International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

June 2015 | Issue 23 | Volume 7 | Number 2 | Pages 39-45

COMPARISON AND CHECKING CLASSIFICATION IN PATTERN RECOGNITION

N. Molla Esmaili Sh. Hoshiary L. Rastgou S. Rajebi

Electrical Engineering Department, Seraj Higher Education Institute, Tabriz, Iran
nasimesmaieli@yahoo.com, shilan_hoshyari@yahoo.com, l.rastgou.ms@seraj.ac.ir, s.rajebi@urmia.ac.ir

Abstract- This paper presents a comparison study of the different parametric and non-parametric pattern classifiers that are commonly used for pattern recognition. Results of individual pattern classifiers, feature selection as well as error estimates for the various data sets used are presented, along with the CPU-time consumption. The evaluation results reported here might be useful when designing practical pattern classification. Four different methods for data classification, are evaluated in our study.

Keywords: Pattern recognition, Classification, Bayesian Method, KNN Method, Parzen Window Method, Perceptron Neural Network (MLP), Selection Feature, FDR Method, Divergence, CCR.

I. INTRODUCTION

Pattern recognition is a branch of learning topics, so that can be said pattern recognition, receiving raw data and makes decisions based on data classification. Pattern recognition has a long history, but be ore the 1960s it was mostly the output of theoretical research in the area of statistics.as with everything else, the advent of computers increased the demand for practical application of pattern recognition, which in turn set new demand for further theoretical developments. Automation in industrial production and the need for information handling and retrieval are becoming increasingly important [1].

Pattern recognition has a wide variety of commercial, medical and industrial applications. Some of the many applications include handwriting recognition, noise classification, face recognition, fingerprint recognition, biomedical image processing applications, mammography, speech recognizers, etc. With such a wide variety of applications it is not possible to come up with a single classifier that can give good results in all the cases. Hence, the classifier(s) adopted are highly dependent on the problem domain [2].

In general, may be two forms of pattern recognition, if using information from previously set of data which takes place as training data for classified, that is known as supervised pattern recognition and if non-availability of previous information about the classification of data is known as unsupervised pattern recognition or clustering. The pattern recognition methods that we review is based on supervised classification. The desired patterns from a data set using prior knowledge about the patterns or statistical information.

The objective of a classification system is to assign a pattern presented to it to a class using the feature vector (list of attribute values). The complexity of the classification problem is dependent on the variability of the feature values for patterns in the same class relative to the difference between feature values for patterns in different classes. Consequently the optimality of a classifier is dataset dependent. Therefore achieving optimal performance for a pattern recognition system is not necessarily consistent with obtaining the best performance for a single classifier. In practice, one might come across a case where no single classifier can make a classification with an acceptable level of accuracy. In such cases it might be better to pool the results of different classifiers to achieve the optimal decision accuracy. Each classifier can operate well on different aspects of the input feature vector [2].

So design pattern recognition system is to determine the optimal decision procedures, the process of identification and classification is required. In this paper, we use both statistical and neural network based classifiers. The statistical techniques for early education in our study include *K* nearest neighbor (KNN), Parzen window and Bayesian classifier and our neural classifier is a simple multilayer perceptron (MLP), each of which are described in the following section.

II. DESCRIPTION

The data that we used for evaluate different classification methods are 768 Standard data which is divided into 12 classes. The way we have chosen to assess this method of classification is Hold Out method, where the test data and training data are completely separated and no test data is equal the training data. The reason for this approach is its high accuracy compared to other methods of evaluation, so the number of data should be used as the training data and the number of data should be used as the testing data. For this purpose, we consider 264 samples and 504 samples respectively as training data and test data of all data extracted.

In this study, to achieve better and more accurate conclusion, we compute separation rate classes for 12 classes. Also the original data set was consisted of 48 features but to reduce the size of calculations, increase the public performance of programs and increase the estimation systems error, so that it doesn't reduce the quality of pattern recognition, we used FDR manner, which is one of the optimal feature selection method.

Actually Fisher discriminant is one of an efficient approaches for dimension reduction in statistical pattern recognition. Since in the classification the bigger the square of the difference between the means of *M* kinds projected sample points is and at the same time the smaller the within-class scatters are, the better the expected line is. More formally, construct the following function, so called Fisher's discriminant ratio [3].

For the multiclass case averaging forms of FDR can be used. One possibility is

$$\text{FDR}=\sum_{i}^{M}\sum_{j\neq i}^{M}\frac{(\mu_i-\mu_j)^2}{\delta_i^2+\delta_j^2} \tag{1}$$

where the subscripts *i, j* refer to the mean and variance corresponding to the feature under investigation for the classes w_i, w_j, respectively [1].

In the separation matrix obtained (FDR) members of the main diagonal of the matrix, represent the separation power of each feature their corresponding index. Much larger than the Fisher's discriminant is a feature, the more powerful features is the ability to classify. So we can calculate the value for each attribute of Fisher's discriminant ratio and rank them in descending order and then select the best features of a certain number.

A. Bayesian Method

This method is a parametric method for classification. Theory Bayesian the basic statistical methods, in this theory, probability theory is a solid foundation for the design of pattern classification. Formation mechanism model based on the framework of this approach is the possibility that set the random values are categorized in classes. General parametric method for classifying data based on the occurrence or non-occurrence of a phenomenon which is given below. Attention to Equation (5) likely each category information belonging to each of the classes is calculated and the class with the highest probability is selected.

Bayesian method is a kind of supervised classification the main feature of this method is that the training volume and parameter estimation requires little to begin with one of the highlights of the gathering. Bayesian classification of multiple features of a phenomenon so that all these features are also able to participate in the rankings, according to Bayes theorem, the conditional probability is established in Equation (3). This method is based on the principle that there is a probability distribution for each entity observing a new data and reasoning about the probability distribution of optimal decisions can be taken.

As follows, if we have l feature in *N* sample $\underline{x}$ and *m* classes w.

$\underline{x}=[x_1,....x_l]$, $w_1, w_2 ..., w_m$,

$\underline{x}\rightarrow w_i : P(\mathrm{w}_i \mid \mathrm{x})\big|\text{maximum}$

For example, if we have two classes:

Decide w_1 if $P(w_1 \mid x) > P(w_2 \mid x)$; otherwise decide w_2.

According to Equation (2):

$$P(w_j \mid x)=\frac{p(x \mid w_j)P(w_j)}{p(x)} \tag{2}$$

we could said:

$$p(x \mid w_1)P(w_1)(><)p(x \mid w_2)P(w_2) \tag{3}$$

For classes with equal probability of occurrence

$$p(x \mid w_1)(><)p(x \mid w_2) \tag{4}$$

Gaussian normal distribution function:

$$p(x)=\frac{1}{\sqrt{2}\pi\delta}\exp\left[-\frac{1}{2}(\frac{x-\mu}{\delta})^2\right] \tag{5}$$

In the flowchart about this method, after detecting the number of data and the number of training data as test data for each class, average and variance have been calculated in Equations (6) and (7), respectively, with the density function probability in Equation (8), the probability of each class calculated on the test data.

$$\mu=\varepsilon[x]=\int_{-\infty}^{+\infty}xp(x)dx \tag{6}$$

$$\delta^2=\varepsilon\left[(x-\mu)^2\right]=\int_{-\infty}^{+\infty}(x-\mu)^2\,p(x)dx \tag{7}$$

Multivariate Gaussian normal distribution function:

$$p(x)=\frac{1}{2\pi^{\frac{d}{2}}|\sum|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x-\mu)^t(x-\mu)\right] \tag{8}$$

where *d* is number of features

$$\sum=\varepsilon\left[(x-\mu)^t(x-\mu)\right]=\int(x-\mu)^t(x-\mu)\,p(x)dx \tag{9}$$

$$\mu_i=\varepsilon[x_i] \tag{10}$$

$$d_{ij}=e\left[(x_i-m_i)(x_j-m_j)\right] \tag{11}$$

And then we determine the probability that any given test result in one class and after class set of test data , we check whether the data in the relevant class has been properly tested or not if that is true, the amount of data correctly to each other by this method is greater [1].

- *Bayesian Algorithm* (Figure 1):
- Receiving image data educational and testing
- Calculate the variance and the mean for each sample
- Calculate the probability distribution function of the data
- Evaluation of test data correctly classified
- Bayesian detection of the assay correctly with other data
- Repeat the fourth step in the wrong classification
- End

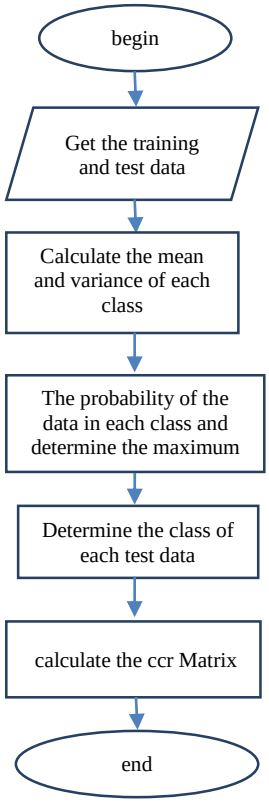


Figure 1. Flowchart of Bayesian classification method

B. *K*-Nearest Neighbor Method

This classifier classifies a pattern *x* by assigning it to the class label that is most frequently represented among it's k nearest neighbor patterns. In the case of a tie, the test pattern is assigned the class with minimum average distance to it. Hence, this method is sensitive to the distance function. For the minimum average distance, the metric employed is the Euclidean distance. This metric requires normalization of all features into the same range. The *k*-nearest neighbor classifier is a conventional nonparametric classifier that is said to yield good performance for optimal values of *K* [4].

Nearest neighbor estimation is a method that implements such a refinement. The method is based on the following observation. Let $R(z) \subset R^N$ be a hyper sphere with volume *V*. The center of *R* (*z*) is *z*. If the number of samples in the training set T_k is N_k, then the probability of having exactly *n* samples with in *R(z)* has a binomial distribution with expectation:

$$E[n] = N_k \int_{y \in R(z)} p(y \mid w_k) dy \approx N_k V_p (z \mid w_k) \tag{12}$$

Suppose that the radius of the sphere around *z* is selected such that this sphere contains exactly samples. It is obvious that this radius depends on the position *z* in the measurement space. Therefore, the volume will depend on *z*. We have to write *V*(*z*) instead of *V*. With that, an estimate of the density is:

$$P(z \mid w_k) = \frac{k}{N_k V(z)} \tag{13}$$

The expression shows that in regions where $p(z \mid w_k)$ is large, the volume is expected to be small. This is similar to having a small interpolation zone. If, on the other hand, $p(z \mid w_k)$ is small, the sphere needs to grow in order to collect the required *k* samples.

The parameter *k* controls the balance between the bias and variance.

We consider the entire training set and use the representation T_s, The total number of Samples is N_s and *i*=1, ..., *k*.

$$k = \arg\max \{ p(z \mid w_i) P(w_i) \} =$$

$$= \arg\max \left\{ \frac{k_i}{V(z) N_i} \frac{N_i}{N_s} \right\} = \arg\max \{ k_i \} \tag{14}$$

The class assigned to a vector *z* is the class with the maximum number of votes coming from *k* samples nearest to *z* [5].

• *K-Nearest Neighbor Algorithm* (Figure 2):

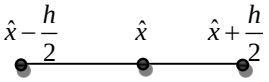
- Get the training and test data
- Set a fixed value for the number of data
- Set an interval for the neighborhood
- Increasing and decreasing the interval
- check out the number of data located in the interval
- If events the constant were to continue, otherwise repetition of the fourth stage
- Class selection has been investigated in a range of best in Class
- End

C. Parzen Window Method

In many respects, it is similar to *K*-Nearest Neighbor method such that both of them are non-parametric methods but with this difference that in Parzen window method Contrary to KNN method, choose a fixed value for volume *V* and determine the corresponding *K* from the data. In this method, instead of continuing to increase the range of features until to a certain number of input data are in the interval, the interval is fixed at a constant value which this fixed interval is called the window. Then number of training data have been generated in the window of each class is measured and the input data will be awarded to the class that the most number of training data of its class, was located in window.

We have in one dimension:

$$P \approx \frac{k_N}{N} \begin{cases} k_N & \text{in } h \\ N & \text{total} \end{cases} \tag{15}$$



$$P(x) = \frac{1}{h} \frac{k_N}{N} \tag{16}$$

$$\left| x - \hat{x} \right| \le \frac{h}{2} \tag{17}$$

$$\phi(x_i) = \begin{cases} 1 & \left| x_{ij} \right| \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

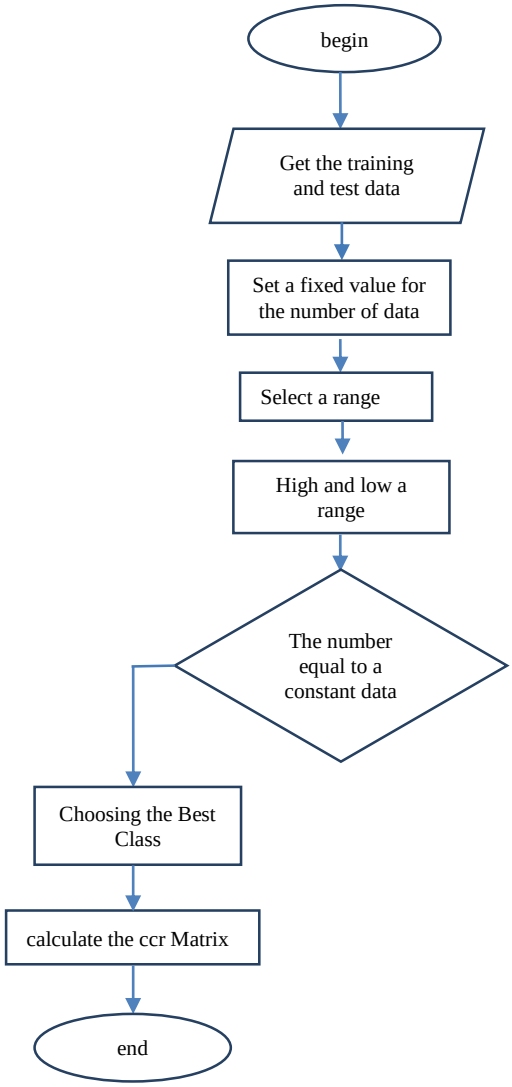


Figure 2. Flowchart of *K*-Nearst Neighbor classification method

where k_N is the number of points falling inside this hypercube and $x_{ij}, j = 1,...,l$ are the components of x_i

$$P(x) = \frac{1}{h^l}\left(\frac{1}{N}\sum_{i=1}^{N}\varphi\left(\frac{(x_i - x)}{h}\right)\right) \quad (19)$$

In other words:

$$P(x) = \frac{1}{\text{volume}} \times \frac{1}{N} \times \text{number of points inside} \quad (20)$$

One of the most important items in this method of classification is the size of the window that is considered that there is no amount specified and predetermined for it, rather should using training data that is available and determine the best value of window size to occur the best possible classification. If the value selected for the size of Parzen window is too large or too small, the classification will have a lot of errors .So that the small size of the window, there was probability of exposure of little training data and if training data are outlier data of a class, they certainly make Parzen window classification will have errors in classifying.

Similarly, if the window is selected too large, the number of training data placed in the window of the different classes will be so high that the probability make the right decision between them will have the errors that to select the best size for Parzen window CRC charts have been used.

It should be noted that the number of available training data also will be very important in determining the size of the window. Obvious matter how of the training data set is great, the classification error will be reduced [1].

• *Parzen Window Algorithm* (Figure 3):

- - Get the training and test data
- - Determine an appropriate window
- - Determine number of training data set of each class in window
- - Select the class that most its data has been in the window
- - End

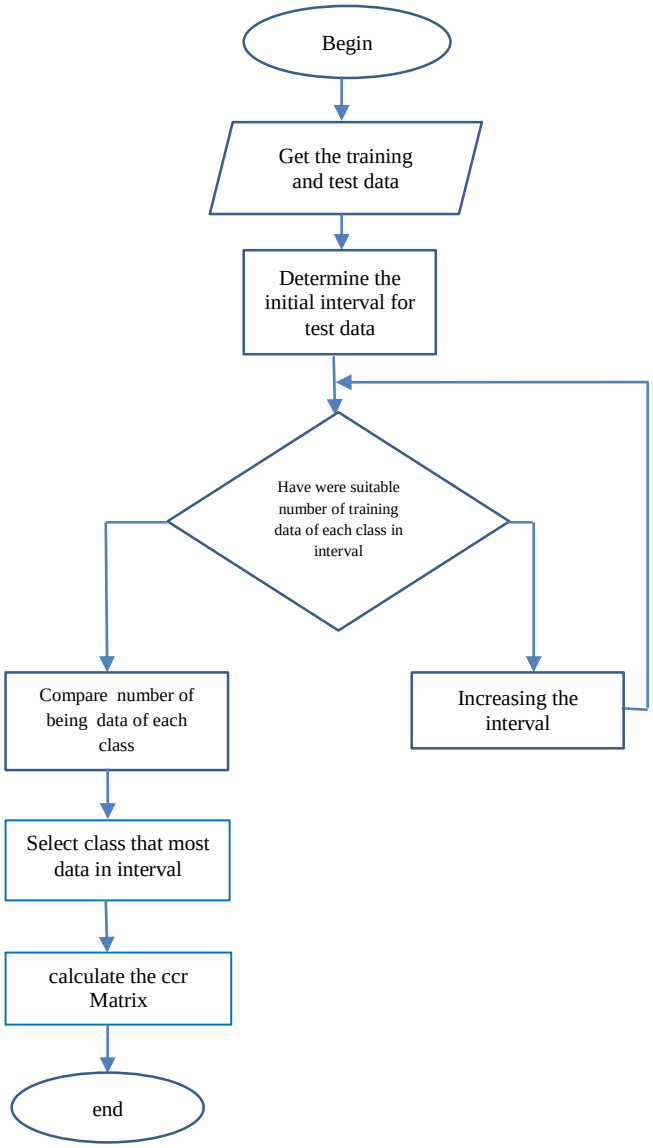


Figure 3. Flowchart of Parzen window classification method

D. Neural Network

A neural network can be defined as a reasoning model based on the structure of human brain. The main property of artificial neural networks is the capacity of learning. Neurons are interconnected through links and every link has a numerical weight associated. Weights are the basic means of realizing the long term memory of artificial neural networks. They express the importance of every neuron. The multilayer perceptron (MLP) is the most widely known and used type of neural network. Most of the times, there are no loops so the output of every neuron does not affect that neuron.

This architecture is called "feed forward". Back-propagation is the most widely known and used supervised learning algorithm. Also called generalized delta algorithm because it extends the means to train a two layer perceptron(delta law), it is based on minimizing the difference between the desired output and the real output, through descent error gradient method .In a back-propagation neural network, the learning algorithm has two phases. First, a training input pattern is presented to the network input layer.

The network then propagates the input pattern from layer to layer until the output pattern is generated by the output layer. If this pattern is different from the desired output, an error is calculated and then propagated backwards through the network from the output layer to the input layer. The weights are modified as the error is propagated. As with any other neural network, a back-propagation one is determined by the connections between neurons, the activation function show in the following function, used by the neurons and the learning algorithm (or the learning law) that specifies the procedure for adjusting weights [6].

The multi-layer perceptron classifier is a basic feed forward artificial neural network. We have used two hidden layers for better classification performance that we show this structure in Figure 5.

$$W_{ij}(k+1) = W_{ij}(k) + \eta \Delta W_{ij}(k) \quad (21)$$

$$\Delta W_{ij}(k) = e_j(k) \times (-1) \times f'(a_2(k) \times W_{ij}) \times a_2(k) \quad (22)$$

where a_2 is transfer function.

If we want to constrain the outputs of a network (such as between 0 and 1), then the output layer should use a sigmoid transfer function, Equation(23), that is shown in Figure 4 .For multiple-layer networks the layer number determines the superscript on the weight matrix. The appropriate notation is used in the two-layer (logsig/purelin).

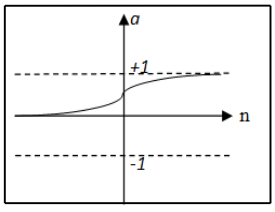


Figure 4. Log-Sigmoid transfer function

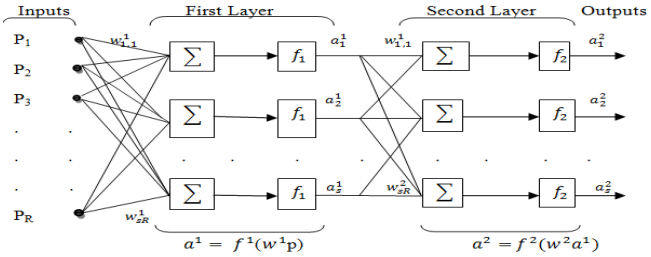


Figure 5. Two-layer perceptron artificial neural network

The sigmoidal function is specific transfer function to be used in the program, ($\alpha = 1$) [7]:

$$F(x(k)) = \frac{1}{1+e^{-\alpha x(k)}} \quad (23)$$

• *Perceptron Neural Network Algorithm* (Figure 6):
- Assigning random values to the weights Gates in [-1,1]
- Perceptron applied to each training data
- If an incorrect evaluation of the second step is repeated if evaluation is Right , algorithm will end

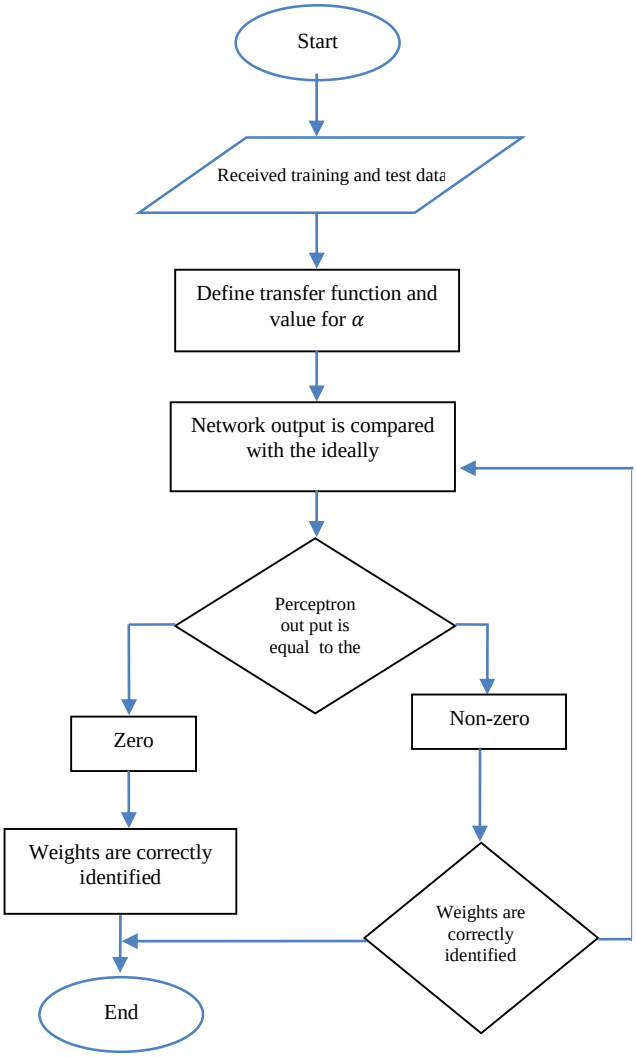


Figure 6. Flowchart of Perceptron Neural Network classification method

E. Divergence

In this study, the separation rate classes for best results we've achieved in this case, with the resulting divergence matrix for 12 classes .according to the following Equation (9) and (10), to calculate the mean and variance of the training data for each class and calculate Divergence Matrix in Equations (24) and (25), that defined amount separation rates classes and said that which classes are closer together and a correct diagnosis is difficult to determine the class for input data. In other words the data is more accurate classification error.

$$D_{ij} = \frac{1}{2}(\mu_i - \mu_j)^t(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) \quad (24)$$

$$d_{ij} = \frac{1}{2}\text{trace}\left\{\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I\right\} \quad (25)$$

where the subscripts *i*, *j* refer to the mean and covariance matrix of a Gaussian distribution corresponding to the feature under investigation for the classes w_i, w_j, respectively [1].

III. SIMULATION RESULTS

In this study, we have examined the Bayesian, *K*-nearest neighbor, Parzen window and multilayer perceptron (MLP) classification on the Standard data image that in MATLAB by simulating in 2009.

In the data are studying, as described above , according to the FDR method, 15 features from 48 existing features, including 95% of data information are selected as the best feature that are presented the effect of feature selection in the different ways classification in Tables 1 and 2.

It can be seen reducing the run time in this methods while performance methods is not very decreased so Important information about the data is not deleted, even Some methods have been better performance.

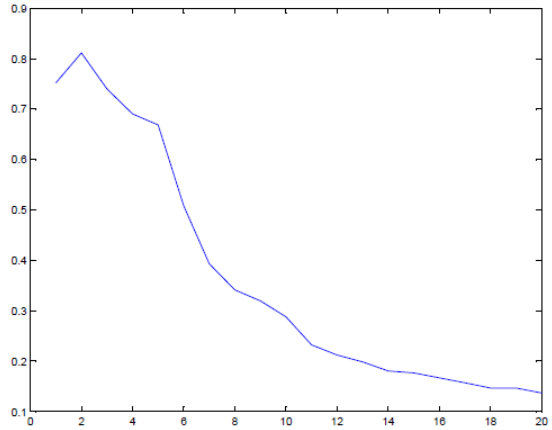
Table 1. Comparison between methods with 48 feature

methods	Bayesian	KNN	Parzen window
Correct Recog (%)	83.93	71.82	74.2
Time of execution (sec)	2.04	1.06	8.93

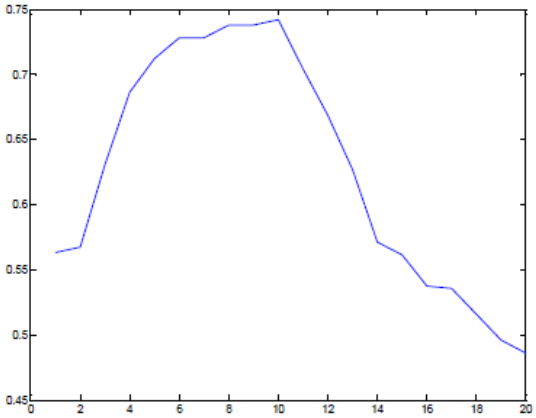
Table 2. Comparison between methods with 15 feature (effect of feature selection on Correct Recog and Time of execution)

methods	Bayesian	KNN	Parzen window
Correct Recog (%)	87.10	78.37	81.15
Time of execution (sec)	0.46	8.00	5.59

In Parzen window method as can be observed in Figure 7, to select the best size for Parzen window about the studied data, we used CRC diagram .We consider 20 as the maximum size for the window that according to the below figure, 10 and 2 is the best size for a window which having the highest rate of Correct classification (ccr) about this data with respectively 48 feature and 15 feature that best feature is selected.



(a)



(b)

Figure 7. CRC diagram in Parzen window method, (a) on data with 15 selected feature, (b) on data with total feature (48)

In Neural Network method, since the sigmoidal transfer function is commonly used in back propagation networks, because it is differentiable, so we use it in this method and get better results. To obtain better comparison results, we were calculated values the percent error or the other way ratio of correct classification (ccr) and running time Related to each of the methods that shown in the Table 3.

Table 3. The comparison classification methods with percentage of correct classification and running time

methods	Bayesian	KNN	Parzen window	multilayer perceptron
Correct Recog (%)	87	51.8	81	82.93
Time of execution (sec)	0.46	74	5.59	285.3

According to the below figure, with increasing epoch in The Multilayer Perceptron neural networks method, reduce the error rate that makes to increase correct classification rate so where that the graph error rate is constant, Figure 8.

As it was said above, in this study, we obtained matrix divergence as a criterion of Separation rates classes that 9, 6 classes, 3, 1, 5 and 1 respectively, are closer together and a correct diagnosis is difficult to determine the class and they will certainly cause errors in the output.

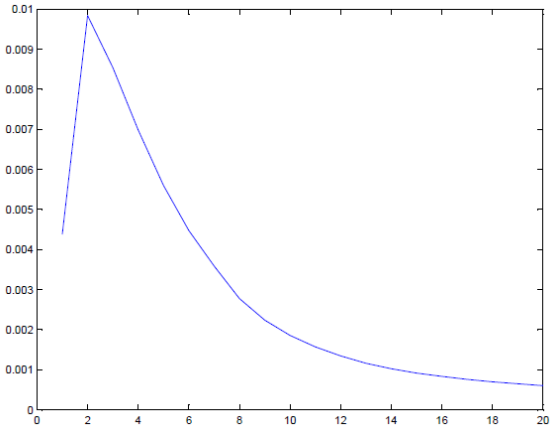


Figure 8. Amount Errors versus Number of Epochs in MLP

IV. CONCLUSIONS

Considering that our goal is to determine the best approach to identify best classification for data but not certainly can say which way is the best way to recognize and outperform all the other classifiers in all the datasets was consistently proved because each method has its advantages and disadvantages with According to the problem and dataset. However, according to our check that we did on certain image data, can say Perceptron Neural networks (MPL),though is more complicated than other methods and its running time is more, but it can be concluded that this method is extensible and development and acts with a less percentage error and high recognition correct classification rate than other methods also by applying feature selection methods ,however reduces the run time in all methods, but we observed that the Bayesian method as a parametric method for this sample data has increased the recognition correct classification rate than other methods.

REFERENCES

[1] S. Theodoridis, K. Koutroumbas, "Pattern Recognition", Fourth Edition, 2008.
[2] A. Joshi, Sh. Bapna, S. Chunduri, "Comparison Study of Different Patter Classifier", 2002.
[3] S. Wang, D. Li, Y. Wei, H. Li, "A Feature Selection Method Based on Fisher's Discriminant Ratio for Text Sentiment Classification", WISM 2009, LNCS 5854, pp. 88-97, 2009.
[4] LA. Nemati, M. Bassiri, "Persian Documents Using Classification Algorithms K-NN", Isfahan University of Technology, Isfahan, Iran.
[5] F. Van der Heijden, R. Duin Dick, www.BookFi.org.
[6] D. Kriesel, "A Brief Introduction to Neural Networks", First Edition, 2007.
[7] H. Demuth, M. Beale, M. Hagan, "Neural Network Toolbox User's Guide", www.mathworks.com.

BIOGRAPHIES

Nasim Molla Esmaili received B.Sc. degree in Electrical Engineering from Shariati Professional and Technical University, Tehran, Iran, in 2012. She is currently M.Sc. student in Electrical Engineering Department, Seraj Higher Education Institute, Tabriz, Iran. Her areas of interest are pattern recognition and digital image processing.

Shilan Hoshiary received B.Sc. degree in ICT Engineering from Roozbeh Higher Education Institute, Zanjan, Iran, in 2013. She is currently M.Sc. student in Electrical Engineering Department, Seraj Higher Education Institute, Tabriz, Iran. Her areas of interest are pattern recognition and digital image processing.

Leila Rastgou received B.Sc. degree in Telecommunications Engineering from Urmia Branch, Islamic Azad University, Urmia, Iran, in 2009. She is currently M.Sc. student in Electrical Engineering Department, Seraj Higher Education Institute, Tabriz, Iran. Her areas of interest are pattern recognition and digital image processing.

Saman Rajebi was born in Tabriz, Iran in 1981. He received his B.Sc. degree in Electronics Engineering from University of Tabriz, Tabriz, Iran in 2003. He received the M.Sc. degree in Communication Engineering from Urmia University, Urmia, Iran. Currently, he is a Ph.D. student at Urmia University. While pursuing his studies for M.Sc., he worked for Urmia ICT Center and Iranian Telecommunication Research Center. He is a fulltime Lecturer at Seraj Higher Education Institute (Tabriz, Iran) and a member of executive committee of ICTPE Conferences. His areas of interests are PLC, microwave applications and modeling and simulation of communication systems and antenna designing.