International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Imbalanced Data set in machine Learning : A Comparative Study

Paarth Gupta[1], Pratyush Kumar[2], Manoj Kumar[3]
[1]DoCSE,SMVDU, Katra, Jammu and Kashmir, India
[2]DoCSE,SMVDU, Katra, Jammu and Kashmir, India
[3]AssistantProfessorDoCSE, Katra, Jammu and Kashmir, India

ABSTRACT

A system should be termed as intelligent only when it has the capability of self-learning. Machine learning being one of the most prominent field of computer science can help the system being able to get into the self-learning mode without the need of explicit programming efforts. The major challenge faced by Machine Learning experts in real life scenarios is uneven data distribution leading to imbalanced data set. Thus the proper distribution of elements in the form of sets plays a major role in achieving the self-learning goal. The uneven distribution of elements can be broadly categorized in the majority (negative) class and the minority (positive) class. The distribution of elements in nearly equal proportions is called as balanced data set. A data set is imbalanced when we have a minority class (I.e. the class which is rarer than the other classes namely the majority class). Dealing minority class is becoming more complex as classification rules tend to be fewer and weaker as compared to majority classes. Recent research findings in the area of machine learning along with the data mining have provided deeper insight into the nature of imbalanced learning along with the newer emerging challenges. Thus, this area of research is still popular among research community. In this paper we are focusing on the challenges and its best fit solutions available. Our aim to find the best fit solution by using different machine learning techniques or algorithms. These algorithms may vary in their approaches to solve the given problem. These approaches can be sampling, clustering, Graphical techniques, and statistical techniques or even with the help of classifiers. This paper provides a discussion on the complication of imbalanced data set and solutions concerning lines of future research for each of them.

Keywords: Machine learning, Imbalanced data, Imbalanced clustering, Sampling, Classifiers, Self-Learning.

I. INTRODUCTION

In the recent research done by the researchers in the field of self-learning, they came out with the few building blocks required for an intelligent system. One of the major building blocks is the concept of machine learning which focusses on the development of the computer programs that can access the data and use it for learning purpose. For the effective working, the complete data is divided into various classes based on various parameters. To achieve the objective of partitioning the data in classes we also prefer to have uniform distribution which we also call as balanced data set. But in real life scenario we might come across some cases in which the distribution of classes is not uniform and hence leads to the problem of imbalanced data set. The imbalanced class problem has become a very common problem with the emergence of machine learning. Its importance grow when researchers realized that their dataset is imbalance and this may cause suboptimal classification performance. Among those classes one is major class (having more number of instances) and another is the minor class. Thus,

generally the standard classifiers selects the instance from the major class because the ratio of elements in major class to smaller is 1:100,1:1000,1:10000 (and sometimes even more) [1]. This problem is very common in many real life situations and applications like [1] [2] detection of fraudulent telephone calls or credit card transactions, medical problems. Class imbalances have been also observed in many other application problems such as detection of oil spills in satellite images, analyzing financial risk, predicting technical equipment failures, managing network intrusion, text categorization andinformationfiltering. Let us consider an example of class imbalance in which instances in training data belonging to one class heavily outnumber the instances of the other class. E.g. we are visiting the hospital for the test of a rare disease, the chances are we might be having the disease or we might not be having the disease. The possibility of having such a disease is very less so the data set formed is imbalanced. Considering we took the test and the possibilities are the test might be positive (having the disease) or negative. Taking the first case if the result is positive and the test went out correctly which means we have that disease. The test might have gone wrong which will at max lead us to some more tests but if the test comes out negative and the test went out correctly which means we are not having the disease. The last and the most important case is that if the test comes out negative but the test did not go correctly this will mislead the patient as he will be thinking that he is fine but the truth is he was have ng that disease but the test didn't go properly. From this situation we can conclude that it is difficult to apply the concept of machine learning when we are dealing with the situation ofminorityclass[1].

In this real life situation, data describing an infrequent but important event, the learning system may have difficulties to learn the concept related to the minority class. In a data set with the class imbalance problem, the most obvious characteristic is the skewed data distribution between classes [3]. From the recent research this is not only the main problem for the data set including skewed there are various points such as small sample size and the existence of within-class sub concepts. Based on the nature of the problem of the imbalanced data set there are many problems occur such as imbalanced class distribution, small sample size, within class-subclass problem. In further section we are going to discuss the complications related to imbalanced data set and their best known solutions.

II. COMPLICATIONANDITSSOLUTION

This section of our research is based on the thought that though the degree of imbalance is one of the factor that hinders learning but is not the only factor that causes the hindrance. As it turns out, data-set complexity is also the primary determiningfactorofclassificationdeterioration.

2.1 Addressing the imbalanced class problem:

preprocessingandcostsensitivelearning:

This section deals with the problem of two-class imbalance problem both for standard learning algorithms and for ensemble techniques. These approaches are basically divided in three differentgroups.

2.1.1Data–LevelMethods

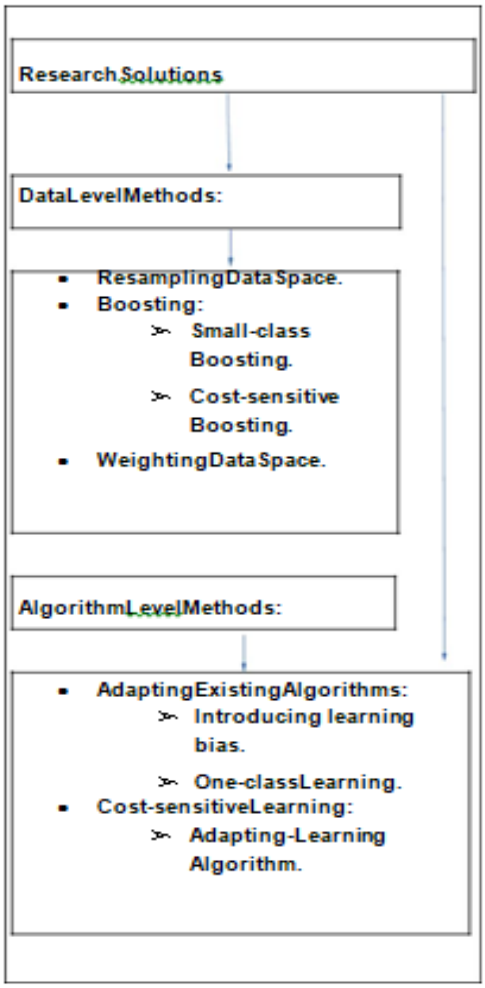
This is one of the easiest methods to handle the above described problem in which we modify the already existing collection of examples just to balance the distribution or in some case we may remove some difficult samples [4][5]. But the only disadvantage of this method is that sometimes all the samples are important and we cannot afford to ignore any one of the existing sample.

2.1.2Algorithm–LevelMethods

This is one of the important methods in which instead of modifying the data we modify the existing algorithms to work efficiently with the minority class. This method gives more accurate result but is less efficient as we might have to change the existing algorithms according to the problem we need to solve and because of this modification the algorithm might take some extra timeaswell[4][6].

2.1.3HybridMethods:

It is basically the combination of the first two methods taking the advantages of both the methods. It is quite a flexible method as we have the flexibility to modify the algorithms according to the need in the problem and we can also remove some of the very rare samples to make the data set balanced and obtain high efficiency [4][7][8].



2.2 Binaryimbalancedclassproblem

This is a problem which was experienced taking various real life applications in consideration like, patient (sick or healthy), access request (valid/authentic or malicious) and so on. In all these scenarios we usually have two options to look forward. These two options are responsible for the construction of two different classes (majority and minority) which are well defined andcanbeeasilydistinguished[9].

The best way to handle such a problem can vary on the type of problem but the best possible solution in most of the case is to try to balance the classes that we already have. To balance the already existing imbalanced classes we usually use the concept of oversampling or undersampling according to the problem we may face. If we use the under sampling method, it creates a subset of the original data-set by eliminating some of the examples of the majority class so that both the class tend to be balance. On the other hand if we use oversampling methods that create a superset of the original data-set by replicating some of the examples of the minority class or creating new ones from the original minority class instances with the same objective [1] [10]. Some of the techniques that uses oversampling or undersampling methods are:

2.2.1 Synthetic Minority Oversampling Technique (SMT)

In this technique the concept of oversampling is used by taking each minority class sample and introducing some synthetic examples along the line segments joining all of the k minority class nearest neighbors. The requirement of amount of oversampling decides the neighbors that is selected from the k-nearest neighbors. While applying the SMT, some majority class examples invade the minority class space and vice versa can also be possible, since the minority class clusters can be expanded by the interpolating of minority class examples. So, there is need of introducing artificial minority class examples deeply

into the majority class space. In this situation induction of classifiers can lead to overfitting, in this scenario we use SMT-ENN (edited nearest neighbor), this is the extended version of SMT. As we have already discussed SMT randomly synthesizes minority instances along a line joining a minority instance and it's selected nearest neighbors, while it ignores the nearby majority class instances [1] [10].

2.2.2Random-Oversampling(ROS)

Random-Oversampling is one of the non-heuristic method which has the primary aim to balance the already existing imbalanced classes through the process of replication of the minority class instances as discussed earlier. Recent research proves that random over-sampling can increase the likelihood of occurring overfitting, since it makes exact copies of the minority class examples. This process also increases the computational task if the data set is already fairly large but imbalanced. The major drawback of this technique is that it increases the duplicate samples in the minority class as it follows the simple process of replication[5][15].

2.2.3Random-Undersampling(RUS)

Random-Undersampling is another non-heuristic method which has the primary aim to balance the already existing imbalanced classes through the process of elimination of majority class samples. The logic behind doing such a thing is that it tries to balance out the dataset. The few disadvantages of this technique are that it might discard some of the potentially useful data from the dataset that could be important for the induction process and another problem with this approach is that in the estimation the probability distribution since the distribution is unknown so we take the help of samples available to us [5] [15].

2.3Multi-classImbalanceClassification

In imbalanced data set classification there are some case where we need multiple class distribution and while distributing we may obtain some imbalanced classes as well. Considering a real life scenario of intrusion detection we may have more than two classes of imbalanced class distributions and this hinder the classification performance. Here we take an example of network intrusion detection problem [11] [12], in this distribution while classification of dataset each record represents either an intrusion or a normal connection. Four kinds of attacks are possible in this problem but in the detection of rare classes among fours attacks have very low identification percentage as compared with the other attacks. This increases the complication in the classification performance of the imbalanced data set. The presence of the multiple imbalanced classes in classification of the data set results in complicated situations, so those methods that tackle the class imbalance problem of binary applications are not diretly applicable. For binary-class applications, we have to change the solutions at data level to change the class size ratio of the two classes, either by performing oversampling on the smaller class or down-sampling on the prevalent class, and to get the optimal distribution we run the learning algorithm many times. But due to increase in sample space practically binary class application is not feasible in presence of multiple classes. So, to resolve this problem we extend the binary classifiers. We also use algorithm to boost up the binary-class application to handle the multiple class imbalance problem of imbalanced data set. Here we use the AdaC2.M1 [12] [13] algorithm which has the task to advance the classification of the multi-class imbalanced class. It is a cost sensitive boosting algorithm. It's very effective in biasing the learning from the data set that is directed by the cost setups generated by GA, and eventually creates a significant improvement in the identification performance of those rare instances.

2.3.1Static-SMT

One of the technique to solve the multi-class imbalance classification is the Static-SMT, a preprocessing mechanism. In this technique the resampling procedure is usually applied in 'n' steps, where n is the number of classes of the problem. It uses the oversampling technique which has been discussed earlier, with each iteration the resampling procedure selects the minimum size class (minority class) and performs the oversampling by adding the duplicates of the instances of the class in the originaldata-set[1][10].

Comparative analysis with the help of experimental study over the multi class classificationmethodologies: In data mining algorithms, because of the imbalance in the classes formed we have to face a lot of problems. One of the problems is related to the boundaries among the classes i.e. the boundaries among the classes may overlap as there is a very little difference in the samples that lie on the boundaries.
The main problem because of the overlap of boundary samples is that this causes reduce in the performance level. Having such a situation, one of the best possible solution is to reduce the gap between binary class and multiple class imbalanced dataset. In order to implement this technique we have two different strategies[14]:

1. We can try to divide the multiclass problem into simpler binary sub-problems.
2. For each sub-problem that we have we can apply the solutions of two-class imbalanceddata-sets.
3. ignoring the examples that do not belong to the relatedclasses[15].

The OVA (One-versus-all approach) builds a single classifier for each of the classes of the problem, considering the examples of the current class to be positives and the remaining instancesnegatives[16].

2.4 Learning Difficulties with Standard Classifier Modeling Algorithms:

In this section, a subset of well-developed classifier learning algorithms over imbalanced data set is discussed. One of the most popularly used algorithm being the decision tree algorithm, which uses the simple knowledge representation to classify various examples into a finite number of classes. The concept is nearly the same as that in the data structure, where the nodes of the tree represent the content or the attributes. Their edges will be representing the possiblevalues

Table 1

Approach	Algorithm	Remark	Reference	Fundamental
OVO + Cost sensitive/ Oversampling.	DecisionTree	Average	Basic algorithm andsimpleOVO	Recursively Splitting thetrainingdata.
	Support Vector Machine	Trulycompetitive	OVO + preprocessing	Binarization and seeking an optimal separating hyperplane to maximize the margin and minimize thetrainingerror.

	K-Nearest Neighbor	Robust performance	Global CS, SL-SMT,SMT	Decidin g the class label of test sample by the most abundant class within the K-NearestNeighbors.
	Associatio n Classifiers	Average	OVO and other Algorithm	Derivin g classification rules from association pattern.

For the conversion of multiple-class classification problem to a set of binary classification problems we have various techniques, some of them being the OVO (pair wise learning) and OVA approaches.

In OVO technique our main focus is to train the classifier for each possible pair of classes, for corresponding node and finally the leaves which have the responsibility to represent the class labels. The only problem with such technique is that it faces a lot of difficulty in the constructionoftree.

2.4.1DecisionTree

A decision tree can be modelled in two phases named as tree building and tree pruning. The step of tree pruning overcome the overfitting of the training samples and it also improves the generalization capability of a decision tree by trimming the branches of the initial tree. In class imbalance problem decision trees may need to create many tests to distinguish the small classes from the large classes efficiently. In other learning processes, the branches for predicting the small classes may be pruned as being susceptible to overfitting. The basic principle behind pruning is that of predicting errors as there is a high probability that some branches that predict the small classes are removed and the new leaf node is labeled with a dominantclass[17].

2.4.2.Backpropagationneuralnetworks

Another most widely used technique to solve such a problem is by the backpropagation (BP) algorithm, which is most widely used model for such classification problems. In a backpropagation neural network, it usually comprises of one input layer, one output layer, and one or more hidden layers. In each layer we have one or more neurons. The first step of this technique involves initializing the weights to random numbers ranging between -1 to 1. Then our aim is to train the BP network using iterative approach.

While applying this technique we can observe that the error for the samples in the prevalent class reduces whereas the samples for the small classesincreases[18][19].

2.4.3.K-nearestneighbor

Another important and simple solution is K-Nearest Neighbor (KNN) which is an instance-based classifier learning algorithm in which we use specific training instances to make predictions without having to maintain a model derived from data. In this algorithm we compute the distance between the test sample and all of the training samples to determine its k-nearest neighbors. As we have already discussed when we have the imbalanced training data, to identify the samples from the smaller class or the

minority class is very difficult. Given a test sample, if we try to calculate k-nearest neighbors it is highly probable to get the result in favor of prevalent class only. Hence, test cases from the small classes are prone to being incorrectlyclassified[20].

III. CONCLUSION

We have discussed the problems because of the imbalanced data set and their challenges along with the appropriate solutions. In data mining community the class imbalance is very pervasive. It is very intrinsic in some applications such as fraud detection, medical diagnosis, and network intrusion detection, modern manufacturing plants, detection of oil spills from radar images of the ocean surface, text classification and direct marketing etc. Some of these applications, such as fraud detection, intrusion detection, medical diagnosis, etc. We come across the experimental studies for the multiple-class imbalanced data-sets with the aim of comparison among the different approaches for the achievement of comparative study. In that section we have highlighted the performance of different algorithm with respect to the binarization techniques with preprocessing of instances such as SMT, SL-SMT, and Global CS algorithm developed for multiple classes. This is going to be the another interesting research issue which is open for learning from imbalanced data set and classification of imbalanced data with multiple class labels. In classic pattern recognition problems there is need of mutually exclusive classes. Classification performance levels decreases when the classes overlap in the feature space. There is a complicated situation occur where the classes are not mutually exclusive. With the applications of these kinds of classes, the class imbalance problem is present. Combined with the multiple class label issue, the class imbalance problem assumes an even more complex situation. Due to the intriguing topics and tremendous potential applications, the classification of imbalanced data will continue to receive more and more attention in both the scientific and the industrial worlds. By this work we tried to provide the basis for the achievement of high quality solutions for imbalanced data-sets with multiple classes, but its significance lies also in the fact that it opens futuretrendsofresearch.

IV. REFERENCES

[1]. Krawczyk, Bartosz. "Learning from imbalanced data: open challenges and future directions." Progress in Artificial Intelligence 5, no.4(2016):221-232.

[2]. Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." International Journal of Pattern Recognition and Artificial Intelligence 23,no.04(2009):687-719.

[3]. Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." ACM Sigkdd ExplorationsNewsletter 6,no.1(2004):20-29.

[4]. FernaNdez, Alberto, Victoria LoPez, Mikel Galar, MaríA Jose Del Jesus, and Francisco Herrera. "Analysing the classification of imbalanced data-sets with multiple classes: Binarizationtechniquesandad-hoc approaches." Knowledge-based systems 42 (2013):97-110.

[5]. Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16(2002):321-357.

[6]. Barandela, Ricardo, Jose Salvador Sanchez, Vicente Garcıa, and Edgar Rangel. "Strategies for learning in class imbalance problems." Pattern Recognition 36, no. 3 (2003): 849-851.

[7]. Domingos, Pedro. "Metacost: A general method for making classifiers cost-sensitive." In Proceedings of the fifth ACM SIGKDD international conference on Knowledge

discovery and data mining, pp. 155-164. ACM, 1999.

[8]. Wozniak, Michal, Manuel Graña, and Emilio Corchado. "A survey of multiple classifier systems as hybrid systems." Information Fusion 16(2014):3-17.

[9]. Krawczyk, Bartosz. "Learning from imbalanced data: open challenges and future directions." Progress in Artificial Intelligence 5, no.4(2016):221-232.

[10]. Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem." Advances in knowledge discoveryanddatamining (2009):475-482.

[11]. Tax, David MJ, and Robert PW Duin. "Using two-class classifiers for multiclass classification." In Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 2, pp. 124-127. IEEE,2002.

[12]. Wang, Shuo, and Xin Yao. "Multiclass imbalance problems: Analysis and potential solutions." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)42, no. 4 (2012):1119-1130.

[13]. Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." International Journal of Pattern Recognition and Artificial Intelligence 23,no.04(2009):687-719.

[14]. Fernandez, Alberto, Mara Jose Del Jesus, and Francisco Herrera. "Multi-class imbalanced data-sets with linguistic fuzzy rule based classification systems based on pairwise learning." In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 89-98.Springer,Berlin,Heidelberg,2010.

[15]. FernaNdez, Alberto, Victoria LoPez, Mikel Galar, MaríA Jose Del Jesus, and Francisco Herrera. "Analysing the classification of imbalanced data-sets with multiple classes: Binarizationtechniquesandad-hoc approaches." Knowledge-based systems 42 (2013):97-110.

[16]. Hastie, Trevor, and Robert Tibshirani. "Classification by pairwise coupling." In Advances in neural information processing systems,pp.507-513.1998.

[17]. Cieslak, David A., T. Ryan Hoens, Nitesh V. Chawla, and W. Philip Kegelmeyer. "Hellinger distance decision trees are robust