Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 8, No. 3, August 2017), Pages: 79-85
www.jacr.iausari.ac.ir

Presenting a Fast Classifier Based on Unsupervised Learning for Diagnosis Diseases

Najmeh Hosseinpour[✉1], **Afzal Ghaseimi**[2]

1) Young Researchers and Elite Club, Andimeshk Branch, Islamic Azad University, Andimeshk, Iran
2) Department of Computer Engineering, Islamic Azad University, Andimeshk Branch, Andimeshk, Iran
nhoseinpor.tax@gmail.com; aghaseimi@gmail.com

Received: 2016/07/19; Accepted: 2016/09/20

Abstract

From long ago, decision support systems (DSS) as a vital tool in many industrials is considered by decision-makers. These systems can aid managers in making better decisions by collecting and interpreting data. Medical decision support systems (MDSS) have critical role in medical practice. They can help physicians for improving the quality of medical diagnosis. Classifiers as main core of MDSS systems play an important role in improving their performance. This paper presents an unsupervised learning-based real time classifier which is able to perform recognizing medical patterns with proper precision and speed. In the training phase, the proposed classifier is capable to obtain reference models related to classes using synergic clustering technique and finding the frequency of attributes. In order to evaluate efficiency of the proposed classifier, the UCI datasets including breast cancer (WBCD), liver disease (ILPD) and diabetic disease (PID) are applied. The obtained results indicate the effectiveness of the proposed method.

Keywords: *Medical Decision Support Systems (MDSS); Machine Learning; Classifier; Clustering*

1. Introduction

From the past, pattern recognition as one of the most important issues in the field of machine learning is interest to researchers and scholars[1, 2]. Indeed, pattern recognition is a mechanism in which a machine is able to discriminate desirable patterns from a set of patterns using prior knowledge about them. Pattern recognizing systems include two phases: training and testing. In the training phase, machine trained on a set of patterns to partition feature space way that maximize the discrimination ability to make proper models. In testing phase, the trained machine can assign an unknown pattern to one of the classes. Pattern recognition has become an important problem in wide variety of fields such as medicine[3-6], biology[7, 8], audio[9] and image processing[10, 11], marketing[12] etc.

Up to now, several methods for pattern classification have been presented. Famous methods such as naïve bayesian (NB)[13], k-nearest neighbor (KNN)[14], support vector machine (SVM)[15], artificial neural networks (ANN)[16] and decision tree (DT)[17] are some of them. Among the most important problems of these methods can address to high time complexity of training and testing phases as well as low accuracy which is not suitable for many applications such as medical field, in which time and accuracy are two important factors.

In this paper, a new fast classifier based on unsupervised learning is presented which is able to perform medical pattern recognition with proper speed and accuracy. The proposed approach, in the training phase, is able to make proper reference models using synergic clustering technique and finding the most frequent features for applying in recognition phase. The efficiency of the proposed classifier is evaluated by WBCD, ILPD and PID UCI datasets[18].

The structure of the paper is organized as following: in section 2, the proposed method is presented; in section 3 the experimental results are shown; finally, the paper end by conclusion.

2. The Proposed Method

In this section, the details of the proposed method for classifying medical patterns are presented. The proposed approach includes two phases: training and recognition phases which their details are introduced in the following.

2.1 Training Phase

The main purpose of this section is finding reference models related to classes and consists of several steps. For this purpose, at first, the normalization of patterns in interval [0 1] are carried out by Eq.1:

$$\mathrm{Normalize}(\mathrm{x_i}) = \frac{(\mathrm{x_i} - \mathrm{x_{i_{min}}}}{(\mathrm{x_{i_{max}}} - \mathrm{x_{i_{min}}})} \quad (1)$$

where x_i is i^{th} feature, $x_{i_{max}}$and $x_{i_{min}}$are the maximum and minimum of i^{th} feature, respectively.

In the following, the normalized patterns are partitioned into K clusters using K-means algorithm. In order to reduce the dimensions of the patterns, vector quantization is carried out. Moreover, it is determined that each class consists of what unique clusters. In the next step, for each one of classes, two matrixes *Max_Freq* and *Min_Dist* are calculated. *Max_Freq* matrix represents the most repeated observations (winners) in the each of pattern features. The shortest distance between the winner observations and training patterns is countered as *Min_Dist* matrix. It should be mentioned that for each class, the unique clusters, winner observations and the shortest distance are considered as class reference model. Pseudo code of the proposed training is illustrated in Figure 1.

```
Procedure of Proposed Training(Inputs Dataset, Output: Max_Freq Min_Dist Matrixes, Unique_Cluster)
Begin
% Pre-processing
Normalize Dataset between [0 1
Codebook=Clustering pre-processed Dataset in to K clusters based on K-means method;
X=Vector Quantization(pre-processed Dataset, Codebook);
New_Dataset=Numeric Quantization(X,0,255);
% Determining unique  clusters of classes
For i=1 to no_classes
    Computing unique clusters of classes, Unique_Cluster(i);
End;
% Calculating of Reference Models
For i=1 to no_class
    For j=1 to no_featurs
        For k=1 to no_frequenceis
            finding frequency matrix Max_Freq(i,j,k) related to $k^{th}$ frequency of jth feature of $i^{th}$ class;
End;
```

```
        End;
    End;
For i=1 to no_class
    For j=1 to no_featurs
        For k=1 to no_frequencies
            Computing minimum distance matrix Min_Dist(i,j,k) via Max_Freq(i,j,k) entity and jth feature of $i^{th}$ class
        End;
    End;
End;
Return class reference models Unique Cluster, Max_Freq and Min_Dist;
End;
```

Figure 1. Pseudo code the proposed training method

2.2 Recognition Phase

This section discusses how to classify unknown testing patterns by synergistic set of calculated obtained reference models. For this purpose, at first, preprocessing operation, involving normalization and vector quantization is performed on the inputted test pattern. In the following, it is checked whether it is belong to unique clusters or not. If belong, the label of the inputted pattern is determined, easily. Otherwise, the label is determined using *Max_Freq* and *Min_Dist* matrixes based on the least amount of distance. Flowchart of the proposed testing method is shown if Figure 2.

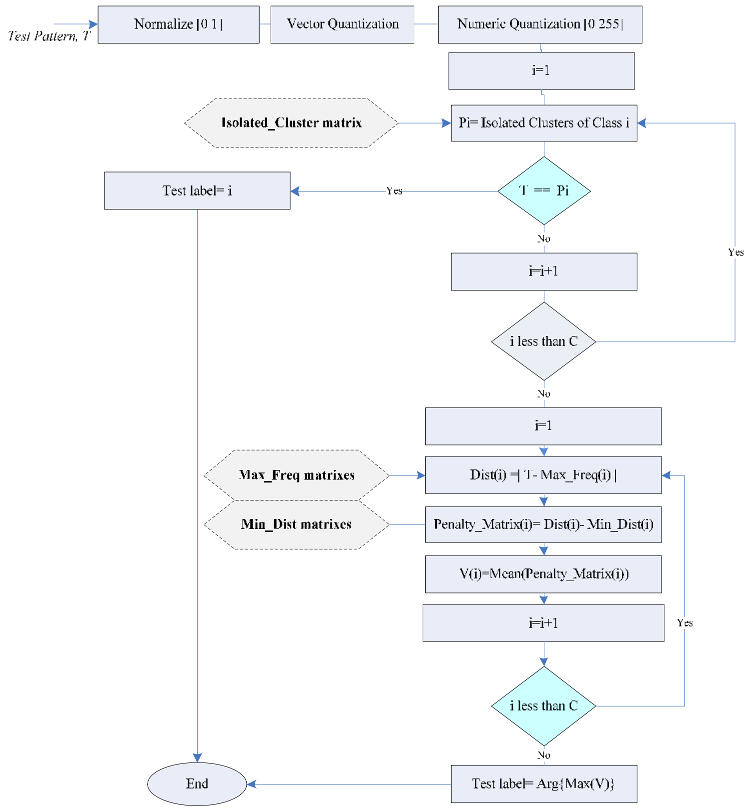

Figure 2. Flowchart of the proposed recognizing method

3. Experimental Results

3.1 Dataset

In order to evaluate the proposed method, three datasets of UCI repository including Wisconsin Breast Cancer Dataset (WBCD), Indian Liver Patient Dataset (ILPD) and PIMA Indian Dataset (PID) are used. Dataset details are shown in Table 1.

Table 1.Description of used datasets.

Dataset	#. of instances	#. of patient	#. of non-patient	#. of attributes
WBCD	699	458	241	10
ILPD	583	416	167	10
PID	768	500	268	8

3.2 Evaluation Metrics

Confusion matrix is useful tool that can be used to evaluate performance of the proposed method. Confusion matrix for two classes is illustrated in Table 2.

Table 2.Confusion matrix for positive and negative records

		Predicted Class	
Actual Class		**Positive**	**Negative**
	Positive	True Positive(TP)	False Negative(FN)
	Negative	False Positive(FP)	True Negative(TN)

True Positive (TP) indicates positive records that are correctly classified while True Negative (TN) demonstrates negative records that have been property classified. Moreover, False Positive (FP) shows negative records that falsely classify while False Negative (FN) indicates positive records that have been incorrectly classified. Accuracy is a measure that indicates the percentage of records which were correctly classified. Also, sensitivity and specificity are two other important measures that are used for evaluating performance classifier. The sensitivity is referred to TP rate while the specificity indicates TN rate.

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{TN}+\text{FN})} \times 100(\%) \tag{2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \times 100(\%) \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \times 100(\%) \tag{4}$$

It should be mentioned the proposed method has been implemented by MATLAB software.

3.3 Results

The evaluation results of the proposed method on breast cancer, liver disease and diabetes datasets by accuracy, specificity and sensitivity metrics has been shown in Table 3.

Table 3.Evaluating the proposed method in 2-fold, 4-fold and 10-fold cross validations

Dataset	Metrics	2-Fold	4-Fold	10-Fold
WBCD	Sensitivity	98.23	97.70	98.03
	Specificity	89.56	88.21	91.50
	Accuracy	95.30	94.26	96.76
ILPD	Sensitivity	72.47	75.13	77.20
	Specificity	48.61	32.84	41.58
	Accuracy	66.32	62.41	66.68
PID	Sensitivity	71.26	71.64	74.32
	Specificity	83.76	85.24	85.45
	Accuracy	79.54	80.50	81.51

As shown in Table 3, the best performance of the proposed method is in 10-fold cross validation that for WBCD, ILPD and PID datasets are 96.76%, 66.68% and 81.51%, respectively.

Also, the proposed method is comprised with three efficient classifiers SVM, KNN and DT. The obtained results have been illustrated in Table 4.

Table 4.Comparison of the proposed method accuracy with SVM, KNN and DT classifiers in 10-flod

Dataset	SVM			KNN			Decision Tree			The proposed method		
	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
WBCD	97.01	56.73	87.16	98.5	94.03	95.82	100	92.54	96.16	100	92.64	96.76
ILPD	80.71	56.16	65.10	71.24	49.50	62.12	80.71	59.16	67.10	71.13	57.05	66.68
PID	70.67	57.34	64.01	73.34	60.01	68.81	80	60.01	72.40	83.20	68.10	81.56

Moreover, up to now several classifiers for disease diagnosis have been presented. In the following, we compare our method with some of them. The comparison results are shown in Table 5.

Table 5.Comparison of the proposed method with related works for WBCD, ILPD and PID datasets

Dataset	Reference	Method	Accuracy (%)
WBCD	S. Bashir et al.[19]	HMV	96.71
	J. Abonyi and F. Szeifert[20]	Supervised fuzzy clustering	95.57
	A. AKGÜNDOGDU[21]	Genetic Programming	96.60
	The proposed method	Unsupervised learning	96.76
ILPD	S. Bashir et al. [19]	HMV	67.54
	S. Karthik et al[22]	Back Propagation	61.5
	F.Wang et al. [23]	Triplet-SVM/Doublet-SVM	64.84/67.9
	The proposed method	Unsupervised learning	66.68
PID	S. Bashir et al.[19]	HMV	77.08
	M. F Ganji and M.S. Abadeh[24]	Fuzzy-ACO	79.48
	M. W Aslam and A. K Nandi[25]	Genetic Programming	78.50
	The proposed method	Unsupervised learning	81.51

4. Conclusion

In this paper, a new real time classifier based on unsupervised learning is suggested which is able to done medical pattern recognition with proper speed and accuracy. The proposed scheme, in the training phase can make proper reference models using synergic clustering technique and finding the most frequent features for applying in recognition phase. The performance of the proposed method is evaluated by WBCD, ILPD and PID datasets. The experimental results show the accuracy of the proposed method on WBCD, ILPD and PID datasets is more than 6% better than average of other mentioned classifiers.

Acknowledgement

The presented paper is the result of a research which was sponsored by the Young Researchers and Elite Club, Andimeshk Branch, Islamic Azad University, Andimeshk, Iran, and the authors hereby express their thanks.

References

[1] T. M. Mitchell, *Machine learning*: McGraw-Hill 1997.

[2] E. Alpaydin, *Introduction to machine learning*: MIT press, 2014.

[3] I. Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective,” *Artificial Intelligence in medicine,* vol. 23, no. 1, pp. 89-109, 2001.

[4] K. Polat, and S. Gunes, “Breast cancer diagnosis using least square support vector machine,” *Digital Signal Processing,* vol. 17, no. 4, pp. 694-701, 2007.

[5] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, “Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques,” *Journal of the American Medical Informatics Association,* vol. 13, no. 5, pp. 516-525, 2006.

[6] P. Sajda, “Machine learning for detection and diagnosis of disease,” *Annu. Rev. Biomed. Eng.,* vol. 8, pp. 537-565, 2006.

[7] A. L. Tarca, V. J. Carey, X.-w. Chen *et al.*, “Machine learning and its applications to biology,” *PLoS Comput Biol,* vol. 3, no. 6, pp. e116, 2007.

[8] C. Kampichler, R. Wieland, S. Calm *et al.*, “Classification in conservation biology: a comparison of five machine-learning methods,” *Ecological Informatics,* vol. 5, no. 6, pp. 441-450, 2010.

[9] M. Shami, and W. Verhelst, “An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech,” *Speech Communication,* vol. 49, no. 3, pp. 201-212, 2007.

[10] L. Zhang, F. Lin, and B. Zhang, “Support vector machine learning for image retrieval,” in International Conference on Image Processing, 2001, pp. 721-724.

[11] E. Rosten, and T. Drummond, “Machine learning for high-speed corner detection,” in European conference on computer vision, 2006, pp. 430-443.

[12] G. Cui, M. L. Wong, and H.-K. Lui, “Machine learning for direct marketing response models: Bayesian networks with evolutionary programming,” *Management Science*, vol. 52, no. 4, pp. 597-612, 2006.

[13] K. M. Leung, “Naive bayesian classifier,” *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

[14] B. V. Dasarathy, *Nearest neighbor (NN) norms:(NN) pattern classification techniques*: McGraw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, California, 1991.

[15] K. Soman, R. Loganathan, and V. Ajay, *Machine learning with SVM and other kernel methods*: PHI Learning Pvt. Ltd., 2009.

[16] W. Webber, R. P. Lesser, R. T. Richardson *et al.*, “An approach to seizure detection using an artificial neural network (ANN),” *Electroencephalography and clinical Neurophysiology,* vol. 98, no. 4, pp. 250-272, 1996.

[17] C. J. Mantas, and J. Abellan, “Credal-C4. 5: Decision tree based on imprecise probabilities to classify noisy data,” *Expert Systems with Applications,* vol. 41, no. 10, pp. 4625-4637, 2014.

[18] C. L. B. a. C. J. Merz. "UCI Repository of Machine Learning Databases," Available from http://www.ics.uci.edu./~mlearn/MLReporsitory.html. .

[19] S. Bashir, U. Qamar, F. H. Khan *et al.*, “HMV: A medical decision support framework using multi-layer classifiers for disease prediction,” *Journal of Computational Science,* vol. 13, pp. 10-25, 2016.

[20] J. Abonyi, and F. Szeifert, “Supervised fuzzy clustering for the identification of fuzzy classifiers,” *Pattern Recognition Letters,* vol. 24, no. 14, pp. 2195-2207, 2003.

[21] A. AKGÜNDOGDU, “ Breast Cancer Classification with Genetic Programming,” *International Journal of Electronics, Mechanical and Mechatronics Engineering (IJEMME),* vol. 2, no. 1, pp. 72:78, 2012.

[22] S. Karthik, A. Priyadarishini, J. Anuradha *et al.*, “Classification and rule extraction using rough set for diagnosis of liver disease and its types,” *Adv. Appl. Sci. Res,* vol. 2, no. 3, pp. 334-345, 2011.

[23] F. Wang, W. Zuo, L. Zhang *et al.*, “A kernel classification framework for metric learning,” *IEEE transactions on neural networks and learning systems,* vol. 26, no. 9, pp. 1950-1962, 2015.

[24] M. F. Ganji, and M. S. Abadeh, “Using fuzzy ant colony optimization for diagnosis of diabetes disease,” in 18th Iranian Conference on Electrical Engineering, 2010, pp. 501-505.

[25] M. W. Aslam, and A. K. Nandi, “Detection of diabetes using genetic programming,” in 18th European Signal Processing Conference 2011, pp. 1184-1188.