A Review on Imbalanced Learning Methods

Varsha S. Babar
M.E student, Department of C.E
Dr. D. Y. Patil School of
Engineering & Technology,
Savitribai Phule Pune Universi

Roshani Ade
Assistant Professor
Dr. D. Y. Patil School of
Engineering & Technology,
Savitribai Phule Pune University

ABSTRACT
Nowadays learning from imbalanced data sets are a relatively a very critical task for many data mining applications such as fraud detection, anomaly detection, medical diagnosis, information retrieval systems. The imbalanced learning problem is nothing but unequal distribution of data between the classes where one class contains more and more samples while another contains very little. Because of imbalance learning problems, it becomes hard for the classifier to learn the minority class samples. The Aim of this paper is to review on various techniques which are used for resolving imbalanced learning problem.This paper proposes a taxonomy for various methods used forhandling the class imbalance problem where each method can be categorized depending on the techniques it uses. To handle imbalanced learning problem significant work has been done, which can be categorized into four categories: sampling-based methods, cost-based methods, kernel-based methods, and active learning-based methods. All these methods resolve the imbalanced learning problem efficiently.

General Terms
Machine learning, imbalance data, Sampling

Keywords
Imbalanced learning, active learning, Cost-sensitive learning

1. INTRODUCTION
Due to the increased calculation power of recent computers and development of technology, the enormous amount of raw data related to the area of machine learning has been created. Consequently, there are various kinds of datasets which have their own data structure and distribution. For some datasets classifiers usually provide imbalanced determination due to their biased class distribution. That is, when dealing with imbalanced data sets, classifiers tend to consider minor examples as major ones.

The imbalanced learning problem is nothing but unequal distribution of data between the classes where one class contains more and more samples while another contains very little. The class which has most of the samples is called majority class and another is called minority class.

Because of imbalance learning problems, it becomes hard for the classifier to learn the minority class samples. In many cases classifier tends to favor majority class samples. This problem can be observed in various data mining applications such as fraud detection [1], medical diagnosis[2], information retrieval systems [3]. This paper gives a brief review of imbalanced learning and various methods used to solve imbalanced learning problem.

When there is an imbalance between classes, then it is called as between class imbalance. This imbalance may present either between binary (two class) classes or among many classes (multiclass). If the majority and minority class samples have more than one concept in which some concepts are rarer than others and regions between some concepts of different classes overlap then it is called within class imbalance. This form of imbalance arises due to the presence of small disjuncts which decreases the classifier performance very greatly. In addition to between class and within class imbalance, there are another two forms of imbalances namely, intrinsic and extrinsic. When imbalance occurs due to the nature of data space, then it is called as intrinsic imbalance. Instead of the nature of data space, imbalance is result of variable factors such as time and storage then it will be extrinsic imbalance.

In many traditional learning methods, all the required data sets should be available at training time. But in many real time application environments such as multimedia systems, sensor networks, web mining, data become available continuously over an infinite time. Therefore new methodologies, algorithms are needed to transform the data into useful information. In this kind of applications, incremental learning becoming very popular nowadays.[4-9].

2. METHODS OF IMBALANCED LEARNING
To handle imbalanced learning problem significant work has been done, which can be categorized into four categories: sampling-based methods, cost-based methods, kernel-based methods, and active learning-based methods. Figure 1 shows these categories. All these methods resolve the imbalanced learning problem efficiently.

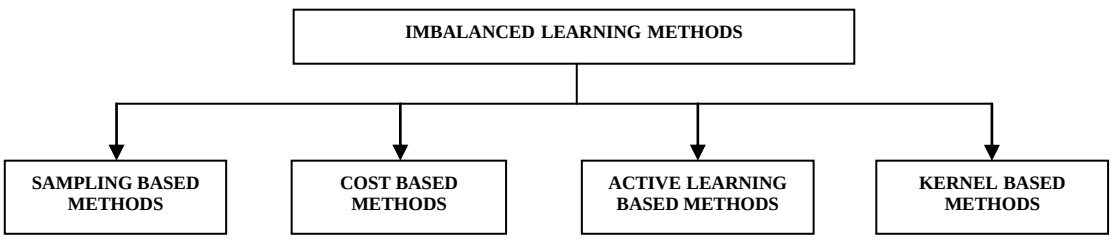

Figure 1: Methods in imbalanced learning

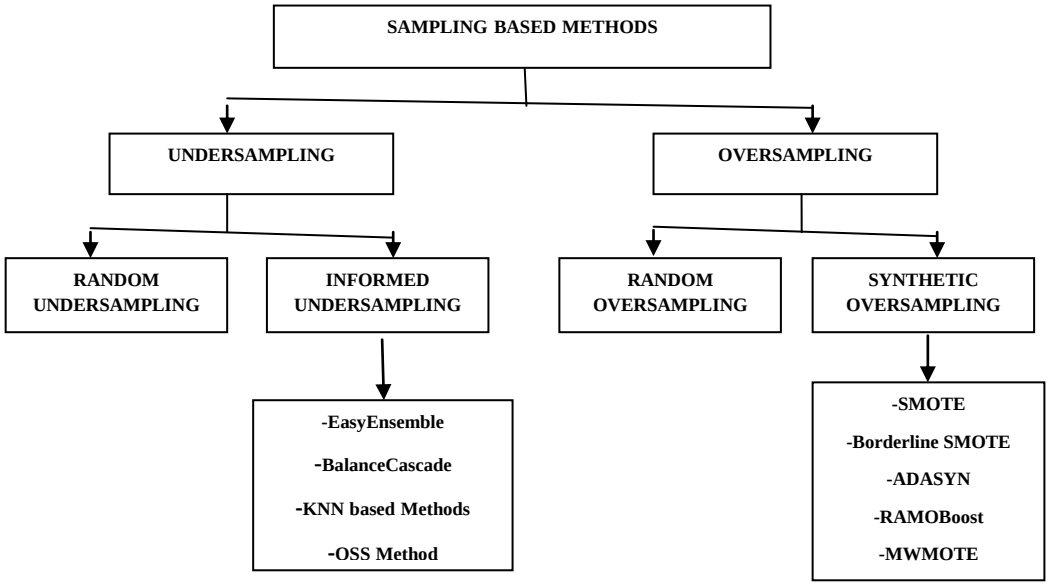

Figure 2: Taxonomy for Sampling based methods

2.1 Sampling based methods

In imbalanced learning sampling methods, the size of the classes is altered, i.e. it may increase the number of samples or it may reduce the samples. The method in which samples get reduced is called as under-sampling, while the method which increases the number of samples is called as oversampling [10].Figure 2 represents a hierarchy of sampling methods.

Under-sampling

Under-sampling methods reduces the number of samples from the majority class in order to balance between majority and minority classes. When majority class samples are reduced randomly it is called as random under-sampling. When samples are reduced on the basis of some statistical knowledge, then it is called as informed under-sampling. The main drawback of random under-sampling is that it may miss some important concepts from the majority class as it randomly removes the samples [11]. To overcome this problem many researchers propose various informed under sampling techniques as follows:

EasyEnsemble

In EasyEnsemble method, majority class is divided into several subsets and the size of each subset is equal to the size of a minority class. Then for each subset, it develops a classifier using whole minority class and majority class subset. Results generated from all the classifiers are combined to get the final decision. To develop a classifier Adaboost is used. EasyEnsemble approach has been shown in Figure 3. As EasyEnsemble uses independent random sampling with replacement, it can be considered as an unsupervised learning algorithm [12].

BalanceCascade

This method follows the supervised learning approach [12]. BalanceCascade method works as follows: Subset of majority class is formed which contains a number of samples equal to the number of minority class sample. When C_1 classifier is trained using the majority class subset and whole minority class, the samples from a majority subset which are correctly classified are removed. This new generated sampled set of majority class is given as an input toC_2. The same procedure is iterated until final classifier is reached. At every classifier, the size of the majority subset gets reduced. In BalanceCascade there is a sequential dependency between classifiers.BalanceCascade differ from EasyEnsemble as it removes true majority samples in order to reduce redundancy.

KNN based methods

To achieve under-sampling, a k-NN based approach has been proposed in order to deal with imbalance data distribution. This k-NN based approach include four different methods, namely NearMiss-1, NearMiss-2, NearMiss-3 and most distant method.[13] In NearMiss-1 method majority samples whose average distances to three closest minority samples are the smallest are selected for the under-sampling. The NearMiss-2 method selects majority samples that are close to all minority samples. This method selects the samples based on their average distances to three farthest minority samples.NearMiss-3 method guarantees every minority sample is surrounded by some majority samples. This method selects a given number of closest majority samples for each minority sample. In most distant method majority samples whose average distances to three closest minority samples are the farthest are selected for the under-sampling. On the basis of experimental results researchers suggested that NearMiss-2 methods performed well as compared to other methods.

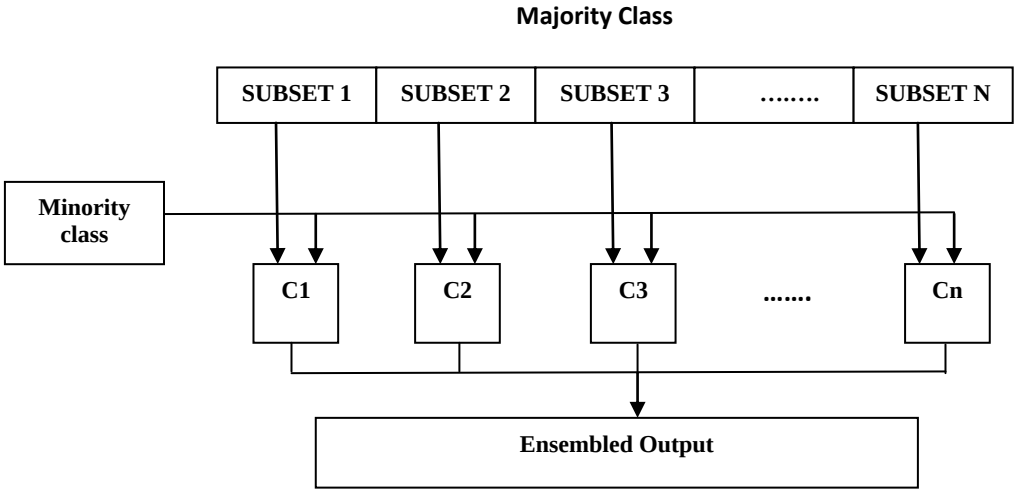

Figure 3: EasyEnsemble approach

One-sided Selection method
This is another type of informed under-sampling method in which only most representative majority samples are kept and remaining samples are removed from the class [14]. In order to choose the most representative samples, OSS first chooses one sample x randomly from majority class.Then taking minority samples and x as a training set OSS uses k-NN algorithm for classification of the remaining samples of majority class. Now, the correctly classified samples are removed from the majority class in order to remove redundancy. Therefore majority class will have only incorrectly classified samples and x in it. At the end OSS removes borderline and noisy samples using data cleaning techniques.

To start the under-sampling process OSS selects only one majority sample randomly. Hence the overall result will be depends on that sample. Also, OSS does not consider the existence of sub concepts within a majority class. To overcome this problem a novel method has been proposed namely ClusterOSS. In this method using clustering methods like k-means, clusters of majority classes are formed. Then samples located near the center of the cluster are selected for an under - sampling process.

There are mainly two differences between OSS and ClusterOSS. OSS uses only one majority sample to start under-sampling process while ClusterOSS uses more than one sample. Another difference is that OSS starts under-sampling randomly, but in a ClusterOSS number of samples and those samples are decided in advance [15]. Experimental resultssuggested that combination of clusterOSS with random under-sampling will give better results.

Oversampling
Oversampling methods add the samples to original imbalanced data set. There are two types of oversampling, random oversampling and synthetic oversampling. In random oversampling minority samples are randomly replicated, but this may lead to over fitting problem. In the synthetic oversampling method, synthetic samples are generated from minority samples. There are various oversampling methods existing in the literature. They are as follows:

SMOTE
In [16] proposed a powerful method, namely synthetic minority oversampling technique (SMOTE) which has been shown a great success in many applications. Initially for each minority sample k-nearest neighbors are determined. Then synthetic sample is generated along the line segment joining minority sample and its nearest neighbor. Firstly SMOTE takes the difference between minority sample and its nearest neighbor. This difference is then multiplied by a random number between 0 and 1, and adds this to original minority sample. In this way synthetic sample is generated. SMOTE generates an equal number of synthetic samples for each minority sample.

To handle the imbalanced learning problem in big data a novel approach, namely, the Enhanced SMOTE algorithm has been proposed. This algorithm works in two steps. In the first step original data set is decomposed into subsets of binary classes using Binarization techniques such as OVA and OVO. Then for each binary class SMOTE is applied. Random Forest is used to classify the data [17].

Borderline-SMOTE
As SMOTE generates synthetic samples for each minority sample it may lead to over generalization [11]. The main objective of Borderline-SMOTE is to identify minority samples located near decision boundary. Then these samples are used further for oversampling. This method focuses on borderline samples because classifier may misclassify them. In [18] two methods borderline-SMOTE1 and borderline-SMOTE2 has been proposed. Both methods give better results on TP rate and F-value as compared to SMOTE.

ADASYN
Haibo He, E.A. Garcia, proposed a novel approach adaptive synthetic sampling to handle imbalanced data set. In synthetic sample generation process, there is no need to consider all minority samples as there may be problem of overlapping [19]. ADASYN uses the weighted distribution of minority samples. It assigns weight to minority sample depending on importance of minority sample. Samples which are difficult to classify got higher weight than others. More samples are generated for the sample having a higher weight. ADASYN

can be integrated with ensemble based learning algorithm to get the good results.

RAMOBoost
Ranked Minority Oversampling in Boosting (RAMOBoost) is a technique which systematically generates synthetic samples depending on sampling weights. It adjusts these weights of minority samples according to their distribution. This method works in two stages. In first stage decision boundary is shifted towards the samples which are difficult to learn from both majority and minority classes. In the second stage to generate synthetic samples a ranked sampling probability distribution is used. If RAMOBoost adopts techniques used in SMOTE-N method, then it can handle datasets having nominal features [20].

MWMOTE
Existing synthetic oversampling methods may have some insufficiencies and inappropriateness in many scenarios [21]. In order to overcome these problems, a new method has been proposed, namely, Majority Weighted Minority Oversampling Technique (MWMOTE). This method works in three steps. In the first step the samples from the minority class which are difficult to learn and which contains more information are selected. In the second step selection weight is assigned to those selected samples. Most important samples got higher weight. In the last step, using selection weights this algorithm generates synthetic samples from selected minority samples. Many oversampling methods use k-NN based approach for sample generation process, but MWMOTE uses a clustering approach which gives better results than previous approaches. MWMOTE attempts to improve sample selection and sample generation process very efficiently. In future MWMOTE can be extended for the multiclass imbalances.

2.2 Cost Based Methods

As sampling based methods tries to remove or add the samples in order to balance between majority and minority classes, Cost based methods use a cost matrix in dealing with imbalanced learning. Cost matrix represents how much cost associated with each misclassification. If any minority sample gets misclassified in majority class, then its cost will be higher than misclassification of majority class sample [11].

When dealing with decision trees, cost sensitive methods can move the decision threshold, can apply pruning schemes based on cost-sensitivity or it can consider cost-sensitivity in the spilt criterion [10]. For building cost sensitive decision trees with missing values, a new splitting criterion has been proposed [22] which is based on tradeoffs between different costs units and the classification ability.

To introduce cost sensitivity in neural networks, output of neural network should make cost sensitive. The error minimization function should be adapted to get the expected cost. Based on cost sensitivity some modifications should be applied to probabilistic estimate [10]. To handle the multiclass imbalance problem, a new method based on ensemble of cost sensitive neural network has been proposed

[23]. To optimize the misclassification cost, this method uses evolutionary search technique.

2.3 Active learning based methods

In semi-supervised learning, there can be pool of data with labeled as well as unlabeled samples. To label the samples manually, it will be very expensive. To improve the classification accuracy, active learning methods focus on acquiring labels for those unlabeled data samples. The active learner chooses the unlabeled samples which are closer to decision boundary, and which are most uncertain. In the traditional approach, there exists a human annotator (oracle) which gives labels to unlabeled samples when the learner queries for labels. There are three main categories of this traditional approach such as pool based active learning [24], stream based active learning [25] and query construction based active learning [26]. Another strategy is proposed for remote sensing image classification [27] in which classifier assigns a rank to each unlabeled sample, and samples that are more important are selected. These selected samples are then labelled manually. A novel method is proposed to deal with noisy labels [28]. There are two procedures used in this method, label integration and sample selection. In label integration process, to get labels from multiple noisy labels a positive label threshold algorithm (PLAT) is introduced. Using the sample selection strategies, learning performance of PLAT can be improved.

2.4 Kernel Based methods

Along with sampling based methods and cost-sensitive methods, many researchers have worked on kernel based methods in order to deal with imbalanced data sets. In [29] a new oversampling strategy based on kernel function has been proposed to train a support vector machine (SVM). Firstly, it generates synthetic sample from minority samples similar to SMOTE, then pre-image of each synthetic sample is identified, and all these pre-images collectively append to the original minority set. To overcome the problem of generalization arises due to various oversampling methods a novel approach of quasi-linear SVM and assembled SMOTE has been proposed [30]. In this approach, using minimum spanning tree data is divided into a number of local linear partitions so that they are linearly separable. Then synthetic samples are generated using assembled SMOTE. Finally, using a quasi linear kernel function SVM classifies data efficiently.

3. DISCUSSIONS AND CONCLUSION

Currently, imbalanced learning becomes a challenging and active research topic in machine learning. This paper gives a brief review on various methods solving imbalanced learning problem. It also provides a brief description of all methods under consideration. To handle imbalanced learning problem, there are several research directions for oversampling techniques such as ADASYN, RAMOBoost and MWMOTE. All these techniques can be generalized to solve multiclass imbalanced learning problem.
Furthermore, these methods can also be modified to facilitate for incremental learning applications. All these oversampling methods can work efficiently on data sets with continuous features, but all these can be extended to handle data sets with nominal features by adopting various techniques used in SMOTE-N method. In ADASYN, RAMOBoost and MWMOTE the Euclidean distance is used as the distance measure, however, there are other alternativesthat are also eligible and worthy of trying and may show improved performance.

In MWMOTE clustering is the key step of sample generation, so finding the new clustering approach will also improve its performance. It can be integrated with other undersampling techniques to investigate whether they together can give better results than a single MWMOTE approach. Finally, this paper will serve as a comprehensive resource for the machine learning researchers and practitioners.

4. ACKNOWLEDGMENTS

The authors would like to thank the Department of Computer Engineering of D.Y.Patil School of Engineering and Technology, Pune for its generous support. They would also like to thank the principal Mr. Uttam B. Kalwane, HOD Mrs. Arti mohanpurkar and all staff members of Computer Engineering department for their valuable guidance.

5. REFERENCES

[1] T.E. Fawcett and F. Provost, "Adaptive Fraud Detection," Data Mining and Knowledge Discovery, vol. 3, no. 1, pp. 291-316, 1997.

[2] P.M. Murphy and D.W. Aha, "UCI Repository of Machine Learning Databases," Dept. of Information and Computer Science, Univ. of California, Irvine, CA, 1994.

[3] D. Lewis and J. Catlett, "Heterogeneous Uncertainty Sampling for Supervised Learning," Proc. Int'l Conf. Machine Learning, pp. 148- 156, 1994.

[4] R. Ade and P.R. Deshmukh, "An incremental ensemble of classifiers as a technique for prediction of student's career choice" Int'l conf. on Networks and soft computing(ICNSC), Aug 2014

[5] Shruti Patil and Roshani Ade, "Software Requirement Engineering Risk Prediction Model", Int'l journal of computer application, Sept 2014

[6] R Ade and P.R. Deshmukh, "Incremental learning in students classification system with efficient knowledge transformation" Int'l conf. on PDGC, Dec 2014

[7] R Ade and P.R. Deshmukh, "Efficient Knowledge Transformation System Using Pair of Classifiers for Prediction of Students Career Choice", Int'l Conf. on Information and communication technologies, Dec 2014

[8] R Ade and P.R. Deshmukh, "Efficient Knowledge Transformation for incremental learning and detection of new concept class in students classification system" Jan 2015

[9] R Ade and P.R. Deshmukh, "Classification of students by using an incremental ensemble of classifiers",Int'l Conf on ICRITO, Oct 2014

[10] H.He, Self-Adaptive Systems for Machine Intelligence,Wiley, Aug 2011

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

[12] X.Y. Liu, J. Wu, and Z.H. Zhou, "Exploratory Under Sampling for Class Imbalance Learning," Proc. Int'l Conf. Data Mining, pp. 965- 969, 2006.

[13] J. Zhang and I. Mani, "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction," Proc. Int'l Conf. Machine Learning (ICML '2003), Workshop Learning from Imbalanced Data Sets, 2003.

[14] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," Proc. Int'l Conf. Machine Learning, pp. 179-186, 1997.

[15] Victor H. Barella, Eduardo p. Costa, and Andre C P L F Carvalho, "ClusterOSS: a new undersampling method for imbalanced learning"

[16] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: Synthetic Minority oversampling Technique," J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[17] Reshma C. Bhagat and Sachin S. Patil, "Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data using Random Forest", IEEE International Advance Computing Conference (IACC), 2015

[18] H. Han, W.Y. Wang, and B.H. Mao, "Borderline-SMOTE: A New Oversampling Method in Imbalanced Data Sets Learning," Proc. Int'l Conf. Intelligent Computing, pp. 878-887, 2005.

[19] H. He, Y. Bai, E.A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," Proc. Int'l Joint Conf. Neural Networks, pp. 1322-1328, 2008.

[20] S. Chen, H. He, and E.A. Garcia, "RAMOBoost: Ranked Minority Oversampling in Boosting," IEEE Trans. Neural Networks, vol. 21, no. 20, pp. 1624-1642, Oct. 2010.

[21] Sukarna Barua, Md. Monirul Islam,Xin Yao, "MWMOTE-Majority Weighted MinorityOversampling Technique for imbalanced data set learning",IEEE Trans. Knowledge anddata engineering, vol. 26, no. 2, February 2014

[22] Xingyi LIU, "Cost-sensitive Decision Tree with Missing Values and Multiple Cost Scales", Int'l Joint Conf. on Artificial Intelligence, 2009

[23] Zhi-Hua Zhou and Xu-Ying Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem", IEEE Trans on knowledge and data engineering, vol. 18, no. 1, January 2006

[24] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in Proc. Conf. Empirical Methods NaturalLang. Process. (EMNLP*)*, Oct. 2008, pp. 1070–1079.

[25] S. Dasgupta, D. Hsu, and C. Monteleoni, "A general agnostic active learning algorithm," in Proc. Adv. Neural Inf. Process. Syst. (NIPS*)*, vol. 20. 2008, pp. 353–360.

[26] C. X. Ling and J. Du, "Active learning with direct query construction," in *Proc. 14th ACM* SIGKDD Int. Conf. Knowl. Discov. DataMining (KDD*)*, Las Vegas, NV, USA, 2008, pp. 480–487.

[27] D.Tuia, F. Ratle, F. Pacifici, M.F. Kanevski and W.J. Emery, "Active Learning Methods for Remote Sensing Image Classification", IEEE Trans. on Geoscience and Remote sensing, vol. 47, issue 7, April 2009

[28] Jing Zhang, Xindong Wu and Victor S. Sheng, "Active Learning with Imbalanced MultipleNoisy Labeling", IEEE Trans. on Cybernetics, vol. 45, no. 5, May 2015

[29] ZhiQiang ZENG and ShunZhi ZHU, "A Kernel-based Sampling to Train SVM with Imbalanced Data Set", Conference Anthology, IEEE, January 2013

[30] Bo ZHOU, Cheng YANG, Haixiang GUO and Jinglu HU, "A Quasi-linear SVM Combined with Assembled SMOTE for Imbalanced Data Classification", Int'l Joint Conf. on Neural Networks, August 2013