International Journal of Research in Advent Technology, Vol.2, No.9, September 2014
E-ISSN: 2321-9637

Survey on Ensemble Alpha Tree for Imbalance Classification Problem

R.Saranya[1], Dr.C.Yamini[2],
Research scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore [1].
Assistant Professor, Sri Ramakrishna College of Arts and Science for Women, Coimbatore [2].
Email: r.saran2708@gmail.com[1],c_yamini@yahoo.com [2]

Abstract- A novel splitting criterion based on alpha divergence is uses to generalize several well-known splitting criteria such as those used in C4.5 and CART. The new method neighborhood cleaning rule (NCL) outperformed simple random and one sided selection is used. The success of SVM is very limited when it is applied to the problem learning from imbalanced datasets in which negative instance heavily outnumber the positive instance. To generate best performing classifier we introduce "budget sensitive" progressive sampling algorithm for selecting training examples. Machine learning algorithms have been used to build classification rules from datasets consisting of hundreds of thousands of instance. The imbalanced learning problem is concerned with the performance of learning algorithm in the presence of underrepresented data and sever class distribution skews. Adapt machine learning algorithm are been used for imbalance class and misclassification cost for the look at under sampling and oversampling which increase or decrease respectively.

Index Terms- Training examples; classification; machine learning.

1. INTRODUCTION:

The literature review is based on the fundamental understanding of knowledge discovery and analysis from raw data to support decision-making processes. The problem of learning from imbalanced data is a relatively new challenge that has attracted growing attention from both academia and industry. The factor of under sampling and oversampling is been processed to improve the SVM by using any one of this algorithm. Two levels of classifiers called ‘stacked-generalization’ or ‘meta learning’ are used to estimate the performance of training set.(learning from the information generated by a set of learners). Machine learning algorithms have been used to build classification rules from data sets consisting of hundreds of thousands of instances.

2. IMPROVING IDENTIFICATION OF DIFFICULT SMALL CLASSES BY BALANCING CLASS DISTRIBUTION

Real-world data sets often have imbalanced class distribution, because many natural processes produce certain observations infrequently. For example, rare diseases in a population may result in medical data with small diagnostic groups. When some classes are heavily under-represented, statistical and machine learning methods are likely to run into problems. Cases from the rare classes are lost among the other cases during learning. The resulting classifiers misclassify new unseen rare cases, and descriptive models may give an inadequate picture of the data. The learning task is even more problematic, if a small class is difficult to identify because of its other characteristics. A small class may, for example, overlap heavily the other classes. In the following, we refer to a small and difficult class as a class of interest. We balanced class distribution with data reduction before the actual analysis, because we aimed to develop a general-purpose method, whose results may be given directly to statistical and machine learning methods. The most well-known data reduction technique comes from the area of statistics, where sampling is used to allow analyses which would be impractical with large populations. Data reduction has been utilized in the area of machine learning especially to accelerate the instance based learning methods. Recently one-sided selection (OSS) which uses instance-based methods to reduce the larger class when class distribution of a two-class problem is imbalanced. In this paper, we describe a new method called neighborhood cleaning rule that utilizes the OSS principle, but considers more carefully the quality of the data to be removed.

3. APPLYING SUPPORT VECTOR MACHINES TO IMBALANCED DATASETS

Support Vector Machines (SVM) was introduced by Vapnik and colleagues and they have been very successful in application areas ranging from image retrieval, handwriting recognition to text classification. However, when faced with imbalanced datasets where the number of negative instances far outnumbers the positive instances, the performance of SVM drops significantly. Application areas such as gene profiling, medical diagnosis and credit card fraud detection have highly skewed datasets with a

very small number of positive instances which are hard to classify correctly, but important to detect nevertheless . An imbalance of 100 to 1 exists in fraud detection domains, even approaching 100,000 to 1 in other applications.

Classifiers generally perform poorly on imbalanced datasets because they are designed to generalize from sample data and output the simplest hypothesis that best fits the data, based on the principle of Occam's razor. This principle is embedded in the inductive bias of many machine learning algorithms including decision trees, which favor shorter trees over longer ones. With imbalanced data, the simplest hypothesis is often the one that classifies almost all instances as negative Another factor is that making the classifier too specific may make it too sensitive to noise and more prone to learn an erroneous hypothesis. Certain algorithms specifically modify the behavior of existing algorithms to make them more immune to noisy instances, such as the IB3 algorithm for KNN, or pruning of decision trees, or soft margins in SVM . While these approaches work well for balanced datasets, with highly imbalanced datasets having ratios of 50 to 1 or more the simplest hypothesis is often the one that classifies every instance as negative. Furthermore, the positive instances can be treated as noise and ignored completely by the classifier. A popular approach towards solving these problems is to bias the classifier so that it pays more attention to the positive instances. This can be done, for instance, by increasing the penalty associated with misclassifying the positive class relative to the negative class. Another approach is to preprocess the data by oversampling the majority class or under sampling the minority class in order to create a balanced dataset. We combine both of these approaches in our algorithm and show that we can significantly improve the performance of SVM compared to applying any one approach. We also show in this paper that even though under sampling the majority class does improve SVM performance, there is an inherent loss of valuable information in this process. Our goal was to retain and use this valuable information, while simultaneously boosting the efficacy of oversampled data. Combined with this dual approach we also used a bias to make SVM more sensitive to the positive class. We specifically chose SVM to attack the problem of imbalanced data because SVM is based on strong theoretical foundations and our empirical results show that it performs well with moderately imbalanced data even without any modifications. Its unique learning mechanism makes it an interesting candidate for dealing with imbalanced datasets, since SVM only takes into account those instances that are close to the boundary, i.e. the support vectors, for building its model. This means that SVM is unaffected by non-noisy negative instances far away from the boundary even if they are huge in number.

4. LEARNING WHEN TRAINING DATA ARE COSTLY: THE EFFECT OF CLASS DISTRIBUTION ON TREE INDUCTION

In many real-world situations the number of training examples must be limited because obtaining examples in a form suitable for learning may be costly and/or learning from these examples may be costly. These costs include the cost of obtaining the raw data, cleaning the data, storing the data, and transforming the data into a representation suitable for learning, as well as the cost of computer hardware, the cost associated with the time it takes to learn from the data, and the "opportunity cost" associated with suboptimal learning from extremely large data sets due to limited computational resources. When these costs make it necessary to limit the amount of training data, an important question is: in what proportion should the classes be represented in the training data? In answering this question, this article makes two main contributions. It addresses (for classification-tree induction) the practical problem of how to select the class distribution of the training data when the amount of training data must be limited, and, by providing a detailed empirical study of the effect of class distribution on classifier performance, it provides a better understanding of the role of class distribution in learning Some practitioners believe that the naturally occurring marginal class distribution should be used for learning, so that new examples will be classified using a model built from the same underlying distribution. Other practitioners believe that the training set should contain an increased percentage of minority-class examples, because otherwise the induced classifier will not classify minority-class examples well. This latter viewpoint is expressed by the statement, "if the sample size is fixed, a balanced sample will usually produce more accurate predictions than an unbalanced split". However, we are aware of no thorough prior empirical study of the relationship between the class distribution of the training data and classifier performance, so neither of these views has been validated and the choice of class distribution often is made arbitrarily—and with little understanding of the consequences. In this article we provide a thorough study of the relationship between class distribution and classifier performance and provide guidelines—as well as a progressive sampling algorithm—for determining a "good" class distribution to use for learning.

There are two situations where the research described in this article is of direct practical use. When the training data must be limited due to the cost of learning from the data, then our results- and the guidelines we establish—can help to determine the class distribution that should be used for the training data. In this case, these guidelines determine how many examples of each class to omit from the training set so that the cost of learning is acceptable. The second scenario is when training examples are costly

to procure so that the number of training examples must be limited. In this case the research presented in this article can be used to determine the proportion of training examples belonging to each class that should be procured in order to maximize classifier performance. Note that this assumes that one can select examples belonging to a specific class. This situation occurs in a variety of situations, such as when the examples belonging to each class are produced or stored separately or when the main cost is due to transforming the raw data into a form suitable for learning rather than the cost of obtaining the raw, labeled, data. Fraud detection provides one example where training instances belonging to each class come from different sources and may be procured independently by class. Typically, after a bill has been paid, any transactions credited as being fraudulent are stored separately from legitimate transactions. Furthermore transactions credited to a customer as being fraudulent may in fact have been legitimate, and so these transactions must undergo a verification process before being used as training data.

In other situations, labeled raw data can be obtained very cheaply, but it is the process of forming usable training examples from the raw data that is expensive. As an example, consider the phone data set, one of the twenty-six data sets analyzed in this article. This data set is used to learn to classify whether a phone line is associated with a business or a residential customer. The data set is constructed from low-level call-detail records that describe a phone call, where each record includes the originating and terminating phone numbers, the time the call was made, and the day of week and duration of the call. There may be hundreds or even thousands of call-detail records associated with a given phone line, all of which must be summarized into a single training example. Billions of call-detail records, covering hundreds of millions of phone lines, potentially are available for learning. Because of the effort associated with loading data from dozens of computer tapes, disk-space limitations and the enormous processing time required to summarize the raw data, it is not feasible to construct a data set using all available raw data. Consequently, the number of usable training examples must be limited. In this case this was done based on the class associated with each phone line—which is known. The phone data set was limited to include approximately 650,000 training examples, which were generated from approximately 600 million call-detail records. Because huge transaction-oriented databases are now routinely being used for learning, we expect that the number of training examples will also need to be limited in many of these cases.

5. A STUDY OF THE BEHAVIOR OF SEVERAL METHODS FOR BALANCING MACHINE LEARNING TRAINING DATA

Most learning systems usually assume that training sets used for learning are balanced. However, this is not always the case in real world data where one class might be represented by a large number of examples, while the other is represented by only a few. This is known as the class imbalance problem and is often reported as an obstacle to the induction of good classifiers by Machine Learning (ML) algorithms. Generally, the problem of imbalanced data sets occurs when one class represents a circumscribed concept, while the other class represents the counterpart of that concept, so that examples from the counterpart class heavily outnumber examples from the positive class. This sort of data is found, for example, in medical record databases regarding a rare disease, were there is a large number of patients who do not have that disease; continuous fault-monitoring tasks where non-faulty examples heavily outnumber faulty examples, and others. In recent years, there have been several attempts at dealing with the class imbalance problem in the field of Data Mining and Knowledge Discovery in Databases, to which ML is a substantial contributor. Related papers have been published in the ML literature aiming to overcome this problem. The ML community seems to agree on the hypothesis that the imbalance between classes is the major obstacle in inducing classifiers in imbalanced domains. However, it has also been observed that in some domains, for instance the Sick data set , standard ML algorithms are capable of inducing good classifiers, even using highly imbalanced training sets. This shows that class imbalance is not the only problem responsible for the decrease in performance of learning algorithms. we developed a systematic study aiming to question whether class imbalances hinder classifier induction or whether these deficiencies might be explained in other ways. Our study was developed on a series of artificial data sets in order to fully control all the variables we wanted to analyze. The results of our experiments, using a discrimination-based inductive scheme, suggested that the problem is not solely caused by class imbalance, but is also related to the degree of data overlapping among the classes. The results obtained in this previous work motivated the proposition of two new methods to deal with the problem of learning in the presence of class imbalance. These methods ally a known over-sampling method, namely Smote , with two data cleaning methods: Tomek links and Wilson's Edited Nearest Neighbor Rule. The main motivation behind these methods is not only to balance the training data, but also to remove noisy examples lying on the wrong side of the decision border. The removal of noisy examples might aid in finding better-defined class clusters, therefore, allowing the creation of simpler models with better generalization capabilities. In addition, in this work we perform a broad experimental evaluation involving ten methods, three of them proposed by the authors, to deal with the class

imbalance problem in thirteen UCI data sets. We concluded that over-sampling methods are able to aid in the induction of classifiers that are more accurate than those induced from under-sampled data sets. This result seems to contradict results previously published in the literature. Two of our proposed methods performed well in practice, in particular for data sets with a small number of positive examples. It is worth noting that Random over-sampling, a very simple over-sampling method, is very competitive to more complex over-sampling methods.

6. NEW APPLICATIONS OF ENSEMBLES OF CLASSIFIERS

Recently, efforts aimed at combining multiple classifiers into one classification system (ensemble of classifiers, multiple classifier systems, mixtures of experts, committees of learners, etc.) have become very popular, and are regarded as one of the most promising current research directions in machine learning and pattern recognition . The main purpose for building up an ensemble is to obtain higher classification accuracy than that produced by its components (individual classifiers that make it up). Ensembles have been defined as consisting of a set of individually trained classifiers whose decisions are combined when classifying new instances. The combination can be made in many different ways. The simplest employs the majority rule in a plain voting system. Despite its simplicity, it is generally regarded as a very robust combination. More elaborate schema use weight voting rules, where each component is associated with a weight. This weight is computed while training the classifier, and must reflect how accurate the individual classifier is, as estimated by its performance on the training set. Other, more sophisticated, architectures have also been proposed, consisting of two levels of classifiers in what has been called 'stacked-generalisation' or 'metalearning' (learning from the information generated by a set of learners). It is widely accepted that improvement in the overall predictive accuracy by the ensemble can occur only if there is diversity among its components, i.e. if the individual classifiers do not always agree. Of course, no benefit arises from combining the predictions of a set of classifiers that frequently coincide in the classifications (strongly correlated classifiers). Although measuring diversity is not straightforward , this classifiers' independence has been sought through different ways, by:

- Manipulating the training patterns (training each classifier on different subsets of the training prototypes): cross-validation, bagging, boosting, etc.
- Manipulating the input features (training each classifier with different subsets of the available features).
- Manipulating the class labels of the training prototypes.
- Incorporating random noise into the feature values or into some parameters of the learning model considered.

Most of the research done up to now has been concerned with the creation of ensembles consisting of classifiers based on the same learning model. Although it is likely that learning with different algorithms will produce diverse classifiers, this diversity is not guaranteed. Moreover, this approach would require the employment of an effective weighted combination, since some of these classifiers would perform much worse than others. Ensembles based on the combination of a set of classifiers are currently used to achieve higher recognition accuracy. In this paper, we explore possibilities to obtain other benefits from the employment of an ensemble. In particular, we present results of experiments carried out to research the convenience of using ensembles for three different tasks:

a) To cope with unbalanced training samples,
b) To get scalability of some pre-processing algorithms,
c) To filter the training sample.
In our research, we have focused on the widely used nearest neighbour rule. This selection has been motivated by the flexibility and other positive characteristics of this classification method.

7. HETEROGENEOUS UNCERTAINTY SAMPLING FOR SUPERVISED LEARNING

Machine learning algorithms have been used to build classification rules from data sets consisting of hundreds of thousands of instances. In some applications unlabeled training instances are abundant but the cost of labeling an instance with its class is high. In the information retrieval application described here the class labels are assigned by a human, but they could also be assigned by a computer simulation or a combination of both . The terms oracle and teacher have been used for the source of labels; we will usually call it the expert. Where one of the constraints on the induction process is a limit on the number of instances presented to the oracle, the choice of instances becomes important. Random sampling may be ineffective if one class is very rare: all of the training instances presented may have the majority class. To make more effective use of the expert's time, methods that we collectively call uncertainty sampling label data sets incrementally, alternating between two phases: presenting the expert a few instances to label, and selecting (from a finite or infinite source) instances whose labels are still uncertain despite the indications contained in previously labeled data. The type of classifier used in uncertainty sampling must be cheap to build and to use. At each iteration a new classifier is built (fortunately from a small sample) and then applied (unfortunately to a large sample). Our uncertainty sampling method also requires an estimate of the

certainty of classifications (a class-probability value); not all induction systems provide this. This paper examines a heterogeneous approach in which a classifier of one type selects instances for training a classifier of another type. It is motivated by applications requiring a type of classifier that would be too computationally expensive to use to select instances.

8. LEARNING FROM IMBALANCED DATA IN PRESENCE OF NOISY AND BORDERLINE EXAMPLES

In some real-life problems, the distribution of examples in classes is highly imbalanced, which means that one of the classes (further called a minority class) includes much smaller number of examples than the other majority classes. Class imbalance constitutes a difficulty for most learning algorithms, which assume even class distribution and are biased toward learning and recognition of the majority classes. As a result, minority examples tend to be misclassified. This problem has been intensively researched in the last decade and several methods have been proposed. They are usually divided into solutions on the data level and the algorithmic level. Solutions on the data level are classifier-independent and consist in transforming an original data distribution to change the balance between classes, e.g., by re-sampling techniques. Solutions on the algorithmic level involve modification of either learning or classification strategies. Some researchers also generalize ensembles or transform the imbalance problem to cost sensitive learning. In this paper we are interested in focused re-sampling techniques, which modify the class distribution taking into account local characteristics of examples. Inspired we distinguish between safe, borderline and noisy examples. Borderline examples are located in the area surrounding class boundaries, where the minority and majority classes overlap. Safe examples are placed in relatively homogeneous areas with respect to the class label. Finally, by noisy examples we understand individuals from one class occurring in safe areas of the other class. We claim that the distribution of borderline and noisy examples causes difficulties for learning algorithms, thus we focus our interest on careful processing of these examples. Our study is related to earlier works of Stefanowski and Wilk on selective pre-processing with the SPIDER (Selective Preprocessing of Imbalanced Data) method . This method employs the Edited Nearest Neighbor Rule (ENNR) to identify the local characteristic of examples, and then it combines removing the majority class objects that may result in misclassifying objects from the minority class with local over-sampling of these objects from the minority class that are "overwhelmed" by surrounding objects from the majority classes. Experiments showed that this method improved the recognition of the minority class and was competitive to the most related approaches SMOTE and NCR. The observed improvements varied over different imbalanced data sets, therefore, in this study we have decided to explore conditions, where the SPIDER method could be more efficient than simpler re-sampling methods. To achieve this goal we have planned controlled experiments with special artificial data sets. According to related works many experiments were conducted on real-life data sets (e.g., coming from UCI). The most well known studies with artificial data are the works of Japkowicz, who showed that simple class imbalance ratio was not the main difficulty. The degradation of performance was also related to other factors, mainly to small disjuncts, i.e., the minority class being decomposed into many sub-clusters with very few examples. Other researchers also explored the effect of overlapping between imbalanced classes – more recent experiments on artificial data with different degrees of overlapping also showed that overlapping was more important than the overall imbalance ratio.

Following these motivations we prepare our artificial data sets to analyze the influence of the presence and frequency of the noisy and borderline examples. We also plan to explore the effect of the decomposition of this class into smaller subclusters and the role of changing decision boundary between classes from linear to non-linear shapes. The main aim of our study is to examine which of these factors were critical for the performance of the methods dealing with imbalanced data. In the experiments we compare the performance of the SPIDER method and the most related focused re-sampling NCR method with the oversampling methods suitable to handle class decomposition and the basic versions of tree or rule-based classifiers.

9. LEARNING FROM IMBALANCED DATA

RECENT developments in science and technology have enabled the growth and availability of raw data to occur at an explosive rate. This has created an immense opportunity for knowledge discovery and data engineering research to play an essential role in a wide range of applications from daily civilian life to national security, from enterprise information processing to governmental decision-making support systems, from micro scale data analysis to macro scale knowledge discovery. In recent years, the imbalanced learning problem has drawn a significant amount of interest from academia, industry, and government funding agencies. The fundamental issue with the imbalanced learning problem is the ability of imbalanced data to significantly compromise the performance of most standard learning algorithms. Most standard algorithms assume or expect balanced class distributions or equal misclassification costs. Therefore, when presented with complex imbalanced data sets, these algorithms fail to properly represent the distributive characteristics of the data and

resultantly provide unfavourable accuracies across the classes of the data. When translated to real-world domains, the imbalanced learning problem represents a recurring problem of high importance with wide-ranging implications, warranting increasing exploration. This increased interest is reflected in the recent instalment of several major workshops, conferences, and special issues including the American Association for Artificial Intelligence (now the Association for the Advancement of Artificial Intelligence) workshop on Learning from Imbalanced Data Sets (AAAI '00) , the International Conference on Machine Learning workshop on Learning from Imbalanced Data Sets (ICML'03), and the Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining Explorations (ACM SIGKDD Explorations '04). With the great influx of attention devoted to the imbalanced learning problem and the high activity of advancement in this field, remaining knowledgeable of all current developments can be an overwhelming task. As can be seen, the activity of publications in this field is growing at an explosive rate. Due to the relatively young age of this field and because of its rapid expansion, consistent assessments of past and current works in the field in addition to projections for future research are essential for long-term development. In this paper, we seek to provide a survey of the current understanding of the imbalanced learning problem and the state-of-the-art solutions created to address this problem. Furthermore, in order to stimulate future research in this field, we also highlight the major opportunities and challenges for learning from imbalanced data.

10. C4.5, CLASS IMBALANCE, AND COST SENSITIVITY: Why Under-Sampling beats Over-Sampling

A study on the two most common sampling schemes used to adapt machine learning algorithms to imbalanced classes and misclassification costs. We look at under-sampling and oversampling, which decrease and increase, respectively, the frequency of one class in the training set to reflect the desired misclassification costs. These schemes are attractive as the only change is to the training data rather than to the algorithm itself. It is hard to justify a more sophisticated algorithm if it cannot outperform existing learners using one of these simple sampling schemes. Here, we study the sampling schemes and how they affect the decision tree learner C4.5, release 8 . We chose C4.5 not only because it is one of the most commonly used algorithms in the machine learning and data mining communities but also because it has become a de facto standard against which every new algorithm is judged. For research into cost sensitivity and class imbalance C4.5 combined with under-sampling or over-sampling is quickly becoming the accepted baseline for comparison. Using our own performance analysis technique, called cost curves, discussed briefly in the next section, we show that under sampling produces a reasonable sensitivity to changes in misclassification costs and class distribution. However, when using C4.5's default settings, over-sampling is surprisingly ineffective, often producing little or no change in performance as these factors are changed. We go on to explore which aspects of C4.5 result in under-sampling being so effective and why they fail to be useful for over-sampling. We have previously shown that the splitting criterion has relatively little effect on cost sensitivity. Observed that costs and class distribution primarily affect pruning. Still, we did not find that this was the main cause of the difference between the two sampling schemes. Oversampling tends to reduce the amount of pruning that occurs. Under-sampling often renders pruning unnecessary. By removing instances from the training set, it stunts the growth of many branches before pruning can take effect. We find that over-sampling can be made cost-sensitive if the pruning and early stopping parameters are set in proportion to the amount of over-sampling that is done. But the extra computational cost of using over-sampling is unwarranted as the performance achieved is, at best, the same as under-sampling.

11. USING RANDOM FOREST TO LEARN IMBALANCED DATA

Many practical classification problems are imbalanced; i.e., at least one of the classes constitutes only a very small minority of the data. For such problems, the interest usually leans towards correct classification of the "rare" class (which we will refer to as the "positive" class). Examples of such problems include fraud detection, network intrusion, rare disease diagnosing, etc. However, the most commonly used classification algorithms do not work well for such problems because they aim to minimize the overall error rate, rather than paying special attention to the positive class. Several researchers have tried to address the problem in many applications such as fraudulent telephone call detection, information retrieval and filtering, diagnosis of rare thyroid deceases and detection of oil spills from satellite images. There are two common approaches to tackle the problem of extremely imbalanced data. One is based on cost sensitive learning: assigning a high cost to misclassification of the minority class, and trying to minimize the overall cost. The other approach is to use a sampling technique: Either down-sampling the majority class or over-sampling the minority class, or both. Most research has been focused on this approach. SHRINK, for imbalanced classification. SHRINK labels a mixed region as positive (minority class) regardless of whether the positive examples prevail in the region or not. Then it searches for the best positive region. They made comparisons to C4.5 and 1-NN, and show that SHRINK has improvement in most cases. It uses the one-sided sampling technique to selectively down sample the majority

class. Over-sample the minority class by replicating the minority samples so that they attain the same size as the majority class. Over-sampling does not increase information; however by replication it raises the weight of the minority samples. Combine over-sampling and down-sampling to achieve better classification performance than simply down-sampling the majority class. Rather than over-sampling with replacement, they create synthetic minority class examples to boost the minority class (SMOTE). They compared SMOTE plus the down-sampling technique with simple down-sampling, one sided sampling and SHRINK, and showed favorable improvement. Apply the boosting procedure to SMOTE to further improve the prediction performance on the minority class and the overall F-measure.

We propose two ways to deal with the problem of extreme imbalance, both based on the random Forest (RF) algorithm. One incorporates class weights into the RF classifier, thus making it cost sensitive, and it penalizes misclassifying the minority class. The other combines the sampling technique and the ensemble idea. It down-samples the majority class and grows each tree on a more balanced data set. A majority vote is taken as usual for prediction. We compared the prediction performance with one-sided sampling, SHRINK, SMOTE, and SMOTE Boost on the data sets that the authors of those techniques studied. We show that both of our methods have favorable prediction performance.

12. CONCULSION

A new method called neighborhood cleaning rule is described that utilizes the OSS principle, but considers more carefully the quality of the data to be removed. Certain algorithms specifically modify the behavior of existing algorithms to make them more immune to noisy instances, such as the IB3 algorithm for kNN, or pruning of decision trees, or soft margins in SVM. The class distribution of training data and classifier performance with respect to accuracy and AUC is analyzed. The decision rules C4.5 produced from uncertainty samples of roughly 1,000 instances chosen by a probabilistic classifier were significantly more accurate than those from random samples ten times larger. New techniques are been used to solve the several issues of the rule based training data with the help of decision making process.

REFERENCE

[1] H. He and E.A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 9, pp. 1263-1284, Sept. 2009.

[2] J. Laurikkala, "Improving Identification of Difficult Small Classes by Balancing Class Distribution," Proc. Eighth Conf. AI in Medicine in Europe: Artificial Intelligence Medicine, pp. 63-66, 2001.

[3] G. Weiss and F. Provost, "The Effect of Class Distribution on Classifier Learning: An Empirical Study," technical report, Dept. of Computer Science Rutgers, Univ., 2001.

[4] K. McCarthy, B. Zarbar, and G. Weiss, "Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes?" Proc. Int'l Workshop Utility-Based Data Mining, pp. 69-77, 2005.

[5] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Data Sets," Proc. 15th European Conf. Machine Learning, 2004.

[6] S. Ertekin, J. Huang, and C.L. Giles, "Learning on the Border: Active Learning in Imbalanced Data Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf., pp. 823-824, 2007.

[7] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study," Intelligent Data Analysis, vol. 6, no. 5, pp. 429-449, 2002.

[8] G. Weiss and F. Provost, "Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction," J. Artificial Intelligence Research, vol. 19, pp. 315-354, 2003.

[9] G.E. Batista, R.C. Prati, and M.C. Monard, "A Study of the Behaviour of Several Methods for Balancing Machine Learning Training Data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20-29, 2004.

[10] Y. Sun, A.K.C. Wong, and M.S. Kamel, "Classification of Imbalanced Data: A Review," Int'l J. Pattern Recognition, vol. 23, no. 4, pp. 687-719, 2009.