Analysis of Data Mining Techniques for HealthCare Data Classification

Dr. M. Balamurugan[1], S. Vijaykumar[2]
[1]Associate Professor, School of Computer Science, Engineering and Applications, Bharathidasan University, Tamilnadu
[2]Research Scholar, School of Computer Science, Engineering and Applications, Bharathidasan University, Tamilnadu

***Abstract*- Medical data classification plays a crucial role in manymedical imaging applications by automating or facilitating the delineation ofimage data. It addresses the problem of diagnosis, analysis and teachingpurposes in medicine. For these several medical imaging data modalitiesand applications based on data mining techniques have been proposed anddeveloped. In this paper, a comparative analysis of applications of datamining techniques has been presented. Thus, the existing literature suggeststhat we do not lose sight of the current and future potential of applications ofdata mining techniques that can impact upon the successful classification ofmedical data into a thematic map. Thus, there is a great potential for the useof data mining techniques for medical data classification, which has notbeen fully investigated and would be one of the interesting directions forfuture research.**

***Index Terms*- Medical data classification, data mining, neural networks,texture classification**

I. INTRODUCTION

With the development of information technology, medical data has been widely availablefrom 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 databases to store the digitized data has increased exponentially (Mitra et al. 2002). An earlyattempt at computerized analysis of medical images were made in the 1960s, later serious andsystematic investigation on computer-aided diagnosis CAD began in the 1980s with afundamental change in the concept for utilization of the computer output, from automatedcomputer diagnosis to computer-aided diagnosis (Doi, 2007).

Effective medical images can play an important role in aiding diagnosis and treatment ofthe diseases and they can also be helpful in the education domain for healthcare students. As aconsequence, the vast amount of medical data accessible calls for developing new tools toeffectively and efficiently manage and retrieve the data of interest to the medical practitionerand to the research community. Thus, the concept of data mining has been created andevolved in ongoing efforts to efficiently harvest useful information from huge datarepositories (Antoni, Zaiane&Coman, 2001). Consequently, a data mining is the non-trivialextraction 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 datamining techniques. Section 3 provides comparative analysis of data mining techniques.Discussion has been made in Section 4 followed by conclusion in Section 5.

II. DATA MINING TECHNIQUES

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

Data mining techniques can be broadly classified into two categories: parametric modelsand non-parametric models. Parametric model based on traditional statistical algorithmsattempt to a fit a mathematical define a priori describe the relationship between input andoutput variables. Typically they assume that domain has an underlying data structures. Incases when there is little knowledge of the underlying distributions, the success of

thesemodels is limited or at least unknown. Non-parametric models on other hand do not usedomain specific knowledge but instead rely heavily on the data to derive the relationship(Han, Kamber & Pei, 2011). Data mining algorithms generally belong somewhere along acontinuum between these two categories of parametric and non- parametric models. Formedical data classification, the emphasis is placed more on non-parametric models (whereideas are motivated from concepts of pattern recognition, image processing, and computervision). 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, 2010in previous years. Many different classification strategies were applied and while in theearlier 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 commonand 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 havebeginning to be applied widely in the detection and differential diagnosis of many differenttypes of abnormalities in medical images obtained in various examinations by use of differentimaging modalities.

III. COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES

This section provides some selected studies on medical image classification aresummarized in Table 1. It is obvious that every classification algorithm provides admirableresults, to date it appears that no one solution is diverse and flexible to obtain generalacceptance in the medical image classification community.

Suguna&Thanushkodi (2010) attempted to probe an improved k-NN using geneticalgorithm was utilized to reduce high calculation complexity with low dependency on thetraining set and no weight difference between each class. Thus, recent studies try to overcomelimitation 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 toclassify brain tissue into normal, benign or malignant tumor. However, SVM tend to performmuch better when dealing with multi-dimensions and continuous features. Moreover, a largesample size is required in order to achieve its maximum prediction accuracy (Kharrat et al.,2010).

Decision tree based classification methods are widely used in data mining for the decisionsupport application. This type of systems use decision support that have to be made byphysicians whether the maximum frequent item set that are found in the transaction tree hasbeen compared with the maximum frequent item of the test images to classify the benign andmalignant 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 toobtain spectral properties of an image. Texture analysis methods can be divided into fourstatistical, geometrical, model-based and signal processing (Kassner& Thornhill, 2010).

IV. 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 challengesinclude data mining methodology, user interaction, performance and scalability; other issuesinclude the exploration of data mining application and their social impacts. Based on Table1and 2, each imaging modality has its own idiosyncrasy with which to contend. With all theefforts, there is still no widely used method to classify medical data. This is due to the factthat medical domain requires high accuracy and especially the rate of false negatives to bevery low. Nevertheless, methods do exist that are more general and can be applied to a varietyof data. However, methods that are specialized to particular applications

can often achievebetter performance by taking into account prior knowledge. Selection of an appropriateapproach to a classification problem can therefore be a difficult dilemma. In consequence,still there is much room for further improvement over current medical data classificationtasks. Therefore, there is a great potential for the use of data mining techniques for medicaldata classification, which has not been fully investigated and would be one of the interestingdirections for future research.

V. CONCLUSION

This paper examines current practices, problems and prospects of medical dataclassification. The emphasis is placed on the summarization of major advanced classification approaches and the techniques used for improving classification accuracy. Since, researchershave gained interest and invested resources to investigate seemingly interesting data miningapplications. A considerable amount of literature has been published on medical dataclassification. While looking into large and growing body of literature, it is appears that datamining techniques have been proven to be successful for classification tasks. Thus, in thispaper a comparative analysis of the recent development in medical data classification hasbeen done. The paper has provided an up to date discussions of medical data classificationtechniques used in the literature. Since, medical data might be in the form of numeric andtextual information that may be interspersed and redundant. Thus, we need efficient, robustand flexible machine learning algorithms. From these evidences on medical dataclassification, it can be seen that there is still much room for further improvement overcurrent medical data classification tasks. More research, however, is needed to identify andreduce uncertainties in medical data classification to improve classification accuracy.

REFERENCES

[1] Antonie, M. L., Zaiane, O. R., &Coman, A. (2001). Application of data mining techniques for medicalimage classification.MDM/KDD, 94-101.Doi, K. (2007). Computer-aided diagnosis inmedical imaging: historical review, current status andfuture potential. Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society, 31(4-5), 198.

[2] Fesharaki, N. J., &Pourghassem, H. (2012). Medical X-ray Images Classification Based on ShapeFeatures and Bayesian Rule. In omputational Intelligence and Communication Networks(CICN), 2012 Fourth International Conference on (pp. 369-373).IEEE.

[3] 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 onSerial Mammograms with Computer-aided Diagnosis: An ROC Study1. *Radiology*, *233*(1), 255-265.

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

[5] Han, J., Kamber, M. & Pei, J. (2011). Data Mining: Concepts and Techniques, Third Edition. MorganKaufmann.

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

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

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

[9] 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, pp55-66.

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

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

[12] Kumar, G. J., & Kumar, G. V. (2008). Biological Early Brain Cancer Detection Using Artificial NeuralNetworks. In Artificial Intelligence and Pattern Recognition(pp. 89-93).

[13] Latifoglu, F., Polat, K., Kara, S., Gunes, S. (2008). Medical diagnosis of atherosclerosis from carotidartery Doppler signals using principal component analysis (PCA), k-NN based weighting preprocessingand 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. IEEEtransactions on neural networks, 13(1), 3-14.Miller, A. S., &Blott, B. H. (1992). Review of neural network applications in medical imaging and

[14] signal processing. Medical and Biological Engineering and Computing, 30(5), 449-464

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

[16] Rajini N. H. &Bhavani, R. (2011). Classification of MRI Brain Images using k- Nearest Neighbor andArtificial Neural Network, IEEE-International Conference on Recent Trends in InformationTechnology, ICRTIT.

[17] Rajendran, P., Madheswaran, M., &Naganandhini, K. (2010). An improved pre-processing techniquewith image mining approach for the medical image classification. In Computing Communicationand Networking Technologies (ICCCNT), 2010 International Conference on(pp. 1-7). IEEE.

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

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

[20] Suguna, N. &Thanushkodi, K. (2010). An improved k-nearest neighbour classification using geneticalgorithm.*I*nternational journal of computer science issues (IJCSI), Volume 7,issue 4,No2.

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

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

[23] Zaidi, H., & El Naqa, I. (2010). PET-guided delineation of radiation therapy treatment volumes: asurvey of image segmentation techniques. European journal of nuclear medicine and molecularimaging, 37(11), 2165-2187.

[24] S. Vijaykumar and S. Saravanakumar, "Future Robotics Memory Management", Advances in DigitalImage Processing and Information Technology, pp. 315–325, 2011.

[25] Vijaykumar, S., S.G. Saravanakumar, Dr. M. Balamurugan: "Unique Sense: Smart Computing Prototype", Procedia Computer Science, Issue. 50, pp. 223 – 228, 2015.