International Proceedings of Economics Development and Research
IPEDR vol.85 (2015) © (2015) IACSIT Press, Singapore

A Categorization Scheme to Data Mining: Practices Used in Public Health

David Nadler Prata[1], Mayara Kaynne Fragoso Cabral [1,2 +], Thatiane de Oliveira Rosa [1,2], Leandro Guimar ães Garcia[1], Patrick Letouze[1]

[1] Post-Graduate in Computational Modelling Systems/ Federal University of Tocantins, Palmas, Brazil

[2] Department of Information and Communication, Federal Institute of Education, Science and Technology of Tocantins – Brazil)

Abstract. Data mining researchers aim to find patterns and new intrinsic knowledge to a particular inquiry context. The health investigation field has been an encouraging promise to apply mining methods and algorithms hopeful to elicit knowledge at least, e.g., from data available in databases, questionnaires, and social networks. Hence, the systematic review presented in this study meant to consistently map the standing research on public health data mining, through the identification of good practices. The efforts result in a categorization of data mining techniques currently used in modern public health. The study associates the core purposes of the enclosed standard procedures and the chosen algorithms applied for data mining objectives.

Keywords: public health, knowledge discovery, data mining, algorithms and methods

1. Introduction

In order to learn new knowledge from large volumes of data, we need to organize, process, and analyze them. For this purpose, we typically customize a process known as KDD (Knowledge Discovery in Database), which is organized [1] into five stages: selection; preprocessing; transformation; data mining; and interpretation / evaluation. The process is shown in Figure 1.

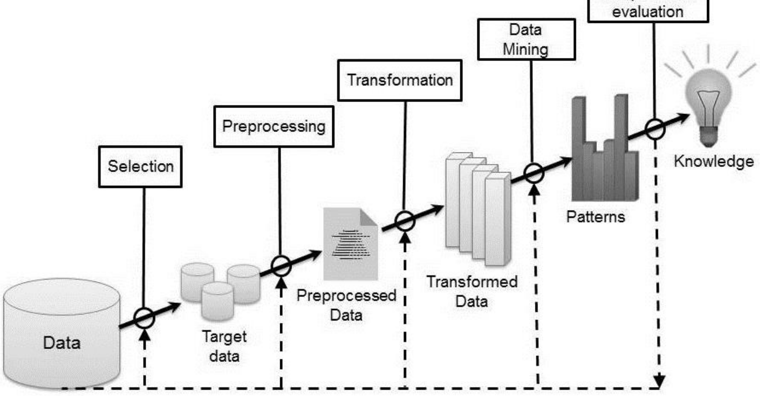

Fig. 1: Comprising steps of the KDD [1] process.

[+] Corresponding author. Tel.: +(63) 8422-9293.
E-mail address: mayarakf@ifto.edu.br.

The focus of this study is to better understand the purposes and practices applied to Data Mining (DM) in Public Health, which is the fourth step of the KDD process. Meanwhile, the DM head stages of KDD should be in place to guide this work.

The most important stage for this work is the selection, which is to understand the application domain by the identification of questions that must be answered. The main purpose is to identify the KDD's objectives for the process carrying out. After understanding these objectives, we can select a data set that must fit the knowledge discovery's commitments [1].

The fourth stage (data mining) applies DM methods in accordance with the objectives set during the first stage of process [1]. In this context, DM aims to analyze large volumes of data, identifying relationships and valid standards [1, 2, 3]. Furthermore, DM searches for new, meaningful and understandable knowledge data to the field [1,2].

According to [1], there are two goals for knowledge discovery through data mining: descriptive and predictive. The first aims to describe existing patterns among the data, allowing humans to interpret and cluster them easily [1]. The second performs interventions to create categorization models that will perform forecast of new data. The importance of using one or the other varies with the defined goals and problems. These objectives are achieved by making use of statistical, mathematical, and artificial intelligence techniques, formulating methods to implement data mining algorithms [1, 2, 3].

Currently, data mining has been applied with great success in many fields such as industry, marketing, economics and health [1, 2, 4]. Among these subjects, the health investigation field has a huge amount of heterogeneous data daily fed and handled by information systems. These data are distributed in different and nonintegrated databases; nonetheless, the data are stored in a decentralized manner. In addition, these bases exhibit great ontology variety, which makes health mining a complex and painstaking process [5]. Therefore, in order to discover useful information, favoring aspects such as: optimization of treatments, prevention and monitoring of outbreaks, and cost savings; led several scientific studies to accomplish with the intention to mine data in health care [6, 7, 8, 9].

Consequently, the main objective of this work is to answer the following questions: What are the algorithms and methods used to mine data from the public health? What is the research map of Data Mining in Public Health?

We have this paper in accordance to the IMRAD structure: introduction, methods, results and discussion. IMRAD is adopted as part of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals of the International Committee of Medical Journals Editors†. We believe that adopting this structure would help search engines in international databases to store and to retrieve information within research papers in order to facilitate meta-analyses and systematic reviews.

2. Methods

The adopted method to perform the research was an exploratory literature review, developed based on indications of EBSE Technical Report [10]. The intention of the research was to identify the knowledge constituted on the issue of data mining applied in Public Health. In order to this achievement, we established six steps based on the recommendations of the Cochrane Handbook [11]. The steps that constitute the present review process are shown in Figure 2.

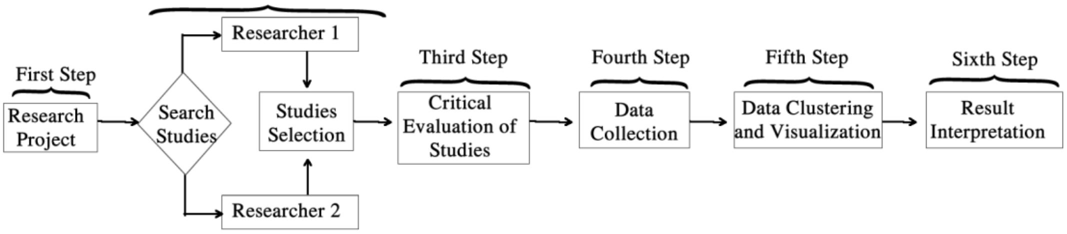

Fig. 2: Preparation steps for the exploratory research.

The keywords used to search the databases were combined with Boolean AND and OR operators, generating the following expression: "Data mining AND (public health OR service health)".

By the use of the keywords in search systems, there were 4,127 articles identified. From the reading of the titles and abstracts of these articles, we download only papers, which apparently met all the inclusion criteria, rejecting those who met at least one exclusion criteria (Table 1).

Tab. 1: Criteria for inclusion and exclusion of articles.

Criteria of Inclusion	**Criteria of Exclusion**
It is related to the theme	It is not related to the theme
It states the data-mining algorithm or used method	It does not state the data-mining algorithm or used method
The study was published between 2009 to September 2013	The study was not published between 2009 to September 2013
It is a publication in a journal, transaction, conference, symposium, workshop or any other scientific event	It is not a publication in a journal, transaction, conference, symposium, workshop or any other scientific event
The access to the paper is provided through CAPES network.	The access to the paper is not provided through CAPES network.
The paper was written in English language.	The paper was not written in English language

The searched studies were classified as identified studies. After reading title and / or summary, the identified studies, which met all the inclusion criteria, were classified as selected. After full reading, the selected studies, which not meet all the inclusion criteria, were classified as excluded. Finally, the studies used in fact to the exploratory study were classified as included studies.

In Table 2, along with the number of included we also present the number of duplicated studies in brackets, in different versions or bases. In such cases, the selection criteria was to choose the must complete and up-to-dated version of the paper.

It is also worthy to note in Table 2, the high exclusion rate among the selected studies due to the fact that many studies, when analyzed in the qualitative and quantitative point of view, did not present the practices employed in DM: algorithm or method.

Tab. 2: Used databases to the exploratory research.

Base	**Identified**	**Selected**		**Included**	
ACM Digital Library	394	6	6,8%	1	4%
IEEE Xplore	2	0	0,0%	0	0%
Emerald	78	1	1,1%	0	0%
Google Scholar	74	34	38,6%	1	4%
ISI Web of Science	96	1	1,1%	0	0%
PubMed	139	24	27,3%	8 (3)	42%
SciELO.org	95	3	3,4%	2 (1)	11,5%
Cochrane Collaboation	18	2	2,3%	0	0%
Wiley InterScience	1745	6	6,8%	4 (1)	19%
Periodicos Capes	1473	7	6,8%	3	11,5%
Microsoft Academic	13	5	5,7%	1 (1)	8%
TOTAL	**4127**	**88**		**19(6)**	

The data extraction process selected 88 articles from full reading, in order to know, understand, and relate the most relevant information. From the used bases, 38.6% of the articles came from the Google Scholar base, resulting in only 4% of inclusion, due to the non-presentation of the algorithm or mining method. However, from the PubMed database 27.3% papers were selected, where 40% of them were included in the review. Afterwards, we proceeded to the collection of data, which is presented in Table 3.

Tab. 3: Information collected and summarized from the selected articles.

Data Extraction	**Description**
Identification of studies	Article title
Author Institution	Identification of institutions to which the authors are bound.
Year/Country	Year of publication and Institution Country that conducted the survey.
Research Method	Case study, survey, questionnaire, experiment, exploratory research.
Problem	Description / summary of the problem.

Objective	Description / summary of the purpose of the research.
Algorithm	Algorithm used to apply the mining technique.
Mining Method	Method used for data mining.
Coverage Database	Target public or region from where the data came from or refer to.
Base Type	Private database, public database with restricted or public access, social networking, questionnaire.
Number of article citations	Number of article citations.
Goal	Classification goal as descriptive or predictive
Classification	Selected or not selected

3. Results and Discussion

The most relevant information to the raised issues of this exploratory study were summarized in Table 4, which relates each study with the article objectives, the data modelling, the mining goals, and the algorithms and methods adopted. Table 4 is sorted by data modelling and methods.

Tab. 4: Relation within articles objectives, data modelling, mining goals, and algorithms and methods.

Study	Article Objectives	Data Modelling	Mining Goals	Algorithm	Algorithm Method	Data Learning
[S12]	Identify relationships between laboratories, medical and prescription drugs	Descriptive Association	Profiles of Hospitals	FP-Growth	Frequent Pattern	Unsupervised
[S13]	Identify signs of adverse effects of drugs	Descriptive Association	Profiles of Medication	Dist Apriori	Frequent Pattern	Unsupervised
[S14]	Identify effects of a drug compared with other medicines	Descriptive Association	Profiles of Medication	Relative Risk Patterns	Relative Risk	Unsupervised
[S15]	Identify adverse reaction in the use of medication	Descriptive Association	Profiles of Medication	Relative Risk Patterns	Relative Risk	Unsupervised
[S17]	Identify disease diagnosis	Descriptive Clustering	Geospatial distribution of disease	SOM	ART-based Neural Network	Unsupervised
[S06]	Identify socioeconomic and geographic aspects of diseases	Descriptive Clustering	Geospatial distribution of disease	SOM	ART-based Neural Network	Unsupervised
[S11]	Identify correlation between the mention of influenza in social networks with the control center data and disease prevention	Descriptive Clustering	Profiles of Diseases	Girvan Newman	Hierarchical	Unsupervised
[S01]	Identify a model of medical treatment	Descriptive Clustering	Profiles of Hospitals	K-means	Partitive	Unsupervised
[S02]	Identify territorial occurrence of diseases	Descriptive Clustering	Geospatial distribution of disease	K-means	Partitive	Unsupervised
[S16]	Identify health status of older people living alone	Descriptive Clustering	Profiles of people	K-means	Partitive	Unsupervised
[S03]	Predict adverse reaction in the use of medication	Predictive Non-Linear classifier	Profiles of Medication	CART	Decision Tree	Supervised
[S04]	Predict financial status of hospitals	Predictive Non-Linear classifier	Profiles of Hospitals	CHAID	Decision Tree	Supervised
[S05]	Predict possible candidates' people for disease	Predictive Non-Linear classifier	Profiles of people	C5	Decision Tree	Supervised
[S07]	Predict childhood obesity	Predictive Non-Linear classifier	Profiles of people	C4.5	Decision Tree	Supervised
[S08]	Predict drug users	Predictive Non-Linear classifier	Profiles of people (users)	CART	Decision Tree	Supervised
[S09]	Predict health conditions of hospitals	Predictive Non-Linear classifier	Profiles of Hospitals	J48	Decision Tree	Supervised
[S10]	Predict infant mortality	Predictive Non-Linear classifier	Profiles of Diseases	J48	Decision Tree	Supervised

[S18]	Predict adverse reactions to medicines	Predictive Linear classifier	Profiles of Medication	Gamma Poisson Shrinker	Bayesian	Supervised
[S19]	Predict risk in the use of medication	Predictive Linear classifier	Profiles of Medication	Gamma Poisson Shrinker	Bayesian	Supervised

Based on Table 4 and in response to the first question raised in this exploratory study: "What are the algorithms and methods used to mine data from public health," we could observe an almost equal percentage of studies between predictive modelling (52,63%) and descriptive modelling (47,36%). Also, there is a slightly greater percentage of studies for clustering (55,55%) against association (44,44%) for descriptive modelling. All predictive algorithms are supervised learning, and the descriptive algorithms are unsupervised. It is worthy acceptable that predictive models need a reasonable precedence and genuine knowledge to learn from databases aiming to accomplish a fairness forecast. For mining goals, we could detect two general types: profiles, and geospatial distribution. Profiles are the general goal of mining with 84,21%, against 15,78% of geospatial distribution, Figure 3.

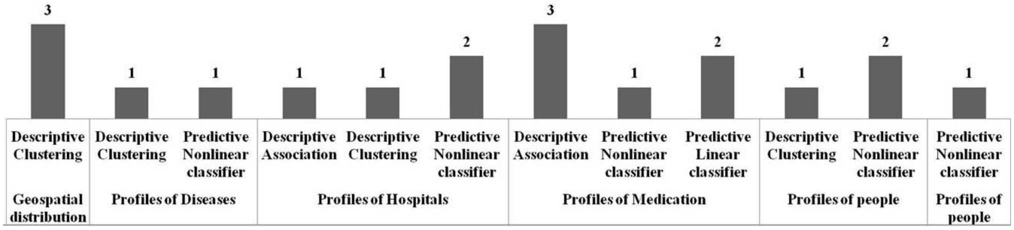

Fig. 3: Data mining algorithms identified in the studies and included detached by mining goal.

For algorithms, we could not find any abnormal frequencies of usage included in these studies, Figure 4, but K-means algorithm achieved a slightly higher percentage of usage than the others, corresponding to 60% for clustering models. Likewise, SOM algorithm was used in 66,66% of geospatial distribution for mining goals.

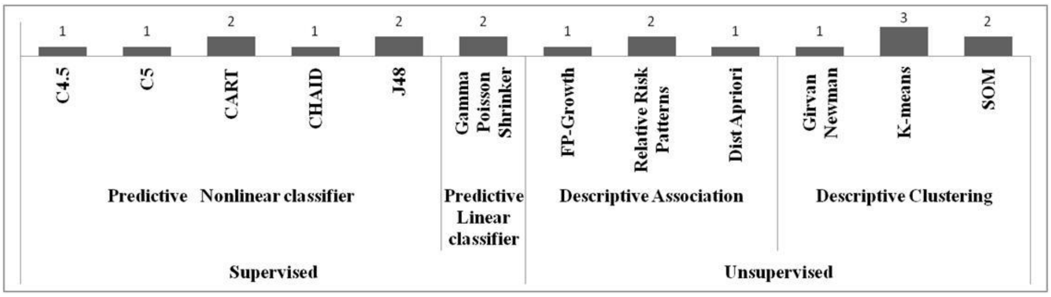

Fig. 4: Data mining algorithms identified in the studies and included detached by data modelling.

K-means and SOM (Self-Organising Map) clustering algorithms have some different aspects that must be considered depending on the work’s purposes. While k-means is better to use when knowing exactly how much (k) clusters the population must be grouped, SOM could converge clusters according to vectors measures and similarity between nodes. SOM has the advantage to perform a non-linear mapping from a high-dimensional data space to a 2-dimensional grid, while keeping the topological relations of the original data.

Relative Risk Patterns (RRP) algorithm was used in 50% of descriptive association models. RRP is a major strategy to demonstrate associations between risk factors and outcome phenotypes. RRP is suitable to usage in clinical data, e.g., to measure the risk of people in developing a disease, or the risk of a side effect from drug treatment.

For the algorithms' methods, decision tree is far the most used method in predictive models, with 77,77%, Figure 5. Moreover, decision tree is the only method for predictive non-linear classifier. Linear models for classification can separate input vectors into classes using linear (hyperplane) decision boundaries. When the classification problem is identified as not linearly separable, non-linear classification must be used. Otherwise, in some cases, non-linear classifiers can provide optimal fit for the data without overfitting.

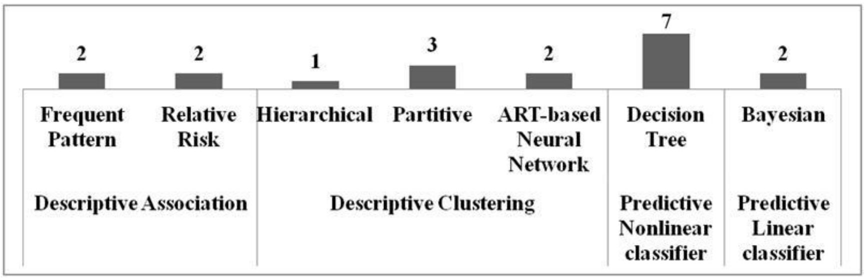

Fig. 5: Data mining methods identified in the studies and included detached by modelling goal.

Many data mining classifiers produces high levels of predictive accuracy but their application to health research and clinical applications could be limited because of the difficult to interpret complex results and to associate it with current knowledge and practices. The graphic knowledge representation of decision trees could help health professional experts to better interpret the elicited evidences, and to postulate causes and effects.

For mining goals, the results showed a profile pattern, Figure 6, highlighting the pursuit to identify patterns for medicines, hospitals, people, and diseases.

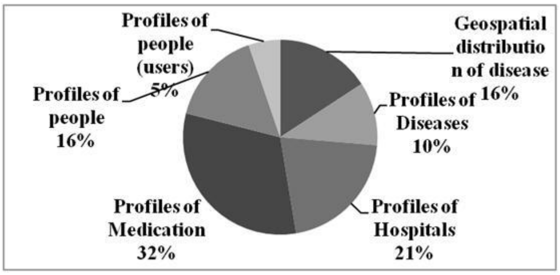

Fig. 6: Mining goals for included studies.

4. Conclusion

The purpose of this exploratory study was to identify used practices (algorithms and methods) and associate them with their mining goals, performing a classification mapping of research in data mining on public health. In this regard, we search for a data mining categorization of models and methods to analyze our data. The result is showed in Figure 7, a categorization scheme for the exploratory research.

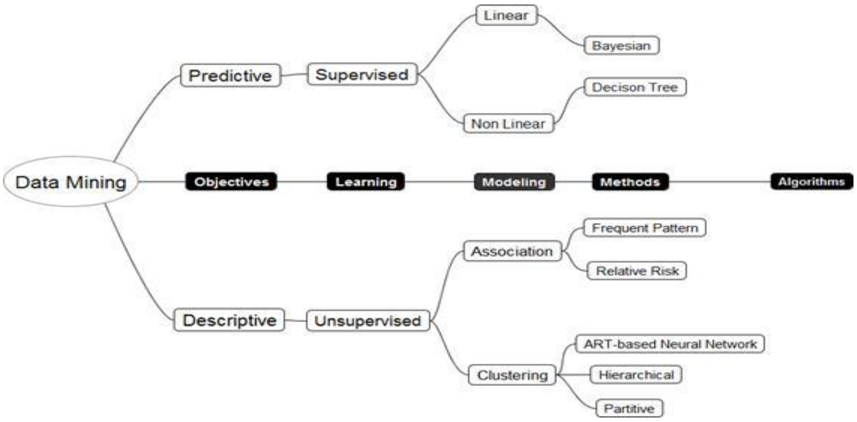

Figure 7. Data Mining Categorization Scheme.

The scheme approach follows the five stages of the KDD process presented earlier. The goal of the first stage is to identify the purposes for the process carrying out, the data mining objectives. Since, we can select a data set based on the learning model, supervised or unsupervised; the data modelling, in this case, could be: linear, non-linear, clustering, or association; and, finally, the algorithm method.

After all, we could associate mining goals to be achieved with the applied methodologies to be used. There is a clean relationship among article objective, data modelling, mining goals, methods and algorithms. This categorization should be useful in future studies for data mining, especially in health field. The scheme could support, e.g., the choice of the methodology to be used depending on the objectives to be achieved. The categorization scheme conveys reliable possibilities of algorithms in accordance with data mining application domains, encouraging researchers from health field by means of good practices.

5. Appendix – Included Studies in Review

[S1] Roa, D.; Rodriguez, N.; Jimenez, A; Bautista, J.; Del Pilar Villamil, M.; Bernal, O., Data mining: A new opportunity to support the solution of public health issues in Colombia, Computing Congress (CCC), *2011 6th Colombian*, **1** (6): 4-6, May 2011 doi: 10.1109/COLOMCC.2011.5936292

[S2] Wei, C. K., Su, S., and Yang, M. C., Application of data mining on the development of a disease distribution map of screened community residents of Taipei County in Taiwan. *J Med Syst*. **36** (3): 2021–2027, 2012

[S3] Chazard, E., Ficheur, G., Bernonville, S., Luyckx, M., and Beuscart, R. Data mining to generate Adverse Drug Events detection rules. *IEEE Trans Inf Technol Biomed*. 2011; **15**: 823–830

[S4] N. Ozgulbas and A.S. Koyuncugil, Financial profiling of public hospitals: An application by Data Mining. *The International Journal of Health Planning and Management*, **22**, DOI: 10.1002/hpm.883, 2007.

[S5] Barakat N, Bradley AP, Barakat MNH: intelligible support vector machines for diagnosis of diabetes mellitus. information technology in biomedicine, *IEEE Transactions on*, USA 2010, **14**:1114-1120.

[S6] Akay, A., Dragomir, A., Yardimci, A., Canatan, D., Yesilipek, A., Pogue, B.: A data-mining approach for investigating social and economic geographical dynamics of β-thalassemia's spread. *IEEE Transaction Information Technology in Biomedicine*, **13**: 774– 783, 2009.

[S7] C. Lazarou, M. Karaolis, A.L. Matalas, D.B. Panagiotakos dietary patterns analysis using data mining method. *An Application to Data from the CYKIDS Study Comput Methods Programs Biomed*, **108** (2): 706–714, 2012.

[S8] Villalon C. Marcelo, Cuellar Caroll. Adolescentes y consumo nocivo de alcohol Chile 2009: Mirando a las políticas públicas. Rev. méd. Chile [online]. 2013, **141** (5): 644-651. Disponível em: <http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0034-98872013000500013&lng=es&nrm=iso>. ISSN 0034-9887. http://dx.doi.org/10.4067/ S0034-98872013000500013.

[S9] N. Lavrač, M. Bohanec, A. Pur, B. Cestnik, B. Debeljak, A. Kobler. Data mining and visualization for decision support and modeling of public health-care resources. *Journal of Biomedical Informatics*, **40** (4): 438–447, 2007.

[S10] Vianna RC, Moro CM, Moyses SJ, Carvalho D, Nievola JC: Data mining and characteristics of infant mortality. *Cad Saude Publica*, **26** (3): 535-542, 2010.

[S11] 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 public health informatics special issue. *International Journal of Environmental Research and Public Health*, **7**, 2010.

[S12] Fernandes, J.; Belo, O., Discovering patterns on medication prescriptions. *2010 5th Iberian Conference on Information Systems and Technologies (CISTI)*, **1** (6): 16-19, June 2010.

[S13] Fan K, Sun X, Tao Y, Xu L, Wang C, Mao X, Peng B, Pan Y. High-Performance Signal Detection for Adverse Drug Events using MapReduce Paradigm. *AMIA Annu Symp Proc*. **2010**: 902–906, 2010.

[S14] Choi, N.-K.; Chang, Y.; Kim, J.-Y.; Choi, Y.-K.; Park, B.-J. Comparison and validation of data-mining indices for signal detection: Using the Korean national health insurance claims database. *Pharmacoepidemiol Drug Saf.* **20**:1278–1286, 2011. doi:10.1002/pds.2237.

[S15] Choi NK, Chang Y, Choi YK, Hahn S, Park BJ. Signal detection of rosuvastatin compared to other statins: Data-mining study using national health insurance claims database. *Pharmacoepidemiol Drug Saf*. **19**: 238–246, 2010.

[S16] Yu-Shiang Hung, Kuei-Ling B. Chen, Chi-Ta Yang, and Guang-Feng Deng. 2012. Mining cluster-based patterns for elder self-care behavior. *In Proceedings of the Tenth Australasian Data Mining Conference - Volume 134 (AusDM '12)*, Yanchang Zhao, Jiuyong Li, Paul J. Kennedy, and Peter Christen (Eds.), **134**: 221-227. Australian Computer Society, Inc., Darlinghurst, Australia, Australia.

[S17] Z. Y. Zhuang, L. Churilov, F. Burstein and K. Sikaris, Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners, Eur. J. Oper. Res., **195** (3): 662-675, 2009.

[S18] Tubert-Bitter P, Bégaud B, Ahmed I. Comparison of two drug safety signals in a pharmacovigilance data mining framework. Stat Methods Med Res. doi:10.1177/0962280212462295

[S19] Berlin, Conny and Blanch, Carles and Lewis, David J. and Maladorno, Dionigi D. and Michel, Christiane and Petrin, Michael and Sarp, Severine and Close, Philippe. Are all quantitative postmarketing signal detection methods equal? Performance characteristics of logistic regression and Multi-item Gamma Poisson Shrinker. Pharmacoepidemiol Drug Saf, **21**: 622–630, 2012.

6. References

[1] Fayyad, Usama M., Piatetsky-Shapiro, Gregory. & Smyth, Padhraic. From data mining to knowledge discovery in databases, *AI magazine*, 1996, **17** (3): 37, doi: http://dx.doi.org/10.1609/aimag.v17i3.1230.

[2] Witten, H. Ian, Frank, Eibe, Hall, Mark A. Data MINING: Practical Machine Learning Tools and Techniques. (3rd eds.) Morgan Kaufmann Inc, Burlington, MA (2011).

[3] Larose, Daniel T. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, New Jersey, USA (2005).

[4] Han, Jiawei, Kamber, Micheline, Pei, Jian. Data Mining: Concepts and Techniques. (3rd eds.) Morgan Kaufmann Inc, San Francisco, USA (2011).

[5] Roa, D., Rodriguez, N., Jimenez, A., Bautista, J., Del Pilar Villamil, M., Bernal, O. (2011), Data mining: A new opportunity to support the solution of public health issues in Colombia, Computing Congress (CCC), *6th Colombian*, 1-6. doi: 10.1109/COLOMCC.2011.5936292.

[6] Wei, C. K., Su, S., and Yang, M. C. (2012), Application of data mining on the development of a disease distribution map of screened community residents of Taipei county in Taiwan. *Journal of Medical Systems*. 2012, **36** (3): 2021–2027. doi: 10.1007/s10916-011-9664-7.

[7] Chazard, E., Ficheur, G., Bernonville, S., Luyckx, M., and Beuscart, R. Data mining to Generate Adverse Drug Events detection rules. *Information Technology in Biomedicine, IEEE Transactions on*. 2011, **15** (6): 823-830. doi: 10.1109/TITB.2011.2165727.

[8] Ozgulbas, N., Koyuncugil, A. S. Financial profiling of public hospitals: An application by data mining, *The International Journal of Health Planning and Management*. 2007, **24** (1): 69-83. doi: 10.1002/hpm.883.

[9] Barakat N., Bradley A. P., Barakat, M.N.H. Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. *Information Technology in Biomedicine, IEEE Transactions on*. 2010, **14** (4): 1114-1120. doi: 10.1109/TITB.2009.2039485.

[10] EBSE Technical Report. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Software Engineering Group School of Computer Science and Mathematics Keele University – UK, and Department of Computer Science University of Durham – UK. Version 2.3, 2009.

[11] Clarke M, Oxman AD, editors. Cochrane Reviewers' Handbook 4.1 [updated March 2001]. in: Review Manager (RevMan) [Computer program]. Version 4.1. Oxford, England: The Cochrane Collaboration, 2001.