A Literature Review of Data Mining Techniques Used in Healthcare Databases

Elma Kolçe (Çela) [1], Neki Frasheri [2]

[1,2] Department of Computer Engineering, Polytechnic University of Tirana, Albania
[1] elmakolce@yahoo.com, [2] nfrasheri@fti.edu.al

Abstract. In this paper we present an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The goal of this study is to identify the most well-performing data mining algorithms used on medical databases. The following algorithms have been identified: Decision Trees, Support Vector Machine, Artificial neural networks and their Multilayer Perceptron model, Naïve Bayes, Fuzzy Rules. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms together results more effective.

Keywords: Data Mining (DM), Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes, Genetic Algorithm, Logistic Regression, Healthcare Database, Diagnosis, Prognosis

1 Introduction

Data mining is defined as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" by Fayyad [1]. Healthcare databases have a huge amount of data but however, there is a lack of effective analysis tools to discover the hidden knowledge. Appropriate computer-based information and/or decision support systems can help physicians in their work. Efficient and accurate implementation of an automated system needs a comparative study of various techniques available. In this paper we present an overview of the current research being carried out using the DM techniques for the diagnosis and prognosis of various diseases, highlighting critical issues and summarizing the approaches in a set of learned lessons. The rest of this paper is organized as follows: First we show the methodology of research used in this study in chapter two, we classify them with different criterions in chapter three, then we identify the most used

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

algorithms for disease diagnosis and prognosis, and finally we show the conclusions of our work.

2 Methodology

The methodology used for this paper was through the survey of journals and publications in the fields of computer science, engineering and health care. European Journal of Scientific Research, International Journal on Computer Science and Engineering, Expert Systems with Applications, Data Science Journal are some of these journals. In order to obtain a general overview on the literature, book chapters, dissertations, working papers and conference papers are also included. The research is focused on most recent publications.

3 Literature review

There are different kinds of studies for DM techniques in medical databases. We identify the following categories:

1. Studies that summarize reviews and challenges in mining medical data in general [6], [24], [25], [31], [32]
2. Studies of DM techniques used for diagnosing and/or prognosing of specific diseases, which can be further classified into three other categories: those which use DM techniques for disease diagnosis [3],[7],[9],[14],[22],[37], for disease prognosis [4],[10],[26],[29],[42],[43], or both diagnosis and prognosis.[13],[36]
3. Studies to investigate factors which have higher prevalence of the risk of a disease[5],[12],[28]
4. Studies that present new technologies and algorithms [18-21], [40], [41] and studies that present new techniques improving old ones, such as [8],[11],[30],[39]
5. Studies that present new frameworks, tool and applications in medicine and healthcare system [2],[15-17],[23],[33-35],[38]

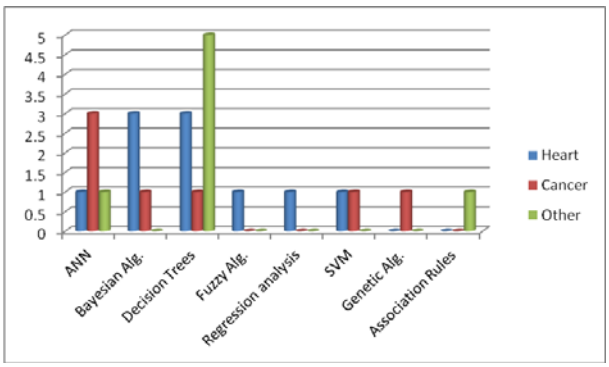

Fig. 1. Efficient Algorithms for Disease Diagnosis

4 Well-performing dm algorithms used for disease diagnosis and prognosis

The graphs in Figures 1 and 2 show the most well-performing algorithms used for disesase diagnosis and prognosis respectively, resulting from the studies in Chapter 3 (excluding studies of categories 1 and 4). We have classified the diseases in Heart Diseases (Cardiovascular disease, Heart Attack, Coronary Arthery Disease, Hypertension), Cancer Diseases (Breast, Prostate, Pancreatic Cancer) and Other Diseases (Asthma, Diabetes, Hepatitis, Kidney Disease, Nerve Diseases, Chronic Disease, Skin Diseases).

As we can see in Fig.1, ANNs are the most well-performing in diagnosing Cancer Diseases, Bayesian Algorithms and Decision Trees in Heart Diseases, and DTS in diagnosing other diseases. On the other side in Fig. 2 we can see that for Cancer and Heart Disease Prognosis, ANNs are the most well-performing and also Bayesian Algorithms the most well-performing in Heart Diseases Prognosis.

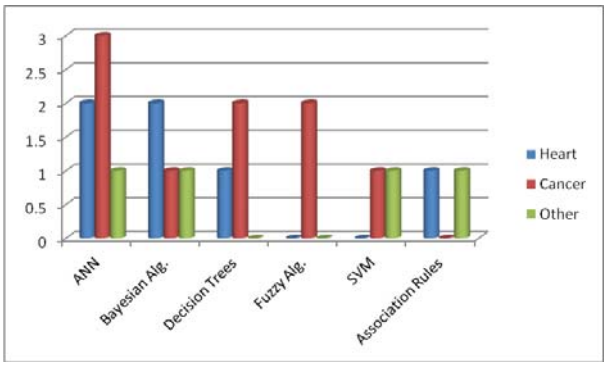

Fig. 2. Efficient Algorithms for Disease Prognosis

5 Conclusions

In this paper we identified and evaluated the most commonly used DM algorithms resulting as well-performing on medical databases, based on recent studies. The following algorithms have been identified: Decision Trees (DT's) C4.5 and C5, Support Vector Machine (SVM), Artificial neural networks (ANNs) and their Multilayer Perceptron model, Bayesian Networks and Naïve Bayes, Logistic Regression, Genetic Algorithms (GAs), Fuzzy Rules, Association Rules.

Analyses show that DTs, ANNs and Bayesian Algorithms are the most well-performing algorithms used for disease diagnosis, while ANNs are also the most well-performing algorithms used for disease prognosis, followed by Bayesian Algorithms, DTs and Fuzzy Algorithms. But it is very difficult to name a single DM algorithm as the best for the diagnosis and/or prognosis of all diseases. Depending on concrete situations, sometime some algorithms perform better than others, but there are cases

when a combination of the best properties of some of the aforementioned algorithms results more effective. The follow-up of our work will aim at dealing with algorithms that have wider spectra of application for groups of diseases.

References

1. Fayyad, U. M. , Piatetsky-Shapiro, G., Smyth, P., Uthurusamy , R. G. R.: Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, CA. (1996)
2. Shantakumar B.Patil, Y.S.Kumaraswamy: Intelligent and EffectiveHeart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.
3. M.Kumari, S. Godara: Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST ISSN : 2229- 4333 Vol. 2, Iss ue 2, June 2011
4. K.Srinivas , B.Kavihta Rani, Dr. A.Govrdhan: Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02,(2010),pp 250-255
5. M. Karaolis, J.A. Moutiris, L. Papaconstantinou, C.S. Pattichis: Association Rule Analysis for the Assessment of the Risk of Coronary Heart Events (2009)
6. R.D. Canlas Jr., Data Mining in Healthcare:Current Applications and Issues (2009)
7. J.Soni, U. Ansari, D. Sharma, S. Soni: Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction (2011)
8. K.S.Kavitha , K.V.Ramakrishnan , M. K. Singh: Modeling and design of evolutionary neural network for heart disease detection, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010, ISSN (Online): 1694-0814, pp. 272-283 (2010)
9. Chi-Ming Chu, Wu-Chien Chien, Ching-Huang Lai, Hans-Bernd Bludau, Huei-Jane Tschai, LuPai, Shih-Ming Hsieh, Nian-Fong Chu, Angus Klar, Reinhold Haux, Thomas Wetter: A Bayesian Expert System for Clinical Detecting Coronary Artery Disease, J Med Sci 2009; 29(4), pp. 187-194 (2009)
10. A.A. Aljumah, M. G.Ahamad, M.K.Siddiqui: Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia, Intelligent Information Management, 3, (2011), pp. 252-261
11. S.H.Ha, S.H.Joo: A Hybrid Data Mining Method for the Medical Classification of Chest Pain, International Journal of Computer and Information Engineering 4:1,pp 33-38 (2010)
12. C. Yang, W. N.Street, Der-Fa Lu, L. Lanning: A Data Mining Approach to MPGN Type II Renal Survival Analysis(2010)
13. S.Gupta, D. Kumar, A.Sharma: Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis (2011)
14. B.D.C.N. Prasad, P.E.S.N. K.Prasad, Y. Sagar: A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis (Asthma) (2011)
15. A.Shukla, R. Tiwari, P. Kaur: Knowledge Based Approach for Diagnosis of Breast Cancer, IEEE International Advance Computing Conference (IACC 2009)
16. E. Savic, J.Potic, Z. Babovic, G. Rakocevic, V. Strineka, M. Dobrota, V. Milutinovic: Sensor Nets and Data Mining in Medical Applications (2011)
17. L. Duan, W. N. Street & E. Xu: Healthcare information systems: data mining methods in the creation of a clinical recommender system, Enterprise Information Systems, 5:2, pp169-181 (2011)

18. T.H. McCormick, C. Rudin, D.Madigan: A Hierarchical Model For Association Rule Mining Of Sequential Events: An Approach To Automated Medical Symptom Prediction
19. S. CHAO, F.WONG: An Incremental Decision Tree Learning Methodology Regarding Attributes In Medical Data Mining (2009)
20. S.Chao , F. Wong: A Multi-Agent Learning Paradigm for Medical Data Mining Diagnostic Workbench
21. I.Ullah: Data Mining Algorithms And Medical Sciences (2012)
22. C. S. Dangare, S.S. Apte: Improved Study of Heart Disease Prediction SystemUsing Data Mining Classification Techniques (2012)
23. D.S.Kumar, G.Sathyadevi, S.Sivanesh: Decision Support System for Medical Diagnosis Using Data Mining (2011)
24. N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik: Data Mining Machine Learning Approaches and Medical Diagnose Systems : A Survey
25. F.Hosseinkhah, H.Ashktorab, R.Veen, M. M. Owrang O.: Challenges in Data Mining on Medical Databases IGI Global pp. 502-511(2009)
26. D.Delen: Analysis of cancer data: a data mining approach (2009)
27. E.Dincer, N.Duru: Prototype of a tool for analysing laryngeal cancer operations
28. Acute Coronary Syndrome Prediction Using Data Mining Techniques- An Application, World Academy of Science, Engineering and Technology 59 pp.474-478 (2009)
29. A.O. Osofisan ,O.O. Adeyemo, B.A. Sawyerr, O. Eweje: Prediction of Kidney Failure Using Artificial Neural Networks (2011)
30. R. Parvathi, S. Palaniammalì: An Improved Medical Diagnosing Technique Using Spatial Association Rules, European Journal of Scientific Research ISSN 1450-216X Vol.61 No.1 pp. 49-59 (2011)
31. F.I.Dakheel, R.Smko, K. Negrat, A.Almarimi: Using Data Mining Techniques for Finding Cardiac Outlier Patients (2011)
32. S.K. Wasan, V. Bhatnagar , H.Kaur: The Impact Of Data Mining Techniques On Medical Diagnostics, Data Science Journal, Volume 5, pp. 119-126 (2006)
33. S.Palaniappan, R. Awang: Intelligent Heart Disease Prediction System Using Data Mining Techniques (2008)
34. M.G. Tsipouras, T.P. Exarchos, D.I. Fotiadis,A.P. Kotsia, K.V. Vakalis, K.K. Naka, L. K. Michalis: Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling (2008)
35. M. L.Jimenez , J. M. Santamarı, R. Barchino, L. Laita, L.M. Laita, L. A. Gonza´lez, A. Asenjo: Knowledge representation for diagnosis of care problems through an expert system: Model of the auto-care deficit situations, Expert Systems with Applications 34 pp.2847–2857 (2008)
36. M.-J. Huang, M.-Y.Chen, S.-C. Lee: Integrating data mining with case-based reasoning for chronicdiseases prognosis and diagnosis, Expert Systems with Applications 32 pp.856–867 (2007)
37. K.Aftarczuk: Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems (2007).
38. T.Sakthimurugan, S.Poonkuzhali: An Effective Retrieval of Medical Records using Data Mining Techniques, International Journal Of Pharmaceutical Science And Health Care. ISSN: 2249-5738. 2(2), pp 72-78 (2012)
39. J.Gao, J. Denzinger, and R.C. James: A Cooperative Multi-agent Data Mining Model and Its Application to Medical Data on Diabetes (2005)
40. A.Habrard, M.Bernard, F. Jacquenet: Multi-Relational Data Mining in Medical Databases, Springer-Verlag (2003), LNAI 278

41. A.Kusiak, Decomposition in Data Mining: A Medical Case Study , B.V. Dasarathy (Ed.), Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology III, Vol. 4384, SPIE, Orlando, FL, April 2001, pp. 267-277
42. S.Floyd: Data Mining Techniques for Prognosis in Pancreatic Cancer (2007)
43. A.Kika, B.Cico, R.Alimehmeti: Using Machine Learning for Preoperative Peripheral Nerve Surgical Prediction (2010)

Keywords Healthcare, Data Analytics, Clinics, Systematic Review, Tools and Techniques. 1 INTRODUCTION. Todayâ€™s healthcare industries are moving from volume-based business into value-based business, which requires an overwork from doctors and nurses to be more productive and efficient.Â Data Mining is described as a process by which data is gathered, analysed and stored in order to produce useful and high quality information and knowledge. This term also includes the way of how this data is gathered, filtering and preparation of the data for use and finally the processing of data to support data analytics and predictive modelling (Russom 2011).Â However, this data is transformed and classified before being ready to use and function (Bakshi 2012).