International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:7, No:9, 2013

An Improved k Nearest Neighbor Classifier Using Interestingness Measures for Medical Image Mining

J. Alamelu Mangai, Satej Wagle, and V. Santhosh Kumar

***Abstract*—**The exponential increase in the volume of medical image database has imposed new challenges to clinical routine in maintaining patient history, diagnosis, treatment and monitoring. With the advent of data mining and machine learning techniques it is possible to automate and/or assist physicians in clinical diagnosis. In this research a medical image classification framework using data mining techniques is proposed. It involves feature extraction, feature selection, feature discretization and classification. In the classification phase, the performance of the traditional kNN k nearest neighbor classifier is improved using a feature weighting scheme and a distance weighted voting instead of simple majority voting. Feature weights are calculated using the interestingness measures used in association rule mining. Experiments on the retinal fundus images show that the proposed framework improves the classification accuracy of traditional kNN from 78.57 % to 92.85 %.

***Keywords*—**Medical Image Mining, Data Mining, Feature Weighting, Association Rule Mining, k nearest neighbor classifier.

I. Introduction

DATA MINING is the process of extracting implicit, non-trivial previously unknown patterns in huge data repositories. The traditional data mining techniques and algorithms are being customized and applied in various fields like computer security, finance, customer relationship management, web content mining, fault diagnosis, medicine etc. Classification and association rule mining are two of the many data mining tasks. Classification models learn to identify a pattern with the help of the training data. This knowledge is used to later to predict the class of a new entity. Association rule mining task generates rules that can be used to identify associations/correlations between the objects in the repository. The rules generated satisfy the user specified threshold *minsup* and *minconf*. For a rule $A \rightarrow B$, support of the rule is *(frequency of A)/(number of observations)*. A high level of support indicates that the rule is frequent enough for the decision making to be interested in it. The confidence of the rule is *(support(AB))/(support(A))*. A high level of confidence implies that the rule is true often enough to justify a decision based on it. Medical image mining refers to the non-invasive methods of investigating the human body with no harmful side effects. This helps the physicians to add to their expertise by providing an automated diagnosis. Retinal fundus images are a class of medical images which gives the details of the inner lining of the eye namely the sensory retina, the retinal pigment epithelium, Bruch's membrane and the choroid. Applying data mining techniques to automate the classification of such medical images is the state of the art research.

Feature extraction is the process of transforming the input data into a format suitable for the mining task. Various types of features can be extracted from images. Some of the approaches of extracting features from images are by applying image transforms, statistical methods, texture based methods. When handling huge data repositories much of the data becomes sparse in nature after feature extraction. It becomes meaningless for some of the data mining techniques in this sparse domain, especially those based on distance measures. Feature Selection helps to avoid this curse of dimensionality by identifying those features that are more predictive of the category of the object. The image feature vectors that are numeric are quantized using discretization. This transforms numeric values into discrete labels. It helps in reducing the classifier modeling time, improves the performance of many classifiers and makes the resulting model simple and easy to interpret. The k nearest neighbor (kNN) is one of the top ten classification models identified by the data mining research community. It predicts the class of a new object based on the classes of its k nearest neighbors. These neighbors then perform a simple majority voting to decide the class of the test instance. In spite of its simplicity, it performs poorly when an equal class probability exists among the k nearest neighbors. In this research a data mining framework using an improved kNN for classifying medical images is proposed.

The rest of the paper is organized as follows: Section II is a survey of the related work, Section III highlights the proposed architecture of medical image classification and Section IV gives the details of the experiments, results and findings. Section V concludes the paper.

II. Related Work

The research issues in image mining, current developments in image mining, image mining frameworks, state-of-the-art techniques and system, some future research directions for image mining are well discussed in [1]. The importance of pre-processing and feature extraction phases in mining large

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).

Satej Wagle is with Birla Institute of Technology & Science Pilani, Dubai Campus,Dubai,345055,UAE (e-mail: satejwagle@gmail.com).

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

collections of multi-media patient records is well discussed in [2]. These form the preparatory phases for any data mining model to achieve quality results. Two data mining techniques namely neural networks and association rule mining are used [3] to classify breast cancer images. These techniques have proved 70% classification accuracy. A hybrid feature selection approach using genetic algorithms and greedy stepwise is proposed [4] to classify mammogram images using decision tree method. A hybrid method of classifying CT scan brain images is proposed in [5] using association rule mining and decision tree method. Rules are generated from the frequent item sets in the feature data base. These rules are then used to model a decision tree based image classifier. Genetic Algorithm is combined with SVM for automatic classification of brain images in [6]. Three different image features namely Haralick, Tamura and Wolds texture features are explored with two classification models namely random forest and decision tree to classify CT scan brain images [7]. The results show that Haralick features combined with Random forest classifier gives better classification accuracy. Early diagnosis of lung cancer using neural networks is proposed in [8] to classify digital X-ray chest films into two categories: normal and abnormal. Features are extracted from MRI brain images using discrete wavelet transform in [9]. The feature set is further reduced using PCA and two classifiers namely kNN k nearest neighbor and ANN artificial neural network are modeled to classify MRI brain images into two categories. The performance of kNN for classifying traditional data sets is improved using a feature weighting scheme [10] and a distance weighted voting scheme [11]. However the authors have experimented this with structured data sets with categorical attributes from the UCI repository. As observed from research literature in medical image mining, the k-Nearest Neighbor classification algorithm has not gained much popularity. Inspite of its simplicity, its performance is degraded due to the simple majority voting scheme it uses to classify a new test image. In this research the performance of kNN to classify retinal fundus images is improved using a combined feature weighting scheme and a distance weighted voting scheme.

III. Proposed Work

The training phase and the testing phase of the proposed data mining framework for classifying medical images is shown in Fig. 1 and Fig. 2 as follows.

A. Image Preprocessing

Due to the non-uniformity in the color distribution of the images among the different subjects, the images are preprocessed for having uniform distribution of gray levels using histogram equalization [10]. This technique adjusts the local variation in contrast by increasing the contrast in lower contrast area and lowering the contrast in high contrast area.

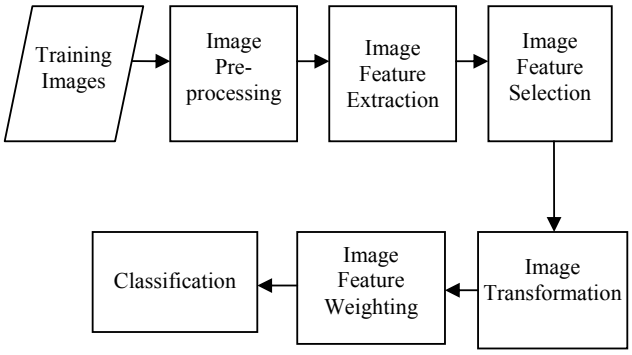

Fig. 1 The Training phase of the Proposed Architecture for Medical Image Classification

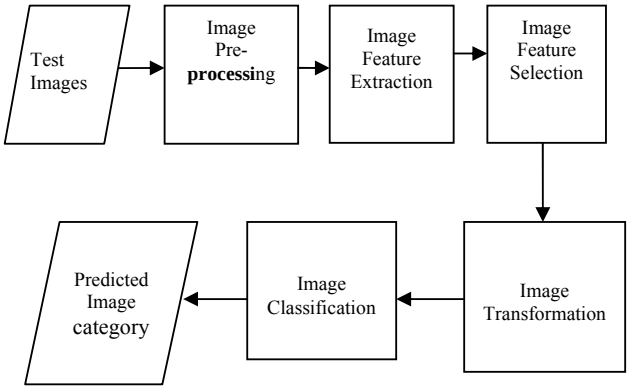

Fig. 2 The Training phase of the Proposed Architecture for Medical Image Classification

B. Feature Extraction

Features are extracted from the retinal fundus images of size 576 x 720 pixels. Localized statistical features are extracted from the images after dividing them into sub-images of size 36 x 90 pixels which yields a feature vector of size 128. Four statistical features such as mean, variance, skewness and kurtosis are extracted from each sub-image. This results in a feature vector of size 512. These features are computed as follows:

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

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

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

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

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 [12]. 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 Weighting

Weights are assigned to each feature using the interestingness measures minsup and minconf used in association rule mining after discretizing the features as follows.

1. Generate association rules with every feature value and image category combination.
2. Calculate the support and confidence of each rule.
3. Find the highest support and the highest confidence of each feature.
4. If the highest support or the highest confidence of a rule is less than minsup and minconf, then the weight of the feature is zero. Otherwise the weight of the feature is *1/(1-highest support of the feature)*.

E. Classification

The class of a new medical image is predicted as follows:

1. Find the k nearest neighbors to the test image using the feature weighted distance formula as $distance(X,Y) = \sqrt{\sum_{i=1}^{n} weight_i((x_i - y_i)^2)}$ where X and Y are any two images and n is the number of features.
2. Calculate the weight of each of the k nearest neighbors as : $distance(X,Y) = \sqrt{\sum_{i=1}^{n} weight_i((x_i - y_i)^2)}$ where x' is the test image, x_k^{NN} is the k^{th} nearest neighbor to the test image, x_1^{NN} is the first nearest neighbor to the test image and x_i^{NN} is the i^{th} nearest neighbor to the test image.
3. Predict the class y' of the test image as

$$y' = \text{argmax}_y \sum_{(x_i^{NN}, y_i^{NN})} w_i' \times \delta(y = y_i^{NN}).$$

$$\text{The function } \delta(y = y_i^{NN}) = \begin{cases} 1, & y = y_i^{NN} \\ 0, & y \neq y_i^{NN} \end{cases}$$

where y_i^{NN} is the category of the i^{th} nearest neighbor, y is the set of all image categories.

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. 3 and Fig. 4.

Fig. 3 Sample normal fundus images

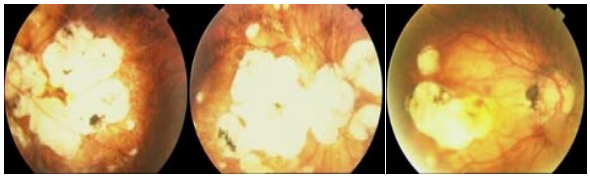

Fig. 4 Sample very severe fundus images

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. 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.

```
@relation 'ga-
weka.filters.unsupervised.attribute.ReplaceMissingVa
lues-weka.filters.unsupervised.attribute.Remove-V-
R192,257,276,284,286,294-295,297,308-309,311-
313,316-318,320-321,324-327,330,332-
334,343,346,357,372,385,407,409-410,426,439,455,457-
459,480-481,513-
weka.filters.unsupervised.instance.Randomize-S42'

@attribute B64 numeric
@attribute C1 numeric
@attribute C20 numeric
@attribute C28 numeric
@attribute C30 numeric
……..
@attribute class {0,1}

@data
1065.5201,13.3043,-0.7566,-0.3199,-0.1932,-0.6762,-
0.7718,0.0298,-1.1593,-0.5015,-0.5806,-
0.1004,0.282,-0.3611,-0.0313,-0.5819,-2.0836,-
1.3548,-0.4647,-1.0692,-0.3817,-1.0075,-
0.5028,0.1765,0.0207,-0.1518,-
1.0057,0.2473,0.0053,0.1604,178.0056,2.7376,1.4559,1
.663,1.2205,3.1125,5.5168,3.0442,2.657,1.9873,6.3285
,2.6745,1
…….
```

Fig. 5 The Image File after Feature Selection

```
@relation 01-Nano-r43

@attribute B64 {3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute C1 {16,17,18,19,20,21,22,23,24,25,26,27}
……
@attribute D97
{698,699,700,701,702,703,704,705,706,707,708,709,710
,711,712}
@attribute class {Normal, Very_Severe}
@data
15, 27, 30, 57, 73, 94, 107, 136, 151, 170, 178,
196, 214, 224, 248, 257, 283, 305, 319, 330, 347,
356, 371, 389, 405, 418, 428, 452, 472, 494, 510,
519, 532, 550, 570, 597, 624, 638, 648, 666, 686,
712, Very_Severe
```

Fig. 6 The Image File after Feature Discretization

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. Feature Weights are then calculated using minsup and minconf interestingness measures. The training file is divided using 70 – 30 % split where 70% of the images are used to calculate the feature weights and also as training set for the classifier. The traditional kNN is run from Weka [13], and the modified kNN is developed using a JAVA program. Table I gives a comparison of the percentage classification accuracy of the traditional and the modified kNN for k = 1.

TABLE I
COMPARATIVE ANALYSIS BASED ON CLASSIFICATION ACCURACY

Classifier	kNN	Proposed kNN	J48	oneR	SVM	NN	Bagging	Boosting
% accuracy	78.57	92.57	85.71	82.14	96.42	96.42	85.71	85.71

The accuracy of the proposed kNN is significantly better than the traditional kNN for medical images. Table I also shows the percentage classification accuracy of some of the other existing classifiers namely decision tree based (J48), rule based (oneR), support vector machine (SVM) neural network based (NN), ensemble bagging and boosting. The kNN is simple and easy to implement of all classification models. If the data is pre-processed well it can be trained to provide better predictions. The confusion matrix in Table II shows the details of correctly and incorrectly classified instances of a classifier. It shows the count of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

TABLE II
THE CONFUSION MATRIX OF A CLASSIFIER

Actual Class	Predicted Class	
	A	B
A	TP	FN
B	FP	TN

The confusion matrix of the traditional kNN is shown in Table III.

TABLE III
THE CONFUSION MATRIX OF TRADITIONAL KNN CLASSIFIER

Actual Class	Predicted Class	
	Very Severe	Normal
Very Severe	6	4
Normal	0	18

The sum of TP and TN would contribute to the accuracy of the classifier. The confusion matrix of the proposed data mining framework for medical image classification using the modified kNN is shown in Table IV.

TABLE IV
THE CONFUSION MATRIX OF PROPOSED KNN CLASSIFIER FOR MEDICAL IMAGE CLASSIFICATION

Actual Class	Predicted Class	
	Very Severe	Normal
Very Severe	8	2
Normal	0	18

As seen in Table IV, the feature weighting scheme and the distance weighted voting scheme have helped to improve the performance of the traditional kNN.

V. CONCLUSION

A data mining framework for classifying medical images is proposed in this research. kNN is a simple machine learning classifier which has been much less explored in medical image mining. In this research the performance of the traditional kNN machine learning classifier is tweaked using a feature weighting scheme and a distance weighted voting scheme. The interestingness measures used in association rule mining is used to calculate the feature weights. This added with distance weighted voting has improved the medical image classification accuracy of the traditional kNN. Our future work is to experiment this framework on other classes of medical images and also to refine our feature space.

ACKNOWLEDGMENT

The authors would like to thank Prof. Dr. R. K. Mittal, the Director of BITS Pilani, Dubai Campus for his encouragement and support in facilitating this research.

REFERENCES

[1] W. Hsu, M. L. Lee, and J. Zhang, "Image Mining : Trends & Developments." in *J of Intellident Information Systems*, vol 19, issue 1, , pp. 7-23, 2002..

[2] A. S. Elmaghraby, M. M. Kantardzic, and M P. Wachowiak, "Data Mining from Multimedia Patient Records", *Data Mining and Knowledge Discovery Approaches based on Rule Induction Techniques*, Springer, pp. 551–595, 2006.

[3] M.-L. Antonie, O. Zaiane, and A. Coman, "Application of Data Mining Techniques for Medical Image Classification", in *Proc of 2nd Intl. Workshop on Multimedia Data Mining*, 2001, pp. 94-101.
[4] M. Vasantha, V. S. Bharathi and R. Dhamodaran, "Medical Image Feature Extraction, Selection and Classification", *In. J of Engineering Science and Technologl*, vol. 2, no. 6, pp. 2071 – 2076, 2010
[5] P. Rajendran, and M. Madheswaran, "Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm", *J. of Computing*, vol 2, no 1, pp. 127–136, 2010.
[6] A. Kharrat, K. Gasmi, M. B. Messaoud, N. Benamrane and M. Abid, "A Hybrid Approach for Automatic Classification of Brain MRI Using Genetic Algorithm and Support Vector Machine", *J. of Sciences*, issue 17, pp. 71-82, 2010.
[7] B.G..Prasad and A. N. Krishna, "Classification of Medical Images using Data Mining Techniques", Advances in Communication, Network & Computing", *Advances in Communication, Network & Computing, Lecture Notes of the Institute of Computer sciences, social informatics and Telecommunication Engineering*, vol.108,, pp. 54-59, 2012.
[8] Z. S. Zubi, and R. A. Saad, "Using some Data Mining Techniques for Early Diagnosis of Lung Cancer", in *Proc. Of 10th WSEAS Intl. conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, 2011, pp. 32-37.
[9] N. H. Rajini, and R.Bhavani, "Classification of MRI Brain Images using k nearest neighbor and artificial neural network" *IEEE-Intl conference on Recent Trends in Information Technology*, pp. 863-868, 2011.
[10] L. Muflikhah, and Adnyana, "Classifying Categorical Data Using Modified K-Nearest Neighbor Weighted by Association Rules", in *Proc of the Int Conf on Future Information Technology*, vol 13, 2011, pp. 347-351.
[11] J. Gou , L. Du , Y. Zhang and T. Xiong, "New Distance-weighted k-nearest Neighbor Classifier", *J. of Informational and Computational Science*, vol 9, no 6, pp. 1426-1429, 2012.
[12] R. C. Ganzalez, amd R. E. Woods, *Digital Image Processing*, 2nd ed, New Jersey: Prentice Hall 2001
[13] 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.