Feature Selection in Unbalanced Data with Approach of Minimum Overlap between Class Samples

Naser Jahangir (Corressponding author)
Department of Computer, Science and Research Branch, Islamic Azad University, Khouzestan, Iran
naser.jahangir@gmail.com

Reza Javidan
Computer Engineering Department,
Beyza Branch, Islamic Azad University
reza.javidan@gmail.com

Mohamad Hoseyn Yektaie
Computer Engineering Department,
abadan, Islamic Azad University
mh.yektaie@gmail.com

Abstract
Each data set with imbalanced distribution in its classes can be considered as imbalanced data. Text classification, image classification, clustering web pages and risk management are only part common uses of these data sets. Different methods for processing imbalanced data have been proposed. Among these methods, feature selection is one of the newest and most effective methods. The goal of feature selection methods is selecting the subset of features that the classification process has the best performance. Feature selection methods include wrapper, embedded, and the filtering methods. Between feature selection methods, filtering method is one of the best methods for imbalanced data processing. For this reason, in this paper a new method for imbalanced data processing that is suitable for imbalanced datasets with small number of samples and high number of features is proposed. In this approach, In order to evaluate the value of each feature, the distribution function of each class has been separately estimated. Then, the importance of each attribute is obtained based on the relationship between different classes of distribution functions. To demonstrate the effectiveness of this method, it is compared with previous well known methods. The results verify that the proposed method offers an optimization and simple model on imbalanced datasets in comparison with other methods.
Keywords: Feature selection, Imbalanced data, Classification

INTRODUCTION

Due to the extensive usage of imbalanced data , processing of these data has been considered by many researchers in most of the real world problems such as learning machines and searching data. the imbalanced dataset is a set of data which its one or some classes have a considerable number of samples in comparison to other classes [1].

the rate of being imbalanced for a set of data may be so much. In the set of data with low number of samples – about some hundreds or even lower number of samples -10 or 20 samples from maximum class may be existed for each one of samples from minimum class .so ,different methods have been represented for processing the unbalanced data. The feature selection method is one of the newest methods which has been considered by the researchers. The purpose of feature selection methods is to select a j-number set of features

, in the way that ,if the classifier is taught with that set of features , it has the best efficiency. J is a parameter which is determined by the user one of the basic step is the process of feature selection in most of the machine learning algorithms; especially , when the number of features in the studied data set is also very high . when the number of features is high, we may face with the phenomenon "aspect hardship ".in this case ,the accounting expense is very much for training a class-divider and an intense decline is observed in the efficiency of class divider[2].

The feature selection Criterion are used for solving the problem of imbalanced data in comparison to other methods .So, a feature selection method has been suggested which has been based on the hypo thesis , so that a good feature is a feature that samples of one class have the lowest rate of overlapping with other classes, samples ;for each amounts of this feature . on the other words ,that feature can separate samples of different classes much more . Therefore ,the feature probability distribution function estimation is used in order to obtain a suitable knowledge of distribution status of samples in different classes and ; then , the significance and privilege of the feature is determined regarding the relation states of each features distribution function in different classes.

This study has been divided in to five parts. In the continues; in the second part , the most important feature selection Criterion , which have been suggested until now , will be explained for processing the imbalanced data .In the third part , the suggestive feature selection method is expressed completely . in the forth part ;firstly , the class-dividers , evaluation Criterion and data sets used in experiments , have been defined. Then , the suggestive feature selection methods have been compared with previous methods. In the fifth part , conclusion and subsequent works have been gathered .

RELATED WORKS

Different methods suggest processing the unbalanced data that the Criterion of the feature selection are more efficient to solve the problems of the unbalanced data compare to the other methods. In this section, the methods of feature selection have explained briefly. The features divide in two general categories: positive features and negative features. The positive features determine the membership of one class and negative features determine the nullity of membership in one class [3]. Some feature's Criterion are one sided and the other are two sided, The Criterion of feature selection is one sided or two sided as regards they select the positive features or combination of positive and negative features, The Criterion of one sided feature selection, consider the grant of each feature with its mark, so they select only the positive features. In contrary the Criterion of two sided feature selection, consider the modulus grant of features for ranking. So the two types of features- positive or negative have the selection facility. Beside the feature selection divide in two general categories; the Criterion of binary feature selection and the Criterion of binary feature selection use only for the binary data but the Criterion of continual feature selection use on the continual data directly and without any preprocessing[3]. In following, some methods of feature selection's Criterion with deliberate [4]. The Criterion of feature selection, CHI-squared Criterion is based on the confusion matrix and attain the independence of one feature with the class label of feature. This criteria, is the two sided Criterion that is generalization on the nominal data but it is not applicatory on the continual data directly [5]. The reason of this behest is that the aptness of attaining from the continual distribution, is zero, Even if we assume that the aptness product of special amount from one distribution is not zero, each special amount of one, Feature usually repeat only one time and it cause to minimize the frequency[5]. In formation Gain (IG) measure the decreasing amount in the role of feature selection's Criterion that observe in nullity discipline of feature related to the class label; Nullity discipline is the random variant like the class label variant that indicate the nullity pragmatism about the examination result of random variant. When the class samples antecede to being similar, the nullity discipline increase in class labels. The conditional nullity discipline determine the nullity pragmatism about the examination result of random variant, while the examination

result of other random variant determine like the features of data series. After computing the conditional nullity discipline, IG compute with the residuum of these dealt from each other. The Criterion of IG feature selection, is the two sided criteria. This criterion like Criterion of CHI squared feature selection is generalizative on the nominal data but it is no useful for the continual data due to the similar reasons for CHI square criteria. Pcc is a statistic test that evaluate the power and quality relations between two random variant. Correlation coefficient put in the range of -1 up 1[6]. The modulus of correlation coefficient, indicate the relation between two random variant and it's power. If the modulus closer to the 1.the relation between the two random variant get stronger. The correlation coefficient mark shows the direction in relation. If the correlation coefficient mark is positive, the two variant will increase and decrease together [6]. But when the mark of this coefficient is negative, the increasing of one variant is in the direction of the other's decreasing and vice versa the two sided Criterion [6].fast (Feature assessment by sliding Thresholds) consider some different decision making boundary for classify the samples and collect the information relate to the data classification with each of the decision making boundary. Each decision making boundary length give the true positive and false positive and create the ROC diagram by using this information. If the area under the chart (AUC) be more, the ability of that feature that create base on the single feature classifier, will be more in separation of the classes [1].

The suggestive feature selection method

in the criterion of suggestive feature selection the estimation of feature probability distribution function is used for a suitable knowledge about the status of samples distribution in different classes. Then , regarding the relation status of distribution function of each feature in different classes , the significance and privilege of each feature is determined. The first step , in this method, is to estimate the probability distribution function of each feature in different classes .The estimation methods of distribution function are divided in to two general categories ; parametric and non-parametric[7].

Parametric methods suppose a special method for distribution and ; in this way , the problem of probability distribution function estimation for a set of samples changes to the problem of determining the considerable distribution parameters. One of the problems of parametric methods is that there is no any pre-determined structure in many cases on which a special model is supposed for data distribution .Also . all the classic parametric distributions are Single exponential ; it means that they have only one maximum point . while , most of that data sets of real world are Multi-exponential or ; in other words , they have some maximum point and ; consequently , the distribution function which is estimated with parametric methods for such data sets , is not suitable[8].

The non-parametric methods don't suppose anything about the figure and structure of samples , distribution function and they calculated the distribution function from samples directly ; for this reason , non-parametric methods are considered more than parametric methods[8].

for this reason ;in the suggestive feature selection criterion , the non-parametric method has been also used in order to estimate the distribution function of features in different classes .The general formula of non-parametric estimations of the probability distribution function has been used in the formula:

$$P(x) \cong \frac{K}{N \times V} \qquad (1)$$

p(x) shows the estimated amount of probability distribution function for sample x , v is the volume around the sample , N is the whole number of samples and also K is the number of samples within volume V. these

concepts have been shown in the following picture .Regarding the condition of parameters K and V , the non-parametric estimation methods are divided in to two general groups: for that group of methods in which the amount of K is supposed constant , each sample is considered as X and V is determined in the way that it certainly includes K sample K methods are called the nearest neighbor estimation . that group of methods which suppose V a constant amount and obtain the number of points existed in volume V in order to estimate the probability distribution function , are called distribution estimation methods based on kerned[8] .

the estimated probability distribution function with KNN method is sensitive to noise existed in sample and it has long tails which causes the integral under the estimated distribution function doesn't become one and ; consequently , the estimated distribution function doesn't have the main conditions of a probability distribution function .Also , the estimated probability distribution function is not continuous[9].

While, estimation methods KDE don't have such these problems. One of the KDE estimation methods is the Parzn window estimation method which considers a super cube with length one in all dimensions and the kerned considered function is defined in this method in the way that its result for each sample which its interval from sample X is lower than 0.5 in all dimensions . it should be equal to one or equal to zero. The Parzn window method has so many problems such as the distribution function resulted from this method is un continues and all the point which are in volume V around X point have a same share in determining the amount of probability distribution function for point X. if the interval of one point from X is farer, it should have lower share in determining it probability distribution function ; so that these problems become extended by using a soft kernel like Gaussian kernel.

Due to the mentioned reasons , for this feature selection method ; the KDE method with Gaussian kernel has been used for estimating the probability distribution function. The suggestive feature selection method has been based on this by thesis that a good feature is a feature that ; regarding its amounts , samples of one class have the lowest overlapping rate with samples of other classes .In other words, that feature can separate samples of different classes more. So , the rate the estimated distributions function overlapping with that features amounts should be calculated in the maximum class and minimum class in order to evaluate one feature with the considerable Criterion .The overlapping figure for feature f in class cl is calculated according to the following formula.

$$\text{(over lapping) (f,cl)} = \int \min(PDF(cl)\ \max(PDF(cl_j))) \quad (2)$$

where $1 \le j \le 1\,\text{num-class},\ j \ne cl$.

the above formula , PDF(CL} shows the estimated probability distribution function for feature f in class cl and number of classes existed in the data set is investigated .By considering the distribution function of one feature in different classes , each one of the probability distribution function in a special dingle of the whole space for samples is a maximum one .Each one of these dingles in which the lable of maximum class changes , is called a part .if the lable of maximum distribution function is considered for all the samples in each part[8] , then the amount of Separation ability of feature f in class cl is calculated according to the formula.

$$\text{Discriminant ability (f , cl)} = \text{(1-overlapping) (f,cl)}. \quad (3)$$

The Separation ability for a feature is equal to the average of its Separation ability assurance is existed for separating samples of different classes and ; so the probability of false classification will be decreased; consequently, that feature will have a higher privilege. By considering this principle that its samples of different classes should be for from each other to a possible extent in order to decrease the probability of

false classification for the test samples, the number of times in which the label of different parts changes should be counted. This number is called num change and the higher the num change, the more scattered the different classed has been observed; so that feature, sprivilege will have lower validity.

Rank(f)= num-classes $\sum_{cl=1}^{num-classes}$ Discriminant ability (f,cl) (4)

The privilege of each feature is calculated according to the following formula. Features which have higher privilege will have lower (better) rank .The best feature is a feature that classifies all samples with complete assurance (the separability probability is equal to one) and samples of its different classes are completely separated from each other (the amount of num change for such feature is equal to two). The highest privilege, which is equal to 0.5 , is related to such this feature .On the contrary ,the worst feature is a feature which the estimated distribution function are all conformed with each other in its different classes ; in other words , amounts of that feature are the same in all classes .Such this feature with assurance probability zero separates samples of different classes and ; therefore , the lowest privilege ,equal to zero , is related to that .The suggestive feature selection Criterion gives chance to positive and negative features to be selected for classification process .The following picture shows the algorithm related to that through a pseudo-code.

Algorithm 1: The steps of the proposed method for feature selection method

Input: D={d_1, d_2, …, d_n} // A data set containing N labeled instances
Input: F={f_1, f_2, …, f_m} // A data set containing m features
Input: CL={cl_1,cl_2,…,cl_n} // A set containing all class labels
Output: $F^{(ranked)}$ // List of ranked features (desired features are receiving lower ranks)

for*f*=1 **to** num_features **do**// except the class label
Step 1. Estimate the probability density function of feature f in each class as
PDF(cli),1<i<num_classes
Step 2. **For** cl=1 **to** k **do**
Step2-1. Compute the overlapping area of feature f in class cl which is the
overlap between PDF of class cl with PDF of other classes,
according to following formula:
Overlapping(f,cl)= ∫Min(PDF(cl),Max(PDF(clj)));
Where 1=<j=<num_classes and J<>cl
Step 2-2 .Compute the non overlapping area of feature in class via the
following
formula which is a good indication for the discriminant ability:
Discriminant Ability(f, cl)=(1 - Overlapping(f, cl))
//Because the area under PDF curve over all sample space is always one
end for
Step 3. Enumerate the number of changes as Number of changes refers to the
number of times that instances' labels toggle from one class to another class
along the PDF of a particular feature. For a given PDF, instance's label is
simply the class having maximum probability (PDF value)in that point.
Step 4. Determine the score of feature based on the following formula:

$$Score(f)= \frac{(1/num\_classes) * \sum \text{Discriminant Ability}(f, cl)}{\text{Num Changes}}$$

end for
Sort features according to their scores in descending order

EXPERIMENTAL RESULTS

the sets used in the experiments include data which are from data set UCL.Talde1 shows a brief explanation of data set used in the experiments. the first column of table no.1 shows the name of the data set. The second

column of table no.1 shows the whole number of samples and the third column shows the whole number of samples of each category and ; at the last column , the whole number of features has been shown.

Table1: describing the data set used in experiments.

#Features	#Instances Per Class	#Class	Name
60	97،111	208	**SONAR**
34	126،225	351	**IONOSPHERE**

In this study , different feature selection Criterions been compared to each other and the best Criterion will be defined .in order to evaluate a special feature selection method ; the classification efficiency , which has been trained with the selected features set by that method , should be compared with the classification efficiency which has been trained with the whole features set. The closest neighbor classificatory is a learning algorithm based on the sample which makes the model directly from samples of educational set in the memory and when it receives a request for classifying a sample of one new test , it calculates the Euclidean internal of the tests sample and all the educational samples and also it returns the lable of the closest educational sample to that sample of test as the result of that samples classification[10] . the learning algorithms , while processing the imbalanced data , often tend to lable the test sample as a sample of maximum class ; because the separating function a useful for maximum class in these algorithms due to the maximum class being dominant .So, usually , other evaluation Criterion are used for measuring the efficiency of learning algorithms in imbalanced data by researchers which they focus on the importance of minimum class. AUC evaluation Criterion and F Criterion which are from the most famous evaluation Criterion used in processing imbalanced data have been measured.

In this study , the efficiency of six superior feature selection Criterion has been measured with evaluation Criterion AUC and criterion F1 which is one of the most famous evaluation Criterion used in the processing of imbalanced data. The efficiency of feature selection Criterion has been tested by training the closest neighbor and by selecting a number of 10,25,50,100,200,500,1000 features. In graphs no-1 and 2; it shows the efficiency average based on evaluation Criterion AUC ,F1 for the classificatory of the closest neighbor.

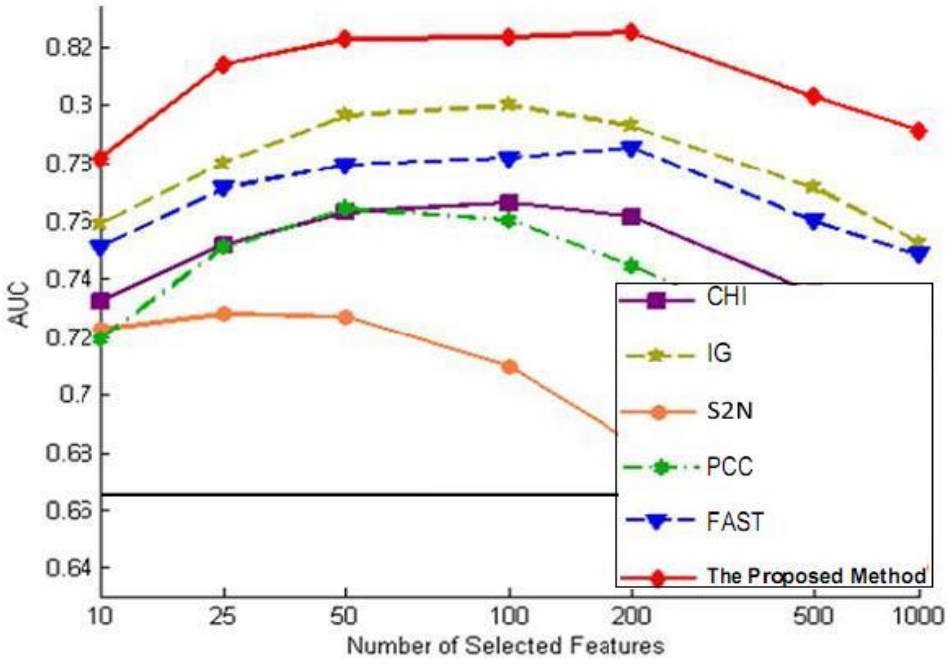

Figure1: efficiency mean are based on the evaluation Criterion AUC for the classifier of the nearest neighbor

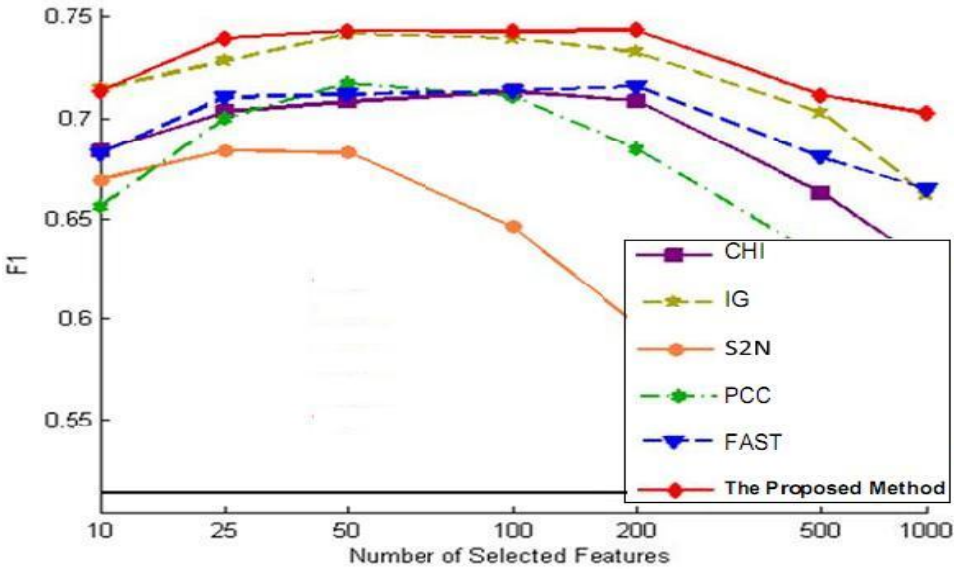

Figure2: efficiency mean are based on the evaluation Criterion F1 for the classifier of the nearest neighbor

As it is observed ;when the number of very low features is considered 10 feature, the best feature selection Criterion are suggestive feature selection and IG Criterion . and for classifying the closest neighbor , after suggestive feature selection Criterion , LG feature selection Criterion will have the best efficiency.In the general condition , for processing data based on imbalanced text , the suggestive feature selection Criterion will be the most suitable choice.On the basis of F1 evaluation Criterion , when only F1 selection with very

low number of features(10 ones) is considered , the efficiency of the suggestive feature selection Criterion and 1G are the same. On the basis of each two evaluation Criterion ; in cases with selecting more features (50 or 100 features) , the purpose is to find the optimal model .The suggestive feature selection Criterion has the best efficiency on the basis of two evaluation Criterion AUC and F1 in the general condition.

Through selecting much more features (more than 100 features) , an intensive decline will be observed in the efficiency of classification; however , in the common condition ,it is expected that the efficiency process of classification becomes rising until time that about a half of primary features haven't been selected and ; after wards , it becomes falling due to the attendance of unrelated and noisy features in order to reach the amount of based method .For researching more for this purpose , the efficiency of different feature selection Criterion has been investigated on two data sets lonosphere and sonar with both evaluation criteria. Graph No.3 to 6 show the results of different classifications on these two data sets with different evaluation Criterion F1 and AUC.

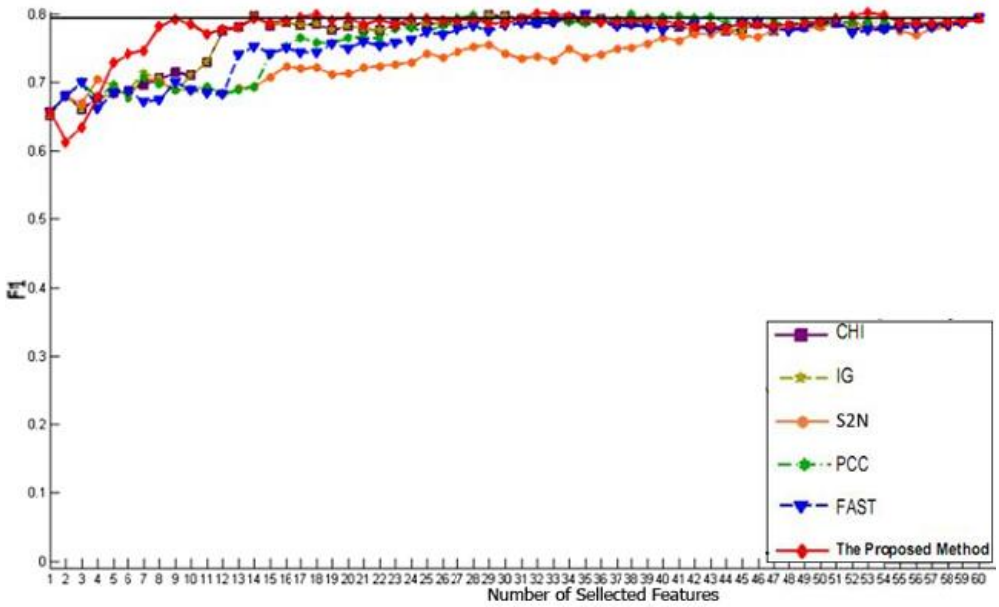

Figure3: efficiency mean are based on the evaluation Criterion F1 for the classifier of the nearest neighbor on Sonar data set

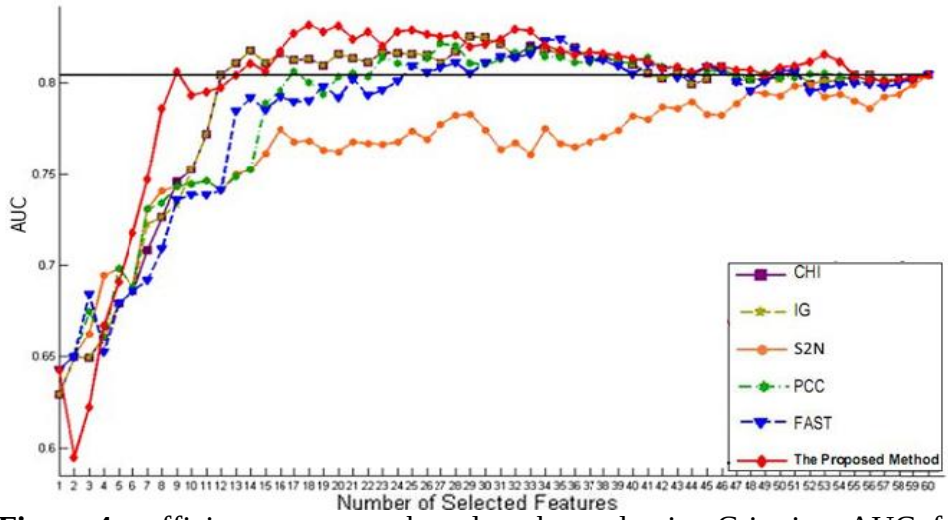

Figure 4 : efficiency mean are based on the evaluation Criterion AUC for the classifier of the nearest neighbor on Sonar data set

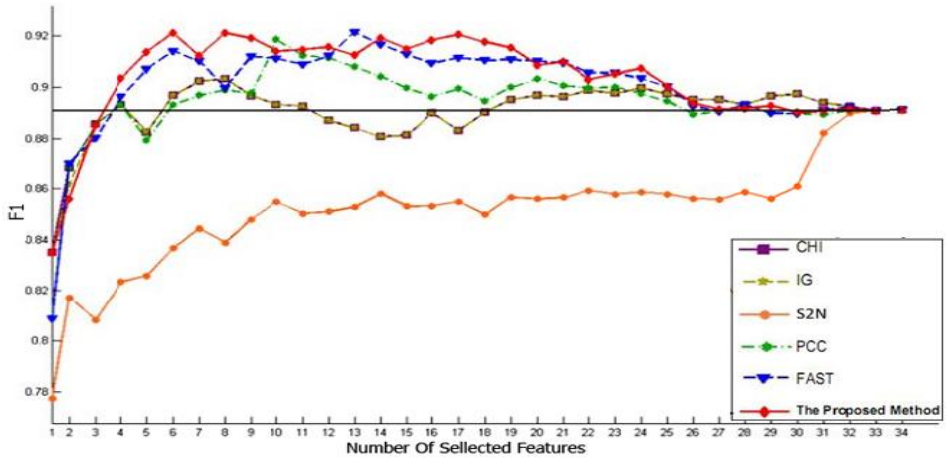

Figure5: efficiency mean are based on the evaluation Criterion F1 for the classifier of the nearest neighbor on IONOSPHERE data set

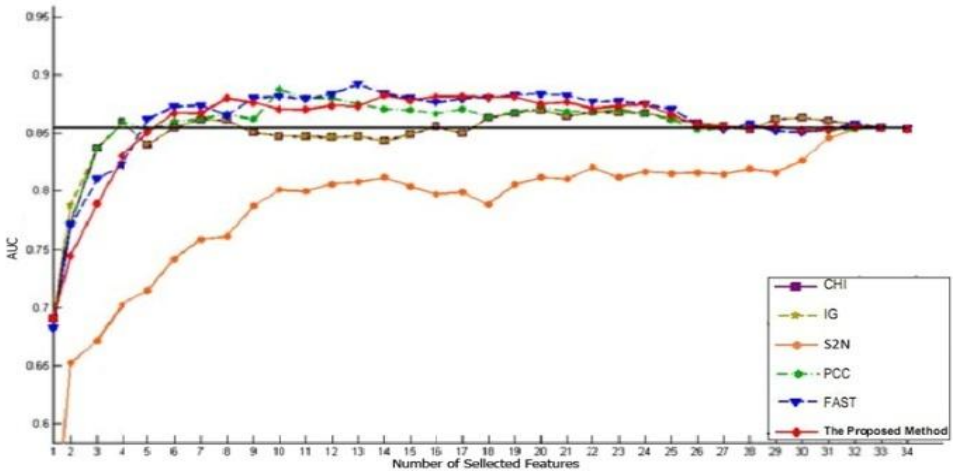

Figure6: efficiency mean are based on the evaluation Criterion AUC for the classifier of the nearest neighbor on IONOSPHERE data set

Aside from the fluctuations which is observed in the efficiency process of classifications on those two data sets, the obtained results confirm the above expectation .The fluctuations which are observed in the efficiency process of these classifications is because of this reason that the selected new feature ; despite being adequate for training the classifications, has repeated information or opposite to the present selected features set and ; consequently , its addition to the present selected features set causes the efficiency to be declined and it is resulted from this matter that the feature selection Criterion don't consider the relation among features for selecting the feature; so that they only measure the usefulness of one feature apart from other features .The reason of this sudden decline in the efficiency is the existence of large number of noisy features in these data sets which make the learning process to face problems .Therefore, the feature selection Criterion of a suitable method are with high dimensions for processing imbalanced data ; because through performing these methods , the volume of imbalanced data set will be intensively decrease and also a very high increase will be observed in the efficiency of classifications due to the elimination of noisy and unrelated features .When the permissible changing range is low for the best efficiency (especially when the permissible changing range is 1 percent) the suggestive feature selection Criterion will have a better efficiency in comparison to the other feature selection Criterion , and this improvement in the efficiency is very note worthy for evaluation Criterion AUC Moreover , through increasing the number of selected features , the difference in the efficiency of the suggestive feature selection Criterion will be more considered .Through increasing the range of permissible changing , the difference in the efficiency of different feature selection Criterion will be decreased .So , the best choice is the suggestive feature selection Criterion and ; after words , feature selection Criterion are respectively 1G and FAST for the classifications of the closest neighbor.

CONCLUSIONS

common techniques of learning machines and data –searching often face problem while processing imbalanced data which are also common in most of the operational fields . For solving this problem , researchers have suggested many new methods for processing the imbalanced data. For decreasing number of features in data sets with very high number of features , the feature selection methods are often used as the common way ; but the importance and effect of feature selection methods has been recently considered by the researchers for solving the problem of processing imbalanced data. In this study , one feature selection Criterion has been suggested as the feature selection Criterion based on features distribution function.

Among the five feature selection Criterion , which have been studied here , the suggestive feature selection Criterion has the best efficiency based on both evaluation Criterion AUC and F1.Since , in this study , much concentration is on the imbalanced data set with high dimensions and low number of samples , the effect of imbalanced data sets samples number and also the number of features should be investigated in the future studies for the dimension of the feature selection criteria.

REFERENCES

[1] X. Chen and M. Wasikowski. "FAST: a roc-based feature selection metric for small samples and Imbalanced data classification problems". In Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Las Vegas, Nevada, USA, August 24 - 27, 2008). KDD '08. ACM, New York, NY, 2008, pp. 124-132.

[2] N.Chawla,N. Japkowicz and A. Kotcz. "Editorial: Special issue on learning from imbalanced data sets". SIGKDD Explorations, vol. 6, no. 1, 2004, pp. 1–6.

[3] Z. Zheng, X. Wu and R. Srihari. "Feature selection for text categorization on imbalanced data". SIGKDD Explorations, vol. 6, 2004, pp. 80–89.

[4] I. GuyonandA.Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, 2003, pp. 1157–1182.

[5] G. Forman. "An extensive empirical study of feature selection metrics for text classification". Journal of Machine Learning Research, vol. 3, 2003, pp. 1289-1305.

[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, 2003, pp. 1157–1182.

[7] A.R. Webb, "Statistical Pattern Recognition", Second Edition, Wiely, 2002.

[8] R.O. Duda, P.E. Hart and D.G. Stork, "Pattern Classification", Second Edition, Wiley, 1997.

[9] C.M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[10] T. Mitchell, Machine Learning. McGraw Hill, 1997.

Bio

Naser Jahangir was born in 1985. He received the B.S. and M.S. degree from Shiraz Branch, Islamic Azad University and Science and Research Branch of Khouzestan, Islamic Azad University, respectively. His research interest is in the area of Data mining.
.

Reza Javidan was born in 1970. He is graduated from MSc Degree in Computer Engineering (Machine Intelligence and Robotics) from Shiraz University in 1996. He received Ph.D. degree in Computer Engineering (Artificial Intelligence) from ShirazUniversity in 2007. His major fields are artificial intelligence, image processing and sonar systems. Dr. Javidan is now assistant professor and lecturer in Department of Computer Engineering in Islamic Azad University, Beyza Branch.