COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES FOR MEDICAL DATA CLASSIFICATION

Lashari S. A. and Ibrahim R.
Universiti Tun Hussein Onn Malaysia, hi120040@siswa.uthm.edu.my, rosziati@uthm.edu.my

ABSTRACT. Medical data classification plays a crucial role in many medical imaging applications by automating or facilitating the delineation of image data. It addresses the problem of diagnosis, analysis and teaching purposes in medicine. For these several medical imaging data modalities and applications based on data mining techniques have been proposed and developed. In this paper, a comparative analysis of applications of data mining techniques has been presented. Thus, the existing literature suggests that we do not lose sight of the current and future potential of applications of data mining techniques that can impact upon the successful classification of medical data into a thematic map. Thus, there is a great potential for the use of data mining techniques for medical data classification, which has not been fully investigated and would be one of the interesting directions for future research.

Keywords: Medical data classification, data mining, neural networks, texture classification

INTRODUCTION

With the development of information technology, medical data has been widely available from different modalities e.g. X-ray, computed tomography (CT), magnetic resonance images (MRI), ultrasound etc. Thus, the explosive growth of data storage and the amount of data bases to store the digitized data has increased exponentially (Mitra et al. 2002). An early attempt at computerized analysis of medical images were made in the 1960s, later serious and systematic investigation on computer-aided diagnosis CAD began in the 1980s with a fundamental change in the concept for utilization of the computer output, from automated computer diagnosis to computer-aided diagnosis (Doi, 2007).

Effective medical images can play an important role in aiding diagnosis and treatment of the diseases and they can also be helpful in the education domain for healthcare students. As a consequence, the vast amount of medical data accessible calls for developing new tools to effectively and efficiently manage and retrieve the data of interest to the medical practitioner and to the research community. Thus, the concept of data mining has been created and evolved in ongoing efforts to efficiently harvest useful information from huge data repositories (Antoni, Zaiane & Coman, 2001). Consequently, a data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data.

The rest of the paper is structured as Section 2 to Section 5. Section 2 discusses data mining techniques. Section 3 provides comparative analysis of data mining techniques. Discussion has been made in Section 4 followed by conclusion in Section 5.

DATA MINING TECHNIQUES

Data mining is a research line that began in 1980 in order to find the knowledge that is hidden in the data. Classification in data mining is used to predict group membership for data instances. Data mining involves the use of sophisticated data analysis tools to discover the relationship in large datasets. Medical databases in general pose a unique problem for pattern extortion because of their complex nature. Data mining approaches are mainly comprised of statistical and machine learning algorithms (Kaur & Wasan, 2006).

Data mining techniques can be broadly classified into two categories: parametric models and non- parametric models. Parametric model based on traditional statistical algorithms attempt to a fit a mathematical define a priori describe the relationship between input and output variables. Typically they assume that domain has an underlying data structures. In cases when there is little knowledge of the underlying distributions, the success of these models is limited or at least unknown. Non-parametric models on other hand do not use domain specific knowledge but instead rely heavily on the data to derive the relationship (Han, Kamber & Pei, 2011). Data mining algorithms generally belong somewhere along a continuum between these two categories of parametric and non- parametric models. For medical data classification, the emphasis is placed more on non-parametric models (where ideas are motivated from concepts of pattern recognition, image processing, and computer vision). These are tasks directly related to medical imaging:

- Image segmentation (Zaidi & ElNaqa, 2010)
- Image data classification (Kharrat et al., 2010; Tu, Shin & Shin, 2009)
- Computer-aided diagnostic characterization (Hadjiiski et al., 2004; Doi, 2007)
- Image annotation (Russell et al., 2008)

There have been few survey articles in the relevant research field (Antonie *et al*. 2001; Perner, 2002, Miller & Blott, 1992; Smitha, Shaji & Mini, 2011; Kassner & Thornhill, 2010 in previous years. Many different classification strategies were applied and while in the earlier years nearest neighbor-based approaches were most common and most successful (Suguna & Thanushkodi, 2010). Later, decision trees (Kharya, 2012), as well as support vector machines (Kharrat et al., 2010, Aarthi et al., 2011) became more and more common and outperformed the nearest neighbor–based approaches. Analogous to feature combination, classifier combination has also been a popular way to improve performance (Kharat, 2012; Suguna & Thanushkodi, 2010.This seems to indicate that data mining techniques have beginning to be applied widely in the detection and differential diagnosis of many different types of abnormalities in medical images obtained in various examinations by use of different imaging modalities.

COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES

This section provides some selected studies on medical image classification are summarized in Table 1. It is obvious that every classification algorithm provides admirable results, to date it appears that no one solution is diverse and flexible to obtain general acceptance in the medical image classification community.

Table 1. Peer Classification Performance

Researcher(s)/ Year	Dataset	Classifier	Property	Overall Accuracy (%)	Comments
Suguna & Thanushkodi, 2010	Dermatology, Cleveland, Heart, HIV, Lung Cancer, Wisconsin	GkNN	Assign patterns to the majority class among *k* nearest neighbor using a performance optimized value for k	97.92	Slow testing, scale depended
Kharya, 2012	MIAS	Decision tree	Finds a set of thresholds for a pattern- dependent sequence of features	93.62	Iterative training producer, overtraining sensitive, need pruning
Rajendran, Madheswaran & Naganandhini, 2010	Brain tumor			90	
Tu, Shin & Shin, 2009	Heart Disease			78.93	
Aarthi et al. 2011	MIAS	SVM	Maximizes the margin between the classes by selecting a minimum number of support vectors.	86.11	Scale dependent , iterative , slow training , nonlinear
Kharrat et al. 2010	Human brain dataset			96.36	
Rajini & Bhavani, 2011	MRI	ANN & k-NN	Iterative MSE optimization of two or more layers of Perception using sigmoid transfer functions	90	Sensitive to training parameters, slow training, may be produce confidence values, overtraining sensitive

Abbreviations:

Magnetic Resonance Image (MRI)
Feed Forward Neural Network (FF-NN)
Mammographic Image Analysis Society (MIAS)
Genetic k-Nearest Neighbor (GkNN)
Back –Propagation Neural Network (BP-NN)

Suguna & Thanushkodi (2010) attempted to probe an improved k-NN using genetic algorithm was utilized to reduce high calculation complexity with low dependency on the training set and no weight difference between each class. Thus, recent studies try to overcome limitation of traditional k-NN, and are able to produce better results. An approach for classification of brain MRI using genetic algorithm with SVM was proposed and able to classify brain tissue into normal, benign or malignant tumor. However, SVM tend to perform much better when dealing with multi-dimensions and continuous features. Moreover, a large sample size is required in order to achieve its maximum prediction accuracy (Kharrat et al., 2010).

Table 2 presents few articles encompass the different diseases detection system where implementation was observed on various types of images with different types of cancer regions.

Table 2. Summary of major approaches using medical image for improving classification accuracy

References	Classifier	Characteristic of Approach
Kumar & Raju, 2008	ANN	Predicting early cancer cells using texture features
Sarhan, 2009	ANN& DCT	Stomach cancer detection system
Rajini & Bhavani, 2011	ANN+ k-NN	An automatic brain MRI diagnostic system with normal and abnormal classes.
Latifoglu et al. 2008	PCA, k-NN weight +artificial immune recognition system	For diagnosis of atherosclerosis disease

Decision tree based classification methods are widely used in data mining for the decision support application. This type of systems use decision support that have to be made by physicians whether the maximum frequent item set that are found in the transaction tree has been compared with the maximum frequent item of the test images to classify the benign and malignant images (Smitha, Shaji & Mini, 2011). Hence the diagnosis can be made easily. Moreover, texture classification can be applied to any modality of digital image and helps to obtain spectral properties of an image. Texture analysis methods can be divided into four statistical, geometrical, model-based and signal processing (Kassner & Thornhill, 2010).

DISCUSSION

Medical data has made a great progress over the past decades in the following three areas (1) development and use of advanced classification algorithms (2) use of multiple features (3) incorporation of ancillary data into classification procedures. However, few challenges include data mining methodology, user interaction, performance and scalability; other issues include the exploration of data mining application and their social impacts. Based on Table 1and 2, each imaging modality has its own idiosyncrasy with which to contend. With all the efforts, there is still no widely used method to classify medical data. This is due to the fact that medical domain requires high accuracy and especially the rate of false negatives to be very low. Nevertheless, methods do exist that are more general and can be applied to a variety of data. However, methods that are specialized to particular applications can often achieve better performance by taking into account prior knowledge. Selection of an appropriate approach to a classification problem can therefore be a difficult dilemma. In consequence, still there is much room for further improvement over current medical data classification tasks. Therefore, there is a great potential for the use of data mining techniques for medical data classification, which has not been fully investigated and would be one of the interesting directions for future research.

CONCLUSION

This paper examines current practices, problems and prospects of medical data classification. The emphasis is placed on the summarization of major advanced classification

approaches and the techniques used for improving classification accuracy. Since, researchers have gained interest and invested resources to investigate seemingly interesting data mining applications. A considerable amount of literature has been published on medical data classification. While looking into large and growing body of literature, it is appears that data mining techniques have been proven to be successful for classification tasks. Thus, in this paper a comparative analysis of the recent development in medical data classification has been done. The paper has provided an up to date discussions of medical data classification techniques used in the literature. Since, medical data might be in the form of numeric and textual information that may be interspersed and redundant. Thus, we need efficient, robust and flexible machine learning algorithms. From these evidences on medical data classification, it can be seen that there is still much room for further improvement over current medical data classification tasks. More research, however, is needed to identify and reduce uncertainties in medical data classification to improve classification accuracy.

ACKNOWLEDGMENT

The authors would like to thank office for Research, Innovation, Commercialization and Consultancy Management (ORICC) and Universiti Tun Hussein Onn Malaysia for supporting this research.

REFERENCES

Antonie, M. L., Zaiane, O. R., & Coman, A. (2001). Application of data mining techniques for medical image classification. *MDM/KDD*, 94-101.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, *31*(4-5), 198.

Fesharaki, N. J., & Pourghassem, H. (2012). Medical X-ray Images Classification Based on Shape Features and Bayesian Rule. In *Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on* (pp. 369-373). IEEE.

Hadjiiski, L., Chan, H. P., Sahiner, B., Helvie, M. A., Roubidoux, M. A., Blane, C. & Shen, J. (2004). Improvement in Radiologists' Characterization of Malignant and Benign Breast Masses on Serial Mammograms with Computer-aided Diagnosis: An ROC Study1. *Radiology*, *233*(1), 255-265.

Hadidi, M. R. A. A., Gawagzeh, M. Y. & Alsaaidah, B. (2012). Solving mammography problems of breast cancer detection using artificial neural networks and image processing techniques. *Indian journal of science and technology*. Volume 5, No.4.pp.2520-2528.

Han, J., Kamber, M. & Pei, J. (2011). Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann.

Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science*, *2*(2), 194-200.

Kotsiantis, S. B.. Supervised machine learning: A review of classification techniques, *Informatica*. 2007, Volume (31), pp. 249-268.

Kassner, A., Thornhill, R. E. (2010). Texture Analysis: A Review of Neurologic MR Imaging Applications, A Journal of NeuroRadiology 31:809

Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. International Journal of computer science and information technology (IJCSIT), Vol. 2, No.2, pp 55-66.

Kharat, K. D., Kulkarni, P. & Nagori, M. B. (2012). Brain tumor classification using neural network based methods. International journal of computer sciences and informatics, ISSN: 2231-5292-, Vol-1, issue 4,pp 85-90.

Kharrat, A., Gasmi, K. Messaoud, M. B., Benamrane, N. & Abid, M. (2010). A hybrid approach for automatic classification of brain MRI using genetic algorithm and support vector machine", *Leonardo Journal of science.*,ISSN-1582-0233, pp. 71-82.

Kumar, G. J., & Kumar, G. V. (2008). Biological Early Brain Cancer Detection Using Artificial Neural Networks. In *Artificial Intelligence and Pattern Recognition* (pp. 89-93).

Latifoglu, F., Polat, K., Kara, S., Gunes, S. (2008). Medical diagnosis of atherosclerosis from carotid artery Doppler signals using principal component analysis (PCA), k-NN based weighting pre-processing and Artificial Immune Recognition System (AIRS). J. Biomed. Inform., 41, 15–23

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE transactions on neural networks*, *13*(1), 3-14.

Miller, A. S., & Blott, B. H. (1992). Review of neural network applications in medical imaging and signal processing. *Medical and Biological Engineering and Computing*, *30*(5), 449-464

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, *77*(1-3), 157-173.

Rajini N. H. & Bhavani, R. (2011). Classification of MRI Brain Images using k- Nearest Neighbor and Artificial Neural Network, IEEE-International Conference on Recent Trends in Information Technology, ICRTIT.

Rajendran, P., Madheswaran, M., & Naganandhini, K. (2010). An improved pre-processing technique with image mining approach for the medical image classification. In *Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on* (pp. 1-7). IEEE.

Sarhan, A. M. (2009). Cancer classification based on micro array gene expression data using DCT and ANN. *Journal of theoretical and applied information technology*, *6*(2), 208-216.

Smitha, P., Shaji, L., Mini, M. G. (2011). A review of medical image classification technique, International confernce on VLSI, Communication & Intrumrnataiom

Suguna, N. & Thanushkodi, K. (2010). An improved k-nearest neighbour classification using genetic algorithm. *International journal of computer science issues (IJCSI), Volume 7,issue 4,No2*.

Tu, M. C., Shin, D., & Shin, D. (2009). A comparative study of medical data classification methods based on decision tree and bagging algorithms. In *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on* (pp. 183-187). IEEE.

Perner, P. (2002). Image mining: issues, framework, a generic tool and its application to medical-image diagnosis. *Engineering Applications of Artificial Intelligence*, *15* (2), 205-216

Zaidi, H., & El Naqa, I. (2010). PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *European journal of nuclear medicine and molecular imaging*, *37*(11), 2165-2187.