THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

December 2010 Vol. 11 No. 1 (ISSN 1727-5997)

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 at http://computer.org/tcsignup/.

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).*

Message from the Editors

In August 2010, I (Vijay) became a member of the Executive Committee of IEEE TC on Intelligent Informatics (TCII). As my primary responsibility as a new member of the EC, I am happy to serve the community as the Editor-in-Chief of the IEEE Intelligent Informatics Bulletin. The research specializations of the TCII members span multiple cognitive and intelligent paradigms, such as knowledge engineering, artificial neural networks, fuzzy logic, evolutionary computing, and rough sets. The TCII Bulletin provides opportunities to the community to communicate with each other regarding their on-going research and professional activities, and to share its experiences for the benefit of all members.

This issue of the TCII Bulletin has four feature articles and one book review. The article by G. Pasi elaborates on the excellent presentation she made recently on preference-modeling and personalization in information retrieval at the WI-IAT 2010 conference, held in Toronto. C. R. Milar´e et al. investigate the influence of rule measures on the performance of evolutionary algorithms used to induce classifiers. D. Yuan et al. discuss methods of identifying interactions among items having large relative-contrasts. Finally, the article by S. Zhang explores an approach for incorporating a certainty factor into the kNN classification strategy. In the Book Review section, N. Hussain and H. Zhou review the 2010 book by Longbing Cao on the exciting topic of Domain-driven Data Mining.

The Bulletin is the result of hard work by the members of the Editorial Board. We express our gratitude to their efforts in working closely with the contributing authors to get this issue out in a timely fashion. But without the indulgence of all of you, the TCII members, it is difficult to ensure that the quality feature articles and other materials in the bulletin meet the Bulletin's goals of being an effective and sought after medium of communication. I strongly encourage your involvement and welcome your suggestions for keeping its contents vibrant and relevant.

The editors look forward to working with many of you. We are also excited to announce that the Editorial Board will be ably assisted by Dr. Xin Li who has graciously agreed to serve in the capacity of the Assistant Managing Editor.

Vijay V. Raghavan
(University of Louisiana at Lafayette)
Editor-in-Chief
William K. Cheung
(Hong Kong Baptist University, Hong Kong)
Managing Editor

Issues in Personalizing Information Retrieval

Gabriella Pasi

***Abstract*—This paper shortly discusses the main issues related to the problem of personalizing search. To overcome the "one size fits all" behavior of most search engines and Information Retrieval Systems, in recent years a great deal of research has addressed the problem of defining techniques aimed at tailoring the search outcome to the user context. This paper outlines the main issues related to the two basic problems beyond these approaches: context representation and definition of processes which exploit the context knowledge to improve the quality of the search outcome. Moreover some other important and related issues are mentioned, such as privacy, and evaluation.**

***Index Terms* — Information Retrieval, Personalization, Context Modeling, User Modeling.**

I. Introduction

IN recent years there has been an increasing research interest in the problem of contextualizing search to the aim of overcoming the limitations of the "one size fits all" paradigm, which is generally applied by Search Engines and Information Retrieval Systems (IRSs). By this paradigm the keyword-based query is considered as the only carrier of the users' information needs. As a consequence, the relevance estimate is system-centered, as the user context is not taken into account. Instead, a contextual Search Engine or IRS relies on a user-centered approach since it involves processes, techniques and algorithms that exploit as much contextual factors as possible in order to tailor the search results to users [6,14,19,27,28,37].

As it will be shown in section II, the key notion of context may have multiple interpretations in Information Retrieval (IR). It may be related to the characteristics and preferences of a specific user or group of users (in this case contextualization can be referred to as personalization), or it may be related to user geographic localization (when for example using a search engine on a smart-phone), or it may refer to the information that qualifies the content of a given document/web page (for example its author, its creation date, its format etc.).
The development and increasing use of tools that either help users to express their topical preferences, or automatically learn them, and the availability of devices and technologies that can detect both users' location (such as GPSs) and monitor users' actions, allow to capture the user's context, related to the considered interpretation or application in the attempt to contextualize search.

To the aim of modeling contextualized IR applications, a significant amount of research has addressed two main problems: how to model the user's context, and how to exploit it in the retrieval process in order to provide context-aware results. Several research works have offered possible solutions to the above problems, related to the considered interpretation of context, giving birth to some specific IR branches such as personalized IR, mobile IR, social IR. Although the specific techniques related to these branches vary (due to the nature of context that needs to be modelled), the common issue of context-based IR is to improve the quality of search by proposing to the user results tailored to the considered context.

In this paper a synthetic overview of some main issues in designing personalized approaches to Information Retrieval is presented. In section II the shift from the system centered approach to the user and context centered approach in IR is discussed. Section III aims at reporting on the issue of defining a formal user model; in section IV the approaches proposed in the literature to exploit the user context in search are classified and shortly described. Finally in section V the important issues of privacy and personalized systems evaluation are discussed.

II. From the System Centered Approach to a User Centered Approach to IR

Most Information Retrieval Systems and Search Engines rely on the so called system-centered approach, where the IRS behaves as a black box, which produces the same answer to the same query, independently on the user context. The notion of context in IR is well described in [36], and it may have several interpretations, ranging from user context (the central notion in context-based IR), to document context, spatio-temporal context, social context, etc. The identification of a specific context allows to identify information that can be usefully exploited to the aim of improving search effectiveness. For example, by user context we generally refer to the information characterizing a person (personal information) and his/her preferences. The personal information may include demographic and professional data; preferences of a person may range from topical preferences, taste preferences, etc. The spatio-temporal context is identified by information such as location, geographic coordinates etc.

If properly acquired, organized, and stored, the context-related information may be used to leverage the process aimed at identifying information relevant to a user need, beyond the mere usage of the user's query. To this aim a context model must be defined by a formal language, which is used to represent the information related to the context.

Gabriella Pasi is with the Università degli Studi di Milano Bicocca, Department of Informatics, Systems and Communication, Milano, Italy (phone: +39-02-64487847; e-mail: pasi@ disco.unimib.it).

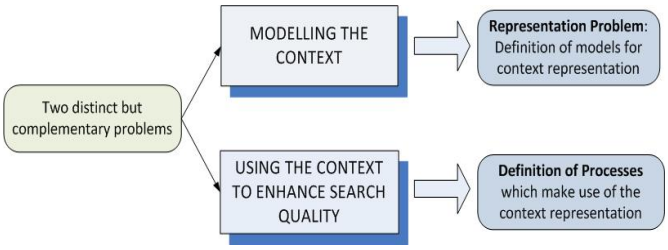


Fig. 1. The main processes involved in personalized IR.

Context-centered IR is an expression which can be used to encompass all tools, techniques and algorithms finalized at producing a search outcome (in response to a user's query), which is tailored to the specific context. This way the "one size fits all" approach is no more valid. When context is referred to the user context, we may talk about personalized IR.

The previous short introduction to the notion of context and its possible use in IR makes it evident that in order to implement a context dependent IR strategy, two main activities must be undertaken, as sketched in Fig.1. The prerequisite activity is of type knowledge representation, and is aimed at the definition of the context model. Such an activity comprises sub-activities such as the identification of the basic knowledge which characterizes the context, the choice of a formal language by which to represent this knowledge, and a strategy to update this knowledge (to adapt the representation to context variations). The second activity is aimed at defining processes (algorithms), which, based on both the knowledge represented in the context representation and the user query, are finalized to produce as a search outcome an estimate of document relevance which takes into account the context dimension(s). In other words, the context is used to leverage the effectiveness of the search outcome. As it will be explained in section III this can be done by different approaches, which can be classified depending on the way in which the contextual information is exploited.

While in this section we have introduced a general definition of context, and of context-centered IR, in the following sections we will focus on personalized IR, i.e. to IR approaches which take advantage of the knowledge represented in a user model, also called user's profile.

III. Modeling the User Context in Personalised IR

In recent years, a great deal of research has addressed the problem of personalizing search, to the aim of taking into account the user context in the process of assessing relevance to user's queries. These research efforts are witnessed both by the numerous publications, and by the existence of conference devoted to personalized approaches to IR, or more generally to IR in context (e.g. the Symposium on Information and Interaction in Context, IIiX [43], the International Conference on User Modeling, Adaptation and Personalization [44], the SIGIR Desktop Search Workshop: Understanding, Supporting, and Evaluating Personal Data Search [45]).

Moreover personalized approaches to IR may be experienced by users through personalized versions of search engines, such as iGoogle, Google Personalized Search (www.google.com/ig).

To personalize search results means to explicitly make use of the user preferences to tailor search results. If for example a query such as "good restaurant in Rome" is formulated by a vegetarian user, the expected results should take this preference into account. To this aim the query evaluation should make explicit use of this information as an additional constraint (besides the query) to estimate document relevance. As another example let us consider a group of users represented by researchers working in an information retrieval lab. If the query "information retrieval" is formulated by the lab director, the query evaluation should produce a different list of documents than the same query formulated by a novice student. In this last example, the user preferences are related to his/her cognitive context, and expertise. The previous simple examples outline that the quality of the search outcome strongly depends on the information beyond the one expressed in a user's query. So the effectiveness of the system strongly depends on the available quantity and quality of information about the user and its preferences. The more accurate the user model is, the more effective the personalized answer can be.

An obvious question rises at this point: how to make this information available to an IR system? To do so three kind of processes should be undertaken: acquisition, representation and updating. The acquisition process is aimed at capturing the information characterizing the user context. The formal representation process is aimed at formally representing the acquired information; this is needed to make it possible that this information be accessed and used by the IRS. The updating process is finalized at learning the changes of the user preferences in time. In the following we shortly discuss each of the above processes. It is clear that the effectiveness of the algorithms which exploit the knowledge of the user context strongly depend on the quality and reliability of the user model (user profile). So the generation of a user model is an important although difficult task.

To capture user's interest two main techniques may be employed: explicit and implicit [17,28]. By the explicit approach the user is asked to be proactive and to directly communicate to the system his/her data and preferences. This can be done by compiling questionnaires, by providing short textual descriptions (to specify topical preferences), and/or by providing a few documents that represents well the user preferences. The texts will be processed by the system to automatically extract their main descriptors. However, an explicit request of information to the user implies to burden the user, and to rely on the user's willingness to specify the required information. This is generally unrealistic. To overcome this problem, several techniques have been proposed in the literature to automatically capture the user's interests, by monitoring the user's actions in the user system interaction, and by inferring from them the user's preferences. The proposed techniques range from click-trough data analysis, query log analysis, desktop information analysis, document display time,

etc [1,11,22,23,30,31,34,37]. The advantage in adopting such techniques is that several sources of knowledge may be considered; the main disadvantage is that automatic processes may be error-prone, as they may introduce noise in the process of identifying the useful information. However, the advantages of using such techniques has revealed much greater than their limitations.

The process of organization and representation of the information obtained by the acquisition phase implies the selection of an appropriate formal language to define the user model. In the literature several representations for the user model have been proposed, ranging from bag of words and vector representations, to graph-based representations, and, more recently, to ontology based representations [8,11,14,16,32,35,40]. The more structured and expressive the formal language is, the more accurate the user model can be. As most current approaches to the definition of user profiles are aimed at defining models based on words or concept features, to the aim of also representing the relations between words/concepts, an external knowledge resource, such as the ODP (Open Directory Project [46], or Wordnet [47]) is required.

An important aspect related to user profiles concerns profile updating; this aspect is generally considered by the research contributions that propose the definition of user models.

IV. Exploiting the user context to enhance search quality

As outlined in section III, the availability of a user model, where the relevant information that characterizes the user context is represented, is necessary to define any process aimed at tailoring, based on this context description, the results proposed as an answer to a user's query. The quality of the personalization process is strongly related to the quality of the user's model, e.g. to its reliability and accuracy.

In the literature several approaches have been proposed, which can be roughly categorized in three main classes [28]:

- approaches to modify/define relevance assessment
- approaches to query modification
- approaches to results re-ranking

Among the approaches belonging to the first category we cite the PageRank based methods, which have proposed modifications of the PageRank algorithm that include user modeling into rank computation, to create personal views of the Web [18,21].

The approaches to results re-ranking are aimed at modifying the ranking score by explicitly matching the user profile against the user query, and then at combining the obtained score with the relevance based score produced by the traditional IRS or search engine. Re-ranking techniques proposed in the literature may differ both in the adopted user model and in the re-ranking strategy. Among the several techniques proposed in the literature we cite [27,31,32].

Query modification techniques are aimed at exploiting the user profile as a knowledge support to select information useful to define more accurate queries via a query expansion or modification technique. Among the techniques proposed in the literature we cite [9,10,26,37].

More recently an interesting problem has been considered to leverage search through a better user knowledge and interaction: this is the problem of visualizing search results in an effective way. One of the biggest problems when using search engines is that, although the information relevant to the user's needs expressed in a query could be probably found in the long ordered list of results, it is quite difficult to locate it. It is in fact well know that users seldom go beyond an analysis of the first two/three pages of search results. An interesting research idea is to enhance results visualization through the knowledge of the user's topical preferences. In [2,4] an approach related to the exploratory search task is proposed, which combines personalized search with a spatial and adaptive visualization of search results

Independently of the decision about how to exploit the knowledge of the user's preferences, an interesting aspect which emerges in context-aware IR is that the availability of a model of context (which may represent both user's preferences, the geographic and social contexts etc.) makes it possible to consider several new dimensions in the relevance assessment process. The birth of Web Search Engines as well as the IR techniques evolution, have implied a shift from topical relevance assessment (which was the only dimension to assess relevance in the first IRSs) to a multi-dimensional relevance assessment, where the considered relevance dimensions encompass topical relevance, page popularity (based on link analysis in web search engines), geographic and temporal dimensions, etc. The availability of a user model (and more generally the availability of more structured context models), make the dimensions available to concur in the process of relevance assessment more numerous. As a consequence, the need of combining the relevance assessments related to each dimensions arise. This problem has been faced so far by adopting simple linear combination schemes, applied independently on the user's preferences over the relevance dimensions. An interesting research direction related to personalized search is to make the user an active player in determining such an aggregation scheme: this could be simply done by making the aggregation dependent on the user's preferences over the single relevance dimensions. In this way, for a same query and a same profile different document rankings can be obtained based on the user's preference over the relevance dimensions. In [12] this approach has been proposed to define user-dependent aggregation schemes defined as linear combinations where weights of relevance dimensions are automatically computed based on the user-specified priority order over the dimensions.

V. The privacy and the evaluation issues

Two important issues that have been addressed in the literature related to personalized IR concern user privacy and the evaluation of the effectiveness of personalized approaches to IR. We start by shortly discussing the privacy issue.

As it has been synthetically discussed in the previous sections, the approaches to personalization strongly rely on user related and user personal information, with the obvious consequent need of preserving users' privacy. In [25,41] very interesting and exhaustive analysis of the privacy issue are presented. As well outlined in these contributions, the user is not incline to make the information that concerns his/her private life available to a centralized system, with the main consequence that often users prefer not to use the personalization facilities. As suggested by the authors, a feasible solution to the privacy issue problem is to design client-side applications.

Systems evaluation is a fundamental activity related to the IR task. The usual approach to evaluate the effectiveness of IRSs is based on the Cranfield paradigm, which is the basic approach undertaken by the TREC (Text Retrieval) Conferences [48]. However, as well outlined in [6], the Cranfield paradigm is not able to accommodate the inherent interaction of users with information systems. The Cranfield evaluation paradigm is in fact based on document relevance assessment on single search results, not suited to interactive information seeking and personalized IR, as it assumes that users are well represented by their queries, and the user's context is ignored. In [36] a good overview of the problem of evaluating the effectiveness of approaches to personalized search is presented. The evaluation of systems that support a personalized access to information encompasses two main aspects, related to the components which play a main role in these systems, i.e. the user model and the personalized search processes. To evaluate a user profile means to assess its quality properties, such as accuracy. With respect to the evaluation of systems' effectiveness, the authors outline in [36] three main approaches undertaken to set up a suited evaluation setting for personalized systems; by the first approach, an attempt to extend the laboratory-based approach to account for the existence of contextual factors were proposed within TREC [5,18]. By the second approach a simulation-based evaluation methodology has been proposed, based on searchers simulations [42]. By the third approach, the one which is most extensively adopted, user-centered evaluations are defined, based on user studies, with the involvement of real users who undertake qualitative system's evaluations [24]. Evaluation is a quite important issue that deserves special attention, and which still needs important efforts to be applied to context-based IR applications.

VI. Conclusions

To conclude this short overview of some main issues related to personalized IR, we want to mention a promising research direction which aims at exploiting the users' social context to produce more effective results in Web Search [33,39]. This is made possible by recent applications and technologies related to the so called Social Web, aimed at making the user active in both content generation and sharing. In [33] an approach to collaborative Web Search has been recently proposed, which based on the search behavior of a community of like-minded users is aimed to adapt results of conventional search engines to the community preferences.

References

[1] E. Agichtein, E. Brill, S. Dumais, R. Ragno, "Learning user interaction models for predicting Web search preferences", in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 3–10.

[2] J. Ahn and P. Brusilovsky, "From User Query to User Model and Back: Adaptive Relevance-Based Visualization for Information Foraging", in *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence*, Silicon Valley, CA, USA, 2007, pp. 706–712.

[3] J. Ahn, P. Brusilovsky, D. He, J. Grady, Q. Li, "Personalized Web Exploration with Task Modles", *in Proc. of the17th international conference on World Wide Web (WWW 2008)*, Bejing, China, 2008, pp. 1-10.

[4] J. Ahn and P. Brusilovsky. "Visual interaction for personalized information retrieval", in Third Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2009), Washington, DC, USA, Oct 2009.

[5] J. Allan J, "Hard track overview in TREC 2003 high accuracy retrieval from documents", in Proc. of the 12th text retrieval conference (TREC-12), National Institute of Standards and Technology, NIST special publication, pp 24–37.

[6] N.J. Belkin, " Some(what) Grand Challenges for Information Retrieval, in *SIGIR Forum*, Vol. 42, No. 1. (2008), pp. 47-54.

[7] P. Brusilovsky, A. Kobsa, W. Nejdl, eds. *The adaptive web: methods and strategies of web personalization*, Springer, LNCS 4321, 2007.

[8] S. Calegari, G. Pasi, "Ontology-Based Information Behaviour to Improve Web Search", in *Future Internet*, 4, 2010, pp. 533-558.

[9] P. Chirita, C. Firan, W. Nejdl, "Summarizing local context to personalize global Web search", in *Proc. of the Annual International Conference on Information and Knowledge Management, CIKM 2006*, pp. 287–296.

[10] P. Chirita, C.Firan, W. Nejdl, "Personalised Query Expansion for the Web", in *Proc. of the ACM SIGIR Conference 2007*, pp. 287–296.

[11] M. Claypool, D. Brown, P. Le, M. Waseda, " Inferring user interest", in *IEEE Intern.Comput.* 32–39, 2001.

[12] C. da Costa Pereira, M. Dragoni, G. Pasi, "Multidimensional Relevance: A New Aggregation Criterion", in Proc. Of the European Conference on Information Retrieval (ECIR 2009), pp. 264-275.

[13] M. Daoud, L. Tamine-Lechani, M. Boughanem, "Towards a graph-based user profile modeling for a session-based personalized search", in *Knowl. Inf. Syst.*, 21 (3), 2009, pp. 365-398.

[14] M. Daoud, L. Tamine, M. Boughanem, "A Personalized Graph-Based Document Ranking Model Using a Semantic User Profile", *in Proc. of UMAP 2010*, pp. 171-182.

[15] Z. Dou, R. Song, J.R. Wen, "A large-scale evaluation and analysis of personalized search strategies", in *Proc. of the International World Wide Web Conference*, 2007, pp. 581–590.

[16] S. Gauch, J. Chaffee, A. Pretschner, "Ontology-Based perosnalized search and browsing", in *Web Intelligence and Agent Systems: an International Journal*, 2003, pp. 219-234.

[17] S. Gauch, M. Speretta, A. Chandramouli, A. Micarelli, "User Profiles for Personalized Information Access", The Adaptive Web 2007, pp. 54-89

[18] T. Havelivala, Topic-sensitive PageRank, ", in Proc. 12th International orld Wide *Web Conference* (WWW 2003), Hawai, 2002.

[19] D. Harman, overview of the 4th text retrieval conference (TREC-4). In *Proc. Of the 4th text retrieval conference (TREC-4)*, National Institute of Standards and Technology, NIST special publication, pp. 1-24.

[20] P. Ingwersen, and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context*, Springer-Verlag Eds, 2005

[21] G. Jeh and J. Widom, "Scaling Personalized Web Search", *in Proc. 12th International World Wide Web Conference* (WWW 2003), Budapest, Hungary, 2003, pp. 271279.

[22] T. Joachims, L. Granka, B. Pang, H. Hembrooke, G. Gay, "Accurately interpreting clickthrough data as implicit feedback", in *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2005, pp. 154–161.

[23] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: A bibliography", in *SIGIR Forum 37*, 2, 18–28, 2003.

[24] D. Kelly, "Methods for evaluating interactive information retrieval systems with users", in Foundations and Trends in Information Retrieval, DOI: 10.1561/15000000123, (1-2), 2009, pp. 1-224..
[25] A. Kobsa, "Privacy Enhanced Personalization", in P. Brusilovsky, A. Kobsa, W. Nejdl, eds. *The Adaptive Web: Methods and Strategies of Web Personalization. Berlin, Heidelberg*, New York, Springer-Verlag, 2007, 628-670.
[26] C Liu, C. Yu, and W. Meng, "Personalized Web Search For Improving Retrieval Effectiveness", in *IEEE Transactions on Knowledge and Data Engineering*, 16(1), January 2004, pp. 28-40.
[27] Z. Ma, G. Pant, and O. Sheng," Interest-based personalized search", in *ACM Transaction on Information Systems, 25*, 5, 2007.
[28] A. Micarelli, F. Gasparetti, F. Sciarrone, S. Gauch, "Personalized Search on theWorld Wide Web", in Lecture Notes in Computer Science, 2007, 4321, pp.195–230.
[29] J. Pitkow, H. Schutze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, "Personalized search", in *Communications of the ACM, 45*, 9, 2002, pp. 50–55.
[30] X. Shen, B. Tan, and C.X. Zhai, "Context sensitive information retrieval using implicit feedback", in *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2005, pp. 43-50.
[31] X. Shen, B. Tan, and C.X. Zhai, "Implicit user modeling for personalized search", in *Proc. of the International Conference on Information and Knowledge Management*, CIKM 2005, pp. 824–831.
[32] A. Sieg, B. Mobasher, R. Burke ,"Web search personalization with ontological user profiles", in *Proc. of the International Conference on Information and Knowledge Management* , CIKM 2007, pp. 525-534.
[33] B. Smyth, "A Community-Based Approach to Personalizing Web Search", in IEEE Computer, 40(8), 2007, pp.42-50.
[34] M. Speretta, S. Gauch: Personalized Search Based on User Search Histories", in *Proc. of the IEEE/WIC/ACM International conference on Web Intelligence, 2005*, pp. 622-628.
[35] M. Speretta, S. Gauch, "Miology: a Web Application for Organizing Personal Domain Ontologies", in *Proc. of the International Conference on Information, Process and Knowledge Management*, 2009, pp. 159-161.
[36] L. Tamine-Lechani, M. Boughanem, M. Daoud, "Evaluation of contextual Information Retrieval effectiveness: overview of issues and research", *Knowledge Information Systems*, 24, 2010, pp. 1-34.
[37] J. Teevan, S. Dumais, and E. Horvitz, " Personalizing search via automated analysis of interests and activities", in *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2005, pp. 449–456.
[38] J. Teevan, S.T. Dumais, and D. J. Liebling, "Personalize or not to personalize: Modeling queries with variation in user intent", in *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2008, pp. 163–170.
[39] J. Teevan, M. Morris, S. Bush, "Discovering and using groups to improve personalized search". *In Proceedings of the ACM International Conference on Web Search and Data Mining*, 2009, pp. 15–24.
[40] J. Trajkova, S. Gauch, "Improving Ontology-Based User Profiles", in *Proc. of the. RIAO 2004 Conference*, pp. 380-390
[41] E. Volokh, "Personalization and Privacy", in *Communications of the ACM*, 43(8), 2000.
[42] R. W. White, I.Ruthven, J. Jose, C. van Rijsbergen, "Evaluating implicit feedback models using seracher simulations", in *ACM Transactions on Information Systems*, 22(3), 2005, pp. 325-361.
[43] http://www.iva.dk/IIiX/
[44] http://www.hawaii.edu/UMAP2010/
[45] http://www.cdvp.dcu.ie/DS2010/
[46] http://dmoz.org/
[47] http://wordnet.princeton.edu/
[48] http://trec.nist.gov/

A Study of the Influence of Rule Measures in Classifiers Induced by Evolutionary Algorithms

Claudia Regina Milaré, Gustavo E.A.P.A. Batista, André C.P.L.F. de Carvalho

***Abstract*—The Pittsburgh representation is a well-known encoding for symbolic classifiers in evolutionary algorithms, where each individual represents one symbolic classifier, and each symbolic classifier is composed by a rule set. These rule sets can be interpreted as *ordered* or *unordered* sets. The major difference between these two approaches is whether rule ordering defines a rule precedence relationship or not. Although ordered rule sets are simple to implement in a computer system, the rule set is difficult to be interpreted by human domain experts, since rules are not independent from each other. In contrast, unordered rule sets are more flexible regarding their interpretation. Rules are independent from each other and can be individually presented to a human domain expert. However, the algorithm to decide a classification of a given example is more complex. As rules have no precedence, an example should be presented to all rules at once and some criteria should be established to decide the final classification based on all fired rules. A simple approach to decide which rule should provide the final classification is to select the rule that has the best rating according to a chosen quality measure. Dozens of measures were proposed in literature; however, it is not clear whether any of them would provide a better classification performance. This work performs a comparative study of rule performance measures for unordered symbolic classifiers induced by evolutionary algorithms. We compare 9 rule quality measures in 10 data sets. Our experiments point out that confidence (also known as precision) presented the best mean results, although most of the rule quality measures presented approximated classification performance assessed with the area under the ROC curve (AUC).**

***Index Terms*—Symbolic classification, evolutionary algorithm, rule quality measures.**

I. Introduction

Evolutionary Algorithms (EAs) have been successfully applied to solve problems in a large number of domains. One of their most prominent features is to perform a global search using multiple candidate solutions, and therefore increasing the possibilities of finding an optimal solution [1]. In contrast, induction of symbolic classifiers can be seen as a search problem in which some performance measure should be optimized, such as accuracy or coverage. Conventional symbolic inducers, for instance decision tree inducers, use a simple greedy search, and EAs present an attractive alternative to better search the hypothesis space.

The Pittsburgh representation [2] is a well-known encoding for symbolic classifiers in EAs, where each individual represents one symbolic classifier, and each symbolic classifier is composed by a rule set. These rule sets can be interpreted as *ordered* or *unordered* sets. In the case of an ordered rule set, rule ordering defines a precedence relationship. For instance, when an example to be classified is presented to an ordered rule set, rules must be analyzed regarding their position in the set. The final classification is given by the class predicted by the *first* rule that covers the example. Although this approach is very simple to implement in a computer system, the rule set is difficult to be interpreted by human domain experts, since rules are not independent from each other. The knowledge expressed in a rule only holds if all preceding rules were not fired.

In contrast, unordered rule sets are more flexible regarding their interpretation. Rules are independent from each other and can be individually presented to a human domain expert. However, the algorithm to decide the classification of a given example is more complex. As rules have no precedence, an example should be presented to all rules at once and some criteria should be established to decide the final classification based on all fired rules. A simple approach to decide which rule should provide the final classification is to select the rule that has the best rating according to a chosen quality measure. Dozens of these measures were proposed in literature. However, it is not clear whether any of them would provide a better classification performance.

This work performs a comparative study of rule performance measures for unordered symbolic classifiers induced by EAs. We compare 9 rule quality measures in 10 data sets. In our experiments, we use the area under the ROC curve (AUC) [3] as the main measure to assess our results. AUC has several advantages over other conventional measure such as error rate and accuracy [4], for instance, AUC is independent from class prior probabilities. Our experiments point out that confidence (also known as precision) presented the best mean results, although most of the rule quality measures presented approximated classification performance.

This paper is organized as follows: Section II describes our EA; Section III presents the rule quality measures used in our experiments; Section IV empirically compares the measures on 10 application domains; and finally, Section V concludes this paper and presents some directions for future work.

II. Our Evolutionary Algorithm

There are two major approaches to represent decision rules as individuals in an EA. These approaches are namely Michigan and Pittsburgh [5]. In short, in Michigan [6] approach each individual codifies only one rule, and in Pittsburgh [2]

C. R. Milaré is with Centro Universitário das Faculdades Associadas de Ensino, São João da Boa Vista, Brazil, e-mail: cmilare@gmail.com.

G. E. A. P. A. Batista and A. C. P. L. F. de Carvalho are with Instituto de Ciências Matemáticas e de Computação – ICMC, Universidade de São Paulo – USP, São Carlos - SP, Brazil, e-mail: {gbatista, andre}@icmc.usp.br.

approach each individual codifies a classifiers, i.e. a rule set. This difference is more than a simple technical detail. Michigan approach is used when we are interested in a single rule with a determined propriety, such as accuracy. Even though the final population has several rules, those rules usually do not have a collective property, such as complementary coverage. Therefore, Michigan approach is frequently used to induce descriptive rules.

In contrast, as Pittsburgh representation codifies a rule set in each individual, a search is performed to optimize some collective property. Thus, Pittsburgh representation is commonly used to induce predictive classifiers, since such classifiers have to combine rules that are individually predictive and collectively complementary, in a way that a large number of examples is covered and correctly classified.

As previously stated, we use the Pittsburgh representation in our EA. One possible criticism regarding this representation is that no search is performed at rule level, i.e., rules are not improved by the search procedure. One possible strategy is to combine Michigan and Pittsburgh in a hybrid representation [7], [8]. However, this approach increases considerably the search space and doubles the number parameters. As consequence, this approach is more computationally intensive, its results are more difficult to analyze (due the larger number of parameters that should be tuned), and more important, the larger search space increases considerably changes of overfitting training data.

In order to use the Pittsburgh representation over a set of predictive rules, we use the Ripper rule induction algorithm [9] to generate an initial rule set. However, for most data sets, the rule set induced by Ripper is usually of restricted number of rules. Thus, we use a bootstrapping sampling strategy to generate multiple training sets. This sampling strategy allow us to increase the number and diversity of the rules.

In more details, in our experiments we use the k-fold stratified cross-validation resampling method for generating k training set $\mathcal{T}_1, \mathcal{T}_2, \ldots \mathcal{T}_k$ and their correspondent test sets from each data set. Next, n bootstrapping samples *with replacement* $\mathcal{T}_{i1}, \mathcal{T}_{i2}, \ldots \mathcal{T}_{in}$ are created from each training set $\mathcal{T}_i, 1 \leq i \leq k$. Each bootstrapped sample has the same number of examples as its corresponding training set, i.e., $|\mathcal{T}_{ij}| = |\mathcal{T}_i|$, $1 \leq j \leq n$.

Each bootstrapped training set $\mathcal{T}_{ij}$ is given as input to Ripper algorithm and n rule sets are induced $\mathcal{R}_{i1}, \mathcal{R}_{i2}, \ldots \mathcal{R}_{in}$. All rule sets are integrated into a unique pool of rules, and the repeated rules are discarded. Figure 1 illustrates this sampling approach.

In a second step, all rules are given as input to our EA. Internally, each rule is associated with a unique identifier. An individual, i.e., a rule set is represented as a set of rule identifiers. Finally, the population is a table that contain all sets of individuals. This representation scheme is very convenient, since conventional evolutionary operations, such as mutation and crossover, can be readily implemented as simple manipulations of the population table. Figure 2 illustrates this representation scheme.

In this work, we analyze how different rule quality measures might influence the classification performance of a rule set

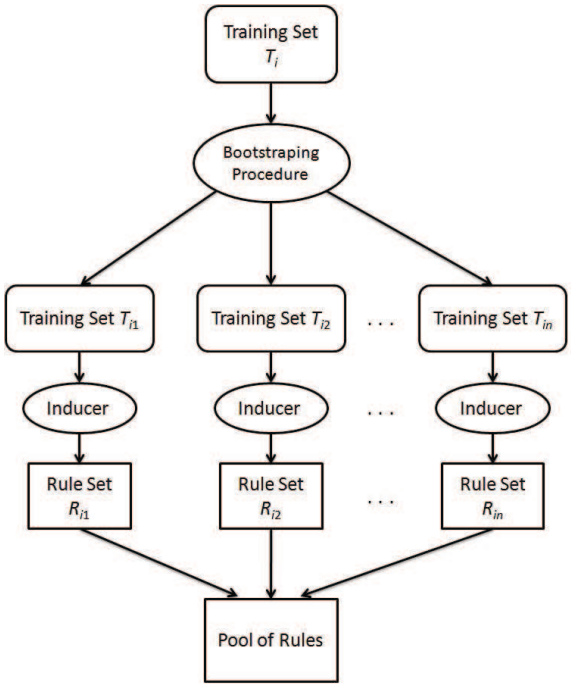


Fig. 1. Approach used to generate multiple rule sets using bootstrapping samples.

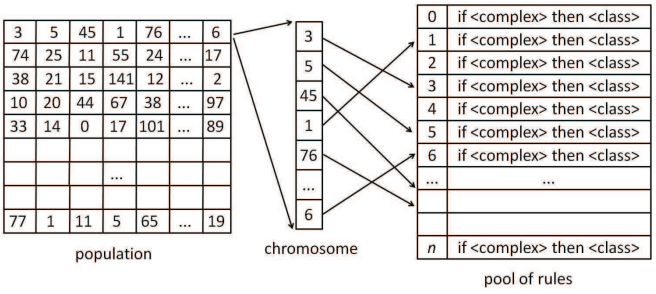


Fig. 2. An entire population represented as a table where each individual is a row (*left*). Each individual is a set of rule identifiers (*middle*). Each rule identifier corresponds to an unique rule in the pool (*right*).

searched by an EA. Given a classifier $\mathcal{C}$, such as the one represented by the individual in Figure 2-*middle*, and an example e to be classified, several rules of $\mathcal{C}$ might cover the example e. Although different classes may be predicted by the fired rules, we want to choose one rule that will provide the final classification. In the next section, we review some popular rule quality measures. These measures are used to decide which rule will provide the class of e.

III. Rule Measures

A *classification rule* is an intelligible representation of a piece of knowledge. A rule R is given in the form

$$B \rightarrow H$$

where B, called *body*, is a conjunction of conditions and H, called *head*, is the class value predicted by the rule.

Given a rule R and an example e, R *covers* e if all conditions of B are verified true in e. A rule *correctly* covers

TABLE I
CONTINGENCY MATRIX.

	H	$\overline{H}$	
B	f_{bh}	$f_{b\bar{h}}$	f_b
$\overline{B}$	$f_{\bar{b}h}$	$f_{\bar{b}\bar{h}}$	$f_{\overline{b}}$
	f_h	$f_{\overline{h}}$	1

an example if the rule covers the example and correctly predicts its class.

When a rule is evaluated against a data set, the examples may be distributed along four sets, B, $\bar{B}$, H and $\bar{H}$. Examples covered by a rule belong to B, while examples having the same class as predicted by the rule belong to H. Their complements, $\bar{B}$ and $\bar{H}$ contain the examples *not covered* and the examples *incorrectly* predicted by the rule. The corollary intersections contain examples correctly covered, incorrectly covered, not covered but correctly predicted, and finally, not covered and incorrectly predicted. These sets are important to construct the rule's *contingency matrix* (see Table I), which is the basis of the Lavrač framework [10].

We use the notation f_{xy} to denote the empirical frequency of an event $x \in \{B, \bar{B}\}$ and an event $y \in \{H, \bar{H}\}$. Therefore, f_{xy} in an empirical estimate of the probability $p(x, y)$. For sake of completeness, we describe all empirical frequencies as follows:

- f_{bh} is the percentage of examples covered and correctly classified by a rule R;
- $f_{b\bar{h}}$ is the percentage of examples covered and incorrectly classified by a rule R;
- $f_{\bar{b}h}$ is the percentage of examples not covered by R, but the class predicted by R is the same class of the example;
- $f_{\bar{b}\bar{h}}$ is the percentage of examples not covered by R, and the class predicted by R is different from the class of the example.

The marginal frequencies f_b, $f_{\bar{b}}$, f_h, $f_{\bar{h}}$ are defined as:

- f_b is the percentage of examples covered by R;
- $f_{\bar{b}}$ is the percentage of examples not covered by R;
- f_h is the percentage of examples correctly classified by R, independently if R covers or not the examples. It is also the prior probability estimate of the class predicted by the rule;
- $f_{\bar{h}}$ is the percentage of examples incorrectly classified by R, independently if R covers or not the examples.

The Lavrač framework allows to define different rule measures under a same organization. In this work, we analyze the influence of 9 rule measures listed in Table II. We briefly describe each measure as follows:

1) **Confidence**, also known as *precision* or *strength*, is the probability that a rule R will provide a correct prediction given that it covered the example. In practice, is probability might be very high for rules that cover a restricted number of examples;
2) **Laplace** is the confidence measure with Laplace correction. Laplace correction is frequently used to improve probabilities estimates when data are scarse. This implementation of Laplace approximates the estimated probability to 0.5 as fewer examples are covered by the rule;
3) **Lift** measures the confidence of R relative to the prior probability of the class predicted by R. This measure is based on the idea that an useful rule should have a confidence higher than a default rule that always predicts the same class;
4) **Conviction** is similar to lift since it also relates confidence with class prior probability. However, conviction is very sensitive to the confidence of a rule. Rules with a confidence value of 1, which it is not rare for low coverage rules, will have an infinite conviction;
5) **Leverage**, also known as *Piatetsky-Shapiro's* measure, is derived from the concept of statistical independence. If two events x and y are independent, then $p(x, y) = p(x) \times p(y)$. Leverage measures how much f_{bh} deviates from $f_b \times f_h$, i.e., the probability estimate assuming the events b and h independent. It is expected that an useful rule has a confidence higher than the prior probability of the class that it predicts, i.e., $\frac{f_{bh}}{f_b} > f_h$. Therefore, we should look for rules that $f_{bh} > f_b \times f_h$;
6) $\mathcal{X}^2$ is a well-known statistical test of independence. It is used to measure the independence between the rule antecedent and consequent. It is closely related to ϕ-**coefficient**, but it takes in account the number of instances in the data set (N);
7) **Jaccard** is a measure of overlapping between the number of cases covered by the rule and the number of cases that belongs to the predicted class. This measure has maximum value $f_b = f_h = f_{hb}$, i.e., when the rule covers all examples of the predicted class and none example from other classes. In contrast, its minimum value is given when $f_{bh} = 0$, i.e., when the rule misclassify every case it predicts;
8) **Cosine** is frequently used in text mining to measure the similarity between two vectors of attributes. For scalar values, cosine is similar to Jaccard measure and assume its minimal and maximal values in the same conditions;
9) ϕ-**coefficient** is a statistical measure of association between two binary variables. This measure is related to the Pearson correlation coefficient and also to the $\mathcal{X}^2$ measure. Since $\phi = \frac{\mathcal{X}^2}{N}$ [12], where N is the number of data instances, the ϕ-coefficient is independent from the data set size.

IV. EXPERIMENTAL EVALUATION

We carried out a number of experiments to evaluate the influence of each rule quality measure in the performance of symbolic classifiers searched by the evolutionary algorithm. The experiments were performed using 10 benchmark data sets, collected from the UCI repository [13]. In addition, we used AUC as the main measure to assess our results. Table III summarizes the main features of these data sets, which are: Identifier – identification of the data set used in the text; #Examples – the total number of examples; #Attributes (quanti., quali.) – the total number of attributes, as well as the number of quantitative and qualitative attributes; Classes (min.,

TABLE II
SUMMARY OF RULE QUALITY MEASURES (ADAPTED FROM [11]).

Measure	Definition	Range
Confidence	$conf = \frac{f_{bh}}{f_b}$	0...1
Laplace	$lapl = \frac{f_{bh}+1}{f_b+2}$	0...1
Lift	$lift = \frac{conf}{f_h}$	$0...+\infty$
Conviction	$conv = \frac{1-f_h}{1-conf}$	$0.5...1...+\infty$
Leverage	$leve = f_{bh} - (f_b \times f_h)$	$-0.25...0...0.25$
$\mathcal{X}^2$	$\mathcal{X}^2 = N \times \sum_{x\in\{b,\bar{b}\},y\in\{h,\bar{h}\}} \frac{(f_{xy} - f_x \times f_y)^2}{f_x \times f_y}$	$0...+\infty$
Jaccard	$jacc = \frac{f_{bh}}{f_b + f_h - f_{bh}}$	0...1
Cosine	$cos = \frac{f_{bh}}{\sqrt{f_b \times f_h}}$	$0...\sqrt{f_{bh}}...1$
ϕ-coefficient	$\phi - coeff = \frac{leve}{\sqrt{f_b \times f_h \times f_{\bar{b}} \times f_{\bar{h}}}}$	$-1...0...1$

TABLE III
DATA SETS DESCRIPTION.

Identifier	#Examples	#Attributes (quanti., quali.)	Classes (min., maj.)
Blood	748	4 (4, 0)	(1, 0) (24.00%, 76.00%)
Breast	699	10 (10, 0)	(benign, malignant) (34.99%, 65.01%)
Bupa	345	6 (6, 0)	(1, 2) (42.02%, 57.98%)
CMC	1473	9 (2, 7)	(1, remaining) (42.73%, 57.27%)
Flare	1066	10 (2, 8)	(C-class, remaining) (17.07%, 82.93%)
Haberman	306	3 (3, 0)	(2, 1) (26.47%, 73.53%)
New-Thyroid	215	5 (5, 0)	(remaining, 1-normal) (30.23%, 69.77%)
Pima	768	8 (8, 0)	(1, 0) (34.89%, 65.11%)
Vehicle	946	18 (18, 0)	(van, remaining) (23.89%, 76.11%)
Yeast	1484	8 (8, 0)	(NUC, remaining) (28.90%, 71.10%)

maj.) % (min., maj.) – the label of the minority and majority classes and the percentage of minority and majority classes. In order to measure the performance of the classifiers using AUC, data sets with more than two classes were transformed in binary classification problems by selecting one of the classes as minority/majority class (as indicated in column Classes) and assigning the examples from the other classes to the majority/minority class.

As previously described, we used the 10-fold stratified cross-validation resampling method for generating training and their correspondent test sets. In addition, 30 bootstrapping samples with replacement were created for each training set. We empirically chose the number of 30 bootstrapping samples since it allowed to create a diverse pool of rules. Increasing this number did not improve our results, but increased the training times.

Each bootstrapped training set was given as input to the Machine Learning algorithm Ripper. The rules from all rule sets were integrated into a unique pool of rules, and the repeated rules were discarded. Next, the pool of rules was given as input to the evolutionary algorithm that outputted a final rule set (a classifier). Finally, AUC was measured over the test set.

TABLE IV
CHROMOSOME SIZE FOR EACH DATA SET BASED ON THE MEAN NUMBER OF RULES PROVIDED BY RIPPER.

Data set	Chromosome size
Blood	6
Breast	6
Bupa	8
CMC	12
Flare	6
Haberman	4
New-Thyroid	4
Pima	10
Vehicle	6
Yeast	8

Our evolutionary algorithm was set to use 40 chromosomes in all experiments. The chromosome size, i.e., the number of rules of an individual classifier was defined according to the average size of the classifiers generated by Ripper in each data set. In the Pittsburgh approach, it is a commonsense to allow variable-sized chromosomes. Therefore, the evolutionary algorithm is free to search for rule sets with different number of rules using a two-point cross-over operator. In our case, we noticed that a search with variable-sized chromosomes resulted in very large rule sets for most domains. These large rule sets had a poor performance in the test set, indicating overfitting. Therefore, we opted to keep all chromosomes with fixed sizes. The chromosome size chosen for each data set is the mean number of rules induced by Ripper in the same data set. Table IV lists the chromosome sizes for each data set.

The fitness function used is the AUC metric measured over the training examples. The selection method is the fitness-proportionate selection. The crossover operator was applied with probability 0.4 and the mutation operator was applied

with probability 0.1. Mutation and crossover rates were chosen based on our previous experience with EAs [14], [15]. The number of generations was limited to 20. Our implementation uses an elitism operator to ensure that the best classifier is kept in the next population. Finally, since evolutionary algorithms perform a stochastic search that might provide different results in each execution, we repeated each experiment 10 times and averaged the results.

Table V presents results obtained. All results represent the mean AUC values calculated over the 10 pairs of training and test sets and averaged for 10 repeated executions. The standard deviations are also showed between parentheses. The second column shows the results obtained by Ripper for all data sets. The next columns show the results obtained by the evolutionary algorithm for each rule quality measure. The best AUC for each data set is emphasized in boldface. We can note that the EA search has improved considerably the AUC when compared with the results obtained by Ripper. However, the best AUC values are scattered throughout the table, indicating that no single measure systematically provides the best results.

Since no single measure provided the best results, we decided to rank the measures considering their mean AUC values. Table VI shows the results for this ranking. The second-to-last column of table shows the sum of ranks obtained by each measure for all the data sets. The last column shows a score based on the sum of ranks for each measure, in a way that the measure that has the lowest sum of ranks scores 1. The measure with lowest score is $confidence$. $Confidence$ obtained the better AUC values for Vehicle and Yeast data sets; the second better AUC values to Breast, CMC, Flare and New-Thyroid data sets; the third better AUC value to Bupa and Pima data set; the fourth better AUC value to Blood; and, the fifth better AUC value to Haberman data set.

In order to analyze whether there is a statistically significant difference among the compared measures, we ran the Friedman test[1]. The Friedman test was run with the null-hypotheses that the performance of all rule measures is comparable. When the null-hypothesis is rejected by the Friedman test, at 95% confidence level, we can proceed with a post-hoc test to detect which differences among the methods are significant. For such, we ran the Bonferroni-Dunn multiple comparisons with a control test.

The null-hypothesis was rejected by the Friedman test at 95% confidence level. So, we ran the Bonferroni-Dunn test using the measure $confidence$ as control. The Bonferroni-Dunn test indicate that the EA allied the measure $confidence$ outperforms Ripper with 95% confidence level. However, there are no statistically significant differences among the rule quality measures.

Our results differ from previously published results. To the best of our knowledge, the most similar work in literature is [11] which compares rule quality measures in the context of association rule classification. Their results indicate that $conviction$ presented the best results. Our experiments indicate that $confidence$ and $lift$ performed slightly better than $conviction$, but with no statistical difference. This difference between the results presented might be motivated by the use of different performance measures, since error rate was used in [11].

[1]The Friedman test is a nonparametric equivalent of the repeated-measures ANOVA. See [16] for a thorough discussion regarding statistical tests in Machine Learning research.

V. Conclusion

In this work, we compared 9 different rule quality measures in 10 different benchmark data sets. The rule measures were used to decide which rule should provide the final classification in an unordered rule set. Our results indicate that the use of different rule measures have a marginal effect over the classification performance assessed by the area under the ROC curve. The $confidence$ measure presented the best mean results, but with no statistical difference to the other rule measures.

As future work, we plan to investigate the use of different rule quality measures as a weighting factor in a voting approach in which all fired rules contribute to the final classification.

Acknowledgment

The authors would like to thank CNPq, CAPES and FAPESP, Brazilian funding agencies, for the financial support.

References

[1] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[2] S. F. Smith, "A learning system based on genetic adaptive algorithms," Ph.D. dissertation, Pittsburgh, PA, USA, 1980.

[3] R. D. J. A. Swets and J. Monahan, "Better decisions through science," *Scientific American*, 2000.

[4] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," HP Labs, Tech. Rep. HPL-2003-4, 2004.

[5] A. A. Freitas, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag, 2002.

[6] J. H. Holland, *Escaping brittleness: The possibilities of general purpose learning algorithms applied to parallel rule-based systems*, ser. Machine learning: An artificial intelligence approach. Morgan Kaufmann, 1986, vol. 2.

[7] D. P. Greene and S. F. Smith, "Competition-based induction of decision models from examples," *Machine Learning*, vol. 13, no. 2–3, pp. 229–257, 1993.

[8] A. Giordana and F. Neri, "Search-intensive concept induction," *Evolutionary Computation*, vol. 3, no. 4, pp. 375–416, 1995.

[9] W. Cohen, "Fast effective rule induction," in *International Conference on Machine Learning*, 1995, pp. 115–123.

[10] N. Lavrač, P. Flach, and R. Zupan, "Rule evaluation measures: A unifying view," in *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP-99)*, vol. 1634. Springer-Verlag, 1999, pp. 74–185.

[11] P. J. Azevedo and A. M. Jorge, "Comparing rule measures for predictive association rules," in *18th European Conference on Machine Learning*, 2007, pp. 510–517.

[12] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004.

[13] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[14] C. R. Milaré, G. E. A. P. A. Batista, A. C. P. L. F. Carvalho, and M. C. Monard, "Applying genetic and symbolic learning algorithms to extract rules from artificial neural neworks," in *Proc. Mexican International Conference on Artificial Intelligence*, ser. LNAI, vol. 2972. Springer-Verlag, 2004, pp. 833–843.

TABLE V
AVERAGE AUC VALUE OBTAINED BY RIPPER AND EVALUATED MEASURES.

	Ripper	$conf$	$lapl$	$lift$	$conv$	$leve$	$\mathcal{X}^2$	$jacc$	cos	ϕ-coeff
Blood	63.34(3.69)	67.42(6.38)	66.89(6.44)	**67.74(6.25)**	67.67(6.16)	66.83(6.66)	66.78(6.30)	67.34(6.28)	67.22(6.32)	67.49(6.27)
Breast	97.33(2.26)	97.60(1.74)	97.06(1.72)	**97.84(1.28)**	97.32(1.34)	97.00(1.44)	96.89(1.70)	96.79(1.81)	96.95(1.91)	96.75(1.90)
Bupa	67.13(6.19)	67.34(4.90)	66.06(6.37)	**69.06(5.20)**	67.25(4.78)	67.19(6.07)	66.49(5.51)	66.21(5.26)	66.77(6.30)	68.42(6.09)
CMC	68.64(2.27)	69.58(3.45)	69.13(3.35)	**69.63(3.18)**	69.30(3.43)	69.11(3.15)	69.32(3.34)	68.87(3.37)	69.08(3.83)	68.96(3.49)
Flare	56.94(2.25)	63.13(4.55)	62.20(4.76)	62.84(4.84)	62.94(4.69)	**63.43(4.85)**	62.26(5.10)	62.56(4.66)	62.45(4.56)	62.30(4.86)
Haberman	60.94(11.31)	63.53(7.95)	63.65(9.05)	62.83(8.08)	63.97(8.17)	62.60(7.95)	63.73(7.97)	**64.09(8.02)**	62.63(7.67)	63.07(7.85)
New-Thyroid	92.50(7.92)	95.06(4.92)	94.86(5.70)	93.95(5.94)	93.36(6.82)	**95.18(5.35)**	94.31(6.08)	94.26(6.23)	94.43(5.85)	94.47(5.40)
Pima	69.98(2.21)	74.12(3.51)	72.28(4.34)	74.19(3.58)	**74.54(2.97)**	71.92(4.28)	72.27(3.84)	72.50(3.36)	72.37(4.22)	72.22(4.05)
Vehicle	92.21(2.55)	**94.42(1.90)**	94.06(1.89)	93.90(1.86)	93.52(2.41)	92.98(2.22)	93.67(2.09)	93.82(1.77)	93.52(2.21)	93.25(1.94)
Yeast	65.99(2.13)	**69.49(3.15)**	68.41(2.81)	69.01(2.72)	69.02(3.37)	68.30(3.32)	68.69(3.33)	67.78(2.66)	68.49(2.61)	68.80(2.85)

TABLE VI
RANKING OF AUC VALUES OBTAINED BY RIPPER AND EVALUATED MEASURES.

Data Set	Blood	Breast	Bupa	CMC	Flare	Haberman	New-Thyroid	Pima	Vehicle	Yeast	Sum	Score
Ripper	10	3	6	10	10	10	10	10	10	10	89	8
$conf$	4	2	3	2	2	5	2	3	1	1	25	1
$lapl$	7	5	10	5	9	4	3	6	2	7	58	4
$lift$	1	1	1	1	4	7	8	2	3	3	31	2
$conv$	2	4	4	4	3	2	9	1	7	2	38	3
$leve$	8	6	5	6	1	9	1	9	9	8	62	6
$\mathcal{X}^2$	9	8	8	3	8	3	6	7	5	5	62	6
$jacc$	5	9	9	9	5	1	7	4	4	9	62	6
cos	6	7	7	7	6	8	5	5	6	6	63	7
$\phi - coeff$	3	10	2	8	7	6	4	8	8	4	60	5

[15] C. R. Milaré, G. E. A. P. A. Batista, and A. C. P. L. F. Carvalho, "A hybrid approach to learn with imbalanced classes using evolutionary algorithms," in *Proc. 9th International Conference Computational and Mathematical Methods in Science and Engineering (CMMSE)*, vol. II, 2009, pp. 701–710.

[16] J. Demšar, "Statistical comparisons of classifiers over multiple data sets." *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

Against-Expectation Pattern Discovery: Identifying Interactions within Items with Large Relative-Contrasts in Databases

Dingrong Yuan, Xiaofang You, Chengqi Zhang

***Abstract*—We design a new algorithm for identifying against-expectation patterns. An against-expectation pattern is either an itemset whose support is out of a range of the expected support value, referred to as an against-expectation itemset, or it is an association rule generated by an against-expectation itemset, referred to as an against-expectation rule. Therefore, against-expectation patterns are interactions within those items whose supports have large relative-contrasts in a given database. We evaluate our algorithms experimentally, and demonstrate that our approach is efficient and promising**

***Index Terms*—Exception, against-expectation pattern, nearest-neighbor graph, correlation analysis.**

I. Introduction

Traditionally, association analysis has focused on techniques aimed at discovering interactions within data. It has mainly involved association rules [1,4,21] and negative association rules [18,23]. These rules can be identified from data by using statistical methods and grouping. In real world applications, data marketers seek to identify interactions and predict profit potential in the relative-contrast of sales. Meanwhile, they recognise that principle items, having large relative-contrasts with respect to their supports expected for a given database, may provide larger profit potential than those with low relative-contrasts. In this paper, we refer to interactions within items that have large relative-contrast as against-expectation patterns. Up until now, the techniques for mining against-expectation patterns have been undeveloped. To rectify this, our paper studies the issue of mining against-expectation patterns in databases.

An against-expectation pattern is either an itemset whose support is out of a range of the expected support value (expectation), referred to here as an against-expectation itemset, or an association rule generated from against-expectation itemsets, referred to as an against-expectation rule.

Dingrong Yuan is with College of Computer Science and Information Technology Guangxi Normal University, Guilin, 541004, China (dryuan@mailbox.gxnu.edu.cn).

Xiaofang You is with the College of Computer Science and Information Technology Guangxi Normal University, Guilin, 541004, China.

Chengqi Zhang is with Faculty of Information Technology, University of Technology Sydney PO Box 123, Broadway NSW 2007, Australia(chengqi@it.uts.edu.au).

If we use extant frequent-pattern-discovery algorithms to mine a market basket dataset, the item ‘apple’ can be identified as a frequent pattern (itemset), even though its support (= 200) is much less than expected sales (= 300). This is because ‘apple’ is a popular fruit, and is frequently purchased. Compared to ‘apple’, ‘cashew’ is an expensive fruit, and is rarely purchased. In the market basket dataset, ‘cashew’ cannot be discovered as a frequent pattern of interest, even though its support (= 20) is much greater than its expected sales (= 5). In an applied context, while the frequent pattern ‘apple’ is commonsense, the purchasing increase of ‘cashew’ is desired in marketing decision-making, and constitutes the against-expectation pattern which is to be mined in this paper. Similarly, the purchasing decrease of ‘apple’ is also an against-expectation pattern desired. These against-expectation patterns assist in evaluating the amount of products purchased in the next time-lag.

Against-expectation patterns are distinct from frequent patterns (or association rules) because: (1) they may be pruned when identifying frequent patterns (or association rules), (2) they can deviate from frequent patterns (or association rules), and (3) against-expectation patterns are hidden in data, whereas traditional frequent patterns (or association rules) are relatively obvious.

Related research includes the following: unexpected patterns [14,15], exceptional patterns [6,8,10,19], and negative association rules [18,23]. The first and second are known as ‘exceptions of rules’, and also as ‘surprising patterns’, whereas ‘negative association rules’ represents a negative relation between two itemsets.

An exception of a rule is defined as a deviational pattern to a well-known fact, and exhibits unexpectedness. For example, while ‘bird(x) → flies(x)’ is a well-known fact, mining exceptional rules aims to find patterns such as ‘bird(x), penguin(x) → ~flies(x)’. The negative relation actually implies a negative rule between the two itemsets, including association rules of forms A → ~B, ~A → B and ~A → ~B, which indicate negative associations between itemsets A and B [18,23].

Hence, against-expectation patterns differ from unexpected patterns, exceptional patterns and negative association rules. Therefore, against-expectation patterns should be regarded as a new kind of pattern.

In addition to those mentioned above, there are also some differences between mining against-expectation patterns and mining the other two patterns change patterns [12], and interesting patterns using user expectation [13]. In these methods, a decision tree is used for mining changes, which is

very distinct from the algorithms employed in this paper. Finding interesting patterns according to user expectation is a kind of subjective measure, while our algorithms aim at objectivity.

As such, while the above achievements provide a good insight into exceptional pattern discovery, this paper will focus on identifying against-expectation patterns. In Section II, we formally define some basic concepts and examine the approach issue of mining against-expectation patterns. In Section III we describe our approach and compare it with existing algorithms. In Section IV, we conduct a set of experiments to evaluate our algorithms. We summarize our contribution in Section V

II. A Framework for Identifying Against-expectation Patterns

In this section we present some basic concepts and describe the issue of mining against-expectation patterns in databases. In particular, we design a new framework for mining against-expectation patterns that consists of interactions within items, with large relative-contrasts referenced to their expectations in a given database. This is based on heterogeneity metrics.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n distinct literals, called items. For a given dataset D over I, we can represent D as follows.

TABLE I
A DATASET D OVER I

TID	i_1	i_2	.	i_n
T_1	a_{11}	a_{12}	.	a_{1n}
T_2	a_{21}	a_{22}	.	a_{2n}
…	…	…	.	…
T_m	a_{m1}	a_{m2}	.	a_{mn}
Support	f_1	f_2	.	f_n

In Table I, T_j is the identifier of transactions in D, a_{jk} is the quantity of item i_k in transaction T_j, f_k is the quantity (support) of i_k in D, or the sum of a_{+k} in *k*th column.

In marketing, data marketers must know the sales expectation of each product. This is used to determine how many items should be bought each month (or during a specified time-lag).

TABLE II
THE EXPECTATION OF ITEMS

	i_1	i_2	…	i_n
Expectation	e_1	e_2	…	e_n

In Table II, e_k is the expected quantity of i_k in D. For example, in Section I, expectation(apple) = 300, support(apple) = 200, expectation(cashew) = 5, and support(cashew) = 20. Therefore, the support of 'cashew' is far larger than its expectation because

$$\textit{increased}(\text{cashew}) = \text{support}(\text{cashew}) - \text{expectation}(\text{cashew})$$

$$= 20 - 5 = 15.$$

This means that the relative contrast, which is the ratio of increment to expectation used for denoting the gap between support and expectation rec(cashew), is three times its expectation. That is,

$$\textit{rec}(\text{cashew}) = \text{increased}(\text{cashew}) / \text{expectation}(\text{cashew})$$

$$= 15/5 = 3.$$

Hence, 'cashew' is an against-expectation pattern. However, the support of 'apple' is far less than its expectation because

$$\textit{increased}(\text{apple}) = \text{support}(\text{apple}) - \text{expectation}(\text{apple})$$

$$= 200 - 300 = -100.$$

This means that the relative contrast *rec*(apple) is

$$\textit{rec}(\text{apple}) = \textit{increased}(\text{apple}) / \text{expectation}(\text{apple}) = -100/300 = -1/3.$$

We also refer to 'apple' as an interesting against-expectation pattern if *rec*(apple) is out of a given range (neighbour) of expectation(apple).

Against-expectation patterns are defined as either those itemsets whose supports are out of a δ-neighbour of their expected values, referred to as *against-expectation itemsets*, or those rules that are interactions within against-expectation itemsets, referred to as *against-expectation rules*. An against-expectation itemset X has its support out of the δ_X-neighbour of its expectation (with a large relative contrast), i.e.

$$\textit{recar}(X) = |\textit{increased}(X)| / \text{expectation}(X)$$

$$= |\text{support}(X) - \text{expectation}(X)| / \text{expectation}(X)$$

$$> \delta_X$$

where δ_X is a user-specified minimum relative-contrast for X. Certainly, recar(X) is a heterogeneity metrics, because δ_X can be different with different X.

An against-expectation rule is of the form

$$X \rightarrow Y$$

which is the interaction between the against-expectation itemsets X and Y; or one of X and Y is a frequent itemset and the other an against-expectation itemset. For example, 'apple → cashew' can be an against-expectation rule between the above two against-expectation itemsets.

An against-expectation rule $X \rightarrow Y$ is interesting if its confidence is greater than, or equal to, a user-specified minimum confidence (minconf).

We classify against-expectation patterns as follows: increment patterns, decrement patterns and negative associations.

- An increment pattern is either an itemset X whose actual support is greater than its expected value, e.g., support(X) – expectation(X) > *e* (where *e* is a user-specified positive value) – this is referred to as an *increment itemset*; or it is a rule that is an interaction within increment itemsets, and is referred to as an *increment rule*.
- A decrement pattern is either an itemset X whose actual support is less than its expected value, e.g., support(X) – expectation(X) < -*e* (where *e* is a user-specified positive value) – referred to as a *decrement itemset*; or it is a rule that is an interaction within decrement itemsets, and is referred to as a *decrement rule*.
- A *mutually-exclusive* correlation is a rule in which its antecedent and action belong to different against-expectation itemsets. That is, either (1) its antecedent is an increment itemset and its action a decrement itemset; or (2) its antecedent is a decrement itemset and its action an increment itemset.
- A *companionate* correlation is a rule in which one of its antecedents and actions is a frequent itemset and the other is an against-expectation itemset. That is, either (1) its antecedent is an increment itemset and its action is a frequent itemset; (2) its antecedent is a frequent itemset and its action an increment itemset; (3) its antecedent is a decrement itemset and its action a frequent itemset; or (4) its antecedent is a frequent itemset and its action a decrement itemset;

The problem of mining against-expectation patterns is a challenging issue because it is *very different from* those problems faced when discovering frequent patterns (or association rules). Because against-expectation patterns can be hidden in both frequent and infrequent itemsets, traditional pruning techniques are inefficient for identifying such patterns. This indicates that we must exploit alternative strategies to

(a) confront an exponential search space consisting of all possible itemsets, frequent and infrequent, in a database;

(b) detect which itemsets can generate against-expectation patterns;

(c) determine which against-expectation patterns are really useful to applications; and

(d) determine the heterogeneity metrics of against-expectation patterns.

One must remember that this process can be expensive and dynamic. The leaders of a new company must make an effort to estimate expectation by analyzing the environment and possible customers. For an old company, data relating to items recorded in a previous time-lag can be used as expectations. Thus, we can check the support of items and identify the against-expectation patterns in that time-lag. We can also predict the support of items and the against-expectation patterns in the next time-lag. In the following, for simplification, we define a time-lag as a month. From this point, our against-expectation patterns are similar to change patterns [12,13], though discovering against-contrast patterns is based on the expected support of items.

The technique to be developed in this paper consists of a two-step approach as follows:

(1) Generating a set of interesting items (i.e., items with a large relative-contrast), and

(2) Identifying interactions within these interesting items based on the Nearest-Neighbor Graph and Correlation Analysis techniques.

III. Algorithm Description

Traditional mining algorithms assume that an association rule is interesting as long as it satisfies minimum support and minimum confidence. But a number of researchers have proved that this can generate many uninteresting rules. Meanwhile, some interesting rules can be missed without taking into account the item's own change trend.

Example 1. Let min_supp = 33.3%. Consider two transaction databases D1 and D2, as shown in Tables III and IV.

TABLE III
D1: Transactions in January

toothbrush, toothpaste
bread, jam
cashew
apple, banana
toothbrush, toothpaste, bread
apple, cola, shampoo
apple, banana, shampoo
bread, jam, apple
apple
apple, banana, cola

TABLE IV
D2: TRANSACTIONS IN FEBRUARY

toothbrush, toothpaste, bread, jam, shampoo
bread, cashew, apple, shampoo, jam
cashew, shampoo, bread, jam
apple, banana, cola, shampoo
toothbrush, toothpaste, shampoo
bread, jam, shampoo
toothbrush, toothpaste, cola, shampoo
apple, shampoo, bread, jam
apple, banana, shampoo
cashew, apple, banana, cola, shampoo
toothbrush, toothpaste, apple, shampoo
bread, jam, apple, shampoo

Using existing association rule mining algorithms, we can identify certain association rules in D1, or D2, or D1∪D2. For example, association rules 'toothbrush → toothpaste', 'toothbrush → shampoo' and 'bread → jam' are of interest in D2 because

$$supp(toothbrush \rightarrow toothpaste) = 33.3\%$$
$$conf(toothbrush \rightarrow toothpaste) = 100\%$$
$$supp(toothbrush \rightarrow shampoo) = 33.3\%$$
$$conf(toothbrush \rightarrow shampoo) = 100\%$$
$$supp(bread \rightarrow jam) = 50\%$$
$$conf(bread \rightarrow jam) = 100\%$$

Let us consider the support and increment (relative contrast) of the items 'apple' and 'cashew' as shown in Table V.

TABLE V
COMPARING APPLE WITH CASHEW

	Jan's supp	Feb's sup	increment
Apple	50%	58.3%	16.7%
cashew	10%	25%	300%

Obviously, whenever it is January or February, 'apple' is always frequent and 'cashew' is always infrequent. But the increment (relative contrast) of 'cashew' is far higher than that of 'apple'. Therefore, 'cashew' is of much more interest when considering the relative contrast (i.e., change trend). These interesting itemsets cannot be found by using existing association rule mining algorithms. Certainly, we can identify 'cashew' as a frequent itemset, using Apriori-like algorithms, by decreasing the min_supp to 25%. However, decreasing the min_supp can lead to the generation of a great many uninteresting itemsets. It is not an intelligent way to proceed. In particular, Apriori-like approaches do not provide information concerning the change trend of itemsets. We are, therefore, encouraged to develop new techniques for discovering against-expectation patterns.

Our proposed approach for identifying against-expectation patterns uses the stock of merchandise in the previous month as an expectation, in order to identify against-expectation patterns in the present month.

A. Finding interesting itemsets using square deviation

This subsection presents techniques for finding a candidate set consisting of particularly interesting items (1-itemset) based on the square deviation, and is aimed at mining against-expectation patterns. Generally, if a piece of merchandise (an item) has a small relative contrast referenced to its expectation, the sales trend of this merchandise is in control. And it is uninteresting when mining against-expectation patterns. For efficiency, the item should be deleted from the candidate set.

Stocked items can reflect a decision makers' anticipation of the sale trend for each item in the future. Therefore, in Definition 1, stock is considered to be equal to the expectation mentioned above. The stage below is necessary for determining changing trends, and its span can be decided according to practical situations, such as week, month or year.

Definition 1. Let x_{ij} be the stock of i-th merchandise at j-th stage, where $0 \le j \le m$, $0 \le I \le n$; and p_{ij} denote the increment ratio between the j-th stage and the (j+1)-th stage of the i-th merchandise. Then we have

$$p_{ij} = \frac{x_{i(j+1)} - x_{ij}}{x_{ij}}.$$

Therefore, its incremental ratio math expectation is

$$E(p_i) = \frac{\sum_{j=1}^{m-1} p_{ij}}{m-1}$$

and its square deviation is

$$D(i) = \sum_{j=1}^{m} (p_{ij} - E(p_i))^2.$$

Let propdenote the threshold given. Then merchandise i is of interest if $D(i) \ge$ prop.

The algorithm to identify all interesting items (merchandise) is constructed in Fig. 1.

Input Stock: set of the number of goods, Prop: minimum math expectation;
Output MID: set of candidate 1-itemsets
(1) **for** each j in Stock
(2) **begin**
(3) calculate $E(p_i)$, D(i);
(4) **if** (D(i) ≥ Prop)
(5) MID ← MID ∪ {i};
(6) **end;**
(7) **Output** all items of interest in MID;
(8) **Return**;

Fig.1. FIM: Generating the candidate set (of 1-itemsets)

In algorithm FIM, all items with large relative contrasts referenced to their expectations are identified using square deviation. In the following Sections III.B and III.C, k-nearest neighbor and correlation analysis will be used to calculate the correlation between two items and generate the candidate set of k-itemsets, which consist of all interesting itemsets potentially useful for generating against-expectation patterns.

B. Nearest neighbor graph based method

Definition 2. Let $A = \{a_0, a_1, a_2, \ldots a_{i-1} a_i, \ldots a_n\}$, $B = \{b_0, b_1, b_2, \ldots b_{i-1} b_i, \ldots b_n\}$ be the stocks of goods A and B. The increments between two adjacent phases for A and B are denoted as

$$A^1=\{a_1\text{-}a_0, a_2\text{-}a_1, \ldots a_i\text{-}a_{i-1}, \ldots a_n\text{-}a_{n-1}\}$$

$$B^1=\{b_1\text{-}b_0, b_2\text{-}b_1, \ldots b_i\text{-}b_{i-1}, \ldots b_n\text{-}b_{n-1}\}.$$

Let $x_{i-1} = (a_i - a_{i-1})/(b_i - b_{i-1})$ and $x_{i-1} \in X$, $X = \frac{A^1}{B^1} = \{x_0, x_1, x_2, \ldots x_{i-1} x_i, \ldots x_n\}$ be known as the relative bargaining quantity of A and B.

The nearest neighbor graph is constructed as follows. For all items in a given database, we first draft a figure with the items as its points. Any two points are connected by an edge, and the edge is associated with the square deviation distance between the two points. For instance, in Definition 2, A and B stand for the points *a* and *b* in a figure respectively. Then the distance between A and B (*a* and *b*) is defined as follows.

$$d_{A,B} = P(X) = P\left(A^1/B^1\right) = \sum_{i=0}^{n-1} \left(X_i - E[X]\right)^2.$$

After finishing the figure accordingly, we then get rid of some of the edges when their distances are larger than the threshold. The rest of the figure is simply the nearest neighbor graph for the database [10]. In this nearest neighbor graph, we refer to certain items as the nearest neighbors of the point a, if they are linked through edges to a. We can reconstruct the nearest neighbor graph by ranking the nearest neighbors for a certain point in decreasing order.

We now describe the algorithm for generating the candidate set of j-itemsets using the above k-nearest neighbor graph.

Input SaleTable: dataset; k: number of nearest neighbors;

Output MID: set of j-itemsets;

(1) StageTable ← SaleTable;
(2) **for** A in MID (firstly generated in Algorithm FIM)
(3) **begin for** B in MID (firstly generated in Algorithm FIM)
(4) **begin** calculate X(A, B);
(5) calculate d(A, B);
(6) **if** (d(A, B) ≥ k)
(7) A.neighbors ← B;
(8) **generate** a candidate itemset i by A and B;
(9) MID ← MID ∪ {i};
(10) **end**
(11) **end**
(12) **Output** all itemsets in MID;
(13) **Return;**

Fig.2. KNNG: Generating the candidate set of *j*-itemsets using the *k*-nearest neighbor graph

From the above description, the algorithm KNNG generates the candidate set of 2-itemsets, using all 1-itemsets in MID, generated first in the Algorithm FIM; the algorithm KNNG generates the candidate set of 3-itemsets, using all 2-itemsets in MID, generated in the Algorithm KNNG; and the algorithm KNNG generates the candidate set of j-itemsets, when using all (j-1)-itemsets in MID, generated in the Algorithm KNNG.

From the definition of against-expectation patterns, all candidate j-itemsets identified by the Algorithms FIM and KNNG are against-expectation itemsets. By way of these against-expectation itemsets, we can generate some against-expectation rules of interest. Here we omit the algorithm for generating against-expectation rules because it is similar to the Apriori algorithm.

C. Correlation-analysis based approach

In this section we design a correlation-analysis based algorithm for generating a candidate set of k-itemsets. The correlation-analysis based algorithm detects whether the correlation between two goods is positive or negative. That is, whether the correlation has a acceleration action or is restricted. In particular, this algorithm can provide pattern quality superior to that of the nearest-neighbor-graph based algorithm.

Definitions and theorems

Definition 3. Let $P(X)$ denote the probability of X occurring in a transaction, and $P(\overline{X})$ the probability of X not occurring in a transaction, where $P(\overline{X})$=1- $P(X)$; and let $P(X \cup Y)$ denote the probability of X and Y both occurring in a transaction. If $P(X \cup Y) \neq P(X)P(Y)$, then X and Y are dependent on each other; otherwise, independent of each other.

Definition 4. If $X_{a1}, X_{a2}, \ldots, X_{a3}$ in a transaction database, are correlated with each other, then $X_{a1}, X_{a2}, \ldots, X_{a3}$ is known as a correlation rule.

Definition 5. Let $X' = X_{a1}, X_{a2}, \ldots, X_{a3}$ and $X = \{X_1, X_2, \ldots, X_{n-1}, X_n\}$, then X' is referred to as a subset of X, and X is referred to as a superset of X'.

Each itemset consists of certain attributes, and a superset consists of all the attributes in its subsets. Therefore, reducing time cost is possible by getting rid of those itemsets that contain at least one uncorrelated subset (a non-correlation rule).

Definition 6. The correlation-analysis quantity of goods A and B is defined as

$$Corr(A,B) = \frac{P(A \cup B)}{P(A) \times P(B)}.$$

For n goods, each denoted as Gi, where 1<=i<=n, the correlation-analysis quantity is

$$Corr(G_1, G_2, ..., G_n) = \frac{P(\bigcup_{i=1}^{n} G_i)}{\prod_{i=1}^{n} P(G_i)}$$

The more closed to 1 the correlation-analysis quantity of *A* and *B* is, the better the independency of *A* and *B*. If $Corr(A,B) > 1$, then *A* and *B* are positively correlated, and if $Corr(A,B) < 1$, then *A* and *B* are negatively correlated .

Algorithm design

The correlation-analysis based algorithm is a post-process based upon a transaction database. It is suitable for static data classification, and is an improvement on Apriori-like algorithms for mining association rules using a support-confidence framework.

Reducing the time complexity of our correlation-analysis based algorithm can be implemented by way of a pruning algorithm that utilizes the closure property of correlation itemsets. Using a pruning algorithm, we can obtain all minimum correlation itemsets of interest. The pruning process is shown in Fig. 3 where, if $\mathrm{Corr}(0,1) \neq 1$, then 0 and 1 are correlated, as are all its supersets. Pruning its left subtree at (0,1) is actually a minimum correlation itemset.

The final output looks like this:

A: Positive correlated item: (*B*), (*E*,*F*), i.e., (*A*,*B*) and (*A*,*E*,*F*) are two minimum correlation itemsets concerning item *A*.

Negative correlated item: (*D*), (*G*).

So, it is easy to detect those interesting items that correlate with A

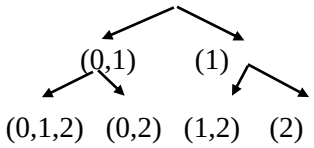


Fig. 3. Set-enumeration tree

We now construct the correlation-analysis based algorithm as follows.

Input TD: set of transactions; ProRange: a threshold for independency;

Output PosCorr: set of all positive correlation itemsets;
NegCorr: set of all negative correlation itemsets;

(1) PosCorr ← ∅; NegCorr ← ∅;

(2) **Construct** a set-enumeration tree for a 1-itemset in MID for TD (generated in Algorithm FIM);

(3) **Scan** each node A in the set-enumeration tree

(4) **If** (Corr > 1 + ProRange)
PosCorr ← PosCorr ∪ {A};

(5) **If** (Corr < 1-ProRange)
NegCorr ← NegCorr ∪ {A};

(6) **If** (|1-ProRange| ≤ Corr ≤ |1+ProRange|)
Delete A from the tree;

(7) **Output** PosCorr and NegCorr;

(8) **Return**;

Fig. 4. CA: Generating the candidate set of k-itemsets using correlation analysis

From the above description, the algorithm CA generates the candidate sets PosCorr and NegCorr, using all 1-itemsets in MID generated in the Algorithm FIM. The candidate sets consist of j-itemsets of interest.

From the definition of against-expectation patterns, all candidate j-itemsets identified by the Algorithms FIM and CA are against-expectation itemsets. Through these against-expectation itemsets, we can generate some against-expectation rules of interest. Like in the KNNG algorithm, we omit the algorithm for generating against-expectation rules.

In our CA algorithm, a threshold $proRange$ is used to measure independency. For example, items in X are independent when:

$$|1 - proRange| \le Corr(X) \le |1 + proRange|.$$

In addition, positive correlation is defined as

$$Corr(X) > |1 + proRange|$$

and negative correlation is defined as

$$Corr(X) < |1 - proRange|.$$

Note that we can use, for example, Chi-square, as our another metrics for measuring the correlation of itemsets by statistical means.

IV. Experiments

To evaluate our two metrics, we have conducted extensive experiments on a DELL Workstation PWS650 with 2G main memory, 2.6G CPU, and WINDOWS 2000. We evaluate our approaches using the databases generated from http://www.kdnuggets.com/ (Synthetic Classification Data Sets from the Internet).

To evaluate our algorithm, we used several databases of different sizes, where the largest database included 40000 transactions involving over 1000 items. But for narrating

convenience, we choose below to analyze the result of the smallest of our databases.

The database consists of 100 transactions involving over 15 items. Experimental results of the k-nearest neighbor, correlation analysis, and Apriori are shown in Tables VI to VIII.

TABLE VI
K-NEAREST NEIGHBOR

MID	Nearest neighbor	MID	Nearest neighbor
0	5,7,13	8	3,4,6,13
1	5,7,11,13	9	12,13
2	5,13	10	13
3	8,5,6,4,13	11	1,13,14
4	6,5,3,8,13	12	9,13
5	0,1,2,3,4,7,13	13	0,1,2,3,4,5, 6,7,8,9,10,11,12
6	4,3,8,13	14	11
7	0,1,5,13		

TABLE VII
CORRELATION ANALYSIS

Positive correlated itemsets	Negative correlated itemsets
(0,1),(0,4),(0,8),(0,10),(1,4) (1,8),(1,10),(1,14),(2,3),(2,5) (2,10),(2,13),(2,14),(3,5),(3,6) (3,11),(3,12),(4,8),(4,10),(4,14) (5,10),(5,14),(6,7),(6,11),(6,12) (7,12),(8,9),(8,12),(10,13) (12,13),(0,7,9),(0,7,13),(0,9,13) (5,9,13),(6,8,13),(6,9,13) (9,10,14)	(0,3),(0,5),(0,11),(1,2),(1,3) (1,5),(1,6),(1,11),(1,12),(2,4) (2,8),(2,11),(2,12),(3,4),(3,8) (4,5),(4,6),(4,11),(5,6),(5,7) (5,8),(5,12),(6,10),(6,14)(7,14) (8,11),(9,11),(10,11),(10,12) (12,14)

TABLE VIII
COMPARISON BETWEEN SUPPORT AND VARIANCE

Merchandise ID	Support	Variance
0	0.43	1. 374705
1	0.43	0. 7838024
2	0.20	1. 7
3	0.50	1. 310159
4	0.40	1. 099296
5	0.24	7. 706056
6	0.35	1. 471311
7	0.55	3. 420516
8	0.60	1. 004835
9	0.83	0. 7899691
10	0.60	1. 109732
11	0.09	1. 05
12	0.18	25. 85556
13	0.74	1. 00284
14	0.12	2. 422

A. Algorithm Analysis

Simulation database

(1) Our algorithms are satisfactory from maturity

From Table VIII we can see clearly that there is no relation between the merchandise's variance and support. For example, in this experiment the information for merchandise 7 and 12 is shown.

(1) supp(7apple)=0.55>>supp(12cashew)=0.18
(2) variance(7apple)=3.2886<<variance(12cashew)=23.1667

Here, (1) means the sale of apples is higher than that of cashews, but in this experiment cashews are obviously more significant than apples. Meanwhile, the variance calculated by our algorithm has shown that, (2) means that the item cashew will be found more easily than apple by our algorithm. Our algorithm can find this type of against-expectation itemset, so the algorithm for finding interesting itemsets satisfies our needs.

Two points can confirm this. First, we use the increment rate of each of two adjacent phases to calculate the variance. This way, we can generate some against-expectation itemsets, which are always omitted easily because they are infrequent. Second, because the pruning structure is based on calculating the variance, we can keep all the interesting itemsets we need. We just require an appropriate threshold.

In addition, the k-nearest neighbor graph is needed to find the nearest neighbors of each interesting goods item, so it can be successful in employing a suitable threshold. Further, the correlation analysis must not miss interesting itemsets during its operation of visiting and pruning the tree.

(2) Our algorithm is very accurate

This experiment demonstrates that the desired result can be obtained. First, there are no uninteresting association rules. Although

$$supp(8 \to 9) = supp(10 \to 13) = 60\%$$
$$conf(8 \to 9) = conf(10 \to 13) = 100\%$$

Meanwhile, with the k-nearest neighbor graph, we get:

8: 3, 4, 6, 13,

9: 12, 13,

10: 13,

13: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,

It can easily be seen that 8 toothbrush is closest to 3 toothpaste, and has nothing to do with 9 bread. 10 and 13 are beer and milk respectively, and their result is $Corr(8,9) = 1.2 < Corr(10,13) = 1.4$. So toothbrush and bread is an

uninteresting rule. Although supp(9→13) = 73%, it is not useful in decision-making because of the closed correlation between the two. Meanwhile, the k-nearest neighbor graph reveals that they are not very closed, and the correlation analysis reveals that they are not independent.

The most important is that the k-nearest neighbor graph can avoid being misled by the actual sale of merchandises, making the increment rate of the analysis objective. In addition, the output patterns of the two algorithms are very clear.

(3) **The advantages of the two algorithms still exist, even if the database changes**

In addition to the above experimental analysis, we have performed several experiments on databases of different sizes to illustrate the efficiency of the algorithms. In Tables IX, X and XI, we have chosen the same transactions, the same items ((T, I) = (100, 10), (T, I) = (100, 15), (T, I) = (150, 10), (T, I) = (150, 15)), and the different (or 'various') average items of transactions (A). In Table IX, we present the performance of the Correlation analysis, based on the same transaction (T) and different items (I), the same item and different transactions, and the average number of items of transactions (A) = 5. We do the same in Table X and Table XI, (A) = 8, (A) = 10 respectively. Table X shows the runtime of the k-nearest neighbor graph based on different simulation databases. We evaluat the number of positive correlation items (pc), negative correlation items (nc), and overall correlation items (oc), and also the runtime (rt).

TABLE IX
CORRELATION ANALYSIS ON (A)=5

(T, I)	pc	nc	oc	rt
(100,10)	38	9	47	0.156
(100,15)	45	43	88	100.906
(150,10)	32	8	40	0.172
(150,15)	49	49	98	101.141

TABLE X
CORRELATION ANALYSIS ON (A)=8

(T, I)	pc	nc	oc	rt
(100,10)	48	0	48	0.172
(100,15)	125	17	142	100.89
(150,10)	56	0	56	0.218
(150,15)	131	17	148	100.672

TABLE XI
CORRELATION ANALYSIS ON (A)=10

(T, I)	pc	nc	oc	rt
(100,10)	50	0	50	0.234
(100,15)	203	10	213	103.359
(150,10)	40	0	40	0.312
(150,15)	259	10	269	101.141

TABLE XII
K-NEAREST NEIGHBOR GRAPH

(T, I, A)	rt	(T, I, A)	rt
(100,10,5)	0	(150,10,5)	0
(100,10,8)	0	(150,10,8)	0
(100,10,10)	0	(150,10,10)	0.016
(200,10,5)	0.015	(100,15,5)	0
(200,10,8)	0	(100,15,8)	0
(200,10,10)	0	(100,15,10)	0

The runtime equals 0 because the system's capability is restricted, so we have chosen runtime = 0 when runtime < 0.001.)

Tables IX, X, XI and XII reveal the following results:

1: The number of correlation itemsets is most deeply influenced by the number of items (I). The greater the number of items, the more correlation itemsets there are.

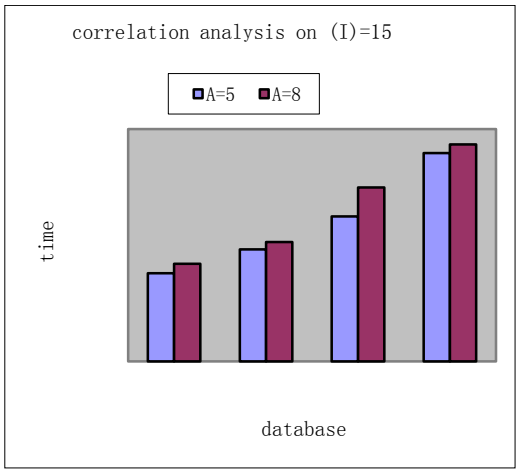
2: The length of the correlation itemsets is affected by the average length of items (A). Long average length results in long correlation itemsets.

3: Compared with the other two factors, running time places most stress on the number of items. Therefore, the algorithm is more effective on a dense database than on a sparse one. Meanwhile, from Table XII, we can see clearly that the k-nearest neighbor graph algorithm can work satisfactorily.

4: The number of negative correlation items is influenced by the average length of items (A), as well as the number of items (I). As the average length of items (A) is invariant, the more the number of items is, the more the number of itemsets. As the number of items (I) is invariant, the less the length of the item, and the more the number of negative correlation items.

Simulation on larger databases

We have performed these experiments on databases which involve 2000, 3000, 5000 and 7500 transactions on 10 or 15 items, where the average length of items is 5 or 8. The results are shown in Fig. 5.



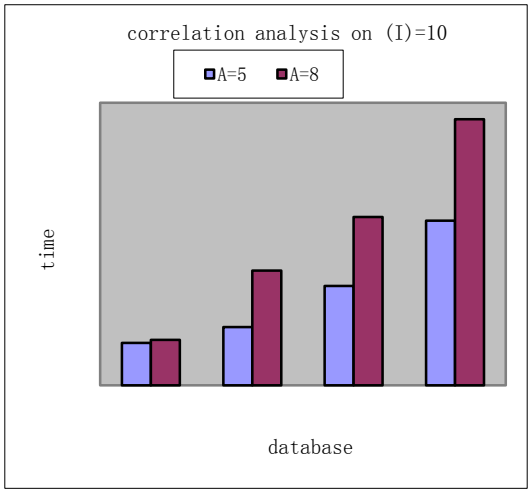
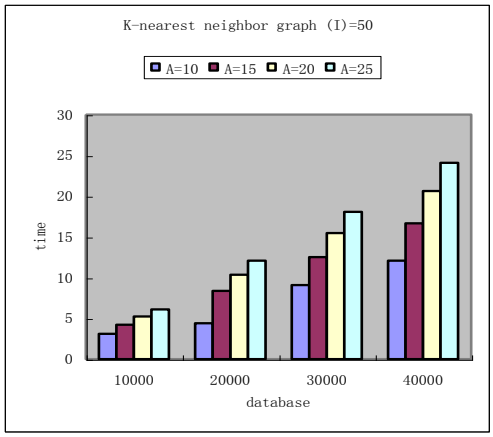
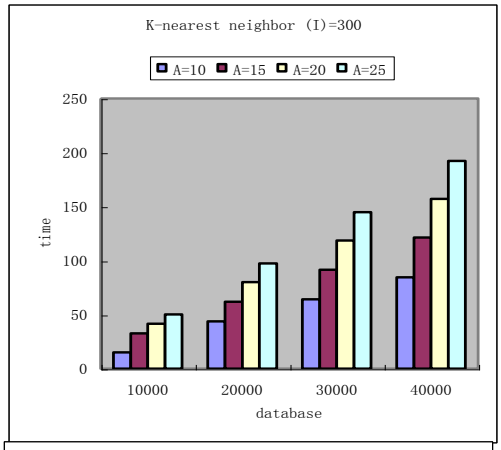
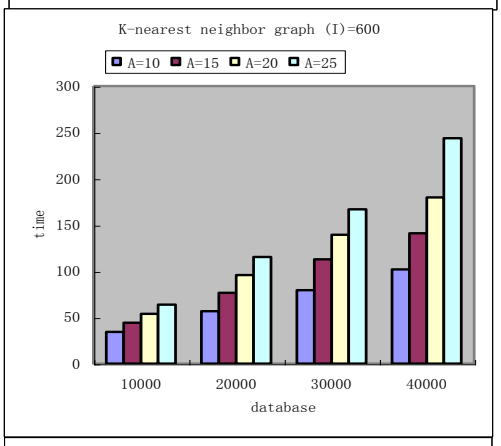


Fig. 5. Runtime of the correlation analysis graph based on different databases







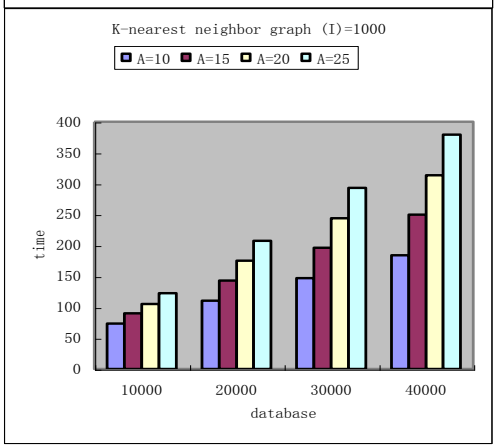


Fig. 6. The runtime of the k-nearest neighbor graph based on different factors

From Fig. 6, we can see clearly that the consuming runtime increases with the growth of the size of the database, namely, the number of transactions(T), items(I), and the average length of items(A).

Meanwhile the rules, which were found in the simulation small databases, as shown in Tables XI–XII, are still appropriate.

B. Comparison between two algorithms

The k-nearest neighbor graph and correlation analysis are evaluated, either one of which can compensate for the drawbacks of the association rule with the support-confidence model, and is useful for static classification.

The former enhances correctness by introducing relative bargain quantity, and considering the increment to be the

quotient. Further, the manner of its result output can be understood, because the sequence of all the nearest neighbors of each of the goods goes from strong to weak.

Correlation analysis can enhance correctness and reduce time costs through pruning. Introducing the fuzzy theorem makes the result more reasonable. Finally, it is convenient for decision makers to distinguish positive correlation from negative correlation for each interesting item.

V. Conclusion

We have designed a new algorithm for identifying against-expectation patterns. These patterns are interactions within items, with large relative-contrasts referenced to their expectations in a given database. This is based on heterogeneity metrics. The techniques for mining against-expectation patterns were previously undeveloped. We have experimentally evaluated our algorithms and demonstrated that our approach is efficient and promising.

Acknowledgment

This work was supported in part by the Australian Research Council (ARC) under grant DP0988016, the Nature Science Foundation (NSF) of China under grant 90718020, the China 973 Program under grant 2008CB317108, and the Guangxi NSF under grant GKZ0640069.

References

[1] R. Agrawal, T. Imielinski, and A. Swami (1993), Mining association rules between sets of items in large databases. In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 207-216.

[2] Brock Barber and Howard J. Hamilton (2003): Extracting Share Frequent Itemsets with Infrequent Subsets. Data Min. Knowl. Discov. 7(2): 153-185.

[3] S. Bay and M. Pazzani (2001), Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 5(3): 213-246.

[4] S. Brin, R. Motwani and C. Silverstein (1997), Beyond Market Baskets: Generalizing Association Rules to Correlations. In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 265-276.

[5] Y.B. Cho, Y.H. Cho and S. Kim (2005), Mining changes in customer buying behavior for collaborative recommendations. *Expert Systems with Applications*, 28(2): 359-369.

[6] L. Egghe and C. Michel (2003), Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. *Information Processing and Management*, 39: 771-807.

[7] Ping-Yu Hsu, Yen-Liang Chen and Chun-Ching Ling (2004), Algorithms for mining association rules in bag databases**.** *Information Sciences*, Volume 166, Issues 1-4: 31-47.

[8] F. Hussain, H. Liu, E. Suzuki, and H. Lu (2000), Exception Rule Mining with a Relative Interestiness Measure. In: *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD), pp. 86-97.

[9] S. Hwang, S. Ho, and J. Tang (1999), Mining Exception Instances to Facilitate Workflow Exception Handling. In: *Proceedings of the International Conference on Database Systems for Advanced Applications* (DASFAA), pp. 45-52.

[10] G. Karypis, E. Han, and V. Kumar (1999), CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. *IEEE Computer*, pp. 68-75.

[11] Xuemin Lin, Yijun Li and Chi-Ping Tsang (1999), Applying on-line bitmap indexing to reduce counting costs in mining association rules. *Information Sciences*, Volume 120, Issues 1-4: 197-208.

[12] B. Liu, W. Hsu, H. Han and Y. Xia (2000), Mining changes for real-life applications. In: Second International Conference on Data Warehousing and Knowledge Discovery, 337-346.

[13] B. Liu, W. Hsu, L. Mun and H. Lee (1999), Finding interesting patterns using user expectations. In: IEEE Transactions on Knowledge and Data Engineering, (11)6, 817-832.

[14] H. Liu, H. Lu, L. Feng, and F. Hussain (1999), Efficient Search of Reliable Exceptions. In: *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD), pp. 194-204.

[15] B. Padmanabhan and A. Tuzhilin (1998), A Belief-Driven Method for Discovering Unexpected Patterns. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining* (KDD), pp. 94-100.

[16] B. Padmanabhan and A. Tuzhilin (2000), Small is beautiful: discovering the minimal set of unexpected patterns. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining* (KDD), pp. 54-63.

[17] G. Piatetsky-Shapiro (1991), Discovery, analysis, and presentation of strong rules. In: *Knowledge discovery in Databases*, G. Piatetsky-Shapiro and W. Frawley (Eds.), AAAI Press/MIT Press, pp229-248.

[18] Savasere, E. Omiecinski, and S. Navathe (1998), Mining for Strong Negative Associations in a Large Database of Customer Transactions. In: *Proceedings of the International Conference on Data Engineering* (ICDE), pp. 494-502.

[19] E. Suzuki and M. Shimura (1996), Exceptional Knowledge Discovery in Databases Based on Information Theory. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining* (KDD), pp. 275-278.

[20] K. Wang, S. Zhou, A. Fu and X. Yu (2003). Mining Changes of Classification by Correspondence Tracing. SIAMDM'03, 2003.

[21] G. Webb (2000). Efficient search for association rules. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining* (KDD), pp. 99-107.

[22] G. Webb, S. Butler and D. Newlands (2003), On detecting differences between groups. *KDD'03*, pp. 256-265.

[23] X. Wu, C. Zhang and S. Zhang (2002). Mining both positive and negative association rules. In: *Proceedings of 21st International Conference on Machine Learning* (ICML), pp. 658-665.

[24] Shichao Zhang, Jingli Lu and Chengqi Zhang (2004), A fuzzy logic based method to acquire user threshold of minimum-support for mining association rules. *Information Sciences*, Volume 164, Issues 1-4: 1-16.

*K*NN-CF Approach: Incorporating Certainty Factor to *k*NN Classification

Shichao Zhang

***Abstract*—*K*NN classification finds *k* nearest neighbors of a query in training data and then predicts the class of the query as the most frequent one occurring in the neighbors. This is a typical method based on the majority rule. Although majority-rule based methods have widely and successfully been used in real applications, they can be unsuitable to the learning setting of skewed class distribution. This paper incorporates certainty factor (CF) measure to *k*NN classification, called *k*NN-CF classification, so as to deal with the above issue. This CF-measure based strategy can be applied on the top of a *k*NN classification algorithm (or a hot-deck method) to meet the need of imbalanced learning. This leads to that an existing *k*NN classification algorithm can easily be extended to the setting of skewed class distribution. Some experiments are conducted for evaluating the efficiency, and demonstrate that the *k*NN-CF classification outperforms standard *k*NN classification in accuracy**.

***Index Terms*—Classification, *k*NN classification, imbalanced classification.**

I. Introduction

GIVEN its simplicity, easy-understanding and relatively high accuracy, the k-nearest neighbor (*k*NN) approach has successfully been used in diverse data analysis applications [4,10,31,35] such as information retrieval, database, pattern recognition, data mining and machine learning. In information retrieval application proposal, the *k*NN approach is used to, for instance, similarity searching [42], text categorization, ranking and indexing [2,61]. In database application proposal, the *k*NN approach is used to, such as, approximate query and high dimensional data query [11,49]. In pattern recognition application proposal, the *k*NN approach is used to, for example, segmentation and prediction [13,45]. In data mining and machine learning application proposal, the *k*NN approach is used to, for example, clustering and classification [14,22,23,26], manifold learning [50,57,58], and missing data imputation for data preparation [64,65]. Therefore, it has recently been recognized as one of top 10 algorithms in data mining [60].

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.

The *k*NN method provides k data points in a given dataset most relevant to a query in a data analysis application. Without other information, the k most relevant data are k nearest neighbors of the query in the dataset. And then predicts the class of the query as the most frequent one occurring in the neighbors. This is a typical method based on the majority rule. Majority-rule based methods have widely and successfully been used in real applications. They can, however, be unsuitable to the learning setting of skewed class distribution. This is illustrated with Example 1 as follows.

Eample 1. Consider some data drawn from a dataset with skewed class distribution, as listed in Table I, or charted in Fig. 1. In Table I, X1 and X2 are two attributes, C is the class attribute (or decision attribute), "+" and "–" stand for the two classes, "?" denotes the unlabeled class.

TABLE I

Data from the questionnaire survey

X1	3	4	4	4	6	7	7	8	4	5	4
X2	7	3	4	10	2	4	9	5	6	5	5
C	+	+	+	+	+	+	+	+	–	–	?

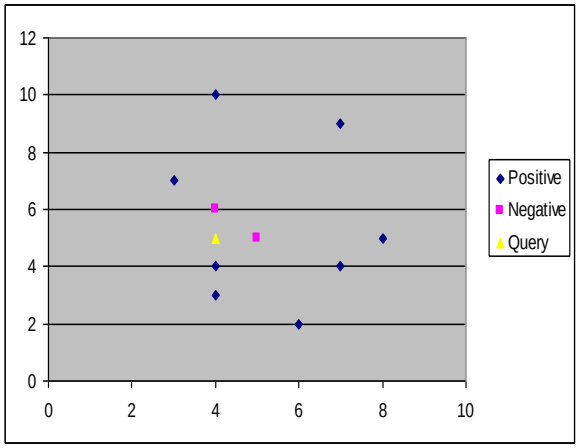


Fig. 1. Plotting the data in Table

Assume k = 5. For the query (4, 5, ?), we can obtain its 5 nearest neighbors in Table I, (3, 7, +), (4, 3, +), (4, 4, +), (4, 6, –), (5, 5, –). According to the *k*NN classification, "+" is the most frequent one occurring in the neighbors. Consequently, the class of the query (4, 5, ?) is predicted as "+". The first

feedback seems to be that the class of (4, 5, ?) should be predicted as "–", although the majority rule predicts it as "+".

To attack the above actual and challenging issue, this paper incorporates certainty factor (CF) measure to *k*NN classification, denoted as *k*NN-CF classification. The main idea is as follows. We have p(C = +) = 0.8 and p(C = –) = 0.2 in the dataset in Table I. The selected 5 nearest neighbors can be taken as a new evidence, E, and p(C = +|E) = 0.3 and p(C = –|E) = 0.2. Clearly, compared with their prior probabilities, the conditional probability of "–" is lifted much more than that of "+". Accordingly, it is reasonable to predict "–" as the class of (4, 5, ?). The CF measure can capture this ad hoc nature. And we will be incorporated to the a *k*NN classification in this paper, called ***k*NN-CF classification**.

This *k*NN-CF strategy can be applied on the top of a *k*NN classification algorithm (or a hot-deck method, or an instance-based algorithm) to meet the need of imbalanced learning. This leads to that an existing *k*NN classification algorithm can easily be extended to the learning setting of skewed class distribution. Some experiments are conducted for evaluating the efficiency, and demonstrate that the *k*NN-CF classification outperforms standard *k*NN classification in accuracy.

The rest of this paper is organized as follows. Section II briefly recalls work mainly related to *k*NN classification, imbalanced classification and certainty factor measure. The *k*NN-CF classification is presented in Section III. We evaluate the *k*NN-CF classification with real datasets downloaded from UCI in Section IV. This paper is concluded in Section V.

II. PRELIMINARY

This section presents some basic concepts and briefly recalls related work in *k*NN classification, imbalanced classification and certainty factor measure.

A. Research into kNN Approach

KNN approach has recently been recognized as one of top 10 algorithms in data mining [60], due to its high classification accuracy in problems with unknown and nonnormal distributions [16,26,31] and its wide applications [35,4]. While NN (nearest neighbor) classification suffers from the issue of overfitting, a more sophisticated approach, *k*NN classification [21], finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood [1,12]. There are three key elements of this approach: a set of labeled objects, e.g., a set of stored records, a distance or similarity metric to compute distance between objects, and the value of k, the number of nearest neighbors. To classify an unlabeled object, the distance of this object to the labeled objects is computed, its k-nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the object [60].

The *k*NN classification has the remarkable property that under very mild conditions, the error rate of a *k*NN classifier tends to the Bayes optimal as the sample size tends to infinity. However, there are several key issues that affect the performance of *k*NN, mainly including the choice of k, predicting the class labels of new data, distance measure selection, and lazy learning. For details, please read the paper [60].

Therefore, many techniques have recently been developed for improving the *k*NN classification. Among them, distance measure selection is relatively hot research topic [16,24,27,36]. Currently a variety of measures such as Euclidean, Hamming, Minkowsky, Mahalanobis, Camberra, Chebychev, Quadratic, Correlation, Chi-square, hyperrectangle distance [41], Value Difference Metric [47], and Minimal Risk Metric [5] are available. However, no distance function is known to perform consistently well, even under some conditions [54]. This makes the use of *k*NN highly experience dependent. Various attempts have been made to remedy this situation. Among those, notably DANN carries out a local linear discriminant analysis to deform the distance metric based on say 50 nearest neighbors [24]. LFM-SVM also deforms the metric by feature weighting, where the weights are inferred from training an SVM on the entire data set [15]. H*k*NN applies the collection of 15-70 nearest neighbors from each class to span a linear subspace for that class, and then classification is done based not on distance to prototypes but on distance to the linear subspaces [53]. There are other kinds of distance defined by the property of data. Examples are tangent distance on the USPS zip code data set [44], shape context based distance on the MNIST digit data set [3], distances between histograms of textons on the CUReT data set [52], and geometric blur based distances on Caltech-101 [70]. Furthermore these measures can be extended by kernel techniques such as to estimate a curved local neighborhood [37], which can make the space around the samples further or closer to the query, depending on their class-conditional probability distributions. More recently, a new measure named neighborhood counting is proposed to define the similarity between two data points by using the number of neighborhoods [54]. Because features of high dimensional data is often correlated so that measure easily becomes meaningless, some approaches are designed to deal with this issue such as an approach that applies variable aggregation to define the measure [16,26]. Besides all kinds of measures above, the other strategy can be also applied to select nearest neighbors. For example, an approach is proposed that considers the geometrical placement of neighbors more than actual distances to appropriately characterize a sample by its neighborhood [43]. This approach is effective in some case, but it is conflict with our intuition when data is on manifold.

Other main improvements of *k*NN classification include, such as fuzzy set theory and evidential reasoning [68], measures for finding the better nearest neighbors [16,26,54], and local mean classifiers (LMC) [8,31,34,62].

B. Research into Imbalanced Classification

The class imbalance (or skewed class distribution) is relatively a new issue in data mining and machine learning. While it was recognized that the imbalance can cause suboptimal classification performance, there are many research reported on imbalanced learning since two workshops "AAAI'2000 Workshop on Learning from Imbalanced Data Sets" and "ICML'2003 Workshop on Learning from

Imbalanced Data Sets". The main efforts include, for example, the detection of software defects in large software systems [33], the identification of oil spills in satellite radar images [28], the detection of fraudulent calls [20], and the diagnoses of rare medical conditions such as the thyroid disease [40].

In the setting of skewed class distribution, it is obvious that the rare instances in these applications are of critical importance. And classification learning should be able to achieve accurate classification for the rare classes. Typically the rare instances are much harder to identify than the majority instances. Different from traditional classification desired a high overall accuracy, the purpose of imbalanced learning is to achieve accurate classification for the rare class without sacrificing the performance for other classes.

While existing classification algorithms work well on the majority classes, there have been several attempts to adjust the decision bias favourable to the minority class. Holte et al. [25] modified the bias of CN2 classifier, by using the maximum generality bias for large disjuncts and a selective specificity bias for small disjuncts. Another piece of work is by Ting [51], where a hybrid approach for addressing the imbalanced problem was proposed. This method adopted C4.5 as the base learner, then an instance-based classifier was used if small disjuncts were encountered. Similar approaches were proposed by [6,7], using a combination of the genetic algorithm and the C4.5 decision tree. However, their experimental results show no statistically significant difference from the base C4.5 learner.

Re-sampling techniques have been a popular strategy to tackle the imbalanced learning problem, including random over-sampling and under-sampling, as well as intelligent re-sampling. Chawla, et al. [9] proposed a synthetic minority over-sampling technique to over-sample the minority class. Kubat and Matwin [29] tried to under-sample the majority class. Another related work by Ling and Li [32] combined over-sampling of the minority class with undersampling of the majority class. However, no consistent conclusions have been drawn from these studies [55]. The effect of re-sampling techniques for active learning was analysed in [69]. They found over-sampling is a more appropriate choice than under-sampling which could cause negative effects on active learning. A bootstrap-based over-sampling approached was proposed, and it was shown to work better than ordinary over-sampling in active learning for word sense disambiguation.

The second strategy tackling the imbalanced distribution problem is cost-sensitive learning [17,19,66]. Domingos [18] proposed a re-costing method called MetaCost, which can be applied to general classifiers. The approach made error-based classifiers cost-sensitive. His experimental results showed that MetaCost reduced costs compared to cost-blind classifier using C4.5Rules as the baseline.

Ensemble learning has also been studied for imbalanced classification. Sun et al. [48] tried to use boosting technique for imbalanced learning, and three cost-sensitive boosting algorithms were introduced. These boosting algorithms were investigated with respect to their weighting strategies towards different types of samples. Their effectiveness in identifying rare cases on several real-world medical datasets with imbalanced class distribution were examined. An empirical study by Seiffert et al. [73] compared the performance between re-weighting and re-sampling boosting implementations in imbalanced datasets. They found that boosting by re-sampling outperforms boosting by re-weighting, which is often the default boosting implementation.

A potential strategy is the instance-based learning that will be built in Section III. The ubiquitous instance-based learning paradigm is rooted in the *k*NN algorithm. Most research efforts in this area have been on trying to improve the classification efficiency of *k*NN [1,59]. Various strategies have been proposed to avoid an exhaustive search of all training instances and to achieve accurate classification. However, to the best of our knowledge, no work has been reported adjusting the induction bias of *k*NN for imbalanced classification.

C. Research into Certainty Factor Measure

The certainty-factor (CF) model is a method for managing uncertainty in rule-based systems. Shortliffe and Buchanan [46] developed the CF model in the mid-1970s for MYCIN, an expert system for the diagnosis and treatment of meningitis and infections of the blood. Since then, the CF model has become the standard approach to uncertainty management in rule-based systems. A certainty factor is used to express how accurate, truthful, or reliable you judge a predicate to be. Note that a certainty factor is neither a probability nor a truth value. Therefore, it is slightly dodgy theoretically, but in practice this tends not to matter too much. This is mainly because the error in dealing with certainties tends to lie as much in the certainty factors attached to the rules (or in conditional probabilities assigned to things) as in how they are manipulated.

A certainty factor is a number between -1 and 1 (or in [-1, 1]) that represents the change in our belief on some hypothesis. A positive number means an increase in the belief and a negative number the contrary. A value of 0 means that there is no change in our belief on the hypothesis.

The CF measure has successfully been used to identify both positive and negative association rules in datasets [71,72]. This leads to the fact that a framework was built for complete association analysis (both positive and negative association rules).

III. KNN-CF CLASSIFICATION

For simplifying the description, we adopt the CF measure in [72] for building the framework of the kNN-CF classification. Before presenting the *k*NN-CF classification, a simple measure, called FR (frequency ratio), is introduced in Section III.A.

A. KNN Classification Based on FR measure

Let D be a training set, C = {c1, c2, …, cm} a set of labels, Q a query, N(Q, k) the set of k nearest neighbors, f(C= ci, D) the frequency of ci in D, and f(C= ci, N(Q,k)) the frequency of ci in N(Q, k). We define the FR measure as follows.

$$\mathrm{FR}(C=c_i) = \frac{f(C=c_i, N(Q,k))}{f(C=c_i, D)} \quad . \qquad (1)$$

The FR strategy for *k*NN classification is defined as follows. We first obtain

$$S_{FR} = \{ \arg\max_{1\le i\le m} \{FR(C = c_i)\} \}. \quad (2)$$

Because there may be one more classes satisfy $\arg\max_{1\le i\le m} \{FR(C = c_i)\}$, $|S_{FR}|$ can be greater than 1. Accordingly, we can predict the class c of Q with Formula (3) as follows.

$$c = \arg\max_{c_j \in S_{FR}} \{c_j\} \quad . \quad (3)$$

We illustrate the use of FR measure with the data in Example 1. From Table I and the 5 nearest neighbors of the query, the frequency of classes "+" and "–" can be computed as follows:

f(C=+, D) = 8

f(C=–, D) = 2

f(C=+, N(Q,5)) = 3

f(C=–, N(Q,5)) = 2.

Consequently, we can obtain

$$\text{FR(C= +)} = \frac{f(C=+, N(Q,5))}{f(C=+, D)} = \frac{3}{8} = 0.375$$

$$\text{FR(C= –)} = \frac{f(C=-, N(Q,5))}{f(C=-, D)} = \frac{2}{2} = 1.$$

Therefore, we can predict the class c of the query Q as follows.

$$S_{FR} = \{ \arg\max_{1\le i\le m} \{FR(C = c_i)\} \} = \{-\},$$
$$c = \arg\max_{c_j \in S_{FR}} \{c_j\} = -.$$

From the above, although class "+" is the most frequent one occurring in N(Q, 5), its frequency ratio, FR(C= +) = 0.375, is much low than that of class "–", FR(C= –) = 1. Therefore, it is reasonable to predict "–" as the class of (4, 5, ?).

The FR is a simple and efficient strategy. It is similar the "lift" measure in data mining and machine learning, which is a measure of the performance of a model at predicting or classifying cases, measuring against a random choice model (adopted from Wikipedia).

Certainly, we can replace FR with the odds ratio. The use of odds ratio to *k*NN classification is similar to that of FR strategy.

B. KNN Classification Based on CF measure

With the assumption in Section III.A, we incorporate the CF measure to *k*NN classification as follows. Assume p(C= ci |D) is the ratio of ci in training set D, p(C= ci |N(Q,k)) is the ratio of ci in the set of k nearest neighbors, N(Q, k). If p(C= ci |N(Q,k)) ≥ p(C= ci |D), the CF is computed with (4) as follows.

$$\text{CF(C= ci, N(Q,k))} = \frac{p(C = c_i \mid N(Q,k)) - p(C = c_i \mid D)}{1 - p(C = c_i \mid D)} . \quad (4)$$

If p(C= ci |N(Q,k)) < p(C= ci |D), the CF is computed with (5) as follows.

$$\text{CF(C= ci, N(Q,k))} = \frac{p(C = c_i \mid N(Q,k)) - p(C = c_i \mid D)}{p(C = c_i \mid D)} . \quad (5)$$

According to the explanation of CF, CF(C= ci, N(Q,k)) is valued in [-1, 1]. If CF(C= ci, N(Q,k)) > 0, our belief on that the class of the query should be predicted as C= ci is increased. CF(C= ci, N(Q,k)) < 0, our belief on that the class of the query should be predicted as C= ci is decreased. CF(C= ci, N(Q,k)) = 0, our belief on that the class of the query should be predicted as C= ci is the same as that in the training set D.

The CF strategy for kNN classification is defined as follows. We first obtain

$$S_{CF} = \{ \arg\max_{1\le i\le m} \{CF(C = c_i, N(Q,k))\} \}. \quad (6)$$

Because there may be one more classes satisfy $\arg\max_{1\le i\le m} \{CF(C = c_i, N(Q,k))\}$, $|S_{CF}|$ can be greater than 1. Accordingly, we can predict the class c of Q with Formula (7) as follows.

$$c = \arg\max_{c_j \in S_{CF}} \{c_j\} . \quad (7)$$

Also, we illustrate the use of CF measure with the data in Example 1. Because f(C=+, D) = 8, f(C=–, |D) = 2, f(C=+, N(Q,5)) = 3 and f(C=–, N(Q,5)) = 2, we have p(C=+|D) = 0.8, p(C=–|D) = 0.2, p(C=+| N(Q,5)) = 0.6 and p(C=–| N(Q,5)) = 0.4. Because p(C=+| N(Q,5)) < p(C=+|D), we should calculate the CF of "+" with (3) as follows

$$\text{CF(C= +, N(Q,5))} = \frac{p(C = + \mid N(Q,5)) - p(C = + \mid D)}{p(C = + \mid D)} = \frac{0.6 - 0.8}{0.8} = -0.25$$

Because p(C=–| N(Q,5)) > p(C=–|D), we should calculate the CF of "–" with (2) as follows

$$\text{CF(C= –, N(Q,5))} = \frac{p(C=-\mid N(Q,5)) - p(C=-\mid D)}{p(C=-\mid D)} = \frac{0.4-0.2}{1-0.2} = 0.25$$

Therefore, we can predict the class c of the query Q as follows.

$$S_{CF} = \{ \arg\max_{1\le i\le m} \{CF(C=c_i, N(Q,k))\} \} = \{-\}$$
$$c = \arg\max_{c_j \in S_{CF}} \{c_j\} = -.$$

From the above, although class "+" is the most frequent one occurring in N(Q, 5), its frequency ratio, FR(C= +) = 0.375, is much low than that of class "–", FR(C= –) = 1. Therefore, it is reasonable to predict "–" as the class of (4, 5, ?).

From CF(C= +, N(Q,5)) = –0.25 and CF(C= –, N(Q,5)) = 0.25, it is reasonable to predict "–" as the class of (4, 5, ?).

C. Analysis

*K*NN classification is a lazy learning technique, or instance-based learning/reasoning method. Different from model-based algorithms (training models from a given dataset and then predicting a query with the models), it needs to store the training data (or cases) in memory and to compute the most relevant data to answer a given query. The answer to the query is the class represented by a majority of the k nearest neighbors. This is the majority rule. Although *k*NN classification with majority rule is simple and effective in general, there are still some limitations from an applied context, for example, cost-sensitive learning and imbalanced classification applications. Therefore, there are great many improvement efforts. We briefly discuss them from three directions as follows.

The first direction is the distance weighted *k*NN rule. Almost all improvement efforts belong to this direction. This direction is actually a selection of the k nearest neighbors for a given query. This is because different distance functions or weighting techniques (or both) can generate different k nearest neighbors only. Whatever the distance functions or weighting techniques are selected, the goal is to find a machine that highlights some attributes and decreases the impact of the rest on the query. This looks like a mapping that transforms the original space to a new space more suitable to a learning task. It is much clear when we apply the λ-cutting rule to such an algorithm. With the λ-cutting rule, the distance weighted *k*NN classification will be carried out on only those data points that the attributes are stretched out or drawn back, or a subspace consisting of attributes with the impact values equal to or greater than λ, or a combination among them.

A lately selection of the nearest neighbors is the SN (Shelly Neighbors) method that uses only those neighbors that form a shell to encapsidate the query, drawn from the k nearest neighbors [64,65]. The SN approach is actually a quadratic selection of the k nearest neighbors.

The second direction is the semi-lazy learning. This direction is actually a procedure of reducing time and space complexity. The *k*NN classification approach usually involves storing the training data in memory and completely search the training data for the k nearest neighbors. If we can properly divide the training set into n subsets and search for the k nearest neighbors from only the nearest subsets, its time and space complexity must be decreased to an acceptable computation level.

Last direction is the prediction of the query (the decision phase with the the k nearest neighbors). The usually used methods include the majority rule, weighting machine, and the Bayesian rule. The *k*NN-CF classification is a new technique that is designed against the issue of imbalanced classification.

From Section III.B, it is simple and understandable to incorporate the CF measure to *k*NN classification. It advocates to take into account the certainty factor of a classification decision when using *k*NN classification approach.

For imbalanced classification, the uncertainty is often occurred in the junction between the majority class and minority class. In this setting, the majority class certainly wins minority class in general. The Example 1 has also illustrated this uncertainty. This may lead to high cost (or risk) in many real applications, such cancer detection. The main objective of introducing the CF measure to *k*NN classification is to distinguish those classes with increased certainty factor from the classes with decreased ones.

The *k*NN-CF classification is only an idea to improve the decision phase with the the k nearest neighbors. There are some challenging issues. For example, it should be a research topic to study a new method for addressing, such as Case-1 and 2 in Figs. 2 and 3 respectively.

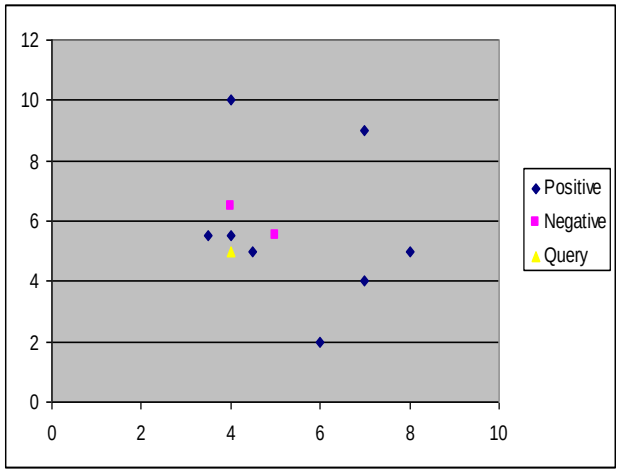


Fig. 2. Case-1 faced by the *k*NN-CF classification

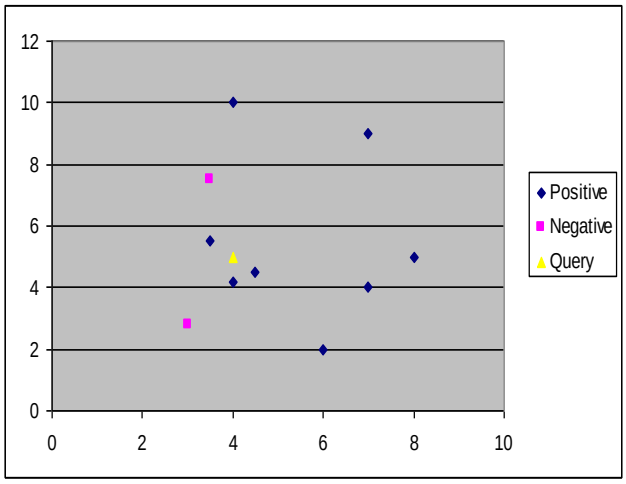


Fig. 3. Case-2 faced by the *k*NN-CF classification

From Zhang [64,65], seems the SN approach is suitable to deal with the above issues. All the above issues are studied against the joint between a minority class across and a majority. While the joint is of uncertainty, the rest are of certainty. Let ci be a majority class, cj a minority class, and Q a query. We can easily prove the following corollaries.

Corollary 1. The FR strategy is equivalent to the majority rule for kNN classification when FR(C= ci) ⩾FR(C= cj).

Corollary 2. The CF strategy is equivalent to the majority rule for *k*NN classification when CF(C= ci, N(Q,k)) ⩾CF(C= cj, N(Q,k)).

IV. Experiments

In order to show the effectiveness of the FR and CF strategies, two sets of experiments were done on real datasets with the algorithm implemented in C++ and executed using a DELL Workstation PWS650 with 2G main memory, and 2.6G CPU.

A. Settings of experiments

The first set of experiments was conducted for examining the efficiency against data points with pure minority class, or with pure majority class. The second set of experiments was conducted on for examining the efficiency against data points randomly drew from a dataset. Because the FR strategy is equivalent to the CF strategy, we only compare the CF strategy with Standard *k*NN approach in the following experiments. In the two sets of experiments, for simplifying the description, we always compared the proposed approaches with standard *k*NN classification. We adopt the recall and precision to evaluate the efficiecy by taking into account four distributions of minority and majority classes: 10% : 90%; 20% : 80%; 30% : 70%; 40% : 60%. For evaluating the recall and precision, all queries are randomly generated from those data points that their classes are known in a dataset. The datasets are summarized in Table II.

TABLE II

The summary of Datasets

Data set	No. of instances	Class dist. (N/P)	No. of features	No. of classes
Breast-w	683	444/239	9	2
Haberman	306	225/81	3	2
Parkinsons	195	147/48	22	2
Transfusion	748	570/178	4	2
Magic	19020	12332/6688	11	2
Ionosphere	351	225/126	33	2
Pima	768	500/268	8	2
Spambase	4601	2788/1813	57	2
SPECTF	267	212/55	44	2
wdbc	569	357/212	30	2

B. The first group of experiments

We examine the efficiency against data points with pure minority class, or with pure majority class. The results are showed in Tables III - VI as follows.

TABLE III

Standard *k*NN and *k*NN-CF classifications are used to predict the class of data points that are randomly generated from the known data points with the minority class for distributions: 10% : 90% and 20% : 80%

	10%:90%		20%:80%	
	kNN	kNN-CF	kNN	kNN-CF
Breast-w	80.3	93.2	92.5	96.5
Haberman	0	25.4	13.1	36.2
Parkinsons	55	73.8	79	89.5
Transfusion	5.3	16.3	25.8	45.3
Magic	38.9	53.4	53.3	68.7
Ionosphere	3.4	23.3	36.8	57.4
Pima	1.5	22.6	30	51.8
Spambase	57.9	69.4	71.4	83.6
SPECTF	6.6	38.3	23.7	75.8
wdbc	86.3	92.4	93.2	93.8
Average	33.52	50.81	51.88	69.86

TABLE IV

Standard *k*NN and *k*NN-CF classifications are used to predict the class of data points that are randomly generated from the known data points with the minority class for distributions: 30% : 70% and 40% : 60%

	30%:70%		40%:60%	
	*k*NN	*k*NN-CF	*k*NN	*k*NN-CF
Breast-w	95	98.1	97.8	98.9
Haberman	22.9	51.3	38.2	65.3
Parkinsons	85.5	94.8	87.2	95.4
Transfusion	40.2	62.9	50.1	71.3
Magic	64.1	77	70.1	81.4
Ionosphere	56.9	69	64.9	71.5
Pima	52.5	71.5	60	77.9
Spambase	79.2	88.5	87.6	92.4
SPECTF	59.6	89.6	73.9	100
wdbc	90.8	93.2	93.9	96.4
Average	64.67	79.59	72.37	85.05

TABLE V

Standard *k*NN and *k*NN-CF classifications are used to predict the class of data points that are randomly generated from the known data points with the majority class for distributions: 10% : 90% and 20% : 80%

	10%:90%		20%:80%	
	*k*NN	*k*NN-CF	*k*NN	*k*NN-CF
Breast-w	98.9	97.9	98.4	98.2
Haberman	97.9	94.1	91.5	78.2
Parkinsons	100	98	98.1	91.6
Transfusion	97.9	90.6	93.9	82.5
Magic	98.4	96	97	92.1
Ionosphere	100	98.5	97.5	97.3
Pima	98.6	92.2	93.7	82.9
Spambase	98.6	96.9	97.4	94.2
SPECTF	96.7	84	86.1	64.9
wdbc	100	99.7	99.4	98.4
Average	98.7	94.79	95.3	88.03

TABLE VI

Standard *k*NN and *k*NN-CF classifications are used to predict the class of data points that are randomly generated from the known data points with the majority class for distributions: 30% : 70% and 40% : 60%

	30%:70%		40%:60%	
	*k*NN	*k*NN-CF	*k*NN	*k*NN-CF
Breast-w	97.6	96.3	97.8	97.6
Haberman	87.3	70	75.8	51.7
Parkinsons	96	88	89.2	80.6
Transfusion	86.3	72.1	80.8	62.1
Magic	93.6	86.3	92.3	82.2
Ionosphere	98.2	97.1	97.9	97.3
Pima	89.5	75.5	81.4	64.1
Spambase	94.5	89.8	92.8	84.4
SPECTF	68.9	46.5	62.8	46.1
wdbc	99	96.2	98.2	95
Average	91.09	81.78	86.9	76.11

C. The second group of experiments

We examine the efficiency with queries randomly generated from a dataset. The results are showed in Tables VII - XIV as follows.

TABLE VII

For dataset Breastw, the efficiency of standard *k*NN and *k*NN-CF classifications when 10% : 90%

Running times	*k*NN		*k*NN-CF	
	Precision	Recall	Precision	Recall
100	0.882353	0.9375	0.882353	0.9375
200	0.75	0.882353	0.772727	1
500	0.923077	0.765957	0.851064	0.851064
1000	0.869565	0.851064	0.847619	0.946809

TABLE VII

For dataset Breastw, the efficiency of standard *k*NN and *k*NN-CF classifications when 20% : 80%

Running times	*k*NN		*k*NN-CF	
	Precision	Recall	Precision	Recall
100	1	0.894737	0.947368	0.947368
200	0.914286	0.820513	0.916667	0.846154
500	0.954545	0.903226	0.936842	0.956989
1000	0.926471	0.931034	0.908257	0.975369

TABLE IX

. For dataset Breastw, the efficiency of standard *k*NN and *k*NN-CF classifications when 30% : 70%

Running times	*k*NN		*k*NN-CF	
	Precision	Recall	Precision	Recall
100	0.866667	0.962963	0.870968	1
200	0.927536	0.955224	0.90411	0.985075
500	0.95	0.956835	0.951049	0.978417
1000	0.973244	0.960396	0.936909	0.980198

TABLE X

For dataset Breastw, the efficiency of standard *k*NN and *k*NN-CF classifications when 40% : 60%

Running times	*k*NN		*k*NN-CF	
	Precision	Recall	Precision	Recall
100	0.945946	1	0.945946	1
200	0.952381	0.987654	0.931034	1
500	0.95977	0.954286	0.935135	0.988571
1000	0.944444	0.968912	0.918465	0.992228

TABLE XI

For dataset Ionosphere, the efficiency of standard *k*NN and *k*NN-CF classifications when 10% : 90%

Running times	*k*NN		*k*NN-CF	
	Precision	Recall	Precision	Recall
100	0.888889	0.615385	0.9	0.692308
200	1	0.111111	0.9	0.5
500	1	0.1	1	0.36
1000	1	0.172414	0.782609	0.413793

TABLE XII

For dataset Ionosphere, the efficiency of standard *k*NN and *k*NN-CF classifications when 20% : 80%

Running times	*k*NN		*k*NN-CF	
	Precision	Recall	Precision	Recall
100	1	0.8	1	0.866667
200	0.884615	0.469388	0.9	0.734694
500	0.933333	0.482759	0.944444	0.586207
1000	0.666667	0.134715	0.801802	0.46114

TABLE XIII

For dataset Ionosphere, the efficiency of standard *k*NN and *k*NN-CF classifications when 30% : 70%

Running times	*k*NN		*k*NN-CF	
	Precision	Recall	Precision	Recall
100	0.931034	0.72973	0.916667	0.891892
200	0.939394	0.563636	0.944444	0.618182
500	0.949367	0.517241	0.948454	0.634483
1000	0.920455	0.514286	0.932039	0.609524

TABLE XIV

For dataset Ionosphere, the efficiency of standard *k*NN and *k*NN-CF classifications when 40% : 60%

Running times	*k*NN		*k*NN-CF	
	Precision	Recall	Precision	Recall
100	1	0.657895	1	0.684211
200	0.959184	0.580247	0.967742	0.740741
500	0.971429	0.676617	0.943396	0.746269
1000	0.95082	0.659091	0.939481	0.740909

From Tables III-XIV, the *k*NN-CF is much better than standard *k*NN classification at predicting the minority class. This indicates that the CF strategy is promising to reduce the misclassification cost for real applications, such as disease diagnosis and risk-sensitive learning.

V. Conclusions and Open Problems

In this paper we have incorporated the certainty factor to *k*NN classification that clearly distinguishes whether the belief of the class of a query is increased, given its k nearest neighbors. We have experimentally illustrated the efficiency of the proposed approach, *k*NN-CF classification. For future study, we list some open problems in *k*NN-CF classification as follows.

1. Improve the discernment of *k*NN-CF classification with a means, such as the SN approach in [64,65].
2. Extending the *k*NN-CF classification to cost/risk-sensitive learning.
3. *k*NN-CF classification with missing values.
4. *k*NN-CF classification with cold-deck instances [38].
5. The evaluation of *k*NN-CF classification.

ACKNOWLEDGMENT

I am grateful for both the suggestions and the experiments carried out by my ex-student, Mr Manlong Zhu. Many thanks for the comments on the early version of this paper from Mr Xiaofeng Zhu (my ex-student), 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, and the Guangxi NSF (Key) grants.

REFERENCES

[1] D.W. Aha, D.F. Kibler and M.K. Albert (1991). Instance-based learning algorithms. *Machine Learning*, Vol. 6: 37–66.

[2] V. Athitsos, J. Alon, S. Sclaroff and G. Kollios (2008). BoostMap: An Embedding Method for Efficient Nearest Neighbor Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell*. 30(1): 89-104.

[3] S. Belongie, J. Malik, and J. Puzicha (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell*., 24(4): 509-522.

[4] E. Blanzieri and F. Melgan (2008). Nearest Neighbor Classification of Remote Sensing Images With the Maximal Margin Principle. *IEEE Trans.Geoscience and Remote Sensing*, 46(6): 1804-1811.

[5] E. Blanzieri and F. Ricci (1999). Probability Based Metrics for Nearest Neighbor Classification and Case-Based Reasoning. *Lecture Notes in Computer Science*, Vol. 1650: 14-29.

[6] D.R. Carvalho and A.A. Freitas (2002). A genetic-algorithm for discovering small-disjunct rules in data mining. *Appl. Soft Comput*., Vol. 2(2): 75–88.

[7] D. Carvalho and A. Freitas (2004). A hybrid decision tree/genetic algorithm method for data mining. *Inf. Sci*., Vol. 163(1-3): 13–35.

[8] J. Chai, H. Liu, B. Chen and Z. Bao (2010). Large margin nearest local mean classifier. *Signal Processing*, 90: 236-248.

[9] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16: 321–357.

[10] Chen, J., and Shao, J., (2001). Jackknife variance estimation for nearest-neighbor imputation. *J. Amer. Statist. Assoc*. 2001, Vol. 96: 260-269.

[11] Cheung, K., Fu, A. (1998). Enhanced Nearest Neighbour Search on the R-tree. *SIGMOD Record*, Vol 27: 16-21.

[12] T. Cover and P. Hart (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, Vol. 13: 21–27.

[13] Daugman, J. (2003). Demodulation by Complex-Valued Wavelets for Stochastic Pattern Recognition. Int'l J. Wavelets, *Multiresolution and Information Processing*, pp 1-17.

[14] Dietterich, T., and Sutton, R. (1995). An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. *Machine Learning*, 19: 5-28.

[15] C. Domeniconi and D. Gunopulos (2001). Adaptive nearest neighbor classification using support vector machines. In *NIPS*, pp 665-672.

[16] C. Domeniconi, J. Peng and D. Gunopulos (2002). Locally adaptive metric nearest-neighbor classification. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(9): 1281-1285.

[17] P. Domingos (1998). How to get a free lunch: A simple cost model for machine learning applications. *Proc. AAAI98/ICML98, Workshop on the Methodology of Applying Machine Learning. AAAI Press*, pp. 1–7.

[18] P. Domingos (1999). Metacost: A general method for making classifiers cost-sensitive. *In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pp. 155–164.

[19] C. Elkan (2001). The foundations of cost-sensitive learning. *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978.

[20] T. Fawcett and F. J. Provost (1997). Adaptive fraud detection. *Data Min. Knowl. Discov*., 1(3): 291–316.

[21] Fix E, Hodges JL, Jr (1951). Discriminatory analysis, nonparametric discrimination. USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, February 1951.

[22] B. J. Frey and D. Dueck (2007). Clustering by Passing Messages Between Data Points. Science, 315: 972-976.

[23] Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). KNN Model-Based Approach in Classification. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pp 986-996.

[24] T. Hastie and R. Tibshirani (1996). Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell*., 18(6): 607-615.

[25] R.C. Holte, L. Acker and B.W. Porter (1989). Concept learning and the problem of small disjuncts. *IJCAI-89*, pp. 813–818.

[26] T.M. Huard and S. Robin (2009). Tailored Aggregation for Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11): 2098-2105.

[27] P. Jing, D.R. Heisterkamp and H.K. Dai (2001). LDA/SVM driven nearest neighbor classification. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog*., pp. 58-63.

[28] M. Kubat, R. Holte, and S. Matwin (1998). Machine learning for the detection of oil spills in satellite radar images. *MachineLearning*, 2-3, pp. 195–215.

[29] M. Kubat and S. Matwin (1997). Addressing the curse of imbalanced training sets: One-sided selection. *In Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179–186.

[30] W. Lam and Y. Han (2003). Automatic Textual Document Categorization Based on Generalized Instance Sets and a Metamodel. *IEEE Trans. Pattern Anal. Mach. Intell*., 25(5): 628-633.

[31] B. Li, Y.W. Chen and Y. Chen (2008). The Nearest Neighbor Algorithm of Local Probability Centers. *IEEE Trans. Syst., Man, Cybern. (B)*, 38(1): 141-153.

[32] C. Ling and C. Li (1998). Data mining for direct marketing: Problems and solutions. *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 73–79.

[33] T. Menzies, J. Greenwald and A. Frank (2007). Data mining static code attributes to learn defect predictors. *IEEE Transactionson Software Engineering*, 33: 2–13.

[34] Mitani, Y. and Hamamoto, Y. (2006). A local mean-based nonparametric classifier. *Pattern Recognition Letters*, 27(10): 1151-1159.

[35] K. Ni and T. Nguyen (2009). An Adaptable k-Nearest Neighbors Algorithm for MMSE Image Interpolation. *IEEE Transactions on Image Processing*, 18(9): 1976-1987.

[36] V. Pascal and B. Yoshua (2003). Manifold Parzen windows. *Proc. NIPS*, pp. 825-832.

[37] J. Peng, D. Heisterkamp and H.K. Dai (2004). Adaptive Quasiconformal Kernel Nearest Neighbor Classification. *IEEE Trans. Pattern Analysis and Machine lntelligence*, 26(5): 656-661.

[38] YS Qin, SC Zhang (2008). Empirical Likelihood Confidence Intervals for Differences between Two Datasets with Missing Data. *Pattern Recognition Letters*, Vol 29(6): 803-812.

[39] J.R. Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.

[40] J.R. Quinlan, P.J. Compton, K.A. Horn and L. Lazarus (1986). Inductive knowledge acquisition: a case study. *Proceedings of the second Australian Conference on the Applications of Expert Systems*, pp. 183–204.

[41] S. Salzberg (1991). A Nearest Hyperrectangle Learning Method. *Machine Learning*, Vol. 6: 251-276.

[42] H. Samet (2008). K-Nearest Neighbor Finding Using MaxNearest-Dist. *IEEE Trans. Pattern Anal. Mach. Intell*. 30(2): 243-252.

[43] Sanchez, J.S., Pla, F. and Ferri, F. J. (1997). On the use of neighbourhood based non-parametric classifiers. *Pattern Recognition Letters*, 18: 1179-1186.

[44] P. Simard, Y. LeCun and J.S. Denker (1993). Efficient pattern recognition using a new transformation distance. *Proceedings of NIPS*, pp 50-58.

[45] Singh, S., Haddon, J., Markou, M. (1999). Nearest Neighbour Strategies for Image Understanding. *Proc. Workshop on Advanced Concepts for Intelligent Vision Systems (ACIVS'99)*, Baden-Baden, pp 2-7.

[46] Shortliffe, E. and Buchanan, B. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23: 351–379.

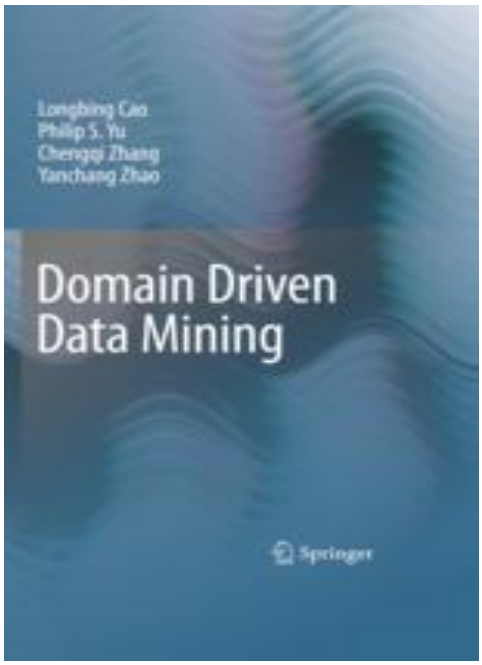
[47] C. Stanfill and D. Waltz (1986). Toward Memory-Based Reasoning. Comm. ACM, Vol. 29: 1213-1229.

[48] Y. Sun, M.S. Kamel, A. Wong and Y. Wang (2007). Costsensitive boosting for classification of imbalanced data. *Pattern Recogn.*, Vol. 40(12): 3358–3378.

[49] Tao, Y., Papadias, D., Lian, X. (2004). Reverse knn search in arbitrary dimensionality. *VLDB-2004*, pp 744-755.
[50] J.B. Tenenbaum and V. de Silva and J.C. Langford (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290: 2319-2323.
[51] K. Ting (1994). The problem of small disjuncts: its remedy in decision trees. *In Proceedings of the Tenth Canadian Conference on Artificial Intelligence*, pp. 91–97.
[52] M. Varma and A. Zisserman (2005). A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1-2): 61-.81.
[53] P. Vincent and Y. Bengio (2001). K-local hyperplane and convex distance nearest neighbor algorithms. *In NIPS*, pp 985-992.
[54] H. Wang (2006). Nearest neighbors by neighborhood counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28: 942-953.
[55] G.M. Weiss (2004). Mining with rarity: a unifying framework. *SIGKDD Explorations*, 6(1): 7–19.
[56] G.M. Weiss and H. Hirsh (2000). A quantitative study of small disjuncts. *AAAI/IAAI*, pp. 665–670.
[57] G. Wen, L. Jiang and J. Wen (2008). Using Locally Estimated Geodesic Distance to Optimize Neighborhood Graph for Isometric Data Embedding. *Pattern Recognition*, 41: 22-26.
[58] G. Wen, et al. (2009). Local relative transformation with application to isometric embedding. *Pattern Recognition Letters*, 30(3): 203-211.
[59] D.R. Wilson and T.R. Martinez (2000). Reduction techniques for instance-based learning algorithms. *Machine Learning*, 383, pp. 257–286, Kluwer Academic Publishers.
[60] Wu, XD., et al. (2008). Top 10 Algorithms in Data Mining, Knowledge and Information Systems, 14(1): 1-37.
[61] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, Vol 1: 67-88.
[62] Y. Zeng, Y. Yang and L. Zhao (2009). Nonparametric classification based on local mean and class statistics. *Expert Systems with Applications*, 36: 8443-8448.
[63] H. Zhang, A. Berg, M. Maire and J. Malik (2006). SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. *Proceeding of CVPR-2006*, pp 2126-2136.
[64] Zhang, SC. (2008). Parimputation: From imputation and null-imputation to partially imputation. *IEEE Intelligent Informatics Bulletin*, Vol 9(1), 2008: 32-38.
[65] Zhang, SC. (2010). Shell-Neighbor Method And Its Application in Missing Data Imputation. *Applied Intelligence, DOI*: 10.1007/s10489-009-0207-6.
[66] 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.
[67] Zhang, S.C., Zhang, C.Q. and Yang, Q. (2004). Information Enhancement for Data Mining. *IEEE Intelligent Systems*, March/April 2004: 12-13.
[68] H. Zhu and O. Basir (2005). An Adaptive Fuzzy Evidential Nearest Neighbor Formulation for Classifying Remote Sensing Images. *IEEE Trans. Geoscience and Remote Sensing*, 43(8): 1874-1889.
[69] J. Zhu (2007). Active learning for word sense disambiguation with methods for addressing the class imbalance problem. *In Proceedings of ACL*, pp. 783–790.
[70] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. *Proceedings of Computer Vision and Pattern Recognition(2005)* , volume 1, pages 26–33, 2005.
[71] X.Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules.*ACM Transactions on Information Systems*, 22(3):381 – 405, July 2004.
[72] S. Zhang and X. Wu. Fundamental Role of Association Rules in Data Mining and Knowledge Discovery. *WIREs Data Mining & Knowledge Discovery*, 2011
[73] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, Building useful models from imbalanced data with sampling and boosting. *In Proceedings of 21st Int. FLAIRS Conference*, May 2008, pp. 306–311.

Domain Driven Data Mining

BY LONGBING CAO, PHILIP S. YU, CHENGQI ZHANG AND YANCHANG ZHAO.



REVIEWED BY
NORLAILA HUSSAIN
HELEN ZHOU

Data mining is a powerful paradigm of extracting information from data. It can help enterprises focus on important information in their data warehouse. Data mining is also known as *Knowledge Discovery in Databases* (KDD). It involves the extraction of hidden pattern to predict future trends and behaviors which allow businesses to make proactive, knowledge-driven decisions.

The current vast development in ubiquitous computing, cloud computing and networking across every sector and business has made data mining emerging as one of the most active areas in information and communication technologies (ICT) as data and its deep analysis becomes an important issue for enhancing the soft power of an organization, its production systems, decision making and performance.

However, there is a large gap has been identified by many studies between academic deliverables and business expectations, as well as between data miners and business analysts. The limited decision-support power of data mining in the real world has prevented it from playing a strategic decision-support role in ICT. The main concerns include the actionability, workability, transferability, and the trustworthy, dependable, repeatable, operable and explainable capabilities of data mining algorithms, tools and outputs.

Nevertheless, these challenges create opportunities for promoting a paradigm shift from data-centered hidden pattern mining to domain-driven actionable knowledge delivery. These real-world concerns and complexities of the KDD methodologies and techniques have motivated Cao et al. (2010) to propose domain driven data mining (D^3M) as effective and practical methodologies for actionable knowledge discovery in order to narrow down and bridge the gap between the academia and the business people. This proposal is elaborated in great length in their latest book "Domain Driven Data Mining".

Domain driven data mining involves the study of effective and efficient methodologies, techniques, tools, and applications which can discover and deliver actionable knowledge that can be passed on to business people for direct decision-making and action-taking.

The book begins by highlighting the gap that exists between academia and business in Chapter 1. This gap includes the large numbers of algorithms published by academia versus only a few are deployed in a business setting. In addition, despite the large number of patterns mined or identified, only a few satisfy business needs and lack of recommended decision-support actions. They stressed that the algorithms models and resulting patterns and knowledge are short of workable, actionable and operable capabilities. The authors went on to summarize the main challenges and technical issues surrounding the traditional data mining and knowledge discovery methodologies.

To address the issues highlighted, the authors introduce the main components and methodological framework of D^3M methodologies in Chapter 2. Based on authors' real world experiences and lessons learned in a capital market, significance results were discovered when domain factors are considered in data mining. An overall picture of D^3M focusing on the concept map, the key methodological components, the theoretical underpinning and the process model were outlined.

The discussions on domain-driven data mining methodologies are further elaborated in Chapter 3 to 5. The authors elaborated the importance of involving and consolidating relevant ubiquitous intelligence (i.e. data intelligence, human intelligence, domain intelligence, network and web intelligence, and organizational and social intelligence) surrounding data mining applications for actionable knowledge discovery and delivery. The definitions, aims, aspects and techniques for involving this ubiquitous intelligence into data mining are identified in Chapter 3.

A key concept in D^3M that is highlighted is *actionable knowledge discovery* (AKD). It involves and synthesizes domain intelligence, human intelligence and cooperation, network intelligence and in-depth data intelligence to define, measure, and

evaluate business interestingness and knowledge actionability. The authors stressed the importance of AKD as an important concept for bridging the gap between technical-based approaches and business impact-oriented expectations on patterns discovered from data mining. This concept is elaborated in Chapter 4.

Four types of system frameworks for actionable knowledge delivery are then introduced in Chapter 5. The frameworks include *PA-AKD* (a two-step AKD process), *UI-AKD* (based on unified interestingness), *CM-AKD* (a multi-step AKD process), and *MSCM-AKD* (based on multiple data sources). The authors describe the flexibility of the proposed frameworks which can cover many common problems and applications and are effective in extracting knowledge that can be used by business people for immediate decision-making.

Chapters 6 to 8 outline several techniques supporting domain-driven data mining. Chapter 6 presents a comprehensive and general approach named *combined mining* for handling multiple large heterogeneous data sources targeting more informative and actionable knowledge. The authors describe this approach as a framework for mining complex knowledge in complex data where many mutative applications can be designed such as combined pattern mining in multiple data sources. They focus on providing general frameworks and approaches to handle multi-feature, multi-source and multi-method issues and requirements.

In Chapter 7, the authors introduce agent-driven data mining for D^3M. The basic concept, driving forces, technical means, research issues and case studies of agent-driven data mining are discussed. The authors suggest the interaction and integration between agents and data mining are necessary as agent technology can greatly complement data mining in complex data mining problems in situations such as data processing, information processing, user modeling and interaction, infrastructure and services.

Chapter 8 elaborates the technique of post analysis and post mining. This technique helps to refine discovered patterns and learned models and present useful and applicable knowledge to users. It uses visualization techniques which present the knowledge desired by the end users and which is easy to read and understand. The authors discuss interesting measures, pruning, selection, summarization, visualisation and maintenance of patterns.

To assist readers in understanding D^3M further, the authors continue to illustrate the use of domain driven data mining in the real world. In Chapter 9, the authors describe how domain-driven data mining is applied to identify actionable trading strategies and actionable market microstructure behavior patterns in capital markets. They elaborate some case studies in which this methodology has been used for smart trading, and mining for deeply understanding of exceptional trading behaviour on capital market data.

Chapter 10 utilizes domain-driven data mining in identifying actionable combined associations and combined patterns in social security data. It illustrates the use of domain driven data for better understanding government service quality, causes and effects of government service problems, customer behaviour and demographics, and government officer-customer interactions. The case study introduces several examples using the *MSCM-AKD* framework in identifying combined associations and combined associations clusters for debt prevention.

The final chapter summarizes some of the open issues and discusses trends in domain-driven data mining research and development. The authors highlighted several fundamental problems that need further investigation such as supporting social interaction and cognition in data mining and making data mining trustful and business-friendly. They also suggest the need for next-generation data mining and knowledge discovery that is far beyond the data mining algorithms as there are many open issues and opportunities arise when problem-solving is viewed from the domain-driven perspective.

Overall, the book is well-written and reading it has been an enjoyable one. The authors present interesting issues and opportunities for further exploration of data mining in the future. The main focus of the book is to demonstrate some new techniques to amplify the decision-support power of data mining and they have certainly succeeded.

THE BOOK:

CAO, PHILIP S. YU, CHENGQI ZHANG AND YANCHANG ZHAO (2010) DOMAIN DRIVEN DATA MINING, 1ST EDITION. 2010, XIII, 237 P. SPRINGER.
ISBN: 978-1-4419-5736-8

ABOUT THE REVIEWERS:

NORLAILA HUSSAIN
School of Engineering and Advanced Technology, Massey University, New Zealand. Contact her at: N.Hussain@massey.ac.nz

HELEN ZHOU
School of Electrical Engineering, Manukau Institute of Technology, Auckland, New Zealand. Contact her at: helen.zhou@manukau.ac.nz

Related Conferences, Call For Papers/Participants

TCII Sponsored Conferences

.

WI 2011
The 2011 IEEE/WIC/ACM International Conference on Web Intelligence
Lyon, France
August 22- 27, 2011
http://wi-iat-2011.org/

Web Intelligence (WI) explores the fundamental roles, interactions as well as practical impacts of Artificial Intelligence engineering and Advanced Information Technology on the next generation of Web systems. 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, discovery and data mining are examples. Advanced Information Technology includes wireless networks, ubiquitous devices, social networks, and data/knowledge grids, as well as cloud computing, service oriented architecture. The 2011 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2011) will take place at the Campus Universitaire de la Doua, Lyon, France. WI 2011 will be co-located with the 2011 IEEE/ACM/WIC International Conference on Intelligent Agent Technology (IAT2011). WI 2011 will include a summer school providing in-depth background on subjects that are of broad interest to Web intelligence and Intelligent Agent Technology communities.

IAT 2011
The 2011 IEEE/WIC/ACM International Conference on Intelligent Agent Technology
Lyon, France
August 22- 27, 2011
http://wi-iat-2011.org/

The 2011 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2011) will be co-located with the 2011 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2011). The IEEE/WIC/ACM 2011 joint conferences will take place at the Campus Universitaire de la Doua, Lyon, France.

IAT 2011 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 2011 will foster the development of novel paradigms and advanced solutions in agent based computing. The joint organization of IAT 2011 and WI 2011 will provide an opportunity for technical collaboration beyond the two distinct research communities.

ICDM 2010
The Tenth IEEE International Conference on Data Mining
Sydney, Australia
December 13-17, 2010
http://datamining.it.uts.edu.au/icdm10/

The IEEE International Conference on Data Mining series (ICDM) has established itself as the world's premier research conference in data mining. It provides an international forum for presentation of original research results, as well as exchange and dissemination of innovative, practical development experiences. The conference covers all aspects of data mining, including algorithms, software and systems, and applications. In addition, ICDM draws researchers and application developers from a wide range of data mining related areas such as statistics, machine learning, pattern recognition, databases and data warehousing, data visualization, knowledge-based systems, and high performance computing. By promoting novel, high quality research findings, and innovative solutions to challenging data mining problems, the conference seeks to continuously advance the state-of-the-art in data mining. Besides the technical program, the conference features workshops, tutorials, panels and, since 2007, the ICDM data mining contest.

Topics related to the design, analysis and implementation of data mining theory, systems and applications are of interest. These include, but are not limited to the following areas: data mining foundations, mining in emerging domains, methodological aspects and the KDD process, and integrated KDD applications, systems, and experiences. A detailed listing of specific topics can be found at the conference website.

BIBM 2010
IEEE International Conference on Bioinformatics & Biomedicine
Hong Kong
December 18-21, 2010
http://www.math.hkbu.edu.hk/BIBM2010/

IEEE BIBM 2010 will provide a general forum for disseminating the latest research in bioinformatics and biomedicine. It is a multidisciplinary conference that brings together academic and industrial scientists from computer science, biology, chemistry, medicine, mathematics and statistics.

BIBM will exchange research results and address open issues in all aspects of bioinformatics and biomedicine and provide a forum for the presentation of work in databases, algorithms, interfaces, visualization, modeling, simulation, ontology and other computational methods, as applied to life science problems, with emphasis on applications in high throughput data-rich areas in biology, biomedical engineering.

IEEE BIBM 2010 intends to attract a balanced combination of computer scientists, biologists, biomedical engineers, chemist, data analyzer, statistician.

ICTAI 2011
The Twenty-Third IEEE International Conference on Tools with Artificial Intelligence
Boca Raton，USA
October 1-3, 2011

The annual IEEE International Conference on Tools with Artificial Intelligence (ICTAI) provides a major international forum where the creation and exchange of ideas related to artificial intelligence are fostered among academia, industry, and government agencies. The conference facilitates the cross-fertilization of these ideas and promotes their transfer into practical tools, for developing intelligent systems and pursuing artificial intelligence applications. The ICTAI encompasses all technical aspects of specifying, developing and evaluating the theoretical underpinnings and applied mechanisms of the AI based components of computer tools (i.e. algorithms, architectures and languages).

Related Conferences

AAMAS 2011
The Tenth International Conference on Autonomous Agents and Multi-Agent Systems
Taipei, Taiwan
May 2- 6, 2011
http://www.aamas2011.tw/

The AAMAS conference series was initiated in 2002, with the merger of three highly respected individual conferences, and has now reached its tenth anniversary. The aim of the joint conference is to provide a single, high-profile, internationally renowned forum for research in the theory and practice of autonomous agents and multiagent systems. The conference is sponsored by the International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS). AAMAS 2011 will take place at Taipei International Convention Center (TICC) in Taipei, Taiwan on May 2-6 2011. AAMAS is the premier scientific conference for research on autonomous agents and multiagent systems.

AAMAS is the leading scientific conference for research in autonomous agents and multiagent systems. The AAMAS conference series was initiated in 2002 by merging three highly-respected meetings: International Conference on Multi-Agent Systems (ICMAS); International Workshop on Agent Theories, Architectures, and Languages (ATAL); and International Conference on Autonomous Agents (AA). The aim of the joint conference is to provide a single, high-profile, internationally-respected archival forum for scientific research in the theory and practice of autonomous agents and multiagent systems. AAMAS 2011 is the Tenth conference in the AAMAS series, following enormously successful previous conferences, and will be held at Taipei International Convention Center (TICC), Taipei, Taiwan. See the IFAAMAS web site for more information on the AAMAS conference series.

SDM 2011
2011 SIAM International Conference on Data Mining
Hilton Phoenix East/Mesa,
Mesa, Arizona, USA
April 23- 30, 2011
http://www.siam.org/meetings/sdm11/

Data mining is an important tool in science, engineering, industrial processes, healthcare, business, and medicine. 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 theoretical and 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.

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.

AAAI 2011
The Twenty-Fifth AAAI Conference on Artificial Intelligence
San Francisco, California
August 7-11, 2011
http://www.aaai.org/Conferences/AAAI/aaai11

The Twenty-Fifth Conference on Artificial Intelligence (AAAI 2011) will be held in San Francisco, California at the Hyatt Regency San Francisco, from August 7-11, 2011. The purpose of the AAAI 2011 conference is to promote research in AI and scientific exchange among AI researchers, practitioners, scientists, and engineers in related disciplines. Details about the AAAI 2011 program will be published on http://www.aaai.org/Conferences/AAAI/aaai11 as they become available.

IJCAI 2011
The Twenty-Second International Joint Conference on Artificial Intelligence
Barcelona, Catalonia, Spain
July 16-22, 2011
http://ijcai-11.iiia.csic.es

The Twenty-Second International Joint Conference on Artificial Intelligence IJCAI 2011 will be held in Barcelona, Spain, July 16-22, 2011. Submissions are invited on significant, original, and previously unpublished research on all aspects of artificial intelligence. The theme of IJCAI 2011 is "Integrated and Embedded Artificial Intelligence" (IEAI) with a focus on artificial intelligence that crosses discipline boundaries within AI, and between AI and other disciplines. Building systems often requires techniques from more than one area (e.g. both machine learning and natural language processing, or both planning and preference representation). In addition, larger systems often have AI components embedded within that provide intelligent functionalities such as learning and reasoning. The conference will include a special track dedicated to such work.

AMT 2011
The 2011 International Conference on Active Media Technology
Lanzhou, China
September 7-9, 2011
http://wi-consortium.org/conferences/amtbi11/

In the great digital era, we are witnessing many rapid scientific and technological developments in human-centred, seamless computing environments, interfaces, devices, and systems with applications ranging from business and communication to entertainment and learning. These developments are collectively best characterized as Active Media Technology

(AMT), a new area of intelligent information technology and computer science that emphasizes the proactive, seamless roles of interfaces and systems as well as new media in all aspects of digital life. An AMT based system offers services to enable the rapid design, implementation and support of customized solutions.

The first International Conference on Active Media Technology (AMT 2001) was held in Hong Kong in 2001, the second International Conference on Active Media Technology (AMT 2004) was held in Chongqing, China in May 29-31 of 2004, the third International Conference on Active Media Technology (AMT 2005) was held in Kagawa, Japan in May 2005, the fourth International Conference on Active Media Technology (AMT 2006) was held in Brisbane, Australia in June 7-9, 2006, the fifth International Conference on Active Media Technology (AMT 2009) was held in Beijing, China in October 22-24, 2009, and the sixth International Conference on Active Media Technology (AMT 2010) was held in Toronto, Canada in August 28-30, 2010. Following the success of AMT 2001, AMT 2004, AMT 2005, AMT 2006, AMT 2009, and AMT 2010 the seventh International Conference on Active Media Technology (AMT 2011) will be held in Lanzhou, China in September 7-9, 2011.

Active Media Technology 2011 will be jointly held with the 2011 International Conference on Brain Informatics (BI 2011). The WIC has decided to organize AMT'11 in memoriam of Herbert A. Simon. The two conferences will have a joint opening, keynote, reception, and banquet. Attendees only need to register for one conference and can attend workshops, sessions, exhibits and demonstrations across the two conferences.

BI 2011
The 2011 International Conference on Brain Informatics
Lanzhou, China
September 7-9, 2011
http://wi-consortium.org/conferences/amtbi11/

Brain Informatics (BI) has recently emerged as an interdisciplinary research field that focuses on studying the mechanisms underlying the human information processing system (HIPS). It investigates the essential functions of the brain, ranging from perception to thinking, and encompassing such areas as multi-perception, attention, memory, language, computation, heuristic search, reasoning, planning, decision-making, problem-solving, learning, discovery, and creativity. The goal of BI is to develop and demonstrate a systematic approach to achieving an integrated understanding of both macroscopic and microscopic level working principles of the brain, by means of experimental, computational, and cognitive neuroscience studies, as well as utilizing advanced Web Intelligence (WI) centric information technologies. BI represents a potentially revolutionary shift in the way that research is undertaken. It attempts to capture new forms of collaborative and interdisciplinary work. In this vision, new kinds of BI methods and global research communities will emerge, through infrastructure on the wisdom Web and knowledge grids that enables high speed and distributed, large-scale analysis and computations, and radically new ways of sharing data/knowledge.

Brain Informatics 2011 provides a leading international forum to bring together researchers and practitioners from diverse fields, such as computer science, information technology, artificial intelligence, Web intelligence, cognitive science, neuroscience, medical science, life science, economics, data mining, data and knowledge engineering, intelligent agent technology, human computer interaction, complex systems, and system science, to explore the main research problems in BI lie in the interplay between the studies of human brain and the research of informatics. On the one hand, one models and characterizes the functions of the human brain based on the notions of information processing systems. WI centric information technologies are applied to support brain science studies. For instance, the wisdom Web and knowledge grids enable high-speed, large-scale analysis, simulation, and computation as well as new ways of sharing research data and scientific discoveries. On the other hand, informatics-enabled brain studies, e.g., based on fMRI, EEG, MEG significantly broaden the spectrum of theories and models of brain sciences and offer new insights into the development of human-level intelligence on the wisdom Web and knowledge grids.

Brain Informatics 2011 will be jointly held with the 2011 International Conference on Active Media Technology (AMT 2011). The WIC has decided to organize BI'11 in memoriam of Herbert A. Simon. The two conferences will have a joint opening, keynote, reception, and banquet. Attendees only need to register for one conference and can attend workshops, sessions, exhibits and demonstrations across the two conferences.

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

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