Application of Data Mining Techniques for Medical Data Classification: A Review

Saima Anwar Lashari[1,*], *Rosziati* Ibrahim[1], *Norhalina* Senan[1] *and N. S. A. M.* Taujuddin[2]

[1]Faculty of Computer Science and Information Technology
[2]Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

Abstract. This paper investigates the existing practices and prospects of medical data classification based on data mining techniques. It highlights major advanced classification approaches used to enhance classification accuracy. Past research has provided literature on medical data classification using data mining techniques. From extensive literature analysis, it is found that data mining techniques are very effective for the task of classification. This paper analysed comparatively the current advancement in the classification of medical data. The findings of the study showed that the existing classification of medical data can be improved further. Nonetheless, there should be more research to ascertain and lessen the ambiguities for classification to gain better precision.

1 Introduction

Data Mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile computing [1]. Of late, data mining has been applied successfully in healthcare fraud and detecting abuse cases [2]. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients [3]. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. Successful data mining applications have provided the impetus for the relevant parties to fully utilized them as they have realised that data mining is crucial in the acquisition of valuable information for all sectors involved in healthcare-related industries.

Healthcare insurers are able to identify fraud and abuse cases, health administrators are able to make better decisions especially in managing their customers, and healthcare practitioners are able to deliver better services and treatments. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods [4]. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data. Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data. Insights gained from data mining can influence cost, revenue, and operating efficiency while maintaining a high level of care. Healthcare organizations that perform data mining are better positioned to meet their long-term needs [5].

In recent years, computers and their peripherals have been made cheaper and more readily available and in line with the development of information technology, various kinds of advanced data mining techniques have hit the market. These new age data mining techniques embrace traditional and more recent sophisticated classification algorithms. Both classification techniques are for handling complex datasets such as multidimensionality, user inference and prior knowledge, web data, spurious data points that cause overfitting of models, improvement in human ability, noisy datasets cleaning, mining multimedia datasets and incremental datasets. Interdisciplinary data mining techniques and approaches can be used for all the above mentioned databases for forecasting the impact and discovering meaningful relationships in the data with the purpose of extracting useful information for knowledge generation [6].

To employ data mining algorithms to medical data, researchers' comprehension on the type of data mining algorithms and their functions should be clear. Descriptive (or unsupervised learning) and predictive (or supervised learning) algorithm are the two categories of data mining algorithms [2] [4]. Descriptive data mining group's data by determining the objects' similarity (or records) and detecting patterns that are unknown, or associations in the data whereby users are able to recognize a massive data pool. Descriptive data mining is investigative. It includes clustering, association, summarization, and sequence discovery [5]. Prediction data mining that comprises classification, regression,

* Corresponding author: saima@uthm.edu.my

time series analysis, and prediction [8] implies predicting rules (also known as classification/prediction models) from (training) data and the rules are employed to unpredicted/unclassified data [9].

In assisting researchers to comprehend understand the importance of data mining, and the application of data mining techniques, three of the most widely-used data mining algorithms (classification, clustering, and association) are discussed below, complete with guidelines for their respective use.

Thus, a variety of models have been fitted in order to determine hidden patterns in the data. The approach that is able to produce the most accurate output and relationships pattern in the observed datasets is considered to be the most efficient in the particular model. Such approach fulfils the objective of data mining. Current data mining practices utilizes a range of model functions including classification, regression, clustering, discovering association rules and sequence analysis [8].

However, the challenge increases as the interest in data mining grows rapidly. In order to handle these problems without using the traditional statistical methods, soft computing has emerged to be one of the encouraging data mining solutions in this area [9].

The digital medical data is not only enormous in amount, but also complex in its structure for traditional software and hardware. Some of the contributing factors to the failure of traditional systems in handling these datasets include:

a) The vast variety of structured and unstructured data such as handwritten doctor notes, medical records, medical diagnostic images (magnetic resonance imaging (MRI), computed tomography (CT)), and radiographic films [10].
b) Existence of noisy, heterogeneous, complex, longitudinal, diverse, and large datasets in healthcare informatics [11].
c) Necessity of improving medical issues such as quality of care, sharing, security of patients' data, and the reduction of the healthcare cost, which are not sufficiently addressed in traditional systems [8].

Hence, solutions are needed in order to manage and analyze such complex, diverse, and huge datasets in a reasonable time complexity and storage capacity for enhanced insight and decision-making. Therefore, this paper aims to highlight an assessment of the current and common methods for medical data categorisation. The existing approaches to the categorisation of medical records based on data mining techniques are being revised by highlighting the diverse categorisation algorithms for clinical imaging applications.

This paper is arranged into five sections. Section 2 explains classification in data mining. Comparative analysis of data mining techniques is provided in Section 3, followed by a discussion of the results in Section 4 and the paper ends with a conclusion in Section 5.

2 Classification in Data Mining

In data mining, categorization is formulated to make a forecast of the memberships in a group for data instances. This process utilizes complex analysis of data to determine data connections in huge datasets. Due to its complex features, medical databases provide complications for pattern extortion [7]. There are two approaches to data mining: statistical and machine learning algorithms. The processes in data mining are classified into descriptive and predictive (Fig. 1). Descriptive mining tasks provide the general data properties in the database. For Predictive mining tasks, inference is made on the data for predictions [9] whereby forecast is made on explicit values based on patterns identified by known results. Descriptive data mining, without having any predefined target, provides characteristics and descriptions for the data set.

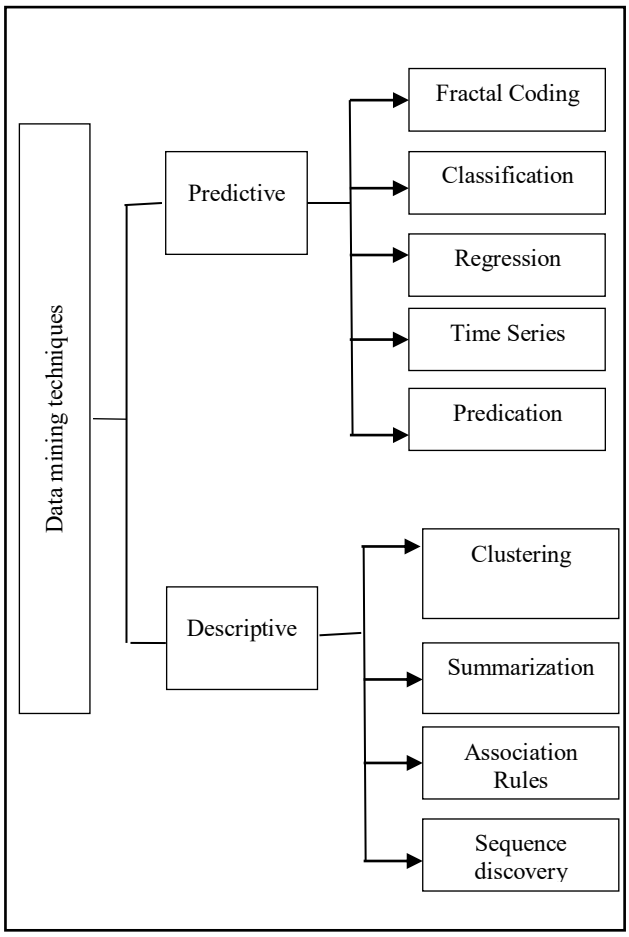

Fig. 1. Data mining techniques

Data mining techniques are effective and predictive for future patterns because: a) it is user friendly and prediction is based on past circumstances b) it operates by learning from past data c) data from numerous resources is managed and only required data is extracted d) models are easily updated by relearning, past information and change in trends. These are what makes it reliable and practical in the medical image categorization [4].

The three learning approaches in data mining algorithms are supervised (the algorithm works with a set of examples with known labels whose values are

nominal in classification task, or numerical in regression task), unsupervised (unknown labels in the dataset, and the algorithm typically aims at grouping examples according to the similarity of their attribute values, characterizing a clustering task, or semi-supervised (whereby learning is conducted when there is availability of a small subset of labelled examples, concurrent with a large number of unlabelled examples).The task of categorising is regarded as a supervised technique in which each instance belongs to a class, specified by the value of a special goal attribute or the class attributes [14].

3 Comparative analysis of data mining techniques

Solutions resulted from classification algorithm are commendable but as of now, none is diverse and flexible to be accepted generally in the medical data classification community. Categorical variables in medical data are occasionally useful to arrive at decisions and to generalize information. Categorical data (e.g. classification of disease and non-disease groups) is handy for data mining technique and also easy to extract medical information. Table 1 represents data mining techniques representation briefly.

Table 1. Data Mining Techniques Representation

Approach	Representation	Pattern Recognition Models Recognition Function	Error Estimation
Template Matching	Sample, pixels, curves	Correlation, distance measurement	Classification error
Statitical	Featuers	Discriminatio n function	Classification error
Neural networks	Samples, voxel pixels, featuers	Network function	Mean square error

With the advancement in data mining for diagnosis and prognosis of different diseases, a significant number of attempts have been proposed for a wide variety of medical image classifications. However, the different assumptions and hypothesis have been made in these methods differ considerably. The section reviews the rapidly expanding body of work on the development and application of classification methods to problems of fundamental importance in medical image classification. Meanwhile, in conducting comparative studies, classification researches rely heavily on stored repositories of data (such as UCI repository [15] it allows new algorithm ideas to test its plausibly on known problems. Table 2 represents summary of data mining techniques.

Table 2. Summary of data mining techniques

Data mining technique	Advantage	Disadvantage
SVM	▪ Better Accuracy as compare to other classifier. ▪ Easily handle complex nonlinear data points. ▪ Over fitting problem is not as much as other methods.	▪ Computationally expensive. ▪ The selection of right kernel function. For every dataset different kernel function shows different results. ▪ Training process take more time. ▪ Suitable for the problem of binary class. It solves the problem of multi class by breaking it into pair of two classes such as one against-one and one-against all.
Decision Tree	▪ No requirements of domain prior knowledge in the construction of decision tree. ▪ It minimizes the ambiguity of complicated decisions. ▪ It is easy to interpret and it also handles both numerical and categorical data.	▪ It generates categorical output. ▪ Performance of classifier depends upon the type of dataset.
ANN	▪ Easily identify complex relationships between dependent and independent variables. ▪ Able to handle noisy data	▪ Local minima. ▪ Over-fitting.
Bayesian Belief Network	▪ It makes computations process easier. ▪ Have better speed and accuracy for huge datasets.	It does not give accurate results in some cases where there exists dependency among variables.

K-NN	▪ It is easy to implement. ▪ Training is done in faster manner	▪ It requires large storage space. ▪ Sensitive to noise. ▪ Testing is slow.

Support Vector Machine (SVM)
Artificial Neural Network (ANN)
k- nearest neighbour (K-NN)

Data mining techniques, which are a recent application in the medical domain, are applied in mining medical data, which comprises association rule mining for finding frequent patterns, prediction, classification and clustering. To date, there have been many research on this and intelligent and decision support systems have been developed to make more accurate diagnosis and prediction of diseases especially in predicting heart diseases, lung and breast cancer and remote health monitoring.

Table 3 provides the summary of medical data classification regarding the resolved difficulties that are solved, convenience in medical data mining or implementation of the tools. Therefore, selected researches on classification performance of different classifiers are summarised. Here is shown the effort made for data classification. Nevertheless, it is obvious that benchmarking to determine the best classification algorithm for medical data classification is still lacking.

Table 3. Summary of medical data mining techniques

Author(s), Year	**Medical Dataset**	**Technique (s)**	**Comments**
Khanmohammadi & Chou, 2016	Six datasets (UCI Repository)	GMBD	Discretization process was more concise for representation of continuous variable
Aswal & Ahuja 2016	six datasets (UCI Repository)	K-NN, SVM	Classification of bio medical data
Long, 2015	heart disease	rough sets based attribute reduction and interval type-2 fuzzy logic system (IT2FLS)	heart disease diagnostic system using rough sets based on attribute reduction and IT2FLS
Zuo et al. 2013	Parkinson Disease	Fuzzy K-NN approach	Familiarised an adaptive Fuzzy K-NN approach for diagnosing the disease
Ghofrani et al. 2014	X-Ray dataset	K-NN & SVM	Slow testing, scale depended
Ramana *et al.* 2011	Liver dataset	KNN & SVM	High accuracy
Kharya, 2012	MIAS	Decision tree	Iterative training producer, overtraining sensitive, need pruning
Rajini & Bhavani, 2011	MRI	ANN & k-NN	Sensitive to training parameters, slow training,
Cheng et al. 2009	Dermatology	ANN & Decision Tre	dermatologic diagnosis
Polat et al. 2007	Breast Cancer and Liver Disorders dataset	Fuzzy-AIRS	Modelling and analysis of medical data
Dangare & Apte, 2012	heart disease	Naive Bayes Decision Trees Neural Networks	accuracy for different classification methods with 13 input attributes & 15 input attributes values
Xing et al. 2007	coronary heart disease	Decision Tree Algorithms such as C4.5, C5, and CART	Prediction models
Shim, et al. 2003	Liver diseases	Classification using BYY	-
Tang et al. 2009	Diabetes, Cancer	k-Nearest Neighbour	Classification of Disease using K-NN

Acronym:
Genetic k-Nearest Neighbour (GkNN)
Support Vector Machine (SVM)
k- nearest neighbour (K-NN)
Gaussian mixture model based discretization (GMBD)
Bayesian Ying Yang (BYY)

This part of the paper is to present the necessities in the healthcare industry, and the potential techniques to be utilised. Here are guidelines on the usage of the different data mining techniques: Identification of the unnecessary attributes which impedes the processing task is crucial before the application of the classification technique. Besides acting as noise and disturbing the process, they also affect the classifier performance. To identify these, statistical methods are employed. In contrast, the feature selection methods are engaged to select functional attributes in order to improve the precision and success of the classification model [18]. Thus, the researchers have found that there is no classifier that generates the best result for each dataset. To check the performance of classifiers, each dataset is

parted into two division – training and testing. A classifier, which is tested using a testing dataset, is chosen based on its performance in comparison to other available classifiers. Cross validation method is conducted for both training and testing dataset to ensure accuracy.

The main emphasis of classification rules is to discover the class of attributes, but it does not take into account the relationships of attributes. While association is useful for identifying the relationship or association among numerous attributes and generates association rules which in turn helpful for domain experts to remove insignificant association rules and consider only those rules which are useful for making vital decision

4 Discussion

Based on Table 1, each data mining modality has its individual characteristics with which to contend. With all the efforts, there is still no extensively used method to classify medical data. The most explainable reason is that exceptionally precise data and especially very low rate of false negatives are required in the medical domain. There are general methods, however, that can be applied to a variety of data. Specialized methods for particular applications that consider prior knowledge can enable the achievement of better performance. There have been progress in medical data over the past decades in the following three areas:

1) Development and use of advanced classification algorithms
2) Use of multiple features
3) Integration of medical data into classification procedures.

Nevertheless, there are challenges. These include data mining methodology, user interaction, performance and scalability. Other issues are the exploration of data mining application and their social impacts. The selection of an appropriate approach to a classification problem can therefore be a difficult dilemma. Consequently, there is the possibility of further improvement in the current medical data classification tasks. As data mining techniques for medical data classification has not been fully investigated, there is a great potential for further work and interesting directions for future research can be created.

5 Conclusion

In this paper, it is observed that data mining techniques have been employed for medical data classification. There are voluminous records in this medical data domain and because of this, it has become a requisite to use data mining techniques to help in decision support and prediction in the field of healthcare to identify diseases. Therefore, medical data mining contributes to business intelligence which is useful for diagnosing of diseases. This paper throws light into data mining techniques that are used for medical data for various diseases which are identified and diagnosed for human health. For future works, despite many opportunities and approaches for big data analytics in healthcare are presented in this work, there are many other directions to be explored, concerning various aspects of healthcare data, such as the quality, privacy, timeliness, and so forth. This section provides an outlook of big data analytics in healthcare informatics from a broader view, which covers the topics of healthcare data characteristics (e.g., high complexity, large scale, etc.), data analytics tasks (e.g. longitudinal analysis, visualization, etc.), and objectives (e.g. real-time, privacy protection, collaboration with experts, etc.). The future of health informatics will benefit from the exponentially increasing digital health data.

Due to their practicality, data mining applications may benefit the healthcare industry immensely. Nonetheless, it should be cautioned that this benefit depends on how clean the data is. To ensure the success of data mining applications, the capture, storage, preparation and mining of data must be critically considered whereby there should be a standard practice of clinical vocabulary and data-sharing across healthcare establishments. Additionally, the scope and nature of healthcare data can be expanded beyond quantitative data. Text mining can be conducted to harvest data from handwritten medical notes and records by the doctors. Combining data and text mining may enrich the process. Besides, there have also been development in data mining through digital diagnostic images. The researchers highly anticipate that findings from this study can provide advancement to data mining and healthcare resources and enable all relevant parties to be benefitted.

This work is supported by Research, Innovation, Commercialization, and Consultancy (ORICC), with the grant post-doctoral fund, Vote No D001, Universiti Tun Hussein Onn Malaysia.

References

[1] Yoo, Illhoi, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. "Data mining in healthcare and biomedicine: a survey of the literature." *Journal of medical systems* 36, no. 4 (2012): 2431-2448.

[2] Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[3] Chang, Chun-Lang, and Chih-Hao Chen. "Applying decision tree and neural network to increase quality of dermatologic diagnosis." *Expert Systems with Applications* 36, no. 2 (2009): 4035-4041.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[5] D. Chauhan and V. Jaiswal, “An efficient data mining classification approach for detecting lung cancer disease,” in *Communication and Electronics Systems (ICCES), International Conference on*, 2016, pp. 1–8.

[6] M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. Treister, “Transforming Health Care Through Big Data Strategies for leveraging big data in the health care industry,” *Inst. Heal. Technol. Transform. http//ihealthtran. com/big-data-in-healthcare*, 2013.

[7] Ramana, Bendi Venkata, M. Surendra Prasad Babu, and N. B. Venkateswarlu. "A critical study of selected classification algorithms for liver disease diagnosis." *International Journal of Database Management Systems* 3, no. 2 (2011): 101-114.

[8] Sharmila, K., and S. A. Vethamanickam. "Survey on data mining algorithm and its application in healthcare sector using Hadoop platform." *International Journal of Emerging Technology and Advanced Engineering* 5, no. 1 (2015): 567-71.

[9] Kharya, Shweta. "Using data mining techniques for diagnosis and prognosis of cancer disease." *arXiv preprint arXiv:1205.1923* (2012).

[10] Dangare, Chaitrali S., and Sulabha S. Apte. "A data mining approach for prediction of heart disease using neural networks." (2012).

[11] Shim, Jeong-Yon, and Lei Xu. "Medical data mining model for oriental medicine via BYY binary independent factor analysis." In *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on*, vol. 5, pp. V-V. IEEE, 2003.

[12] Ghofrani, Fatemeh, Mohammad Sadegh Helfroush, Habibollah Danyali, and Kamran Kazemi. "Improving the performance of machine learning algorithms using fuzzy-based features for medical x-ray image classification." *Journal of Intelligent & Fuzzy Systems* 27, no. 6 (2014): 3169-3180.

[13] S. A. Lashari, R. Ibrahim, N. Senan, I. T. R. Yanto, and T. Herawan, “Application of Wavelet De-noising Filters in Mammogram Images Classification Using Fuzzy Soft Set,” in *International Conference on Soft Computing and Data Mining*, 2016, pp. 529–537.

[14] S. A. Lashari, R. Ibrahim, and N. Senan, “De-noising analysis of mammogram images in the wavelet domain using hard and soft thresholding,” in *Information and Communication Technologies (WICT), 2014 Fourth World Congress on*, 2014, pp. 353–357.

[15] Khanmohammadi, Sina, and Chun-An Chou. "A Gaussian mixture model based discretization algorithm for associative classification of medical data." *Expert Systems with Applications* 58 (2016): 119-129.

[16] Xing, Yanwei, Jie Wang, and Zhihong Zhao. "Combination data mining methods with new medical data to predicting outcome of coronary heart disease." In *Convergence Information Technology, 2007. International Conference on*, pp. 868-872. IEEE, 2007.

[17] Rajini, N. Hema, and R. Bhavani. "Classification of MRI brain images using k-nearest neighbor and artificial neural network." In *Recent Trends in Information Technology (ICRTIT), 2011 International Conference on*, pp. 563-568. IEEE, 2011.

[18] Polat, Kemal, Seral Şahan, Halife Kodaz, and Salih Güneş. "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism." *Expert Systems with Applications* 32, no. 1 (2007): 172-183.

[19] Aswal, Shobha, and Neelu Jyothi Ahuja. "Experimental analysis of traditional classification algorithms on bio medical dtatasets." In *Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on*, pp. 566-568. IEEE, 2016.

[20] Tang, Ping-Hung, and Ming-Hseng Tseng. "Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification." In *Machine Learning and Cybernetics, 2009 International Conference on*, vol. 5, pp. 3070-3075. IEEE, 2009.

[21] Long, Nguyen Cong, Phayung Meesad, and Herwig Unger. "A highly accurate firefly based algorithm for heart disease prediction." *Expert Systems with Applications* 42, no. 21 (2015): 8221-8231.

[22] Zuo, Wan-Li, Zhi-Yan Wang, Tong Liu, and Hui-Ling Chen. "Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach." *Biomedical Signal Processing and Control* 8, no. 4 (2013): 364-373.