IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 4, April 2016.*

# MINDEX_IB: A Feature Selection method for Imbalanced Dataset

**Hemlata Pant[1], Dr. Reena Srivastava[2]**

Research Scholar, School of Engineering, BBD University, Lucknow, India[1]

Dean, School of Computer Applications, BBD University, Lucknow, India[2]

**Abstract**: It is common to have an unbalanced class distribution in many classification problems. The class imbalance problem is even more severe when the dimensionality is high. One commonly used strategy to improve the classification performance is feature selection. Feature selection is a technique to select a subset of relevant features that allow a classifier to reach optimal performance. Most of the approaches for feature selection methods for imbalanced datasets mainly focus on an imbalanced dataset with two classes and does not work significantly well with a multiclass imbalanced dataset. In this paper, we propose a filter feature selection algorithm called MINDEX_IB, for unbalanced data sets. MINDEX_IB is a filter approach based measure. The proposed measure focuses on efficient partitioning of the attribute domain. Here, partitioning is done via micro-clustering i.e. the process of making micro-clusters. MINDEX_IB outperforms other feature selection algorithms in terms of number of features selected, accuracy and also in terms of performance measures for the imbalanced dataset such as F-measure and AUC evaluation measure.

**Keywords**: Feature selection, Imbalanced dataset, Classification, Filter based approach.

## I. INTRODUCTION

The class imbalance problem present in real world applications is one of the greatest challenges in the field of machine learning and data mining. The class imbalance problem addresses the issue that occur when in a dataset a class or classes have significantly more samples than the other classes of the dataset. Imbalanced classes are seen in a variety of domains including text classification, risk management, web categorization, medical diagnosis/monitoring, biological data analysis, credit card fraud detection, oil spill identification from satellite images and have major economic, commercial, and environmental concerns.

The majority of current research to address class-imbalance problem can be grouped into two categories: sampling techniques [1] [2] [3] [4] and algorithmic methods[5] [6] [7] [8] [9]. The sampling methods perform leveling the class samples by under-sampling the larger class or by over- sampling the smaller one or by combination of these techniques, so that they are no longer imbalanced. Algorithmic methods include an adjust of the operation of the algorithm to treat the unbalanced data.

Pattern recognition, data mining, and machine learning applications have to deal with data sets which are having hundreds and thousands of variables or features present in it. Processing such huge data sets becomes a challenging task. Feature selection addresses this problem and improves the performance of the learning algorithm by removing irrelevant, redundant, or noisy data. There are two wide categories of Feature selection algorithms filter based methods and wrapper methods [10] [11]. In wrappers approach, in order to assess the relative usefulness of subsets of variables, feature selection method uses the prediction performance of a given learning machine. The filters approach on the other hand, is based on the intrinsic properties of the data, rather than being biased toward a particular classifier.

The class imbalance problem is even more severe when the dimensionality is high. While feature selection has been extensively studied [12] [13] [14] [15] [16] [17]. In particular the importance of feature selection to class imbalance problem was recently realized and the machine learning and data mining research community has shown increased attention towards it. To address the imbalanced dataset the feature selection methods should focus on the attributes that are helpful in the identification of minor classes.

## II. RELATED WORK

There is large number of feature selection algorithms; however, very few researches have been targeted particularly towards imbalanced class distributions.

In particular in 2008, Xue-wen Chen and Michael Wasikowski has proposed an approach namely FAST [18] ,a method which is based on the area under a ROC curve generated by moving the decision boundary of a single feature classifier with thresholds placed using an even-bin distribution.

In 2011, Mina Alibeigi et.al. proposed unsupervised feature selection method Based on the distribution of features attributed to imbalanced data sets [19]. This method removes redundant features from the original feature space based on the distribution of features.

In 2013, Ilnaz Jamali and Sattar Hashemi proposed Feature Selection method FSSH based on Shapley Value [20]. This method first constructs some coalitions. According to the AUC value of the coalition, the marginal importance for each feature is computed. The average of marginal importance of each feature is computed and is called as the shapely value. All features are ranked from maximum to minimum value according to their Shapley value.

In 2011, German Cuaya et.al proposed a minority class feature selection method FSMC for unbalanced data sets [21]. FSMC selects attributes that have minority class distributions significantly different from the majority class distributions.

In 2014 D. Tiwari proposed feature selection algorithm for imbalanced datasets by modifying the original RELIEFF algorithm to address the class imbalance problem [22]. This method assigns higher weight to attributes while dealing with minority classes which results in higher weight of attributes which cater to minority samples.

Most of the approaches available for feature selection from imbalanced datasets mainly focus on an imbalanced dataset with two classes and does not work significantly well with a multiclass imbalanced dataset. In this paper, we propose a filter feature selection algorithm named MINDEX_IB which is a filter approach based measure that focuses on efficient partitioning of the attribute domain via micro-clustering i.e. the process of making micro-clusters.

## III.MINDEX_IB

As discussed earlier MINDEX_IB is a filter approach based measure that focuses on efficient partitioning of the attribute domain via micro-clustering i.e. the process of making micro-clusters. A micro-cluster is a small hyper rectangle (enclosing some tuples) in a d-dimensional Euclidean space where, d is the number of the attributes used to identify the micro-clusters. If instead of d attributes only one attribute is used, then the micro-clusters formed will give partitions imposed by the involved attribute. The process of main algorithm is as follows.

Algorithm MINDEX_IB first finds the micro-clusters formed with the help of Micro-cluster identification with a single attribute for the concerned attribute [23]. For each Micro-cluster the algorithms finds the number of instances of each class present in each micro cluster and the class having highest number of instances in a micro-cluster becomes the class label of that micro cluster. Finally the algorithm computes the MINDEX_IB of the attribute as the number of micro-clusters having a minority class as its class label. The attributes having higher MINDEX_IB are considered as more relevant.

### *3.1 Algorithm MINDEX_IB*

Input Relation R, Attribute Ai

Output Mindex_IB for attribute Ai. i.e. the no. of micro clusters having minority class as its class label when micro-clusters are formed with respect to attribute Ai .

Algorithm
1. Find the Micro-cluster set M for attribute Ai with the help of Micro-cluster identification with a single attribute as discussed in [23].
2. For each micro cluster in M,
   a) Find the no of instances of each class present in the micro cluster.
   b) Set the class label of each micro-cluster as the class to which highest number of instances present in that micro cluster.
3. Compute MINDEX _IB for Ai as the number of micro clusters having minority class as its class label.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

In order to evaluate the performance of our algorithm MINDEX_IB, we used data sets from the UCI repository, namely, ECOLI, IONOSPHERE, OZONE, and DIABETES as shown in table I. The first, second and third column of table I show the name of the data set, number of features and number of classes in dataset respectively. The last column in each row gives the information about the number of instances per each class in the dataset.

TABLE I: DATA SET USED

| NAME | #FEATURES | #CLASS | #INSTANCES PER CLASS |
|---|---|---|---|
| ECOLI | 7 | 8 | 143, 77, 52, 35, 20, 5, 2, 2 |
| IONOSPHERE | 34 | 2 | 126, 225 |
| OZONE | 73 | 2 | 2463,73 |
| DIABETES | 8 | 2 | 268,500 |

We used five different classifiers to analyze performance measures for these data sets in terms of accuracy, F measure and AUC evaluation measure. Table II to table V, separately illustrate the experimental results on each of the introduced data sets.

TABLE II: PERFORMANCE OF MINDEX_IB ON ECOLI DATA SET IN TERMS OF THE NUMBER OF SELECTED FEATURES, ACCURACY, F- MEASURE AND AUC EVALUATION MEASURE WITH DIFFERENT CLASSIFIERS.

| Classifier | F Measure | ROC | Accuracy | NO. Of Features Selected |
|---|---|---|---|---|
| NB | 0.85 | 0.96 | 85.42 | 4 |
| PART | 0.81 | 0.90 | 80.95 | 4 |
| J48 | 0.81 | 0.90 | 81.85 | 4 |
| SMO | 0.80 | 0.93 | 83.04 | 4 |
| BAGGING | 0.81 | 0.95 | 82.14 | 4 |

TABLE III: PERFORMANCE OF MINDEX_IB ON IONOSPHERE DATA SET IN TERMS OF THE NUMBER OF SELECTED FEATURES, ACCURACY, F- MEASURE AND AUC EVALUATION MEASURE WITH DIFFERENT CLASSIFIERS

| Classifier | F Measure | ROC | Accuracy | NO. Of Features Selected |
|---|---|---|---|---|
| NB | 0.89 | 0.94 | 88.60 | 4 |
| PART | 0.89 | 0.89 | 89.46 | 4 |
| J48 | 0.90 | 0.91 | 90.60 | 4 |
| SMO | 0.80 | 0.76 | 80.91 | 4 |
| BAGGING | 0.91 | 0.92 | 91.17 | 4 |

TABLE IV: PERFORMANCE OF MINDEX_IB ON DIABETES DATA SET IN TERMS OF THE NUMBER OF SELECTED FEATURES, ACCURACY, F- MEASURE AND AUC EVALUATION MEASURE WITH DIFFERENT CLASSIFIERS

| Classifier | F Measure | ROC | Accuracy | NO. Of Features Selected |
|---|---|---|---|---|
| NB | 0.74 | 0.79 | 75.00 | 1 |
| PART | 0.71 | 0.70 | 73.05 | 1 |
| J48 | 0.71 | 0.70 | 73.05 | 1 |
| SMO | 0.73 | 0.67 | 74.61 | 1 |
| BAGGING | 0.70 | 0.76 | 72.01 | 1 |

TABLE V: PERFORMANCE OF MINDEX_IB ON OZONE DATA SET IN TERMS OF THE NUMBER OF SELECTED FEATURES, ACCURACY, F- MEASURE AND AUC EVALUATION MEASURE WITH DIFFERENT CLASSIFIERS

| Classifier | F Measure | ROC | Accuracy | NO. Of Features Selected |
|---|---|---|---|---|
| NB | 0.91 | 0.85 | 86.79 | 6 |
| PART | 0.96 | 0.69 | 97.20 | 6 |
| J48 | 0.96 | 0.49 | 97.12 | 6 |
| SMO | 0.96 | 0.50 | 97.12 | 6 |
| BAGGING | 0.96 | 0.84 | 97.16 | 6 |

Table VI and VII give the performance of NB classifier for Ecoli and Ionosphere data sets respectively in term of AUC evaluation statistics with different feature selection methods including MINDEX_IB.

TABLE VI: THE PERFORMANCE OF NB CLASSIFIER FOR ECOLI DATA SET IN TERM OF AUC EVALUATION STATISTICS

| Feature Selection Method | AUC | NO. Of Features Selected |
|---|---|---|
| Baseline | 0.96 | 7 |
| CfsSubsetEval | 0.96 | 6 |
| ConsistencySubsetEval | 0.96 | 6 |
| SFS with Entropy | 0.96 | 6 |
| PDF based method | 0.96 | 6 |
| MINDEX_IB | 0.96 | 4 |

TABLE VII: THE PERFORMANCE OF NB CLASSIFIER FOR IONOSPHERE DATA SET IN TERM OF AUC EVALUATION STATISTICS

| Feature Selection Method | AUC | NO. Of Features Selected |
|---|---|---|
| Baseline | 0.935 | 34 |
| CfsSubsetEval | 0.935 | 14 |
| ConsistencySubsetEval | 0.926 | 7 |
| SFS with Entropy | 0.82 | 42 |
| PDF based method | 0.958 | 12 |
| MINDEX_IB | 0.936 | 4 |

Table VIII and IX give the average performance of PART, J48, Bagging, SMO classifier across Diabetes and Ozone data set respectively in term of Classification accuracy statistics with different feature selection methods including MINDEX_IB.

TABLE VIII: THE AVERAGE PERFORMANCE OF PART, J48, BAGGING, SMO CLASSIFIER ACROSS DIABETES DATA SET IN TERM OF CLASSIFICATION ACCURACY STATISTICS

| Feature Selection Method | Accuracy (%) | NO. Of Features Selected |
|---|---|---|
| CfsSubsetEval | 69.61 | 4 |
| Filtered subseteval | 70.23 | 3 |
| PrincipalComponents | 70.42 | 8 |
| FSMC | 70.05 | 1 |
| MINDEX_IB | 73.18 | 1 |

TABLE IX: THE AVERAGE PERFORMANCE OF PART, J48, BAGGING, SMO CLASSIFIER ACROSS OZONE DATA SET IN TERM OF CLASSIFICATION ACCURACY STATISTICS

| Feature Selection Method | Accuracy (%) | NO. Of Features Selected |
|---|---|---|
| CfsSubsetEval | 94.96 | 18 |
| Filtered subseteval | 94.96 | 18 |
| PrincipalComponents | 94.01 | 19 |
| FSMC | 95.38 | 8 |
| MINDEX_IB | 97.12 | 1 |

## V. CONCLUSION

Feature selection methods are used to find highly relevant features from a given feature set. It in turns helps the classifiers to reach optimal performance. In this paper, we proposed a feature selection approach MINDEX_IB for imbalance datasets. This is a partition based method that uses the concept of micro-clustering for partitioning the attribute domain and then concludes the relevance of attribute from the statistics obtained from the micro-cluster.

Experimental results show that Mindex_IB selects much less number of attributes in turn making the classification algorithm more efficient. Mindex_IB is comparable to other existing methods in terms of classification Accuracy, F-measure and AUC evaluation measure.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 4, April 2016.*

The Feature selection method MINDEX_IB in the present form requires a reprocessing step if the dataset is not in the numerical form. Thus, the work can be extended to deal directly with nonnumeric databases.

## REFERENCES


[1] Kubat M, Matwin S (1997) "Adressing the curse of imbalanced training sets: onesided selection". Proc. 14th Intl. Conf. on Machine Learning, pp. 179–186.. (1994)

[2] LIA et al "generative oversampling for mining imbalanced datasets". In proceeding of 2007 international conference on data mining, pp, Las vegas,Nevada,USA 2007,CSREA Press.

[3] Chien-I Lee, Cheng-Jung Tsai, Tong-Qin Wu, Wei-Pang Yang "An approach to mining the multi-relational imbalanced database". Expert Systems with Applications 34 (2008) 3021–3032

[4]. S.-J. Yen, and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions". Expert Systems with Applications, Vol. 36, pp. 5718–5727, 2009.

[5] Hongyu Guo and Herna L. Viktor "Mining Imbalanced Classes in Multirelational Classification".Proceedings of the 6th Multi-Relational Data Mining Workshop(PKDD/MRDM'07), in conjunction with 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Warsaw, Poland, September 2007.

[6] Y. Murphey, H. Wang, G. Ou, and L. Feldkamp, "OAHO: an effective algorithm for multi-class learning from imbalanced data". In International Joint Conference on Neural Networks (IJCNN), Aug. 2007, pp. 406–411.

[7] A. S. Ghanem, S. Venkatesh, and G. West, "Learning in Imbalanced Relational Data". In Proc. ICPR. IEEE Computer Society, December 2008.

[8] Amal S. Ghanem and Svetha Venkatesh, Geoff West "Multi-Class Pattern Classification in Imbalanced Data". In International Conference on Pattern Recognition, 2010.

[9] Jeatrakul, P. and Wong, K.W. "Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm". In: Annual International Joint Conference on Neural Networks, IJCNN 2012, 10 - 15 June, Brisbane, Australia.

[10] Somol, P., Baesens, B., Pudil, P., and Vanthienen, J. "Filter- versus wrapper-based feature selection for credit scoring". International Journal of Intelligent Systems 20, 10 (2005), 985–999.

[11] Yusta, S. C. "Different metaheuristic strategies to solve the feature selection problem".Pattern Recognition Letters 30 (2009), 525–534.

[12] Pudil, P., Novovicova, J., and Kittler, J., 1994. "Floating search methods in feature selection". Pattern Recognition Letters, 15,1119 – 1125.

[13] Weston, J. et al. 2000. "Feature selection for support vector machines". In Advances in Neural Information Processing Systems.

[14] Chen, X. 2003. "An improved branch and bound algorithm for feature selection". Pattern Recognition Letter, 24, 1925-1933.

[15] I., and Elisseeff, A. 2003. "An introduction to variable and feature selection". JMRL special Issue on variable and Feature Selection 3, 1157-1182.

[16] Yu, L. and Liu, H. 2004. "Efficient feature selection via analysis of relevance and redundancy". Journal of Machine Learning Research, 5, 1205-1224.

[17] Chen, X. and Jeong, J. 2007. "Minimum reference set based feature selection for small sample classifications". Proc. of the 24th International Conference on Machine Learning 153-160.

[18] X. Chen and M. Wasikowski, "FAST: A ROC-based feature selection metric for small samples and imbalanced data classfication problems," in Proceedings of the 14th ACM SIGKDD.

[19] Mina Alibeigi et.al "Unsupervised Feature Selection Based on the Distribution of Features Attributed to Imbalanced Data Sets" International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (2) : Issue (1) : 2011.

[20] Ilnaz Jamali, Sattar Hashemi , The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue: The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13], San Francisco State University, CA,U.S.A., March 2013 Doi: 10.7321/jscse.v3.n3.89 e-ISSN: 2251-7545 588 "A Feature Selection Method for Imbalance Data sets".

]21] German Cuaya et al." A Minority Class Feature Selection Method", National Institute of Astrophysics, Optics and Electronics,Computer Science Department,Luis Enrique Erro 1, 72840 Tonantzintla, M´exico.

[22] D. Tiwari,International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), Vol. 2, Issue 2, Ver. 3 (April - June 2014)" Handling Class Imbalance Problem Using Feature Selection".

[23] Reena Srivastava and M M Gore. Mindex: A Relevance Measure for IntervalScaled Attributes. Published in the Proceedings of 4th Indian International Conference on Artificial Intelligence (IICAI), held in Tumkur, India, from 16-18 Dec., 2009. Pages: 2140-2157, ISBN 978-0-9727412-7-9.