A Novel Approach for Classifying Medical Images Using Data Mining Techniques

J. Alamelu Mangai, Jagadish Nayak and V. Santhosh Kumar

***Abstract*—**The ever increasing amounts of patient data in the form of medical images, imposes new challenges to clinical routine such as diagnosis, treatment and monitoring. Hence research in medical data mining is the state of the art. In this paper, a novel approach for automatic classification of fundus images is proposed. The method uses image and data pre-processing techniques to improve the performance of machine learning classifiers. Further a discretization method is proposed to improve the accuracy of the classifiers. Experiments were done on retinal fundus images using the proposed method on three classifiers Naïve Bayes NB, k nearest neighbor kNN, and support vector machine SVM. Results in terms of classification accuracy and area under the Roc curve AUC show that NB outperform the other classifiers as per the proposed method.

***Keywords*—** Medical image mining, feature selection, discretization, NB, kNN, SVM, AUC .

I. Introduction

DATA MINING is the process of extracting implicit, non-trivial knowledge from huge data repositories. This technique although not fully matured, has gained high popularity and applications in various fields like computer security, financial applications, electric load profiling, customer relationship management, e-commerce, fault diagnosis, medicine, etc. Medical imaging refers to a number of non-invasive methods of looking inside the body. It enables a doctor to diagnose, treat and cure patients without causing harmful side effects. Medical images are more important assets of clinical history and their analysis is essential in modern medicine. Fundus images are a class of medical images. The retinal fundus images give details of the inner lining of the eye which includes sensory retina, the retinal pigment epithelium, Bruch's membrane and the choroid. Medical image mining using data mining techniques helps to automate clinical diagnosis and research in this direction is the state of the art.

Classification is a supervised data mining task of assigning a test data to one of pre-defined categories. The Naïve Bayesian (NB) classifier is one of the top ten classification methods and has been widely used for medical image classification. Despite the fact it assumes that all attributes are independent, it has several advantages like simplicity, computationally efficient, requires relatively little data for training, do not have lot of parameters and is robust to missing and noisy data [1]. As it uses all parameters in decision making, it is appealing to physicians, as the decision seems to be natural. The performance of NB classifiers can be improved through discretization. Many studies in data mining and knowledge discovery have shown that induction tasks can benefit from discretization. It is the process of transforming continuous valued parameters to discrete intervals which leads to improved predictive accuracy [2]. Apart from the algorithmic requirements, discretization also helps in increasing the speed and accuracy of induction algorithms. It reduces classifier induction time, makes the results of the induced classifier shorter, compact and easier to understand than those generated using continuous features.

Feature Selection is an essential data pre-processing step, for getting quality mining results from quality data. If information is irrelevant or redundant then knowledge discovery during training phase is more difficult. So, feature selection prior to learning is more beneficial. Reducing the dimensionality of data also helps in reducing the hypothesis space, allows the algorithms to operate faster and even accuracy can be improved in some cases. This is essential for some classifiers like NB, kNN and SVM as, they do not perform implicit feature selection as decision trees. In this paper, machine learning classifiers NB, kNN and SVM are modeled to automatically classify retinal fundus images. The performance of these classifiers is further improved through feature selection and a discretization method, which unlike other existing methods, automatically identifies the number of intervals a parameter needs to be discretized. Section 2 is a survey of the related work, Section 3 presents the proposed model for medical image classification, Section 4 is the detail of experimental setup, results and discussion and Section 5 is the conclusion.

II. Related Work

The NB classifier has been successfully applied for research on medical data. It out-performed six other classification models on eight diagnostic problems [3]. In [4] the authors emphasize that NB is one of the most effective and efficient classification algorithms, through an empirical comparison of

J. Alamelu Mangai is with Birla Institute of Technology & Science Pilani, Dubai Campus, Dubai, 345055, UAE (corresponding author's phone:+971503987928; fax: +97144200844; e-mail:mangai@bits-dubai.ac.ae)

Jagadish Nayak is with Birla Institute of Technology & Science Pilani, DubaiCampus,Dubai,345055,UAE (e-mail: jagadishnayak@bits-dubai.ac.ae).

V. Santhosh Kumar is with Birla Institute of Technology & Science Pilani, DubaiCampus,Dubai,345055,UAE (e-mail: santhoshkumar@bits-dubai.ac.ae)

NB with five popular classifiers (Logistic Regression, nearest neighbor, Decision Tree, Neural Network and Rule Based) on 15 medical data sets. The classifiers are compared based on the area under the Receiver Operating Characteristics (ROC) curve. Also, previous literature says that discretization has the potential to improve classification performance. It has been also used as a variable selection method [5], and the classifier with largest gain in performance is NB The requirements of effective discretization for NB differ from that of other machine learning algorithms. The impact of nine discretization methods namely equal width, equal frequency, fuzzy, entropy minimization, iterative, proportional k-interval, lazy, non-disjoint and weighted proportional on NB is compared [6]. A new method by combining the advantages of non-disjoint and weighted proportional k-interval is proposed to reduce the NB's average classification error. A hybrid feature selection algorithm (CHI-WSS) that combines the filter approach with wrapper subset evaluator is used in [7]. Also experimental results have shown that with minimal description length (MDL) discretization method, the NB classifier seems to be the best performer compared to its variants and other non-NB statistical classifiers. Feature selection and exudates pixel classification using a Naïve Bayes classifier on 6 features extracted from retinopathy images is proposed in [8]. The proposed method in this paper for medical image classification improves the performance of NB using a discretization method which automatically identifies the number of bins each feature needs to be discretized.

III. Proposed Method Of Medical Image Classification

The proposed architecture for classifying medical images is shown in Fig. 1. A detailed description of each phase is described below.

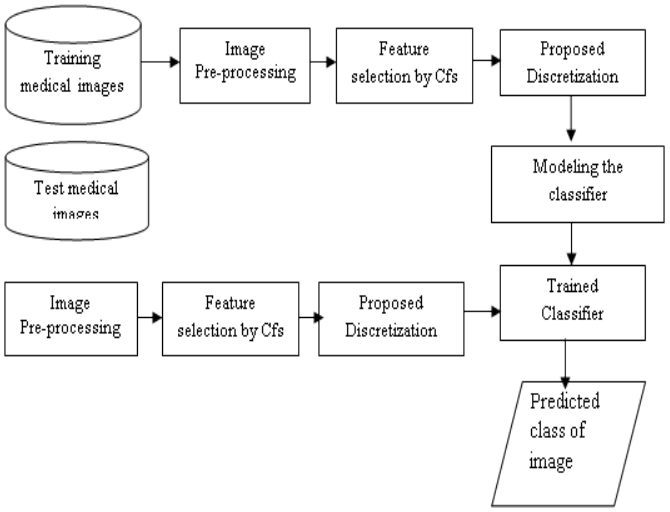

Fig.1 The Proposed Architecture for Medical Image Classification

A. Image Preprocessing

Due to the non-uniformity in the color distribution of fundus images among the different subjects, each image is preprocessed for having uniform distribution of gray levels. The two major reasons for this non uniformity are non uniform illumination and variation in the pigment color in the eye. This is corrected by applying adaptive histogram equalization [9] to the image before it processed further. This technique adjusts the local variation in contrast by increasing the contrast in lower contrast area and lowering the contrast in high contrast area.

B. Feature Extraction

Features are extracted from the retinal fundus images of size 576x720 pixels. Localized statistical features are computed by dividing the image into sub-images. More localization of the image will yield accurate features. In this proposed system fundus images are divided into sub images of size 36x90 pixels, which will result in a feature vector of size 128. We have extracted four statistical features such as mean, variance, skewness and kurtosis. Total size of the feature vector is 512. These features are computed as follows.

a. $Mean\ \overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ where N is the number of data points and n is the order of the moment.

b. $Variance = \frac{\sum_{i=1}^{N} \left(x - \overline{x}\right)^2}{N}$

c. $Skewness = \frac{1}{N}\left(\frac{\left(x - \overline{x}\right)}{\sigma}\right)^3$ and

d. $Kurtosis = \frac{1}{N}\left(\frac{\left(x - \overline{x}\right)}{\sigma}\right)^4 - 3$

After feature extraction, each image is transformed into a feature vector and are represented as $\{f_1, f_2, f_3,\ldots,f_n, Image\ Category\}$ where each f_i is a continuous feature and *Image category* is the pre-defined category of the image.

C. Feature Selection

With no quality data, there is no quality mining results. So, in order to reduce the hypothesis space for the classifiers and to reduce the average classification error, feature selection is performed using CfssubsetEval, a correlation based method [10]. This method evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred.

D. Feature Discretization

The image features are then discretized by a series of split and merge as described below.

Input: Image Feature Vectors *IMFV*, Fundus Image Categories *C*, the threshold *B*, the inconsistency threshold within an interval *Incon*.

Output: Discretized Image Feature Vectors, *DIMFV*.

Method:

a. Repeat the following for each image feature *f* in *IMFV*

a.1 Sort *f* into ascending order.

a.2 Establish the initial cut points Cut_1, wherever two consecutive feature values differ in class label.

a.3 For each bin in Cut_1 do

a.3.1. Find its majority class.

a.3.2. If the number of values in two consecutive bins belonging to their respective majority classes is less than the threshold *B*, then merge them.

a.3.3. Save the new cut points in Cut_2.

a.4 for each bin in Cut_2 do

a.4.1 Find its inconsistency as

$I = (a-b)/a$ where *a* and *b* are the size of bin and the number of values of majority class in the bin, respectively.

a.4.2 If I < *Incon* for any two consecutive bins, merge them, which forms the final set of cut points Cut_3.

a.5 Replace each continuous value of image feature by a corresponding bin label.

b. Output the discretized image feature vector *DIMFV*.

c. Stop.

E. Machine learning Classifiers

Some of the top ten classification algorithms of data mining are Naïve Bayes (NB), k nearest neighbor kNN and support vector machine classification. To classify a new object, the NB initially has a hypothesis that the given data belongs to a particular class. It then calculates the probability of the hypothesis to be true using the Bayes theorem in statistics. If *X* is the object to be classified the Bayes theorem calculates the conditional probability of it belonging to one of the classes C_1, C_2, C_3 etc, $P(C_i \mid X)$ as $P(C_i \mid X) = [P(X|C_i)P(C_i)]$ where, $P(X \mid C_i)$ is the probability of object X belonging to class C_i. $P(X \mid C_i)$ is the probability of obtaining the attribute values *X*, if we know that it belongs to class C_i. $P(C_i)$ is the prior probability of any object belonging to class C_i. Once these probabilities are computed for all classes, then *X* is simply assigned to the class with the highest conditional probability.

The kNN classifier predicts the class of a test object by calculating its k nearest neighbors. It then applies a simple majority voting on these nearest neighbors to predict the class of the test object. The Support vector machine SVM classifiers use the training set to construct a hyper plane that separates the classes. Separating the classes with a large margin minimizes the bound on the expected generalization error. Among these classifiers NB has been widely used in medical domain. Although SVM exhibits good performance in text classification, it is not deeply explored in medical domain. This paper compares the impact of the proposed method on these classifiers for medical images

IV. Experimental Results And Discussion

The experiments were done on retinal fundus images taken from Kasturba Medical College, India. For analysis, 61 normal fundus images and 32 very severe images were considered. Samples of these two categories of images are shown in Fig. 2 and Fig. 3.

Fig. 2. Sample normal fundus images

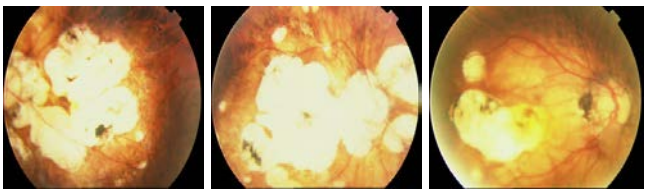

Fig. 3. Sample very severe fundus images

The image preprocessing and feature extraction was done using a MATLAB program. After image preprocessing using adaptive histogram equalization, 512 features were extracted from the images. Each image is then transformed into a 513-dimensional feature vector appended with its category. The files with image features are stored in arff format called attribute relation file format, supported by Weka. Fig. 4, shows a portion of the arff file with all 512 features extracted. where the last column is the image category namely 1 = normal images and 0 = very severe images.

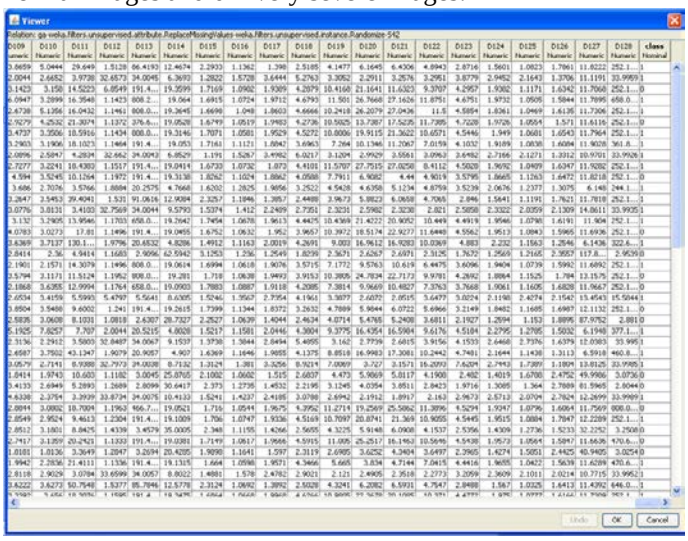

Fig. 4. Image Feature vectors with all features extracted in arff

As a dimensionality reduction step, the irrelevant features which may degrade the classification performance are removed using the Cfs feature selection method in Weka. The description of the image file after feature selection is shown in Fig. 5.

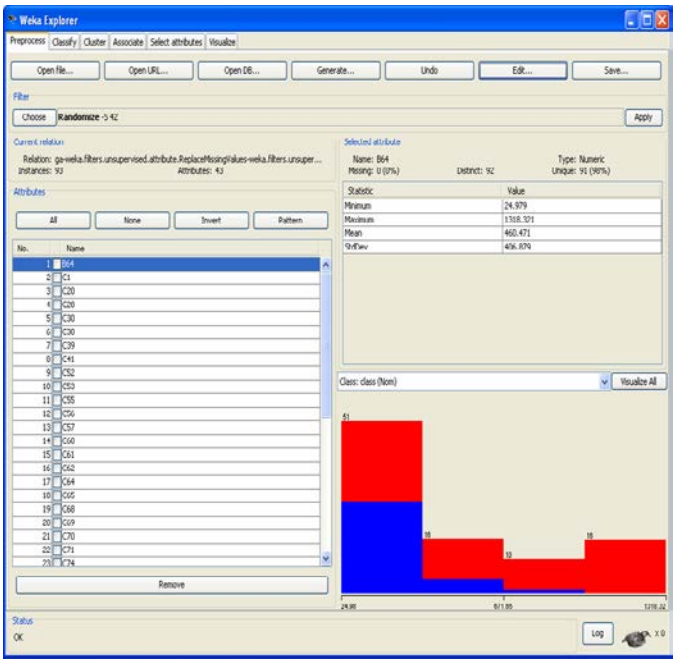

Fig. 5. The Image File after Feature Selection

As shown in Fig. 5, the number of features selected by the Cfs method is only 42 out of 512 original numbers of features. Since the induction time of classifiers with all features being numeric will be more, the features are discretized by the proposed method. This was implemented using a Java program. A portion of the file after discretization in arff is shown in Fig 6.

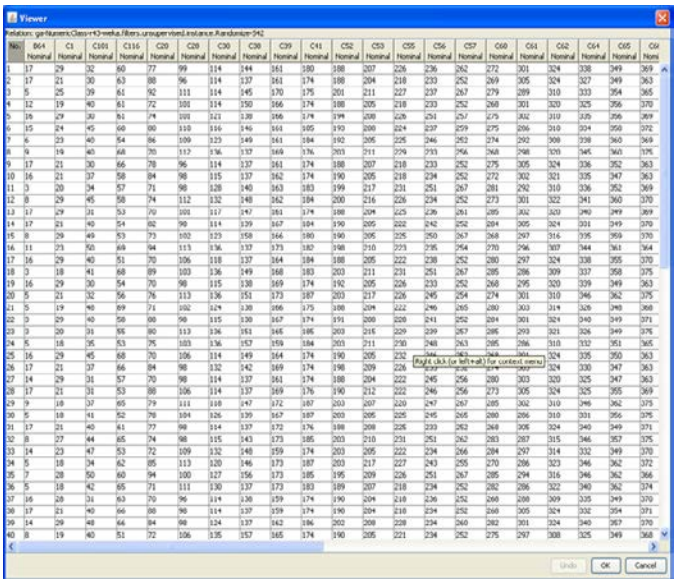

Fig. 6 Image File after Discretization by proposed method.

As shown in Fig. 6, all features which were continuous are transformed into nominal (discrete). In order to automatically generate the labels, numbers are used as discrete labels. The classifiers are then modeled with these discrete features using 10-fold cross validation. They are evaluated using two measures namely percentage classification accuracy and area under the receiver operating characteristics curve (AUC). Table 1 shows the classification accuracy of NB, kNN and SVM on the discretised image file using Weka. The SVM classifier is modeled using polykernel and the value of k for kNN is chosen using cross validation.

TABLE I
COMPARATIVE ANALYSIS BASED ON CLASSIFICATION ACCURACY

Classifier	All numeric features	After Cfs	After Cfs & SB	After Cfs & MDL	After Cfs & proposed discretization
NB	81.72	97.85	93.55	100.00	100.00
kNN	34.41	72.04	77.42	93.55	100.00
SVM	91.40	91.40	92.47	97.85	100.00

The proposed discretization method is compared with two other existing discretization methods namely simple binning (SB) and entropy based minimal description based (MDL). The SB algorithm is run with 10 number of intervals for each feature. Comparing the classification results of all three classifiers using features discretized by all three methods show that, the proposed method of discretization improves their performance.

Receiver Operating Characteristics (ROC) is a technique for comparing the performance of classifiers and is commonly used in medical image decision making. The classifiers are compared using the area under the ROC curve [11]. The AUC value is always between 0 and 1.0. A realistic classifier has AUC above 0.5 and AUC being 1 for a perfect classifier. Table 2 is a comparative analysis of all three classifiers using features discretized by all three methods based on the AUC.

TABLE II
COMPARATIVE ANALYSIS BASED ON AREA UNDER THE CURVE, AUC

Classifier	All numeric features	After Cfs	After Cfs & SB	After Cfs & MDL	After Cfs & proposed discretization
NB	0.94	0.99	0.99	1	1
kNN	0.53	0.82	0.92	0.99	1
SVM	0.88	0.89	0.89	0.97	1

Comparing the results of Table 1 and Table 2, it can be observed that

- Features selected by Cfs significantly improve the classification accuracy for NB and kNN.
- The proposed discretization method improves the predictive accuracy of all three classifiers.
- Based on AUC, the proposed method of medical image classification using NB outperforms kNN and SVM.

V. CONCLUSION

As the volume of medical image data is exponentially increasing over time, patient diagnosis and maintaining clinical records is quite challenging. In this paper, a feature selection and discretization method for automatic classification of retinal fundus images is proposed. Three classifiers NB, kNN and SVM were modeled with these

discrete features. Experimental results show that the feature selection and discretization algorithms significantly improve the classification accuracy of all three classifiers. Comparing the classifiers using AUC shows that NB outperforms the other two classifiers and is better suited for medical image classification. Our future work is to implement this method on other classes of medical images.

REFERENCES

[1] K. M. Al Aidaroos, A. A. Bakar and Z. Othaman, "Naïve Bayes variants in classification learning." in *Proc. Intl. Conf. on Information Retrieval & Knowledge Management*, Malaysia, 2010, pp. 276 – 281.

[2] H. Liu, F Hussain, C L Tan and M. Dash, "Discretization : An enabling technique", *Data Mining and Knowledge Discovery*, vol. 6, issue. 4, pp. 393–423, Oct. 2002.

[3] I. Kononenko, I. Bratko and M. Kukar, "Application of machine learning to medical diagnosis", *Machine Learning and Data Mining: Methods and Applications*, John Wiley & Sons, 1998.

[4] K. M. Al Aidaroos, A. A. Bakar and Z. Othaman, "Medical data classification with Naïve Bayes approach.", *Information Technology Journal*, vol. 11, no. 9, pp. 1166 – 1174, 2012

[5] J. L. Lustgarten, G. Vanathi, H. Grover, S. Visweswaran, "Improving classification performance with discretization on biomedical datasets," in *Proc. of AMIA Annual Symposium*, 2008, pp. 445–449.

[6] Y. Yangand, G. I. Webb, "A comparative study of discretization methods for Naïve Bayes classifiers", in *Proc of 2002 Pacific Rim Knowledge Acquisition Workshop*, 2002, pp. 159-173.

[7] R. Abraham, J. B. Simha and S. Iyengar, "Effective discretization and hybrid feature selection using Naïve Bayesian classifier for medical data mining", *Int. J of Computational Intelligence Research*, vol.5, no. 2, pp. 116-129, 2009.

[8] A. Sopharak, K. T. New, Y. A. Moe, M. N. Dailey, B. Uyyani, "Automatic exudate detection with a Naïve Bayes classifier", in *Int. Conf. on Embedded Systems and Intelligent Technology*, Thailand, 2008, pp. 139-142.

[9] R. C. Ganzalez , R. E. Woods, *Digital Image Processing*, 2nd ed, New Jersey: Prentice Hall 2001.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. "The WEKA data mining software: an update", *ACM SIGKDD Explorations*, vol. 11, issue 1, pp. 10-18, June 2009.

[11] T. Fawcett , "An introduction to ROC analysis", Pattern Recogn. Lett., vol. 27, issue 8, pp. 203-221, June 2006