Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 72 (2015) 306 – 313

Procedia Computer Science

The Third Information Systems International Conference

Data Mining in Healthcare – A Review

Neesha Jothi[a], Nur'Aini Abdul Rashid[b], Wahidah Husain[c], a*

[abc]*School of Computer Sciences, Universiti Sains Malaysia, 11800 Minden, Penang Malaysia*

Abstract

The knowledge discovery in database (KDD) is alarmed with development of methods and techniques for making use of data. One of the most important step of the KDD is the data mining. Data mining is the process of pattern discovery and extraction where huge amount of data is involved. Both the data mining and healthcare industry have emerged some of reliable early detection systems and other various healthcare related systems from the clinical and diagnosis data. In regard to this emerge, we have reviewed the various paper involved in this field in terms of method, algorithms and results. This review paper has consolidated the papers reviewed inline to the disciplines, model, tasks and methods. Results and evaluation methods are discussed for selected papers and a summary of the finding is presented to conclude the paper.

Peer-review under responsibility of organizing committee of Information Systems International Conference (ISICO2015)

Keywords: Data Mining, Data Mining in Healthcare, Health Informactics;

1. Introduction

Across all the fields, data are being collected and accumulated at a vivid pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. At the core of the process is the application of specific data mining methods for pattern discovery and extraction [1]. Among the data mining techniques developed in recent years, the data mining methods are including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining. [2]. As an element of data mining technique research, this paper surveys the

* Corresponding author. Tel.:+604-653-3645; fax: +604-657-4759.
E-mail address: nj14_com042@student.usm.my

Peer-review under responsibility of organizing committee of Information Systems International Conference (ISICO2015)
doi:10.1016/j.procs.2015.12.145

development of data mining technique, through a literature review and the classification of articles from 2005 until 2015 are reviewed. The period is important because, during the time period there is a newly widespread of data mining techniques being used in the healthcare industry where technology has played a significant role especially in the development of methodologies for the collection of data from online databases. The review interest for this literature review, started in the March 2015 with searches made of the keyword indices on the ScienceDirect, Springerlink and IEEE Xplore online databases, for full article containing the phrase "application of data mining techniques in healthcare". For the period from 2005 to 2015, 3840 articles were found. Topic filtering reduced this number to 205 articles, which were related to the keyword. From the 205 articles, 50 articles is used for this review, the papers are collected based on the phrase "application of data mining techniques in healthcare" in no specific categorization. The remaining part of the paper is organized as follows. Section 2.0 discusses the overview of data mining. While section 3.0 discuss the various data mining algorithms used in healthcare.

2. Data Mining An Overview

Data size are generally growing from day to day. The need to understand large, complex, information enriched data sets has now increased in all the varied fields of technology, business and science. With these large amount of data, the ability to extract useful knowledge hidden in these large amount of data and to act on the knowledge is becoming increasingly important in today's competitive world. The process of applying computer based information system (CBIS), including new techniques, for discovering knowledge from data is called data mining [3]. The following subsections will be oriented to define the mentioned attributes of data mining, provide their related instances and insight some figures on their occurrence among the 50 articles mentioned in the section 1.0.

2.1. Disciplines Involved in Data Mining

The data mining baseline is grounded by disciplines such as machine learning [4], artificial intelligence [5], probability [6] and statistics [7]. The disciplines identified among the papers reviewed are summarized in Table 1. Table 1 assets the disciplines mentioned for the papers reviewed.

Table 1. The different disciplines in the papers reviewed

Discipline	Count
Machine learning	40
Artificial intelligence	5
Statistical	3
Probability	2

2.2. Data Mining Models

Generally, there are two kinds of data mining models: predictive model and descriptive model [8]. The predictive model often apply supervised learning functions to predict unknown or future values of other variables of interest [8]. The descriptive model on the other hand, often apply the unsupervised learning functions in finding patterns describing the data that can be interpreted humans [8]. The data mining models identified among the papers reviewed are summarized in Table 2. The predictive models are more commonly used in the healthcare.

Table 2. The two different models in the papers reviews

Model	Count
Predictive	47
Descriptive	3

2.3. Data Mining Tasks

Usually, the implementation of a model is made by a task. For instance, clustering [9], association rules [10], correlation analysis [11], are often used for descriptive models. While classification [12], regression [13] and categorization [14] are used for predictive models. Table 3 shows the task derived from the papers reviewed.

Table 3. The task derived from the papers reviewed

Task	Count
Classification	42
Association rules	5
Clustering	2
Anomaly detection	1

2.4. Data Mining Methods

Having the data mining model and task defined, next would be the data mining methods to build the approach based on discipline involved. The methods used for anomaly detection are, standard support vector data description, density induced support vector data description, Gaussian mixture. While the vector quantization method is widely used for clustering. The methods widely used for classification are statistical, discriminant analysis, decision tree, Markov based, swarm intelligence, k-nearest neighbor, genetic classifiers, artificial neural network, support vector and association rule.

3. Data Mining Algorithms in Healthcare

Healthcare covers a detailed processes of the diagnosis, treatment and prevention of disease, injury and other physical and mental impairments in humans [15]. The healthcare industry in most countries are evolving at a rapid pace. The healthcare industry can be regarded as place with rich data as they generate massive amounts of data including electronic medical records, administrative reports and other benchmarking finding [16]. These healthcare data are however being under-utilized. As discussed in 2.0 data mining is able to search for new and valuable information from these large volumes of data. Data mining in healthcare are being used mainly for predicting various diseases as well as in assisting for diagnosis for the doctors in making their clinical decision. The discussion on the various methods used in the healthcare industry are discussed as follows.

3.1. Anomaly Detection

Anomaly detection is used in discovering the most significant changes in the data set [17]. Bo Lie et al [18] had used three different anomaly detection method, standard support vector data description, density-induced support vector data description and Gaussian mixture to evaluate the accuracy of the anomaly detection on uncertain dataset of liver disorder dataset which is obtained from UCI. The method is evaluated using the AUC accuracy. The results obtained for a balanced dataset by average was 93.59%. While the average standard deviation obtained from the same dataset is 2.63. The uncertain dataset are prone to be available in all datasets, the anomaly detection would be a good way to resolved this matter, however since there is only one paper discussing this method, we cannot comment much on the effectiveness of the method.

3.2. Clustering

The clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data [17]. Rui Velosoa [19] had used the vector quantization method in clustering approach in predicting the readmissions in intensive medicine. The algorithms used in the vector quantization method are k-means, k-mediods and x-means. The datasets used in this study were collected from patient's clinical process and laboratory results. The evaluation for each of the algorithms are conducted using the Davies-Bouldin Index. The k-means obtained the best results while x-means obtained a fair results while the k-mediods obtained the worst results. From the results the work by these researchers provide a useful result in helping to characterize the different types of patients having a higher probability to be readmitted. A more significant comparison on the method cannot be made since this is the only one paper in my review discussing on the vector quantization.

3.3. Classification

Classification is the discovery of a predictive learning function that classifies a data item into one of several predefined classes [17]. The related work in classification will be discussed in the following subsections.

3.3.1. Statistical

The MTS algorithm is being extensively applied in multivariable statistical analysis. The Mahalanobis distance (MD) is used to build statistical judgements to distinguish one group from another and the Mahalanobis space (MS) is used to represent the degree of abnormality of observations from the known reference group. In the statistical classifiers, the authors Su et al. [20], have used the Mahalanobis Taguchi System (MTS) to design the prediction model for pressure ulcers. The class imbalance problems are very much prevalent in the healthcare datasets. Usage of the data mining algorithms are often affected with skewed distribution when using skewed or imbalanced data sets. This problem often leads to the tendency of producting highly predictive classification accuracy over the majority class and poor accuracy over the minority class. Having such a nature to distinguish the degree of abnormality of observations, this method would be a good method to test on the real data set pressure ulcers. This method is also used since the MD is suitably scaled. The test conducted using this algorithms were done in four phases with scaled datasets ranging from 14 to 8, 5, and 2 accordingly. The results obtained in the paper [20] shows that the measurement scale for this algorithm has good a performance based on the huge difference between the normal and abnormal examples. Being an algorithm which is suitable for scaling the MTS proves to have better sensitivity and g-means values in the testing stage. The MTS has enhanced performance in terms of sensitivity.

3.3.2. Discriminant Analysis

Linear discriminant analysis (LDA) is widely used in discriminant analysis to predict the class based on a given set of measurements on new unlabeled observations [17]. Authors Armañanzas et al. [21] and Jen et al [22] have used the linear discriminant analysis in their respective work. Jen et al [21] had the algorithm in predicting the severity staging of Parkinson's disease patient using scores of non-motor symptoms. Their study is intended to quantitatively analyze the inner relationships between both motor and non-motor symptoms. The linear discriminant analysis is the conditional probability density function of the predictors follows a normal distribution based on the given class value. The algorithm's ability to capture statistical dependencies among the predictor variables indicates that this algorithm would be suitable to explore the linear constraint of this study to discovery the synergy between motor and non-motor symptoms. The proposed model obtained an accuracy estimation of 69% compared to other algorithms since the algorithm's performance increases significantly when the dependencies are in linear form. Based on the same nature of the algorithm the authors Armañanzas et al. [21], used the algorithm to evaluate the classification accuracy to seek the most substantial risk factor and establish the initial set of substantial risk factors for chronic illness early warning. From the results of the two works we can safely say that the algorithm has good results and it is suitable to be utilized to identify significant accuracy if the relationships of the healthcare data are in linear form.

3.3.3. Decision Tree

Several study have explored the decision tree method to analyze clinical data. The authors Sharma & Om [23], Wang et al. [24] and Zolbanin et al.[25] have used the decision tree algorithm in their respective work. Having the nature to examine data and make the tree and its rules are used to make a prediction. All the three works have used the decision tree to the data set to improve the prognostic performance, in terms of accuracy. The nature of the data set used in this research are rather balanced set of data set. From the comparative of the works, we conclude that decision tree as cannot be used in proposing prognostic decision to solve imbalanced problems because the decision tree recursively separate observations into branches to construct a tree.

3.3.4. Swarm Intelligence

The authors Yeh et al. [26], Fei 2010 [27], and Abdi & Giveki [28] have used the swarm intelligence method to designed their diagnosis model. The algorithm particle swarm optimization (PSO) is able to efficiently find the optimal or near optimal solutions in large search spaces. All the three authors tried to resolve optimzation problem which often involves features in the classification problems. The classification process will be faster and more accurate if less number of features are used. From the work studied, the PSO based approach proves to improve the overall classification results since PSO is being used to select suitable parameters in the involved classifiers.

3.3.5. K-Nearest Neighbor

Authors García-Laencina et al. [29], Armañanzas et al.[21], Jen et al.[22], Bagui et al.[30], and aŞahan et al. [31] have used the k-nearest neighbour in their respective predictive models. The k-nearest neighbour is an instance based classifier method. The parameter units consists of samples that are used in the method and this algorithm then assumes that all instances relate to the points in the *n*-dimensional space R^N. The algorithm is very expedient as the information in the training data is never lost. However, this algorithm would be suitable if the training data set is large as this algorithm is very time consuming when each of the sample in training set is processed while classifying a new data and this process requires a longer classification time. From the work by the mentioned authors, the classification accuracy is what

they would like to attain instead of classification time as the classification accuracy is more important in the medical diagnosis.

3.3.6. Logistic Regression

Logistic regression (LR) is a method that would use the given set of features either continuos, discrete, or a mixture of both types and the binary target, the LR then computes a linear combination fo the inputs and passes through the logistic function [29]. This method is commonly used because it is easy to implementation and it provides competitive results. Authors García-Laencina et al. [29], Mamiya et al. [32], Su et al. [20], Wang et al.[24], Zolbanin et al. [25], Thompson et al. [33], and Samanta et al. [34] have adopted the LR in their respective research work. The results obtained from all the authors are not very significant, due to the signifiant decrement in the size of the input data sets. The results would have been more significant if the datasets were large in quantity as the boundaries of accuracy would be larger. The LR works well for larger datasets.

3.3.7. Bayesian Classifier

Authors Armañanzas et al. [21], and Bandyopadhyay et al.[35] have used the Bayesian classifier method in their respective predictive model. The Bayesian classifiers is well known for its computational efficient and ability to handle missing data naturally and efficiently. Having this advantage both the authors have recorded a good prediction accuracy from the models designed respectively. By having the models implemented the Bayesian classifier also proves that the model is suitable since the averaging approach has led to improved prediction accuracy and allows authors to extract more features from the data without being overfitting. This method would be a good approach if there data sets are suffering from missing data.

3.3.8. Support Vector

The support vector method (SVM) is proven to be advantageous in handling classification tasks with execellent generalization performance. The method seeks to minimize the upper bound of the generalization error based on the structural risk minimization principle. The SVM training is equivalent to solve a linear constrained quadratic programming problem [36]. The method is very commonly used in medical diagnosis. Authors García-Laencina et al. [29], Zheng et al. [36], Kang et al. [37], and Su et al.[20] have used the method in their model in medical diagnoses. Some of the authors have used the SVM method for comparative study purpose. The SVM method generalization ability is controlled by two different factors, that is the training error and the capacity of the learning machine measured. The training error rate can be controlled by changing the features in the classifiers. From the results obtained from the studies, it clearly shows that the SVM showed greater performance since it maps the features to higher dimensional space**.**

Discussion

From the papers reviewed and discussed, the data mining methods accuracy varies depending on the features of the data sets and the size of data set between the training and testing sets. The common characteristics among the healthcare data sets are highly imbalanced data sets, where by the majority and the minority classifier are not balanced resulting prediction erroneous when run by the classifiers. Another characteristics of healthcare data sets are the missing values. The sample size of the data is often seen as another characteristics as the data available are usually in small scale. There is no one suitable data mining method to resolve all this issues.

Conclusion

The data mining has played in an important role in healthcare industry, especially in predicting various types of diseases. The diagnosis is widely being used in predicting diseases, they are extensively used in medical diagnosing. In conclusion, there is no one data mining method to resolve the issues in the healthcare data sets. In order to obtain the highest accuracy among classifiers which is important in medical diagnosing with the characteristics of data being taken care, we need to design a hybrid model which could resolve the mentioned issues. Our future directions is to enhance the predictions using hybrid models.

Acknowledgements

We would like to express our gratitude to Universiti Sains Malaysia (USM) for supporting this research.

References

[1] R. Agrawal and G. Psaila, "Active data mining," *Current*, pp. 3–8, 1995.

[2] S. H. Liao, P. H. Chu, and P. Y. Hsiao, "Data mining techniques and applications - A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, 2012.

[3] G. E. Vlahos, T. W. Ferratt, and G. Knoepfle, "The use of computer-based information systems by German managers to support decision making," *Inf. Manag.*, vol. 41, no. 6, pp. 763–779, 2004.

[4] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. 2011.

[5] D. K. Bhattacharyya and S. M. Hazarika, *Networks, Data Mining And Artificial Intelligence: Trends And Future Directions*, 1st ed. Narosa Pub House, 2006.

[6] M. Karegar, A. Isazadeh, F. Fartash, T. Saderi, and A. H. Navin, "Data-Mining by Probability-Based Patterns," pp. 353–360, 2008.

[7] H. Thomas and L. Paul, *Statistics: Methods and Applications*, 1st ed. StatSoft, Inc, 2005.

[8] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd ed. Wiley-IEEE Press, 2011.

[9] P. Berkhin, "A Survey of Clustering Data Mining," *Group. Multidimens. Data*, no. c, pp. 25–71, 2006.

[10] T. P. Hong, K. Y. Lin, and S. L. Wang, "Fuzzy data mining for interesting generalized association rules," *Fuzzy Sets Syst.*, vol. 138, no. 2, pp. 255–269, 2003.

[11] D. R. Hardoon, S. Sandor R., and S. John R., "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *J. Neural Comput.*, vol. 16, no. 12, pp. 2639 – 2664, 2004.

[12] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3918 LNAI, pp. 199–204, 2006.

[13] Z. Wu and C. Li, "L0-Constrained Regression for Data Mining," pp. 981–988, 2007.

[14] A. Genkin, D. D. Lewis, and D. Madigan, "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[15] J.-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang, and H. Pan, "Emerging information technologies for enhanced healthcare," *Comput. Ind.*, vol. 69, pp. 3–11, 2015.

[16] N. Wickramasinghe, S. K. Sharma, and J. N. D. Gupta, "Knowledge Management in Healthcare," vol. 63, pp. 5–18, 2005.

[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, pp. 37–54, 1996.

[18] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "SVDD-based outlier detection on uncertain data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 597–618, 2013.

[19] R. Veloso, F. Portela, M. F. Santos, Á. Silva, F. Rua, A. Abelha, and J. Machado, "A Clustering Approach for Predicting Readmissions in Intensive Medicine," *Procedia Technol.*, vol. 16, pp. 1307–1316, 2014.

[20] C. T. Su, P. C. Wang, Y. C. Chen, and L. F. Chen, "Data mining techniques for assisting the diagnosis of pressure ulcer development in surgical patients," *J. Med. Syst.*, vol. 36, no. 4, pp. 2387–2399, 2012.

[21] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, "Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach," *Artif. Intell. Med.*, vol. 58, no. 3, pp. 195–202, 2013.

[22] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, "Application of classification techniques on development an early-warning system for chronic illnesses," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8852–8858, 2012.

[23] N. Sharma and H. Om, "Data mining models for predicting oral cancer survivability," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 2, no. 4, pp. 285–295, 2013.

[24] K.-J. Wang, B. Makond, and K.-M. Wang, "An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data.," *BMC Med. Inform. Decis. Mak.*, vol. 13, p. 124, 2013.

[25] H. M. Zolbanin, D. Delen, and A. Hassan Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," *Decis. Support Syst.*, vol. 74, pp. 150–161, 2015.

[26] W.-C. Yeh, W.-W. Chang, and Y. Y. Chung, "A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8204–8211, 2009.

[27] S. W. Fei, "Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6748–6752, 2010.

[28] M. J. Abdi and D. Giveki, "Automatic detection of erythemato-squamous diseases using PSO-SVM based on association rules," *Eng. Appl. Artif. Intell.*, vol. 26, no. 1, pp. 603–608, 2013.

[29] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Comput. Biol. Med.*, vol. 59, pp. 125–133, 2015.

[30] S. C. Bagui, S. Bagui, K. Pal, and N. R. Pal, "Breast cancer detection using rank nearest neighbor classiÿcation rules," vol. 36, pp. 25–34, 2003.

[31] S. Şahan, K. Polat, H. Kodaz, and S. Güneş, "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis," *Comput. Biol. Med.*, vol. 37, no. 3, pp. 415–423, 2007.

[32] H. Mamiya, K. Schwartzman, A. Verma, C. Jauvin, M. Behr, and D. Buckeridge, "Towards probabilistic decision support in public health practice: Predicting recent transmission of tuberculosis from patient attributes," *J. Biomed. Inform.*, vol. 53, pp. 237–242, 2015.

[33] V. L. S. Thompson, S. Lander, S. Xu, and C. Shyu, "Identifying key variables in African American adherence to colorectal cancer screening : the application of data mining," pp. 1–10, 2014.

[34] B. Samanta, G. L. Bird, M. Kuijpers, R. a. Zimmerman, G. P. Jarvik, G. Wernovsky, R. R. Clancy, D. J. Licht, J. W. Gaynor, and C. Nataraj, "Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms," *Artif. Intell. Med.*, vol. 46, no. 3, pp. 201–215, 2009.

[35] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. O'Connor, *Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data*. 2014.

[36] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1476–1482, 2014.

[37] S. Kang, P. Kang, T. Ko, S. Cho, S. Rhee, and K.-S. Yu, "An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction," *Expert Syst. Appl.*, vol. 42, no. 9, pp. 4265–4273, 2015.