Literature survey on classification of Machine Learning techniques effective on imbalanced dataset

Sravan Kumar Pagolu
AI Tech Systems
University College of Engineering JNTUK Narsaraopet,
Narsaraopet, Andhra Pradesh, India.
sravansmart7733@gmail.com

Abstract— **Unbalanced data set, a problem often found in real world application, can cause seriously negative effect on classification performance of machine learning algorithms. There have been many attempts at dealing with classification of unbalanced data sets. In this paper we present a brief review of existing solutions to the class-imbalance problem proposed both at the data and algorithmic levels. Even though a common practice to handle the problem of imbalanced data is to rebalance them artificially by oversampling and/or under-sampling, some researchers proved that modified support vector machine, rough set-based minority class-oriented rule learning methods, cost sensitive classifier perform good on imbalanced data set. We observed that current research in imbalance data problem is moving to hybrid algorithms.**

Keywords— cost-sensitive learning, imbalanced data set, modified SVM, oversampling, under sampling.

I. INTRODUCTION

A data set is called imbalanced if it contains many more samples from one class than from the rest of the classes. Data sets are unbalanced when at least one class is represented by only a small number of training examples (called the minority class) while other classes make up the majority. In this scenario, classifiers can have good accuracy on the majority class but very poor accuracy on the minority class(es) due to the influence that the larger majority class has on traditional training criteria. Most original classification algorithms pursue to minimize the error rate: the percentage of the incorrect prediction of class labels. They ignore the difference between types of misclassification errors. In particular, they implicitly assume that all misclassification errors cost equally. In many real-world applications, this assumption is not true. The differences between different misclassification errors can be quite large. For example, in medical diagnosis of a certain cancer, if the cancer is regarded as the positive class, and non-cancer (healthy) as negative, then missing a cancer (the patient is actually positive but is classified as negative; thus it is also called ―false negative‖) is much more serious (thus expensive) than the false-positive error.

The patient could lose his/her life because of the delay in the correct diagnosis and treatment. Similarly, if carrying a bomb is positive, then it is much more expensive to miss a terrorist who carries a bomb to a flight than searching an innocent person. The unbalanced data set problem appears in many real world applications like text categorization, fault detection, fraud detection, oil-spills detection in satellite images, toxicology, cultural modelling, medical diagnosis.[1] Many research papers on imbalanced data sets have commonly agreed that because of this unequal class distribution, the performance of the existing classifiers tends to be biased towards the majority class. The reasons for poor performance of the existing classification algorithms on imbalanced data sets are: 1. They are accuracy driven i.e.; their goal is to minimize the overall error to which the minority class contributes very little. 2. They assume that there is equal distribution of data for all the classes. 3. They also assume that the errors coming from different classes have the same cost [2]. With unbalanced data sets, data mining learning algorithms produce degenerated models that do not take into account the minority class as most data mining algorithms assume balanced data set. A number of solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels [3]. At the data level, these solutions include many different forms of re-sampling such as random oversampling with replacement, random under sampling, directed oversampling (in which no new examples are created, but the choice of samples to replace is informed rather than random), directed under sampling (where, again, the choice of examples to eliminate is informed), oversampling with informed generation of new samples, and combinations of the above techniques. At the algorithmic level, solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning. The most common techniques to deal with unbalanced data include resizing training datasets, cost-sensitive classifier, and snowball method. Recently, several methods have been proposed with good performance on unbalanced data. These approaches include modified SVMs, k nearest neighbor (kNN), neural networks, genetic programming, rough set-based algorithms, probabilistic decision tree and learning methods. The next sections focus on some of the method in detail.

II. SAMPLING METHODS

An easy Data level method for balancing the classes consists of resampling the original data set, either by oversampling the minority class or by under-sampling the majority class, until the classes are approximately equally represented. Both strategies can be applied in any learning system, since they act as a preprocessing phase, allowing the learning system to receive the training instances as if they belonged to a well-balanced data set. Thus, any bias of the system towards the majority class due to the different proportion of examples per class would be expected to be suppressed. Hulse et al. [4] suggest that the utility of the resampling methods depends on a number of factors, including the ratio between positive and negative examples, other characteristics of data, and the nature of the classifier. However, resampling methods have shown important drawbacks. Under-sampling may throw out potentially useful data, while over-sampling artificially increases the size of the data set and consequently, worsens the computational burden of the learning algorithm. A. Oversampling The simplest method to increase the size of the minority class corresponds to random over-sampling, that is, a no heuristic method that balances the class distribution through the random replication of positive examples. Nevertheless, since this method replicates existing examples in the minority class, overfitting is more likely to occur. Chawla proposed Synthetic Minority Over-sampling Technique (SMOTE) [5] an over-sampling approach in which the minority class is over-sampled by creating ―synthetic‖ examples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

From the original SMOTE algorithm, several modifications have been proposed in the literature. While SMOTE approach does not handle data sets with all nominal features, it was generalized to handle mixed datasets of continuous and nominal features. Chawla propose SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous) and SMOTE-N (Synthetic Minority Over-sampling Technique Nominal), the SMOTE can also be extended for nominal features. Andrew Estabrook's et al. proposed a multiple resampling method which selected the most appropriate re-sampling rate adaptively [6]. Taiho Jo et al. put forward a cluster-based over-sampling method which dealt with between-class imbalance and within-class imbalance simultaneously [7]. Hongyu Guo et al. found out hard examples of the majority and minority classes during the process of boosting, then generated new synthetic examples from hard examples and add them to the data sets [8].Based on SMOTE method, Hui Han and Wen-Yuan Wang [9] presented two new minority over-sampling methods, borderline-SMOTE1 and borderline-SMOTE2, in which only the minority examples near the borderline are oversampled. These approaches achieve better TP rate and Fvalue than SMOTE and random over-sampling methods. B. Undersampling Under-sampling is an efficient method for classing imbalance learning. This method uses a subset of the majority class to train the classifier. Since many majority class examples are ignored, the training set becomes more balanced and the training process becomes faster. The most common preprocessing technique is random majority under-sampling (RUS), IN RUS, Instances of the majority class are randomly discarded from the dataset. However, the main drawback of under-sampling is that potentially useful information contained in these ignored examples is neglected. There many ways attempts to improve upon the performance of random sampling, such as Tomek links, Condensed Nearest Neighbor Rule and One-sided selection etc. one-sided selection (OSS) is proposed by Rule Kubat and Matwin attempts to intelligently under-sample the majority class by removing majority class examples that are considered either redundant or ‗noisy.' Over-sampling is a method for improve minority class recognition, randomly duplicate the minority data not only without increase any category of a small number of new information, but also will lead to over-fitting.

For some problems like fraud detection which is highly overlapped unbalanced data classification problem, where non-fraud samples heavily outnumber the fraud samples, T. Maruthi Padmaja[10] proposed hybrid sampling technique, a combination of SMOTE to over-sample the minority data (fraud samples) and random undersampling to under-sample the majority data (non-fraud samples) if we eliminate extreme outliers from the minority samples for highly skewed imbalanced data sets like fraud detection classification accuracy can be improved. Sampling methods consider the class skew and properties of the dataset as a whole. However, machine learning and data mining often face nontrivial datasets, which often exhibit characteristics and properties at a local, rather than global level. It is noted that a classifier improved through global sampling levels may be insensitive to the peculiarities of different components or modalities in the data, resulting in a suboptimal performance. David A. Cieslak, Nitesh V. Chawla [11] has suggested that for improving classifier performance sampling can be treated locally, instead of applying uniform levels of sampling globally. They proposed a framework which first identifies meaningful regions of data and then proceeds to find optimal sampling levels within each. There are known disadvantages associated with the use of sampling to implement cost-sensitive learning. The disadvantage with undersampling is that it discards potentially useful data. The main disadvantage with oversampling, from our perspective, is that by making exact copies of existing examples, it makes overfitting likely. In fact, with oversampling it is quite common for a learner to generate a classification rule to cover a single, replicated, example. A second disadvantage of oversampling is that it increases the number of training examples, thus increasing the learning time. Given the disadvantages with sampling, still sampling is a popular way to deal with imbalanced data rather than a cost-sensitive learning algorithm. There are several reasons for this. The most obvious reason is there are not cost sensitive implementations of all learning algorithms and therefore a wrapper-based approach using sampling is the only option. While this is certainly less true today than in the past, many learning algorithms (e.g., C4.5) still do not directly handle costs in the learning process. A second reason for using sampling is that many highly skewed data sets are enormous and the size of the training set must be reduced in order for learning to be feasible.

In this case, undersampling seems to be a reasonable, and valid, strategy. if one needs to discard some training data, it still might be beneficial to discard some of the majority class examples in order to reduce the training set size to the required size, and then also employ a cost sensitive learning algorithm, so that the amount of discarded training data is minimized. A final reason that may have contributed to the use of sampling rather than a cost-sensitive learning algorithm is that misclassification costs are often unknown. However, this is not a valid reason for using sampling over a cost-sensitive learning algorithm, since the analogous issue arises with sampling—what should the class distribution of the final training data be? If this cost information is not known, a measure such as the area under the ROC curve could be used to measure classifier performance and both approaches could then empirically determine the proper cost ratio/class distribution [12].

III. COST-SENSITIVE LEARNING

At the algorithmic level, solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning. Cost-Sensitive Learning is a type of learning in data mining that takes the misclassification costs (and possibly other types of cost) into consideration. There are many ways to implement cost sensitive learning, in [13], it is categorized into three, the first class of techniques apply misclassification costs to the data set as a form of data space weighting, the second class applies cost-minimizing techniques to the combination schemes of ensemble methods, and the last class of techniques incorporates cost sensitive features directly into classification paradigms to essentially fit the cost sensitive framework into these classifiers. Incorporating cost into decision tree classification algorithm which is one of the most widely used and simple classifiers. Cost can be incorporated into it in various ways. First way is cost can be applied to adjust the decision threshold, second way is cost can be used in splitting attribute selection during decision tree construction and the other way is cost sensitive pruning schemes can be applied to the tree. Ref. [14] propose a method for building and testing decision trees that minimizes total sum of the misclassification and test costs.

The algorithm used by them chooses an splitting attribute that minimizes the total cost, the sum of the test cost and the misclassification cost rather than choosing an attribute that minimizes the entropy. Information gain, Gini measures are considered to be skew sensitive [15]. In Ref. [16] a new decision tree algorithm called Class Confidence Proportion Decision Tree (CCPDT) is proposed which is robust and insensitive to size of classes and generates rules which are statistically significant. Ref. [17] analytically and empirically demonstrates the strong skew insensitivity of Hellinger Distance and its advantages over popular alternative metrics. They arrived at a conclusion that for imbalanced data it is sufficient to use Hellinger trees with bagging without any sampling methods. Ref. [18] uses different operators of Genetic algorithms for oversampling to enlarge the ratio of positive samples and then apply clustering to the oversampled training data set as a data clearing method for both classes, removing the redundant or noisy samples. They used AUC as evaluation metric and found that their algorithm performed better. Nguyen ha vo, Yonggwan won [19] extended Regularized Least Square(RLS) algorithm that penalizes errors of different samples with different weights and some rules of thumb to determine those weights. The significantly better classification accuracy of weighted RLS classifiers showed that it is promising substitution of other previous cost-sensitive classification methods for unbalanced data set. this approach is equivalent to up sampling or down-sampling depending on the cost we choose. For example, doubling the cost-sensitivity of one class is said to be equivalent to doubling the number of samples in that class. Ref[20] proposed a novel approach reducing each within group error, BABoost that is a variant of AdaBoost. Adaboost algorithm gives equal weight to each misclassified example. But the misclassification error of each class is not same. Generally, the misclassification error of the minority class will larger than the majority 's. So Adaboost algorithm will lead to higher bias and smaller margin when encountering skew distribution. BABoost algorithm in each round of boosting assigns more weights to the misclassified examples, especially those in the minority class. Yanmin Sun a and Mohamed S. Kamel [21] explored three cost-sensitive boosting algorithms, which are developed by introducing cost items into the learning framework of AdaBoost.

These boosting algorithms are also studied with respect to their weighting strategies towards different types of samples, and their effectiveness in identifying rare cases through experiments on several real world medical data sets, where the class imbalance problem prevails.

IV. SVM AND IMBALANCED DATASETS

The success of SVM is very limited when it is applied to the problem of learning from imbalanced datasets in which negative instances heavily outnumber the positive instances. Even though undersampling the majority class does improve SVM performance, there is an inherent loss of valuable information in this process. Rehan Akbani[22]combined sampling and cost sensitive learning for improving performance of SVM. Their algorithm is based on a variant of the SMOTE algorithm by Chawla et al, combined with Veropoulos et al's different error costs algorithm. TAO Xiao-yan[23] presented A modified proximal support vector machine (MPSVM) which assigns different penalty coefficients to the positive and negative samples respectively by adding a new diagonal matrix in the primal optimization problem. And further the decision function is obtained. The real-coded immune clone algorithm (RICA) is employed to select the global optimal parameters to get the high generalization performance. M. Muntean 1 and H. Vălean [24] provided the Enhancer, a viable algorithm for improving the SVM classification of unbalanced datasets. They improve the Cost-sensitive classification for Support Vector Machines, by multiplying in the training step the instances of the underrepresented classes. Yuchun Tang and Nitesh Chawla [25] also implemented and rigorously evaluated four SVM modeling techniques SVM can be effective if incorporate different ―rebalance‖ heuristics in SVM modeling, including cost-sensitive learning, and over and undersampling. Genetic programming (GP) can evolve biased classifiers when data sets are unbalanced.

The cost sensitive learning uses cost adjustment within the learning algorithm to factor in the uneven distribution of class examples in the original (unmodified) unbalanced data set, during the training process. In GP, cost adjustment can be enforced by adapting the fitness function. Here, solutions with good classification accuracy on both classes are rewarded with better fitness, while those that are biased toward one class only are penalized with poor fitness.

Common techniques include using fixed misclassification costs for minority and majority class examples [26], [27], or improved performance criteria such as the area under the receiver operating characteristic (ROC) curve (AUC) [28], in the fitness function. While these techniques have substantially improved minority class performances in evolved classifiers, they can incur both a tradeoff in majority class accuracy and, thus, a loss in overall classification ability, and long training times due to the computational overhead in evaluating these improved fitness measures. In addition, these approaches can be problem specific, i.e., fitness functions are handcrafted for a particular problem domain only.

V. HYBRID ALGORITHMS

The EasyEnsemble classifier is an under-sampling algorithm, which independently samples several subsets from negative examples and one classifier is built for each subset. All generated classifiers are then combined for the final decision by using Adaboost. in imbalanced problems, some features are redundant and even irrelevant, these features will hurt the generalization performance of learning machines. Feature selection, a process of choosing a subset of features from the original ones, is frequently used as a preprocessing technique in analysis of data. It has been proved effective in reducing dimensionality, improving mining efficiency, increasing mining accuracy and enhancing result comprehensibility. Ref[29] combined the feature selection method with EasyEnsemble in order to improve the accuracy. In ref[30] a hybrid algorithm based on random oversampling, decision tree (DT), particle swarm optimization (PSO) and feature selection is proposed to classify unbalanced data. The proposed algorithm has the ability to select beneficial feature subsets, automatically adjust values of parameter and obtain the best classification accuracy. The zoo dataset is used to test the performance. From simulation results, the classification accuracy of this proposed algorithm outperforms other existing methods Decision trees, supplemented with sampling techniques, have proven to be an effective way to address the imbalanced data problem. Despite their effectiveness, however, sampling methods add complexity and the need for parameter selection. To bypass these difficulties a new decision tree technique called Hellinger Distance Decision Trees (HDDT) which uses Hellinger distance as the splitting criterion is suggested in ref [17].

They took advantage of the strong skew insensitivity of Hellinger distance and its advantages over popular alternatives such as entropy (gain ratio). For imbalanced data it is sufficient to use Hellinger trees with bagging without any sampling methods.

VI. CONCLUSION

This paper provides an overview of the classification of imbalanced data sets. At data level, sampling is the most common approach to deal with imbalanced data. oversampling clearly appears as better than under-sampling for local classifiers, whereas some under-sampling strategies outperform over-sampling when employing classifiers with global learning. Researchers proved that Hybrid sampling techniques can perform better than just oversampling or undersampling. At the algorithmic level, solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning. Solutions based on modified support vector machine, rough set-based minority class oriented rule learning methods, cost sensitive classifier are also proposed to deal with unbalanced data. There are of course many other worthwhile research possibilities that are not included here. Developing Classifiers which are robust and skew insensitive or hybrid algorithms can be point of interest for the future research in imbalanced dataset.

REFERENCES

[1] Szil´ard Vajda, Gernot A. ―Fink Strategies for Training Robust Neural Network Based Digit Recognizers on Unbalanced Data Set 2010‖ 12th International Conference on Frontiers in Handwriting Recognition

[2] C.V. KrishnaVeni,T. Sobha Rani ―On the Classification of Imbalanced Datasets‖ IJCST Vol . 2, SP 1, December 2011

[3] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz ―Editorial: Special Issue on Learning from Imbalanced Data Sets‖ Sigkdd Explorations. Volume 6, Issue 1

[4] Hulse, J., Khoshgoftaar, T., Napolitano, A ―Experimental perspectives on learning from im-balanced data‖ In: Proceedings of the 24th International Conference on Machine learning, pp. 935–942 (2007)

[5] Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. ―SMOTE: Synthetic minority over-sampling technique‖. Journal of Artificial Intelligence Research 16, 321–357 (2002)

[6] Andrew Estabrooks, Taeho Jo and Nathalie Japkowicz ―Multiple Resampling Method for Learning from Imbalanced Data Sets. Comprtational Intelligence 20 (1) (2004) 18-36

[7] Taeho Jo, Nathalie Japkowicz ―Class Imbalances versus Small Disjuncts‖. Sigkdd Explora-tions 6 (1) (2004) 40-49

[8] Hongyu Guo, Herna L Viktor: ―Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach ‖ . Sigkdd Explorations 6 (1) (2004) 30-39

[9] Hui Han, Wen-Yuan Wang, Bing-Huan Mao ― Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning ‖ ICIC 2005, Part I, LNCS 3644, pp. 878-887, 2005.

[10] T. Maruthi Padmaja , Narendra Dhulipalla , Raju S. Bapi , P.Radha Krishna ― Unbalanced Data Classification Using extreme outlier Elimination and

Sampling Techniques for Fraud Detection ‖ 15th International Conference on Advanced Computing and Communications

[11] David A. Cieslak, Nitesh V. Chawla ―Start Globally, Optimize Locally, Predict Globally: Improving Performance on Imbalanced Data ‖ 2008 Eighth IEEE International Conference on Data Mining

[12] Gary M. Weiss, Kate McCarthy, and Bibi Zabar ― Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? ‖

[13] Haibo He, Edwardo A. Garcia, ― Learning from Imbalanced Data ‖ , 2009.

[14] Charles X. Ling, Qiang Yang, Jianning Wang, Shichao Zhang, ―Decision Trees with Minimal Costs ‖ , 2004.

[15] David A. Cieslak, Nitesh V. Chawla, ―Learning DecisionTrees for Unbalanced Data ‖ , 2008.

[16] Wei Liu, Sanjay Chawla, David A. Cieslak, Nitesh V. Chawla, ― A Robust Decision Tree Algorithm for Imbalanced Data Sets ‖ , 2010.

[17] David A. Cieslak, T. Ryan Hoens,Nitesh V. Chawla, W. Philip Kegelmeyer ―Hellinger Distance Decision Trees are Robust and Skew-Insensitive ‖ The Journal of Data Mining and Knowledge Discovery, 2011

[18] Satyam Maheshwari, Prof. Jitendra Agarwal, Dr. Sanjeev Sharma, ― A New Approach for Classification of Highly Imbalanced Datasets Using Evolutionary Algorithms ‖ , 2011.

[19] NGUYEN HA VO, YONGGWAN WON ― Classification of Unbalanced Medical Data with Weighted Regularized Least Squares ‖ Frontiers in the Convergence of Bioscience and Information Technologies 2007

[20] Jie Song, Xiaoling Lu, Xizhi Wu ―An Improved AdaBoost Algorithm for Unbalanced Classification Data ‖ 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery

[21] Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, Yang Wang ―Cost-sensitive boosting for classification of imbalanced data ‖ Pattern Recognition 40 (2007) 3358–3378

[22] Rehan Akbani, Stephen Kwek , Nathalie Japkowicz ―Applying Support Vector Machines to Imbalanced Datasets ‖

[23] TAO Xiao-yan, JI Hong-bing ―A Modified PSVM and its Application to Unbalanced Data Classification ‖ Third International Conference on Natural Computation (ICNC 2007)

[24] M. Muntean , H. Vălean, I. Ileană , C. Rotar ―Improving Classification with Support Vector Machine for Unbalanced Data ‖

[25] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser, ―SVMs Modeling for Highly Imbalanced Classification ‖ IEEE transactions on systems, man, and cybernetics—part b: cybernetics, vol. 39, no. 1, february 2009

[26] J. H. Holmes, ―Differential negative reinforcement improves classifier system learning rate in two-class problems with unequal base rates, ‖ in Proc. 3rd Annu. Conf. Genetic Program pp. 635–644.

[27] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk, ―Reducing misclassification costs, ‖ in Proc. 11th Int. Conf. Mach. Learn., 1994,pp. 217–225.

[28] A. P. Bradley, ―The use of the area under the ROC curve in the evaluation of machine learning algorithms, ‖ Pattern Recognit., vol. 30, no. 7,pp. 1145‑1159, Jul. 1997.

[29] Tian-Yu Liu ―EasyEnsemble and Feature Selection for Imbalance Data Sets ‖ 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing

[30] C.Y. Lee, M.R.Yang, L.Y. Chang, Z. 1. Lee ―A Hybrid Algorithm Applied to Classify Unbalanced Data.