iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** XI **Month of publication:** November 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Ensemble based novel class identification for Class Imbalance under sampled Data

D. Priyadarshini[1], Sulfyth M[2]
[1]Assistant Professor, [2]Research Scholar, Dept. of Computer Science, Sree Narayana Guru College of Arts and Science

Abstract: ***The Classification of data is represented as research challenge in handling the class imbalance problem. Classification problems are represented by highly unbalanced data sets, in which, the number of samples from one class is much smaller than from another. This is known as class imbalance problem and is often reported as an obstacle for constructing a model that can successfully discriminate the minority samples from the majority samples. Under sampling is a popular method in dealing with class-imbalance problems, which uses only a subset of the majority class and thus is very efficient. The main deficiency is that many minority class examples are ignored. Ensemble based under sampling method is proposed for the class imbalance problem. The class imbalance problem is defined in terms of which the ratio of the majority and minority class cardinalities is inversed. The main idea is to severely under sample the majority class thus creating a large number of distinct training sets using normalized information gain. For each training set we then find a decision boundary which separates the minority class from the majority class using the classifier c5.0. By combining the multiple designs through fusion, we construct a composite boundary between the majority class and the minority class using entropy calculation. Experimental results show that both proposed method class-imbalance learning method out performs state of arts approaches higher in terms of precision, recall and f-measure for disproportionate class sample size for different boundaries.***
Index Terms – Class Imbalance Data, Under Sampling, Ensemble Classification, Outlier Data

I. INTRODUCTION

In real-world problems, the data sets are typically imbalanced, i.e., some classes have much more instances than others. Imbalance has a serious impact on the performance of classifiers. Learning algorithms that do not consider class imbalance tend to be overwhelmed by the majority class and ignore the minority class [1]. Training set size, class priors, cost of errors in different classes, and placement of decision boundaries are all closely connected. In fact, many existing methods for dealing with class imbalance rely on connections among these four components. Sampling methods handle class imbalance by varying the minority and majority class sizes in the training set. Cost-sensitive learning deals with class imbalance by incurring different costs for the two classes and is considered as an important class of methods to handle class imbalance [2] Inverse random under sampling (IRUS) method[3] is used for handling the class imbalance problem. The Class imbalance based classification using under sampling method can be applicable to static data's.

II. RELATED WORK

A. Inverse random under sampling (IRUS) method

It is used for handling the class imbalance problem. The class imbalance problem is defined in terms of which the ratio of the majority and minority class cardinalities is inversed. The main idea is to severely under sample the majority class thus creating a large number of distinct training sets. For each training set we then find a decision boundary which separates the minority class from the majority class [3]. By combining the multiple designs through fusion, we construct a composite boundary between the majority class and the minority class. Main pitfalls of the work can be the Effectiveness of the Composite boundaries may lead to cluster inconsistency.

B. Exploratory under-sampling for class-imbalance learning

Under sampling is a popular method in dealing with class-imbalance problems which uses only a subset of the majority class and thus is very efficient. The main deficiency is that many majority class examples are ignored[4]. Easy Ensemble samples several subsets from the majority class, trains a learner using each of them, and combines the outputs of those learners through exploratory analysis. Balance Cascade trains the learners sequentially, where in each step, the majority class examples that are correctly classified by the current trained learners are removed from further consideration. The outlier determination may lead to large

execution time in generating a class for the data and clustering it with similar data points [5].

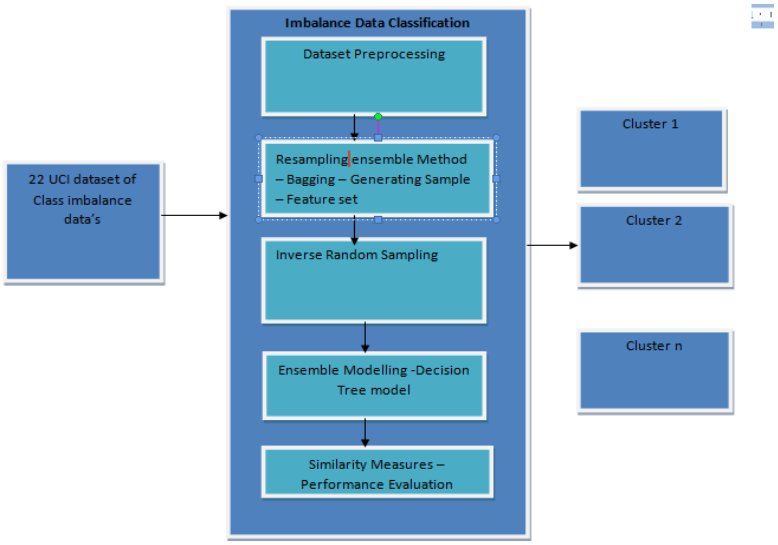
III. OUR MODEL

A. *Feature selection for data classification*

One of the most assumptions of ancient data processing is that knowledge is generated from one, static and hidden platform. However, it is hard to be true for data stream learning [6], where unpredictable changes are likely to eventually happen. Data may be inconsistent if said to occur once the underlying function that generates instances changes over time. The k-means clustering is known to be efficient in clustering large datasets. This clustering is one in all the best and also the best far-famed unsupervised learning algorithms that solve the well-known clustering problem. The K-Means algorithm aims to partition a group of objects supported their attributes/features, into k clusters, wherever k may be a predefined or user-defined constant, into k clusters, where k is a predefined or user-defined constant.

B. *Bagging based resampling*

Bagging is the one of the most popular resampling ensemble method [7]. It is a relatively simple idea: given a training set, bagging generates many bootstrap samples or training subsets. Each bootstrap sample is drawn by randomly generating subsets of samples where each sample is selected with replacement and equal probability.



A prediction method e.g. decision tree or neural network is applied to each bootstrap sample to get base model. A bagged ensemble predicts a new sample by having each of its base models classify that example. The final prediction of the class is normally obtained by majority voting. Diversity also plays an important role in improving the performance of ensemble classifier. The main idea in bagging is that the base models generated from the different bootstrapped training sets disagree often enough that the ensemble performs better than the base models and with the variance being reduced due to aggregation [8].

C. *Outlier Classification –Ensemble Technique*

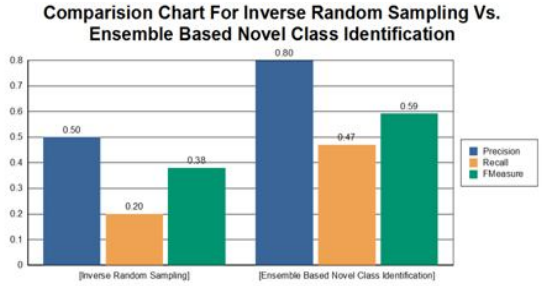
The idea is to severely under sample the majority class thus creating a large number of distinct training sets. For each training set we then find a decision boundary which separates the minority class from the majority class. By combining the multiple designs through fusion, we construct a composite boundary between the majority class and the minority class using the classifier c5.0[9]. The data classifying from the under sampled data is based on the splitting criterion for normalized information gain. The data with normalized information gain will be classified into the new class or majority class. Data pruning is also applied for high normalized data with high info gain using entropy calculation to determine the weight of the feature in the particular class.

IV. EXPERIMENTAL RESULTS

A. *Dataset Description*

Experiments were carried out on 22 public data sets the UCI repository which have different degrees of imbalance. For each data set, it shows the number of attributes (A), the number of samples (Ns), the number of majority samples (N Max) and the number of

minority samples (N min). All the reported results are obtained by 10-Fold Cross Validation and the paired t-test is then used to determine their significance under a value of 0.05.Initially features are extracted from the dataset and it is placed into subsets. Classification of the data points in the subset is carried out using the c5.0 decision tree algorithm. The Classification algorithm is used for reducing the minority class coverage for under sampled data and increase the majority class coverage for oversampled data. The Performance of the under sampling techniques is computed using precision recall and f measure. However it is absorbed that existing inverse random sampling mechanism does not able to classify the minority class data into the different available subset in the majority class. The precision, recall and f- measure is computed using the true positive and false positive rates. The precision and f- measure rate is high for Ensemble based Novel Class Identification using 10 fold cross validations based on the multiple categories. In figure 2, it is clear that proposed system produces the more accuracy compared to existing mechanism.



Algorithm	Precision	Recall	F Measure
Inverse Random Sampling	0.50	0.20	0.38
Ensemble based Novel Class Identification	0.80	0.47	0.59

Performance Evaluation and comparison of the under sampling Technique

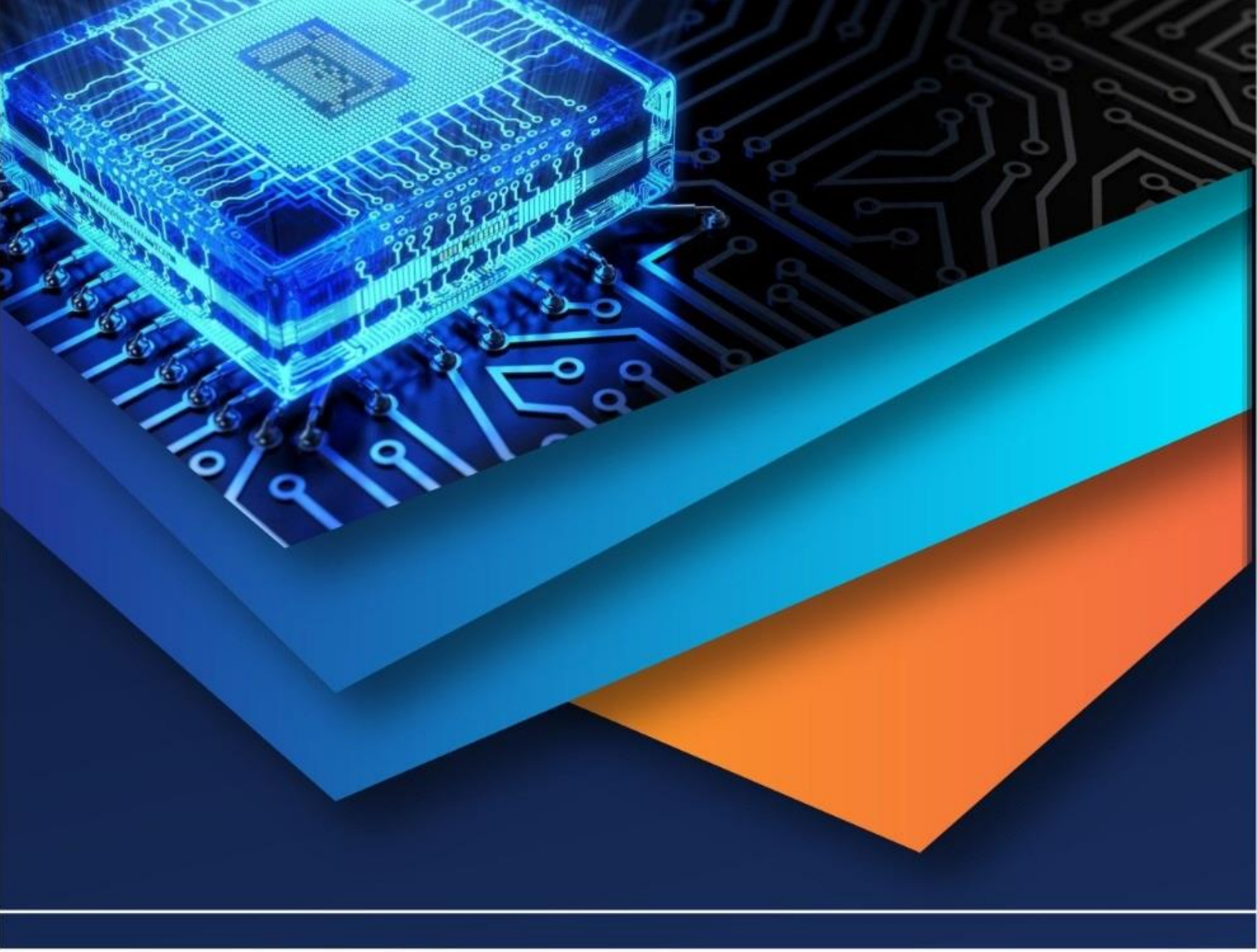
The Proposed Techniques solves imbalance problem by maintaining a very high true positive rate through imbalance inversion and controlling the false positive rate through classifier as depicted in the table 1for multi label data classification. The boundary of the classifier c5.0 would reflect the positive class with high true positive rate since the number of samples from the negative class are much less than the number of samples from the positive class. Further, since the number of samples from the negative class is very small in relation to the dimensionality of the feature space, the capacity of each boundary to separate the classes fully is high.

V. CONCLUSION

We have designed and implemented technique to classify the imbalance data which acted as an outlier data in data cluster named as dynamic sampling mechanism through Ensemble based novel class identification technique. The majority class and minority class is predicted using complex decision boundary. Thus decision boundaries for each cluster is attained separately which is further fused together to obtain the composite boundaries. The Composite boundary is applicable to the data classification using c5.0 algorithm. The Classifier initiates after calculating the information gain to the cluster using entropy. The classified results are analysed using performance metric such as precision, recall and f- measure.

REFERENCES

[1] Ken chen “Efficient Classification of Multi-label and Imbalanced Data using Min-Max Modular Classifiers” Published in International IEEE conference, 2006, pp: 1770-1775.
[2] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Synthetic minority over-sampling technique, Journal of Artificial Intelligence Review 16) (2002) 321–357
[3] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda`, “New Ensemble Methods for Evolving Data Streams,” Proc. ACM SIGKDD 15th Int’l Conf. Knowledge Discovery and Data Mining, pp. 139-148, 2009.
[4] S. Chen, H. Wang, S. Zhou, and P. Yu, “Stop Chasing Trends: Discovering High Order Models in Evolving Data,” Proc. IEEE 24th Int’l Conf. Data Eng. (ICDE), pp. 923-932, 2008.
[5] W. Fan, “Systematic Data Selection to Mine Concept-Drifting Data Streams,” Proc. ACM SIGKDD 10th Int’l Conf. Knowledge Discovery and Data Mining, pp. 128-137, 2004.
[6] J. Gao, W. Fan, and J. Han, “On Appropriate Assumptions to Mine Data Streams,” Proc. IEEE Seventh Int’l Conf. Data Mining (ICDM), pp. 143-152, 2007.
[7] S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari, “Adapted One-versus-All Decision Trees for Data Stream Classification,” IEEE Trans. Knowledge and Data Eng., vol. 21, no. 5, pp. 624-637, May 2009.
[8] G. Hulten, L. Spencer, and P. Domingos, “Mining Time-Changing Data Streams,” Proc. ACM SIGKDD Seventh Int’l Conf. Knowledge Discovery and Data Mining, pp. 97-106, 2001. MASUD ET AL.: CLASSIFICATION AND ADAPTIVE NOVEL CLASS DETECTION OF FEATURE-EVOLVING DATA STREAMS 1495
[9] I. Katakis, G. Tsoumakas, and I. Vlahavas, “Dynamic Feature Space and Incremental Feature Selection for the Classification of Textual Data Streams,” Proc. Int’l Workshop Knowledge Discovery from Data Streams (ECML/PKDD), pp. 102-116, 2006.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY