = P(Y/X) - P(Y'/X')$, the above expression becomes;
$= \frac{\alpha+P(Y'/X')P(Y)-(\alpha+P(Y'/X'))P(Y')-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$
$= \frac{\alpha+P(Y'/X')P(Y)-\alpha P(Y')-P(Y'/X')P(Y')-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$
$= \frac{\alpha(1-P(Y'))+P(Y'/X')(P(Y)-P(Y'))-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$
$= \frac{\alpha(1-P(Y'))+(P(Y')-P(Y))(1-P(Y'/X'))}{(1-P(Y))(1-P(Y'))}$

Because $P(Y) \leq P(Y')$ and $P(Y/X) \geq P(Y'/X')$ which makes $\alpha \geq 0$, we prove that the above expression ≥ 0. Hence, $MB(X,Y)$ - $MB(X',Y') \geq 0$

2) Assuming that $P(Y'/X') \leq P(Y')$. In this situation, we have two cases.
(i) $P(Y/X) \leq P(Y)$
In this case, $MB(X,Y)$-$MB(X',Y')$ = 0. To prove the theorem, we need to prove that $MD(X',Y')$-$MD(X,Y) \geq 0$. From Equation (4), we have
$MD(X',Y')$-$MD(X,Y)$ $= \frac{P(Y')-P(Y'/X')}{P(Y')} - \frac{P(Y)-P(Y/X)}{P(Y)}$
After expanding the above expression and eliminating identical dual terms, we have
$MD(X',Y')$-$MD(X,Y)$ $= \frac{P(Y/X)P(Y')-P(Y'/X')P(Y)}{P(Y)P(Y')}$.
$\geq \frac{P(Y/X)P(Y')-P(Y/X)P(Y)}{P(Y)P(Y')}$.
Again, since $P(Y) \leq P(Y')$, we get
$MD(X',Y')$-$MD(X,Y) \geq 0$.

(ii) $P(Y/X) \geq P(Y)$
In this case, $MD(X,Y)$=0 and $MB(X',Y')$ = 0. To prove the theorem, we need to prove that
$MD(X',Y')$+$MB(X,Y) \geq 0$. Because $P(Y'/X') \leq P(Y')$ and $P(Y/X) \geq P(Y)$, from the equations (3) and (4), it is true that
$MD(X',Y')$+$MB(X,Y) \geq 0$

Combining the results of the above cases, we have
$CF(X,Y)$ - $CF(X',Y') \geq 0$, hence
$CF(X,Y) \geq CF(X',Y')$

■

According to Theorem 1, the CF value of a redundant rule defined by Definition 3 is never higher than that of its corresponding non-redundant rule and thus the elimination of such redundant rules is reliable since it won't reduce the belief to the extracted non-redundant rules.

III. Concise Basis for Non-redundant Approximate Association Rules

Pasquier et al. [13] proposed a condensed basis to represent non-redundant approximate association rules, which is defined as follows:

Definition 4: (Min-max Approximate Basis) Let C be the set of frequent closed itemsets and G be the set of minimal generators of the frequent closed itemsets in C. The min-max approximate basis is:

$$MinMaxApprox = \{g \Rightarrow (c \backslash g) | c \in C, g \in G, \gamma \circ \tau(g) \subset c\}$$

For the 88 rules extracted from the Mushroom dataset mentioned above, there are 71 approximate rules. Based on the Min-max approximate basis, 25 approximate rules, as displayed in Table III, are extracted and considered non-redundant in terms of the redundancy definition given in [13]. However, under Definition 3, some of the rules extracted from the min-max approximate basis are redundant such as rules 22 to 25 which are redundant to rules 17, 11, 10, and 16, respectively.

TABLE III
NON-REDUNDANT APPROXIMATE RULES EXTRACTED FROM MIN-MAX APPROXIMATE BASIS (MUSHROOM DATASET, MINSUPP=0.8, MINCONF=0.8)

	Rules (supp, conf)
1	veil-type-p ⇒ gill-attachment-f (0.97415,0.97415)
2	veil-type-p ⇒ gill-spacing-c (0.8385 ,0.8385)
3	veil-type-p ⇒ veil-color-w (0.97538,0.97538)
4	veil-type-p ⇒ ring-number-o (0.92171,0.92171)
5	veil-type-p ⇒ gill-attachment-f,veil-color-w (0.97317,0.97317)
6	veil-type-p ⇒ gill-attachment-f, ring-number-o (0.89808,0.89808)
7	veil-type-p ⇒ gill-spacing-c, veil-color-w (0.81487,0.81487)
8	veil-type-p ⇒ gill-attachment-f,gill-spacing-c, veil-color-w (0.81265,0.81265)
9	veil-type-p ⇒ gill-attachment-f,veil-color-w, ring-number-o (0.8971,0.8971)
10	gill-attachment-f ⇒ veil-type-p, veil-color-w (0.97317, 0.99899)
11	gill-attachment-f ⇒ veil-type-p, ring-number-o (0.89808,0.92191)
12	gill-attachment-f ⇒ gill-spacing-c,veil-type-p, veil-color-w (0.81265,0.83422)
13	gill-attachment-f ⇒ veil-type-p,veil-color-w, ring-number-o (0.8971,0.9209)
14	gill-spacing-c ⇒ veil-type-p,veil-color-w (0.81487,0.97181)
15	gill-spacing-c ⇒ gill-attachment-f,veil-type-p, veil-color-w (0.81265,0.96917)
16	veil-color-w ⇒ gill-attachment-f, veil-type-p (0.97317, 0.99773)
17	veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81487,0.83544)
18	veil-color-w ⇒ gill-attachment-f,gill-spacing-c, veil-type-p(0.81265, 0.83317)
19	veil-color-w ⇒ gill-attachment-f,veil-type-p, ring-number-o (0.8971,0.91974)
20	ring-number-o ⇒ gill-attachment-f, veil-type-p (0.89808, 0.97436)
21	ring-number-o ⇒ gill-attachment-f,veil-type-p, veil-color-w (0.8971, 0.97329)
22	gill-attachment-f,veil-color-w ⇒ gill-spacing-c, veil-type-p (0.81265,0.83506)
23	gill-attachment-f,veil-color-w ⇒ veil-type-p, ring-number-o (0.8971, 0.92183)
24	gill-attachment-f,ring-number-o ⇒ veil-type-p, veil-color-w (0.8971,0.9989)
25	gill-spacing-c,veil-color-w ⇒ gill-attachment-f, veil-type-p (0.81265, 0.99728)

A. Reliable Approximate Basis

Corresponding to the Min-max approximate basis, we propose a more concise basis called Reliable Approximate basis as defined in Definition 5.

Definition 5: (Reliable Approximate Basis) Let C be the set of frequent closed itemsets and G be the set of minimal generators of the frequent closed itemsets in C. The Reliable approximate basis is:

$$ReliableApprox = \{g \Rightarrow (c\backslash g) | c \in C, g \in G, \gamma\circ\tau(g) \subset c, \neg(g \supseteq ((c\backslash c') \cup g')) \text{ or } conf(g \Rightarrow (c\backslash g)) > conf(g' \Rightarrow (c'\backslash g')) \; where \; c' \in C, g' \in G, g' \subset g, \gamma\circ\tau(g') \subset c'\}$$

The correctness of the above definition can be proved by the following theorems and properties.

Lemma 1: Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C, and $\gamma\circ\tau(g) \subset c$. If $\exists c' \in C, \exists g' \in G$, $\gamma\circ\tau(g') \subset c', g' \subset g, g \supseteq ((c\backslash c')\cup g')$, and $conf(g \Rightarrow c\backslash g) \leq conf(g' \Rightarrow c'\backslash g')$, then $g \Rightarrow c\backslash g$ is redundant to $g' \Rightarrow c'\backslash g'$.

Proof: Let $A = c\backslash c'$ so that $c \subseteq A \cup c'$ and $A \cap c' = \emptyset$. Therefore, we have $c\backslash((c\backslash c')\cup g') \subseteq (A\cup c')\backslash(A\cup g')$. From $\gamma\circ\tau(g') \subset c'$, we have $g' \subset c'$. Since $A \cap c' = \emptyset$, then $A \cap g' = \emptyset$. So, $c\backslash((c\backslash c')\cup g') \subseteq (A\cup c')\backslash(A\cup g') = ((A\cup c')\backslash A)\backslash g') = c'\backslash g'$. That is, $c\backslash((c\backslash c') \cup g') \subseteq c'\backslash g'$. Because $g \supseteq ((c\backslash c') \cup g')$, we have $c\backslash g \subseteq c\backslash((c\backslash c') \cup g') \subseteq c'\backslash g'$, hence, $c\backslash g \subseteq c'\backslash g'$. Since $c\backslash g \subseteq c'\backslash g'$, $g \supset g'$, and $conf(g \Rightarrow c\backslash g) \leq conf(g' \Rightarrow c'\backslash g')$, according to Definition 3, we can conclude that $g \Rightarrow c\backslash g$ is redundant to $g' \Rightarrow c'\backslash g'$. ■

According to Modus tolen inference rule, i.e., if the consequent of an implication is false, the antecedent of the rule must be false, from Lemma 1, we get the following corollary:

Corollary 1: Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C, and $\gamma\circ\tau(g) \subset c$. If $g \Rightarrow c\backslash g$ is a non-redundant rule, then $\forall c' \in C, \forall g' \in G, \gamma\circ\tau(g') \subset c'$ and $g' \subset g$, we have $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$.

Theorem 2: Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C, and $\gamma\circ\tau(g) \subset c$. $g \Rightarrow c\backslash g$ is a non-redundant rule iff $\forall c' \in C, \forall g' \in G, g' \subset g, \gamma\circ\tau(g') \subset c'$, and $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$.

Proof:

1) $\Longrightarrow$. The proof follows the conclusion of Corollary 1.
2) $\Longleftarrow$. (i) Assuming that $\neg(g \supseteq ((c\backslash c') \cup g'))$, we get $g \subset (c\backslash c') \cup g'$, or $g \cap ((c\backslash c') \cup g') = \emptyset$, or $(g\cap((c\backslash c')\cup g') \subset ((c\backslash c')\cup g')) \wedge (g\cap((c\backslash c')\cup g') \subset g)$.

   (1). In the case that $g \subset (c\backslash c') \cup g'$ is true, assuming that $g \Rightarrow c\backslash g$ is redundant, then we get, $\exists c' \in C$, $\exists g' \in G$, and $\gamma\circ\tau(g') \subset c'$ (hence $g' \subset c'$) such that $g' \subseteq g$ and $c'\backslash g' \supseteq c\backslash g$.
   From $c'\backslash g' \supseteq c\backslash g$ and $g' \subseteq c'$, we have $c' \supseteq c'\backslash g' \supseteq c\backslash g$, i.e., $c' \supseteq c\backslash g$. Since $\gamma\circ\tau(g) \subset c$ thus $g \subset c$, obviously we have $c = (c\backslash g) \cup g$ and $(c\backslash g)\cap g = \emptyset$; also $(c\backslash c')\cup c' \supseteq c$ and $(c\backslash c')\cap c' = \emptyset$ are true. Therefore, we have $(c\backslash c')\cup c' \supseteq c = (c\backslash g)\cup g$, i.e.:

   $(c\backslash c') \cup c' \supseteq (c\backslash g) \cup g \quad (a)$
   Because $c' \supseteq c\backslash g$, $(c\backslash c') \cap c' = \emptyset$ and $(c\backslash g) \cap g = \emptyset$, after c' being removed from the left side of (a) and $c\backslash g$ being removed from the right side of (a), the formula (a) becomes $c\backslash c' \subseteq g$. From $g' \subseteq g$, we get $(c\backslash c') \cup g' \subseteq g \cup g' = g$, i.e., $(c\backslash c') \cup g' \subseteq g$ which contradicts to $(c\backslash c') \cup g' \supset g$.
   Therefore, the assumption is false, i.e., $g \Rightarrow c\backslash g$ is non-redundant.

   (2). In the case that $g \cap ((c\backslash c') \cup g') = \emptyset$ is true, $g \cap g' = \emptyset$, thus $g \supset g'$ is always false. Therefore, $g \Rightarrow c\backslash g$ can't be redundant to $g' \Rightarrow c'\backslash g'$.

   (3). In the case that $(g \cap ((c\backslash c') \cup g') \subset ((c\backslash c') \cup g')) \wedge (g \cap ((c\backslash c') \cup g') \subset g)$ is true, there must exist some x such that $x \in c\backslash c'$ and $x \notin g$ or $x \in g'$ and $x \notin g$. The former will make $(c\backslash g) \subset (c'\backslash g')$ false and the latter will make $g \supset g'$ false. Therefore, $g \Rightarrow c\backslash g$ will never be redundant to $g' \Rightarrow c'\backslash g'$

(ii) Assuming that $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$. From Definition 3, we can directly conclude that $g \Rightarrow c\backslash g$ is not redundant.

■

The proposed Reliable Approximate Basis defines a more concise set of approximate rules which are non-redundant, sound and lossless. The algorithm to extract non-redundant approximate rules based on the Reliable Approximate Basis is given below:

Algorithm 1: **ReliableApproxBasis(**$Closure$**)**
Input: $Closure$: a set of frequent closed itemsets
$Generator$: a set of minimal generators

Output: A set of non-redundant approximate rules.
1. $approxRules := \emptyset$
2. for each $c \in Closure$
3. for each $g \in Generator$ such that $\gamma \circ \tau(g) \subset c$
4. if $\forall c' \in Closure$, $\forall g' \in G$ such that $\gamma \circ \tau(g') \subset c'$ and $g' \subseteq g$
5. we have $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$
6. then $approxRules := approxRules \cup \{g \Rightarrow (c\backslash g)\}$
7. end-for
8. end-for
9. Return $appproxRules$

For the Mushroom example dataset, 21 non-redundant approximate rules are extracted based on the Reliable Approximate basis. Rules 22, 23, 24 and 25 in Table III extracted based on the Minmax Approximate basis are considered redundant under the Reliable Approximate basis, respectively, and thus eliminated.

B. Deriving All Approximate Association Rules

Algorithms have been proposed to derive all association rules from the Min-max bases [13] and the Reliable Exact basis [17]. In this section, we provide an algorithm that can derive all approximate rules from the Reliable Approximate basis.

According to the definitions 4 and 5, the Min-max Approximate basis can be described as:

$$\begin{aligned} MinMaxApprox &= \{g \Rightarrow (c\backslash g) | c \in C, g \in G, \gamma \circ \tau(g) \subset c\} \\ &= \{g \Rightarrow (c\backslash g) | c \in C, g \in G, \\ &\quad (\neg(g \supseteq ((c\backslash c') \cup g')) \text{ or } conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')) \\ &\quad for\ all\ c' \in C, g' \in G, \gamma \circ \tau(g) \subset c \\ &\quad or\ \ (g \supseteq ((c\backslash c') \cup g') \text{ and } conf(g \Rightarrow c\backslash g) \leq conf(g' \Rightarrow c'\backslash g')) \\ &\quad for\ some\ c' \in C, g' \in G, \gamma \circ \tau(g') \subset c'\} \\ &= \{g \Rightarrow (c\backslash g) | c \in C, g \in G_c, \\ &\quad \neg(g \supseteq ((c\backslash c') \cup g')) \text{ or } conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g') \\ &\quad for\ all\ c' \in C, g' \in G, \gamma \circ \tau(g) \subset c\} \cup \\ &\quad \{g \Rightarrow (c\backslash g) | c \in C, g \in G_c, \\ &\quad g \supseteq ((c\backslash c') \cup g') \text{ and } conf(g \Rightarrow c\backslash g) \leq conf(g' \Rightarrow c'\backslash g') \\ &\quad for\ some\ c' \in C, g' \in G, \gamma \circ \tau(g') \subset c'\} \\ &= ReliableApprox \cup NonReliableApprox \end{aligned}$$

Where

$$\begin{aligned} NonReliableApprox = \{&g \Rightarrow (c\backslash g) | c \in C, g \in G_c, \\ &g \supseteq ((c\backslash c') \cup g') \text{ and } conf(g \Rightarrow c\backslash g) \leq conf(g' \Rightarrow c'\backslash g') \\ &for\ some\ c' \in C, g' \in G, \gamma \circ \tau(g') \subset c'\} \end{aligned} \quad (6)$$

The following theorem shows that, for $r_2 : g_2 \Rightarrow c_2\backslash g_2$, $c_2 \in C$ and $g_2 \in G$ (i.e., r_2 is a rule in $MinMaxApprox$), if for some $c_1 \in C$ and some $g_1 \in G$, there is $(g_1 \supseteq (c_1\backslash c_2) \cup g_2)$ and $conf(r_1) \leq conf(r_2)$, then we can deduce: $r_1 : g_1 \Rightarrow c_1\backslash g_1$ is a rule in $NonReliableApprox$. This means that, from a rule in $MinMaxApprox$, we could deduce a $NonReliableApprox$ rule.

Theorem 3: Let C be the set of frequent closed itemsets and G be the set of minimal generators. For rules $r_1 : g_1 \Rightarrow c_1\backslash g_1$ and $r_2 : g_2 \Rightarrow c_2\backslash g_2$ where $c_1, c_2 \in C$, $g_1, g_2 \in G$, $\gamma \circ \tau(g_1) \subset c_1$, and $\gamma \circ \tau(g_2) \subset c_2$. r_1 is a NonReliable approximate rule iff $(g_1 \supseteq (c_1\backslash c_2) \cup g_2)$ and $conf(r_1) \leq conf(r_2)$.

Proof:

1) $\Longrightarrow$
According to the definition of Min-max approximate basis, both $r_1 : g_1 \Rightarrow c_1\backslash g_1$ and $r_2 : g_2 \Rightarrow c_2\backslash g_2$ are Min-max approximate rules. If $g_1 \supseteq (c_1\backslash c_2) \cup g_2$ and $conf(r_1) \leq conf(r_2)$, then $\neg(g_1 \supseteq (c_1\backslash c_2) \cup g_2)$ must be false. According to the definition of Reliable approx basis, $r_1 \notin ReliableApprox$ must be true. Therefore, $r_1 \in NonReliableApprox$ is true.

2) $\Longleftarrow$
Assuming that $r_1 : g_1 \Rightarrow c_1\backslash g_1 \in NonReliableApprox$. From Equation (6), we immediately get, $g_1 \supseteq ((c_1\backslash c_2) \cup g_2)$ and $conf(r_1) \leq conf(r_2)$ for some $c_2 \in C$,, and $g2 \in G$.

■

We designed the following algorithm $AllApproxFromReliable$ to derive all approximate rules from the Reliable Approx basis. The algorithm $AllApproxFromReliable$ takes $ReliableApprox$ as the initial value for $MinMaxApprox$. Steps 4-8 generate approximate rules from an approximate basis rule in current $MinMaxApprox$. Steps 9 to 14 deduce $NonReliableApprox$ basis rules and add them into the current $MinMaxApprox$. Therefore, during the process of deriving approximate rules, we generates all $NonReliableApprox$ rules so that $MinMaxApprox$ will be completed progressively during the course. Theorem 3 ensures that we can deduce all $NonReliableApprox$ basis rules. On completion of executing Algorithm 2, $MinMaxApprox$ will contains all $ReliableApprox$ basis rules and also all $NonReliableApprox$ basis rules. Steps 17 to 21 in Algorithm 2 derive all approximate rules from these basis rules, which performs the same task as the steps 11 to 17 in the approximate reconstruction algorithm proposed in [13].

Algorithm 2: **AllApproxFromReliable**($ReliableApprox$)
Input: $ReliableApprox$: reliable approximate basis
Output: $AllApprox$: A set of all approximate association rules
1. $AllExact := \emptyset$, $MinMaxApprox := ReliableApprox$
2. for $i = 2$ to maximum size of closed itemsets
3. for rule $(r_1 : a_1 \Rightarrow c_1, r_1.supp, r_1.conf) \in$

$MinMaxApprox$
and $|c_1| = i$
4. for subset $c_2 \subset c_1$
5. if $(r_2 : a_1 \Rightarrow c_2, r_2.supp, r_2.conf) \notin AllApprox$
6. and $r_2.conf \neq 1$//r_2 is not an exact rule
7. then $AllApprox := AllApprox \cup \{(r_2 : a_1 \Rightarrow c_2, r_1.supp, r_1.conf)\}$
8. end-for
9. for each closed itemset c_3
10. for generator a such that $a \supseteq a_1$ and $a.closure \subset c_3$
11. if $a \supseteq ((c_3 \backslash (c_1 \cup a_1)) \cup a_1)$ and $r_1.conf \geq \frac{c_3.supp}{a.supp}$
12. then $MinMaxApprox := MinMaxApprox \cup \{a \Rightarrow (c_3 \backslash a), c_3.supp, \frac{c_3.supp}{a.supp}\}$
13. end-for
14. end-for
15. end-for
16. end-for
17. for rule $(r_1 : a_1 \Rightarrow c_1, r_1.supp, r_1.conf) \in AllApprox$
18. for each subset $c_3 \subseteq c_2$ where $c_2 = a_1.closure \backslash a_1$, $\frac{(a_1.closure).supp}{a_1.supp} = 1$
19. $AllApprox := AllApprox \cup \{a_1 \cup c_3 \Rightarrow c_1 \backslash c_3, r_1.supp, r_1.conf\}$
20. end-for
21. end-for
22.return $AllExact$

IV. Experiments

We have conducted experiments to evaluate the effectiveness of the proposed Reliable approximate basis. This section presents the experimental results.

A. Datasets

We used the following three datasets from UCI KDD Archive (http://kdd.ics.uci.edu/). The Mushrooms dataset contains 8,124 records each of which describes the characteristics of one mushroom object. Each mushroom object has 23 attributes some of which are multiple value attributes. After converting the multiple value attributes to binary ones, the number of attributes of each object becomes 126. The Annealing dataset contains 898 annealing instances (objects), each has 38 attributes. After converting multiple value attributes to binary ones, each object has 276 attributes. The Flare2 dataset contains 1,066 solar flare instances each of which represents captured features for one active region on the sun. Each flare instance has 50 attributes after the multiple value attributes are converted to binary attributes. The experiment is to find the associations among attributes for the three datasets.

B. Evaluation Results

In this experiment, firstly we confirm that both the MinMax basis and the Reliable basis can deduce all approximate rules. For example, when $Minsupp$ is 0.3, both bases produce 21,377 approximate rules for the Mushroom dataset as showed in Table IV. Secondly, we test the reduction ratio between the size of the $MinMaxApprox$ basis and the size of the $ReliableApprox$ basis for different $Minsupp$ settings.

TABLE IV
Number of approximate rules (Mushroom dataset, minconf=0.5)

Minsupp	Approx rules derived (MinMax,Reliable)	MinMax Approx Basis	Reliable Approx Basis	Reduction Ratio
0.3	21,377	2,634	1,970	25%
0.4	2,528	465	361	22%
0.5	835	175	135	23%
0.6	228	59	52	12%
0.7	161	39	34	13%
0.8	71	25	21	16%

For all tests, the $minconf$ was set to 0.5. Table IV, Table V, and Table VI present the test results for the three datasets, respectively.

The experiment results showed that the reduction is considerable high. For instance, when $Minsupp$ was set to 0.3, for the Annealing dataset, the $MinMax$ basis contains 865 basis rules as showed in Table V, while the $Reliable$ basis contains 554 basis rules, the reduction ratio is 36%. In this case, 5,052 approximate rules can be deduced either from the $MinMax$ basis or from the $Reliable$ basis. For example, the following 9 rules in the $MinMax$ basis are redundant to the reliable rule *steel-A ⇒ product-type-C,strength-000* (0.4844, 0.9886), therefore they are excluded in the $Reliable$ basis:

steel-A,carbon-00 ⇒ product-type-C,strength-000, (0.47327, 0.9884)
steel-A,hardness-00 ⇒ product-type-C,strength-000, (0.30512,0.9821)
steel-A,bore-0000 ⇒ product-type-C,strength-000, (0.4655,0.9882)
steel-A,class-3 ⇒ product-type-C,strength-000, (0.3853,0.9858)
steel-A,carbon-00,bore-0000 ⇒ product-type-C,strength-000, (0.4543, 0.9879)
steel-A,carbon-00,class-3 ⇒ product-type-C,strength-000, (0.3775,0.9854)
steel-A,hardness-00,bore-0000 ⇒ product-type-C,strength-000, (0.3040, 0.9820)
steel-A,bore-0000,class-3 ⇒ product-type-C,strength-000, (0.3731,0.9853)
steel-A,carbon-00,bore-0000,class-3 ⇒ product-type-C,strength-000, (0.3653,0.9850)

The 9 rules listed above have the same consequent but a larger antecedent than that of the reliable rule *steel-A ⇒ product-type-C,strength-000*. Both the support and confidence values , as indicated as (support, confidence) at the end of each rule, of these 9 rules are smaller than that of the reliable rule. Therefore, according to Theory 1, their CF value won't be greater than that of the reliable rule. In real world problem solving, if we know that *steel-A* is true, by applying the rule *steel-A ⇒ product-type-C,strength-000*, we can conclude that *product-type-C,strength-000* is true. We don't have to know *hardness-00*, *class-3*, or *bore-0000*, etc. in order to reach this consequence. That means, all the 9 rules are useless or redundant if we have the rule *steel-A ⇒ product-type-C,strength-000* at hand. Eliminating these redundant rules can greatly reduce the size of the discovered rule set, but the capacity of the rule base in solving problems remains the same.

V. Related Work

Many approaches have been proposed aiming at reducing the number of extracted rules and improving the"usefulness" of the rules as well[1], [3], [7], [15]. Also some work has been done on concisely representing and interpreting multidimensional association rules using granules and multi-tier

TABLE V
NUMBER OF APPROXIMATE RULES (ANNEALING DATASET, MINCONF=0.5)

Minsup	Approx rules derived (MinMax,Reliable)	MinMax Approx Basis	Reliable Approx Basis	Reduction Ratio
0.3	5,052	865	554	36%
0.4	1,835	435	296	32%
0.5	1,186	300	218	27%
0.6	416	137	102	26%

TABLE VI
NUMBER OF APPROXIMATE RULES (FLARE2 DATASET, MINCONF=0.5)

Minsupp	Approx rules derived (MinMax,Reliable)	MinMax Approx Basis	Reliable Approx Basis	Reduction Ratio
0.3	7,604	1216	710	42%
0.4	2,420	644	479	27%
0.5	5,599	1081	730	32%
0.6	5,368	1203	687	43%

structures [10], [18]. But eliminating redundancy of rules is not a focus of these approaches. The approaches proposed in [13] and [19] focus on extracting non-redundant rules. Both of them make use of the closure of the Galois connection [5] to extract non-redundant rules from frequent closed itemsets instead of from frequent itemsets. One difference between the two approaches is the definition of redundancy. The approach proposed in [19] extracts the rules with shorter antecedent and shorter consequent as well among rules which have the same confidence, while the method proposed in [13] defines that the non-redundant rules are those which have minimal antecedents and maximal consequents. Our definition to redundant rules is similar to that of [13]. However, the requirement to redundancy is relaxed, and the less requirement makes more rules to be considered redundant and thus eliminated. Most importantly, we prove that the elimination of such redundant rules does not reduce the belief to the extracted rules and the capacity of the extracted non-redundant rules for solving problems will also not be reduced. The concept of non-derivable itemsets was introduced in [4]. The basic idea is to find lower and upper bounds on the support of an itemset based on the support of its subsets. When these bounds are equal, the itemset is considered derivable. The set of frequent non-derivable itemsets allows for deriving the supports of all other frequent itemsets and as such forms a concise representation from which all other frequent itemsets can be derived. Goethals proposed a method to derive non-derivable rules from the non-derivable itemsets [6]. The amount of the non-derivable rules is much smaller than the size of the entire rule set. However, it was not discussed whether the non-derivable rule set has the same capacity to solve problems as the entire rule set.

VI. CONCLUSION

One challenge problem with association rule mining is the redundancy in the extracted rules. The work presented in this paper aims at improving the quality of association rules by eliminating redundancy. In this paper, we proposed a relaxed definition of redundancy and a concise representation of approximate association rules. We theoretically proved that the proposed Reliable Approximate basis can eliminate considerable amount of redundancy. Based on certainty factor theory, we also proved that the elimination of the redundancy using the proposed Reliable basis does not reduce the belief to the extracted rules. Similar to the Min-max basis, the proposed Reliable approximate basis is not only a concise but also a lossless representation of approximate rules. From the Reliable approximate basis, all approximate rules can be deduced.

REFERENCES

[1] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4:217–240, 2000.

[2] M. J. A. Berry and G. S. Linoff. *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley and Sons, 1997.

[3] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD Conference*, pages 255–264, 1997.

[4] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. pages 74–85. Springer, 2002.

[5] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, 1999.

[6] B. Goethals, J. Muhonen, and H. Toivonen. Mining non-derivable association rules. In *Proceedings of the SIAM International Conference on Data Mining*, pages 239–249, 2005.

[7] J. Han and Y. Fu. Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering*, 11:798–804, 5 2000.

[8] J. Han, L. V. S. Lakshmanan, and R. T. Ng. Constraint-based multidimensional data mining. *IEEE Computer*, 32(8):46–50, 1999.

[9] M. Kryszkiewicz, H. Rybinski, and M. Gajek. Dataless transitions between concise representations of frequent patterns. *Journal of Intelligent Information Systems*, 22(1):41–70, 2004.

[10] Y. Li, W. Yang, and Y. Xu. Multi-tier granule mining for representations of multidimensional association rules. In *the 6th IEEE International Conference on Data Mining (ICDM06)*, pages 953–958, 2006.

[11] R. T. Ng, V. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning otimizations of constrained association rules. In *Proceedings of the SIGMOD conference*, pages 13–24, 1998.

[12] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rultes using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[13] N. Pasquier, R. Taouil, Y. Bastide, G. Stumme, and L. Lakhal. Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, 24(1):29–60, 2005.

[14] E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3/4):351–379, 1975.

[15] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *Proceedings of the KDD Conference*, pages 67–73, 1997.

[16] R. Wille. *Restructuring lattices theory: An approach based on hierarchies of concepts*. I. Rival (editor), Ordered sets. Dordrecht-Boston, 1982.

[17] Y. Xu and Y. Li. Generating concise association rules. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM07)*, pages 781–790, 2007.

[18] W. Yang, Y. Li, J. Wu, and Y. Xu. Granule mining oriented data warehousing model for representations of multidimensional association rules. *International Journal of Intelligent Information and Database Systems*, 2008.

[19] M. J. Zaki. Generating non-redundent association rules. In *Proceedings of the KDD Conference*, pages 34–43, 2000.

[20] M. J. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223–248, 2004.

Parimputation: From Imputation and Null-Imputation to Partially Imputation

Shichao Zhang, *Senior Member, IEEE*

***Abstract*— Missing data imputation is an important step in the process of machine learning and data mining when certain values are missed. Among extant imputation techniques, kNN imputation algorithm is the best one as it is a model free and efficient compared with other methods. However, the value of *k* must be chosen properly in using kNN imputation. In particular, when some nearest neighbors are far from a missing data, the kNN imputation algorithms are often of low efficiency. In this paper, a new imputation framework is designed. The imputation uses the left or right nearest neighbor for a missing data in a given dataset. Furthermore, a parimputation (partially imputation) strategy is proposed for dealing with the issue of missing data imputation. Specifically, some missing data are imputed when there are some complete data in a small neighborhood of the missing data and, other missing data without imputation are given up in applications, such as data mining and machine learning.**

***Index Terms*—Artificial intelligence; Data management; Data processing.**

I. INTRODUCTION

In real applications, missing value imputation is an actual and challenging problem confronted by machine learning and data mining. Therefore, there are great many efforts to missing value imputation. Traditional missing value imputation techniques can be roughly classified into regression imputation (RI) and nearest neighbor imputation (NNI) [33]. And missing values in a dataset are completed by replacing them with some plausible values. The plausible values are generally generated from the dataset using an imputation method.

RI can be classified into deterministic regression imputation (DRI) and stochastic regression imputation (SRI). Using a DRI method, missing values in a dataset are replaced with only the mean of all the known values in the dataset. Using an SRI method, each of missing values is replaced with the mean plus a random value. Experiment results have proven [22] that SRI methods are much better than DRI methods in many practical cases. However, it is usually more difficult to mathematically prove the efficiency for SRI methods.

NNI [33] is one of the hot deck techniques used to compensate for missing data. It has been successfully used in, for example, U.S. Census Bureau and Canadian Census Bureau. Using an NNI method, a missing value in a dataset is replaced with the value of the nearest neighbor in the dataset. kNNI (*k*-nearest-neighbors imputation) is an extension of NNI method (It is an NNI algorithm when $k = 1$). It takes into account *k* nearest neighbors when imputing. Yet, it is difficult to mathematically prove the efficiency for kNNI methods.

While having good randomicity, SRI methods are poor in efficiency when compared with kNNI techniques. However, the value of *k* must be selected properly when using kNNI methods. In particular, the nearest neighbor may be far from a missing data and the kNNI methods are thus of low efficiency. In this paper a new imputation framework is designed. Furthermore, it advocates giving up imputation if there is no close neighbors and only imputing those missing data that the nearest neighbor is not far from them. It is referred to a parimputation (partially imputation) strategy.

The rest of this paper is organized as follows. Section II briefly recalls related work on missing value imputation. In Section III we present an imputation framework. In Section IV, we design the parimputation strategy. We simply evaluate the proposed approach in Section V. This paper is concluded in Section VI.

II. RELATED WORK

The missing data problem is faced in many application domains, such as, statistical analysis, machine learning, data mining, pattern recognition and information retrieval. Because imputation algorithms are designed independent of applications, we only review major related work in the application domains of statistical analysis and data mining in this section.

A. Research into Statistical Imputation for Missing Data

Statistical analysis with missing data has been noted in the literature for more than 70 years. Wilks [28] initiated a study on the maximum likelihood estimation for multivariate normal models with fragmentary data. Thereafter, extensive discussions on this topic continue. A useful reference for general parametric statistical inferences with missing data can be found in [16].

Little and Rubin [15] classified missing data mechanisms into three categories as follows.

1. **Missing Completely at Random** (MCAR): Cases with complete data are indistinguishable from cases with incomplete data. Heitjan [9] provided an example of

Shichao Zhang is with the College of Computer Science and Information Technology, Guangxi Normal University, PR China; the State Key Lab for Novel Software Technology, Nanjing University, PR China; e-mail: zhangsc@mailbox.gxnu.edu.cn.

MCAR missing data and, Graham, Hofer and MacKinnon [12] illustrated the use of planned missing data patterns.

2. **Missing at Random** (MAR): Cases with incomplete data differ from cases with complete data, but the pattern of data missingness is traceable or predictable from other variables in the database rather than being due to the specific variable on which the data are missing.
3. **Nonignorable**: The pattern of data missingness is non-random and it is not predictable from other variables in the database.

In practice it is usually difficult to meet the nonignorable assumption. MAR is an assumption that is more often (MCAR is a special case of MAR), but not always tenable. The more relevant and related predictors one can include in statistical models, the more likely it is that the MAR assumption will be met.

B. Research into Missing Data Imputation in Data Mining

Recently, Magnani [17] has reviewed the main missing data techniques, including conventional methods, global imputation, local imputation, parameter estimation and direct management of missing data. He tried to highlight the advantages and disadvantages for all kinds of missing data mechanisms. For example, he revealed that statistical methods have been mainly developed to manage survey data and proved to be very effective in many situations. However, the main problem of these techniques is its strong model assumptions.

Batista and Monard [3] have analyzed the performance of 10-NNI as an imputation method, comparing its performance with other three missing data imputation methods: mean or mode imputation, C4.5 and CN2. This work proposed the advantages of the method: it can predict both qualitative attributes and quantitative attributes, and it does not create explicit modes (like a decision tree or a rules) because it is a lazy model. Their experiments showed that the method provides very good results than the other three methods, even for a large amount of missing data. A main drawback is that the algorithm must search through all the data set limiting in large databases only based on MCAR. Different imputations for industrial databases have also been studied in [12].

Yuan [30] reviewed three methods of multiple imputation for missing data, including regression method, propensity score method and MCMC (Markov Chain Monte Carlo) method. Also, he used standard statistical methods to evaluate the efficiency of multiple imputation.

Allison [2] has evaluated two algorithms for producing multiple imputations or missing data using simulated data based on the software of SOLAS. Software using a propensity score classifier with the approximate Bayesian boostrap was found to produce badly biased estimates of regression coefficients when data on predictor variables are MAR or MACR. Allison has also showed that listwise deletion produces unbiased regression estimates whenever the missing data mechanism depends only on the predictor variable, not on the response variable.

Other missing data imputation methods include a new family of reconstruction problems for multiple images from minimal data [11], a method for handling inapplicable and unknown missing data [8], different substitution methods for replacement of missing data values [20], robust Bayesian estimator [26], and nonparametric kernel classification rules derived from incomplete (missing) data [18].

III. An Imputation Framework for Dealing with Missing Values

Let X be a d-dimensional vector of factors and let Y be a response variable influenced by X. In practice, one often obtains a random sample (sample size = n) of incomplete data associated with a population (X, Y, δ),

$$(X_i, Y_i, \delta_i),\ i = 1, 2, \ldots, n$$

Where all the X_i's are observed and $\delta_i = 0$ if Y_i is missing, otherwise $\delta_i = 1$. Suppose that (X_i, Y_i) satisfies the following model:

$$Y_i = m(X_i) + \varepsilon_i,\ i = 1, 2, \ldots, n$$

Where $m(.)$ is an unknown function, and the unobserved ε_i (with population ε) are i.i.d. random errors with mean 0 and unknown finite variance σ^2, and are independent of the i.i.d. random variables X_i's.

To impute the missing values, $m(.)$ must be estimated. The $m(.)$ are often measured the statistical parameters of the response variable Y such as $\mu = EY$, $\theta = F(y)$ and θ_q, i.e. the mean, the distribution function and the q-th quantile of Y, where y is a fixed point in $\Re$, and $0<q<1$. EY stands for the average level of Y, the distribution function $F(y)$ is the probability of Y being smaller than or equal to the given y, and θ_q is the level of Y that satisfies $P(Y \le \theta_q) = q$. The median of Y (the case of $q = 1/2$) is the most important case of quantiles. The inference for them is a very important issue in practice.

In the situation where $m(.)$ is a linear function, i.e. Y and X fit a linear model, Wang and Rao [29] have compared the adjusted empirical likelihood methods and the normal approximation methods in terms of coverage accuracies and average lengths of the confidence intervals. They have indicated that the adjusted empirical likelihood methods perform competitively, the use of auxiliary information provides improved inferences and the deterministic imputation method performs well in making inference for the mean of Y. Qin et al. [21] have showed that one must use random imputation methods in making inference for distribution functions and quantiles of Y.

Yet in many complex practical situations, $m(.)$ (an unknown function) is not a linear function. When we do not know the form of $m(.)$, i.e. the nonparametric situation, Wang and Rao [29] have considered empirical likelihood inference on

the mean of response *Y* when *Y* is missing at random (MAR) . They have only used the deterministic imputation method to infer the mean of *Y*, and left the inference for distribution functions and quantiles of *Y* unsolved.

To avoid estimating $m(.)$, NNI method replaces a missing value in a dataset with the value of the nearest neighbor in the dataset. Further, kNNI method is proposed, which replaces a missing value in a dataset with the mean of *k* nearest neighbors when imputing. NNI or kNNI algorithms have experimentally been proved more efficient than other existing imputation methods [33]. They have widely been used in applications. However, as mentioned before, (1) it must seek a proper *k* when using kNN imputation methods; and (2) when some nearest neighbors are far from a missing datum, the kNN imputation algorithms are often of low efficiency. The first issue is tackled in this section and the second one will be dealt with in next section.

A. Imputation Model

This subsection builds a new imputation model that uses the left or right nearest neighbor for a missing data in a given dataset.

For a 2-dimensional imputation problem, let T_1 = $(X_i, Y_i, 1)$, T_2 = $(X_j, Y_j, 1)$, T = $(X_l, Y_l, 0)$ in a dataset, where T_1 and T_2 are the left and right nearest neighbors of an incomplete data *T* with respect to the factor *X*, respectively. That is, for any complete data T_3 = $(X_k, Y_k, 1)$ in the dataset, we have either

$$X_k \le X_i \text{ or } X_k \ge X_j$$

With T_1 and T_2, we can replace Y_l with the mean of Y_i and Y_j, or

$$Y_l = \frac{1}{2}(Y_i + Y_j)$$

In the same reason, for an (*n*+1)-dimensional imputation problem, we select such 2*n* complete data, T_1^-, T_1^+, …, T_n^-, T_n^+ from a given dataset, where T_i^-, T_i^+ are the left and right nearest neighbors of an incomplete data *T* with respect to the factor X_i , respectively. Formally, let *T* = $(X_{l1}, X_{l2}, ..., X_{ln}, Y_l, 0)$ in the dataset, *NN* is a set of all nearest neighbors of *T* in the dataset, and *T*'s left and right nearest neighbors with respect to the factor X_i are as follows:

T_i^- = $(X_{i1}^-, X_{i2}^-, ..., X_{in}^-, Y_{i-}, 1)$, *i* = 1, 2, …, *n*

T_i^+ = $(X_{i1}^+, X_{i2}^+, ..., X_{in}^+, Y_{i+}, 1)$, *i* = 1, 2, …, *n*

where T_i^- or T_i^+ may not exist in *NN*. They satisfy that, for a nearest neighbor $(X_{j1}, X_{j2}, ..., X_{jn}, Y_{j+}, 1)$ in *NN*, either $X_{ji} \le X_{ii}^-$ if there is a T_i^- in *NN*, or $X_{ji} \ge X_{ii}^+$ if there is a T_i^+ in *NN*.

With these nearest neighbors, we can replace Y_l with the mean of all the Y_{i-} and Y_{i+}. Or

$$Y_l = \frac{1}{2n}\sum_{i=1}^{n}(Y_{i-} + Y_{i+}) \quad (2)$$

B. Model Enhancement

In Section III.A we have proposed a simple and easy-implemented imputation model. From the selection of the left and right nearest neighbors of a missing datum with respect to the factor X_i, there are three cases as follows.

1. There may be no left or right nearest neighbor for a missing data in a given dataset, with respect to the factor X_i.
2. A complete data may be selected multiple times in the set of left /right nearest neighbors of a missing data in a given dataset, with respect to the factor X_i.
3. Some left or right nearest neighbors of a missing data in a given dataset, with respect to the factor X_i may be far from the missing data.

For the first case, we can simply give up all the missed left or right nearest neighbors when estimating the missing data. The second case shows that fact: the more times a complete data is selected, the closer to the missing data the complete data is.

For the third case, we can use weighting technique to weaken their impact to the missing data when estimating the missing data. The weight of a left or right nearest neighbor of a missing data can be determined as follows.

For a left or right nearest neighbor T_i = $(X_{i1}, X_{i2}, ..., X_{in}, Y_i, 1)$ of a missing data *T* = $(X_{l1}, X_{l2}, ..., X_{ln}, Y_l, 0)$, we obtain

$$d_i = \sqrt{(X_{i1} - X_{l1})^2 + ... + (X_{in} - X_{ln})^2}$$

Hence, we can get the weight w_i of T_i as follows.

$$w_i = 1 - \frac{d_i}{d_1 + d_2 + ... + d_m} \quad (3)$$

Where, "*m*" is the number of the selected left or right nearest neighbors of the missing data. With these weights, we can estimate Y_l as follows.

$$Y_l = \sum_{i=1}^{n} (w_{i-} Y_{i-} + w_{i+} Y_{i+}) \quad (4)$$

Further, we can waive all the left or right nearest neighbors that are far from the missing data according to d_i or w_i. In other words, we can select those left or right nearest neighbors that are very close to the missing data. After filtering some nearest neighbors, it is easy to estimate Y_l by improving Eqns. (3) and (4).

C. Imputation Framework

From the above, our new approach, called ENI (encapsidated-neighbor imputation), is similar to kNNI method. There are two main differences between ENI and kNNI as follows:

1. The ENI approach takes into account the left and right nearest neighbors of a missing data, whereas the kNNI method selects *k* nearest neighbors.
2. In ENI approach, the number of the selected nearest neighbors is a variable determined by data when imputing missing data, whereas the kNNI method uses a fixed *k*.

With the ENI approach, the process of missing data imputation is as follows.

Let X be a n-dimensional vector of factors, Y a response variable influenced by X, a dataset of incomplete data associated with a population (X, Y, δ) be as follows

$$(X_i, Y_i, \delta_i), \ i = 1, 2, \ldots, N$$

1. For each incomplete data $T = (X_{l1}, X_{l2}, ..., X_{\ln}, Y_l, 0)$, search all the left or right nearest neighbor of Y: $T_1, T_2, \ldots, T_m$;
2. Use the Eqn (3) to calculate the weight w_i of T_i, i = 1, 2, …, m;
3. Estimate Y_l with Eqn (4);
4. Repeat Steps 1-3 until no incomplete data in the dataset.

This process is simple and easy to be understood and implemented.

IV. PARIMPUTATION: PARTIALLY IMPUTATION

From Section III, the ENI method takes into account all the left or right nearest neighbors of missing data when imputing them. However, like kNNI algorithms, it still suffers from the fact: sometimes all the left or right nearest neighbors can be far from a missing data in a dataset. When this case happens and the missing data is imputed with ENI or kNNI method, the results from the dataset can be inaccurate. To deal with this issue, this paper advocates a parimputation strategy: some missing data are imputed when there are some complete data in a small neighborhood of the missing data and, other missing data without imputation are given up in applications, such as data mining and machine learning.

For understanding the strategy, in this section, we first review the known value strategy and the null strategy that have been widely used in machine learning and data mining applications for dealing with missing data [22], and then propose the parimputation strategy, regarded as a new strategy.

A. Known Value Strategy for Missing Data

In cost-sensitive learning, the first tree building and test strategy for "missing is useful" is called the Known Value Strategy [14] [31]. It utilizes only the known attribute values in the tree building for each test example. For each test example, a new (and probably different) decision tree is built from the training examples with only those attributes whose values are known in the test example. That is, the new decision tree only uses attributes with known values in the test example, and thus, when the tree classifies the test example, it will never encounter any missing values.

The Known Value Strategy was proposed in [14] but its ability of handling unknown values was not studied. Clearly, the strategy utilizes all known attributes and avoids any missing data directly.

In [14], an internal node strategy was also proposed. It keeps examples with missing values in internal nodes, and does not build branches for them during tree building. When classifying a test example, if the tree encounters an attribute whose value is unknown, then the class probability of training examples falling at the internal node is used to classify it. As unknown values are dealt with using internal nodes, this strategy is called as the Internal Node Strategy.

As there might be several different situations where values are missing, leaving the classification to the internal nodes may be a natural choice. This strategy is also quite efficient as only one tree is built for all test examples.

B. Null Strategy

As values are missing for a certain reason – unnecessary and too expensive to test – it might be a good idea to assign a special value, often called "null" in databases [6], to missing data. The null value is then treated just as a regular known value in the tree building and test processes. This strategy has also been proposed in machine learning [1].

One potential problem with the Null Strategy is that it does not deliberately utilize the known values, as missing values are treated just as a known value. Another potential drawback is that there might be more than one situation where values are missing. Replacing all missing values by one value (null) may not be adequate. In addition, subtrees can be built under the "null" branch, suggesting oddly that the unknown is more discriminating than known values. The advantage of this strategy is its simplicity and high efficiency compared to the Known Value Strategy, as only one decision tree is built for all test examples.

Also, C4.5 [23][24] does not impute missing values explicitly, and it is shown to be quite effective [3]. And C4.5's missing-value strategy is applied directly in cost-sensitive trees. During training, an attribute is chosen by the maximum cost reduction discounted by the probability of missing values of

that attribute. During testing, a test example with missing value is split into branches according to the portions of training examples falling into those branches, and goes down to leaves simultaneously. The class of the test example is the weighted classification of all leaves.

C. Parimputation Strategy

As described previously, the parimputation is a strategy for dealing with the issue of missing data imputation. The parimputation strategy is proposed for addressing those missing data in a given dataset that all the left or right nearest neighbors are far from them.

From the observed part of an incomplete datum in a dataset, if there are some complete data in a small neighborhood of the incomplete data, we refer it to a predictable missing data; otherwise, we refer it to an unpredictable missing data. With the observed part of an unpredictable missing data in a dataset, seeking the unpredictable missing data is similar to that of detecting outliers (or isolation points) in machine learning and data mining. This means that there are many well-established outlier detection techniques (such as [10] [25]) that can be applied to determining whether a missing data is unpredictable.

With the parimputation strategy, we can deal with the missing data in two ways as follows:

1. Impute all the predictable missing data in a dataset; remove all the unpredictable missing data from the dataset, and then discover patterns from the dataset that contains complete data and imputed data.
2. Impute only the predictable missing data in a dataset; and then discover patterns from the dataset with the known value strategy, or the null strategy.

From the above, the parimputation strategy is simple and easy to be understood and implemented.

V. Experiments

In order to show the effectiveness of the proposed methods, extensive experiments were done on a real dataset with the algorithm implemented in C++ and executed using a DELL Workstation PWS650 with 2G main memory, and 2.6G CPU.

A. Algorithm Design

As mentioned previously, the ENI is simple and easy to be understood and implemented. However, the description is only used to state the problem. We should select the left and right nearest neighbors of a missing data from a set of nearest neighbors of the missing data. There three cases as follows.

(1) There is no nearest neighbor in the set, i.e., the missing data is unpredictable and it is not imputed in our experiments.
(2) The number of left and right nearest neighbors of a missing data is often lesser than k. This means that there are only few data observed in the set of nearest neighbors.
(3) The number of left and right nearest neighbors of a missing data is greater than k. This means that there are plenty data observed in the set of nearest neighbors.

These indicate that the number of selected left and right nearest neighbors is variable when imputing missing data. In particular, we can only select the left and right nearest neighbors from the k nearest neighbors that are selected for a kNNI algorithm.

Because the goal of this paper is to introduce a new imputation strategy, we simply evaluate the ENI in next subsection with compared with the kNN method for imputing continuous missing target attributes in terms of imputation accuracy.

B. Experimental Results

The first set of experiments was conducted on a real dataset of a class in a high school. The dataset contains 711 instances in total and 12 attributes for each instance (non missing attribute values). The average score was selected as the target attributes (response variable, Y) and, the Math (X_1), Chinese (X_2) and English (X_3) as the factors, where $Y = X_1 + X_2 + X_3$. We used the missing mechanisms MCAR and MAR on Y at different missing rates of 5%, 10% and 20%. Then the ENI and kNNI algorithms were utilized to fill out the missing values of Y. Our experiments have demonstrated that the ENI is much better than kNNI method at the efficiency for this linear function.

The second set of experiments was conducted on a real dataset, *Abalone*, downloaded from UCI machine learning repository. We selected 1528 instances where 7 attributes were picked as the factors and another one as the response variable. We use the missing mechanisms MCAR and MAR on Y at different missing rates of 5%, 10% and 20%. Then the ENI and kNNI algorithms were utilized to fill out the missing values of Y. The experimental results are listed in Tables 1-3.

From Tables 1, 2 and 3, the ENI is much better than kNN method. In particular, when the missing rate is 20%, the ENI is better than kNN method in each imputation times. This demonstrates that using only the left and right nearest neighbors can improve the imputation performance kNNI methods.

This research is focused on the case that only one attribute is with missing values. If several attributes are with missing values. The use of the ENI is as follows.

(1) Select such an attribute as the response variable that the number of its missing values is minimal among the attributes with missing values.

(2) Use the ENI to impute the missing values based on all complete attributes [1] (without missing values).

(3) Repeat Steps (1) and (2) until all predictable missing values are imputed.

[1] From the second imputation, the imputed attributes are taken as complete attributes.

Table 1. When the missing rate is 5%, there are 76 instances with missing data in *Abalone*.

Imputation times	1	2	3	4	5	6	7	8	9	10
ENI	**41**	**42**	**42**	35	**44**	**47**	**39**	**38**	**43**	**48**
kNNI	35	34	34	**41**	32	29	37	**38**	33	28

Table 2. When the missing rate is 10%, there are 153 instances with missing data in *Abalone*.

Imputation times	1	2	3	4	5	6	7	8	9	10
ENI	**80**	**78**	**86**	75	**80**	**78**	**82**	76	**88**	**90**
kNNI	73	75	67	**78**	73	75	71	**77**	65	63

Table 3. When the missing rate is 20%, there are 306 instances with missing data in *Abalone*.

Imputation times	1	2	3	4	5	6	7	8	9	10
ENI	**170**	**165**	**162**	**158**	**174**	**168**	**155**	**159**	**154**	**163**
kNNI	136	141	144	148	132	138	151	147	152	143

VI. Conclusions

In this paper we have proposed a new imputation, called ENI. It is different from the kNNI method because

1. The ENI approach takes into account the left and right nearest neighbors of a missing data, whereas the kNNI method selects *k* nearest neighbors.
2. In ENI approach, the number of the selected nearest neighbors is variable when imputing missing data, whereas the kNNI method uses a fixed *k*.

From the extrapolation, the ENI approach is more reasonable than the kNNI method. Further, a parimputation strategy has been advocated for dealing with the unpredictable missing data in a dataset. The experimental results have demonstrated that the ENI is much better than the kNNI method.

The future work is to apply the ENI approach and the parimputation strategy to real machine learning and data mining applications, so as to improve the methods.

VII. Acknowledge

Thanks for the experiments carried out by my student, Mr Manlong Zhu. Thanks for the comments on the early version of this paper from Dr Yongsong Qin, Mr Xiaofeng Zhu, and Dr. William K. Cheung.

This work was supported in part by the Australian Research Council (ARC) under grant DP0985456, the Nature Science Foundation (NSF) of China under grant 90718020, the China 973 Program under grant 2008CB317108, the Research Program of China Ministry of Personnel for Overseas-Return High-level Talents, the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (07JJD720044), and the Guangxi NSF (Key) grants.

References

[1] Ali, K.M. and Pazzani, M.J. (1993). Hydra: A noise-tolerant relational concept learning algorithm. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (IJCAI93), pp. 1064-1071. Morgan Kaufmann, 1993.

[2] Allison, P. (2001). *Missing Data*. Sage Publication, Inc, 2001. [place of publication]

[3] Batista G. and Monard, M.C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, Vol. 17, pp. 519-533, 2003.

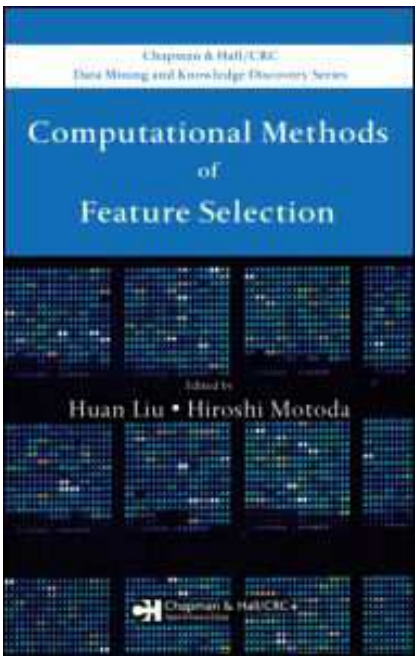
[4] Caruana, R, (2001). A Non-parametric EM-style algorithm for Imputing Missing Value. *Artificial Intelligence and Statistics*, January 2001.

[5] Chen, J., and Shao, J., (2001). Jackknife variance estimation for nearest-neighbor imputation. *J. Amer. Statist.* Assoc. 2001, Vol. 96: 260-269.
[6] Date, C.J. and Darwen, H. (1989). The default values approach to missing information. In: *Relational Database Writings 1989-1991*, pp. 343-354, 1989.
[7] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, series B, Vol. 39, pp. 1–38.
[8] Gessert, G., (1991). Handling Missing Data by Using Stored Truth Values. *SIGMOD Record,* 2001, Vol. 20(3): 30-42.
[9] Heitjan, D.F. Annotation (1997). What can be done about missing data? Approaches to imputation. *American Journal of Public Health*, 87(4): 548-550
[10] John, G. H. (1995). Robust Decision Trees: Removing Outliers from Databases. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 174–179. Menlo Park, CA: AAAI Press.
[11] Kahl, F., et al., (2001). Minimal Projective Reconstruction Including Missing Data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, Vol. 23(4): 418-424.
[12] Lakshminarayan, K., et al., (1996). Imputation of Missing Data Using Machine Learning Techniques. *KDD-1996*: 140-145.
[13] Lakshminarayan K. et al. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, 11: 259-275.
[14] Ling, C.X., Yang, Q., Wang, J. & Zhang, S. (2004). Decision trees with minimal costs. *ACM International Conference Proceeding Series*, ICML 2004.
[15] Little, R.J.A. and Rubin, D.A. (1987). *Statistical analysis with missing data*. New York: John Wiley and Sons.
[16] Little R. and Rubin D. (2002). *Statistical Analysis with Missing Data*. Wiley, 2002.
[17] Magnani, M. (2004). *Techniques for Dealing with Missing Data in Knowledge Discovery Tasks*, (available at http://magnanim.web.cs.unibo.it/index.html).
[18] Pawlak, M. (1993). Kernel classification rules from missing data. *IEEE Transactions on Information Theory*, 39(3): 979-988.
[19] Pearson R.K. (2006). The Problem of Disguised Missing Data. *ACM SIGKDD Explorations Newsletter*, Vol. 8, No. 1: 83 - 92.
[20] Pesonen, E., Eskelinen, M. and Juhola, M., (1998). Treatment of missing data values in a neural network based decision support system for acute abdominal pain. *Artificial Intelligence in Medicine*, 13(3): 139-146.
[21] Qin, Y.S., Zhang, S.C., Zhu, X.F., Zhang, J.L. and Zhang, C.Q. (2007). Semi-parametric Optimization for Missing Data Imputation. *Applied Intelligence*, 27(1): 79-88.
[22] Qin, Z.X. (2007). *Multiple costs and their combination in cost-sensitive learning*. PhD Thesis, University of Technology Sydney, 2007.
[23] Quinlan, J. (1989). Unknown Attribute values in Induction. In *Proc 6th Int workshop on machine learning*: Ithaca, pp 164-168.
[24] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
[25] Ramaswamy, S., Rastogi, R. & Shim, K. (2000). Efficient Algorithms for Mining Outliers from Large Data Sets. *Proceedings of the ACM SIGMOD Conference on Management of Data*, Dallas, TX, 427–438.
[26] Ramoni, M. and Sebastiani, P. (2001). Robust Learning with Missing Data. *Machine Learning*, 2001, Vol. 45(2): 147-170.
[27] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, Wiley: New York, 1987.
[28] Wilks, S. (1932). Moments and distributions of estimates of population parameters from fragments samples. *Ann. Math. Statist.* 3: 163-203.
[29] Wang, Q.H. and Rao, R.N.K. Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist*. 30: 896-924.
[30] Yuan Y.C. (2001). Multiple imputation for missing data: concepts and new development SAS/STAT 8.2. (Available at http://www.sas.com/statistics) SAS Institute Inc. Cary, NC.
[31] Zhang, SC, Qin, YS, Zhu, XF, Zhang, JL, and Zhang, CQ. (2006). Optimized Parameters for Missing Data Imputation. *PRICAI06*, 2006: 1010-1016.
[32] Zhang, S.C., Qin, Z.X., Sheng, S.L. and Ling, C.L. (2005). "missing is useful": Missing values in cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17 No. 12: 1689-1693.
[33] Zhang, S.C., Qin, Y.S., Zhang, J.L., Zhu, X.F., Zhang, C.Q. (2008). Missing Value Imputation Based on Data Clustering. *Transactions on Computational Science Journal*, LNCS 4750, pp 128-138.

Shichao Zhang: Shichao Zhang is a professor and the dean of College of Computer Science and Information Technology at the Guangxi Normal University, Guilin, China. He holds a PhD degree in Computer Science from Deakin University, Australia. His research interests include data analysis and smart pattern discovery. He has published about 50 international journal papers, including 7 in IEEE/ACM Transactions, 2 in Information Systems, 6 in IEEE magazines; and over 40 international conference papers, including 3 AAAI, 2 ICML, 1 KDD, and 1 ICDM papers. He has won 6 China NSF/863/973 grants, 1 Overseas-Returning High-level Talent Research Program of China Hunan-Resource Ministry, 3 Australian large ARC grants. He is a senior member of the IEEE; a member of the ACM; and serving as an associate editor for IEEE Transactions on Knowledge and Data Engineering, Knowledge and Information Systems, and IEEE Intelligent Informatics Bulletin.

Computational Methods of Feature Selection

HUAN LIU AND HIROSHI MOTODA



REVIEWED BY LONGBING CAO AND DAVID TANIAR

Feature selection selects a subset of relevant features, and also removes irrelevant and redundant features from the data to build robust learning models. Feature selection is very important, not only because of the curse of dimensionality, but also due to emerging data complexities and quantities faced by multiple disciplines, such as machine learning, data mining, pattern recognition, statistics, bioinformatics, and text mining.

In recent years, we have seen extensive and productive efforts on feature selection. The research has been expanding from simple to complex feature types, from supervised to unsupervised and semi-supervised feature selection, and from simple to more advanced techniques, both in depth and in breadth.

"Computational Methods of Feature Selection", edited by H. Liu and H. Motoda, two leading experts in the field, collects recent research works from various disciplines on computational methods in feature selection and extraction. The collection reflects the advancements in recent years, following the editors' pioneer book on feature selection published in 1998. Consequently, these publications provide a comprehensive roadmap of feature selection research, and this is certainly very helpful to a very wide audience from beginners to professionals, and from practitioners to researchers.

The collection features a state-of-the-art survey, technique advancement, practical guides, promising directions, and case studies. It ranges from presenting background and fundamentals relevant to feature selection, to recent results in extending feature selection, weighting and local methods for feature selection, as well as feature selection progress in text mining and bioinformatics, organized in five independent parts. The content, carefully selected from invited contributors, relevant workshops and tutorials, covers areas such as text classification, web mining, bioinformatics and high-dimensional data.

The book starts with an introduction and background material, which consists of four chapters. A very insightful and enjoyable overview on background and the basics of feature selection are presented in Chapter 1. An evolutionary picture is drawn on feature selection development from supervised to unsupervised and semi-supervised learning in order to handle the increasing mixture of labeled, unlabeled and partially labeled data. An overview of unsupervised feature selection is presented in Chapter 2, which highlights the identification of the smallest feature subset that best uncovers interesting and natural clusters based on certain criteria. Filter and wrapper methods are introduced to select features in unlabeled data, while subspace clustering and co-clustering/bi-clustering are discussed from the local unsupervised feature selection perspective.

Randomization is widely used in feature selection when appropriate choices can be managed. A survey on randomized feature selection is presented in Chapter 3. Two types of randomization, namely Las Vegas and Monte Carlo algorithms, are introduced, followed by an overview of three complexity classes defining the probabilistic requirements in analyzing randomized algorithms. The work features six illustrations of randomized features and prototype algorithms for feature selection problems. Another factor that may facilitate feature selection is causal relationship discovery for cutting down dimensionality and deep understanding of the underlying mechanism. Chapter 4 addresses non-causal and causal feature selection. With a definition of probabilistic causality, the causal Bayesian network is used to analyze feature relevance and further to design a causal discovery algorithm by finding the Markov blanket.

Recent advancements in feature selection are highlighted in Part II in a number of strategies. Firstly, Chapter 5 describes how an active feature value is acquired to estimate feature relevance in domains where feature values are expensive to measure. A sampling benefit function is derived from a statistical formulation of the problem, followed by an active sampling algorithm, which is shown to outperform random sampling by a mixture model for the joint class-feature distribution to reduce the number of feature samples. Secondly, from the feature extraction perspective, Chapter 6 presents the notion of the decision border, in which a labeled vector quantizer, that can efficiently be trained by the Bayes risk weighted vector quantization (BVQ) algorithm, is devised to extract the best linear approximation. It is shown that the approach gives comparable results to the SVM-based decision boundary and performs better than the multi-layer perceptron-based method.

Chapter 7 further explains how independent probe variables are used in the same distribution in generating a probe. Feature relevance is compared with the relevance of its randomly per-

muted probes for classification using random forests. The approach is promising in terms of data types and quantity, performance, and computational complexity. Finally, in Chapter 8, an incremental ranked usefulness is used to decide whether or not a feature is relevant in massive data, and then to select the best non-consecutive features from the ranking. The approach chooses a small subset of features with similar predictive performance to others in dealing with high-dimensional data.

Strategies and methods related to weighting and local methods are addressed in Part III. Firstly, the Relief family algorithms are described in Chapter 9. Relief is extended to a more realistic variant ReliefF to deal with incomplete data for classification, and is further extended to the Regressional ReliefF for regression problems. The variety of the Relief family shows its general applicability as a non-myopic feature quality measure. Feature selection in K-means is usually not automated. Chapter 10 proposes techniques to automatically determine the important features in K-means clustering. This is done through calculating the sum of the within-cluster dispersions of the feature, and renewing the weights in an iterative process.

In contrast to maximum benefit-based active feature sampling, Chapter 11 focuses on local feature relevance and weighting by designing adaptive metrics or parameter estimates that are local in an input space. Chapter 12 presents a mathematical interpretation of the Relief algorithms. It is proven to be equivalent to solving an online convex optimization problem with a margin-based objective function. New feature weighting algorithms are then proposed to find the nearest neighbor classifier.

In Part IV, text feature selection is addressed by a survey, a new feature selection score, and constraint-guided and aggressive feature selection approaches. Firstly, Chapter 13 presents a comprehensive overview of feature selection for text classification, including feature generation, representation, and selection, with illustrative examples, from a pragmatic viewpoint. Text feature generators, such as word merging, word phrases, character N-grams, and multi-field records are introduced. An introduction to classification feature filtering is also provided. Secondly, Chapter 14 introduces a new feature selection score, namely posterior inclusion probability under Bernoulli and Poisson distributions. The score is defined as the posterior probability of including a given feature over all possible models, in which each model corresponds to a different set of features that includes the given feature. The advantage of the score is that the selected features are easy to interpret while maintaining comparable performance to other typical score metrics, such as information gain.

Two different pairwise constraint-guided dimensionality reduction approaches, through projecting data into a lower space and co-clustering of features and data, are introduced in Chapter 15. Investigations are also conducted on improving semi-supervised clustering performance in high-dimensional data. In Chapter 16, an aggressive feature selection method is proposed, which can filter more than 95% features for text mining. To handle feature redundancy, information gain-based ranking for text classification is also proposed using a mutual information measure and inclusion index.

The last section covers feature selection in bioinformatic data, which may not be effectively handled by general feature selection approaches. This part consists of four chapters. Chapter 17 introduces the challenges of micro-array data analysis and presents a redundancy-based feature selection algorithm. A Markov blanket based filter method is proposed to approximate the selection of discriminative and non-redundant genes. In Chapter 18, a scalable method based on sequence components and domain knowledge is developed to generate automatic features on biological sequence data. The algorithm can construct features, explore the space of possible features, and identify the most useful ones. Chapter 19 proposes an ensemble-based method to find robust features for biomarker discovery. Ensembles are obtained by choosing different alternatives at each stage of data mining from normalization to binning, feature selection, and classification. Finally, a penalty-based feature selection method is proposed in Chapter 20 to produce a sparse model by utilizing the grouping effect. As a generalization of a penalized least squares method, lasso, the proposed approach is promising in handling high-dimensional data for various purposes, such as regression and classification problems.

Overall, we enjoyed reading this book. It presents state-of-the-art guidance and tutorials on methodologies and algorithms in computational methods in feature selection. Enhanced by the editors insights, and based on previous work by these leading experts in the field, the book forms another milestone of relevant research and development in feature selection. The selected chapters also present interesting open issues and promising directions for further exploration of feature selection in the next decade. With such a research roadmap, it is highly exciting to foresee the next generation of feature selection methodologies and techniques inspired by this collection.

About the reviewers:

LONGBING CAO: Data Sciences and Knowledge Discovery Laboratory, Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia. Contact him at lbcao@it.uts.edu.au, and http://www-staff.it.uts.edu.au/∼lbcao/ and datamining.it.uts.edu.au for more information.

DAVID TANIAR: Clayton School of Information Technology, Monash University, Australia. Contact him at David.Taniar@infotech.monash.edu.au, and http://users.monash.edu.au/∼dtaniar/ for more information.

RELATED CONFERENCES, CALL FOR PAPERS/PARTICIPANTS

TCII Sponsored Conferences

.

WI 2009
The 2009 IEEE/WIC/ACM International Conference on Web Intelligence
Milan, Italy
September 15-18, 2009

Web Intelligence (WI) is a new research paradigm aimed at exploring the fundamental interactions between AI-engineering and advanced Information Technology (AIT) on the next generation of Web systems, services, etc. Here AI-engineering is a general term that refers to a new area, slightly beyond traditional AI: brain informatics, human level AI, intelligent agents, social network intelligence and classical areas, such as knowledge engineering, representation, planning, and discovery and data mining are examples. AIT includes wireless networks, ubiquitous devices, social networks, and data/knowledge grids, as well as cloud computing. WI research seeks to explore the most critical technology and engineering to bring in the next generation Web systems. The 2009 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2009) will be jointly held with the 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology. It will be organized by the University of Milano Bicocca, Milano, Italy, and it will be sponsored by the IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), the Web Intelligence Consortium (WIC), and ACM-SIGART.

WI 2009 is planned to provide a leading international forum for researchers and practitioners (1) to present the state-of-the-art of WI technologies; (2) to examine performance characteristics of various approaches in Web-based intelligent information technology; and (3) to cross-fertilize ideas on the development of Web-based intelligent information systems among different domains. By idea-sharing and discussions on the underlying foundations and the enabling technologies of Web intelligence, WI 2009 will capture current important developments of new models, new methodologies and new tools for building a variety of embodiments of Web-based intelligent information systems. A doctoral mentoring program will be also organized.

IAT 2009
The 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology
Milan, Italy
September 15-18, 2009

The 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2009) will be jointly held with the 2009 IEEE/WIC/ACM International Conference on Web Intelligence at the University of Milano Bicocca, Milano, Italy, and it will be sponsored by the IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), the Web Intelligence Consortium (WIC), and ACM-SIGART.

IAT 2009 will provide a leading international forum to bring together researchers and practitioners from diverse fields, such as computer science, information technology, business, education, human factors, systems engineering, and robotics, to (1) examine the design principles and performance characteristics of various approaches in intelligent agent technology, and (2) increase the cross fertilization of ideas on the development of autonomous agents and multi-agent systems among different domains. By encouraging idea-sharing and discussions on the underlying logical, cognitive, physical, and sociological foundations as well as the enabling technologies of intelligent agents, IAT 2007 will foster the development of novel paradigms and advanced solutions in agent based computing. The joint organization of IAT 2009 and WI 2009 will provide an opportunity for technical collaboration beyond the two distinct research communities. A doctoral mentoring program will be also organized.

ICDM'09
The Ninth IEEE International Conference on Data Mining
Venice, Italy
October 28-30, 2009

The International Conference on Data Mining (ICDM 2009) aims to bring together researchers, scientists, engineers, and scholar students to exchange and share their experiences, new ideas, and research results about all aspects of Data Mining, and discuss the practical challenges encountered and the solutions adopted.

All full paper submissions will be peer reviewed and evaluated based on originality, technical and/or research content/depth, correctness, relevance to conference, contributions, and readability. The full paper submissions will be chosen based on technical merit, interest, applicability, and how well they fit a coherent and balanced technical program. The accepted full papers will be published in the refereed conference proceedings. Prospective authors are kindly invited to submit full text papers including results, tables, figures and references. Full text papers (.doc, .rft, .ps, .pdf) will be accepted only by electronic submission.

You are cordially invited to submit a paper and/or a proposal to organize a workshop and actively participate in this conference. Proposals are invited for workshops to be affiliated with the conference scope and topics. The conference seeks proposals for workshops on foundational and emerging topics in areas relevant to Data Mining. The conference workshops provide a challenging forum and vibrant opportunity for researchers and industry practitioners to share their research positions, original research results and practical development experiences on specific new challenges and emerging issues. The workshop topics should be focused so that the Participants can benefit from interaction with each other and the cohesiveness of the topics.

The refereed conference proceedings will be published prior to the conference in both Hard Copy Book and CD-ROM, and distributed to all registered participants at the conference. The refereed conference proceedings are reviewed and indexed by Open Science Index (OSI), Google Scholar, Directory of Open Access Journals (DOAJ),

EBSCO, Ulrich's Periodicals Directory, German National Library of Science and Technology and University Library Hannover (TIB/UB), Electronic Journals Library (Elektronische Zeitschriftenbibliothek, EZB), Genamics, GALE and INTUTE.

ICDM 2009 has teamed up with the International Journal of Computational Intelligence (IJCI) for publishing a Special Journal Issue on Advances in Data Mining. All submitted papers will have opportunities for consideration for this Special Journal Issue. The selection will be carried out during the review process as well as at the conference presentation stage. Submitted papers must not be under consideration by any other journal or publication. The final decision will be made based on peer review reports by the guest editors and the Editor-in-Chief jointly.

Related Conferences

AAMAS'09
The Eighth International Conference on Autonomous Agents and Multi-Agent Systems
Budapest, Hungary
May 10-15, 2009
http://www.conferences.hu/AAMAS2009/

AAMAS is the leading scientific conference for research in autonomous agents and multi-agent systems. The AAMAS conference series was initiated in 2002 as a merger of three highly respected individual conferences: the International Conference in Autonomous Agents, the International Workshop on Agent Theories, Architectures, and Languages, and the International Conference on Multi-Agent Systems. The aim of the joint conference is to provide a single, high-profile, internationally respected archival forum for research in all aspects of the theory and practice of autonomous agents and multi-agent systems.

AAMAS 2009 is the Eighth conference in the AAMAS series, following enormously successful previous conferences at Bologna, Italy (2002), Melbourne, Australia (2003), New York, USA (2004), Utrecht, The Netherlands (2005), Hakodate, Japan (2006), Honolulu, USA (2007) and Estoril, Portugal (2008). AAMAS-09 will be held at the Europa Congress Center, Budapest, Hungary.

AAMAS 2009 encourages the submission of *original* papers covering theoretical, experimental, methodological, and application issues in autonomous agents and multiagent systems. Authors are discouraged to submit papers describing work that has already been published in previous AAMAS workshops with post-proceedings publications and/or published as short papers in previous AAMAS proceedings, unless that the authors *clearly* demonstrate significant new content with respect to the previous publication.

ISWC'09
The Eighth International Semantic Web Conference
Washington, USA
October 25-29, 2009
http://iswc2009.semanticweb.org/

ISWC is a major international forum where visionary and state-of-the-art research of all aspects of the Semantic Web are presented. ISWC'06 follows the 1st International Semantic Web Conference (ISWC'02 which was held in Sardinia, Italy, 9-12 June 2002), the 2nd International Semantic Web Conference (ISWC'03 which was held in Florida, USA, 20 - 23 October 2003), 3rd International Semantic Web Conference (ISWC'04 which was held in Hiroshima, Japan, 7 - 11 November 2004), 4th International Semantic Web Conference 2005 (ISWC'05 which was held in Galway, Ireland, 6 - 10 November, 2005), 5th (ISWC'06 which was held in Athens, GA, USA 5 - 9 November, 2006), 6th (ISWC'07 which was held in Busan, Korea 11 - 15 November, 2007), and 7th (ISWC'08 which was held in Karlsruhe, Germany).

SDM'09
2009 SIAM International Conference on Data Mining
Sparks, Nevada, USA
April 30-May 2, 2009
http://www.siam.org/meetings/sdm09/

Data mining and knowledge discovery is rapidly becoming an important tool in all walks of human endeavor including science, engineering, industrial processes, healthcare, business, medicine and society. The datasets in these fields are large, complex, and often noisy. Extracting knowledge requires the use of sophisticated, high-performance and principled analysis techniques and algorithms, based on sound statistical foundations. These techniques in turn require powerful visualization technologies; implementations that must be carefully tuned for performance; software systems that are usable by scientists, engineers, and physicians as well as researchers; and infrastructures that support them. For the main conference the program committee seeks outstanding papers in all areas pertaining to data mining and knowledge discovery.

This conference provides a venue for researchers who are addressing these problems to present their work in a peer-reviewed forum. It also provides an ideal setting for graduate students and others new to the field to learn about cutting-edge research by hearing outstanding invited speakers and attending tutorials (included with conference registration). A set of focused workshops are also held on the last day of the conference. The proceedings of the conference are published in archival form, and are also made available on the SIAM web site.

IAAI'09
The Twenty-First Innovative Applications of Artificial Intelligence Conference
Pasadena, California, USA
July 14–16, 2009
http://www.aaai.org/Conferences/IAAI/iaai09.php

The Twenty-First Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-09) will focus on successful applications of AI technology. The conference will use technical papers, invited talks, and panel discussions to explore issues, methods, and lessons learned in the development and deployment of AI applications; and to promote an interchange of ideas between basic and applied AI.

IAAI-09 will consider papers in two tracks: (1) emerging applications or methodologies and (2) deployed application case studies. Submissions should clearly identify which track they are intended for, as the two tracks are judged on different criteria. Applications are defined as deployed once they are in production use by their final end users (not the people who created the application) for sufficiently long that experience can be reported (usually greater than three months of use by the end-users). All submissions must be original.

IJCAI'09
The Twenty-First International Joint Conference on Artificial Intelligence
Pasadena, California, USA
July 11–17, 2009
http://ijcai-09.org/index.html

The IJCAI-09 Program Committee invites submissions of technical papers for IJCAI-09, to be held in Pasadena, CA, USA, July 11-17, 2009. Submissions are invited on significant, original, and previously unpublished research on all aspects of artificial intelligence.

The theme of IJCAI-09 is "The Interdisciplinary Reach of Artificial Intelligence," with a focus on the broad impact of artificial intelligence on science, engineering, medicine, social sciences, arts and humanities. The conference will include panel discussions, invited talks and other events dedicated to this theme.

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903

Non-profit Org.
U.S. Postage
PAID
Silver Spring, MD
Permit 1398