International Journal of Computer Architecture and Mobility
(ISSN 2319-9229) Volume 3 -Issue 8, October 2015

Neighborhood Covering Reduction by Ada-Boost Ensemble Learning Approach

Pavitra Sharma
Oriental University, Indore
sharma.pavitra07@gmail.com

***Abstract*—Predictive accuracy, computational speed, scalability and robustness are criteria for the evaluation of classification and predictive methods. Existing rule learning techniques having some limitation related performance. The known nearest neighbor methods are robust with the variable data set but they are sensitive to input data set. In this work, we introduced ada boost ensemble learning for reduction of randomized attributes. We focused on to nearest neighborhood classifier for defining attribute. Random attribute selection can be used to obtain a reduces representation of the data while minimizing the loss of information content because stream data content large data set with the variations and this approach gives the effective results for high dimensionality of the data. Experimental results show that if we combine the nearest neighbor classification with the randomized selection of attributes for stream data than this will be effective method handling possible uncorrelated errors and noisy data.**

I. INTRODUCTION

Currently, many applications use extremely large data sets of high dimensionality, thus classifying, understanding or compressing this information becomes a very difficult task. Data and web mining, text categorization, financial forecasting, and biometrics are some examples of domains in which huge amounts of information have to be employed. The processing of a very large data set suffers from a major difficulty: high storage and time requirements. On one hand, a large data set cannot be fully stored in the internal memory of a computer. On the other hand, the time needed to learn from such a whole data set can become prohibitive. These problems are especially critical in the case of using some distance-based learning algorithm, such as the Nearest Neighbour rule due to the apparent necessity of indiscriminately storing all the training instances [1, 2].

II. LITERATURE SURVEY

In the last decade, we have witnessed great progress in rough set theory and its applications. Initiative work in the beginning of 1980s has become a powerful tool to deal with imperfect and inconsistent data, and extract useful knowledge from a given dataset.

Freund et al. reach the same conclusion by showing that boosting (called arcing by Breiman) can reduce both bias and variance in an example using stumps. Breiman built a new framework for randomization methods in terms of strength and correlation, and provides an upper bound for the generalization error of a random forest in terms of them. C. J. Merz, Using correspondence analysis to combine classifier it is idea of ensemble methodology is to build a predictive model by integrating multiple models.

III. PREPARE

The overwhelming amount of data that is available nowadays in any field of research poses new problems for data mining and knowledge discovery methods. This huge amount of data makes most of the existing algorithms inapplicable to many real-world problems. Two approaches have been used to face this problem: scaling up data mining algorithms and data reduction. Nevertheless, scaling up a certain algorithm is not always feasible. It has been shown that different groups of learning algorithms need different instance selectors in order to suit their learning/search bias. This may render many instance selection algorithms useless, if their philosophy of design is not suitable for the problem at hand. Our algorithm does not assume any structure of the data or any behaviour of the classifier, adapting the instance selection to the performance of the classifier [5].

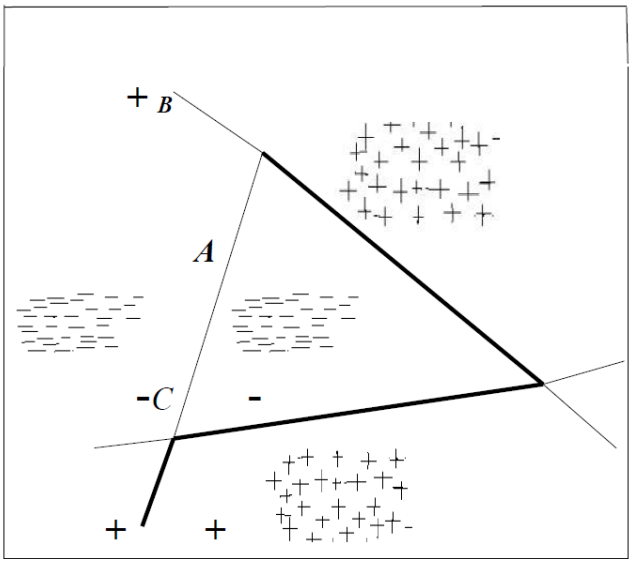


Fig. 1. An example of single linear classifier.

IV. PROPOSED WORK

The purpose of supervised learning is to classify patterns (also known as instances) into a set of categories which are also referred to as classes or labels. Commonly, the classification is

based on a classification models (classifier) that is induced from an exemplary set of pre classified patterns.

The key idea of Boosting algorithm is to transfer a nearest neighbour classifier to a strong one by integration and train for attribute reduction. Adaboost algorithm is a kind of Boosting algorithms, which is an adaptive Boosting one. Adaboost algorithm can adjust weight distribution of the training samples adaptively, and consistently select the best nearest neighbour classifier of sample weight distribution, to integrate all nearest neighbour classifier and vote by a certain weight to form a strong classifier. Adaboost algorithm combines nearest neighbour classification with index selection, and reaches the key indexes selection on the basis of forecast accuracy.[10]

1) Given train sample set neighbour classifier space H, $x \in X$, $S = \{(x_1,y_1),(x_2,y_2),\ldots\ldots\ldots\ldots.(x_n,y_n)\}$X is a sample space, $y = \{1,2,3,\cdots,K\}$ is a class label set. Initiating sample probability distribution $D_t(i) = 1/n$,i=1,2,…,n.

2) For t=1,2……. T, T is the feature numbers needed.

To every nearest classifier h of H, we can do below:

a) Dividing sample space X, we can get $x_1, x_2, \ldots\ldots\ldots\ldots\ldots.x_m$

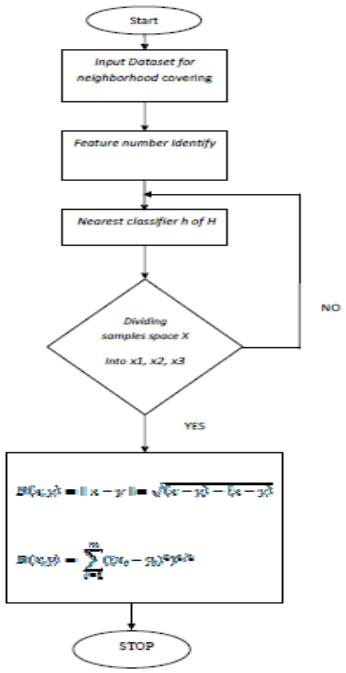
b) Under the training sample probability distributionD, we can calculate

$$D(x,y) = \| x - y \| = \sqrt{(x-y)-(x-y)}$$

$$D(x,y) = \sum_{i=1}^{m}((x_i - y_i)^2)^{1/2}$$

Improving accuracy, robustness and understand ability is the objective of classification modelling. Regarding instability and performance limitation of existing rule learning techniques, we introduce an ensemble classifier based on randomized neighbourhood reduction and neighbourhood covering reduction.

Flow Chart



Algorithm

start

Input for neighborhood covering with feature

for (i=1:n)

if num_cur=0

array cur[]

else

array_cur(:,num_cur)= data_array(:,feature_slct(num_cur)) # add the new feature

end

end

for i=1:length

array_tmp=array_cur # compute the significance of features

Compute the significance of newly added feautre and add it to efc_tmp

for i=1:m1 # find the neighborhood of xi

if d>1

sqare_distance=sqare_distance+1 #for categorical features

else

sqare_distance=sqare_distance+d # for numerical features

end

select the best feature

```
        if (length(feature_lft)>=N)
        randN=unidrnd(N)
        else
        randN=unidrnd(length(feature_lft))
        end
        if ( num_cur>0 & max_efc-sig(num_cur)<efc_ctrl)
        num_cur=n-1
        else
        sig(num_cur+1)=max_efc
        feature_slct(num_cur+1)=feature_lft(max_sequence1
)
        feature_lft(max_sequence1)=[]
    num_cur=num_cur+1
    end
```

IV. IEXPERIMENT AND RESULT ANALYSIS

The performance for our algorithm can be measured either by its efficiency or effectiveness. Efficiency describes the time taken in the learning the classifier and/or the time taken to classify the test cases. Efficiency becomes very important when it comes to experimental comparison between different learning algorithms. The best measurement criterion for the single label problem is the classification accuracy. Our work is based on the real datasets for evaluating the performance of the algorithms and collected it from from the UCI Machine Learning Repository [10] .Table 1 shows the characteristics of the real datasets used in our experiments. It shows the number of items, number of transactions and size for the each datasets. For all experiments, we used four datasets with different characteristics. Thus, the advantages and disadvantages of the algorithms can be observed. We used four real datasets which are Balance Wine, Iris, Wdbc and pima.

Table 1 Description of datasets

Dataset	Features	Class	Instances
Wine	14	3	178
Iris	5	3	150
Wdbc	31	2	569
Pima	9	2	768

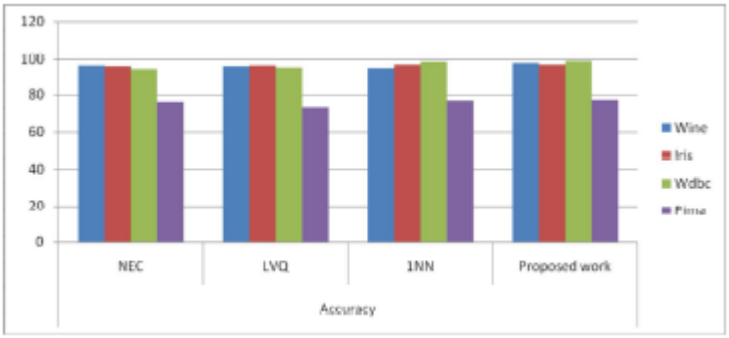
Result Analysis

Below are the results of some of the more significant tests performed using by nearest neighbour classification based on ensemble method called as adaboosting ensemble method. In this section, we first compare the accuracy of these algorithms on different datasets. Then we present the comparative results with some classical feature selection algorithms. Finally, we discuss the influence of parameters in the neighbourhood model.

Table 2 Accuracy of the data

Dataset	Accuracy			
	NEC	LVQ	1NN	Proposed work
Wine	96.6	96.0	94.9	97.6
Iris	96.0	96.6	96.8	96.9
Wdbc	94.6	95.4	98.7	99.0
Pima	76.0	73.3	76.6	76.9

Furthermore, we collect ten classification tasks from UCI machine learning repository. The description of these data sets is given in Table 1. The accuracies based on nearest neighbour rule (NN), neighbourhood classifier (NEC)[9], LearningVector Quantization (LVQ) [8] and proposed work are presented in Table 2.



CONCLUSION

Reducing redundant or irrelevant attributes can expand classification performance in most of cases and decrease cost of classification. We design a feature evaluating function, called neighbourhood dependency, which reflects the percentage of samples in the decision positive region. Theoretical arguments show that the significance of features monotonically increases with the feature subset. This property is important for integrating the evaluating function into some search strategies. Then adaboosting ensemble feature selection algorithms are constructed based on the dependency function. We have compared the accuracy of some classical feature selection algorithms on different datasets and we are able to obtain improved accuracy over single nearest neighbour rule (NN), neighbourhood classifier (NEC), LearningVector Quantization (LVQ) which have been using in the past. We have checked accuracy in terms of both effectiveness and correctness of approximation the test sample, all these goal have been achieved by ensemble the nearest neighbour classifier using adaptive boosting technique while reduced subset of features or attributes are taken into consideration for improving the accuracy.

Analysis indicates that the performance of any ensemble methods is dependent on the characteristics of the data set being examined. In fact, results show that Boosting ensembles can perform efficiently with trimmed input data stream while considering only those features which are needed to take the decision for recognizing the output.

References

[1] C. J. Merz, "Using correspondence analysis to combine classifier," Mach. Learn., vol. 36, no. 1, pp. 33–58, Jul. 1999.
[2] XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou ·Michael Steinbach · David J. Hand · Dan Steinberg, "Classification Algorithms for Data Mining:A Survey" , Springer-Verlag London, 2007
[3] Nicolás García-Pedrajas, "Constructing Ensembles of Classifier by Means of Weighted Instance Selection", IEEE Transactions on Neural Networks,vol.20,Feb. 2009.
[4] T.Windeatt, "Accuracy/diversity and ensembleMLP classifier design", IEEE Trans. Neural Netw., vol. 17, no. 5, pp. 1194–1211, Sep. 2006.
[5] Cover T, Hart P, "Nearest neighbour pattern classification", IEEE Trans Inform Theory 13(1):21–27,1967
[6] Dasarathy BV (ed) , " Nearest neighbour (NN) norms: NN pattern classification techniques", IEEE Computer Society Press, 1991
[7] Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press 1996
[8] Xindong Wu et.al, "Top 10 Algorithms of Data Mining", Springer-Verlag London, 2007.
[9] Ahmed S, Coenen F, Leng PH (2006) Tree-based partitioning of date for association rule mining. Knowl Inf Syst 10(3):315–331
[10] Alkoot, F., & Kittler, J. (2002a). Moderating k-NN classifier. Pattern Analysis and Applications, 5,326–332.
[11] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," Mach. Learn., vol.36, no. 1/2, pp. 105–142, Jul./Aug. 1999.
[12] H. Liu and H. Motoda, "On issues of instance selection," Data Mining Knowl. Disc., vol. 6, pp. 115–130, 2002.
[13] Yang T, Li Q G. Reduction about approximation spaces of covering generalized rough sets [J] . International Jo urnal of Approximate Reasoning . 2010, 51: 335- 345
[14] Yang T, Li Q G. Reduction about appro ximation spaces of covering generalized rough sets [J] . International Jo urnal of Approximate Reasoning . 2010, 51: 335- 345
[15] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional spaces. In den Bussche,J. V. and Vianu, V., editors, Database Theory - ICDT 2001, 8th International Conference,London, UK, January 4-6, 2001, Proceedings, volume 1973 of Lecture Notes in Computer Science, pages 420–434. Springer.
[16] Kibriya, A. M. and Frank, E. (2007). An empirical comparison of exact nearest neighbour algorithms. In Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), volume 4702 of Lecture Notes in Computer Science, pages 140–151. Springer.