THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

December 2014 Vol. 15 No. 1 (ISSN 1727-5997)

Feature Articles

Special Issue

Selected Extended Abstracts: IEEE International Conference on Health Informatics 2014 Doctoral Consortium

Announcements

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

The IEEE Intelligent Informatics Bulletin

Aims and Scope

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

1) Letters and Communications of the TCII Executive Committee
2) Feature Articles
3) R&D Profiles (R&D organizations, interview profile on individuals, and projects etc.)
4) Book Reviews
5) News, Reports, and Announcements (TCII sponsored or important/related activities)

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

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

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

Patient Centered Healthcare Informatics

Christopher C. Yang

***Abstract*—The healthcare system is undergoing a transformation from reactive care to proactive and preventive care. Patients or health consumers are actively acquiring knowledge to manage their health and seeking supports from their peers in addition to receiving healthcare support from healthcare professionals. 74% of American adults use the internet, of which 80% have looked online for healthcare information [6]. With the popularity of social media, many health consumers are also exchanging informational and emotional support with peers who have similar health conditions or diseases. The large volume of health consumer contributed content provides valuable resources for healthcare informatics research. It is worth to note that the information in the health consumer contributed content is timelier than the traditional resources such as electronic health records, centralized reporting systems, and pharmaceutical databases because health consumers often discuss their concerns with peers before any of them are reported in the traditional resources [22]. In this article, we review a few important healthcare informatics research issues that are centered on the patient contributed content and concerns.**

***Index Terms* — Social media analytics, healthcare informatics, consumer health vocabulary, social support, drug safety signal detection, topic detection, recommendation systems.**

I. Introduction

HEALTH and wellbeing plays an important role in our societies. Improving the health and wellbeing of people is a main goal accomplished through both government and private healthcare organizations [21]. In the recent years, we have also observed the increasing effort of health consumers or patients in managing their own health proactively and preventively. Health consumers are actively educating themselves about health and wellness in order to maintain a healthy body or prevent diseases. Patients are going to Internet to acquire knowledge about their health conditions or treatments by identifying authoritative information from popular health web sites such as WebMD and PubMed. When the resources are limited, health consumers and patients are also going to social media sites such as MedHelp and PatientsLikeMe to seek and offer supports with their peers who have similar health conditions or diseases [23]. Many patients are sharing experiences with their peers and offering advice and opinions to support one another. In this article, we are focusing on five specific issues: (a) consumer health expressions, (b) social support, (c) community topic detection and recommendation systems, (d) drug safety signal detection, and (e) symptom profiling and clustering.

Christopher C. Yang is with in College of Computing and Informatics at Drexel University, Philadelphia, PA 19104, USA, Email: chris.yang@drexel.edu

II. Consumer Health Expressions

Despite the fact that patients and health consumers are actively seeking and exchanging healthcare information on the Internet, identifying the relevant and useful information is very challenging to most patients and health consumers. It is because health consumers and health professionals often use different vocabularies to express health related topics [10,29,30,31]. While health professionals are trained to use professional language, which can be easily identified from the healthcare professional ontologies such as UMLS and MeSH, to describe the health issues, health consumers use a variation of vocabularies to express their health concerns depending on their cultural, educational, social, and economic backgrounds. The language gap creates a huge barrier between the communications of health consumers and health professionals as well as between the communications of health consumers with substantially different backgrounds. For example, a patient experiencing nose bleeding may not be able to find relevant authoritative information when the scientific publications are using the professional term “epistaxis” to describe the symptom. Similarly, some patients may express the symptom as nose bleeding while some others may express it as bloody nose. The variation of expressions adopted by health consumers makes it difficult to communicate and search information online simply by keyword matching.

Many researchers have devoted to develop Consumer Health Vocabularies (CHVs) to capture the expressions used by health consumers and map these vocabularies to healthcare professional ontologies. Zeng et al. has developed the first generation of CHV [31]. However, a substantial amount of manual effort is required in their effort. In addition, the consumer health expressions are evolving from time to time. As a result, CHV needs to be maintained continuously in order to capture new expressions that have not been included in CHVs yet.

In the recent years, more efforts have been made to utilize the health consumer contributed content in social media to identify the new consumer health expression semi-automatically or automatically. Jiang and Yang have utilized co-occurrence analysis to extract consumer health expressions by expanding to the original CHV [7,8,9]. Co-occurrence analysis is used because two or more words that tend to occur in similar linguistic context tend to resemble each other in meaning. In the co-occurrence analysis, we find that most health consumer

expressions are bi-grams. By using the expanded consumer health expressions, we can extract up to ten times more relevant online discussion threads on a particular health issue such as the adverse drug reaction heart disease. By mapping the expanded consumer health expressions to UMLS, health consumers can identify authoritative information more effectively or health professionals or researchers can extract the patient concerns from health consumer contributed content in a timely manner. The extracted content through using expanded consumer health expressions are also useful for many knowledge discovery applications such as drug safety signal detection, symptom analysis, topic detection and recommendation systems, which will be discussed in the later sections.

III. Social Support

Social support has been proven to be important in healthcare intervention. Social media provides a platform for health consumers to make connection with others without time and geographical constraints. Hence the limitations of the traditional social support groups that meet regularly at dedicated locations can be removed to support a broader community. In our previous study, it is found that a substantial amount of informational support and nurturant support are found in healthcare social media such as MedHelp and QuitNet [1,2,3,4].

Informational support offered by health consumers provides information related to treatment or coping with diseases. The information includes advice to cope with situations, referrals to other resources, facts that reassessing situations, opinions on issues but not necessarily based on facts, and personal experiences. In a previous study, we found that most informational supports are found in online discussion forum setting [1]. Among all different types of informational support, personal experiences are the most popular, followed by advice and opinions [1]. Referrals and facts are relatively less popular.

Nurturant supports are expressions that show signs of listening, expressing sympathy or the importance of relationship. There are three major types nurturant supports, including (a) esteem support that gives positive comments to validate the recipient's self-concept and alleviate feelings, (b) emotional support that gives expressions to support the recipient's feeling or reciprocates emotion, and (c) networking support that focuses on connecting recipients to others with similar situation to broaden social networks. More nurturant supports are found in private settings such as the journal section of the personal profile pages, where health consumers often discuss their own health status [2]. Among all types of nurturant support, emotional support is the most popular followed by networking support and esteem support [2].

The interaction patterns in informational support and nurturant support are different [33,35]. Health consumers with health status in the later stages tend to offer informational support to other health consumers with health status in the earlier stage. For example, in the QuitNet forum, health consumers who have quitted smoking for a longer time tend to offer informational support to health consumers who have just started to quit or have quitted for a relatively shorter time. However, health consumers tend to offer nurturant support to other health consumers who have similar health status.

IV. Community Topic Detection and Recommendation Systems

The discussions in healthcare social media sites are valuable resources to discover the timely patient concerns. We have applied dynamic stochastic blockmodeling and temporal Dirichlet process to detect hidden communities [12]. Such detection model is able to detect the evolving communities since the discussion groups may expand, shrink, split, or merge as the discussions are going on. By monitoring the emerging topics and evolving communities, it is helpful to capture the issues raised in social media [15]; and hence, makes helpful recommendations [13], provides timely support to health consumers and identifies new research problems. ACTONNECT is a web-based search engine that aims to enable patients, clinicians, researchers, and others to conduct searches of health information gleaned from dozens of patient forums and social media sites and share their results graphically [18]. It has received the first place conceptual model in the PCORI Patient-Research Matching Challenge in 2013

Although there is a large number of users in healthcare social media sites, the online social networks are usually sparse. Each user may only interact with a limited number of peers while missing many other peers who have common interests or healthcare concerns. As a result, health consumers are often not connected to those peers who may offer them the best information or the nurturant support they need. Through understanding the user intent and the social support types the health consumers are involved (as discussed in the previous section), we are able to match health consumers with one another to enrich their interactions in healthcare social media sites. In light of this, we investigate an automatic process of classifying user intent and social support types with the human annotated content as the training data set [33,34]. In the classifier, we adopt content analysis and health status as features. The result is promising and it shows that the classification performance can be improved when health status are adopted.

By analyzing the interaction patterns, we have also proposed the UserRank algorithm to rank the user influence in healthcare social media sites [14]. The health consumers who are most active in social media are not necessarily the most influential [20]. Instead, the influence is a measure of how much impact a health consumer has made to the community. By identifying the influential users as well as the explicit and implicit relationships [19], we can utilize the healthcare social network to disseminate the timely and important information to the target users.

V. Drug Safety Signal Detection

Drug safety signal detection is important in postmarketing drug safety surveillance because many potential adverse drug reactions cannot be identified in premarketing review process. 5% of hospital admissions are attributed to adverse drug reactions and many deaths are reported every year. Current drug safety detection techniques relies heavily on resources such as centralized reporting systems, electronic health records, and pharmaceutical databases. However, there is a high under-reporting ratio in the centralized reporting system such as FDA Adverse Event Report System (FAERS) due to the

nature of passiveness. Many adverse drug experiences reported by the health consumers are not necessarily recorded in electronic health records by the health professionals unless sophisticated evaluations are made. In the recent years, there is an increasing effort of detecting the drug safety signals using social media as the resources.

Yang et al. have adopted the expanded health consumer expressions (as discussed in Section II) to discover the discussions on adverse drug reactions on social media sites [24]. By extending the previous effort, Yang et al. have developed association mining [25] and heterogeneous network mining techniques [27] to detect the adverse drug reactions of particular drugs [17] and to detect the drug-drug interactions of any given two drugs [28]. Not only social media data is promising in detecting drug safety signals, they have also conducted temporal analysis and found that the techniques can detect the adverse drug reactions earlier than FDA alters by several years [26]. It can be explained by the fact that health consumers are actively discussing the adverse drug experiences on social media sites before any traditional resources have records of such adverse drug reactions. The heterogeneous network mining techniques also indicate the meaningful paths that involve users, drugs, adverse drug reactions and diseases, which are helpful to present the relationships of drug-drug interactions. For example, some drug-drug interactions cannot be observed by their direct relationships but the interactions can be detected when two drugs are prescribed to patients who have multiple diseases. The heterogeneous network mining is also potential for investigating drug repositioning or off-label use of drugs.

VI. Symptom Profiling and Clustering

There have been many clinical longitudinal studies trying to understand how symptoms are developing over time and how symptoms are correlated. In particular, in cancer treatments, a symptom cluster is defined as three or more concurrent and related symptoms frequently found in patients. Symptom clustering is drawing attention in the recent years. It is because co-existing symptoms may share a common underlying etiology [5,10]. For example, biomarkers such as serum cortisol, melatonin, and serotonin are all related to a cluster of symptoms including fatigue, sleep, and depressive moods during chemotherapy. Examining co-existing symptoms is more efficient and effective than coping with symptoms one by one. It is found that understanding the co-variation in symptoms is helpful in the discovery of physiological mechanisms that lead to the manifestation of disease and side effects of treatment. Previous studies also suggest that intervention improves multiple symptoms concurrently. As a result, there are both clinical and physiological interests in studying symptom clustering.

Clinical studies usually require a lot of effort in recruiting subjects and the same group of subjects may not always be available for a longitudinal study due to the time and geographical constraints. Social media data is an alternative source for symptom clustering. Yang et al. have recently conducted a comparative study of symptom clustering on clinical and social media data for breast cancer [16]. In the study, it is found that there is a substantial agreement between the results derived from the social media data and from the clinical study data. However, there are also some significant discrepancies. In general, we find that there are a couple of clusters with a large number of symptoms and there are also clusters with only one single symptom when the clinical data is used. It can be explained by the fact how the data is collected. In the clinical study, each subject was given a long list of symptoms and was asked to check the symptoms that each had experienced. In such case, the subjects were able to examine the symptoms one by one and checked all those that they had experienced regardless if they had serious concerns on the checked symptoms. On the other hand, the users in healthcare social media sites were voluntarily discussing the symptoms that they concerned. As a result, general symptoms were not discussed as frequent in social media. Clusters of symptoms can be easily identified and the symptoms are more evenly distributed to the clusters when social media is used. The highly correlated symptoms are grouped into the same clusters. In the future, we are also interested to investigate the symptom profiles of patients and how they are correlated.

VII. Conclusion

In this article, we have discussed five emerging issues of patient-centered healthcare informatics research by harnessing healthcare social media. Health consumer expressions are essential to understand the concepts that health consumers are concerning. It is a continuous effort in expanding the health consumer expressions as the vocabularies used in social media are evolving. Social support, which helps to engage users interactions, plays an important role in healthcare social media sites. The underlying social network are useful in understanding the interaction patterns and identifying the influential users; and hence valuable for disseminating timely and important healthcare information. Not only the health consumer expressions are evolving, the hidden communities in healthcare social media sites are evolving. By capturing the dynamic communities, we are able to understand the evolving issues raised by health consumers. Recommendations can also be made effectively to the health consumers in order to enrich the user interactions. Social media data is also useful in knowledge discovery applications such as drug safety signal detection and symptom profiling and clustering. It can supplement the traditional resources such as centralized reporting system, electronic health records, and clinical and pharmaceutical databases. There are also many opportunities of harnessing the social media platforms in patient-centered healthcare informatics research that have not been explored yet. By integrating healthcare sensor data and mobile applications with social media data, a large volume of healthcare data can be collected and more sophisticated analysis can be developed to understand the impact of medical treatments and medications on patients' health conditions. Patient-centered healthcare management system can also be developed to support health consumers in managing their own health and wellbeing. Health consumers are becoming more proactive and preventive. They want to be equipped with knowledge and personalized data analytics to make their own healthcare decisions. As we continue in these efforts, a smart and connected health era may not be too far away.

REFERENCES

[1] K. Chuang and C. C. Yang, "Informational Support Exchanges on Different Computer-Mediated Communication Formats in a Social Media Community of Alcoholism," *Journal of the American Society for Information Science and Technology*, vol.65, no.1, 2014, pp.37-52.

[2] K. Chuang and C. C. Yang, "Interaction Patterns of Nurturant Support Exchanged in Online Health Social Networking," *Journal of Medical Internet Research*, vol. 14, no.3, 2012.

[3] K. Chuang and C. C. Yang, "A Study of Social Positions in an Online Alcoholism Community," *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Washington, D.C., April 2 – 5, 2013.

[4] K. Chuang and C. C. Yang, "A Study of Informational Support Exchanges in MedHelp Alcoholism Community," *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, College Park, MD, April 3-5, 2012.

[5] M. Dodd, S. Janson, N. Facione et al.m "Advancing the Science of Symptom Management," *J Adv irs*. 33, 668-676, 2001.

[6] S. Fox. (2011, October 17). *The Social Life of Health Information*. Available: http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info/Summary-of-Findings.aspx

[7] L. Jiang and C. C. Yang, "Expanding Consumer Health Vocabularies by Learning Consumer Health Expressions from Online Health Social Media," *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction 2015*, Washington, D. C., April 1 - 3, 2015.

[8] L. Jiang and C. C. Yang, "Using Co-occurrence Analysis to Expand Consumer Health Vocabularies from Social Media Data," *Proceedings of IEEE International Conference on Healthcare Informatics*, Philadelphia, PA, September 8 - 11, 2013.

[9] L. Jiang, C. C. Yang, J. Li, "Discovering Consumer Health Expressions from Consumer-Contributed Content," *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Washington, D.C., April 2 – 5, 2013.

[10] T. B. Patrick, H. K. Monga, M. E. Sievert, J. Houston Hall, and D. R. Longo, "Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes," Journal of medical Internet research, vol. 3, p. E24, 2001.

[11] J. K. Payne, B. F. Piper, I. Rabinowitz, M. B. Zimmerman, "Biomarkers, Fatigue, Sleep, and Depressive Symptoms in Women With Breast Cancer: a Pilot Study," *Oncology Nursing Forum*, 33, 775–783, 2006.

[12] X. Tang and C. C. Yang, "Detecting Social Media Hidden Communities using Dynamic Stochastic Blockmodel with Temporal Dirichlet Process," *ACM Transactions on Intelligent Systems and Technology*, accepted for publication.

[13] X. Tang, M. Zhang, and C. C. Yang, "Leveraging User Interest to Improve Thread Recommendation in Online Forum," *Proceedings of International Conference on Social Intelligence and Technology*, State College, PA, May 8 – 9, 2013.

[14] X. Tang and C. C. Yang, "Ranking User Influence in Healthcare Social Media," *ACM Transactions on Intelligent Systems and Technology*, vol.3, no.4, 2012.

[15] X. Tang, M. Zhang, and C. C. Yang, "User Interest and Topic Detection for Personalized Recommendation," *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, Macau, December 4-7, 2012.

[16] C. C. Yang, E. Ip, N. Avis, Q. Ping, and L. Jiang, "A Comparative Study of Symptom Clustering on Clinical and Social Media Data," *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction 2015*, Washington, D. C., April 1 - 3, 2015

[17] C. C. Yang, H. Yang, and L. Jiang, "Postmarketing Drug Safety Surveillance using Publicly Available Health Consumer Contributed Content in Social Media," *ACM Transactions on Management Information Systems*, vol.5, no.1, April, 2014.

[18] C. C. Yang, S. Lin, M. C. Kim, L. Jiang, "ACTONNECT: A Platform to Support Patients and Researchers Collaboration," *Proceedings of IEEE International Conference on Healthcare Informatics 2014 (ICHI 2014)*, Verona, Italy, September 15-17, 2014.

[19] C. C. Yang, X. Tang, H. Yang, and L. Jiang, "Identifying Implicit and Explicit Relationships in Social Commerce Activities," *International Journal of Electronic Commerce*, vol.18, no.2, Winter, 2013-2014, pp.73-96.

[20] C. C. Yang and X. Tang, "Estimating User Influence in the MedHelp Social Network," *IEEE Intelligent Systems*, vol.27, no.5, 2012, pp.44-50.

[21] C. C. Yang, G. Leroy, and S. Ananiadou, "Smart Health and Wellbeing," *ACM Transactions on Management Information Systems*, vol.4, no.4, 2013.

[22] C. C. Yang, R. Chiu, S. Lin, and A. Kumar, "Patient-Centered Research and Social Media," *IEEE Life Sciences Newsletter*, 2013.

[23] C. C. Yang, J. Yen, and J. Liu, "Social Intelligence and Technology," *IEEE Intelligenct System*, vol.29, no.2, March-April, 2014.

[24] C. C. Yang, L. Jiang, H. Yang, and X. Tang, "Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media," *Proceedings of ACM SIGKDD Workshop on Health Informatics*, Beijing, 2012.

[25] C. C. Yang, H. Yang, L. Jiang, and M. Zhang, "Social media mining for drug safety signal detection," *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, 2012, pp. 33-40.

[26] H. Yang and C. C. Yang, "Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis," *ACM Transactions on Intelligent Systems and Technology*, accepted for publication.

[27] H. Yang and C. C. Yang, "Drug-Drug Interactions Detection from Online Heterogeneous Healthcare Network," *Proceedings of IEEE International Conference on Healthcare Informatics 2014 (ICHI 2014)*, Verona, Italy, September 15-17, 2014.

[28] H. Yang and C. C. Yang, "Harnessing Social Media for Drug-Drug Interactions Detection," *Proceedings of IEEE International Conference on Healthcare Informatics*, Philadelphia, PA, September 8 - 11, 2013.

[29] Q. Zeng, S. Kogan, N. Ash, R. A. Greenes, and A. A. Boxwala, "Characteristics of consumer terminology for health information retrieval," Methods Inf Med, vol. 41, pp. 289-298, 2002.

[30] Q. T. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," Journal of the American Medical Informatics Association : JAMIA, vol. 13, pp. 24-29, 2006.

[31] Q. T. Zeng, T. Tse, G. Divita, A. Keselman, J. Crowell, and A. C. Browne, "Exploring Lexical Forms: First-Generation Consumer Health Vocabularies," AMIA Annual Symposium Proceedings, pp. 1155-1155, 2006.

[32] M. Zhang and C. C. Yang, "Using Content and Network Analysis to Understand the Social Support Exchange Patterns and User Behaviors of Online Smoking Cessation Intervention Program" *Journal of the*

American Society for Information Science and Technology, accepted for publication.

[33] M. Zhang and C. C. Yang, "Classifying User Intention and Social Support Types in Online Healthcare Discussions," *Proceedings of IEEE International Conference on Healthcare Informatics 2014 (ICHI 2014)*, Verona, Italy, September 15-17, 2014.

[34] M. Zhang and C. C. Yang, "Classification of Online Health Discussions with Text and Health Features Sets," *Proceedings of AAAI International Workshop on the World Wide Web and Public Health Intelligence 2014 (W3PHI 2014)*, Quebec City, Canada, July 27, 2014.

[35] M. Zhang and C. C. Yang, "Social Support and Exchange Patterns in an Online Smoking Cessation Intervention Program," *Proceedings of IEEE International Conference on Healthcare Informatics*, Philadelphia, PA, September 8 - 11, 2013.

Trends in Interactive Knowledge Discovery for Personalized Medicine: Cognitive Science meets Machine Learning

Andreas Holzinger, *Member, IEEE*

***Abstract*—A grand goal of future medicine is in modelling the complexity of patients to tailor medical decisions, health practices and therapies to the individual patient. This trend towards personalized medicine produces unprecedented amounts of data, and even though the fact that human experts are excellent at pattern recognition in dimensions of ≤ 3, the problem is that most biomedical data is in dimensions much higher than 3, making manual analysis difficult and often impossible. Experts in daily medical routine are decreasingly capable of dealing with the complexity of such data. Moreover, they are not interested the data, they need knowledge and insight in order to support their work. Consequently, a big trend in computer science is to provide efficient, useable and useful computational methods, algorithms and tools to discover knowledge and to interactively gain insight into high-dimensional data. A synergistic combination of methodologies of two areas may be of great help here: Human–Computer Interaction (HCI) and Knowledge Discovery/Data Mining (KDD), with the goal of supporting human intelligence with machine learning. A trend in both disciplines is the acquisition and adaptation of representations that support efficient learning. Mapping higher dimensional data into lower dimensions is a major task in HCI, and a concerted effort of computational methods including recent advances from graph-theory and algebraic topology may contribute to finding solutions. Moreover, much biomedical data is sparse, noisy and time-dependent, hence entropy is also amongst promising topics. This paper provides a rough overview of the HCI-KDD approach and focuses on three future trends: graph-based mining, topological data mining and entropy-based data mining.**

***Index Terms*—HCI-KDD, interactive knowledge discovery, machine learning, graph-based data mining, topological data mining, entropy-based data mining**

I. Introduction

EXPERTS in the life sciences have to deal with large amounts of complex, high-dimensional, heterogenous, noisy, and weakly structured data sets [1], [2], and large amounts of unstructured information [3].

This "Big Data" [4] in the medical domain is driven by the trend towards precision P4-medicine (Predictive, Preventive, Participatory, Personalized) [5], [6], and has resulted in an explosion in the amount of generated data sets, in particular "-omics" data, for example from genomics, proteomics, metabolomics, epigenetics, transcriptomics, lipidomics, fluxomics, phenomics, microbiomics, etc. [7], [8], [9]. The trend is in moving from a reactive to a proactive medicine and P4-medicine is closely related to systems approaches to disease and content analytics tools [10], [11]. The well-known challenges with such data include the complexity of feature dimensions (scaling and mapping problems), the heterogeneity of the data (problems of data integration, data fusion), the change over time, and most of all the classic medical data problem: uncertainty of the data quality, false, incomplete data and the danger of modelling artifacts. The often mentioned problem of large amounts of data is rather an advantage with machine learning approaches: Big data actually can provide benefits, as in the biomedical domain, we look often at only a few hundred training examples, so there is the danger of random guessing. Having millions of training samples will raise the precision. The issue of large data sets connects to this question: "What constitutes predictable structures in the world?" as something might be predictable but not comprehensible [12]. Machine learning researchers study algorithms being capable of learning from data and because learning is an important aspect of intelligent behavior, machine learning has become a modern and central aspect of research in artificial intelligence. The most obvious example of learning occur in humans, so there is a natural bridge between research in machine learning and cognitive science, which is strongly related to HCI.

The paradigmatic shift, from classical science, where you first have the question and then collect the data, to data sciences, where you first have the data and then ask questions [13]. The main challenge in this new approach is to ask relevant questions so to find relevant *structural* patterns and/or *temporal* patterns ("knowledge") in such data, because those are often hidden and not directly accessible to the expert [14].

This paper is organized as follows: In section 2 some key terms are briefly explained. In section 3 the basic idea of the HCI-KDD approach is presented, along with the seven research areas involved, however, in the following we concentrate briefly on only three of them: In section 4 on graph-based data mining, in section 5 on topological data mining and in section 6 on entropy-based data mining, concluding by emphasizing that the *combination* of such approaches may bring added values. In the limited space given, such vast topics can only be touched, so the goal of this tutorial is to provide a coarse overview, to motivate and stimulate further research and to encourage to test crazy ideas.

II. Glossary and Key Terms

- **Algebraic Topology:** is concerned with computations of homologies and homotopies in topological spaces [15].

A. Holzinger is lead of the research unit HCI-KDD at the Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Austria, Web: hci-kdd.org e-mail: a.holzinger@hci-kdd.org

- **Alpha Shapes:** family of piecewise linear simple curves in the Euclidean plane associated with the shape of a finite set of points [16]; i.e. α-shapes are a generalization of the convex hull of a point set: Let $\mathbf{S}$ be a finite set in $\mathbb{R}^3$ and α a real number $0 \leq \alpha \leq \infty$; the u-shape of $\mathbf{S}$ is a polytope that is neither necessarily convex nor necessarily connected. For $\alpha \to \infty$ the α-shape is identical to the convex hull of $\mathbf{S}$ [17]; important e.g. in protein-related interactions [18].
- **Betti Number:** can be used to distinguish topological spaces based on the connectivity of n-dimensional simplicial complexes: In dimension k, the rank of the k-th homology group is denoted β_k, useful in the presence of noisy shapes, because Betti numbers can be used as shape descriptor admitting dissimilarity distances stable under continuous shape deformations [19].
- **Graph mining:** is the application of graph-based methods to structural data sets [20], a survey on graph mining can be found here [21].
- **Homomorphism:** is a function that preserve the operators associated with the specified structure.
- **Homotopy:** Given two maps $f, g : X \to Y$ of topological spaces, f and g are homotopic, $f \simeq g$, if there is a continuous map $H : X \times [0, 1] \to Y$ so that $H(x, 0) = f(x)$ and $H(x, 1) = g(x)$ for all $x \in X$ [22].
- **Homology:** (and cohomology) are algebraic objects associated to a manifold, which give one measure of the number of holes of the object. Computation of the homology groups of topological spaces is a central topic in topology; if the simplicial complex is small, the homology group computations can be done manually; to solve such problems generally a classic algorithm exists [23].
- **Human–Computer Interaction:** study, design and development of the interaction between end users and computers; this classic definition goes back to the work of Alan Newell and Herbert Simon (refs), and HCI research has in the last decades focused almost exclusively on ergonomics of the user interface, while the HCI-KDD approach concentrates almost exclusively on human–data interaction.
- **Information Entropy:** is a measure of the uncertainty in a random variable. This refers to the Shannon entropy, which quantifies the expected value of the information contained in a message.
- **Manifold:** is a fundamental mathematical object which locally resembles a line, a plane, or space.
- **Network:** Synonym for a graph, which can be defined as an ordered or unordered pair (N, E) of a set N of nodes and a set E of edges [24]. Engineers often mention: Data + Graph = Network, or call at least directed graphs as networks; however, in theory, there is no difference between a graph and a network.
- **Pattern discovery:** subsumes a plethora of machine learning methods to detect complex patterns in data sets [25]; applications thereof are, for instance, graph mining [26] and string matching [27].
- **Persistent Homology:** Persistent homology is an algebraic tool for measuring topological features of shapes and functions. It casts the multi-scale organization we frequently observe in nature into a mathematical formalism [28].
- **Simplicial Complex:** is made up of simplices, e.g. a simplicial polytope has simplices as faces and a simplicial complex is a collection of simplices pasted together in any reasonable vertex-to-vertex and edge-to-edge arrangement. A graph is a 1-dim simplicial complex.
- **Small world networks:** are generated based on certain rules with high clustering coefficient [24], [29] but the distances among the vertices are rather short in average, hence they are somewhat similar to random networks and they have been found in several classes of biological networks, see [30].
- **Topological Entropy:** is a nonnegative real number that is a measure of the complexity of a dynamical system [31].

III. The HCI-KDD approach

The HCI-KDD approach [32] is a beneficial synergistic combination of methodologies and approaches of two areas that offer ideal conditions towards unraveling some of the ”big data” problems mentioned above: Human-Computer Interaction (HCI) and Knowledge Discovery/Data Mining (KDD), with the goal of supporting human intelligence with computational intelligence - by bringing the human into the loop. This approach appreciates both what humans can do best and what computers can do best. A good example for demonstrating the strengths of humans over sophisticated computers is GO, which is a board game from China more than 2,000 years old. It still remains a challenge for computers [33], [34]. Humans are very good at pattern recognition in the low-dimensional space, although humans do not see in three spatial dimensions directly, but via sequences of planar projections. Humans spend a lot of their life time to learn how to infer three-dimensional spatial data from these paired planar projections. Years of practice have tuned a remarkable ability to extract global structures from representations in lower dimension [35]. Kernels in machine learning have a high relevance for understanding issues of generalization and similarity in cognitive science. It is very interesting that most similarity measures considered by psychologists were examples of positive definite kernels, for which a rich body of mathematical theory exists [36]. Consequently, kernel methods can be seen as a unifying theoretical tool showing how several competing and seemingly incommensurate theories in cognitive science (exemplar models versus perceptron models) can be put together [37]. Arguably, the problem of learning represents a gateway to understanding intelligence in both brains and machines, to discovering how the human brain works and to develop intelligent algorithms, which learn from data and improve their competencies - the same as children do [38].

On the other hand, computers can be very beneficial in dealing with high-dimensional data, where we can make use of the benefits of computational topology [39], e.g. by replacing

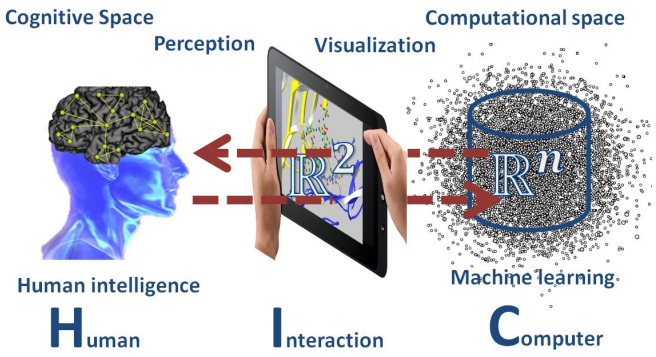

Fig. 1. This image, created originally by A. Holzinger as logo for his group hci-kdd.org, shall emphasize the importance of the manipulating data in the high-dimensional computational space in $\mathbb{R}^n$ and highlights the reality that current devices only allow data visualization in $\mathbb{R}^2$. Consequently, a major challenge for Human–Computer Interaction is to map data from high-dimensional spaces into lower-dimensional spaces.

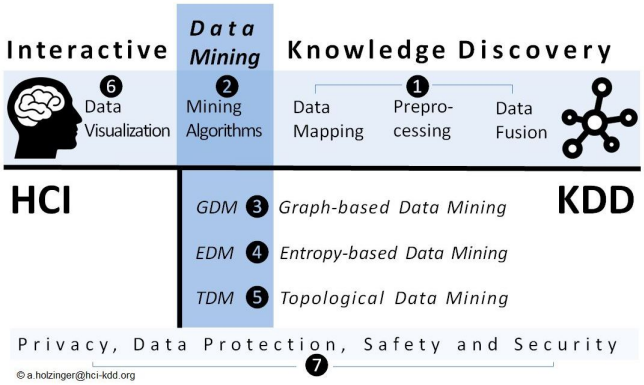

Fig. 2. The big picture of the HCI-KDD approach: KDD encompasses the whole horizontal process chain from data to information and knowledge; actually from physical aspects of raw data, to human aspects including attention, memory, vision, interaction etc. as core topics in HCI, whilst DM as a vertical subject focuses on the development of methods, algorithms and tools for data mining (Image taken from the hci-kdd.org website, as of December, 19, 2014)

a set of point cloud data with a simplicial complex, which converts the data into global topological objects. To combine the most desirable of these formidable talents might highly benefit the knowledge discovery process [32], [40] however, the most critical and most difficult part is in interaction and visualization (see Figure 1).

The original idea of the HCI-KDD [41] approach (Figure 2) is in combining aspects of the best of two worlds: Human–Computer Interaction (HCI), with emphasis on perception, cognition, interaction, reasoning, decision making, human learning and human intelligence, and Knowledge Discovery/Data Mining (KDD), dealing with data processing, computational statistics, artificial intelligence and particularly with integrative machine learning [42]. The most important aspect is the human-in-the-loop approach. Meanwhile it is acknowledged that in many domains computational approaches can not be completely automated - especially in the biomedical domain. The domain knowledge of the expert is of extreme importance and the grand goal is to enable them to interactively manipulate their data, so that they can interactively ask questions to their data sets. An early example for such an approach was given in the medical radiology domain: The clinically useful information in an image typically consists of gray level variations in highly localized regions of the x-ray image and to extract such regions automatically by standard image processing techniques is a hard problem. To bring the physician-in-the-loop means that the expert delineates the pathology bearing regions and a set of anatomical landmarks in the image. To the so marked regions, low-level computer vision tools and image processing algorithms can be applied to extract attributes related to the variations in gray scale [43]. A more recent emphasise of interaction of that kind can be found in [44] and [45].

Whilst interactive knowledge discovery encompasses the horizontal process ranging from physical aspects of data (left in Figure 2) to the human aspects of information processing (right in Figure 2), data mining can be seen vertically and deals specifically with methods, algorithms and tools for finding patterns in the data. In the HCI-KDD approach, seven (the new magical number 7) essential research areas can be determined as outlined in Figure 2, including: Area 1: Data integration, data fusion and data mapping; Area 2: mining algorithms and Area 6: data visualization [46], [47], [48]. The remainder of this paper focuses on three hot topics, ***Area 3: Graph-based Data Mining (GDM)*** [49], [50], [51], [52]. ***Area 4: Entropy-based Data Mining (EDM)*** [53], [54], and ***Area 5: Topological Data Mining (TDM)*** [55].

In the biomedical domain as in some other domains issues of Area 7: privacy, data protection, safety and security are mandatory [56].

IV. Graph-based Data Mining

Graphs have been used in the life sciences for quite a time and there is a new trend to combine graph theory, machine learning, and statistical data analysis to arrive at a new field, network analysis, to explore complex biomedical graph data. Large-scale generation of genomics, proteomics, metabolomic etc. and signaling data allows the construction of networks that provide a new framework for understanding the molecular basis of physiological and pathological states. Networks and network-based methods have been used in biology to characterize genomic and genetic mechanisms as well as protein signaling; diseases are researched as abnormal perturbations of critical cellular networks. Onset, progression, and intervention in complex diseases including cancer and diabetes can be analyzed today using network approaches. Once the system is represented by a graph = network, methods of graph theory can be applied to find novel insights, important system properties, in structure, time and function. Various statistical and machine learning methods have been developed for this purpose and have already been applied to networks [57], [58]. Graph theory provides powerful tools to map data structures and to find novel connections between single data objects [24], [59]. A mapping of already existing and in medical practice approved *knowledge spaces* as a conceptual graph (as e.g. demonstrated in [50] and a subsequent visual and

graph-theoretical analysis can bring novel insights on hidden patterns in the data, which exactly is the goal of knowledge discovery. Another benefit of a graph-based data structure is in the applicability of methods from network topology and network analysis and data mining, for example the small-world phenomenon [60], [61], and cluster analysis [62], [63].

The first question is "How to get a graph?", or simpler "How to get point sets?", because point cloud data sets (PCD) can be used as primitives for such approaches. The answer to this question is not trivial [64], apart from "naturally available" point clouds, e.g. from laser scanners [65], protein structures [66], or text mapped into a set of points (vectors) in $\mathbb{R}^n$ [67]. Looking at the last example, graphs are intuitively more informative as example words/phrase representations [68], and graphs are the best studied data structures in computer science, with a strong relation to logical languages [69]. The beginning of graph-based data mining approaches was two decades ago, some pioneering work include [70]–[72]. According to [69] there are five theoretical bases of graph-based data mining approaches such as (1) subgraph categories, (2) subgraph isomorphism, (3) graph invariants, (4) mining measures and (5) solution methods. Furthermore, there are five groups of different graph-theoretical approaches for data mining such as (1) greedy search based approach, (2) inductive logic programming based approach, (3) inductive database based approach, (4) mathematical graph theory based approach and (5) kernel function based approach [73]. However, the main disadvantage of graph-theoretical text mining is the computational complexity of the graph representation, consequently the goal of future research in the field of graph-theoretical approaches for text mining is to develop efficient graph mining algorithms which implement effective search strategies and data structures [68].

In [74] a graph-theoretical approach for text mining is used to extract relation information between terms in "free-text" electronic health care records that are semantically or syntactically related. Another field of application is the text analysis of web and social media for detecting influenza-like illnesses [75].

Moreover there can be content-rich relationship networks among biological concepts, genes, proteins and drugs developed with topological text data mining like shown in [76]. According to [77] network medicine describes the clinical application field of topological text mining due to addressing the complexity of human diseases with molecular and phenotypic network maps.

A recent example is PEGASUS, an open source graph mining library, which performs typical graph mining tasks such as computing the diameter of a graph,the radius of each node and finding connected components. PEGASUS is implemented on the HADOOP platform, the open source version of MAPREDUCE. Many graph mining operations (Page Rank, spectral clustering, diameter estimation, connected components etc.) are a repeated matrix-vector multiplication; in PEGASUS the authors use a primitive, called generalized iterated matrix-vector multiplication, which is optimized and achieved good performances tested with a Web graph with 6,7 billion edges [78].

V. Topological Data Mining

Closely related to graph-based methods are topological data mining methods; for both we need point cloud data sets - or at least distances - as input. A set of such primitives forms a space, and if we have finite sets equipped with proximity or similarity measure functions $sim_q \colon S^{q+1} \rightarrow [0,1]$, which measure how "close" or "similar" $(q+1)$-tuples of elements of S are, we speak about a *topological space*. A value of 0 means totally different objects, while 1 corresponds to equivalent items. Interesting are manifolds, which can be seen as a topological space, which is locally homeomorphic (that means it has a continuous function with an inverse function) to a real n-dimensional space. In other words: X is a d-manifold if every point of X has a neighborhood homeomorphic to $\mathbb{B}^d$; with boundary if every point has a neighborhood homeomorphic to $\mathbb{B}$ or $\mathbb{B}^d_+$ [79].

A topological space may be viewed as an abstraction of a metric space, and similarly, manifolds generalize the connectivity of d-dimensional Euclidean spaces $\mathbb{B}^d$ by being locally similar, but globally different. A d-dimensional chart at $p \in X$ is a homeomorphism $\phi : U \rightarrow \mathbb{R}^d$ onto an open subset of $\mathbb{R}^d$, where U is a neighborhood of p and open is defined using the metric. A d-dimensional manifold (d-manifold) is a topological space X with a d-dimensional chart at every point $x \in X$ [80].

For us also interesting are simplicial complexes ("simplicials") which are spaces described in a very particular way, the basis is in Homology. The reason is that it is not possible to represent surfaces precisely in a computer system due to limited computational storage; thus, surfaces are sampled and represented with triangulations. Such a triangulation is called a simplicial complex, and is a combinatorial space that can represent a space. With such simplicial complexes, the topology of a space from its geometry can be separated. Zomorodian [80] compares it with the separation of syntax and semantics in logic.

Topological techniques originated in pure mathematics, but have been adapted to the study and analysis of data during the past two decades. The two most popular topological techniques in the study of data are *homology* and *persistence*. The connectivity of a space is determined by its cycles of different dimensions. These cycles are organized into groups, called homology groups. Given a reasonably explicit description of a space, the homology groups can be computed with linear algebra. Homology groups have a relatively strong discriminative power and a clear meaning, while having low computational cost. In the study of persistent homology the invariants are in the form of persistence diagrams or barcodes [81].

In data mining it is important to extract significant features, and exactly for this, topological methods are useful, since they provide robust and general feature definitions with emphasis on global information, for example Alpha Shapes [17].

A recent example for topological data mining is given by [82]: Topological text mining, which builds on the well-known vector space model, which is a standard approach in text mining [83]: a collection of text documents (corpus) is mapped into points (=vectors) in $\mathbb{R}^n$. Moreover, each word

can be mapped into so-called term vectors, resulting in a very high dimensional vector space. If there are n words extracted from all the documents then each document is mapped to a point (*term vector*) in $\mathbb{R}^{\kappa}$ with coordinates corresponding to the weights. This way the whole corpus can be transformed into a point cloud data set. Instead of the Euclidean metric the use of a similarity (proximity) measure is sometimes more convenient; the *cosine similarity measure* is a typical example: the cosine of the angle between two vectors (points in the cloud) reflects how "similar" the underlying weighted combinations of keywords are. Amongst the many different text mining methods (for a recent overview refer to [84]); topological approaches are promising, but need a lot of further research.

Due to finding meaningful topological patterns greater information depth can be achieved from the same data input [85]. However, with increasing complexity of the data to process also the need to find a scalable shape characteristic is greater [86]. Therefore methods of the mathematical field of topology are used for complex data areas like the biomedical field [86], [81]. Topology as the mathematical study of shapes and spaces that are not rigid [86], pose a lot of possibilities for the application in knowledge discovery and data mining, as topology is the study of connectivity information and it deals with qualitative geometric properties [87].

One of the main tasks of applied topology is to find and analyse higher dimensional topological structures in lower dimensional spaces (e.g. point cloud from vector space model as discussed in [85]). A common way to describe topological spaces is to first create simplicial complexes, because a simplicial complex structure on a topological space is an expression of the space as a union of simplices such as points, intervals, triangles, and higher dimensional analogues. Simplicial complexes provide an easy combinatorial way to define certain topological spaces [87]. A simplical complex K is defined as a finite collection of simplices such that $\sigma \in K$ and τ, which is a face of σ, implies $\tau \in K$, and $\sigma, \sigma' \in K$ implies $\sigma \cap \sigma'$ can either be a face of both σ and σ' or empty [88]. One way to create a simplical complex is to examine all subsets of points, and if any subsets of points are close enough, a p-simplex (e.g. line) is added to the complex with those points as vertices. For instance, a Vietoris-Rips complex of diameter ϵ is defined as $VR(\epsilon) = \sigma | diam(\sigma) \leq \epsilon$, where $diam(\epsilon)$ is defined as the largest distance between two points in σ [88]. Figure 2 shows the Vietoris-Rips complex with varying ϵ for four points with coordinates (0,0), (0,1), (2,1), (2,0). A common way a analyse the topological structure is to use persistent homology, which identifies cluster, holes and voids therein. It is assumed that more robust topological structures are the one which persist with increasing ϵ. For detailed information about persistent homology, it is referred to [88].

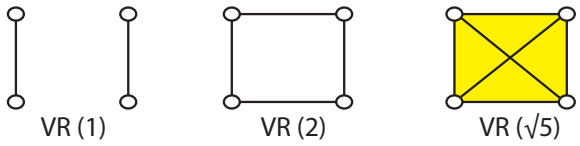

Fig. 3. Vietoris-Rips complex of four points with varying ϵ [88].

VI. Entropy-based Data Mining

In the real medical world, we are confronted not only with complex and high-dimensional data sets, but usually with sparse, noisy, incomplete and uncertain data, where the application of traditional methods of knowledge discovery and data mining always entail the danger of modeling artifacts. Originally, information entropy was introduced by Shannon (1949), as a measure of *uncertainty in the data*. To date, there have emerged many different types of entropy methods with a large number of different purposes and applications. Here we mention only two:

Graph Entropy was described by [89] to measure structural information content of graphs, and a different definition, more focused on problems in information and coding theory, was introduced by Körner in [90]. Graph entropy is often used for the characterization of the structure of graph-based systems, e.g. in mathematical biochemistry, but also for any complex network [91]. In these applications the entropy of a graph is interpreted as its structural information content and serves as a complexity measure, and such a measure is associated with an equivalence relation defined on a finite graph; by application of Shannons Eq. 2.4 in [92] with the probability distribution we get a numerical value that serves as an index of the structural feature captured by the equivalence relation [92].

Topological Entropy (TopEn), was introduced by [93] with the purpose to introduce the notion of entropy as an invariant for continuous mappings: Let (X, T) be a topological dynamical system, i.e., let X be a nonempty compact Hausdorff space and $T : X \to X$ a continuous map; the TopEn is a nonnegative number which measures the complexity of the system [94].

Hornero et al. [95] performed a complexity analysis of intracranial pressure dynamics during periods of severe intracranial hypertension. For that purpose they analyzed eleven episodes of intracranial hypertension from seven patients. They measured the changes in the intracranial pressure complexity by applying ApEn, as patients progressed from a state of normal intracranial pressure to intracranial hypertension, and found that a decreased complexity of intracranial pressure coincides with periods of intracranial hypertension in brain injury. Their approach is of particular interest to us, because they proposed classification based on ApEn tendencies instead of absolute values.

Pincus et al. took in [96] heart rate recordings of 45 healthy infants with recordings of an infant one week after an aborted sudden infant death syndrom (SIDS) episode. They then calculated the ApEn of these recordings and found a significant smaller value for the aborted SIDS infant compared to the healthy ones.

Holzinger et al. (2012) [97] experimented with point cloud data sets in the two dimensional space: They developed a model of handwriting, and evaluated the performance of entropy based slant and skew correction, and compared the results to

other methods. This work is the basis for further entropy-based approaches, which are very relevant for advanced entropy-based data mining approaches.

VII. Conclusion, Open Questions and Future Outlook

Advances in knowledge discovery in complex, high-dimensional data sets need a concerted effort of various topics, ranging from data preprocessing, data fusion, data integration and data mapping to interactive visualization within a low-dimensional space. For this reason, graph-based and topological methods are very useful, since they provide robust and general feature definitions and may support a "global information view". A promising area of future research is in graph-theoretical approaches for text mining, in particular to develop efficient graph mining algorithms which implement robust and efficient search strategies and data structures [68]. Such approaches can be combined with techniques from machine learning, e.g. multi-agents and evolutionary algorithms [98], [99], [49]. However, there remain many open questions, for example about the graph characteristics and the isomorphism complexity [69]. Not only such specific questions are challenging, there are some grand challenges directly involved, e.g. there is much work available on feature selection

As [37] pointed out, there is a large literature on feature selection in machine learning, especially in conjunction with kernel methods, but there are many more methods that could potentially be useful for identifying features, or corresponding similarity measures and in many situations in the real-world a human category learner has to learn the right features (or the right similarity measure), at the same time as he or she learns the categories [100] and machine learning methods can provide hypotheses on how a human learner might achieve this.

It is interesting that much work in cognitive science and machine learning has focused on either supervised or unsupervised learning, i.e. scenarios where either the category labels for all of the stimuli or for none of the stimuli are provided. However, in the real world semi-supervised learning can be beneficial [101].

A definitive challenge when mining high-dimensional data is in measuring distances, e.g. for clustering, outlier detection, similarity measures etc.) as interesting patterns might occur in different subspaces.

A further promising research route is to combine such methods with entropy-based approaches, which have extensively been applied for analyzing sparse and noisy time series data, but so far have not yet been applied to weakly structured data in combination with techniques from computational topology. Consequently, the inclusion of entropy measures for discovery of knowledge in high-dimensional biomedical data is a big future issue, opening a lot of challenging research routes [53].

The grand vision for the future is to effectively support human learning with machine learning. The human brain is an extremely complex organ and can perform many tasks efficiently and effective by (human) learning, particularly when humans are faced with problems that they were faced throughout human evolution (recognizing the Grizzly bear behind you), so we have to keep in mind that our brain can be seen as a statistical decison-making organ, however, only those tasks, which were most important during evolution, are handled most optimal.

The HCI-KDD network of excellence is proactively supporting this vision in bringing together experts with diverse background, but sharing a common goal. A recent output of the network can be found here [102] (for more information please refer to www.hci-kdd.org).

Acknowledgment

This is an expanded version of my Extravaganza Tutorial at the WIC 2014 conference in Warsaw. The author is grateful for the friendly support of Dominik Slezak, Jerzy Stefanowski, Juzhen Dong, Andrzej Skowron, William K.W. Cheung, and for fruitful discussions with members of the HCI-KDD network. Moreover, I thank my Institutes both at Graz University of Technology and the Medical University of Graz, my group, my colleagues and my students for the enjoyable academic freedom, enabling me to think about crazy ideas.

References

[1] A. Holzinger, M. Dehmer, and I. Jurisica, "Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions," *BMC Bioinformatics*, vol. 15, no. Suppl 6, p. I1, 2014.

[2] A. Holzinger, *Biomedical Informatics: Discovering Knowledge in Big Data*. New York: Springer, 2014.

[3] A. Holzinger, C. Stocker, B. Ofner, G. Prohaska, A. Brabenetz, and R. Hofmann-Wellenhof, *Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an assistive technology in the biomedical domain*. Heidelberg, Berlin, New York: Springer, 2013, pp. 13–24.

[4] X. D. Wu, X. Q. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[5] A. Breso, C. Saez, J. Vicente, F. Larrinaga, M. Robles, and J. M. Garcia-Gomez, "Knowledge-based personal health system to empower outpatients of diabetes mellitus by means of p4 medicine," *Methods in molecular biology (Clifton, N.J.)*, vol. 1246, 2015.

[6] P. B. T. R. P. A. H. Klaus Donsa, Stephan Spat, *Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges*. Heidelberg, Berlin: Springer, 2015, pp. 235–260.

[7] B. Huppertz and A. Holzinger, "Biobanks a source of large biological data sets: Open problems and future challenges," in *Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics, Lecture Notes in Computer Science, LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Heidelberg, Berlin: Springer, 2014, pp. 317–330.

[8] J. E. Sackman and M. Kuchenreuther, "Marrying big data with personalized medicine," *Biopharm International*, vol. 27, no. 8, pp. 36–38, 2014, iSI Document Delivery No.: AM9UC Times Cited: 0 Cited Reference Count: 7 Sackman, Jill E. Kuchenreuther, Michael Advanstar communications inc Duluth.

[9] E. D. Perakslis and J. Shon, "Translational informatics in personalized medicine: an update for 2014," *Personalized Medicine*, vol. 11, no. 3, pp. 339–349, 2014.

[10] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (p4) cancer medicine," *Nature Reviews Clinical Oncology*, vol. 8, no. 3, pp. 184–187, 2011.

[11] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine*, vol. 363, no. 4, pp. 301–304, 2010.

[12] O. Chapelle, B. Schoelkopf, and A. Zien, *Semi-supervised learning*. Cambridge: MIT press Cambridge, 2006.

[13] C. A. Mattmann, "Computing: A vision for data science," *Nature*, vol. 493, no. 7433, pp. 473–475, 2013.

[14] D. Otasek, C. Pastrello, A. Holzinger, and I. Jurisica, "Visual data mining: Effective exploration of the biological universe," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Heidelberg, Berlin: Springer, 2014, p. 1934.

[15] A. Hatcher, *Algebraic Topology*. Cambridge: Cambridge University Press, 2002.

[16] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Transactions on Information Theory*, vol. 29, no. 4, pp. 551–559, 1983.

[17] H. Edelsbrunner and E. P. Mucke, "3-dimensional alpha-shapes," *ACM Transactions on Graphics*, vol. 13, no. 1, pp. 43–72, 1994.

[18] L. P. Albou, B. Schwarz, O. Poch, J. M. Wurtz, and D. Moras, "Defining and characterizing protein surface using alpha shapes," *Proteins-Structure Function and Bioinformatics*, vol. 76, no. 1, pp. 1–12, 2009.

[19] P. Frosini and C. Landi, "Persistent betti numbers for a noise tolerant shape-based approach to image retrieval," *Pattern Recognition Letters*, vol. 34, no. 8, pp. 863–872, 2013.

[20] D. Cook and L. B. Holder, *Mining Graph Data*. Wiley-Interscience, 2007.

[21] D. Chakrabarti and C. Faloutsos, "Graph mining: Laws, generators, and algorithms," *ACM Computing Surveys (CSUR)*, vol. 38, no. 1, p. 2, 2006.

[22] G. W. Whitehead, *Elements of homotopy theory*. Springer, 1978.

[23] J. R. Munkres, *Elements of algebraic topology*. Addison-Wesley Reading, 1984, vol. 2.

[24] S. Dorogovtsev and J. Mendes, *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, 2003.

[25] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification. Second Edition*. New York et al.: Wiley, 2000.

[26] D. J. Cook and L. B. Holder, "Graph-based data mining," *IEEE Intelligent Systems and their Applications*, vol. 15, no. 2, pp. 32–41, 2000.

[27] D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[28] H. Edelsbrunner and J. Harer, *Persistent homology - a survey*, ser. Contemporary Mathematics Series. Providence (RI): Amer Mathematical Soc, 2008, vol. 453, pp. 257–282.

[29] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[30] F. Emmert-Streib and M. Dehmer, "Networks for systems biology: Conceptual connection of data and function," *IET Systems Biology*, vol. 5, pp. 185–207, 2011.

[31] D. Koslicki, "Topological entropy of dna sequences," *Bioinformatics*, vol. 27, no. 8, pp. 1061–1067, 2011.

[32] A. Holzinger, "Human–computer interaction & knowledge discovery (hci-kdd): What is the benefit of bringing those two fields to work together?" in *Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127*, A. Cuzzocrea, C. Kittl, D. E. Simos, E. Weippl, and L. Xu, Eds. Heidelberg, Berlin, New York: Springer, 2013, pp. 319–328.

[33] D. Lichtenstein and M. Sipser, "Go is polynominal space hard," *Journal of the ACM (JACM)*, vol. 27, no. 2, pp. 393–401, 1980.

[34] C. S. Lee, O. Teytaud, M. H. Wang, and S. J. Yen, "Computational intelligence meets game of go @ ieee wcci 2012," *Ieee Computational Intelligence Magazine*, vol. 7, no. 4, pp. 10–12, 2012.

[35] S. Edelman and R. Shahbazi, "Renewing the respect for similarity," *Frontiers in Computational Neuroscience*, vol. 6, 2012.

[36] B. Schoelkopf, K. Tsuda, and J.-P. Vert, *Kernel methods in computational biology*. Cambridge (MA): The MIT press, 2004.

[37] F. Jkel, B. Schlkopf, and F. A. Wichmann, "Does cognitive science need kernels?" *Trends in cognitive sciences*, vol. 13, no. 9, pp. 381–388, 2009.

[38] T. Poggio and S. Smale, "The mathematics of learning: Dealing with data," *Notices of the AMS*, vol. 50, no. 5, pp. 537–544, 2003.

[39] R. Ghrist, "Barcodes: the persistent topology of data," *Bulletin of the American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 2008.

[40] A. Holzinger, *Extravaganza Tutorial on Hot Ideas for Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Heidelberg, Berlin: Springer, 2014, pp. 502–515.

[41] ——, "On knowledge discovery and interactive intelligent visualization of biomedical data - challenges in humancomputer interaction and biomedical informatics," in *DATA 2012*. Rome, Italy: INSTICC, 2012, pp. 9–20.

[42] A. Holzinger and I. Jurisica, "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Heidelberg, Berlin: Springer, 2014, p. in print.

[43] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick, "Assert: A physician-in-the-loop content-based retrieval system for hrct image databases," *Computer Vision and Image Understanding*, vol. 75, no. 12, pp. 111–132, 1999.

[44] C. Bauckhage, M. Hanheide, S. Wrede, T. Kaster, M. Pfeiffer, and G. Sagerer, "Vision systems with the human in the loop," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 14, p. 302161, 2005.

[45] G. Schirner, D. Erdogmus, K. Chowdhury, and T. Padir, "The future of human-in-the-loop cyber-physical systems," *Computer*, vol. 46, no. 1, pp. 36–45, 2013.

[46] A. Holzinger, M. Bruschi, and W. Eder, "On interactive data visualization of physiological low-cost-sensor data with focus on mental stress," in *Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127*, A. Cuzzocrea, C. Kittl, D. E. Simos, E. Weippl, and L. Xu, Eds. Heidelberg, Berlin: Springer, 2013, p. 469480.

[47] B. L. W. Wong, K. Xu, and A. Holzinger, "Interactive visualization for information analysis in medical diagnosis," in *Information Quality in e-Health, Lecture Notes in Computer Science, LNCS 7058*, A. Holzinger and K.-M. Simonic, Eds. Springer Berlin Heidelberg, 2011, pp. 109–120.

[48] M. Wiltgen, A. Holzinger, and G. Tilz, *Interactive Analysis and Visualization of Macromolecular Interfaces Between Proteins*. Berlin, Heidelberg, New York: Springer, 2007, pp. 199–212.

[49] M. Preuss, M. Dehmer, S. Pickl, and A. Holzinger, "On terrain coverage optimization by using a network approach for universal graph-based data mining and knowledge discovery," in *Active Media Technology - 10th International Conference, AMT 2014, Warsaw, Poland, August, 11-14, 2014. Proceedings. Lecture Notes in Computer Science LNCS*. Heidelberg, Berlin: Springer, 2014, p. in print.

[50] A. Holzinger, B. Ofner, and M. Dehmer, "Multi-touch graph-based interaction for knowledge discovery on mobile devices: State-of-the-art and future challenges," in *Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics, Springer Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Berlin, Heidelberg: Springer, 2014, p. 241254.

[51] A. Holzinger, B. Malle, R. Aigner, and N. Giuliani, "On graph extraction from image data," in *Active Media Technology AMT 2014, Lecture Notes in Computer Science LNCS 8610*, D. Slezak, G. Schaefer, T. S. Vuong, and Y.-S. Kim, Eds. Heidelberg, Berlin: Springer, 2014, p. in print.

[52] A. Holzinger, B. Ofner, C. Stocker, A. C. Valdez, A. K. Schaar, M. Ziefle, and M. Dehmer, *On Graph Entropy Measures for Knowledge Discovery from Publication Network Data*. Heidelberg, Berlin: Springer, 2013, pp. 354–362.

[53] A. Holzinger, M. Hortenhuber, C. Mayer, M. Bachler, S. Wassertheurer, A. Pinho, and D. Koslicki, "On entropy-based data mining," in *Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics. Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Heidelberg, Berlin: Springer, 2014, pp. 209–226.

[54] A. Holzinger, C. Stocker, M. Bruschi, A. Auinger, H. Silva, H. Gamboa, and A. Fred, "On applying approximate entropy to ecg signals for knowledge discovery on the example of big sensor data," in *Active Media Technology, Lecture Notes in Computer Science, LNCS 7669*, R. Huang, A. Ghorbani, G. Pasi, T. Yamaguchi, N. Yen, and B. Jin, Eds., 2012.

[55] A. Holzinger, "On topological data mining," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Heidelberg, Berlin: Springer, 2014, pp. 331–356.

[56] P. Kieseberg, H. Hobel, S. Schrittwieser, E. Weippl, and A. Holzinger, "Protecting anonymity in the data-driven medical sciences," in *Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics, Springer Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Berlin, Heidelberg: Springer, 2014, pp. 303–318.

[57] M. Dehmer and S. C. Basak, *Statistical and Machine Learning Approaches for Network Analysis*. Wiley Online Library, 2012.

[58] M. Dehmer, F. Emmert-Streib, and A. Mehler, *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*. Boston: Birkhaueser, 2011.

[59] S. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[60] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[61] J. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, pp. 845–845, 2000.

[62] W. Koontz, P. Narendra, and K. Fukunaga, "A graph-theoretic approach to nonparametric cluster analysis," *IEEE Transactions on Computers*, vol. 100, no. 9, pp. 936–944, 1976.

[63] T. Wittkop, D. Emig, A. Truss, M. Albrecht, S. Boecker, and J. Baumbach, "Comprehensive cluster analysis with transitivity clustering," *Nature protocols*, vol. 6, no. 3, pp. 285–295, 2011.

[64] A. Holzinger, B. Malle, M. Bloice, M. Wiltgen, M. Ferri, I. Stanganelli, and R. Hofmann-Wellenhof, "On the generation of point cloud data sets: the first step in the knowledge discovery process," in *Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics, Springer Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Berlin, Heidelberg: Springer, 2014, pp. 57–80.

[65] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[66] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack, "A graph-theory algorithm for rapid protein side-chain prediction," *Protein science*, vol. 12, no. 9, pp. 2001–2014, 2003.

[67] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, p. 620, 1975.

[68] C. Jiang, F. Coenen, R. Sanderson, and M. Zito, "Text classification using graph mining-based feature extraction," *Knowledge-Based Systems*, vol. 23, no. 4, pp. 302–308, May 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095070510900152X

[69] T. Washio and H. Motoda, "State of the art of graph-based data mining," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, p. 59, Jul. 2003. [Online]. Available: http://portal.acm.org/citation.cfm?doid=959242.959249

[70] D. J. Cook and L. B. Holder, "Substructure discovery using minimum description length and background knowledge," *J. Artif. Int. Res.*, vol. 1, no. 1, pp. 231–255, Feb. 1994. [Online]. Available: http://dl.acm.org/citation.cfm?id=1618595.1618605

[71] K. Yoshida, H. Motoda, and N. Indurkhya, "Graph-based induction as a unified learning framework," *Applied Intelligence*, vol. 4, no. 3, pp. 297–316, Jul. 1994. [Online]. Available: http://link.springer.com/10.1007/BF00872095

[72] L. Dehaspe and H. Toivonen, "Discovery of frequent DATALOG patterns," *Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 7–36, Mar. 1999. [Online]. Available: http://link.springer.com/article/10.1023/A%3A1009863704807

[73] D. Windridge and M. Bober, "A kernel-based framework for medical big-data analytics," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Heidelberg, Berlin: Springer, 2014, pp. 196–207.

[74] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," in *Proceedings of the 2006 ACM symposium on Applied computing - SAC '06*. New York, New York, USA: ACM Press, Apr. 2006, p. 235. [Online]. Available: http://dl.acm.org/citation.cfm?id=1141277.1141330

[75] C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh, "Text and structural data mining of influenza mentions in Web and social media." *International journal of environmental research and public health*, vol. 7, no. 2, pp. 596–615, Feb. 2010.

[76] H. Chen and B. M. Sharp, "Content-rich biological network constructed by mining PubMed abstracts." *BMC bioinformatics*, vol. 5, no. 1, p. 147, Oct. 2004. [Online]. Available: http://www.biomedcentral.com/1471-2105/5/147

[77] A. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011. [Online]. Available: http://www.nature.com/nrg/journal/v12/n1/abs/nrg2918.html

[78] U. Kang, C. E. Tsourakakis, and C. Faloutsos, "Pegasus: mining peta-scale graphs," *Knowledge and Information Systems*, vol. 27, no. 2, pp. 303–325, 2011.

[79] J. W. Cannon, "The recognition problem: what is a topological manifold?" *Bulletin of the American Mathematical Society*, vol. 84, no. 5, pp. 832–866, 1978.

[80] A. Zomorodian, *Computational Topology*, ser. Chapman & Hall/CRC Applied Algorithms and Data Structures series. Boca Raton (FL): Chapman and Hall/CRC, 2010, pp. 1–31, doi:10.1201/9781584888215-c3.

[81] C. Epstein, G. Carlsson, and H. Edelsbrunner, "Topological data analysis," *Inverse Problems*, vol. 27, no. 12, p. 120201, Dec. 2011. [Online]. Available: http://iopscience.iop.org/0266-5611/27/12/120201

[82] H. Wagner and P. Dlotko, "Towards topological analysis of high-dimensional feature spaces," *Computer Vision and Image Understanding*, vol. 121, pp. 21–26, 2014.

[83] M. Kobayashi and M. Aono, "Vector space models for search and cluster mining," in *Survey of Text Mining: Clustering, Classification, and Retrieval*, M. W. Berry, Ed. New York: Springer, 2004, pp. 103–122.

[84] A. Holzinger, J. Schantl, M. Schroettner, C. Seifert, and K. Verspoor, *Biomedical Text Mining: Open Problems and Future Challenges*. Heidelberg, Berlin: Springer, 2014, pp. 271–300.

[85] H. Wagner, P. Dlotko, and M. Mrozek, "Computational topology in text mining," in *Computational Topology in Image Context*, ser. Lecture Notes in Computer Science, M. Ferri, P. Frosini, C. Landi, A. Cerri, and B. Fabio, Eds. Springer Berlin Heidelberg, 2012, vol. 7309, pp. 68–78.

[86] M. Nicolau, A. J. Levine, and G. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 17, pp. 7265–70, Apr. 2011. [Online]. Available: http://www.pnas.org/content/108/17/7265.short

[87] G. Carlsson, "Topology and Data," *Bull. Amer. Math. Soc.*, vol. 46, pp. 255–308, 2009. [Online]. Available: http://comptop.stanford.edu/u/preprints/topologyAndData.pdf

[88] X. Zhu, "Persistent homology: An introduction and a new text representation for natural language processing." in *IJCAI*, F. Rossi, Ed. IJCAI/AAAI, 2013.

[89] A. Mowshowitz, "Entropy and the complexity of graphs: I. an index of the relative complexity of a graph," *The bulletin of mathematical biophysics*, vol. 30, no. 1, pp. 175–204, 1968.

[90] J. Körner, "Coding of an information source having ambiguous alphabet and the entropy of graphs," in *6th Prague Conference on Information Theory*, 1973, pp. 411–425.

[91] A. Holzinger, B. Ofner, C. Stocker, A. C. Valdez, A. K. Schaar, M. Ziefle, and M. Dehmer, "On graph entropy measures for knowledge discovery from publication network data," in *Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127*, A. Cuzzocrea, C. Kittl, D. E. Simos, E. Weippl, and L. Xu, Eds. Heidelberg, Berlin: Springer, 2013, pp. 354–362.

[92] M. Dehmer and A. Mowshowitz, "A history of graph entropy measures," *Information Sciences*, vol. 181, no. 1, pp. 57–78, 2011.

[93] R. L. Adler, A. G. Konheim, and M. H. McAndrew, "Topological entropy," *Transactions of the American Mathematical Society*, vol. 114, no. 2, pp. 309–319, 1965.

[94] R. Adler, T. Downarowicz, and M. Misiurewicz, "Topological entropy," *Scholarpedia*, vol. 3, no. 2, p. 2200, 2008.

[95] R. Hornero, M. Aboy, D. Abasolo, J. McNames, W. Wakeland, and B. Goldstein, "Complex analysis of intracranial hypertension using approximate entropy." *Crit Care Med*, vol. 34, no. 1, pp. 87–95, 2006.

[96] S. M. Pincus, "Approximate entropy as a measure of system complexity." *Proceedings of the National Academy of Sciences*, vol. 88, no. 6, pp. 2297–2301, 1991.

[97] A. Holzinger, C. Stocker, B. Peischl, and K.-M. Simonic, "On using entropy for enhancing handwriting preprocessing," *Entropy*, vol. 14, no. 11, pp. 2324–2350, 2012.

[98] K. Holzinger, V. Palade, R. Rabadan, and A. Holzinger, "Darwin or lamarck? future challenges in evolutionary algorithms for knowledge discovery and data mining," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401*, A. Holzinger and I. Jurisica, Eds. Heidelberg, Berlin: Springer, 2014, p. in print.

[99] A. Holzinger, D. Blanchard, M. Bloice, K. Holzinger, V. Palade, and R. Rabadan, "Darwin, lamarck, or baldwin: Applying evolutionary algorithms to machine learning techniques," in *The 2014 IEEE/WIC/ACM*

International Conference on Web Intelligence (WI 2014). IEEE, 2014, pp. 449–453.

[100] P. G. Schyns, R. L. Goldstone, and J.-P. Thibaut, "The development of features in object concepts," *Behavioral and Brain Sciences*, vol. 21, no. 1, pp. 1–17, 1998.

[101] K. Vandist, M. De Schryver, and Y. Rosseel, "Semisupervised category learning: The impact of feedback in learning the information-integration task," *Attention, Perception, and Psychophysics*, vol. 71, no. 2, pp. 328–341, 2009.

[102] A. Holzinger and I. Jurisica, *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401*. Heidelberg, Berlin: Springer, 2014.

Classification Rules in Methods of Clustering

Sadaaki Miyamoto, *Member, IEEE,*

***Abstract*—While classification rules are essential in supervised classification methods, they are not noticed well in methods of clustering. Nevertheless, some clustering techniques have clear rules of classification, while they are not obvious in other methods. This paper discusses classification rules or classification functions in the former class including K-means, fuzzy c-means, and the mixture of distributions, and shows theoretical properties that exhibit the nature of a method in this class. In contrast, linkage methods of agglomerative hierarchical clustering do not appear to have classification rules. We show, however, the single linkage method has the rule of nearest neighbor classification, while other linkage methods not. An advanced method using positive-definite kernels is also discussed.**

***Index Terms*—Agglomerative hierarchical clustering, K-means, fuzzy K-means, mixture of distributions, classification rules, inductive property.**

I. Introduction

DATA clustering, or simply clustering, is becoming one of major tools for analyzing large scale data in this world of the 'big data'. Many years ago, clustering techniques have supplementary roles to supervised classification. Due to the increase of necessities to survey and examine huge and unorganized data collections, we are confronting with more unsupervised cases, and thus unsupervised classification is being noted to be important.

Although there are various methods of unsupervised classification, we discuss solely clustering which has a long history in this class of methods. At least its age is more than 60 years, and on the other hand new methods are developed and various applications are actively studied.

Most papers on methods of clustering have a simple structure:

1) Propose a new algorithm.
2) Apply it to a number of examples and compare results with those by typical existing methods.
3) Show that the proposed method is superior to the compared old methods.

Many studies have been like this, but a fundamental question is: is this way of discussion really useful?

Such discussions may expand methods of clustering, but do not serve deeper understanding of methods of clustering, for which theoretical studies are needed.

Theoretical considerations are minor in foregoing literature but important studies have been done: a typical example is K-means++ [2] where the efficiency of the K-means [9] is improved and theoretical properties on the efficiency of the algorithm is discussed.

S. Miyamoto is with the Department of Risk Engineering, University of Tsukuba, Ibaraki 305-8573 Japan email: (see http://www.risk.tsukuba.ac.jp/miyamoto/index.html).

In this paper we do not consider the efficiency of algorithms, but we study theoretical properties of well-known classes of methods.

What we focus upon is classification rules in some methods of clustering. A classification rule obviously exists in a method of supervised classification, whereas it is ambiguous or unclear in clustering, since clustering implies generation of classes on a set of given objects and nothing more, and thus to have a classification rule does not seem to be a matter of interest. Classification rules are, however, essential to understand theoretical properties of methods of clustering, which we will show in this paper.

We consider two well-known classes of methods for this purpose: first class is the K-means and related methods; second class is the agglomerative hierarchical clustering including different linkage methods.

Some methods in these classes have clearly defined classification rules, while others not. Note also that classification rules may include fuzzy rules or probabilistic rules.

Most discussions in this paper is methodological and examples are simple and for the purpose of illustration.

The rest of this paper is organized as follows. Chapter 2 discusses the K-means and related methods. Not only fuzzy K-means [3], [5] but also the model of mixture of distributions [10] are considered to be related methods to the K-means. Chapter 3 studies agglomerative hierarchical clustering where the single linkage and other linkage methods [6] are contrasted. Chapter 4 finally concludes the paper.

To save space, we omit the proofs of the propositions; they are not difficult and readers may refer to the literature, e.g., [13].

II. K-Means and Related Methods

We begin with giving notations. $X = \{x_1, x_2, \ldots, x_N\}$ is the set of objects for clustering, in which x_k $(k = 1, 2, \ldots, N)$ is a point in $\boldsymbol{R}^p$, $x_k = (x_k,^1, \ldots, x_k^p)^\top \in \boldsymbol{R}^p$. $\boldsymbol{R}^p$ is the p-dimensional Euclidean space with the Euclidean norm $\|x\| = \sqrt{x^\top x}$.

Clusters of X denoted by $G_1, \ldots, G_K$ are subsets of X that form a partition of X:

$$\bigcup_{i=1}^{K} G_i = X, \quad G_i \cap G_j = \emptyset \ \ (i \neq j). \tag{1}$$

However, this property holds only for hard clusters. When we consider fuzzy clusters and probabilistic clusters, the above property should be weakened.

A. The Basic K-Means

The name of K-means comes from the well-known paper of MacQueen [9], but the basic algorithm of the K-means

mentioned in the literature is simpler than the one described in [9]. Actually the name of K-means indicates a class of related algorithms instead of a single algorithm.

We first describe a prototypical procedure for K-means:

A Prototype Procedure for K-Means

1) Generate initial clusters randomly.
2) Determine a prototype vector for each cluster.
3) Allocate each object to the nearest prototype
4) If clusters are convergent, stop. Else go to Step 2).

The above procedure is not an algorithm in a strict sense, since how prototypes are determined is not described.

The reason why we describe this prototype is that different algorithms are expressed as variations of this prototypical procedure; hence they are regarded as members of a family related to K-means prototype. Note also that the number of clusters K should be decided beforehand.

1) The hard K-means: The K-means, which is also called hard K-means, uses centroids, in other words, centers of gravity as the prototypes:

$$v_i = \frac{1}{|G_i|} \sum_{x_k \in G_i} x_k \tag{2}$$

where $|G_i|$ is the number of elements in G_i.

Hence the algorithm **KM** of the K-means becomes as follows:

KM1: Generate initial clusters $G_1, \ldots, G_K$ randomly.

KM2: Calculate cluster prototypes v_i $(1 = 1, \ldots, K)$ by (2).

KM3: Allocate every object $x_k \in X$ to the cluster of the nearest prototype:

$$x_k \to G_i \iff i = \arg \min_{1 \le j \le K} \|x_k - v_i\|^2 \tag{3}$$

KM4: If clusters are convergent, stop. Else go to Step **KM2**.

Another way to determine cluster prototype is in Kohonen's SOM [8]: the VQ (vector quantization) algorithm can be used for clustering, where a learning scheme

$$v_i^{(t+1)} = v_i^{(t)} + \alpha(t)(x_k - v_i^{(t)}) \tag{4}$$

is used for cluster prototypes. Note that t is the number of iterations and $\alpha(t)$ is the learning parameter; x_k in this equation is the last element allocated to cluster G_i.

2) Fuzzy K-means: Fuzzy K-means [5], [3] is a variation of the K-means, where cluster prototypes are fuzzy centroids v_i. Instead of the nearest allocation, fuzzy nearest allocation using membership u_{ki} is used:

$$u_{ki} = \left[\sum_{j=1}^{K} \left(\frac{\|x_k - v_i\|^2}{\|x_k - v_j\|^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{5}$$

$$v_i = \frac{\sum_k (u_{ki})^m x_k}{\sum_k (u_{ki})^m} \tag{6}$$

where $m > 1$ is a fuzzifying parameter. It has been shown that as $m \to 1$, the solutions approach to those of the K-means [13]. The algorithm of fuzzy K-means repeats (5) and (6) until convergence, and hence fuzzy K-means can be regarded as a variation of the K-means.

3) Mixture of Distributions: Usually the model of the mixture of distributions is different from the K-means and fuzzy K-means, but this model can be related to those in a sense. Let us take the most typical case of the Gaussian mixture.

Let us also suppose that covariances for clusters are known and given by

$$\sigma^2 I = \frac{1}{2\lambda} I, \tag{7}$$

where λ is a given positive parameter and I is the identity matrix. This assumption appears a bit strange but is convenient for our purpose.

Then the parameter estimation is only for the averages of the Gaussian distribution. Let v_i is the mean vector and let $u_{ki} = P(G_i|x_k)$. Using the EM algorithm [10], we have

$$u_{ki} = \frac{\exp(-\lambda \|x_k - v_i\|^2)}{\sum_{j=1}^{K} \exp(-\lambda \|x_k - v_j\|^2)}, \tag{8}$$

$$v_i = \frac{\sum_k u_{ki} x_k}{\sum_k u_{ki}}. \tag{9}$$

The algorithm repeats (8) and (9) until convergence. Note that these equations are similar to those for fuzzy K-means. Thus the Gaussian mixture in this restricted form is regarded as a variation of K-means.

4) Fuzzy K-means and Gaussian mixture: We can observe relations between the Gaussian mixture and fuzzy K-means in more detail. For this purpose we review the formulation of fuzzy K-means, which is an alternate optimization of the following objective function:

$$J(U, V) = \sum_{i=1}^{K} \sum_{k=1}^{N} (u_{ki})^m \|x_k - v_i\|^2, \quad (m > 1), \tag{10}$$

with simplified notations of membership matrix $U = (u_{ki})$ and matrix V collecting the prototypes: $V = (v_1, \ldots, v_K)$. A constraint is imposed upon U:

$$M = \{\, U = (u_{ki}) : \sum_{j=1}^{c} u_{kj} = 1, \forall j;\ u_{kj} \ge 0, \forall k, j \,\}. \tag{11}$$

The alternate optimization means that, with a give random initial value of U and/or V, we optimize $J(U, V)$ with respect to U with the previously determined V, and then we optimize $J(U, V)$ with respect to V with the previously determined U, until convergence. As a result the solutions (5) and (6) are obtained and thus we repeat (5) and (6).

We introduce here another objective function:

$$J_E(U, V) = \sum_{i=1}^{K} \sum_{k=1}^{N} \left\{ u_{ki} \|x_k - v_i\|^2 + \lambda^{-1} u_{ki} \log u_{ki} \right\}, \tag{12}$$

where $\lambda > 0$. This function has been considered as a variation of fuzzy K-means by a number of researchers [12], [13]. By

the alternate optimization described above using $J_E(U,V)$ instead of $J(U,V)$, and with the same constraint (11), we have the solutions (8) and (9). Thus the restricted form of the Gaussian mixture is equivalent to fuzzy K-means using $J_E(U,V)$, which is sometimes called entropy-based fuzzy K-means.

General Gaussian mixture model has the covariance matrix in addition to the mean vector. For this general case we still have similar relations in which generalized entropy-based fuzzy K-means express a generalization of the solutions derived from the EM algorithm [7], of which we omit the details for simplicity. See, e.g., [7] or [13].

B. Classification Rules in K-Means and Related Methods

We consider the method of hard K-means again and study classification rule associated with it.

1) Voronoi regions as classification rule: Let us introduce characteristic function $u_i(x)$ for cluster G_i: For each $x_k \in X$,

$$u_i(x_k) = 1, \quad x_k \in G_i, \tag{13}$$
$$u_i(x_k) = 0, \quad x_k \notin G_i. \tag{14}$$

Thus $u_i \colon X \to \{0,1\}$. For such a function defined on a discrete set X, it is difficult to observe a mathematical property. However, extending u_i to the whole space $\boldsymbol{R}^p$ is straightforward, as we will see below.

A key for this extension is the Voronoi region (see Fig. 1) which is actually referred to in vector quantization [8]. Thus the K-means is understood as the algorithm to generate Voronoi regions with centers of the cluster prototypes.

Let us denote the Voronoi regions be $W_i(V)$ $(i = 1, \ldots, K)$ with centers $V = (v_1 \ldots, v_K)$:

$$W_i(V) = \{x \in \boldsymbol{R}^p \colon \|x - v_i\| \le \|x - v_j\|, \ \forall j, \ j \ne i \ \}. \tag{15}$$

Assume that the K-means algorithm is repeated and the converged cluster prototypes are $\bar{V}$. Also suppose that the obtained clusters are $G_1, \ldots, G_K$. We then have

$$G_i = W_i(\bar{V}) \cap X, \quad i = 1, \ldots, K. \tag{16}$$

Thus the extended function $u_i \colon \boldsymbol{R}^p \to \{0,1\}$ is:

$$u_i(x) = 1, \quad x \in G_i, \tag{17}$$
$$u_i(x) = 0, \quad x \notin G_i. \tag{18}$$

They are respectively derived from (13) and (14) by replacing object symbol x_k with variable symbol x.

2) Mixture of distributions and fuzzy K-means: Let us move to the discussion of the mixture of distributions. The basic model of the mixture of distributions is as follows:

$$P(G_i|x) = \frac{p(x|G_i)P(G_i)}{\sum_{j=1}^{K} p(x|G_j)P(G_j)} \tag{19}$$

where $p(x|G_i)$ is the probability density with the condition of class G_i; $P(G_i)$ is the prior probability of class G_i. As the result $P(G_i|x)$, the posterior probability that x belongs to class G_i, is calculated. This equation is common between supervised and unsupervised classifications. Thus $P(G_i|x)$ is actually the probabilistic allocation rule, whereby the probabilistic membership of x_k to G_i is obtained by substituting $x = x_k$.

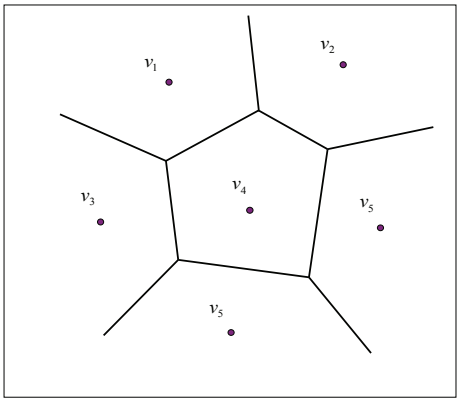

Fig. 1. A simple example of Voronoi regions with six centers on a plane.

Let us turn to fuzzy K-means and consider what we have in relation to the mixture of distributions. As in the case of the K-means, let us replace object symbol x_k by variable x in (5). We then have $u_{ki} \to U_i(x)$:

$$U_i(x) = \left[\sum_{j=1}^{K} \left(\frac{\|x - v_i\|^2}{\|x - v_j\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{20}$$

The entropy-based method has another fuzzy rule

$$U_i^E(x) = \frac{\exp(-\lambda \|x - v_i\|^2)}{\sum_{j=1}^{K} \exp(-\lambda \|x - v_j\|^2)}, \tag{21}$$

by replacing x_k in (8) by x. This equation can also be derived as $U_i^E(x) = P(G_i|x)$ using the Gaussian mixture with fixed variances (7) as above.

3) Theoretical properties: The Voronoi regions are relatively clear but the properties of the probabilistic and/or fuzzy allocation rules are not trivial, which we study in this section.

We first note that the probabilistic or fuzzy rules are closely related to Voronoi regions.

When we use probabilistic or fuzzy clustering, we often want to have hard reallocations of the objects. In such a case reallocation rule using the maximum of fuzzy memberships is natural:

$$u_i(x) = 1 \iff i = \arg \max_{1 \le j \le K} U_i(x), \tag{22}$$

where $u_i(x)$ is the final hard allocation rule and $U_i(x)$ is a fuzzy or probabilistic allocation rule.

We then have the following propositions.

Proposition 1: The characteristic function $u_i(x)$ derived from (22) defines the Voronoi region: $W_i(V)$, where V is the collection of prototypes derived from (6) or (9). The same property holds for $U_i^E(x)$.

Proposition 2: $U_i(x)$ given by (20) satisfies

$$\max_{x \in \boldsymbol{R}^p} U_i(x) = U_i(v_i) = 1, \tag{23}$$

$$\lim_{\|x\| \to \infty} U_i(x) = \frac{1}{K}. \tag{24}$$

On the other hand, $U_i^E(x)$ which has been derived from the entropy-based fuzzy K-means does not have the above properties in Proposition 2. Generally,

$$U_i^E(v_i) < 1 \tag{25}$$

and moreover

$$\sup_{x \in \boldsymbol{R}^p} U_i^E(x) > U_i^E(v_i). \tag{26}$$

The behavior of $U_i^E(x)$ as $\|x\| \to \infty$ is more complicated than that of $U_i(x)$. We need a new definition for this purpose.

Definition 1: A set of points $v_1, \ldots, v_K$ is called to have a general position when no three points of them are on a same line.

When cluster centers $v_1, \ldots, v_K$ are in a general position, no two boundaries of the Voronoi regions are parallel.

In addition, note that some Voronoi regions are bounded in the sense that a sufficiently large sphere can include the region, while others are unbounded.

We now have the next proposition.

Proposition 3: Assume that cluster prototypes $v_1, \ldots, v_K$ are in a general position. Let $V = (v_1, \ldots, v_K)$. If the Voronoi region $W_i(V)$ is bounded, then the corresponding fuzzy rule satisfies

$$\lim_{\|x\| \to \infty} U_i^E(x) = 0, \tag{27}$$

whereas if the Voronoi region $W_j(V)$ is unbounded, then the corresponding fuzzy rule satisfies

$$\lim_{\|x\| \to \infty} U_j^E(x) = 1, \tag{28}$$

provided that x moves inside the region $W_j(V)$.

4) Implications of fuzzy rules: We thus observe theoretical properties of probabilistic or fuzzy rules. We see they give Voronoi regions when clusters are made hard. This means that fundamental property of allocating objects are same for hard and fuzzy K-means. In other words, cluster boundaries are piecewise linear. If we want to have clusters with nonlinear boundaries which we call here nonlinear clusters for simplicity, we should use other methods.

There are two methods to have nonlinear cluster boundaries: one is to use additional variables [13] which we omit here to save space. Another is to use positive-definite kernel functions which we describe below.

C. Kernel-Based Clustering

The development of support vector machines [15], [14] stimulated the use of positive definite kernels [14]. Application of kernels to K-means clustering is described as follows.

1) High-dimensional mapping: Remember that X is a subset of $\boldsymbol{R}^p$. Assume H be another Euclidean space which may be finite or infinite dimensional. The norm of H is denoted by $\|\cdot\|_H$ and its inner product is $\langle \cdot, \cdot \rangle_H$. Let $\Phi \colon \boldsymbol{R}^p \to H$ a function which is called a high-dimensional mapping. We assume that $\Phi(x)$ itself is unknown but its inner product $\langle \Phi(x), \Phi(y) \rangle_H$ is represented by an explicit function:

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_H. \tag{29}$$

A typical example is the Gaussian kernel:

$$K(x, y) = \exp(-\lambda \|x - y\|^2). \tag{30}$$

Let us consider a variation of the objective function of fuzzy K-means:

$$J(U, W) = \sum_{j=1}^{K} \sum_{k=1}^{N} (u_{kj})^m \|\Phi(x_k) - w_i\|_H^2 \tag{31}$$

where $m \geq 1$: when $m > 1$, the above function is for kernel-based fuzzy K-means; when $m = 1$, it implies hard K-means. Note that $W = (w_1, \ldots, w_K)$ is the collection of prototypes in H.

The alternate optimization with respect to U and W cannot be carried out, since

$$w_i = \frac{\sum_k u_{ki} \Phi(x_k)}{\sum_k u_{ki}} \tag{32}$$

cannot be calculated due to unknown $\Phi(x_k)$.

The alternate optimization is hence replaced by the iterative calculation of U and $D(x_k, w_i) = \|\Phi(x_k) - w_i\|_H^2$:

$$\begin{aligned} D(x_k, w_i) &= \|\Phi(x_k) - w_i\|_H^2 \\ &= K(x_k, x_k) - \frac{2}{\sum_{k=1}^{K} (u_{ki})^m} \sum_{l=1}^{N} (u_{li})^m K(x_l, x_k) \\ &\quad + \frac{1}{(\sum_{k=1}^{K} (u_{ki})^m)^2} \sum_{j=1}^{N} \sum_{l=1}^{N} (u_{ji} u_{li})^m K(x_l, x_j), \end{aligned} \tag{33}$$

while the membership is given by the same equations as before:

$$u_{ki} = \left[\sum_{j=1}^{K} \left(\frac{D(x_k, w_i)}{D(x_k, w_j)} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (m > 1) \tag{34}$$

$$u_{ki} = \begin{cases} 1 & \iff i = \arg \min_{1 \leq j \leq K} D(x_k, w_i) \\ 0 & \text{otherwise} \end{cases} \quad (m = 1) \tag{35}$$

Note that not only the function $\Phi(x)$ but also the space H need not be explicitly given in this derivation.

2) Fuzzy classification rule: The fuzzy classification rule of kernel-based clustering is derived from replacing x_k by x

in (33) and (34):

$$D(x,w_i) = \|\Phi(x) - w_i\|_H^2 = K(x,x) - \frac{2}{\sum_{k=1}^K (u_{ki})^m} \sum_{l_1}^N (u_{li})^m K(x,x_l) + \frac{1}{(\sum_{k=1}^K (u_{ki})^m)^2} \sum_{j=1}^N \sum_{l_1}^N (u_{ji}u_{li})^m K(x_l,x_j), \tag{36}$$

$$U_i(x) = \left[\sum_{j=1}^K \left(\frac{D(x,w_i)}{D(x,w_j)} \right)^{\frac{1}{m-1}} \right]^{-1} \qquad (m>1) \tag{37}$$

The hard classification rules are omitted, since they are easy to derive from (35).

3) Another algorithm of kernel-based clustering: In addition to the above method, there is another algorithm. Let $\mathcal{K} = (K(x_i,x_j))$ be $N \times N$ matrix derived from a given kernel function. Note that $\mathcal{K}^{\frac{1}{2}}$ is well-defined using the square root of the positive eigenvalues, as all eigenvalues are non-negative. Assume that $\boldsymbol{e}_i$ $(1 \leq i \leq N)$ be ith elementary vector: $\boldsymbol{e}_1 = (1,0,0,\ldots)$, $\boldsymbol{e}_2 = (0,1,0,\ldots)$, and so on.

Let $y_k = \mathcal{K}^{\frac{1}{2}}\boldsymbol{e}_k$, and substitute y_k into x_k $(x_k \leftarrow y_k)$. Repeat the ordinary formula (5) and (6) until convergence. In short, we use the ordinary algorithm to $Y = (y_1,\ldots,y_N)$. Then what we have is the same as the ones by (34):

Proposition 4: Let $\hat{u}_{ki}$ be the solution by putting $x_k = y_k = \mathcal{K}^{\frac{1}{2}}\boldsymbol{e}_k$ in (5) and (6). Then we have $\hat{u}_{ki} = u_{ki}$, where u_{ki} is the solution of (34).

The proof is omitted, but readers can easily check that the solutions are the same.

4) Inductive and non-inductive clustering: Where is the difference between the methods to derive $\hat{u}_{ki}$ and u_{ki} in the last proposition?

Note that $\Phi\colon \boldsymbol{R}^p \to H$, whereas the map $x_k \mapsto \mathcal{K}^{\frac{1}{2}}\boldsymbol{e}_k$ is defined on X with the range $\boldsymbol{R}^N$. In this way, although the solutions are the same, the domains of definition and the ranges are different. This means that the former method has the fuzzy rule of classification $U_i(x)$ defined on $\boldsymbol{R}^p$, while the latter does not have a classification rule outside of X.

We have seen that a family of methods related to the K-means has classification rules defined over the whole object space. Thus when a new object occurs after clustering, each method can classify it to a certain cluster.

The last algorithm, in contrast, does not have such a classification rule. It simply generates clusters of X but a new object cannot be classified.

The former method is called here *inductive clustering*, while the latter is called *non-inductive*. This name is after Vapnik's concept of *inductive inference* and *transductive inference* in semi-supervised learning [4]. Thus methods related to the K-means are inductive, while the last algorithm is that of non-inductive clustering.

We see another class of methods and consider whether they are inductive or not.

III. Agglomerative Hierarchical Clustering

Another class of methods popular in various application fields is agglomerative hierarchical clustering which outputs dendrograms [1], [6]. This class of methods is very old but the number of users in applications are maybe as large as those of the K-means.

We assume that a distance between points $D(x,y)$ in this section is defined in some way, and it need not be an Euclidean distance or squared Euclidean distance.

The general procedure of agglomerative hierarchical clustering is in the following, where $D(G,G')$ is a distance between clusters, which will be defined after the procedure. Note also that the procedure has a given real parameter α.

1) Let initial clusters be individual objects $G_i = \{x_i\}$, $i = 1,\ldots,N$, and let the number of clusters be $K = N$.
2) Find the pair of clusters of minimum distance:
$$(G_p.G_q) = \arg \min_{1 \leq i,j \leq K} D(G_i,G_j) \tag{38}$$
3) If $D(G_p,G_q) > \alpha$, then stop the merging and output clusters $\mathcal{G}(\alpha) = \{G_h,\ldots,G_l\}$ and the clustering process as a dendrogram and stop.
4) Merge: $G_r = G_p \cup G_q$. Delete $G_p.G_q$ and add G_r to the collection $\mathcal{G}$ of clusters. Reduce the number of clusters $K = K - 1$.
5) If $K = 1$, then output the trivial cluster $\mathcal{G} = \{X\}$ and the clustering process as a dendrogram and stop.
6) Update distances $D(G_r,G_j)$ for all other clusters G_j in $\mathcal{G}$. Go to step 2).

Note that this procedure has two different kinds of outputs: an output is $\mathcal{G} = \{G_h,\ldots,G_l\}$ and another is a dendrogram (a dendrogram is undefined here, but readers can refer to standard texts of clustering like Everitt et al. [6]).

There are different methods of updating $D(G_r,G_j)$ which are called linkage methods. We consider three linkage methods below.

Single linkage: The distance is defined to be the minimum of distances between two points in the two clusters:

$$D(G,G') = \min_{x \in G, y \in G'} D(x,y) \tag{39}$$

Frequently, an updating formula which calculates $D(G_r,G_j)$ from $D(G_p,G_j)$ and $D(G_q,G_j)$ is used:

$$D(G_r,G_j) = \min\{D(G_p,G_j), D(G_q,G_j)\} \tag{40}$$

Complete linkage: The distance is defined to be the maximum of distances between two points in the two clusters which can be contrasted with the single linkage:

$$D(G,G') = \max_{x \in G, y \in G'} D(x,y) \tag{41}$$

The corresponding updating formula is as follows:

$$D(G_r,G_j) = \max\{D(G_p,G_j), D(G_q,G_j)\} \tag{42}$$

Average linkage: The distance is defined to be the average of distances between every combination of two points in the two clusters:

$$D(G,G') = \frac{1}{|G||G'|} \sum_{x \in G, y \in G'} D(x,y) \tag{43}$$

where $|G|$ is the number of elements in G. The corresponding updating formula is as follows:

$$D(G_r, G_j) = \frac{|G_p|}{|G_r|} D(G_p, G_j) + \frac{|G_q|}{|G_r|} D(G_q, G_j). \quad (44)$$

$$|G_r| = |G_p| + |G_q| \quad (45)$$

There are other two linkage methods of the centroid method and the Ward method, but we omit the detail of them.

It appears that these linkage methods do not have the inductive property, in other words, they do not have a particular classification rule for classifying another object. As we see in the next section, however, the single linkage method has a sound classification rule.

A. Inductive property of the single linkage

Unlike other linkage methods, the single linkage method is known to have a number of good theoretical properties: It is essentially equivalent to the minimum spanning tree of a weighted graph [1] and the max-min transitive closure of a fuzzy relation [11].

The single linkage method is closely related to the nearest neighbor classification rule, as shown by the definition of the distance (39).

Let us redefine the collection of clusters $\mathcal{G}$ of the output; as it has parameter α and it is applied to set X, we write the output as:

$$\mathcal{G}(\alpha; X) = \{G_h, \ldots, G_l\}. \quad (46)$$

Suppose that we have a new object y to some cluster. We find the nearest neighbor $z \in G_i$ of y and allocate y to G_i. This rule is written here as

$$\mathcal{G}(\alpha; X) \leftarrow y. \quad (47)$$

Note that

$$\mathcal{G}(\alpha; X) \leftarrow y = \{G_h, \ldots, G_i \cup \{y\}, \ldots G_l\}. \quad (48)$$

We have the following.

Proposition 5: Let

$$\alpha > \min_{x \in X} D(x, y). \quad (49)$$

We then have

$$\mathcal{G}(\alpha; X \cup \{y\}) = \mathcal{G}(\alpha; X) \leftarrow y. \quad (50)$$

In other words, clusters obtained from the single linkage with adding y to X before the algorithm starts and the allocation of y after clusters of X are obtained leads to the same result, provided that (49) holds. Note that if α is too small and (49) does not hold, then $\{y\}$ forms an isolated cluster in the left hand of (50).

This means that the single linkage clustering includes the nearest neighbor allocation rule as its essence. Hence we can say that the single linkage method has an inductive property.

Figure 2 is a complicated figure in which 20 points with numbers $1 - 20$ on the plane are objects for clustering. The segments connecting these points forms minimum spanning trees for the three clusters. They have been derived from the minimum spanning tree connecting all points by deleting three longest segments from it. Thus we have three clusters. It is known that these three clusters are obtained using the single linkage. Curved arcs are with the radius of a certain value of α, and the regions inside the arcs mean that when y is given within a region, y is allocated to the respective cluster, and (49) and (50) are satisfied. The dotted segments show unions of the Voronoi regions for the three clusters. If y is given in a region of the union of the Voronoi region, y will be allocated to the respective cluster, but (50) is not satisfied in general.

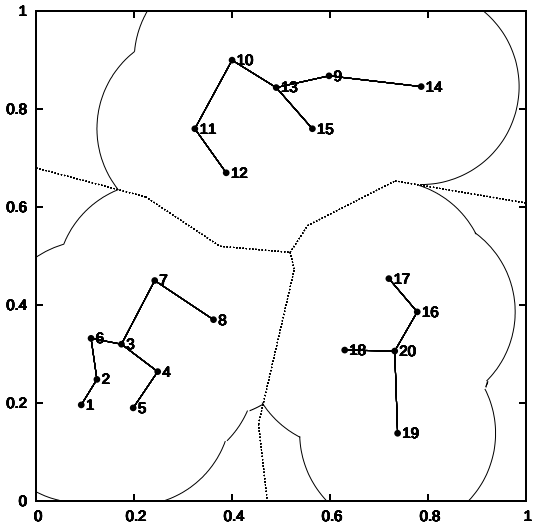

Fig. 2. Three clusters generated from 20 points by the single linkage. The regions surrounded by curves satisfy (50), while those regions outside of the curves and within the dotted lines produce isolated points that finally belong to the respective clusters.

What about the other linkage methods? We can define a furthest neighbor allocation rule related to the complete linkage method and an average allocation rule related to the average linkage method. However, a result like the one in Proposition 5 is not derived. Hence we cannot say the other two methods have the inductive property.

IV. Conclusion

Methods related to the K-means, kernel-based K-means, and agglomerative hierarchical clustering have been overviewed. The discussion is focused upon classification rules which include fuzzy rules and probabilistic rules. Methods with such rules are called inductive, while those without classification rules are called non-inductive.

We omitted complicated discussions of exceptional cases, e.g., when an object is on a prototype, or it is on the boundary of more than one Voronoi regions for simplicity, as such detailed discussion for exceptional cases will not alter the essential part of the present results.

The motivation for such clustering rules is mainly methodological: investigation of theoretical properties of rules will help deeper understanding of the method under consideration. However, such a methodological consideration will help us

when we want to choose a suitable method of clustering in a variety of applications.

Other subjects related to classification rules related to clustering discussed in this paper were omitted due to page limitation. Readers will find other methods and applications in [13].

ACKNOWLEDGMENT

The authors would like to thank the editors for inviting the author to this work. This study has partly been supported by the Grant-in-Aid for Scientific Research, JSPS, Japan, no.26330270.

REFERENCES

[1] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, 1973.
[2] D. Arthur, S. Vassilvitskii, k-means++: the advantages of careful seeding, *Proc. of SODA 2007*, pp.1027-1035.
[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, 1981.
[4] O. Chapelle, B. Schölkopf, A. Zien, eds., *Semi-Supervised Learning*, The MIT Press, 2006.
[5] J. C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. of Cybernetics*, Vol.3, pp.32-57, 1974.
[6] B. S. Everitt, S. Landau, M. Leese, D. Stahl, *Cluster Analysis, 5th ed.*, Wiley, 2011.
[7] H. Ichihashi, K. Miyagishi, K. Honda, Fuzzy *c*-means clustering with regularization by K-L information, *Proc. of 10th IEEE Intern. Conf. on Fuzzy Systems*, Vol.2, pp.924-927, 2001.
[8] T. Kohonen, *Self-Organizing Maps 2nd ed.*, Springer, 1997.
[9] J. B. MacQueen, Some methods of classification and analysis of multivariate observations, In: Proc. of 5th Berkeley Symposium on Math. Stat. and Prob., pp.281-297, 1967.
[10] G. McLachlan, D. Peel, *Finite Mixture Models*, Wiley, 2000.
[11] S. Miyamoto, *Fuzzy Sets in Information Retrieval and Cluster Analysis*, Springer, 1990.
[12] S. Miyamoto, M. Mukaidono, Fuzzy *c*-means as a regularization and maximum entropy approach, *Proc. of the 7th IFSA World Congress*, Vol.2, pp.86-92, 1997.
[13] S. Miyamoto, H. Ichihashi, K. Honda, *Algorithms for Fuzzy Clustering*, Springer, 2008.
[14] B. Schölkopf, A. J. Smola, *Learning with Kernels*, The MIT Press, 2002.
[15] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

Recognition of Upper Limb Movements for Remote Health Monitoring

Dwaipayan Biswas
Faculty of Physical Sciences and Engineering
University of Southampton
Email: db9g10@ecs.soton.ac.uk

***Abstract*—In this paper we present two methodologies based on a systematic exploration to recognize three fundamental movements of the human forearm (extension, flexion and rotation) performed during an archetypal activity of daily-living (ADL) - 'making-a-cup-of-tea' by four healthy subjects and stroke survivors. The recognition methodologies have been further implemented in hardware (ASIC/FPGA) which can be embedded on a resource constrained WSN node for real-time detection of arm movements. We propose that these techniques could be used as a clinical tool to assess rehabilitation progress in neurodegenerative pathologies such as stroke or cerebral palsy by tracking the number of times a patient performs specific arm movements (e.g. prescribed exercises) with the paretic arm throughout the day.**

***Index Terms* — Clustering, WSN, CORDIC, ASIC, FPGA, movement recognition.**

I. Introduction

With a large number of stroke survivors in the world suffering from physical and cognitive disabilities, there is a strong requirement to improve the ambulatory care model within the home settings for achieving enhanced rehabilitation at reduced costs [1]. In this research work, we look into the domain of upper limb rehabilitation by detecting specific upper limb movements during activities of daily living (ADL). The three movements investigated, along with examples of their daily occurrence, were: extension/flexion of the forearm (reach and retrieve object); rotation of the forearm about the elbow (drinking action); and rotation of the arm about the median axis (opening a door, using a key or pouring action). These movements were chosen since they comprise a significant proportion of the activities performed with our upper limb in daily life [2]. The development of wireless low-cost miniaturized, wearable sensors has enabled recording of kinematic movement in natural environments over long durations thereby aiding in unobtrusive patient monitoring using a minimal number of sensors. In view of its long term operability it is imperative to choose low complexity data processing techniques that are executed on the sensor nodes itself, to yield energy efficient solutions [3]. We implement two approaches to recognize the arm movements – (1) clustering and minimum distance classifier and (2) tracking the orientation of the inertial sensor and mapping the transition in the orientations to the investigated movements. We further implement the two arm recognition methodologies on an ASIC and FPGA, which can be embedded on a resource constrained wireless sensor node (WSN) for real-time operations. These are discussed briefly in the following section.

For this study, movements are performed by four healthy subjects and four stroke survivors, in two phases – the subjects perform multiple trials of the three enlisted movements in a controlled environment representative of a '*training*' or '*exercise*' phase. The subjects then perform repeated trials of an archetypal activity of '*making-a-cup-of-tea*', which includes multiple occurrences of extension, flexion and rotation of the forearm, representative of the '*testing*' or *ADL* phase. Data was collected from a wireless tri-axial accelerometer and gyroscope, placed on the dominant wrist during the experiments.

II. Methodology And Results

A. Clustering & Minimum Distance Classifier

The *training* data are represented by a ranked set of 30 time-domain features. Using the sequential forward selection technique, for each set of feature combinations three clusters are formed using *k*-means clustering (*k*=3) followed by 10 runs of 10-fold cross validation on the *training* data to determine the best feature combinations. The movements from the *ADL* phase are associated with each cluster label using a minimum distance classifier in a multi-dimensional feature space, comprised of the best ranked features, using the Euclidean or Mahalonobis distance as the metric [2]. The process is further illustrated in Fig. 1.

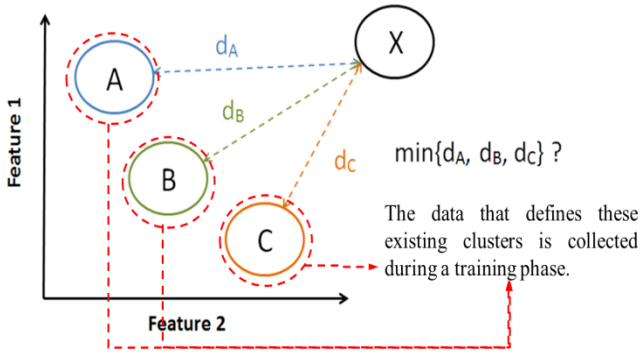

Fig. 1. Illustration of the minimum distance classifier. A, B and C represent the three movements performed in the training phase and X represents the test dataset in the respective feature space.

The three movements were detected with an overall average

accuracy of 88% using the accelerometer data and 83% using the gyroscope data across all healthy subjects and arm movement types. The average accuracy across all stroke survivors was 70% using accelerometer data and 66% using gyroscope data.

B. Sensor Orientation

The three movements are recognized by accurately mapping six predefined standard orientations of a tri-axial accelerometer located near the wrist, to the corresponding arm movements investigated. The arm movements are inferred by detecting transitions between the sensor orientations incurred during an activity. A sample transition between two pre-defined orientations, as shown in Fig. 2, demonstrates a drinking activity. Our experimental results show that the proposed methodology can independently recognize the three investigated movements with accuracies in the range of 91-99% for healthy subjects and 70%-85% for stroke patients [4].

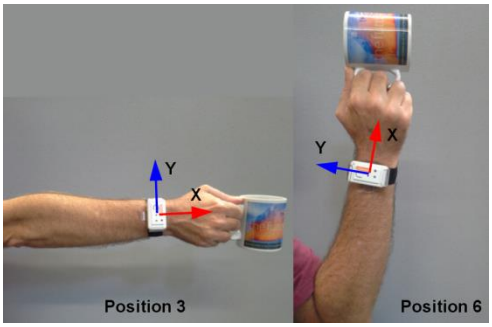

Fig. 2. Transition from Position 3 to Position 6, corresponding to a drinking type of activity, with the sensor worn on the right arm.

C. Hardware design

The clustering based approach has been implemented as an ASIC chip which can be embedded on a wireless sensor node (WSN) platform for long-time continuous detection of arm movements in real-time. The feature computation, cluster formation on the *training* data (being relatively time and memory intensive) were done in an offline mode (in software). The computation of the selected features on the *testing* data and the minimum distance computation (Euclidean) of the features from the pre-computed cluster centroids was done in hardware for real-time implementation. The arithmetic operations involved in computing the features on the *testing* data were realized using the different transcendental functions of the CoOrdinate Rotation Digital Computer (CORDIC) algorithm [5]. The design was synthesized using ST130 nm technology library at 20 MHz clock frequency to test its functionality at high speed. The synthesized design occupied an area of 347K (2-input NAND gate equivalent) with a dynamic power consumption of 25.9 mW.

The approach based on sensor orientation was coded in HDL and synthesized on the Altera DE2-115 FPGA board. For real-time operation as shown in Fig. 3, interfacing between the streaming sensor unit, host PC and the FPGA was achieved through a combination of Bluetooth, RS232 and an application software developed in C# using the .NET framework to facilitate serial port controls. The synthesized design used 1804 logic elements and recognised the performed arm movement in 41.2 µs, @50 MHz clock on the FPGA. The detected movements were displayed on a seven segment display in real-time.

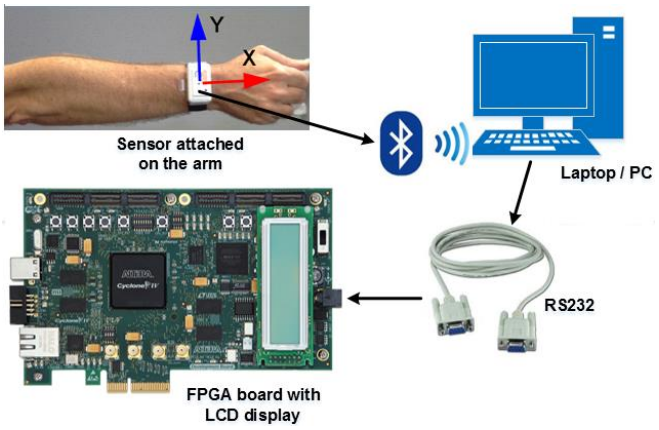

Fig. 3. Setup for the real-time recognition of arm movements using the sensor orientation approach.

III. Conclusion And Future Work

In view of the achieved results, the clustering based approach and the orientation based approach can be conveniently used for detecting arm movements in daily life. The clustering based approach can be conveniently used to include any other category of movements depending on the clinical requirements and hence has been developed as an ASIC. The implementation of the orientation approach does not use any memory element and avoids the overheads of complex data processing involved in any standard activity recognition system. Although implemented on FPGA, the salient features of the architecture makes it amenable for developing it as a low-power ASIC chip which can be embedded on a sensor platform along with other vital components such as A/D converter and a de-noising circuit to detect real-time arm movements for long-term continuous monitoring. Enumerating occurrences of these movements over time can indicate rehabilitation progress since the patient is more likely to repeat these movements as their motor functionality improves.

References

[1] J. Birns et al., "Telestroke: a concept in practice," *Age and Ageing*, vol. 39, no. 6, pp. 666–667, 2010.

[2] D. Biswas et al., "Recognition of elementary upper limb movements in an activity of daily living using data from wrist mounted accelerometers," *in Proc. IEEE Int. Conference on Health Informatics (ICHI)*, Verona, pp. 232-237, Sept. 2014.

[3] K. Maharatna, E. B. Mazomenos, J. Morgan, and S. Bonfiglio, "Towards the development of next-generation remote healthcare system: Some practical considerations," *in Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, Seoul, pp. 1–4, May. 2012.

[4] D. Biswas et al., "Recognition of elementary arm movements using orientation of a tri-axial accelerometer located near the wrist," *Physiological Measurement*, vol. 35, no. 9, pp. 1751-1768, Aug. 2014.

[5] P. K. Meher et al., "50 years of CORDIC: Algorithms, architectures, and applications," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 56, no. 9, pp. 1893–1907, 2009.

Safe and Reliable Interoperability of Medical Devices using Data-Dependent Controller Synthesis

Franziska Bathelt-Tok*
*Software Engineering for Embedded Systems
Technische Universität Berlin
Email:franziska.bathelt-tok@tu-berlin.de

In the medical sector, there is a high demand for innovative methods to ensure a vendor independent, safe, and reliable communication between heterogeneous medical devices when patients should be provided with the highest possible level of care. Due to the increasing amount of medical devices that should interact, the complexity of such systems grows continuously. To accommodate this trend and ensure a correct functionality, we aim at providing an automated and correct composition of the involved components. The concepts in the field of service composition developed for service-oriented architectures (SOAs) are highly promising to reach such an automation, while ensuring reliability and safety of the system.

The core idea of SOAs is to develop services with sophisticated functionality by composing simpler services appropriately. Services that are independently developed by different providers often cannot communicate with each other directly due to interface incompatibilities. To overcome these incompatibilities an adaption or unification of the interfaces becomes necessary. But, the internal adaption of their interfaces is not realizable due to the provision as closed-source services. To enable their interaction nevertheless, it is necessary to develop a connector that communicates with each service to be composed and, by that, overcomes the arising incompatibilities. This component, called controller, is responsible for the routing and modification of messages as well as for the compliance of behavioral requirements.

As medical devices have similar characteristics, they can be understood as services. Thus, the problem of enabling the interoperability of medical devices can be shifted to the more general problem of service composition.

However, existing approaches dealing with automated service composition, like [3], do not provide a formal data-treatment nor consider data-dependent behavior. Thus, they are insufficient or much manual effort is necessary to enable a safe interoperability of medical devices using service composition.

In our work, we address the problem of providing a controller synthesis process under consideration of a formal data-treatment to enable its application in the medical area. Our approach, allows to

1) express and ensure complex, data-dependent specifications the composed system has to fulfill,
2) detect and consider data-dependent behavior,
3) synthesize a controller automatically,
4) ensure correctness-by-construction w.r.t. data-dependent, functional and safety-critical requirements

As base for our approach, we assume that a formal representation as algebraic Petri net (APN) [5] of each service is known. This is not a restrictive assumption, because ongoing research in the field of data mining focuses on the extraction of formal models out of textual descriptions. As medical devices have to pass a certification process, they are always available. Furthermore we assume that the interface matching as well as the set of requirements are expressed by a subset of the computation tree logic (CTL). These requirements specify the interaction between the single devices and define the behavior of the composed system, respectively. Our main target is to synthesize a correct controller, so that the composed system fulfills all given requirements. For this, we have developed a three-step approach, which is shown in Figure 1.

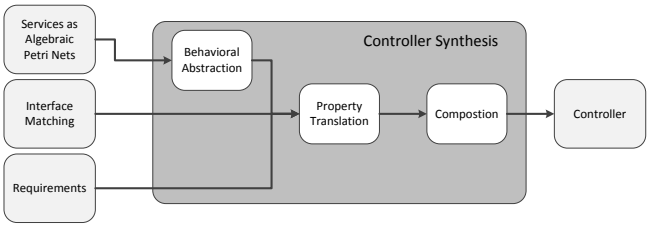

Fig. 1. Basic Idea of the Controller Synthesis Approach [1]

Based on the inputs, we firstly determine restrictions of the data domains during the *behavioral abstraction* step. In this step, the behavior of each service is reduced to the observable behavior at the interface ports. The idea is shown in Figure 2.

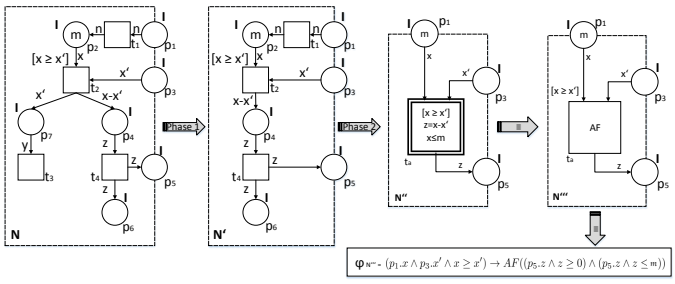

Fig. 2. Basic Idea of the Behavioral Abstraction

For each net the reduction process is done in two phases: Starting from a net representing the behavior of a service, all parts of the net that are not involved in the communication that is realized by interface places (p_1, p_3, p_5) are omitted. As result of this elimination step, we get a reduced net (N'), which is the second net in Figure 2. Based on this net, the rest of the behavior is automatically analyzed to determine restrictions of

the data domains at the (output) interface ports. For this, we summarize all internal states and transitions into one single hierarchical transition [4]. Data-dependencies occurring along this abstraction step must be extracted and stored. As result we get a CTL-formula that must be conjunct with the CTL-formulae which are representing the interface matching and the requirements. For this purpose, we have defined and proven transformation rules that describe the mapping of atomic and more complex elements of the formulae to corresponding elements of APNs. To enable this mapping, in the *property translation* step, the second step of our approach, we have extended the syntax and semantics of APNs by path operators the CTL-formulae comprise. In Figure 3 the transformation of a formula into an extended version of APN is exemplified.

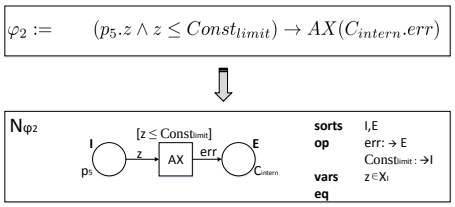

Fig. 3. Basic Idea of the Transformation from CTL-formulae into APNs

Atomic elements of the formulae are transformed into elements of APNs, e.g. $p_5.z$ is transformed into a place with label p_5 and an arc with inscription z. Propositions that comprise relational operators on the left-hand side of the implication, e.g. $z \leq Const_{limit}$, are translated into guards. Variables, like z, and constants, like err and $Const_{limit}$, become part of the algebraic specification as variables (**vars**) and operations (**op**), respectively. Additionally, CTL-specific operators, e.g. AX, are represented as label for the transition, which exemplifies the extension of APN. In this way, the set of CTL-formulae is transformed into a set of extended APNs.

Afterwards, in the *composition* step, we compose these nets to an overall Petri net in which all properties are fulfilled. This is due to the fact that the composition of the extended APNs is equivalent to the conjunction of the CTL-formulae which are represented by the corresponding nets. Figure 4 visualizes the functionality of our composition algorithm.

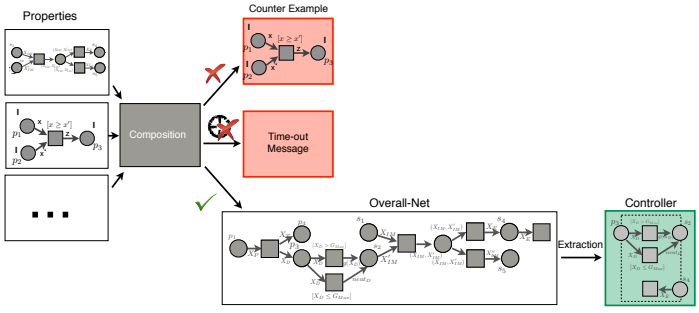

Fig. 4. Functionality of the Composition

Nevertheless, the composition is not always successful. This can have two reasons: First, it is possible that formulae are contradictory. In this case, the composition algorithm discloses a counterexample to the user to clarify which requirement is not compatible with the others. The user can decide whether this requirement is negligible and whether the controller synthesis process should be continued without this requirement. Second, there might be an infinite run time of the composition process. This is caused by undecidable problems that can occur during the analysis of the reachability graphs of the nets to be composed. To avoid this, we have defined a time interval depending on the complexity of the system. If the composition process does not terminate within this time limit, a timeout notification is sent to the user, who can change the specification and give it another try. If the composition succeeds, we get an overall net that comprises all properties, i.e., all extended APNs that represent the corresponding CTL-formulae.

In the last step of our synthesis process, the requested controller can be extracted from the resulting overall net, which includes the behavior of all services. As the purpose of the controller is to ensure the correct communication between the services, we have to extract all parts of the overall net that are responsible for that. For this, we use the interface matching to mark the interface places and separate the net structure between these places. The extended APN that comprises the separated net structures is the controller. After extracting the controller, we substitute the extended transitions, i.e., that are labeled with CTL-specific operators, by transitions an original APN consists of. The main advantage of reverting the extension is the usability of existing tools developed for original APN. The resulting controller, which is represented by an (original) APN, will be combined with the APNs that represent the service behavior and that have been used as input for our approach. Because it is correct-by-construction, it ensures that the composed system, consisting of the services and the synthesized controller, fulfills all requirements that have been specified by the customer.

As each of these three steps is fully automatic and proven as correct, we get an automated synthesis of service controllers that are correct w.r.t. data-dependent, functional and safety-critical requirements. Currently, we are evaluating this approach using a case study which has been provided by Dr. Oliver Blankenstein (Endocrinology, Charité Berlin). This case study deals with the development of an artificial pancreas, where a glucose sensor and an insulin pump have to interact. A sketch of this can be found in [2]. In future work, we aim to introduce a priority of the requirements so that a controller can be synthesized that ensures at least the most important properties. This means, properties that are absolutely necessary are combined first. Thus, a first draft of the controller is provided which can be extended by less important properties. This may reduce the risk of a time-out.

References

[1] F. Bathelt-Tok and S. Glesner. Towards the automated synthesis of data dependent service controllers. In *Service-Oriented Computing ICSOC 2013 Workshops*, Lecture Notes in Computer Science. Springer, 2014.

[2] F. Bathelt-Tok, S. Glesner, and O. Blankenstein. Data-dependent controller synthesis to enable reliable and safe interoperability of medical devices. PervasiveHealth '14, 2014.

[3] C. Gierds, A. J. Mooij, and K. Wolf. Reducing adapter synthesis to controller synthesis. *IEEE T. Services Computing*, 5(1):72–85, 2012.

[4] K. Jensen. *Coloured Petri Nets. Basic Concepts, Analysis Methods and Practical Use*, volume Volume 1. Springer-Verlag, 1997.

[5] W. Reisig. Petri nets and algebraic specifications. *Theoretical Computer Science*, 80:1–34, 1991.

Probabilistic Multi-Label Learning for Medical Data

Damien Zufferey
AiSlab Group, Institute of Information Systems, University of Applied Sciences and Arts Western Switzerland
Email: damien.zufferey@hevs.ch
DIVA Group, Department of Informatics, University of Fribourg, Switzerland

***Abstract*—We report on a probabilistic approach for the classification of chronically ill patients. We rely on multi-label learning for its ability to represent in a natural way classification problems involving coexistence of diseases. We use a public clinical database for the evaluation of our proposed algorithm. Preliminary results show the benefits of our approach.**

I. Introduction

Multi-label learning (MLL) is a growing research topic that has received, in last few years, significant contributions from machine learning community [1]. MLL differs from classical machine learning by tackling the learning problem from a different perspective which looks like natural for many problems of the real life, such as this application in the medical domain: prediction of gene function [2]. In our case, we are interested in applying MLL on clinical data for the identification of chronic diseases. This research is motivated by the problem of classifying patients affected by multiple co-morbidities to enhance decision support for physicians. We proposed an algorithm [3] based on bag of words (BoW) and supervised dimensionality reduction methods for the classification of chronically ill patients. We here extend our work following a probabilistic approach as described in the Section IV. For the evaluation of our work, the public-access intensive care unit database (MIMIC-II) [4] has been used.

II. Background

Let X be the domain of observations and L be the finite set of labels. Given a training set $T = \{(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)\}$ $(x_i \in X, Y_i \subseteq L)$ i.i.d. drawn from an unknown distribution D, the goal is to learn a multi-label classifier $h : X \rightarrow 2^L$. However, it is often more convenient to learn a real-valued scoring function of the form $f : X \times L \rightarrow \mathbb{R}$. Given an instance x_i and its associated label set Y_i, a working system will attempt to produce larger values for labels in Y_i than those not in Y_i, i.e. $f(x_i, y_1) > f(x_i, y_2)$ for any $y_1 \in Y_i$ and $y_2 \notin Y_i$. By the use of this function $f(\cdot, \cdot)$, we can obtain a multi-label classifier: $h(x_i) = \{y | f(x_i, y) > \delta,\ y \in L\}$, where δ is a threshold to infer from the training set. The function $f(\cdot, \cdot)$ can also be adapted to a ranking function $rank_f(\cdot, \cdot)$, which maps the outputs of $f(x_i, y)$ for any $y \in L$ to $\{1, 2, ..., |L|\}$ such that if $f(x_i, y_1) > f(x_i, y_2)$ then $rank_f(x_i, y_1) < rank_f(x_i, y_2)$.

III. Data Set

The MIMIC-II clinical database [4] is publicly and freely available after registration. The last release of the database contains around 33,000 patients. We choose to skip the neonates and the children in order to concentrate only on the adult population ($\geq$ 16 years old) which consists of around 24,000 patients, where we extracted a subset of 19,773 patients with chronic diseases. Regarding the restriction to the adult population, we motivate this decision by the divergence which exists between these two groups in term of medical conditions and treatment plans. The average age of the patients in the database is 67 years old. The distribution of the population is: 56% of man and 44% of women. The clinical data we consider are the laboratory tests and the items registered in the chart. By chart, we mean a logbook per patient which records the results of heterogeneous examinations, such as: fluid assessment, physiological measure, or severity score which evaluates vital functions. According to the length of the stay, a patient will make several laboratory tests and various examinations. Thus, clinical data of patients are time series. In order to attenuate the amount of missing values, we take a subset of items, from the laboratory tests and from the chart, that are present at least for 80% of the patients. We end up with 76 items from the laboratory test and from the chart.

As labels we consider 10 chronic diseases where their distributions amongst the 19,773 extracted patients are presented in the Table I. We use the coding scheme of the International Classification of Disease revision 9 (ICD-9)[1] available in the MIMIC-II database for building the 10 chronic diseases.

Label / Chronic disease	No. of patients	%
Hypertensive disease	12,309	62.3%
Fluid electrolyte disease	6,177	31.2%
Diabetes mellitus	6,056	30.6%
Lipoid metabolism disease	5,965	30.2%
Kidney disease	5,828	29.5%
COPD	4,253	21.5%
Thyroid disease	2,246	11.4%
Hypotension	1,962	9.9%
Liver disease	1,088	5.5%
Thrombosis	931	4.7%

TABLE I. Distribution of labels / chronic diseases in the 19,773 extracted patients of the MIMIC-II database.

IV. Method

A. Feature extraction

Laboratory events and chart events of each patient are summarized into one feature vector. Due to the heterogeneity and the different frequencies of the selected medical data, we propose the following approach for the feature extraction according to the type of the measured values:

[1] http://www.who.int/classifications/icd

1) Numerical values: consist of measured values such as blood pressure, creatinine and temperature. When they appear one time, such as the height at the patient admission, they are taken in the feature vector as they are. When they appear several times, the following summary features are computed: mean, median, standard deviation and range.

2) Categorical values: consist of observed values such as cardiovascular function assessment score and urine color. For a patient which did several times a particular examination where results are discrete values which can be divided into mutually exclusive classes, we can represent this information as an histogram. Then, the relative frequency of each category of the histogram is used as feature. There is also the case where only one observation exists for each patient, such as the gender at the patient admission, in that case, we encode as feature the value in a binary variable.

B. Model

We propose a generative model for MLL using Gaussian Mixture Model (GMM) [5] as the based classifier, called ML-GMM hereafter. We use a combination of Label Power-set (LP) and Binary Relevance (BR) as the transformation methods [1]. In particular, we apply the LP method, which considers all possible combinations between labels to transform the MLL problem to a multi-class formulation which can be naturally solved using GMMs. To handle the intractability of LP, we consider, according to a predefined constant n, only combinations of labels that have at least n observations in the training set. In contrast, combinations of labels that have less than n observations are grouped together to form the class *"other"*. Then, in the case of the class *"other"* is relevant, we apply the BR method, which considers each label independently as relevant or not, to transform the MLL problem to a set of binary problems, according to the number of labels. The manner how ML-GMM combines LP and BR allows handling the MLL problem efficiently and at the same time preserving dependency information between labels.

V. Experiment

In the classical learning approach of multiclass problems, the evaluation is done through common metrics such as accuracy, precision, and recall. In multi-label problems, the evaluation is much more complicated and needs extended evaluation metrics. One of the commonly used evaluation metrics is the Hamming loss [6], which is described below.

Let a testing set $S = \{(x_1, Y_1), (x_2, Y_2), ..., (x_m, Y_m)\}$.

Hamming loss evaluates how many times an observation-label pair is misclassified. The score lies between 0 and 1, where 0 corresponds to the best result:

$$hloss_S(h) = \frac{1}{m}\sum_{i=1}^{m}\frac{|h(x_i)\triangle Y_i|}{|L|}, \quad (1)$$

where $\triangle$ represents the symmetric difference.

To evaluate our proposed algorithm (ML-GMM), we divided the MIMIC-II dataset containing 19,773 patients into 3 subsets of 6,591 patients. We used the first two subsets (training and validation) in a grid-search for finding optimal parameters for our algorithm. Finally, using the third subset (testing), we computed the reported results, as presented in the Figure 1. We can see that the Hamming loss is reducing when we allow the algorithm to handle a larger combination of classes using the LP method. The best performance is achieved when considering 14 classes in LP method. Note that the algorithm is purely BR when the number of classes is 0.

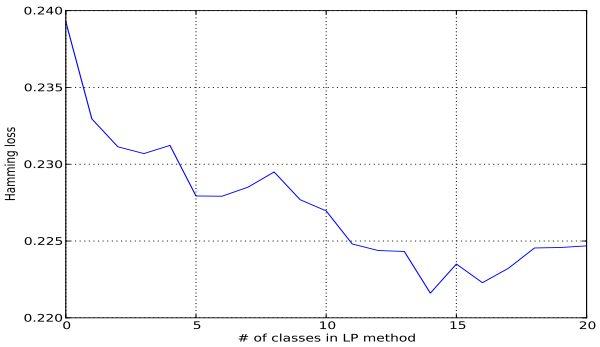

Fig. 1. Results in term of Hamming loss of our ML-GMM algorithm on the MIMIC-II database.

VI. Conclusion

In this abstract, we proposed an approach for a MLL-based algorithm for the classification of chronically ill patients. Our solution elegantly combines the LP method for its ability to consider correlations between labels, and the BR method for its ability to scale well with a large number of labels. For future work, we will conduct additional experiments to evaluate our algorithm by considering additional evaluation metrics and to compare with existing state-of-the-art multi-label algorithms.

References

[1] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084 – 3104, 2012, best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312001203

[2] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya, "Hierarchical multi-label prediction of gene function," *Bioinformatics*, vol. 22, no. 7, pp. 830–836, 2006. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/22/7/830.abstract

[3] S. Bromuri, D. Zufferey, J. Hennebert, and M. Schumacher, "Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms," *Journal of Biomedical Informatics*, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046414001270

[4] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database," *Critical Care Medicine*, vol. 39, pp. 952–960, May 2011.

[5] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, vol. 1.

[6] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038 – 2048, 2007.

Leefplezier: Personalized Well-being

Frank Blaauw
Distributed Systems Group
Johann Bernoulli Institute
University of Groningen
f.j.blaauw@umcg.nl

Lian van der Krieke
Psychiatry (ICPE)
UMC Groningen
University of Groningen
j.a.j.van.der.krieke@umcg.nl

Peter de Jonge
Psychiatry (ICPE)
UMC Groningen
University of Groningen
peter.de.jonge@umcg.nl

Marco Aiello
Distributed Systems Group
Johann Bernoulli Institute
University of Groningen
m.aiello@rug.nl

***Abstract*—Each person is different and should be treated as such. Comparing personal data to group averages can give basic ideas about personal conditions, but does not suffice for providing 'true' personalized feedback. In health psychology, a paradigm shift is taking place from a general population approach towards a more person-centered one. Instead of comparing a person with population averages, the focus shifts towards comparing people with themselves over time. The 'Leefplezier' project elaborates on this focus shift, by helping to sustain or improve the well-being of elderly people. The participating elderly people are asked to keep track of various psychological factors for a period of time, by means of repetitive questionnaires via a mobile phone application. At the end of this period, feedback is automatically generated, based on the resulting time series dataset and by means of automated vector autoregression.**

I. Introduction

In the Netherlands, a large care group organization (Espria[1]) has initiated a project for assessing, improving and enhancing well-being of its elderly members and elderly people in general, by researching various factors influencing well-being. Limitations exist in the traditional group-based (*nomothetic*) method of conducting research, with regards to the flexibility and generalizability of such research. The alternative is to *personalize* research, by using a person-centered (*idiographic*) approach [1]. With the Leefplezier project, we aim to apply an idiographic approach and to assess and provide feedback for each individual in isolation, instead of focussing on the group of elderly as a whole. To do so, researchers from the *University Medical Center Groningen* (UMCG) and *University of Groningen* (RUG) cooperate with Espria to develop a system for measurement and analysis that provides meaningful and personal feedback. Since we intend to measure fluctuations of certain psychological, physiological and other factors over time, each participant should be measured multiple times per day, for several days in a row. The goal of each measurement is to quantify various psychological or physiological factors, which might influence well-being, together with the subjective well-being at that moment in time. The resulting temporal data contains well-being and factors that might influence it. This allows us to determine whether and how much certain factors influence fluctuations of well-being. In this project, we focus on novel and effective ways to analyze the collected data, and to provide the participants with insightful, personalized feedback about their well-being. To determine causal relations between well-being and its potential indicators, we generate feedback with automated vector autoregression (VAR), a statistical technique from the domain of time series analysis in econometrics [2]. This feedback provides insight into the factors that influence well-being and might help to sustain or enhance well-being.

[1]Website: http://www.espria.nl

II. Methodology

The Leefplezier project is divided into two main phases: (i) gathering of general information, and (ii) capturing and analyzing experiences on a daily basis. In the first phase, we gather information about the target population. We develop a web application that allows participants to fill out a series of questionnaires that help us to derive general knowledge about and insight into the well-being of the participants from a cross sectional point of view, that is, from the point of view of the population sample. The selected questionnaires measure various psychological constructs, such as well-being, mood, anxiety, personality, depression, stress and various general demographic factors. For each questionnaire, the application provides feedback, including a comparison of the individual results to the group averages. This comparison aligns with the more traditional, nomothetic approach, viz., generating a population average and generalizing individuals to this average.

In the second phase, we capture and analyze personal experiences over time. Instead of considering the population as a whole, we analyze each individual participant in isolation. We use a technique for repeatedly conducting questionnaires with a method known as *Ecological Momentary Assessment* (EMA), a way to capture experiences on a daily basis as questionnaire data, known as a diary study [3]. These data enable for the creation of personalized statistical models representing the (causal) relationships between the participant's wellbeing and influencing factors, using an automated VAR modeling application known as *Autovar* [4], [5]. VAR models show, for one participant in isolation, how certain measured factors influence each other over time. For each participant a VAR model is generated based on their measurements. The strongest effects in the model are presented as feedback to the participant (e.g. 'An increase in factor α precedes a decrease in factor β'). To facilitate these measurements and provide such feedback we design and develop a mobile phone application.

III. Initial Evaluation

The scientific contribution of this project can be summarized as finding novel ways to effectively and automatically analyze time-series data gathered using repetitive questionnaires, and to provide insightful and interactive feedback to the participants. Besides the automated feedback, intensive momentary assessments are only occasionally carried out on

a large scale, especially in combination with elderly people. To the best of our knowledge, our application will be the first to provide automated personalized feedback based on the application of VAR models to EMA data of elderly persons.

In December 2013, we started a pilot study for the Leefplezier project called *HowNutsAreTheDutch* (Dutch: HoeGekIsNL[2]) [6]. HowNutsAreTheDutch focusses on all people aged 18 or over in the Netherlands. The project design for HowNutsAreTheDutch is comparable to the Leefplezier project; it was built for the same two research phases and provides comparable functionality, albeit web-based (i.e., no actual mobile application was developed). This pilot study currently has 12,690 participants that have completed 61,773 questionnaires in the cross-sectional analysis (the first phase, recorded on November 25th, 2014). At the end of May 2014, we launched a diary study as the second phase of HowNutsAreTheDutch (the daily questionnaires), to which approximately 600 participants have subscribed (recorded on November 25th, 2014). Analysis of the data from phase one has yielded various interesting results. For instance, we measured that approximately 10% of the population suffers from severe depressive symptoms, but at the same time, 75% rate their happiness 6 out of 10, or higher (in fact, 25% rate it 8 out of 10, or higher). Furthermore, we observed large individual variation in positive and negative affect measured with the Dutch version of the Positive and Negative Affect Schedule (PANAS) [7]. Analysis on a subset of the HowNutsAreTheDutch participants ($n = 6895$) shows that participants with identical levels of negative affect varied substantially in their positive affect. The distribution of positive and negative affect of these participants is depicted in Figure 1. The gray level signifies the density of the participant distribution, ranging from light-gray (sparse) to black (dense). The dots in the image depict the measured combinations of positive and negative affect, the lines depict the average of positive (horizontal line) and negative affect (vertical line). This supports our approach in which we focus on individuals, rather than averages, because it would be unrealistic to come up with one generalized measure for all participants.

The first phase of the Leefplezier project started in May 2014. Currently 677 people have registered for the project (recorded on November 25th, 2014). At the end of January 2015, the second phase of the project will be launched.

IV. Future work and Discussion

The Leefplezier project is still in its infancy. Although phase one has successfully been completed, the most important phase (phase 2) is yet to be started. During the pilot study, progress was made with regards to the data analysis, but it shall be interesting to see whether the general analysis tools will also work for our elderly population. The pilot study showed that nomothetic research would not be sufficient, because of the large variance between participants. Applying idiographic research would in this case be more suitable.

The next step is to refine the process and mobile application in such a way that most people will understand and feel comfortable working with the application. This is covered by a usability analysis. Furthermore, we will research and apply other techniques to analyze the data and perform simulations based on the established models. For instance, when a personal model has been determined, we could allow participants to virtually influence the factors in the model and simulate how these changes propagate through the model to the other factors. For example, if one would increase factor α at time step $t = 0$, what would happen to the other factors in the model at time step $t = 1..n$?

[2]Website: https://www.hoegekis.nl

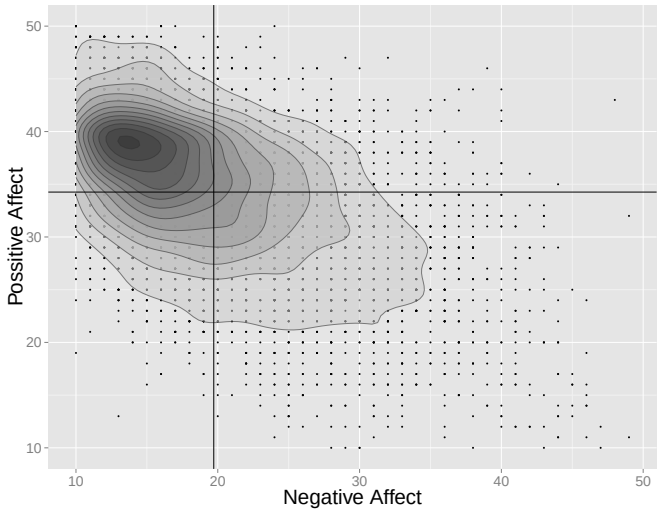

Fig. 1. Distribution of positive and negative affect in HowNutsAreTheDutch participants ($n = 6895$), measured using the PANAS questionnaire [7].

Acknowledgments

The Leefplezier project is funded by an unrestricted grant from Espria, a VICI grant (no: 91812607) from the Netherlands organization for scientific research (NWO-ZonMW) and by the University Medical Center Groningen Research Award 2013, both received by Peter de Jonge.

References

[1] D. H. Barlow and M. K. Nock, "Why can't we be more idiographic in our research?" *Perspectives on Psychological Science*, vol. 4, no. 1, pp. 19–21, 2009.

[2] C. A. Sims, "Macroeconomics and reality," *Modelling Economic Series.Clarendon Press, Oxford*, vol. 48, no. 1, pp. 1–48, 1980.

[3] S. Shiffman and A. A. Stone, "Ecological momentary assessment: A new tool for behavioral medicine research," *Technology and methods in behavioral medicine*, pp. 117–131, 1998.

[4] J. A. J. van der Krieke, "Patients in the driver's seat - A role for e-mental health?" Ph.D. dissertation, University of Groningen, 2014. [Online]. Available: http://irs.ub.rug.nl/ppn/371485681

[5] A. Emerencia, "Computing a Second Opinion: Automated Reasoning and Statistical Inference applied to Medical Data," Ph.D. dissertation, University of Groningen, 2014. [Online]. Available: http://irs.ub.rug.nl/ppn/376463287

[6] F. Blaauw, L. van der Krieke, E. Bos, A. Emerencia, B. F. Jeronimus, M. Schenk, S. de Vos, R. Wanders, K. Wardenaar, J. T. W. Wigman, M. Aiello, and P. de Jonge, "HowNutsAreTheDutch: Personalized feedback on a national scale," in *Expanding the Boundaries of Health Informatics Using AI (HIAI'14): Making Personalized and Participatory Medicine A Reality*, 2014.

[7] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

Analysis of Medical Treatments Using Data Mining Techniques

Xin Xiao, Silvia Chiusano
Dipartimento di Automatica e Informatica, Politecnico di Torino - Torino, Italy Email:
{xin.xiao,silvia.chiusano}@polito.it

***Abstract*—Since in health care systems the amount of data is continuously growing, data mining techniques can be applied to analyse these large collections and gain interesting insights. However, some critical issues should be properly addressed. For example, data collections on patient treatments are usually characterized by an inherent sparseness and variable distribution, due to the large variety of possible treatments performed by patients affected by a given disease. To effectively extract interesting knowledge from such collections, we present a framework coupling a *clustering* and a *classification* algorithm. The clustering approach is named *multiple-level* because we apply the clustering algorithm in a multiple-level fashion, by focusing on different dataset portions and locally identifying groups of patients with similar profile and examination history. The classification algorithm is used to characterize the discovered clusters. This paper also describes future research issues and possible developments of the proposed framework.**

***Index Terms*—Data mining, cluster analysis, classification analysis, medical records, patient examination history.**

I. Introduction

Nowadays, large amount of medical data, storing the patient medical history, is collected during health care. The analysis of these medical data collections is a challenging task for health care systems since a huge amount of interesting knowledge can be automatically mined to effectively support both physicians and health care organizations.

Data mining techniques [1], which focus on studying effective and efficient algorithms to transform large amounts of data into useful knowledge, have been widely exploited on medical data by analyzing different pathologies or different aspects of the same disease. For example, previous studies addressed food analysis [2] and investigated risk factors associated with diabetes and pre-diabetes [3], while current issues in medicine knowledge discovery are discussed in [4].

Analysing real world health care data collections may impose new challenges. These collections can have *large volume* and *high dimensionality* due to the large cardinality of patient records and the variety of medical treatments usually adopted for a given pathology. In addition, they are usually characterized by a *variable data distribution* and *inherent sparseness*. Consequently, innovative data mining approaches are needed to efficiently gain interesting insights from such collections.

In this paper we present a framework to discover, in a patient data collection with a variable distribution, cohesive and well-separated groups of patients with a similar profile (i.e., patient age and gender) and examination history (given by the set of examinations performed by patients). The framework couples a clustering approach (named "multiple-level clustering") for cluster set computation, and a classification algorithm used to both characterize the cluster content and measure the effectiveness of the clustering process. Health care organizations can exploit the discovered knowledge for example to check the coherence between the adopted treatments and existing medical guidelines for a given disease, as well as enrich the existing guidelines or assess new ones.

The paper is organized as follows. The framework is presented in Section II, while future research issues and possible developments of the framework are discussed in Section III.

II. The proposed data analysis framework

The presented framework is depicted in Figure 1 and summarized in the following.

To deal with the inherent sparseness and variable distribution of patient data collections, a density-based *multiple-level clustering* approach is adopted in the framework. We named the approach "multiple-level" because it performs multiple runs over the considered data collection. This strategy aims at progressively partitioning the initial data collection into (quite) homogeneous subsets, thus easing the computation of cohesive clusters on each of them. Specifically, at each iteration a different dataset portion is analyzed, and clusters are locally identified on it.

A novel distance measure has been defined to cluster patients according to the three aspects characterizing them, i.e., patient age, gender and examination history. For the data representation, the patient examination history tailored to the Vector Space Model (VSM) is used, and the examination frequency is weighted using the TF-IDF (Term Frequency (TF) - Inverse Document Frequency (IDF)) score [1]. TF-IDF has been used in text mining to analyse document collections, with the aim of weighting the relevance of words in documents. In our context, we used this approach to weight the relevance of examinations in the patient examination history, highlighting peculiar examinations for each patient. The examination frequency can vary significantly from standard tests to specific examinations used to diagnose disease complications. TF-IDF allows focusing on examinations specific for each patient and discarding examinations done by most patients.

The discovered cluster set is evaluated through the Silhouette index [1], and with the support of domain experts to describe the cluster content from a medical perspective. Specifically, a class label is assigned to each cluster.

Starting from the labeled cluster set, a *classification model* is created both to characterize the content of clusters and measure the effectiveness of the clustering process. This model can be used to automatically assign a new patient to a given class based on her/his profile and examination history. In addition, when the adopted classification algorithm provides a readable model (e.g., decision trees [1]), this model can give useful insights to domain experts on some peculiar properties characterizing patients in each class (in terms of gender, age and undergone examinations).

As a first attempt, the proposed framework has been used in [5] to analyse a real dataset of diabetic patients provided by an Italian Local Health Center. The DBSCAN algorithm [1] has been adopted for the multiple-level clustering approach, while decision trees to compute the classification model.

Results showed that the multiple-level clustering approach progressively discovers clusters containing patients with increasing disease severity. For example, clusters computed in the first iteration of the approach contain patients mainly undergoing routine tests to monitor diabetes conditions. Instead, clusters computed in the next iterations contain patients tested using an increasing number of examinations to diagnose several diabetes complications. The cluster set is characterized by good Silhouette values, and the classification model computed from it is very accurate (above 90% of accuracy), with good precision and recall values for most class labels.

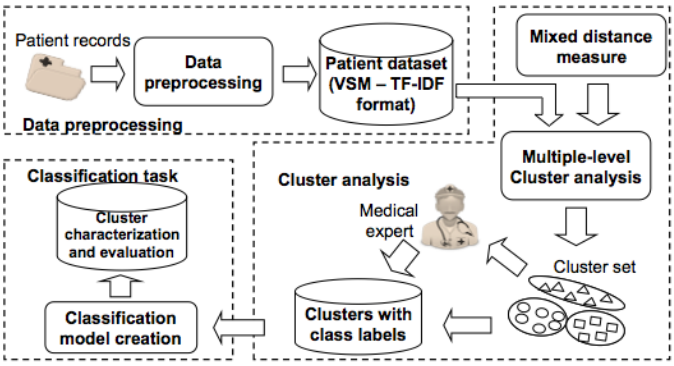

Fig. 1. The proposed framework

III. Future work

This section describes some research directions for the analysis of medical data, and specifically possible developments of the framework presented in Section II.

(i) Evaluating alternative clustering and classification algorithms. Different clustering and classification algorithms can be selected for integration in the proposed framework. Based on the target application scenario, the proper algorithms can be adopted by considering different issues as final number of clusters and average cluster size, classification model accuracy and readability, and computational cost.

(ii) Analysing additional information on patient treatments. Besides patient examination history, additional aspects of the medical treatments such as prescribed drugs can also be considered. This analysis can help discovering correlations between prescribed drugs and disease complications, as well as detecting and preventing drug misusing. The information on prescribed drugs can be used to characterize patient clusters computed using the framework presented in Section II, or to drive the clustering process for discovering groups of patients with similar examination histories and drug therapies.

(iii) Using data taxonomy in the data analysis process. Taxonomies can be used to generalize examinations and drugs into their corresponding categories. They can be used to drive the process of clustering patient data, or to reduce the data dimensionality problem by considering medical data described at different abstraction levels.

(iv) Considering the temporal dimension in the patient examination history. Since medical data includes frequently occurring temporal patterns in patient records, the temporal dimension in medical data analysis can be a critical issue. Temporal mining approaches can be used to discover clusters of patients who have similar temporal relations among the examinations. The temporal analysis can help to identify examination pathways commonly followed by patients, and potentially check and improve predefined guidelines.

References

[1] Pang-Ning T. and Steinbach M. and Kumar V., *Introduction to Data Mining*. Addison-Wesley, 2006.

[2] M. Phanich, P. Pholkul, and S. Phimoltares, "Food recommendation system using clustering analysis for diabetic patients," in *IEEE International Conference on Information Science and Applications (ICISA)*, 2010, pp. 1–8.

[3] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung Journal of Medical Sciences*, no. 0, 2012.

[4] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," *Expert Systems with Applications*, vol. 41, pp. 4434–4463, 2014.

[5] G. Bruno, T. Cerquitelli, S. Chiusano, and X. Xiao, "A clustering-based approach to analyse examinations for diabetic patients," in *IEEE International Conference on Healthcare Informatics*, 2014, pp. 45–50.

RELATED CONFERENCES, CALL FOR PAPERS/PARTICIPANTS

TCII Sponsored Conferences

WI 2015
The 2015 IEEE/WIC/ACM International Conference on Web Intelligence
Singapore
December 6-9, 2015
http://wi-iat15.ntulily.org/wi/

The 2015 IEEE/WIC/ACM International Conference on Web Intelligence (WI'15) and the 2015 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'15) will be held in Singapore December 6-9, 2015, hosted by the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY) at Nanyang Technological University (NTU) and the Living Analytics Research Centre at Singapore Management University (SMU). The two co-located conferences are sponsored by IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), Web Intelligence Consortium (WIC), ACM-SIGART, and the Memetic Computing Society.

Following the great successes of WI-IAT'01 held in Maebashi, Japan, WI-IAT'03 in Halifax, Canada, WI-IAT'04 in Beijing, China, WI-IAT'05 in Compiegne, France, WI-IAT'06 in Hong Kong, WI-IAT'07 in Silicon-Valley, USA, WI-IAT'08 in Sydney, Australia, WI-IAT'09 in Milano, Italy, WI-IAT'10 in Toronto, Canada, WI-IAT'11 in Lyon, France, WI-IAT'12 in Macau, China, WI-IAT'13 in Atlanta, USA, and WI-IAT'14 in Warsaw, Poland, the WI-IAT'15 will provide a global forum for scientists, engineers and educators to present the latest WI-IAT technologies, discuss how to develop future intelligent systems for complex applications. We will celebrate the common theme of Big Data in Global Brain and Social Networks. We will also encourage industry sessions representing both local and international companies developing solutions and running projects related to those areas.

WI-IAT'15 will have various workshops, WI-IAT technical sessions, tutorials and panels. WI-IAT'15 will have keynotes, a social reception together with the poster session and industry demo, and a banquet. Attendees only need to register once to attend all technical events at WI- IAT'15.

Web Intelligence focuses on scientific research and applications by jointly using Artificial Intelligence (AI) (e.g., knowledge representation, planning, knowledge discovery and data mining, intelligent agents, and social network intelligence) and advanced Information Technology (IT) (e.g., wireless networks, ubiquitous devices, social networks, semantic Web, wisdom Web, and data/knowledge grids) for the next generation of Web-empowered products, systems, services, and activities.

IAT 2015
The 2015 IEEE/WIC/ACM International Conference on Intelligent Agent Technology
Singapore
December 6-9, 2015
http://wi-iat15.ntulily.org/iat/

The 2015 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'15) will be held in Singapore on December 6-9, 2015, hosted by the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY) at Nanyang Technological University (NTU) and the Living Analytics Research Centre at Singapore Management University (SMU). Co-located with the 2015 IEEE/WIC/ACM International Conference on Web Intelligence (WI'15), IAT'15 is sponsored by IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), Web Intelligence Consortium (WIC), ACM-SIGAI, and Memetic Computing Society.

IAT'15 will feature special thematic workshops, technical sessions, tutorials and panels. Conference programme will further include a welcome reception, keynotes, demos, poster sessions, and a banquet. Attendees only need one registration to attend all technical events.

IAT 2015 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. Intelligent Agent Technology explores advanced intelligent systems and their broad applications in computer science and engineering, big data mining, biomedical informatics, health informatics, social networks, education, robotics, security, etc.

ICDM 2015
The Twenty-Second IEEE International Conference on Data Mining
Atlantic City, NJ, USA
November 14-17, 2015
http://icdm2015.stonybrook.edu/

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. 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.

ISMIS 2015
The 22nd International Symposium on Methodologies for Intelligent Systems
Lyon, France
October 21-23, 2015
http://liris.cnrs.fr/ismis15/

ISMIS is an established and prestigious conference for exchanging the latest research results in building intelligent systems. Held twice every three years, the conference provides a medium for exchanging scientific research and technological achievements accomplished by the international community.

The scope of ISMIS is intended to represent a wide range of topics on applying Artificial Intelligence techniques to areas as diverse as decision support, automated deduction, reasoning, knowledge based systems, machine learning, computer vision, robotics, planning, databases, information retrieval, etc. The focus is on research in intelligent systems. The conference addresses issues involving solutions to problems that are complex to be solved through conventional approaches and that require the simulation of intelligent thought processes, heuristics and applications of knowledge. The integration of these multiple approaches in solving complex problems is of particular importance. ISMIS provides a forum and a means for exchanging information for those interested purely in theory, those interested primarily in implementation, and those interested in specific research and industrial applications.

ICHI 2015
The IEEE International Conference on Healthcare Informatics
Dallas, USA
October 21-23, 2015
http://multimedia.utdallas.edu/ichi2015/

ICHI 2015 is the premier community forum concerned with the application of computer science principles, information science principles, information technology, and communication technology to address problems in healthcare, public health, and everyday wellness. The conference highlights the most novel technical contributions in computing-oriented health informatics and the related social and ethical implications. ICHI 2015 will be held in Dallas, Texas USA on October 21-23, 2015. It will be a forum for demo and paper contributions from researchers, practitioners, developers, and users to explore and disseminate cutting-edge ideas and results, and to exchange techniques, tools, and experiences.

Related Conferences

AAMAS 2015
The 14th International Conference on Autonomous Agents and Multi-Agent Systems
Istanbul, Turkey
May 5-8, 2015
http://www.aamas2015.com/en/

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: the International Conference on Multi-Agent Systems (ICMAS); the International Workshop on Agent Theories, Architectures, and Languages (ATAL); and the 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.

AAAI 2015
The 29th AAAI Conference on Artificial Intelligence
Austin Texas, USA
January 25-30, 2015
http://www.aaai.org/Conferences/AAAI/aaai15

The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15) will be held January 25–30, 2015 at the Hyatt Regency in Austin, Texas, USA. The purpose of this conference is to promote research in artificial intelligence (AI) and scientific exchange among AI researchers, practitioners, scientists, and engineers in affiliated disciplines. AAAI'15 will have a diverse technical track, student abstracts, poster sessions, invited speakers, tutorials, workshops, and exhibit/competition programs, all selected according to the highest reviewing standards. AAAI'15 welcomes submissions on mainstream AI topics as well as novel crosscutting work in related areas.

SDM 2015
The Fifteenth SIAM International Conference on Data Mining
Vancouver, British Columbia, Canada
Apr 30 - May 2, 2015
http://www.siam.org/meetings/sdm15/

Data mining is the computational process for discovering valuable knowledge from data. It has enormous application in numerous fields, including science, engineering, healthcare, business, and medicine. Typical datasets in these fields are large, complex, and often noisy. Extracting knowledge from these datasets requires the use of sophisticated, high-performance, and principled analysis techniques and algorithms, which are based on sound theoretical and statistical foundations. These techniques in turn require implementations on high performance computational infrastructure that are carefully tuned for performance. Powerful visualization technologies along with effective user interfaces are also essential to make data mining tools appealing to researchers, analysts, and application developers from different disciplines.

The SDM 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 presentations and tutorials (included with conference registration). A set of focused workshops is 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.

IJCAI 2015
The Twenty-Third International Joint Conference on Artificial Intelligence
Buenos Aires, Argentina
July 25-31, 2015
http://ijcai-15.org/

IJCAI is the International Joint Conference on Artificial Intelligence, the main international

gathering of researchers in AI. Held biennially in odd-numbered years since 1969, IJCAI is sponsored jointly by IJCAI and the national AI societie(s) of the host nation(s). IJCAI is a not-for-profit scientific and educational organization incorporated in California. Its major objective is dissemination of information and cutting-edge research on Artificial Intelligence through its Conferences, Proceedings and other educational materials.

IJCAI Board of Trustees in its historical meeting held on Thursday, July 21, 2011 in Barcelona, Catalonia, Spain, decided that IJCAI conferences will be held annually in the future. Following the success of IJCAI-13 held in Beijing, China, IJCAI'15 will be held in Buenos Aires, Argentina. IJCAI'16 will be held in New York, USA and IJCAI'17 in Melbourne, Australia.

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

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