International Journal of Computing and Business Research (IJCBR)

ISSN (Online) : 2229-6166

Volume 5 Issue 4 July 2014

IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW

Dr.D.Ramyachitra[#1], *P.Manikandan*[#2]

[#1] *Assistant Professor, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, India, Coimbatore-641 046*

[#2] *Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, India, Coimbatore-641 046*

Abstract--- Imbalanced data set problem occurs in classification, where the number of instances of one class is much lower than the instances of the other classes. The main challenge in imbalance problem is that the small classes are often more useful, but standard classifiers tend to be weighed down by the huge classes and ignore the tiny ones. In machine learning the imbalanced datasets has become a critical problem and also usually found in many applications such as detection of fraudulent calls, bio-medical, engineering, remote-sensing, computer society and manufacturing industries. In order to overcome the problems several approaches have been proposed. In this paper a study on Imbalanced dataset problem and the solution is given.

Keywords:- Imbalance Problems, Imbalanced datasets, Imbalance Techniques, Characteristics, Machine Learning.

1. INTRODUCTION

The class imbalance problem has received significant attention in areas such as Machine Learning and Pattern Recognition in recent years. A two-class data set is implicit to be imbalanced when one of the classes in the minority one is heavily under-represented in contrast to the other class in the majority one. This concern is mainly essential in real world applications where it is costly to misclassify examples from the minority class, such as detection of fraudulent telephone calls, diagnosis of rare diseases, information retrieval, text categorization and filtering tasks [8]. Several approaches have been earlier proposed to deal with this problem, which can be categorized into two groups: 1. Create innovative algorithms or change existing ones to take the class-imbalance problem into consideration is identified as the internal approaches and, 2. Pre-process the data in order to diminish the effect caused by their class imbalance is identified as External approaches.

The internal approaches have the disadvantage of being algorithm explicit, whereas external approaches are independent of the classifier used and are, more flexible. For this reason CO^2RBFN is applied to solving imbalanced classification problems [9]. In most applications, the exact classification of minority class is more important than majority class. For example, in predicting protein–protein interactions, the numbers of non-interacting proteins are greater than number of interacting proteins. Also in medical analysis problem, the numbers of disease cases are usually smaller than non-disease cases**.** [6].

The high activity of advancement in the imbalanced learning problem remains knowledgeable of all current developments and can be a difficult task. The ability of imbalanced data to significantly compromise the performance of most standard learning algorithm is the fundamental issue with the imbalanced learning problem. The imbalanced dataset problem occurs in different kinds of fields. In order to highlight the implications of the imbalanced learning problem, this paper presents some of the fields such as, medical diagnosis, text classification, detection of oil spill in radar images, information retrieval that had problems on imbalanced dataset that are represented in Fig 1.

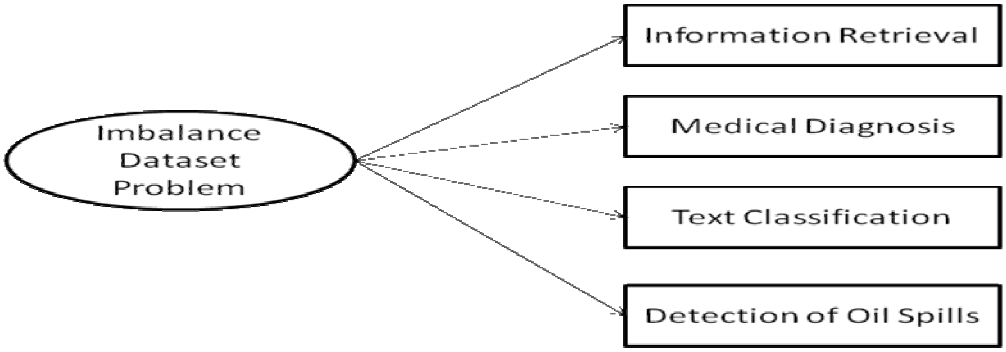

Fig 1: Example fields of Imbalanced dataset.

Information retrieval

IR probabilistic models from the perception of pattern classification and showed that they are generative in nature. The applicability of different classifiers such as SVMs and MEs are explored to IR. The striking theoretical properties and their major utility to IR lie in their ability to learn automatically a variety of features that influence relevance. The experiments on ad-hoc retrieval demonstrate that using the same type of features, SVMs perform as well as LMs on most of runs. The ability of SVMs in learning a variety of features in the home-page finding task, and outperform the baseline runs that utilize solitary content-based features by about 50% in MRR. [4]

Medical diagnosis

Another significant issue is that medical datasets used for machine learning should be representative of the general incidence of the studied disease. In the medical diagnostic area *REMED* competitive algorithm can be used.

However, *REMED* does not pretend to be the solution of machine learning in medical diagnostic, but a good approach with the preferred features to decipher medical analytical responsibilities, the comprehensibility of diagnostic knowledge, good performance, the capacity to enlighten decisions, and the capability of the algorithm to reduce the number of tests necessary to obtain reliable diagnosis [1].

Text classification

Sampling strategies such as *Oversampling* and *Subsampling* are trendy in tackling the problem of class imbalance. The three types of classifiers such as SVM, Knn and Naive-Bayes, are applied to search on the PubMed scientific database. In the classification of biomedical texts three types of dictionaries are used. The experiments are conducted with three different dictionaries such as NLPBA, BioCreative, and an ad-hoc subset of the UniProt database, using the mentioned classifiers and sampling strategies. Best consequences were obtained with NLPBA and Protein dictionaries and the SVM classifier using the *Subsampling* balancing method. These outcome were compared with those obtained using the TREC Genomics 2005 public corpus [3].

Detection of oil spills

Imbalanced dataset problem arises more frequently in applications and considerably reduces the performance of standard techniques. Numerous methods for coping with imbalanced classes have been proposed, but they are scattered. At the very least, a large scale comparative study is needed to assess the relative merits of these methods and how they work in permutation. Many creature methods, the SHRINK algorithm for example, can undoubtedly be improved by further research. It seems important to study small imbalanced training sets separately from large ones. Learning from batched examples is another issue which requires further research. Learning from batched examples is related to the issues of learning in the occurrence of circumstance, as the batches often characterize the unknown context in which the training examples were collected [2].

The remaining sections of this paper are organized as follows. Section 2 describes the imbalanced dataset characteristics, techniques, tools and databases. Section 3 describes the performance metrics. Finally Section 4 gives the conclusion.

2. IMBALANCED DATASET CHARACTERISTICS AND TECHNIQUES

2.1 IMBALANCED DATASET CHARACTERISTICS

Any data set that shows an unequal distribution between its classes can be considered imbalanced [5]. However, the common perceptive in the society is that imbalanced data communicate to data sets exhibiting significant, and in some cases extreme, imbalances. Particularly, this form of imbalance is referred to as a between-class imbalance; not uncommon are between class imbalances on the classification of 100:1, 1,000:1, and 10,000:1, where in every case, one class severely out represents another [22], [2], [23].

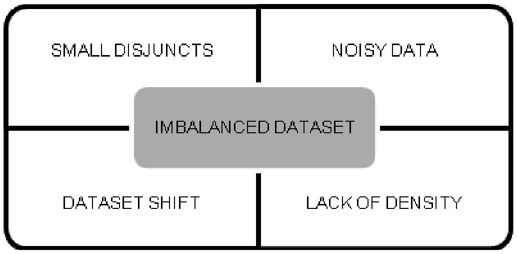

Fig 2: Imbalanced dataset characteristics

The problem related to using data intrinsic distinctiveness in this classification problem. This will facilitate to develop the current models with respect to: the lack of density in the training data, the presence of small disjuncts, the identification of noisy data, the dataset shift between the training and the test distributions and the significance of the borderline instances are represented in Fig 2 [11].

Small disjuncts

Presence of the imbalanced classes is closely related to the problem of small disjuncts. The small disjuncts problem occurs, when the concepts are represented within small clusters, which arise as a direct result of underrepresented sub concepts [13]. Small disjuncts problem becomes accentuated for those classification algorithms which are based on a divide-and conquer approach .This methodology consists in subdividing the original problem into smaller ones, such as the procedure used in decision trees, and can lead to data fragmentation , that is, to obtain several partitions of data with a few representations of instances[20].

Lack of density

One of the most problems that can arise in classification is the small sample size [17]. This issue is related to the ‘‘lack of density’’ or ‘‘lack of information’’, where induction algorithms do not have enough data to make generalizations about the distribution of samples, a situation that becomes more difficult in the presence of high dimensional and imbalanced data. Combination of the imbalanced data and the small sample size problem presents a new challenge to the research community [19].

Noisy data

The scenario of imbalanced data, the presence of noise has a greater impact on the minority classes than on usual cases [20]; since the positive class has fewer examples to begin with, it will take fewer ‘‘noisy’’ examples to impact the learned sub concept. In [14], the authors presented a similar study on the significance of noise and imbalance data using bagging and boosting techniques. Their results show the goodness of the bagging approach without replacement, and they recommend the use of noise reduction techniques prior to the application of boosting procedures.

Dataset shift

The dataset shift problem [12] is defined as the case where training and test data follow different distributions. It’s a familiar problem that can affect all kind of categorization problems, and it often appears due to sample selection bias issues Dataset shift issue is especially relevant when dealing with imbalanced categorization, because in highly imbalanced domains, the minority class is mostly sensitive to singular classification errors, due to the typically low number of examples it presents [15].

2.2 TECHNIQUES FOR IMBALANCED DATASET PROBLEMS:

Standard machine learning algorithms fail to classify imbalanced data because the classification error in majority class dominates the classification error in minority class. This domination results in pushing the separating function away from the majority class to decrease the classification error during weight adjusting process. Consequently, the testing data in minority class are misclassified added often than those in the majority class.

The techniques to handle imbalanced data problem can be categorized as data level, algorithmic level, cost-sensitive level, feature selection level, ensemble level and it can explained in the Fig 3.

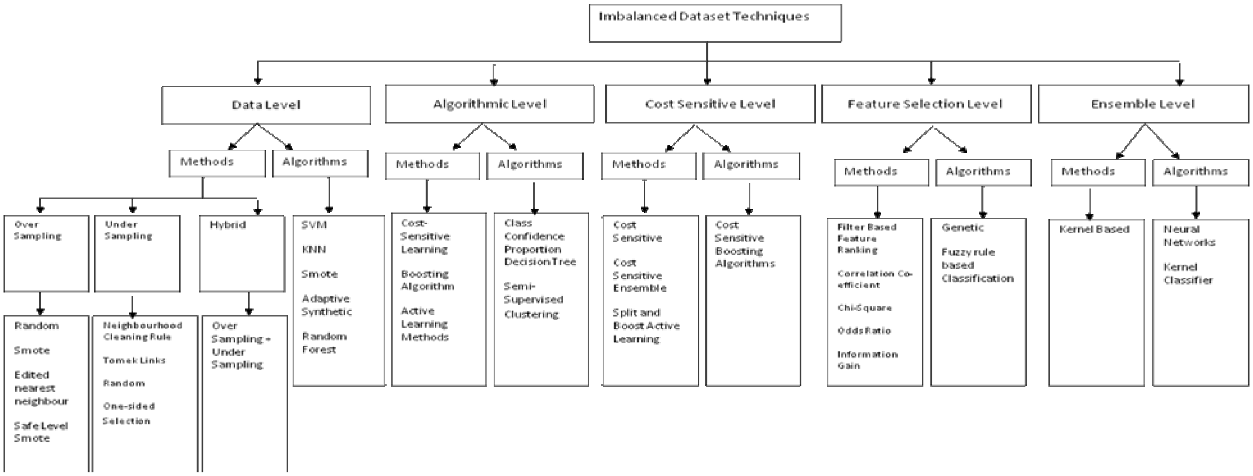

Fig 3: Classification of Techniques for solving Imbalanced Dataset.

Data level approaches

It works, in a pre-processing stage, directly on the data space, and tries to re-balance the class distributions. They are self-determining of the actual classification stage, and hence can be employed flexibly. The most admired approaches employ an oversampling strategy that introduces artificial objects into the data space. The best known technique here is SMOTE [24], although more recently, improved alternatives such as ADASYN [25] or RAMO [26] are exists. Oversampling methods however may also lead to other problems, such as class distribution shift when running too much iteration [27].

Methods:

- Sampling Methods

Jong Myong Choi proposes an iterative sampling methodology that was used to produce smaller learning sets by removing unnecessary instances. It integrates informative and the representative under-sampling mechanisms to speed up the learning procedure for imbalanced data learning with a SVM. For large-scale imbalanced datasets, sampling methodology provides a resourceful and effective solution for imbalanced data learning with an SVM [35].

- Adaptive sampling methods and synthetic data generation.

The intention is to provide a balanced distribution from over-sampling and/or under-sampling techniques to improve overall classification. In regards to synthetic sampling, the synthetic minority over-sampling technique (SMOTE) created synthetic data in minority class by selecting some of the nearest minority neighbors of a minority data and generating synthetic minority data along with the lines between the minority data and the nearest minority neighbors. Adaptive sampling methods were proposed to generate synthetic data. The idea of Borderline-SMOTE technique was to find out the borderline minority samples. Then, synthetic samples were generated along the line between the borderline samples and their nearest neighbors of the same class [43].

Algorithms:

- Support Vector Machine

TAO Xiao-yan, et al., 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 crisis. And more 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 [47].

Lili Diao, Chengzhong Yang, et al., propose an undersampling method to compress and balance the training set used for the conventional SVM classifier with minimal information thrashing. The key surveillance is that they can build a trade-off between training set size and information loss by carefully defining a similarity measure between data samples. Their experiments show that the SVM classifier provides an enhanced performance by applying the compressing and balancing approach [49].

- K- nearest neighbor

Wei Liu and Sanjay Chawla proposed a novel *k*-nearest neighbor (*K*NN) weighting strategy is proposed for handling the problem of class imbalance. They proposed CCW (class confidence weights) that uses the probability of attribute values given class labels to weight samples in *K*NN. The major benefit of CCW is that it is able to correct the inherent bias to majority class in accessible *k*NN algorithms on several distance dimensions. Theoretical study and complete experiments confirm their claims [41].

Yang Yong proposed sample sampling method based on the K-means cluster and the genetic algorithm. He used K-means algorithm to cluster and group the minority kind of sample, and in each cluster he use the genetic algorithm to gain the new sample and to carry on the valid authentication. At last, through using KNN and SVM sorter he proved the technique validity in the simulation experiment [51].

- SMOTE

Chawla, et al., proposed a Synthetic Minority Over-sampling Technique (SMOTE) [39] approach in which the minority class is over-sampled by creating synthetic examples rather than by over-sampling with alternation. The minority class is over-sampled by enchanting each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class adjacent neighbors. Depending upon the quantity of over-sampling required, the neighbors from the k nearest neighbors are randomly chosen.

- ADASYN

Haibo He, et al., presents a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The basic thought of ADASYN is to utilize a weighted distribution for different minority class examples according to their level of complexity in learning, where more synthetic data is created for minority class examples that are harder to learn compared to those minority examples that are easier to study. As a consequence, the ADASYN approach get better learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. The Simulation analyses on several machine learning data sets show the effectiveness of this method across five evaluation metrics [42].

- Random Forest

Dengju Yao, et al., proposed the random forest algorithm based on sampling with replacement. They extracted multiple example subsets randomly with replacement from majority class, and the example number of extracted example subsets is as the same with minority class example dataset. Then, the multiple new training datasets were constructed by combining the each exacted majority example subset and minority class dataset respectively, and multiple random forest classifiers were training on these training dataset. For a forecast example, the group was determined by majority voting of multiple random forest classifiers. The tentative outcome on five groups UCI datasets and a real clinical dataset show that these method could deal with the class-imbalanced data problem and the improved random forest algorithm outperformed original random forest and other methods in literatures [38].

Classifier level approaches

Trying to adapt existing algorithms to the problem of imbalanced datasets and bias them towards favoring the minority class known as Classifier level approaches. Here, some more in-depth knowledge about the nature of the used predictors and factors that cause its failure in minority class recognition is required. One possibility is to perform one-class classification, which can learn the concepts of the minority class by treating majority objects as outliers [29].

Methods:

- Cost-sensitive learning

In the Cost-sensitive learning method cost is associated with misclassifying patterns. The cost matrix is used for arithmetic illustration of the consequence of classifying examples from one class to another [33]. No consequence 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 goal of cost-sensitive learning method is to minimize the overall cost on the training dataset. Charles [52] presents 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 [53] recommended a more general method to make a learning system a cost-sensitive.

- Boosting Method

Boosting is a technique to progress the performance of weak classifiers. AdaBoost [54] is the most known boosting algorithm, and which is an ensemble erudition model. In each iteration, the weights are modified with the objective of correctly classifying examples in the subsequently iteration. At the end, all customized models contribute in a weighted vote to classify unlabeled examples. This method is more helpful to deal with class imbalance problem because minority class examples are mainly expected to be misclassified and hence given higher weights in subsequent iterations.

- Active learning methods

Traditional active learning methods were used to solve unbalanced training data. Recently, various approaches on active learning from imbalanced data sets were proposed. Active learning method based on SVM effectively selected the instances from a random set of training data, therefore significantly to reduce the computational cost when dealing with large imbalanced data sets [46].

Algorithms:

- Class Confidence Proportion Decision Tree (CCPDT)

Wei Liu, Sanjay Chawla, et al., propose a new decision tree algorithm, the Class Confidence Proportion Decision Tree (CCPDT), which is robust and insensitive to size of classes and generates rules which are statistically significant. To generate rules which are statistically significant they design a novel and efficient top-down and bottom-up approach which uses Fisher's exact test to prune branches of the tree which are not statistically considerable. Together these two changes defer a classifier that performs statistically better than not only traditional decision trees but also trees learned from data that has been balanced by well known sampling techniques. Their claims are confirmed through the extensive experiments and comparisons against C4.5, CART, HDDT and SPARCCC [34].

- Semi supervised Clustering

Mingwei Leng, et al., proposes an active semi supervised clustering algorithm that uses an energetic method for data selection to minimize the amount of labeled information, and it operates multithreshold to enlarge labeled datasets on multidensity and imbalanced datasets. In these three standard datasets and one synthetic dataset are used to demonstrate the algorithm, and the tentative outcomes show that the semi supervised clustering algorithm has a higher accuracy and a more stable performance in comparison to other clustering and semi supervised clustering algorithms, particularly when the datasets are multidensity and imbalanced [37].

Cost-sensitive approaches

It can use both data and modifications of the learning algorithms. A higher misclassification cost is assigned for minority class objects and classification performed so as to reduce the overall learning cost. Costs are often specified in form of cost matrices. The lack of knowledge on how to set the actual values in the cost matrix is the main drawback of cost-sensitive methods, since in most cases this is not known from the data nor given by an expert [48].

Methods:

- Cost-sensitive methods.

Cost-sensitive learning methods used the cost matrix to consider the costs associated with misclassifying samples. Cost-sensitive neural network with threshold-moving technique was proposed to adjust the output threshold toward inexpensive classes, such that high-cost samples are unlikely to be misclassified. Three cost-sensitive boosting methods, AdaC1, AdaC2, and AdaC3 were proposed and cost items were used to weight updating strategy in boosting algorithm [44].

- Cost sensitive Ensemble Method

Yong Zhang and Dapeng Wang propose a cost-sensitive ensemble method based on cost-sensitive support vector machine (SVM), and query-by-committee (QBC) to solve imbalanced data classification. In this method it first divides the majority-class dataset into several sub datasets according to the proportion of imbalanced samples and trains sub classifiers using AdaBoost method. Then, the method generates candidate training samples by QBC active learning method and uses cost-sensitive SVM to learn the training samples [30].

- Split and Boost active learning methods

Yong Zhang, et al., proposes a split and boost active learning method (Split Boost), based on cost-sensitive SVM, to solve imbalanced data classification. The Split Boost method first splits the majority class dataset into several sub-datasets according to the proportion of imbalanced samples, and instructs sub-classifiers with AdaBoost method. Then, the Split Boost generates candidate training samples by QBC active learning method and uses cost-sensitive SVM to learn the training samples. By using 6 class-imbalance datasets, the experimental results show that the method has top AUC, F-measure, and G-mean than various obtainable class imbalance learning methods [36].

Algorithms:

- Cost sensitive boosting algorithm

There are three ways to introduce cost items into the weight update formula of AdaBoost: inside the exponent, outside the exponent, and both inside and outside the exponent. Three modifications of Equation as

Modification I:

$$D^{t+1}(i) = \frac{D^{t}(i)\exp\left(-\alpha_t C_i h_t(x_i) y_i\right)}{Z_t} \tag{1}$$

Modification II:

$$D^{t+1}(i) = \frac{C_i D^{t}(i)\exp\left(-\alpha_t h_t(x_i) y_i\right)}{Z_t} \tag{2}$$

Modification III:

$$D^{t+1}(i) = \frac{C_i D^{t}(i)\exp\left(-\alpha_t C_i h_t(x_i) y_i\right)}{Z_t} \tag{3}$$

Each modification can be taken as a new boosting algorithm denoted as AdaC1, AdaC2 and AdaC3, respectively. As these algorithms use cost items, they can also be regarded as cost sensitive boosting algorithms. For the AdaBoost algorithm, the preference of the weight update parameter is crucial in converting a weak learning algorithm into a strong one [57].When the cost items are introduced into the weight updating formula of the

AdaBoost algorithm, the updated data distribution is affected by the cost items. Without re-inducing the weight update parameter, which takes the cost items into consideration for each cost-sensitive boosting algorithm, the boosting efficiency is not guaranteed.

Feature selection level approaches

The main idea of feature selection is to choose a subset of input features by eliminating features with little or no predictive information according to some measure. To adopt feature selection within the imbalanced problem, there are two approaches that exist. The initial one is based on adapting class-probability estimates. The next approach is based on the beginning of new feature selection measures [18].

Methods:

- Information gain

The Information gain [15] measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. The Information gain method is also known as Expected Mutual Information. The Expected Likelihood Estimation (ELE) smoothing technique was used to handle singularities when estimating those probabilities [56].

- Odds Ratio

The Odds ratio (OR) method measures the odds of the word occurring in the positive class normalized by that of the negative class. The basic thought is that the distribution of features on the relevant documents is different from the distribution of features on the nonrelevant documents.

Algorithms:

- Genetic Algorithms

Jong Myong Choi proposes a metaheuristic approach (Genetic Algorithm) for under sampling of an imbalanced dataset in the context of a SVM classifier. The objective of their approach is to locate an optimal learning set from imbalanced datasets without empirical studies that are normally required to find an optimal class distribution. Tentative outcome with real datasets indicate that this metaheuristic under-sampling performed fine in rebalancing class distributions. For large-scale imbalanced datasets, their method provides a capable and valuable solution for imbalanced data learning with an SVM [35].

Satyam Maheshwari et al., proposed Genetic Algorithms (GA) for over-sampling to enlarge the ratio of optimistic samples, and next apply clustering to the over-sampled training dataset as a data cleaning method for together classes, eliminating the unnecessary or noisy samples. This approach was experimentally analyzed and the experimental result shows an improvement in the classification measured as the area under the receiver operating characteristics (ROC) curve [33].

- Fuzzy rule based classification system

Alberto Fernandez, et al., propose the use of a hierarchical fuzzy rule based classification system, which is stand on the modification of a simple linguistic fuzzy model by means of the extension of the structure of the knowledge base in a hierarchical way and the use of a genetic rule selection process in order to get a compact and accurate model. The excellent performance of this approach is shown through an extensive experimental study carried out over a large collection of imbalanced data-sets [55].

Ensemble level approaches

Ensemble methods improve the performance of the overall system. The efficiency of ensemble methods is highly reliant on the independence of the error committed by the base learner. The performance of ensemble methods strongly depends on the accuracy and the diversity of the base learner. The easiest approach to generate diverse base classifier is by manipulating the training data [21].

Methods:

- Kernel-based methods.

In kernel-based methods, there have been many works to apply sampling and ensemble techniques to the support vector machine (SVM) concept. Different error costs were suggested for different classes to bias the SVM to shift the decision boundary away from positive instances and make positive instances more densely distributed. Other methods developed an ensemble system by modifying the data distribution and the SVM with asymmetric misclassification costs in order to boost the performance [45].

Algorithms:

- Neural Network

Adel Ghazikhani, et al., proposes an online ensemble of Neural Network (NN) classifiers. Ensemble models are the most frequent methods that are used for classifying non-stationary and imbalanced data streams. The main contribution is a two-layer approach for handling class imbalance and non-stationary. In the first layer, cost-sensitive learning is embedded into the training phase of the NNs, and in the second layer a new method for weighting classifiers of the ensemble is used. This method is evaluated on 3 synthetic and 8 real-world datasets and these results show statistically significant improvement compared to online ensemble methods with similar features [40].

Maciej A. Mazurowski, Piotr A. Habas, et al., investigates the effect of class imbalance in training data when developing neural network classifiers for computer-aided medical diagnosis. The investigation is performed in the medical data, specifically tiny training sample size, huge number of features, and correlations among features. Two ways of neural network training is explored: classical backpropagation (BP) and particle swarm optimization (PSO) with clinically relevant training criterion. A tentative learning is performed using simulated data and the conclusions are further validated on real clinical data for breast cancer diagnosis. Their results show that classifier performance deteriorates with even modest class imbalance in the training data. Further, they showed that BP is generally preferable over PSO for imbalanced training data especially with small data sample and large number of features [50].

- Kernel classifier identification

Xia Hong proposed a kernel classifier identification algorithm is based on a new regularized orthogonal weighted least squares (ROWLS) estimator and the model selection criterion of maximal leave-one-out area under curve (LOO-AUC) of the receiver operating characteristics (ROCs). It is clearly shown that, owing to the orthogonalization method, the LOO-AUC can be deliberated using an analytic formula based on the new regularized orthogonal weighted least squares parameter estimator, without actually dividing the evaluation data set. This algorithm can achieve minimal computational expense via a set of forward recursive updating formula in searching model terms with maximal incremental LOO-AUC value. Numerical models are used to exhibit the efficacy of the algorithm [32].

GENEREL STEPS FOR SOLVING IMBALANCED DATASET PROBLEM

The general steps that can be involved in solving the imbalanced dataset problem can be described in Fig 4. The Fig 4 clearly shows that the imbalanced problems occur in the case of minority or majority classes. In the minority class the data are oversampled for balanced the dataset. In case of majority class the data are under sampled for balanced the dataset. And finally we can find the balanced accuracy with/without threshold value for solving the imbalanced problem.

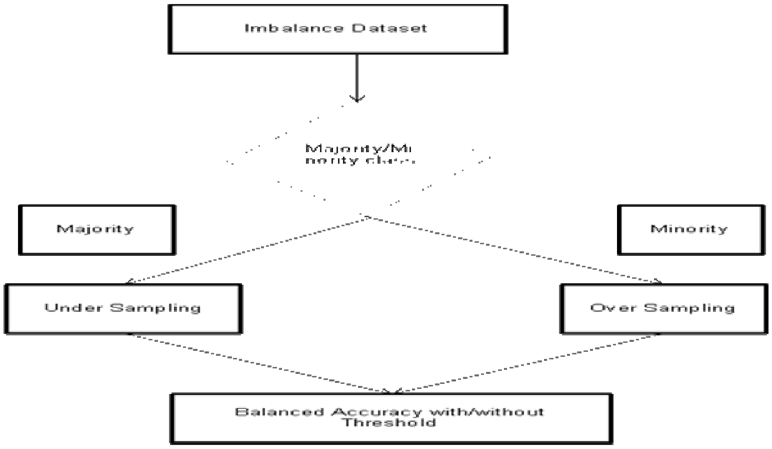

Fig 4: Steps to solve imbalance dataset problem.

2.4 DATABASES AND MACHINE LEARNING TOOLS

☞ Databases

➢ NCBI

The National Center for Biotechnology Information (NCBI) is a branch of the United States National Library of Medicine (NLM), and in addition it's a branch of the National Institutes of Health. The NCBI houses a series of databases related to biotechnology and biomedicine. The most important databases include GenBank for DNA sequences and the PubMed, a bibliographic database used for the biomedical literature. Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine.

➢ UNIPROT

The task of UniProt database is to meet the expense of the scientific community with a widespread, high-quality and freely accessible resource of protein sequence and functional information. The UniProtKB includes complete and reference proteome sets. The UniRef Sequence clusters can be used to speed up sequence similarity searches. The UniParc can be used for Sequence archive, and also used to keep track of sequences and their identifiers.

➢ UCI REPOSITORY

The UCI repository maintains 269 data sets as a service to the machine learning community. We may view all the data sets through the searchable interface. The previous web site is tranquil accessible, for those who desire the previous format.

➢ PUBMED

The PubMed database includes over 22 million citations for biomedical literature from MEDLINE, online books, and life science journals. The PubMed citations and abstracts include the fields of biomedicine and health, covering portions of the behavioral sciences, life sciences, chemical sciences, and bioengineering. It also provides access to additional relevant web sites and links to the other NCBI molecular biology resources. It's a free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI). The Publishers of journals can submit their citations to NCBI and then provide access to the full-text of articles at journal web sites using LinkOut.

☞ Tools

- Weka

The Weka is a collection of machine learning algorithms for data mining responsibilities. The algorithms can also be applied straightly to a dataset or called from our own Java code. And also Weka is open source software issued under the GNUs General Public License. The Weka contains tools for classification, data pre-processing, clustering, association rules, regression, and visualization.

- Rapid Miner

The Rapid Miner is a software platform developed by the company of the same name that provides an integrated environment used for predictive analytics, machine learning, business analytics, data mining and text mining,. The Rapid Miner is developed on a business source model which means the core and earlier versions of the software are available under an OSI-certified open source license.

- Keel

The KEEL is an open source (GPLv3) Java software tool to assess evolutionary algorithms for Data Mining problems as well as, classification, clustering, and pattern mining regression, and so on. It contains a big collection of classical knowledge removal algorithms, preprocessing methods, computational intelligence based erudition algorithms, as well as evolutionary rule learning algorithms based on different approaches and hybrid models such as evolutionary neural networks, genetic fuzzy systems and so on.

3. EVALUATION METRICS

The performance of classifiers in learning from imbalanced data can be evaluated using the four criteria's. They are (1) Minimum Cost criterion (MC), (2) the criterion of Maximum Geometry Mean (MGM) of the accuracy on the majority class and the minority class, (3) the criterion of the Maximum Sum (MS) of the accuracy on the majority class and the minority class, and (4) the criterion of Receiver Operating Characteristic (ROC) analysis [18]. The various kinds of performance metrics for imbalanced dataset can be shown in Fig 5.

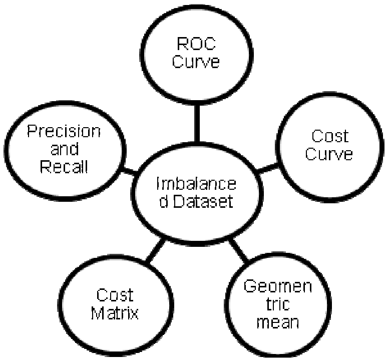

Fig 5: Performance metrics for imbalanced dataset.

In the confusion matrix Table 1, *TN* is the number of negative examples correctly classified (True Negatives), *FP* is the number of negative examples incorrectly classified as positive (False Positives), *FN* is the number of positive examples incorrectly classified as negative (False Negatives) and TP is the number of positive examples correctly classified (True Positives).

Actual/predicted	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Table 1. Confusion Matrix [21]

Accuracy = (TP+TN)/(TP+FN+FP+TN) (4)

FP rate = FP/ (TN+FP) (5)

TP rate = Recall = TP/ (TP+FN) (6)

Precision = TP/ (TP+FP) (7)

3.1 ROC Curves

Receiver Operating Characteristic (ROC) curve is a standard method for summarizing classifier performance over a range of tradeoffs between true positive and false positive error rates. The Area under the Curve (AUC) is an accepted performance metric for a ROC curve. ROC curves can be reflection of as representing the family of best decision boundaries for relative costs of **TP** and **FP.** On an ROC curve the,

X-axis represents %FP = FP/ (TN + FP) (8)

and the,

Y-axis represents %TP =TP/ (TP + FN) (9)

3.2 Precision and Recall

Commencing from the confusion matrix in Table 1, we can derive the expression for *precision* and *recall* as,

Precision = TP / (TP + FP) (10)

Recall = TP / (TP + FN) (11)

The main objective for learning from imbalanced datasets is to improve the *recall* without hurting the *precision. O*n the other hand, *recall* and *precision* goals can be often conflicting, since when increasing the true positive for the minority class, the number of false positives can also be increased; this will reduce the precision [16].

3.3 Geometric Mean

It's one of standard performance measures used in an imbalanced dataset classifier. The reason of using G-mean is to balance the ratio of prediction between majority and minority class. The proportion of G-mean shows that how good an imbalanced dataset classifier predicts the classes. The G-mean is deliberate as follows:

$$\text{G - Mean} = \sqrt{\mathrm{TNR} * \mathrm{TPR}} \qquad (12)$$

where,

$$\mathrm{TNR} = \mathrm{TN} / (\mathrm{TN} + \mathrm{FP}) \qquad (13)$$

$$\mathrm{TPR} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \qquad (14)$$

where,

TP, *FN*, *FP* and *TN* can be defined as follows. True Positive (*TP*) refers to correctly prediction of the majority class. False Negative (*FN*) refers to wrongly prediction of the minority class as majority class. False Positive (*FP*) refers to wrongly prediction of majority class as minority class. True Negative (*TN*) refers to correctly prediction of minority class.

3.4 Cost curve

Drummond and Holte introduced the Cost curve metrics, and they have also provided a detailed comparison between ROC curve and cost curve in Drummond and Holte (2004). Basically, cost curve looks at how classifiers perform across a range of different misclassification cost. It can be seen as different slope line tangent to the ROC curve, therefore every ROC curve has a corresponding cost curve [57].

3.5 Cost matrix

Cost matrix occasionally, the costs are known for the trouble at dispense, i.e. the misclassification cost of a positive or negative example. In this case, we can utilize the known cost to penalize the resulting confusion matrix to arrive at a significant performance assessment. A cost matrix looks the same as a confusion matrix, but it shows the cost of misclassification. The drawback of using confusion matrix-based evaluation is that they are only looking at the performance on a “spot”, which means we cannot tell how different class distribution or different cost will change the performance. So researchers may prefer to visually see the performance over a range of situations using one of the graphical evaluation tools, such as a ROC curve.

4. CONCLUSION

Thus this paper represents some of the characteristics of imbalanced dataset problem, techniques, algorithms, and also gives an idea of classification of the imbalanced dataset. Also, this paper has given a survey of the problems of the imbalanced dataset. In this paper we discuss several types of solutions and algorithms to the imbalance dataset. And also we had gone through the evaluation metrics of the imbalance dataset. And finally we conclude that, the solution for solving the imbalanced dataset problem is the data level process. Because, the data level process provides better results by using the oversampling algorithm for preprocessing and for balancing we use several algorithms that can be mentioned above. Thus this paper might be useful for the researchers to know about the imbalance dataset problems and also its solutions.

References

1) Luis Mena, Jesus A. Gonzalez, “Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic “, Proceedings of the 19th International FLAIRS Conference (FLAIRS-2006), Melbourne Beach, Florida, May 11-13, 2006.

2) Miroslav Kubat, Robert C. Holte, Stan Matw, “Machine Learning for the Detection of Oil Spills in Satellite Radar Images”, Kluwer Academic Publishers, Boston, Manufactured in the Netherlands.

3) L. Borrajo, R. Romero, E. L. Iglesias and C. M. Redondo Marey, “Improving imbalanced scientific text classification using sampling strategies and dictionaries”, Journal of Integrative Bioinformatics, 8(3):176, 2011.

4) Ramesh Nallapati,” Discriminative Models for Information Retrieval”, nmramesh@cs.umass.edu.

5) Haibo He, Member, IEEE, and Edwardo A. Garcia,” Learning from Imbalanced Data”, IEEE Transactions on Knowledge And Data Engineering, Vol. 21, No. 9, September 2009.

6) Putthiporn Thanathamathee , Chidchanok Lursinsap, “Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques”, Pattern Recognition Letters 34 (2013) 1339–1347.

7) Bartosz Krawczyk, Michał Wozniak, Gerald Schaefer, “Cost-sensitive decision tree ensembles for effective imbalanced classification”, Applied Soft Computing xxx (2013) xxx–xxx.

8) V. García, J.S. Sánchez, R.A. Mollineda, R. Alejo, J.M. Sotoca, “The class imbalance problem in pattern classification and learning”, Pattern Analysis and Learning Group, Dept.de Llenguatjes i Sistemes Informàtics, Universitat Jaume I.

9) María Dolores Pérez-Godoy , Alberto Fernández, Antonio Jesús Rivera, María José del Jesus, “Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets”, Pattern Recognition Letters 31 (2010) 2375–2388.

10) Putthiporn Thanathamathee , Chidchanok Lursinsap, ” Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques”, Pattern Recognition Letters 34 (2013) 1339–1347.

11) Victoria Lopez, Alberto Fernandez, Salvador Garcia, Vasile Palade, Francisco Herrera," An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics", Information Sciences 250, 2013, 113–141.

12) Alaiz-Rodriguez. R, Japkowicz. N, "Assessing the impact of changing environments on classifier performance", Proceedings of the 21st Canadian Conference on Advances in Artificial Intelligence (CCAI'08), Springer-Verlag, Berlin, Heidelberg, 2008, pp. 13–24.

13) Jo. T, Japkowicz. N, "Class imbalances versus small disjuncts", ACM SIGKDD Explorations Newsletter 6 (1), 2004, 40–49.

14) Khoshgoftaar T.M, Van Hulse. Napolitano J, A, "Comparing boosting and bagging techniques with noisy and imbalanced data", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 41 (3), 2011, 552–568.

15) Moreno-Torres J.G, Herrera. F, "A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction", Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA'10), 2010, pp. 501–506.

16) Nitesh V. Chawla," Data mining for imbalanced datasets: an overview", *IN 46530, USA.*

17) Raudys S.J, Jain. A.K., "Small sample size effects in statistical pattern recognition: recommendations for practitioners", IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (3), 1991, 252–264.

18) Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, "Handling imbalanced datasets: A review", GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006.

19) Wasikowski. M, Chen X.-W, "Combating the small sample class imbalance problem using feature selection", IEEE Transactions on Knowledge and Data Engineering 22 (10), 2010, 1388–1400.

20) Weiss. G.M, “Mining with rarity: a unifying framework”, SIGKDD Explorations 6 (1), 2004, 7–19.

21) Yongqing Zhang, Danling Zhang, Gang Mi, Daichuan Ma, Gongbing Li , Yanzhi Guo, Menglong Li, Min Zhu, “Using ensemble methods to deal with imbalanced data in predicting protein–protein interactions”, Computational Biology and Chemistry 36 , 2012, 36–41.

22) H. He and X. Shen, “A Ranked Subspace Learning Method for Gene Expression Data Classification,” Proc. Int’l Conf. Artificial Intelligence, pp. 358-364, 2007.

23) R. Pearson, G. Goney, and J. Shwaber, “Imbalanced Clustering for Microarray Time-Series,” Proc. Int’l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II, 2003.

24) N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16(2002) 321–357.

25) S. Chen, H. He, E.A. Garcia, Ramoboost: Ranked minority oversampling in boosting, IEEE Transactions on Neural Networks 21 (10) (2010) 1624–1642.

26) H. He, Y. Bai, E.A. Garcia, S. Li, Adasyn: adaptive synthetic sampling approach for imbalanced learning, in: Proceedings of the International Joint Conference on Neural Networks, 2008, pp. 1322–1328.

27) B. Krawczyk, G. Schaefer, M. Wozniak, Breast thermogram analysis using a cost-sensitive multiple classifier system, in: Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2012),2012, pp. 507–510.

28) Y. Sun, A.K.C. Wong, M.S. Kamel, Classification of imbalanced data: a review, International Journal of Pattern Recognition and Artificial Intelligence 23 (4)(2009) 687–719.

29) B. Krawczyk, M. Wozniak, Combining diverse one-class classifiers, in: E. Cor-chado, V. Snasel (Eds.), Ajith Abraham, Michal Wozniak, Manuel Grana, and Sung-Bae Cho, editors, Hybrid Artificial

Intelligent Systems, Volume 7209 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, 2012, pp.590–601.

30) Yong Zhang and Dapeng Wang," A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets", Hindawi Publishing Corporation ,Abstract and Applied Analysis Volume 2013, Article ID 196256, 6 pages ,http://dx.doi.org/10.1155/2013/196256.

31) Digna R. Velez, Bill C. White, Alison A. Motsinger, William S. Bush, Marylyn D. Ritchie, Scott M. Williams, and Jason H. Moore, "A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Datasets using Multifactor Dimensionality Reduction", Genetic Epidemiology 31: 306–315 (2007).

32) Xia Hong," A Kernel-Based Two-Class Classifier for Imbalanced Data Sets", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 18, NO. 1, JANUARY 2007.

33) Satyam Maheshwari, Prof. Jitendra Agrawal, Dr. Sanjeev Sharma, "A New approach for Classification of Highly Imbalanced Datasets using Evolutionary Algorithms", International Journal of Scientific & Engineering Research Volume 2, Issue 7, July-2011, ISSN 2229-5518.

34) Wei Liu, Sanjay Chawla, David A. Cieslak, Nitesh V. Chawla, "A Robust Decision Tree Algorithm for Imbalanced Data Sets", Copyright © by SIAM.

35) Jong Myong Choi," A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines", *Iowa State University*, cjm7331@gmail.com.

36) Yong Zhang, Yuting Zhang, Dapeng Wang, "A split and boost active learning method for class-imbalance data sets", School of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China.

37) Mingwei Leng, Jianjun Cheng, Jinjin Wang, Zhengquan Zhang,Hanhai Zhou, and Xiaoyun Chen, "Active Semi supervised Clustering Algorithm with Label Propagation for Imbalanced and Multidensity Datasets", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2013, Article ID 641927, 10 pages, http://dx.doi.org/10.1155/2013/641927.

38) Dengju Yao, Jing Yang, and Xiaojuan Zhan, “An Improved Random Forest Algorithm for Class-Imbalanced Data Classification and its Application in PAD Risk Factors Analysis”, *The Open Electrical & Electronic Engineering Journal,* 2013, *7,* (Supple 1: M7) 62-70.

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

40) Adel Ghazikhani, Reza Monsefi, Hadi Sadoghi Yazdi, “Ensemble of online neural networks for non-stationary and imbalanced data streams”, Neurocomputing122 (2013)535–544.

41) Wei Liu and Sanjay Chawla, “Class Confidence Weighted *k*NN Algorithms for Imbalanced Data Sets”, School of Information Technologies, University of Sydney.

42) Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning”, *2008 International Joint Conference on Neural Networks (IJCNN 2008).*

43) Estabrooks, A., Jo, T., Japkowicz, N., “A multiple resampling method for learning from imbalanced data sets”, Computational Intelligence 20, 18–36, 2004.

44) Sun. Y, Kamel. M, Wong. A, Wang.Y, “Cost-sensitive boosting for classification of imbalanced data”, Pattern Recognition 40, 3358–3378, 2007.

45) Akbani, R., Kwek, S., Japkowicz, N., 2004, “Applying support vector machines to imbalanced datasets. In: Proceedings of the 15th European Conference on Machine Learning (ECML)”, pp.39.39-50.

46) Ertekin, S., Huang, J., Bottou, L., Giles, C., 2007, “Learning on the border: active learning in imbalanced data classification. In: Proceedings of the sixteenth ACM conference on information and knowledge management, pp.127-136.

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

48) Elkan. C, "The Foundations of Cost-Sensitive Learning", In Proceedings of the Seventeenth International joint Conference on Artificial Intelligence, pp.73-978, 2001.

49) Lili Diao, Chengzhong Yang and Hao Wang, "Training SVM email classifiers using very large imbalanced dataset", Journal of Experimental & Theoretical Artificial Intelligence, Vol. 24, No. 2, June 2012, 193–210.

50) Maciej A. Mazurowski, Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker, Georgia D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", Neural Networks 21 (2008) 427–436.

51) Yang Yong, "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm", Energy Procedia 17 (2012) 164 – 170.

52) Charles Elkan, "The Foundations of Cost-Sensitive Learning," the Seventeenth International Joint Conference on Artificial Intelligence, pp. 973-978.

53) 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.

54) 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.

55) Alberto Fernández, María José del Jesus, Francisco Herrer, "Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets", International Journal of Approximate Reasoning 50 (2009) 561–577.

56) Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization", The Fourteenth International Conference on Machine Learning, pages, 412{420, 1997.

57) Chris Drummond · Robert C. Holte, “Cost curves: An improved method for visualizing classifier performance”, Mach Learn (2006) 65:95–130, DOI 10.1007/s10994-006-8199-5.