A REVIEW ON CLASS IMBALANCE LEARNING IN KNOWLEDGE DISCOVERY

[1]R. BuliBabu, [2]Dr. Mohammed Ali Hussain

[1] Research Scholar, Department of Computer Science, Bharathair University, Coimbatore, India.
[2] Professor, Department of Electronics and Computer Engineering, KLEF University, India

***Abstract*—Data Mining and Knowledge Discovery is an ever growing field from past few decades. The data sources used for knowledge discovery are rapidly changing due to tremendous developments in the field of science and technology. Class Imbalance learning and its shortcomings for the knowledge discovery with traditional algorithms are exposed over the last few years. In this paper, we provide a comprehensive review of current methods for constructing models for learning from class imbalanced data. Our focus is to provide a critical review of the nature of the problem, the state-of-the-art technologies, and the current assessment metrics used to evaluate learning performance under the imbalanced learning scenario.**

***Index Terms*— Classification, class imbalance, under-sampling, over-sampling, Class Imbalance Learning (CIL)**

1. Introduction

In Machine Learning community, and in Data Mining works, Classification has its own importance. Classification is an important part and the research application field in the data mining [1]. With ever-growing volumes of operational data, many organizations have started to apply data-mining techniques to mine their data for novel, valuable information that can be used to support their decision making [2]. Organizations make extensive use of data mining techniques in order to define meaningful and predictable relationships between objects [3]. Decision tree learning is one of the most widely used and practical methods for inductive inference [4]. This paper presents an updated survey of various decision tree algorithms in machine learning. It also describes the applicability of the decision tree algorithm on real-world data.

The rest of this paper is organized as follows. In Section 2, we presented the basics of data mining and classification. In Section 3, we present the imbalanced data-sets problem, and In Section 4 we present the various data balancing techniques used for class imbalanced learning. In Section 5 we present the various evaluation criteria's used for class imbalanced learning. In Section 6, we presented updated survive of class imbalance learning methods. Finally, in Section 7,we make our concluding remarks.

2. Data Mining

Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the owner [5]. There are many different data mining functionalities. A brief definition of each of these functionalities is now presented. The definitions are directly collated from [6]. Data characterization is the summarization of the general characteristics or features of a target class of data. Data Discrimination, on the other hand, is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. Association analysis is the discovery of association rules showing attribute value conditions that occur frequently together in a given set of data.

Classification is an important application area for data mining. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model can be represented in various forms, such as classification rules, decision trees, mathematical formulae, or neural networks. Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.

Outlier Analysis attempts to find outliers or anomalies in data. A detailed discussion of these various functionalities can be found in [6]. Even an overview of the representative algorithms developed for knowledge discovery is beyond the scope of this paper. The interested person is directed to the many books which amply cover this in detail [5], [6].

The Classification Task

Learning how to classify objects to one of a pre-specified set of categories or classes is a characteristic of intelligence that has been of keen interest to researchers in psychology and computer science. Identifying the common ―core characteristics of a set of objects that are representative of their class is of enormous use in focusing the attention of a person or computer program. For example, to determine whether an animal is a zebra, people know to look for stripes rather than examine its tail or ears. Thus, stripes figure strongly in our *concept* (generalization) of zebras. Of course stripes alone are not sufficient to form a class

description for zebras as tigers have them also, but they are certainly one of the important characteristics. The ability to perform classification and to be able to *learn* to classify gives people and computer programs the power to make decisions. The efficacy of these decisions is affected by performance on the classification task.

In machine learning, the classification task described above is commonly referred to as *supervised learning*. In supervised learning there is a specified set of classes, and example objects are labeled with the appropriate class (using the example above, the program is told what a zebra is and what is not). The goal is to generalize (form class descriptions) from the training objects that will enable novel objects to be identified as belonging to one of the classes. In contrast to supervise learning is *unsupervised learning*. In this case the program is not told which objects are zebras. Often the goal in unsupervised learning is to decide which objects should be grouped together—in other words, the learner forms the classes itself. Of course, the success of classification learning is heavily dependent on the quality of the data provided for training—a learner has only the input to learn from. If the data is inadequate or irrelevant then the concept descriptions will reflect this and misclassification will result when they are applied to new data. The popular approach of classification examples are C4.5 [7], CART [8] and REP [9].

3. Problem of Imbalanced Datasets

A dataset is class imbalanced if the classification categories are not approximately equally represented. The level of imbalance (ratio of size of the majority class to minority class) can be as huge as 1:99[10]. It is noteworthy that class imbalance is emerging as an important issue in designing classifiers [11], [12], [13]. Furthermore, the class with the lowest number of instances is usually the class of interest from the point of view of the learning task [14]. This problem is of great interest because it turns up in many real-world classification problems, such as remote-sensing [15], pollution detection [16], risk management [17], fraud detection [18], and especially medical diagnosis [19]–[22].
There exist techniques to develop better performing classifiers with imbalanced datasets, which are generally called Class Imbalance Learning (CIL) methods. These methods can be broadly divided into two categories, namely, external methods and internal methods. External methods involve preprocessing of training datasets in order to make them balanced, while internal methods deal with modifications of the learning algorithms in order to reduce their sensitiveness to class imbalance [23]. The main advantage of external methods as previously pointed out, is that they are independent of the underlying classifier.

4. Data Balancing Techniques

Whenever a class in a classification task is underrepresented (i.e., has a lower prior probability) compared to other classes, we consider the data as imbalanced [24], [25]. The main problem in imbalanced data is that the majority classes that are represented by large numbers of patterns rule the classifier decision boundaries at the expense of the minority classes that are represented by small numbers of patterns. This leads to high and low accuracies in classifying the majority and minority classes, respectively, which do not necessarily reflect the true difficulty in classifying these classes. Most common solutions to this problem balance the number of patterns in the minority or majority classes.

Either way, balancing the data has been found to alleviate the problem of imbalanced data and enhance accuracy [24],[25], [26]. Data balancing is performed by, e.g., oversampling patterns of minority classes either randomly or from areas close to the decision boundaries. Interestingly, random oversampling is found comparable to more sophisticated oversampling methods [26]. Alternatively, under-sampling is performed on majority classes either randomly or from areas far away from the decision boundaries. We note that random under-sampling may remove significant patterns and random oversampling may lead to over-fitting, so random sampling should be performed with care. We also note that, usually, oversampling of minority classes is more accurate than under-sampling of majority classes [26].Synthetic minority oversampling technique (SMOTE) [29] is an oversampling method, where new synthetic examples are generated in the neighborhood of the existing minority-class examples rather than directly duplicating them. In addition, several informed sampling methods have been introduced in [30]. The bottom line is that when studying problems with imbalanced data, using the classifiers produced by standard machine learning algorithms without adjusting the output threshold may well be a critical mistake. This skewness towards minority class (positive) generally causes the generation of a high number of false-negative predictions, which lower the model's performance on the positive class compared with the performance on the negative (majority) class. A comprehensive review of different CIL methods can be found in [27-28]. The following two sections briefly discuss the external-imbalance and internal-imbalance learning methods.

5. Evaluation Criteria's for Class Imbalance Learning

5.1. Evaluation Criteria

To assess the classification results we count the number of true positive (TP),true negative (TN), false positive (FP) (actually negative, but classified as positive) and false negative (FN) (actually positive, but classified as negative) examples. It is now well known that error rate is not an appropriate evaluation criterion when there is class imbalance or unequal costs. In this paper, we use AUC, Precision, F-measure, TP Rate and TN Rate as performance evaluation measures.

Let us define a few well known and widely used measures for C4.5[7] as the baseline classifier with the most popular machine learning publicly available datasets at Irvine [31]. Apart from these simple metrics, it is possible to encounter several more complex evaluation measures that have been used in different practical domains. One of the most popular techniques for the evaluation of classifiers in imbalanced problems is the Receiver Operating Characteristic (ROC) curve, which is a tool for visualizing, organizing and selecting classifiers based on their tradeoffs between benefits (true positives) and costs (false positives).

The most commonly used empirical measure, accuracy does not distinguish between the number of correct labels of different classes, which in the framework of imbalanced problems may lead to erroneous conclusions. For example a classifier that obtains an
accuracy of 90% in a dataset with a degree of imbalance 9:1, might not be accurate if it does not cover correctly any minority class instance.

$$ACC = \frac{TP + TN}{TP + FN + FP + FN}$$

Because of this, instead of using accuracy, more correct metrics are considered. A quantitative representation of a ROC curve is the area under it, which is known as AUC. When only one run is available from a classifier, the AUC can be computed as the arithmetic mean (macro-average) of TP rate and TN rate:

The Area under Curve (AUC) measure is computed by,

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2}$$

Or

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2}$$

On the other hand, in several problems we are especially interested in obtaining high performance on only one class. For example, in the diagnosis of a rare disease, one of the most important things is to know how reliable a positive diagnosis is. For such problems, the precision (or purity) metric is often adopted, which
can be defined as the percentage of examples that are correctly labeled as positive:

The Precision measure is computed by,

$$\Pr ecision = \frac{TP}{(TP) + (FP)}$$

The F-measure Value is computed by,

$$F - measure = \frac{2 \times \Pr ecision \times \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call}$$

To deal with class imbalance, sensitivity (or recall) and specificity have usually been adopted to monitor the classification performance on each class separately. Note that sensitivity (also called true positive rate, TP rate) is the percentage of positive examples that are correctly classified, while specificity (also referred to as true negative rate, TN rate) is defined as the proportion of negative examples that are correctly classified:

The True Positive Rate measure is computed by,

$$TruePositiveRate = \frac{TP}{(TP) + (FN)}$$

The True Negative Rate measure is computed by,

$$TrueNegativeRate = \frac{TN}{(TN) + (FP)}$$

5.2. Benchmark datasets used in Class imbalance Learning

Table 1 summarizes the benchmark datasets used in almost all the recent studies conducted on class imbalance learning. The details of the datasets are given in table 1. For each data set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority) and IR is given. This table is ordered by the IR, from low to high imbalanced data sets.

TABLE I
SUMMARY OF BENCHMARK IMBALANCED DATASETS

Datasets	# Ex.	# Atts.	Class (_,+)	IR
Glass1	214	9	(build-win-non_float-proc;remainder)	1.82
Ecoli0vs1	2207		(im;cp)	1.86
Wisconsin	683	9	(malignant;benign)	1.86
Pima	768	8	(tested-positive; tested-negative)	1.90
Iris0	150	4	(Iris-Setosa;remainder)	2.00
Glass2	214	9	(build-win-float-proc;remainder)	2.06
Yeast1	1484	8	(nuc;remainder)	2.46
Vehicle1	846	18	(Saab;remainder)	2.52
Vehicle2	846	18	(Bus;remainder)	2.52
Vehicle3	846	18	(Opel;remainder)	2.52
Haberman	306	3	(Die; Survive)	2.68
Glass3	214	9	(non-window glass;remainder)	3.19
Vehicle0	846	18	(Van;remainder)	3.23
Ecoli1	336	7	(im;remainder)	3.36
Thyroid2	2155		(hypo;remainder)	4.92
Thyroid1	2155		(hyper;remainder)	5.14
Ecoli2	336	7	(pp;remainder)	5.46
Segment0	230819		(brickface;remainder)	6.01
Glass6	214	9	(headlamps;remainder)	6.38
Yeast3	1484	8	(me3;remainder)	8.11
Ecoli3	336	7	(imU;remainder)	8.19
Page-blk15	472	10	(remainder;text)	8.77
Ecoli4	200	7	(p,imL,imU;om)	9.00
Yeast2	5148		(cyt;me2)	9.08
Ecoli05	222	7	(cp,omL,pp;imL,om)	9.09
Ecoli06	202	7	(cp,imS,imL,imU;om)	9.10
Glass4	172	9	(build-win-non_float-proc, tableware,build-win-float-proc; ve-win-float-proc)	9.12
Yeast4	506	8	(mit,me1,me3,erl; vac,pox)	9.12
Yeast5	1004	8	(mit, cyt,me3,vac,erl;me1,exc,pox)	9.14
Yeast6	1004	8	(mit, cyt,me3,exc;me1,vac,pox, erl)	9.14
Ecoli7	203	6	(cp,imU,omL;om)	9.15
Ecoli8	244	7	(cp,im;imS,imL,om)	9.17
Ecoli9	224	7	(cp,imS,omL,pp;imL,om)	9.18
Glass5	929		(build-win-float-proc,containers; tableware)	9.22
Ecoli10	205	7	(cp,imL,imU,omL;om)	9.25
Ecoli11	257	7	(cp,imL,imU,pp;om,omL)	9.28
Yeast7	528	8	(me2;mit,me3,exc, vac, erl)	9.35
Ecoli12	220	6	(cp,omL,pp;om)	10.00
Vowel	988	13	(hid;remainder)	10.10
Glass6192	9		(ve-win-float-proc; build-win-float-proc, build-win-non_float-proc,headlamps)	10.29
Glass7	2149		(Ve-win-float-proc;remainder)	10.39
Ecoli13	336	7	(cp,im,imU,pp;imS,imL,om,omL)	10.59
Led7digit	4437		(0, 2, 4, 5, 6, 7, 8, 9;1)	10.97
Glass7108	9		(build-win-float-proc,headlamps;	11.00

tableware)
Ecoli14240 6 (cp,im;om) 11.00
Glass8205 9(build-win-float-proc,11.06
containers,headlamps,
build-win-non_float-proc;
ve-win-float-proc)
Ecoli15332 6 (cp,im,imU,pp;om,omL) 12.28
Cleveland 177 13(0; 4) 12.62
Ecoli16280 6 (cp,im,imU,omL;om) 13.00
Ecoli17336 7 (om;remainder) 13.84
Yeast8459 8 (nuc; vac) 13.87
Shuttle1 1829 9 (Rad Flow;Bypass) 13.87
Glass4 214 9 (containers;remainder) 15.47
Page-blk47210 (graphic; horiz.line,picture) 15.85
Abalone731 8 (18; 9) 16.68
Glass9184 9 (tableware; build-win-float-proc,19.44
build-win-non_float-proc,
headlamps)
Shuttle2129 9 (FpvOpen;Bypass)20.5
Yeast9693 8 (vac; nuc,me2,me3,pox) 22.10
Glass10214 9 (tableware;remainder) 22.81
Yeast10 482 8 (pox;cyt) 23.10
Yeast111484 8 (me2;remainder) 28.41
Yeast12 947 8 (vac; nuc,cyt,pox,erl) 30.56
Yeast13 14848 (me1;remainder) 32.78
Ecoli18281 7 (pp,imL;cp,im,imU,imS) 39.15
Yeast14 14848 (exc;remainder) 39.15
Abalone19 4174 8(19;remainder) 128.87

The imbalance ratio (IR)is obtained by dividing the number of positive samples over the number of negative samples. A dataset is termed balance if the imbalance ratio is one. The complete details regarding all the datasets can be obtained from UCI Machine Learning Repository [31].

6. Recent Advances on Class Imbalance Learning

Currently, the trends of research in class imbalance learning are presented in this section. The recent research directions for class imbalance learning are as follows:

Dariusz Brzezinski *et al.* [32] have compare several techniques that can be used in the analysis of imbalanced credit scoring data sets. They progressively increase class imbalance in each of these data sets by randomly under-sampling the minority class of defaulters, so as to identify to what extent the predictive power of the techniques is adversely affected. Ana C. Lorenaet al. [33] have investigated the use of different supervised machine learning techniques to model the potential class imbalance distribution of 35 plants pieces from Latin America.

Victoria Lópezet al. [34] have proposed an evolutionary framework, which uses an Iterative Instance Adjustment for Imbalanced Domains. Their method iteratively learns the appropriate number of examples that represent the classes and their particular positioning. Their learning process contains three key operations in its design: a customized initialization procedure, an evolutionary optimization of the positioning of the examples and a selection of the most representative examples for each class. NeleVerbiestet al. [35] have proposed an improved SMOTE in the presence of class noise. Their approach cleans the data before applying SMOTE such that the quality of the generated instances is better and cleans the data after applying SMOTE, such that instances (original or introduced by SMOTE) that badly fit in the new dataset are also removed.

Peng Caoet al. [36] have proposed an effective wrapper approach incorporating the evaluation measure directly into the objective function of cost-sensitive neural network to improve the performance of classification, by simultaneously optimizing (Particle Swarm Optimization) the best pair of feature subset, intrinsic structure parameters and misclassification costs. Yetian Chen [37] has reported two classification tasks based on data from scientific experiment. The first task is a binary classification task which is to maximize accuracy of classification on an evenly-distributed test data set, given a fully labeled imbalanced training data set. The second task is also a binary classification task, but to maximize the F1-score of classification on a test data set, given a partially labeled training set.

Doucette et al. [38] have proposed a 'Simple Active Learning Heuristic' (SALH) in which a subset of exemplars is sampled with uniform probability under a class balance enforcing rule for fitness evaluation. Aditya Krishna Menon et al. [39] have study consistency with respect to one such performance measure, namely the arithmetic mean of the true positive and true negative rates(AM), and establish that some practically popular approaches, such as applying an empirically determined threshold to a suitable class probability estimate or performing an empirically balanced form of risk minimization, are in fact consistent with respect to the AM (under mild conditions on the underlying distribution).

Shuo Wang et al. [40-41] have defined class imbalance online, and proposed two learning algorithms OOB and
UOB that build an ensemble model overcoming class imbalance in real time through re sampling and time decayed metrics. In their further work improved the re sampling strategy inside OOB and UOB, and look into their performance in both static and dynamic data streams. They find that UOB is better at recognizing minority-class examples in static data streams, and OOB is more robust against dynamic changes in class imbalance status. In their further work they proposed a multi-objective ensemble method MOSOB that combines OOB and UOB. MOSOB finds the Pareto-optimal weights for OOB and UOB at each time step, to maximize minority-class recall and majority-class recall simultaneously. They concluded that MOSOB performs well in both static and dynamic data streams.

Bing Yang et al. [42] have given a close attention to the uniqueness of uneven data distribution in imbalance classification problems. Without change the original imbalance training data, they indicated the advantages of proximal classifier for imbalance data classification. In order to improve the accuracy of classification, they proposed a new model named LSNPPC, based the classical proximal SVM models which find two nonparallel planes for data classification. M'hamed B. Abidine et al. [43] have proposed a new version of the multi-class Weighted Support Vector Machines(WSVM) method to perform automatic recognition of activities in a smart home environment. WSVM is capable of solving the class imbalance problem by improving the class accuracy of activity classification compared to other methods like CRF, *k*-NN and SVM.
Hala S. Own et al. [44] have proposed a novel weighted rough set as a Meta classifier framework for 14classifiers to find the smallest and optimal ensemble, which maximize the overall ensemble accuracy. They also proposed a new entropy-based method to compute the weight of each classifier. Each classifier assigns a weight based on its contribution to classification accuracy. The powerful reduction technique in rough set guarantees high diversity of the produced reduct ensembles. The higher diversity between the core classifiers has a positive impact on the performance of minority class as well as on the overall system performance.

Bao-Gang Huet al. [45] have investigated on twelve performance measures, such as F measure, G-means in terms of accuracy rates, and of recall and precision, balance error rate (BER), Matthews correlation coefficient (MCC), Kappa coefficient, etc. A new perspective is presented for those measures by revealing their cost functions with respect to the class imbalance ratio. Basically, they are described by four types of cost functions. The functions provide a theoretical understanding why some measures are suitable for dealing with class-imbalanced problems. Based on their cost functions, they are able to conclude that G-means of accuracy rates and BER are suitable measures because they show "proper" cost behaviors in terms of "a misclassification from a small class will cause a greater cost than that from a large class".

Nicola Lunardonet al. [46] have decsried ROSE, a package which provides functions to deal with binary classification problems in the presence of imbalanced classes. Artificial balanced samples are generated according to a smooth edbootstrap approach and allow for aiding both the phases of estimation and accuracy evaluation of a binary classifier in the presence of a rare class. Xiaowan Zhang et al. [47] have proposed a novel cost-free learning approach which seeks to maximize normalized mutual information of the targets and the decision outputs of classifiers. Using the strategy, they can handle binary/multi-class classifications with/without abstaining. While the degree of class imbalance is changing, the proposed strategy is able to balance the errors and rejects accordingly and automatically. Another advantage of their strategy is its ability of deriving optimal rejection thresholds for abstaining classifications and the "equivalent" costs in binary classifications. They also explored the connection between rejection thresholds and ROC curve.

Andrea Dal Pozzolo et al. [48] have demonstrated regarding how Hellinger Distance Decision Trees (HDDT)can be successfully applied in unbalanced and evolving stream data. Using HDDT allows us to remove instance propagations between batches with several benefits such as improved predictive accuracy, speed and single pass through the data. They used a Hellinger weighted ensemble of HDDTs to combat concept drift and increase accuracy of single classifiers. Taghi M. Khoshgoftaaret al. [49] have presented a comprehensive empirical investigation using neural network algorithms to learn from imbalanced data with labeling errors. In particular, they investigates the impact of class noise and class imbalance on two common neural network learning algorithms, while considers the ability of data sampling (which is commonly used to address the issue of class imbalance) to improve their performances.

7. Conclusion

In this paper, the state of the art methodologies to deal with class imbalance problem has been reviewed. In recent years, several methodologies integrating solutions to enhance the induced classifiers in the presence of class imbalance by the usage of evolutionary techniques have been presented. This study summarizes the recent developments in the field of class imbalance learning.

References

[1] Juanli Hu, Jiabin Deng, Mingxiang Sui, A New Approach for Decision Tree Based on Principal Component Analysis, Proceedings of Conference on Computational Intelligence and Software Engineering, page no:1-4, 2009.

[2] Huimin Zhao and Atish P. Sinha, An Efficient Algorithm for Generating Generalized Decision Forests, IEEE Transactions on Systems, Man, and Cybernetics —Part A : Systems and Humans, VOL. 35, NO. 5, Page no: 287-299, Septmember 2005.

[3] D. Liu, C. Lai and W. Lee; A Hybrid of Sequential Rules and Collaborative Filtering for Product Recommendation, Information Sciences 179 (20), Page no: 3505-3519, 2009.

[4] M. Mitchell. Machine Learning. McGraw Hill, New York, 1997.

[5] David Hand, HeikkiMannila, and Padhraic Smyth. Principles of Data Mining. MIT Press, August 2001.

[6] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, April 2000.

[7] J. Quinlan. C4.5 Programs for Machine Learning, San Mateo, CA:Morgan Kaufmann, 1993.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

[9] J. Quinlan. Induction of decision trees, Machine Learning, vol. 1, pp. 81C106, 1986.

[10] J. Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg, "Fast asymmetric learning for cascade face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 369–382, Mar. 2008.

[11] N. V. Chawla, N. Japkowicz, and A. Kotcz, Eds., *Proc. ICML Workshop Learn. Imbalanced Data Sets*, 2003.

[12] N. Japkowicz, Ed., *Proc. AAAI Workshop Learn. Imbalanced Data Sets*, 2000.\

[13] G. M.Weiss, "Mining with rarity: A unifying framework," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 7–19, Jun. 2004.

[14] N. V. Chawla, N. Japkowicz, and A. Kolcz, Eds., *Special Issue Learning Imbalanced Datasets, SIGKDD Explor. Newsl.*,vol. 6, no. 1, 2004.

[15] W.-Z. Lu and D.Wang, "Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme," *Sci. Total. Enviro.*, vol. 395, no. 2-3, pp. 109–116, 2008.

[16] Y.-M. Huang, C.-M. Hung, and H. C. Jiau, "Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem," *Nonlinear Anal. R. World Appl.*, vol. 7, no. 4, pp. 720–747, 2006.

[17] D. Cieslak, N. Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets," in *IEEE Int. Conf. Granular Comput.*, 2006, pp. 732–737.

[18] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Netw.*, vol. 21, no. 2–3, pp. 427–436, 2008.

[19] A. Freitas, A. Costa-Pereira, and P. Brazdil, "Cost-sensitive decision trees applied to medical data," in *Data Warehousing Knowl. Discov. (Lecture Notes Series in Computer Science)*, I. Song, J. Eder, and T. Nguyen, Eds.,

[20] K.Kilic¸,O¨ zgeUncu and I. B. Tu¨rksen, "Comparison of different strategies of utilizing fuzzy clustering in structure identification," *Inf. Sci.*, vol. 177, no. 23, pp. 5153–5162, 2007.

[21] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Comput.Med. Imag. Grap.*, vol. 31, no. 6, pp. 362–373, 2007.

[22] X. Peng and I. King, "Robust BMPM training based on second-order cone programming and its application in medical diagnosis," *Neural Netw.*, vol. 21, no. 2–3, pp. 450–457, 2008.Berlin/Heidelberg, Germany: Springer, 2007, vol. 4654, pp. 303–312.

[23] RukshanBatuwita and Vasile Palade (2010) FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning, IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 18, NO. 3, JUNE 2010, pp no:558-571.

[24] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study," Intelligent Data Analysis, vol. 6, pp. 429-450, 2002.

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

[26] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," SIGKDD Explorations, vol. 6, pp. 20-29, 2004.1

[27] D. Cieslak and N. Chawla, "Learning decision trees for unbalanced data," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer-Verlag, 2008, pp. 241–256.

[28] G.Weiss, "Mining with rarity: A unifying framework," *SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 7–19, 2004.

[29] N. Chawla, K. Bowyer, and P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[30] J. Zhang and I. Mani, "KNN approach to unbalanced data distributions: A case study involving information extraction," in *Proc. Int. Conf. Mach. Learning, Workshop: Learning Imbalanced Data Sets*, Washington, DC, 2003, pp. 42–48.

[31] A. Asuncion D. Newman. (2007). *UCI Repository of Machine Learning Database* (School of Information and Computer Science), Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci.edu/ ~mlearn/MLRepository.html.

[32] Iain Brown ⇑, Christophe Mues" An experimental comparison of classification algorithms for imbalanced creditscoring data sets", Expert Systems with Applications 39 (2012) 3446–3453

[33] Ana C. Lorena, Luis F.O. Jacintho, Marinez F. Siqueira, Renato De Giovanni, Lúcia G. Lohmann, André C.P.L.F. de Carvalho, Missae Yamamoto" Comparing machine learning classifiers in potential distribution modelling", Expert Systems with Applications 38 (2011) 5268–5275.
[34] VictoriaLópez, IsaacTriguero, CristóbalJ.Carmona, SalvadorGarcía, FranciscoHerrera" Addressingimbalancedclassification withinstance generation techniques:IPADE-ID",Neurocomputing, 126(2014), 15–28",
[35] NeleVerbiesta, EnislayRamentol, Chris Cornelisa, Francisco Herrera" Preprocessing noisy imbalanced datasets using SMOTE enhanced withfuzzy rough prototype selection" Applied Soft Computing 22 (2014) 511–517.
[36] Peng Cao, DazheZhao andOsmarZaiane,"A PSO-based Cost-Sensitive Neural Network for Imbalanced Data Classification",adfa, p. 1, 2011. © Springer-Verlag Berlin Heidelberg 2011.
[37] Yetian Chen" Learning Classifiers from Imbalanced,Only Positive and Unlabeled Data Sets".
[38] Doucette and Malcolm I. Heywood"GP Classification under Imbalanced Data sets: Active Sub-sampling and AUC Approximation" M. O'Neill et al. (Eds.): EuroGP 2008, LNCS 4971, pp. 266–277, 2008. Springer-Verlag Berlin Heidelberg 2008.
[39] Aditya Krishna Menon, HarikrishnaNarasimhan, Shivani Agarwal, Sanjay Chawla" On the Statistical Consistency of Algorithms for Binary Classification under Class Imbalance", Appearing in Proceedings of the 30 thInternational Conference on Machine Learning, Atlanta, Georgia, USA, 2013.
[40] Shuo Wang, Leandro L. Minku and Xin Yao" A Multi-Objective Ensemble Method for Online Class ImbalanceLearning", 2014 International Joint Conference on Neural Networks (IJCNN)July 6-11, 2014, Beijing, China.
[41] Shuo Wang, Member, IEEE, Leandro L. Minku, Member, IEEE, and Xin Yao, Fellow, IEEE" Resampling-Based Ensemble Methods forOnline Class Imbalance Learning", DOI 10.1109/TKDE.2014.2345380, IEEE Transactions on Knowledge and Data EngineeringIEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.
[42] Bing Yang and Ling Jing," A Novel Nonparallel Plane Proximal SVM for Imbalance Data Classification", JOURNAL OF SOFTWARE, VOL. 9, NO. 9, SEPTEMBER 2014.
[43] M'hamed B. Abidine and BelkacemFergani" A New Multi-Class WSVM Classification toImbalanced Human Activity Dataset" JOURNAL OF COMPUTERS, VOL. 9, NO. 7, JULY 2014.
[44] Hala S. Own1, Ajith Abraham" A Novel-weighted Rough Set-based MetaLearning for Ozone Day Prediction"ActaPolytechnicaHungarica Vol. 11, No. 4, 2014.
[45] Bao-Gang Hu, *Senior Member, IEEE*, Wei-Ming Dong," A study on cost behaviors of binary classificationmeasures in class-imbalanced problems",
[46] Nicola Lunardon, Giovanna Menardi, and Nicola Torelli" ROSE: A Package for Binary ImbalancedLearning" The R Journal Vol. 6/1, June ISSN 2073-4859.
[47] Xiaowan Zhang and Bao-Gang Hu, Senior Member, IEEE" A New Strategy of Cost-Free Learningin the Class Imbalance Problem" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 12, DECEMBER 2014, 1041-4347 _ 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
[48] Andrea Dal Pozzolo, Reid Johnson, Olivier Caelen,Serge Waterschoot, Nitesh V Chawla and GianlucaBontempi" Using HDDT to avoid instances propagation in unbalanced andevolving data streams"
[49] Taghi M. Khoshgoftaar, *Member, IEEE*, Jason Van Hulse, *Member, IEEE*, and Amri Napolitano"Supervised Neural Network Modeling: An EmpiricalInvestigation Into Learning From ImbalancedData With Labeling Errors" IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 21, NO. 5, MAY 2010.