A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases

Syed Zahid Hassan and Brijesh Verma
School of Computing Sciences
Central Queensland University, Australia
z.hassan@cqu.edu.au, b.verma@cqu.edu.au

Abstract

This paper presents a novel hybrid data mining approach for knowledge extraction and classification in medical databases. The approach combines self organizing map, k-means and naïve bayes with a neural network based classifier. The idea is to cluster all data in soft clusters using neural and statistical clustering and fuse them using serial and parallel fusion in conjunction with a neural classifier. The approach has been implemented and tested on a benchmark medical database. The preliminary experiments are very promising.

1. Introduction

Over the past several years, there has been an influx amount of medical data generated and subsequently collected in the health information systems. According to Damien McAullay [1], "there are 5.7 million hospitals admissions, 210 million doctor's visits, and a similar number of prescribed medicines dispensed in Australia annually.

Lately, this abundance of medical data has resulted in a large number of concerted efforts to inductively discover 'useful' knowledge from the collected data, and indeed interesting results have been reported by researchers [1], [2], [3]. However, despite the noted efficacy of knowledge discovery methods— known as Data Mining (DM) algorithms— the challenge facing healthcare practitioners today is about data usability and impact—i.e. the use of 'appropriate' data mining algorithms with the right data to discover knowledge that can lead towards medical decision-making.

Notably, recent advances in computational intelligence and data mining techniques have made it possible to transform any kind of raw data into high level knowledge. However, the main problem is the limitations associated with individual techniques that affect the overall classification results. Consequently, the need of a hybrid data mining approach is widely recognized by the data mining community. The numbers of hybrid data mining endeavours have been initiated in the past however the developed hybrid approaches are mainly based on combination of various classifiers. They do not take full advantage of soft clustering and learning abilities offered by statistical and intelligent clustering and fusion techniques.

This paper proposes a novel hybrid data mining approach which is an effective combination of statistical and intelligent techniques in conjunction with a neural fusion, in order to utilize the strengths of each individual technique and compensate for each other's weaknesses.

This paper is divided into five sections. Section II, presents a review of data mining algorithms and their applications in medical domain. Section III discusses the proposed approach in detail. Section IV presents the preliminary experiments and results, evaluates the performance of the proposed approach by performing quantitative analysis and discussion on the results achieved. The conclusions and future directions are presented in Section V.

2. Literature review

The statistical, intelligent and hybrid algorithms have been used for data mining in medical domain. The unsupervised learning algorithms have drawn prominent attention in medical data classification due to the nature of its problem domain, where the databases consist of complex, large and unlabelled data samples. The success of various unsupervised learning algorithms, such as self-organizing map (SOM), k-means, k-NN, etc. have been reported in various medical data mining applications ranges from feature selection, extraction, classification to data

0-7695-2976-3/07 $25.00 © 2007 IEEE
DOI 10.1109/ISDA.2007.48

visualization. For example, self-organizing map (SOM) is used to identify the clusters in breast cancer diagnosis [4], to predict biopsy outcomes [5] and to model selection of mammography features [6].

The various hybrid models have also been implemented and tested in different application domain [7], [8], [9], [10], and [11]. A neural network and fuzzy logic based hybrid models have been widely reported [7]. The model presented in [7] focuses on an integration of the merits of neural and fuzzy approaches to build intelligent decision-making systems. It has the benefits of both 'neural networks' like massive parallelism, robustness, and learning in data-rich environments, and 'fuzzy logic', which deal with the modeling of imprecise and qualitative knowledge in natural/linguistic terms as well as the transmission of uncertainty are possible through the use of fuzzy logic. This hybrid approach has shown a high rate of success when applied in various complex domains of medical applications [8], [12], and [13]. For example, in [12], a neural fuzzy approach is presented to measure radiotracers in vivo. In this application, fuzzy logic is the core part of the system, which deals with the modeling of imprecise knowledge (image degradation) due to the photon scattering through the collimated gamma rays.

The Neural Networks and Evolutionary Algorithms (NN-EA) hybrid approach has also received prominent attention in medical domain [9]. In general EA is used to determine the NN weights and architecture. In most cases NN are tuned (not generated) by the EA, but there are also appreciation when NN are tuned as well as generated by EA. In [9], EA-NN hybrid approach is presented to diagnose breast cancer: benign and malignant. The other hybrid combination, Fuzzy Logic and Genetic Algorithms (FL-GA) has also been deployed successfully in various control engineering applications and complex optimisation problems [7]. GA is used for solving fuzzy logical equations in medical diagnostic expert systems [13].

The another interesting hybrid combination examined in the literature is Decision Trees and Fuzzy Logic (DT-FL) combination, where fuzzy logic is used to model uncertainty and missing decision attributes before these attributes are subjected to decision trees for classification and diagnosis tasks [14]. With regards to medical applications, this approach showed some great accuracy in diagnosing coronary stenosis [11] and segmentation of multi-spectral magnetic resonance images (MRI) [14]. Some authors have also proposed the combination of Decision Trees and Evolutionary Algorithms (DT-EA). In this hybrid approach, decision trees are generally used to extract relevant features from large datasets whereas EA algorithms are used to generalize the data [15]. This approach overcomes the limitation of the EA which requires more time to process complex tasks. In [15], an evolutionary modular MLP is combined with the ID3 decision tree algorithm, for the staging of cervical cancer.

The hybridization of Fuzzy Decision Tree (FDT) and Neural Network has also been investigated [16]. With the induction of fuzzy decision trees, it happened to perform well and generate comprehensive results, but learning accuracy was not very good. A new hybrid methodology with neural networks based FDT weights training was proposed in [16], which lead to the development of hybrid intelligent systems with higher learning accuracy. This approach has been successfully tested on various databases and interesting results have been reported.

3. Proposed approach

This section describes the proposed hybrid approach in detail. The proposed hybrid approach deploys the variety of clustering algorithms on the training data sets and then combines clusters produced by them in two forms: parallel data fusion and serial data fusion, as shown in Figures 1 and 2. The parallel fusion incorporates a multilayer perceptron for learning of soft clusters and classifies them into appropriate classes, as demonstrated in Figure 3. In serial approach; we first monitored the individual classifier performance and then trained each classifier with the classified patterns of other classifiers and noticed its affect on overall system performance. The algorithms used vary in their method of search and representation which ensures to achieve diversity in the errors of the learned models.

More specifically, the two types of hybrid combinations are investigated in this paper: parallel hybrid data mining and serial hybrid data mining, whereby each hybrid combination consists of four parts: A) input data B) hybrid clustering C) fusion of clusters and D) data visualization, as shown in Figures (1) and (2).

3.1. Input data

The input data contain raw data as well as extracted features which are used as an input to the clustering algorithms. The input data are normalized between 0-1.

3.2. Clustering algorithms

Three learning algorithms SOM, k-mean and naïve bayes have been combined in conjunction with a MLP.

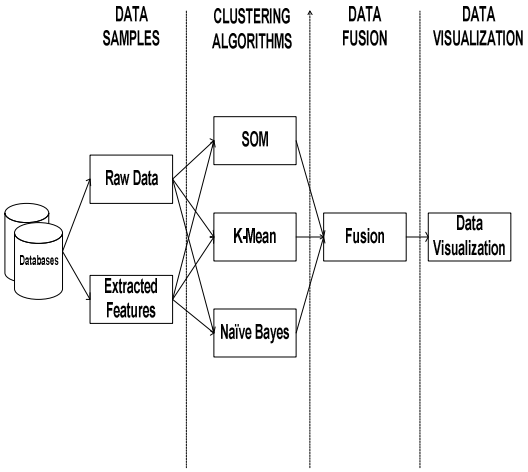

Figure 1. Parallel hybrid data mining approach

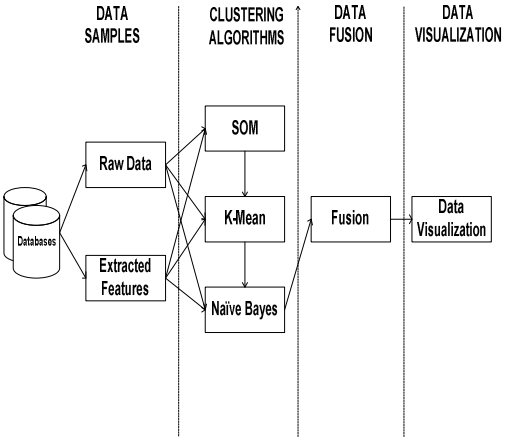

Figure 2. Serial hybrid data mining approach

SOM is a self organising map based on Kohonen neural network. SOM consisted of 16 neurons partitioned in a single layer in a 2-D grid of 4 x 4 neurons. We construed and assigned the random reference input vectors (neuron weights) to each partition. For each input, the Euclidean distance between the input and each neuron was calculated.

The reference vector with minimum distance is identified. After the most similar case is determined, all the neighbourhood neurons, connected with the same link, adjust their weight with respect to the reference vector to form a group in two dimensional grid. The whole process is repeated several times, decreasing the amount of learning rate to increase the reference vector, until the convergence is achieved. The SOM visualization offers the clear partition of data into discernable clusters.

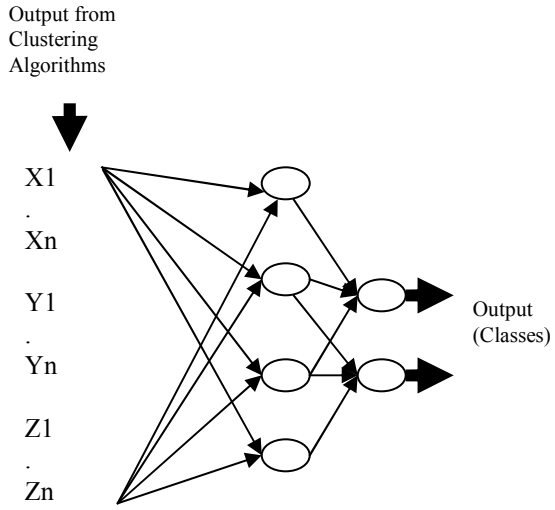

Figure 3. Neural fusion

The k-mean is a well known clustering algorithm. In k-mean algorithm, we randomly partitioned the input data into k-cluster centers along with its all closest features. With each input feature, it calculated the mean point of each feature and constructs a new partition by associating data-entities to one of the k clusters. Cluster features are moved iteratively between k clusters and intra-and-inter-cluster similarity. Distances are measured at each move. Features remained in the same cluster if they were closer to it otherwise in new cluster. The centers for each cluster are recalculated after every move. The convergence achieved when moving object increased intra-cluster distances and decreases inter-cluster dissimilarity.

Naïve bayes clustering is based on probability distribution. It accepts raw data or features as input and creates soft clusters which are later combined with the outputs from other two clustering techniques mentioned above and passed to the fusion module.

3.3. Data fusion

The working of decision fusion model can be understood by its fusion hypothesis: which assumes reliable the cluster is for decision making.

The outputs of clustering algorithms are combined using serial fusion and parallel fusion. The parallel fusion is based on a multilayer perceptron (MLP) as

shown in Figure 3. A simple majority voting method is also used and compared with MLP.

3.4. Data visualization

This process involves the designing of a data mining visualization model, coupled with all data mining clustering methods and generates the knowledge which is derived by the data mining inference engine, in the form of charts, graphs and maps. These coupling of DM algorithms and visualization algorithms provide added value. The visualization techniques may range from simple scatter plots to histogram plots over parallel to two dimensions coordinates.

4. Implementation and experimental results

The proposed approach has been implemented in order to evaluate the performance and accuracy. The experiments were conducted on a benchmark dataset. The dataset and experimental results are described below.

4.1. Benchmark database

The dataset of digital mammograms is used in this research and it is taken from Digital Database for Screening Mammography (DDSM) established by University of South Florida. The main reason to choose DDSM for the experiment purposes is that it is a benchmark dataset so the final results can be compared with published results by other researchers. The DDSM database contains approximately 2,500 case studies, whereby each study includes two images of each breast, along with some associated patient information (age at time of study, breast density rating, subtlety rating for abnormalities, keyword description of abnormalities) and image information (scanner, spatial resolution etc). The database contains a mixture of normal, benign, benign without call-back and cancer volumes selected and digitized. Images containing suspicious areas have associated pixel-level information about the locations and types of suspicious regions.

The data set consists of six features (measurements) from 300 mammograms cases: 150 benign and 150 malignant. The features include: Patients Age, Density, Shape, Margin, Assessment Rank and Subtlety.

4.2. Experiments results and discussion

The experimental results are presented below in Tables 1 and 2.

Table 1
Results showing the improvement in classification accuracies

Algorithm	Classification Error [%]	Root Mean Square Error	Classification Accuracy [%]
SOM	12	0.2777	88
K-Mean	16	0.2433	84
Naïve Bayes	10	0.3022	90
Proposed Hybrid Approach	7.6	0.2572	92.3077

Table 2
Detailed accuracy by classes
TP = true positive rate; FP = false positive rate and F-Measure= frequency measure over class accuracy

	Classes	TP Rate	FP Rate	F-Measure
Individual Algorithm	Benign (SOM)	0.8	0.04	0.87
	Malignant (SOM)	0.96	0.2	0.88
	Benign (Naïve Bayes)	0.94	0.14	0.90
	Malignant (Naïve Bayes)	0.86	0.06	0.89
Proposed Hybrid Approach	Benign	0.88	0.038	0.92
	Malignant	0.96	0.115	0.92

From the comparative results shown in Table I, it is observed that the proposed hybrid approach, combination of statistical and intelligent techniques provides better results than the stand alone individual technique. It is also noticed that the proposed approach outperforms all individual approaches in all main output categories (see Table I): classification accuracy, misclassification accuracy and error rates. Out of the total of 100 digital mammogram cases of the test dataset, SOM made 12% misclassifications; K-Mean made 16% misclassifications, Naïve bayes misclassified 10% cases and proposed approach made 7.6923% misclassifications. This corresponds to

classification accuracies achieved by SOM, k-Mean, Naïve Bayes and proposed approached are 88%, 84%, 90% and 92.307%, respectively.

The experiments were also performed comparing the accuracies of algorithms by individual class: benign and malignant. For each class, the ROC analysis attributes, such as TP rate, FP rate, and F-measure, are measured with particular algorithm as shown in Table II. It is noticeable that the attributes frequency measures for both classes benign and malignant are quite high with the proposed hybrid approach.

We created a Confusion matrix to evaluate individual classifier performance by displaying the correct and incorrect pattern classifications. Typical Confusion Matrix can be represented as:

Confusion Matrix

a	b	<-------- Classified as
x1	x2	a = Malignant
y1	y2	b = Benign

Where row (x1 and x2) represents the actual patterns and column (x1 and y1) represents the classified patterns for class a (Malignant). The difference between the actual patterns and the classified patterns can be used to determine the performance of a classifier. To explicate it further, we draw the Confusion Matrix for each classifier to evaluate how many patterns in a given class are classified correctly/incorrectly. Note: There were 300 mammogram cases were used: 200 cases for training purposes and 100 for testing purposes.

SOM Confusion Matrix

a	b	<-------- Classified as
48	2	a = Malignant
10	40	b = Benign

This SOM classifier successfully classified 88 cases out of 100 cases presented. The row values (48, 2) are the actual cases for the class malignant, and row values (10, 40) represent the actual class benign. However, the classified outputs are represented by column a (48, 10) and column b (2, 40). The comparison of these rows and columns, between actual pattern and classified patterns, can provide interesting insights. For instance: for the malignant class accuracy, we notice that the original malignant patterns were (48, 2) and the classifier indicates (48, 10). Thus, it classified 48% cases correctly as a malignant class and misclassified 2 cases. It is also noticeable that those two patients will be given clear when they were supposed to be treated like a cancer patients. Similarly, for the benign class accuracy, the actual cases are (10, 40) and whereas the classifier indicates (2, 40). The 40% cases were classified correctly as a class benign and 10% cases were misclassified. In this scenario, those 10 patients who are not the victim of cancers will be treated like a cancer patient despite it being the opposite scenario. However, the overall outcome is much more favourable: 48% classified correctly as a malignant class and 40% classified correctly as a benign class.

K-Mean Confusion Matrix

a	b	<-------- Classified as
38	11	a = Malignant
5	46	b = Benign

By applying the above-mentioned confusion matrix method on the K-mean classifiers, the 38% cases were classified correctly as a class malignant (11 cases were misclassified) and 46% cases classified correctly as a class benign (misclassified 5 cases), overall achieved 84% classification accuracy.

Naïve Bayes Confusion Matrix

a	b	<-------- Classified as
43	7	a = Malignant
3	47	b = Benign

The naïve bayes classified 43% and 47% cases correctly as a class malignant and benign respectively, with the ratio of 2 misclassified cases of a class malignant and 3 cases for a class benign, overall computed 90% accuracy.

From the decision-making perspective, it's also noticeable that by fusing the outputs of all clustering algorithms, based on simple voting method we can get the final clusters which are more accurately classified. In this voting approach, the winner cluster is the one with the most votes from the classifiers.

The experiments show that proposed hybrid data mining approach; is useful for the analysis of clinical parameters and their combinations for the cancer diagnosis. More experiments are still in progress with different hybrid combinations.

5. Conclusions

We have presented a novel hybrid data mining approach, an approach that combines intelligent and statistical data mining algorithms such as SOM, K-Means and Naïve Bayes in conjunction with a serial fusion and a multilayer perceptron based parallel fusion. The approach was implemented and tested on DDSM benchmark database. The proposed hybrid approach achieved over 92% classification accuracy on test set which is very promising. The proposed approach is also able to visualize the data which helps in interpretation of the results.

The results presented in this paper were obtained from serial fusion as shown in Figure 2. In our future research, we are planning to investigate parallel fusion using neural based fusion.

6. References

[1] Damien McAullay, Graham J. Williams, Jie Chen, Huidong Jin (2005), "A Delivery Framework for Health Data Mining and Analytics". In proceedings of the Twenty-eighth Australasian conference on Computer Science - Volume 38, Pages: 381 – 387. Newcastle, Australia.

[2] Babic, A and P. Kokok (1999), "Knowledge Discovery for Advanced Clinical data Management and Analysis". Medical Informatics Europe'99, Ljubljana, Amsterdam: IOS Press.

[3] S. S. R. Abidi, K. M. Hoe and A. Goh, (2001), "Analyzing Data Clusters: A Rough Set Approach to Extract Cluster Defining Symbolic Rules". Fisher, Hand, Hoffman, Adams (Eds.) Lecture Notes in Computer Science: Advances in Intelligent Data Analysis, 4th Intl. Symp of Intelligent Data Analysis IDA'01, Berlin:Springer Verlag.

[4] Markey M. K, L, J. Y., Tourassi, G. D, Floyd, C. E., (2003). "Self-organizing Map for Cluster Analysis of a Breast Cancer Database". Artificial Intelligence in Medicine 2003; Vol 2, pp:113-27.

[5] Chen, D., Chang, R. F., and Huang, Y. L, (2000), "Breast Cancer Diagnosis Using Self-organizing Map for Sonography". Ultrasound in Medicine Biology 2000. Vol 26, pp:405–11.

[6] West, D., and West, V. (2000), "Model Selection for a Medical Diagnostic Decision Support System: A Breast Cancer Detection Case". Artificial Intelligence in Medicine 2000;. Vol 20, pp:183–204.

[7] George, D. and Derek, A. (2004), "Linkens: Adaptive Systems And Hybrid Computational Intelligence In Medicine". Artificial Intelligence in Medicine. Vol 3, pp: 151-155.

[8] Chin Te, C., L. Win Li, et al. (1998), "A combination of neural network and fuzzy logic algorithms for adaptive control of arterial blood pressure". Biomedical Engineering, Applications Basis and Communication, 10 (3), pp. 139-150.

[9] Hussein A. Abbass, (2002), "An Evolutionary Artificial Neural Networks Approach For Breast Cancer Diagnosis". Artificial Int. in Medicine, Vol. 25, Issue. 3, pp. 265-281

[10] Wang, Z. S.; Chen, J. D. Z., (2000), "Robust ECG R-R Wave Peak Detection Using Evolutionary-Programming-Based Fuzzy Inference System (EPFIS) And Its Application To Assessing Brain-Gut Interaction". 1st International Conference on Advances In Medical Signal and Information Processing, (IEEE Conference Publication No. 476), pp. 265-274

[11] Cios, K.J., Goodenday, L.S., Sztandera, L.M., (1994), "Hybrid Intelligence System For Diagnosing Coronary Stenosis. Combining Fuzzy Generalized Operators With Decision Rules Generated By Machine Learning Algorithms". IEEE Engineering in Medicine and Biology Magizine. Vol 13, Issue 5 , pp. 723–729.

[12] Clarke, L. P. and Q. Wei (1998), "Fuzzy-Logic Adaptive Neural Networks For Nuclear Medicine Image Restorations". Proceedings of the 20th AnnualInternational Conference of the IEEE Engineering in Medicine and Biology Society, Vol.20, Biomedical Engineering Towards the Year 2000 and Beyond, IEEE Piscataway NJ USA, pp. 3384-3390

[13] Haiming, Q., D. J. Tyler, et al. (1999), "Neurofuzzy Adaptive Controlling Of Selective Stimulation For FES: A Case Study". IEEE Transaction on Rehabilitation Engineering. 7(2), pp. 183-192.

[14] Jzau-Sheng Lin, Kuo-Sheng Cheng, Chi-Wu Mao, (1996), "Segmentation of Multispectral Magnetic Resonance Image Using Penalized Fuzzy Competitive Learning Network", Computing and Biomedicine. Vol 29 (4) pp.314-326

[15] Pabitra Mitra, Sushmita Mitra , et al., "Evolutionary Modular MLP with Rough Sets and ID3 Algorithm for Staging of Cervical Cancer", Neural Computing & Application Vol. 10, Issue 1, pp 67-76.

[16] Tsang, X.Z. Wang, D.S. Yeung (2000), "Improving Learning-Accuracy of Fuzzy Decision Trees by Hybrid Neural Networks". In IEEE Transactions on Fuzzy Systems, Vol. 8, Issue 5, pp: 601 - 614.

[17] Neat, G. W. and Kaufman, H, et al. (1998). "A Hybrid Adaptive Control Approach For Drug Delivery Systems". Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.88CH2566-8). PP: 25-31. IEEE New York NY USA.

[18] Pattaraintakorn, P.; Cercone, N.; Naruedomkul, K. (2005), "Hybrid Intelligent Systems: Selecting Attributes For Soft-Computing Analysis". 29th Annual International Computer Software and Applications Conference(COMPSAC'05). Vol 1, Issue 2, pp: 319 – 325.

[19] Chee-Peng Lim; Jenn-Hwai Leong; Mei-Ming Kuan; (2005), "A Hybrid Neural Network System For Pattern Classification Tasks With Missing Features". Pattern Analysis and Machine Intelligence, IEEE Transactions on Volume 27, Issue 4, April 2005 Page(s):648 – 653.

[20] Goonatilake, S. and Khebbal, S. (1995), *Intelligent Hybrid Systems*. John Wiley and Sons, Chichester.