ISSN 2229-5518

A New approach for Classification of Highly Imbalanced Datasets using Evolutionary Algorithms

Satyam Maheshwari, Prof. Jitendra Agrawal, Dr. Sanjeev Sharma

Abstract— Today's most of the research interest is in the application of evolutionary algorithms. One of the examples is classification rules in imbalanced domains. The problem of Imbalanced data sets plays a major challenge in data mining community. In imbalanced data sets, the number of instances of one class is much higher than the others, and the class of fewer representatives is of more interest from the point of the learning task. Traditional Machine Learning algorithms work well with balanced data sets, but not able to deal with classification of imbalanced data sets. In the present paper we use different operators of Genetic Algorithms (GA) for over-sampling to enlarge the ratio of positive samples, and then apply clustering to the over-sampled training dataset as a data cleaning method for both classes, removing the redundant or noisy samples. The proposed approach was experimentally analyzed and the experimental results shows an improvement in the classification measured as the area under the receiver operating characteristics (ROC) curve.

Index Terms— classification, data mining, evolutionary algorithm, imbalanced datasets, re-sampling, samplings, support vector machine.

1 INTRODUCTION

THE problem of imbalanced data-sets occurs when the majority class has a large percent of the samples, while minority class occupies a small part of all samples. Such a condition pose challenges for classical machine learning algorithms that are designed to optimize oTverall classification accuracy. Imbalanced datasets exists in many domains such as medical applications [1], risk management [2], face recognition [3] and information technology, and so on. In these domains, minority class is of more interest than majority class. In imbalanced data sets, the traditional way of maximizing overall performance will often fail to learn anything useful about the minority class, because of the dominating effect of the majority class. A learner can probably achieve 99% accuracy with ease, but still fail to correctly classify any rare examples. Therefore, analyzing the imbalanced data sets (IDS) problem requires new and more adaptive methods than those used in the past.

In this paper we over-samples the minority class by mutation and crossover operators to decrease the imbalance ratio and then using clustering for both classes to delete redundant and noisy samples. Thus, by combining the both method the samples of interest are remained, improving the computational efficiency.

The contribution is organized as follows: Section 2 introduces the problem of imbalanced data sets, describing its feature, how to deal with this problem. Next, in section 3 we will expose the related work done in this field. Section 4 describes the characteristics of our proposal. Section 5 contains the measure of performance evaluation of imbalanced datasets. Section 6 analyses the experimental results. Finally, conclusion and future work will be pointed out in section 7.

2 NATURE OF THE PROBLEM

Learning from imbalanced data is an important topic that has recently appeared in Machine Learning Community [4]. Imbalanced data sets can occur in many real-world applications, such as detection of fraudulent telephone calls [5], text classification [6], information retrieval and filtering tasks [7], data mining for direct marketing [8], and so on. The problem of imbalanced datasets in classification occurs when the number of instances of one class is much lower than that of the other classes. Specifically, when the datasets has only two classes, this happen when one class is represented by a high number of examples, while the other is represented by only a few and usually the minority class represents the concept of interest.

Traditional classifier algorithms are more biased towards the majority class (Negative Samples), since the rules that predict the higher numbers of examples are positively weighted during the learning process in favors of the accuracy metric. Consequently, the samples that belong to the minority class (Positive Samples) are more misclassified than often those belongings to the majority class [20].

Imbalanced datasets faces many challenges; the first challenge is measure of performance. To overcome this

- *Satyam Maheshwari is currently pursuing Masters Degree program in computer technology and application in soit RGPV Bhopal, India,E-mail:satyam.vds@gmail.com*
- *Prof. Jitendra Agrawal, Dr. Sanjeev sharma are with the Department of SOIT RGPV Bhopal,India, E-mail: jitendra@rgtu.net,sanjeev@rgtu.net*

problem, Evaluation metrics are used to guide the learning process towards the desired solution. The second challenge is lack of data. If a class may have very few samples, then it is very difficult to construct accurate decision boundaries between classes. The third challenge is noise. Noisy data have a serious impact on minority classes than on majority classes. Furthermore, classical machine learning algorithms tend to treat samples from minority class as a noise.

3 RELATED WORK

The state-of-the-art research methodologies to handle imbalanced learning problem can be broadly categorized in to two approaches, which have been proposed both at the data level, such as over-sampling and under-sampling and at the algorithmic level, such as recognition-based approaches, cost-sensitive learning and boosting.

3.1 Data-Level Approaches

In this approach, the objective is to re-balance the class distribution by re-sampling the data space. The ways for dealing with class imbalance is to alter the class distributions toward a more balanced distribution. These solutions include many different forms of re-sampling such as over-sampling and under-sampling. The over-sampling approach increase the number of minority class samples to reduce the degree of imbalanced distribution. The under-sampling is also a non-heuristic method aim to balance the data sets by eliminating examples of majority class.

3.1.1 Tomek Links [9]

This method can be defined as follows: Consider the two examples a and b which belongs to different classes, and d (a,b) is the distance between a and b. A(a,b) pair is called a Tomek Link if there is not an example c, such that d(a,c)<d(a,b) or d(b,c)<d(a,b). If two examples form a Tomek link, then either one of these examples is noise or both examples are border-line. This method can be used as an under-sampling method. In the under-sampling method, examples of the majority class are eliminated.

3.1.2 One-side selection (OSS) [10]

In this method samples of majority class are removed that are considered either noisy or redundant. OSS method uses the Tomek links followed by the application of Condensed Nearest Neighbor Rule (CNN) [11]. This method can be used as an under-sampling method. It is an efficient method because it reduces possibilities of noise, but it is bit slower because it uses Tomek link.

3.1.3 Synthetic minority over-sampling technique (SMOTE) [12]

This is an over-sampling method. In this method new minority samples are formed by interpolation among several minority class examples that lie together. The minority samples are over-sampled to create synthetic samples rather than by just over-sampling with replacement. This method is more useful than random over-sampling because it creates new minority samples artificially.

3.1.4 SMOTE+Tomek link [13]

The drawback of SMOTE and Tomek link are removed by hybrid sampling technique. This method is used for better-defined class clusters among majority and minority classes.

3.2 Algorithm-Level Approaches

It this level, solutions try to adapt existing classifier learning algorithms to strengthen learning with regard to the small class. Two common methods Boosting and Cost-sensitive learning are used in this approach.

3.2.1. Cost-sensitive learning

In this learning method cost is associated with misclassifying examples. The cost matrix is used for numerical representation of the penalty of classifying examples from one class to another. No penalty is assigned for correct classification of either class and the cost of misclassifying minority samples is higher than the majority samples, i.e., C (Majority, Minority)> C (Minority, Majority). The objective of cost-sensitive learning method is to minimize the overall cost on the training dataset. Charles [14] gives a theorem that shows how to change the proportion of positive and negative samples in order to make optimal cost-sensitive classifications for a concept-learning problem. Pedro [15] suggested a more general method to make a learning system a cost-sensitive.

3.2.2. Boosting algorithm

Boosting is a technique to improve the performance of weak classifiers. AdaBoost [16] is the most common boosting algorithm, which is an ensemble learning model. In every iteration, weights are modified with the objective of correctly classifying examples in the next iteration. At the end, all modified models participate in a weighted vote to classify unlabeled examples. This method is more effective to deal with class imbalance problem because minority class examples are most likely to be misclassified and therefore given higher weights in subsequent iterations.

4 EVOLUTIONARY-SVM ALGORITHM

4.1 Over-sampling the minority class

The SMOTE algorithm [12] generates an arbitrary number of "artificial" minority examples to shift the classifier learning bias toward the minority class. The minority class examples are over-sampled by creating the artificial examples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing new artificial examples by joining any or all of the k minority class nearest neighbors. The neighbors from the k-nearest neighbors are randomly selected based on the amount of over-sampling is required. Synthetic examples are generated by the following ways: (i) First, we take the difference between the sample X under consideration and its nearest neighbor selected randomly from the k minority class nearest neighbors. (ii) Secondly, we take the difference

between the feature vector under consideration of its nearest neighbor. (iii) Third, we multiply this difference by a random number between 0 and 1, and add it to the sample under consideration. This causes the selection of a random point between specific samples. This method can effectively force the decision region of the minority class to become more general to dataset. The new samples are defined as follows:

$$X_{new} = X + rand(0,1) * (\tilde{X} - X) \quad (1)$$

Accordingly, artificial examples can be generated through repeating the above steps.

4.2 Data cleaning using clustering method

If the datasets have skewed class distribution, then amount of over-sampling required is too large. This may cause the minority class to become the majority class. In this circumstance, a data cleansing method is needed for both classes instead of randomly under-sampling the majority class. In this paper we use clustering method which reduces redundant or noisy samples from the dataset.

5 EVALUATION MEASURES

Evaluation measures play an important role in both assessing the classification performance and guiding the classifier modeling. Most of the studies in imbalanced domains mainly concrete on two-class problem as multi-class problem can be simplified to two-class problem. By convention, the class label of the minority class is positive, and the class label of the majority class is negative. The following table (Table 1) shows confusion matrix of a two-class problem. The first column of the table is the actual class label of the examples, and the first row presents their predicted class label. In the matrix, TP shows the true positive samples, FP shows the false positive samples, TN shows the true negative samples, and FN shows the false negative samples respectively.

5.1 F-measure

The performance metric used in this work is the F- measure. This metric uses recall and precision for performance measurement. If only the performance of positive class is considered, two measures are important: True Positive Rate and Positive Predicted Value

$$\text{Recall= TPrate} = \frac{TP}{TP + FN} \quad (2)$$

Positive Predicted Value is defined as precision denoting the percentage of relevant objects that are identified for retrieval:

TABLE 1 A CONFUSION MATRIX FOR A TWO-CLASS CLASSIFICTION

	Predicted Positive	Predicted Negative
Actual Positive	True Positive(TP)	False Negative(FN)
Actual Negative	False Positive(FP)	True Negative(TN)

$$\text{Precision= PPvalue} = \frac{TP}{TP + FP} \quad (3)$$

F-measure is a combination of recall and precision. This represents a harmonic mean between recall and precision. In practice, high F-measure value ensures that both recall and precision are reasonably high.

$$\text{F-measure} = \frac{2 * \Pr ecision * \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call} \quad (4)$$

5.2 AUC analysis

The area under a ROC (Receiver Operating Characteristic) curve (AUC) provides a single measure of a classifier's performance for evaluating which model is better on average. AUC can also be applied to evaluate the imbalanced data sets [17]. From ROC graph it is possible to calculate an overall measure of quality; the AUC is the fraction of the total area that falls under the ROC curve. This measure is equivalent to several to other statistical measures for evaluating classification and ranking models. The AUC effectively factors in the performance of the classifier over all costs and distributions. The area also has a nice interpretation as the probability that the classifier ranks a randomly chosen positive instance above a randomly chosen negative one.

6 EXPERIMENTAL STUDY

We study the performance of our algorithm employing a large collection of imbalanced datasets with a high imbalanced ratio (IR > 10). We have considered 4 different datasets from UCI repository [18] with different IR, as shown in Table 2. This table is in ascendant order according to the IR. Multi-class datasets are modified to obtain two-class imbalanced problems, defining the joint of one or more classes as positive and the joint of one or more classes as negative.

We work under the framework of WEKA [19], which is an open-source data mining suite. In the WEKA we create a new classifier in the folder SmWork. E-SVM is a modified code of Sequential Minimal Optimization (SMO) as SVM training algorithm [21]. We include JAR file of E-SVM using APACHE ANT or ECLIPSE in the WEKA GUI Explorer. E-SVM algorithm over-samples the minority samples for uniform distribution of data but if datasets is highly skewed, then much over-sampling is required. However, there may be probability that minority samples become majority samples and the new samples generated are not completely con-

tent to data distribution. To overcome from this circumstance we apply data cleansing method using clustering to reduce redundant or noisy samples. The snapshot of WEKA is shown in figure 1:

TABLE 2 DATA DISTRIBUTION

Datasets	Positive samples	Negative samples	IR (Negative/Positive)
yeast(9)	5	1479	295.8
page-blocks(2)	28	5445	194.46
glass(5)	9	205	22.78
balance-scale(2)	49	576	11.75

This paper compares three methods, which are SVM, SMOTE-SVM and E-SVM (Evolutionary SVM with clustering). The comparison is based on a five-folder cross validation model, i.e., 5 random partitions of data with a 20%, and the combination of 4 of them (80%) as training and the remaining one as test. The amount of over-sampling is 100% for all methods; the crossover constant is o.4 and similarity threshold is 0.9.

TABLE 3 THE RESULTS OF EXPIREMENT ON IMBALANCED DATASETS

Datasets	F-measure			AUC		
	SVM	SMOTE-SVM	E-SVM	SVM	SMOTE-SVM	E-SVM
Yeast(9)	0.532	0.541	0.553	0.723	0.762	0.764
Page-blocks(2)	0.985	0.954	0.962	0.746	0.731	0.749
Glass(5)	0.524	0.548	0.654	0.739	0.794	0.834
Balance-scale(1)	0.846	0.843	0.824	0.854	0.898	0.867
Mean	0.721	0.722	0.784	0.765	0.796	0.803

The results from table 3 show that our algorithm obviously improves the performance of classification. Therefore, E-SVM (evolutionary over-sampling with clustering) is a novel method for highly imbalanced datasets

Fig. 1. Snapshot of WEKA displaying new added classifier E-SVM

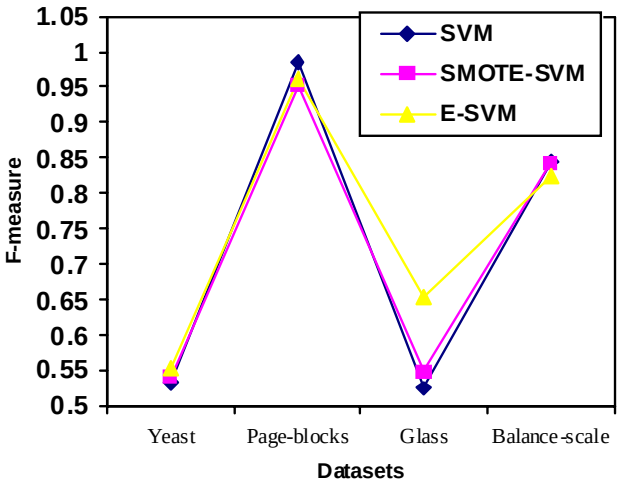

Fig. 2. The Chart of F-measure

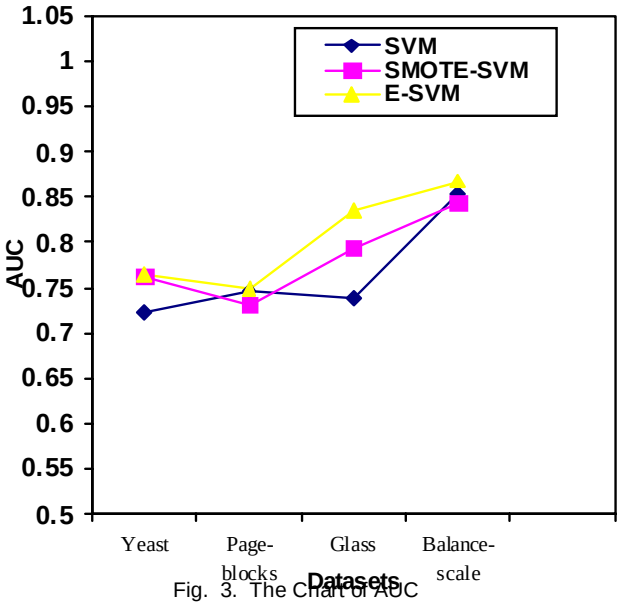

Fig. 3. The Chart of AUC

Figure 2 and 3 represent the comparison of F-measure and AUC. The three algorithms, in different shapes are SVM, SMOTE-SVM and E-SVM.

7 CONCLUSION

In this paper, we proposed a novel E-SVM (evolutionary over-sampling with clustering) method for SVM classification on IDS. To improve the computational efficiency of the algorithm, it is proposed by combining over-sampling the minority samples and data clustering to removes redundant or noisy samples. To verify the effectiveness of the proposed algorithm, four different UCI datasets are adopted to validate this approach. The results indicate that the proposed approach can receive better performance than the previous approaches.

ACKNOWLEDGMENT

We thank Dr. R.C Jain and Sh. Sunil Joshi for discussing and giving us advice on its implementation.

REFERENCES

[1] M.A. Mazurowski, P.A. Habas, J.M. Zurda, J.Y. Lo, L.A. Baker, and G.D. Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks, 21(2-3):427-436,2008.

[2] Y. M. Huang, C. M. Hung, and H. C. Jiau. Evaluation of neural networks and data mining methods on a credit assessment tasks for class imbalance problem. Nonlinear Analysis: Real World Applications, 7(4):720-747, 2006.

[3] Y.H Liu and Y.T. Chen. Face recognition using total margin-based adaptive fuzzy support vector machines. IEEE Transactions on Neural Networks, 18(1):178-192,2207.

[4] Chawla N.V., Japowicz N., Kolcz A., Editorial:Special issue on learning from imbalanced data sets. SIGKDD Explorations 6(1), 1-6 ,2004.

[5] Phua C, Alahakoon D, Lee V. Minority report in fraud detection: Classification of skewed data[J]. SIGKDD Explore,,6(1):50-59,2004.

[6] Del Castillo M D, Serrano J I. A multi strategy approach for digital text categorization from imbalanced documents[J]. SIGKDD Explorer,6(1):70-79,2004.

[7] Turney P D. Learning algorithms for keyphrase extraction [J]. Information Retrieval,,2(4):303-336,2000.

[8] Ling C X, Li C. Data mining for direct marketing: Problems and solutions[J]. Knowledge Discovery and Data Mining,.73-79,1998.

[9] Ivan Tomek (1976). "Two Modifications of CNN". IEEE Transactions on Systems,Man, and Cybernatics, Vol. 6, No. 11, pp.769-722,1976.

[10] Miroslav Kubat,Matwin Stan, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection", the 14th International Conference on Machine Learning, pp. 179-186.

[11] Peter E. Hart 1968. "The Condensed Nearest Neighbor Rule". The IEEE Transactions on Information Theory, Vol. 14, No. 3, pp.515-516,1968.

[12] Chawla N V, Hall L O , Bowyer k W, et al. SMOTE: Synthetic Minority Oversampling Technique[J]. Journal of Artificial Intelligence Research, 16:321:357,2002.

[13] Gustavo E.A. P.A. Bastista, Prati Ronaldo C., Monard Maria Carolina. " A Study of the Behavior of Several Methods for Balancing machine Learning Training Data". ACM SIGKDD Explorations Newsletter, Vol. 6, No. 1, pp.20-29,2004.

[14] Charles Elkan, "The Foundations of Cost-Sensitive Learning," the Sevnteenth International Joint Conference on Artificial Intelligence, pp. 973-978.

[15] Pedro Domingos, "Metacost: A General Method for Making Classifiers Cost-Sensitive," the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 155-164.

[16] Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". Journal of Computer and System Science, 55(1):119-139, 1997.

[17] Andrew P Bradley. "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms". Pattern Recognition, Vol. 30, No. 7,pp.1145-1159,1997.

[18] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. University of California, Irvine, School of Information and Computer Sciences. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

[19] The Weka Machine Learning Workbench. http://www.cs.waikato.ac.nz/ml/weka.

[20] Weiss G., Mining with rarity: a unifying framework. SIGKDD Explorations 6(1), 7-19 2004.

[21] Platt J.C. Fast training of support vector machines using sequential minimal optimization. In: scholkopf B, Burges C, Smola A, eds. Advances in Kernal Methods Support Vector Learning. Cambridge, MA: MIT press. 185-208, 1999.