ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

TAXONOMY OF DATA MINING AND MACHINE LEARNING APPROACHES FOR HEALTHCARE SYSTEMS

N.Satyanandam[1], Dr.Ch.Satyanarayana[2]

Associate Professor, Dept. of CSE, Bhoj Reddy Engineering College for Women, Hyderabad,India[1]

Associate Professor, Dept. of CSE, JNTUK, Kakinada, India[2]

ABSTRACT: This article present the classification of data mining and machine learning approaches for healthcare systems. Machine Learning (ML) field has gained its thrust in any domain of research and now recently has become a reliable utensil in the medical domain. The pragmatic sphere of automatic learning is used in various levels such as medical decision support, medical imaging, extraction of medical knowledge and for overall patient management care. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better, well organized therapeutic mind. It describes a ML based methodology for building an application that is capable of identifying and disseminating healthcare information. Moreover our methodology will extract sentences from published medical papers that mention diseases and treatments and identifies semantic relations that exist between diseases and treatments. Here in this manuscript we also discussed a comprehensive outline of data mining and machine learning approaches for healthcare systems.

Keywords:Data Mining, Machine Learning, Knowledge Management, Healthcare Systems.

I. INTRODUCTION

People care intensely about their physical condition. Life is more hectic than has ever been the medicine that is practiced today is an Evidence-Based Medicine in which medical expertise is not only based on years of practice but on the latest discoveries as well. Utensils can help us to supervise and better keep track of our health, when it comes to healthcare knowledge and management. The traditional healthcare system is also becoming one that embraces the internet and the electronic world. Electronic Health Records (EHR) are becoming the standard in the healthcare domain. Health information recording and clinical data repositories are instant access to patient diagnoses, allergies, and lab test results that enable healthier and time-efficient medical decisions. Medication management rapidly access to drug reactions, immunizations, supplies..etc. Decision support has a facility to capture and exercise the eminence medical data for decisions in the workflow of healthcare. In order to embrace the views, the EHR system has better, faster and more reliable to access data in terms of performance. Major investigations shows that building the process of identifying and disseminating consistent information is complex task.

The remaining sections of the paper are organized as follows: In Section 2, a classification of the data mining and machine learning is presented. Current state of the art is given in Section 3 and the conclusions are summed up in Section 4.This document is a template. An electronic copy can bedownloaded from the conference website. For questions on paper guidelines, please contact the conference publicationscommittee as indicated on the conference website. Information about final paper submission is available from the conference website.

II. CLASSIFICATION OF DATA MINING AND MACHINE LEARNING

Data mining is a moderately novel technology that has not fully matured. Despite of this, there are number of industries that are using it on a regular basis. Many of the organizations are combining data mining with statistics, pattern recognition and other important tools. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions [20].

Machine learning algorithms can be classified as supervised learning or unsupervised learning. In supervised learning, training examples consist of input/output pair patterns. In unsupervised learning, training examples contain only the input patterns and no explicit target output is associated with each input. The unsupervised learning algorithms need to use the input values to discover meaningful associations or patterns. Many successful machine learning systems have been developed over the past three decades in the computer science and statistics communities.

Chen and Chau in 2004 categorized five major paradigms of machine learning research, namely probabilistic and statistical models, symbolic learning and rule induction, neural networks, evolution-based models and analytic learning and fuzzy logic. As of their prognostic power, data mining techniques have been widely used in diagnostic and health care applications. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. The resulting model represents formalized knowledge, which can often provide a good diagnostic opinion. We will briefly review research in each of these areas and discuss their applicability in biomedicine [13, 15, 16, 20].

Classification is the most widely used technique in medical data mining. Dreiseitl et al.in 2001 compare five classification algorithms for the diagnosis of pigmented skin lesions. Their results show that logistic regression, artificial neural networks, and support vector machines performed comparably, while k-nearest neighbors and decision trees performed worse [6,8,10]. This is more or less consistent with the performances of these classification algorithms in other applications (Yang and Liu 1999).

Classification techniques are also applied to analyze various signals and their relationships with particular diseases or symptoms. For example, Acir and Guzelis (2004) apply support vector machines in automatic spike signal detection in Electro Encephalo Grams (EEG), which can be used in diagnosing neurological disorders related to epilepsy. Kandaswamy et al. (2004) use artificial neural network to classify lung sound signals into six different categories (normal, wheeze and rhonchus) to assist diagnosis. Data mining is also used to extract rules from health care data. For example, it has been used to extract diagnostic rules from breast cancer data (Kovalerchuk et al., 2001). The rules generated are similar to those created manually in expert systems and therefore can be easily validated by domain experts. Data mining has also been applied to clinical databases to identify new medical knowledge (Prather et al., 1997 and Hripcsak et al., 2002) [20].

III. **CURRENT STATE OF THE ART**

Natural Language Processing (NLP) and Machine Learning (ML) practices are shows that what demonstration of information and what classification algorithms are suitable to use for identifying and classifying relevant medical information in short texts. We distinguish the fact that tools are able to identify reliable information in the medical domain, construction blocks for a healthcare system with the most recent discoveries. In this inspect we focus on diseases and treatment information and the relation that exists between these two entities. The approach used to solve the two proposed tasks is based on NLP and ML techniques. In a standard supervised ML setting, a training set and a test set are required. The training set is used to train the ML algorithm and the test set to test its performance.

Healthcare information systems collect massive amounts of textual and numeric information about patients, visits, prescriptions, physician notes and more. The information encapsulated within electronic clinical records could lead to improved healthcare quality, promotion of clinical and research initiatives, fewer medical errors and lower costs. However, the documents that comprise the health record vary in complexity, length and use of technical vocabulary. This makes knowledge discovery complex. Commercial text mining tools provide a unique opportunity to extract critical information from textual data archives. Here, they share their experience of a collaborative research project to develop predictive models by text mining electronic clinical records.

M. Schurr, from the Section for Minimal Invasive Surgery of the Eberhard-Karls-University of Tuebingen, gave an invited talk on endoscopic techniques and the role of ML methods in this context. He referred to current limitations of endoscopic techniques, which are related to the restrictions of access to the human body, associated to endoscopy.

Observation: Technical limitations include, restrictions of manual capabilities to manipulate human organs through a small access, limitations in visualizing tissues and restrictions in getting diagnostic information about tissues. To alleviate these problems, international technology developments focus on the creation of new manipulation techniques involving robotics and intelligent sensor devices for more precise endoscopic interventions. It is also acknowledged that this new generation of sensor devices contributes to the development and spread of intelligent systems in medicine by providing ML methods with data for further processing.

Observation: Potentials of new imaging ideology such as fluorescence imaging or laser scanning microscopy and machine learning methods are very elevated. The clinical idea behind these developments is early detection of malignant lesions in stages were local endoscopic therapy is possible. Technical developments in this field are very promising however, clinical results are still pending and ongoing research will have to clarify the real potential of these expertiges for clinical use.

Moustakis and Charissis work (Moustakis and Charissis 1999) surveyed the role of ML in medical decision making and provided an extensive literature review on various ML applications in medicine that could be useful to practitioners interested in applying ML methods to improve the efficiency and quality of medical decision making systems.

Observation: In the above work the point of getting away from the accuracy measures as sole evaluation criteria of learning algorithms was stressed. The issue of comprehensibility i.e. how well the medical expert can understand and thus use the results from a system that applies ML methods are very important and should be carefully considered for the evaluation.

Alexopoulos, Dounias and Vemmos (Alexopoulos et al., 1999) was focused on the application of inductive ML methods in medical diagnosis of stroke. Their approach was based on the See5 algorithm, which is an updated version of the C4.5 algorithm.

Observation: The experiments reported in this approach showed that the capability to learn from examples and to handle missing information by constructing a decision tree, which was possible to be transformed to if/then rules. Special attention was given to the determination of the complexity and comprehensibility of the acquired decision rules, in collaboration with medical experts.

HOLþ et al., 1999, ML methods like the Magnus assistant decision tree learner and the Bayesians classifier were used for the diagnosis and prognosis of first cerebral paroxysm. Despite the fact that best predictions were obtained using the naïve Bayesian classifier, the most interesting results from a medical point of view were achieved using the Magnus assistant decision tree learner. Data and attributes which were considered by expert neurologists as obvious and meaningless turned out to be very important for automatic diagnosis and prognosis.

Observation: In this case, ML methods provided a different estimation of some clinical attributes and motivated clinicians to generate new hypotheses and ultimately to improve their standard diagnostic and prognostic processes.

An interactive system for the ascertainment of visual perception disorders was presented in the article (Ruseckaite et.al 1999). The system performed data analysis and extracted interesting dependencies between visual perception disorder and damage of the brain by applying a modified version of the ML algorithm Charade.

Observations: Here in their approach preliminary results indicated the effectiveness of the rehabilitating persons with certain brain anomalies.

Bourlas, Giakoumakis and Papakonstantinou (Bourlas et al., 1999) extended previous work on medical expert systems for ECG diagnosis by incorporating ML methods to continuously improve the knowledge base of a medical expert system. Their new system exhibited continual learning capabilities using an extended version of the ID3 algorithm to extract from time to time a set to diagnosis rules based on a training set of ECGs.

Observations: The extracted rules were merged into the older ones and the replicas were removed. In order to optimize the performance of the system, a knowledge management subsystem is provided for monitoring the recital of the final rules, in terms of their diagnostic accuracy and modified the knowledge base.

The work of Neves, Alves, Nelas, Romeu and Basto (Neves et al., 1999) demonstrated the need for Health Care Unit's medical imaging models and introduced the concept of a generic and deductive/inductive model of operation, which supports scheduling, forecasting, and accounting.

Observation: Following this approach, several agents concurred in generating hypotheses, each one of them having a different role in evaluating parts of the data and neural networks were used to discover associations in the dataset.

Asteroth and Möller (Asteroth et al., 1999) investigated the use of neural network-based approximation of structural information to the identification of individualized models of the human cardiovascular system. This approach allowed them to achieve robust real-time identification.

The problem of identifying the structure of a population of patients with brain disorder was investigated in the article (Pranckeviciene, 1999). The similarity among patients Electro EncephaloGrams (EEGs) was evaluated by a single layer neural network. Experiments indicated that this approach successfully revealed similarities in the electrical activity of the brain of different patients.

In the article (Jankowski et.al, 1999) the use of incremental neural networks was suggested for approximation and classification tasks. Their model was based on neurons with a new form of rotated bi-radial transfer functions and was dynamically generated to match the complexity of the training data.

Observations: The model demonstrates superior generalization performance when compared with other popular methods for the classification of medical data in the simulations.

In the article of Karkanis, Magoulas, Grigoriadou, and Schurr (Karkanis et al., 1999) a simple scheme consisting of a feature extraction stage and a classification stage was applied. Second order gray level statistics were used for the texture description and a multi-layer feed forward neural network was employed to detect abnormalities in colonoscopy images with high accuracy.

In the article of Karkanis, Galousi, and Maroulis (Karkanis et al., 1999a) outlined a new approach to texture classification applied on lung endoscopic images. Feature selection was based on the texture spectrum of the image and a clustering method was used to distinguish the features with the most discriminative ability.

IV. CONCLUSION

This paper appraised contemporary art in data mining and machine learning approaches. Assessment illustrates that there is huge progress in utilization of DM and ML for healthcare systems. Along with that it is also clear and noticed that the new dimensions in usage of DM, ML and invention of new approaches and strategies give a greater scope for research in healthcare systems. By seeing the growing fame it extracts sentences from published medical papers that mention diseases and treatments, identifies semantic relations that exist between diseases and treatments, which would predict the usability and maintainability in an efficient manner. So we are sanguine regarding future work in this particular direction.

REFERENCES

[1].Baim P.W., A Method for Attribute Selection Inductive Learning Systems, IEEE Trans. on PAMI, Vol.10, No. 6, 1988, pp.888-896.

[2].Bevk M., Kononenko I., Zrimec T., Relation between energetic diagnoses and GDV images, Proc. New Science of Consciousness: 3rd Int Conf. on Cognitive Science, Ljubljana, October 2000, pp. 54-57.

[3]Bratko I., Mozetiˇc I., Lavraˇc N., KARDIO: A study in deep and qualitative knowledge for expert systems, Cambridge,MA: MIT Press, 1989.

[4].Bratko I., Mulec P., An Experiment in Automatic Learning of Diagnostic Rules, Informatica, Ljubljana, Vol.4, No.4, 1980, pp. 18-25.

[5].Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) Classification and Regression Trees, Wadsforth International Group.

[6].Catlett J., On changing continuous attributes into ordered discrete attributes, Proc. European Working Session on Learning-91, Porto, March 4-6, 1991, pp. 164-178.

[7].Cestnik B., Estimating Probabilities: A Crucial Task in Machine Learning, Proc. European Conf. on Artificial Intelligence, Stockholm, August, 1990, pp. 147-149.

[8]Cestnik B., Kononenko I.&Bratko I., ASSISTANT 86 : A knowledge elicitation tool for sophisticated users, in: I.Bratko, N.Lavrac (eds.): Progress in Machine learning, Wilmslow: Sigma Press, 1987.

[9].Chan K.C.C. &Wong A.K.C., Automatic Construction of Expert Systems from Data: A Statistical Approach, Proc. IJCAI Workshop on Knowledge Discovery in Databases, Detroit, Michigan, August, 1989, pp.37-48.

[10].Clark P. & Boswell R., Rule Induction with CN2: Some Recent Improvements, Proc. European Working Session on Learning-91, Porto, Portugal, March, 1991, pp.151-163.

[11].Craven M.W. and Shavlik J.W., Learning symbolic rules using artificial neural networks, Proc. 10th Intern. Conf.on Machine Learning, Amherst, MA, Morgan Kaufmann, 1993, pp.73-80.

[12].Diamond G.A. and Forester J.S., Analysis of probability as an aid in the clinical diagnosis of coronary artery disease, New England J. of Medicine, 300:1350, 1979.

[13].Elomaa T., Holsti N., An Experimental Comparison of Inducing Decision Trees and Decision Lists in Noisy Domains, Proc. 4th European Working Session on Learning, Montpeiller, Dec. 4-6, 1989, pp.59-69.

[14].Good I.J., Probability and the Eeighing of Evidence. London: Charles Griffin, 1950.Good I.J., The Estimation of Probabilities.

[15]Jiawei Han and MichelineKamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd edition.

[16].Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[17]Bharati M. Ramageri / Indian Journal of Computer Science and Engineering,Vol. 1 No. 4 301-305.

[18]"Knowledge Management, Data Mining and Text Mining in Medical Informatics" Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, Medical Informatics and Clinical Epidemiology, Portland, Oregon 97239-3098.

[19]. Abidi, S. S. R. (2001). "Knowledge Management in Healthcare:Towards 'Knowledge driven' Decision support Services," International Journal of Medical Informatics, 63, 5-18.

[20]N.Satyanandam,Dr.Ch.Satyanarayana,Md.Reyazuddin and Amjan Shaik "Data Mining, Machine Learning Approaches and Medical Diagnose Systems: A Survey" International Journal of Computer and Organization Trends" volume 2, Issue 3,2012.

BIOGRAPHY

N.Satyanandamis working as an Associate Professor in the Department of CSE, Bhoj Reddy Engineering College for Women, Hyderabad, India. He received B.Tech(CSSE) in 1996 and MBA (MM) in 1999 both from Andhra University and M.Tech(CSE) in 2004 from Jawaharlal Nehru Technological University, Hyderabad. He has published 1 research paper in International Journal. His main research interests are Data Mining and Warehousing, Digital Image Processing, Computer Networks, Software Engineering and Natural Language Processing. He is a Member of ISTE.

Dr. Ch. Satyanarayana is working as an Associate Professor in the Department of CSE, JNTUK, Kakinada, India. He received B.Tech(CSE) in 1996 and M.Tech(CST) in 1998 both from Andhra University. He has been working in Jawaharlal Nehru Technological University for the last 12 years. He has published 27 research papers in various International Conferences and Journals. His main research interests are Pattern Recognition, Image Processing, Speech Processing, Computer Graphics, Data Mining and Warehousing and Compiler Writing. He is a member of different technical bodies like ISTE, IETE and CSI.
readers.