Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

An under-Sampled Approach for Handling Skewed Data Distribution using Cluster Disjuncts

Syed Ziaur Rahman[1]

[1] Andhra University

Received: 11 December 2013 Accepted: 5 January 2014 Published: 15 January 2014

Abstract

In Data mining and Knowledge Discovery hidden and valuable knowledge from the data sources is discovered. The traditional algorithms used for knowledge discovery are bottle necked due to wide range of data sources availability. Class imbalance is a one of the problem arises due to data source which provide unequal class i.e. examples of one class in a training data set vastly outnumber examples of the other class(es). Researchers have rigorously studied several techniques to alleviate the problem of class imbalance, including resampling algorithms, and feature selection approaches to this problem. In this paper, we present a new hybrid frame work dubbed as Majority Under-sampling based on Cluster Disjunct (MAJOR_CD) for learning from skewed training data. This algorithm provides a simpler and faster alternative by using cluster disjunct concept. We conduct experiments using twelve UCI data sets from various application domains using five algorithms for comparison on six evaluation metrics. The empirical study suggests that MAJOR_CD have been believed to be effective in addressing the class imbalance problem.

Index terms— classification, class imbalance, cluster disjunct, under sampling, MAJOR_CD.

1 Introduction

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 [1]. It is noteworthy that class imbalance is emerging as an important issue in designing classifiers [2], [3], [4]. Furthermore, the class with the lowest number of instances is usually the class of interest from the point of view of the learning task [5]. This problem is of great interest because it turns up in many real-world classification problems, such as remote-sensing [6], pollution detection [7], risk management [8], fraud detection [9], and especially medical diagnosis [10]- [13].

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 [14]. The main advantage of external methods as previously pointed out, is that they are independent of the underlying classifier.

Whenever a class in a classification task is under represented (i.e., has a lower prior probability) compared to other classes, we consider the data as imbalanced [15], [16]. 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.

Resampling techniques can be categorized into three groups. Under-sampling methods, which create a subset of the original data-set by eliminating instances (usually majority class instances); oversampling methods, which create a superset of the original dataset by replicating some instances or creating new instances from existing ones; and finally, hybrids methods that combine both sampling methods. Among these categories, there exist several different proposals; from this point, we only center our attention in those that have been used in under sampling. Either way, balancing the data has been found to alleviate the problem of imbalanced data and enhance accuracy [15], [16], [17]. 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 [17]. 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, selective under sampling of majority classes is more accurate than oversampling of minority class. In this paper, we are laying more stress to propose an external class imbalance learning method for solving the class imbalance problem by performing selective under sampling of majority class.

This paper is organized as follows. Section II presets the problem of cluster disjuncts. Section III briefly reviews the data balancing problems and its measures and in Section IV, we discuss the proposed method of MAJOR_CD (Majority Under-sampling based on Cluster Disjunct) for class imbalance learning. Section V presents the imbalanced datasets used to validate the proposed method, while In Section VI, we present the experimental setting and In Section VII discuss, in detail, the classification results obtained by the proposed method and compare them with the results obtained by different existing methods and finally, in Section VIII we conclude the paper.

2 II.

3 Problem of Cluster Disjunct

In Class Imbalance learning, the numbers of instances in the majority class are outnumbered to the number of instances in the minority class. Furthermore, the minority concept may additionally contain a sub concept with limited instances, amounting to diverging degrees of classification difficulty [18][19]. This, in fact, is the result of another form of imbalance, a within-class imbalance, which concerns itself with the distribution of representative data for sub concepts within a class [20][21] **??**22].

The existence of within-class imbalances is closely intertwined with the problem of small disjuncts, which has been shown to greatly depreciate classification performance [20][21] **??**22][23]. Briefly, the problem of small disjuncts can be understood as follows: A classifier will attempt to learn a concept by creating multiple disjunct rules that describe the main concept [18][19], [23]. In the case of homogeneous concepts, the classifier will generally create large disjuncts, i.e., rules that cover a large portion (cluster) of examples pertaining to the main concept. However, in the case of heterogeneous concepts, small disjuncts, i.e., rules that cover a small cluster of examples pertaining to the main concept, arise as a direct result of underrepresented sub concepts [18][19], [23]. Moreover, since classifiers attempt to learn both majority and minority a concept, the problem of small disjuncts is not only restricted to the minority concept. On the contrary, small disjuncts of the majority class can arise from noisy misclassified minority class examples or underrepresented subconcepts. However, because of the vast representation of majority class data, this occurrence is infrequent. A more common scenario is that noise may influence disjuncts in the minority class. In this case, the validity of the clusters corresponding to the small disjuncts becomes an important issue, i.e., whether these examples represent an actual subconcept or are merely attributed to noise. To solve the above problem of cluster disjuncts we propose the method cluster disjunct minority oversampling technique for class imbalance learning.

4 III.

5 Literature Review

In this section, we first review the major research about clustering in class imbalance learning and explain why we choose under-sampling as our technique in this paper.

The different imbalance data learning approaches are as follows: [25] have proposed a method named EPLogCleaner that can filter out plenty of irrelevant items based on the common prefix of their URLs.

M.S.B. PhridviRaj et al. [26] have proposed an algorithm for finding frequent patterns from data streams by performs only one time scan of the database initially and uses the information to find frequent patterns using frequent pattern generation tree. Chumphol Bunkhumpornpat et al. **??**27] have a new over-sampling technique called DBSMOTE is proposed. DBSMOTE technique relies on a density-based notion of clusters and is designed to oversample an arbitrarily shaped cluster discovered by DBSCAN. DBSMOTE generates synthetic instances along a shortest path from each positive instance to a pseudo centroid of a minorityclass cluster. Matías Di Martino et al. [28] have presented a new classifier developed specially for imbalanced problems, where maximum F-measure instead of maximum accuracy guide the classifier design.

V. Garcia et al. [29] have investigated the influence of both the imbalance ratio and the classifier on the performance of several resampling strategies to deal with imbalanced data sets. The study focuses on evaluating how learning is affected when different resampling algorithms transform the originally imbalanced data into artificially balanced class distributions. Table 2 presents recent algorithmic advances in class imbalance learning available in the literature. Obviously, there are many other algorithms which are not included in this table. A profound comparison of the above algorithms and many others can be gathered from the references list.

María Dolores Pérez-Godoy et al. [30] have proposed CO2RBFN, a evolutionary cooperativecompetitive model for the design of radial-basis function networks which uses both radial-basis function and the evolutionary cooperative-competitive technique on imbalanced domains. CO2RBFN follows the evolutionary cooperative-competitive strategy, where each individual of the population represents an RBF (Gaussian function will be considered as RBF) and the entire population is responsible for the definite solution.

This paradigm provides a framework where an individual of the population represents only a part of the solution, competing to survive (since it will be eliminated if its performance is poor) but at the same time cooperating in order to build the whole RBFN, which adequately represents the knowledge about the problem and achieves good generalization for new patterns. **??**

RUSBoost A new hybrid sampling/boosting [29] Algorithm.

6 CO2RBFN

A evolutionary cooperative-competitive [30] model for the design of radial-basis function networks which uses both radial-basis function and the evolutionary cooperative-competitive technique.

7 Improved

Adapt the 2-tuples based genetic tuning [33] FRBCSs approach to classification problems showing the good synergy between this method and some FRBCSs.

8 BSVMs

A model assessment of the interplay [37] between various classification decisions using probability, corresponding decision costs, and quadratic program of optimal margin classifier.

Der-Chiang Li et al. [31] have suggested a strategy which over-samples the minority class and under-samples the majority one to balance the datasets. For the majority class, they build up the Gaussian type fuzzy membership function and a-cut to reduce the data size; for the minority class, they used the mega-trend diffusion membership function to generate virtual samples for the class. Furthermore, after balancing the data size of classes, they extended the data attribute dimension into a higher dimension space using classification related information to enhance the classification accuracy.

Enhong Che et al. [32] have described a unique approach to improve text categorization under class imbalance by exploiting the semantic context in text documents. Specifically, they generate new samples of rare classes (categories with relatively small amount of training data) by using global semantic information of

9 Global Journal of Computer Science and Technology

Volume XIV Issue VII Version I classes represented by probabilistic topic models. In this way, the numbers of samples in different categories can become more balanced and the performance of text categorization can be improved using this transformed data set. Indeed, this method is different from traditional re-sampling methods, which try to balance the number of documents in different classes by re-sampling the documents in rare classes. Such re-sampling methods can cause overfitting. Another benefit of this approach is the effective handling of noisy samples. Since all the new samples are generated by topic models, the impact of noisy samples is dramatically reduced.

Alberto Fernández et al. [33] have proposed an improved version of fuzzy rule based classification systems (FRBCSs) in the framework of imbalanced data-sets by means of a tuning step. Specifically, they adapt the 2-tuples based genetic tuning approach to classification problems showing the good synergy between this method and some FRBCSs. The proposed algorithm uses two learning methods in order to generate the RB for the FRBCS. The first one is the method proposed in [34], that they have named the Chi et al.'s rule generation. The second approach is defined by Ishibuchi and Yamamoto in [35] and it consists of a Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML) algorithm.

J. Burez et al. [36] have investigated how they can better handle class imbalance in churn prediction. Using more appropriate evaluation metrics (AUC, lift), they investigated the increase in performance of sampling (both random and advanced under-sampling) and two specific modeling techniques (gradient boosting and weighted random forests) compared to some standard modeling techniques. They have advised weighted random forests, as a cost-sensitive learner, performs significantly better compared to random forests.

Che-Chang Hsu et al. [37] have proposed a method with a model assessment of the interplay between various classification decisions using probability, corresponding decis ion costs, and quadratic program of optimal margin classifier called: Bayesian Support Vector Machines (BSVMs) learning strategy. The purpose of their learning method is to lead an attractive pragmatic expansion scheme of the Bayesian approach to assess how well it is

aligned with the class imbalance problem. In the framework, they did modify in the objects and conditions of primal problem to reproduce an appropriate learning rule for an observation sample. In [38] Alberto Fernández et al. have proposed to work with fuzzy rule based classification systems using a preprocessing step in order to deal with the class imbalance. Their aim is to analyze the behavior of fuzzy rule based classification systems in the framework of imbalanced data-sets by means of the application of an adaptive inference system with parametric conjunction operators. Jordan M. Malof et al. [39] have empirically investigates how class imbalance in the available set of training cases can impact the performance of the resulting classifier as well as properties of the selected set. In this K-Nearest Neighbor (k-NN) classifier is used which is a well-known classifier and has been used in numerous case-based classification studies of imbalance datasets.

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 falsenegative predictions, which lower the model's performance on the positive class compared with the performance on the negative (majority) class.

IV.

10 Methodology

In this section, we follow a design decomposition approach to systematically analyze the different imbalanced domains. We first briefly introduce the framework design for our proposed algorithm.

The working style of under-sampling tries to remove selective majority instances. Before performing selective under-sampling on the majority subset, the main cluster disjuncts has to be identified and the borderline and noise instances around the cluster disjuncts are to be removed. The number of instances eliminated will belong to the 'k' cluster disjuncts selected by visualization technique. The remaining cluster disjunct instances of the majority subset have to combined with minority set to form improved dataset. Credit History (d). Housing

The algorithm 1: MAJOR_CD can be explained as follows,

The inputs to the algorithm are majority subclass "p" and minority class "n" with the number of features j. The output of the algorithm will be the average measures such as AUC, Precision, F-measure, TP rate and TN rate produced by the MAJOR_CD methods. The algorithm begins with initialization of k=1 and j=1, where j is the number of cluster disjuncts identified by applying visualization technique on the subset "n" and k is the variable used for looping of j cluster disjuncts. The 'j' value will change from one dataset to other, and depending upon the unique properties of the dataset the value of k can be equal to one also i.e no cluster disjunct attributes can be identified after applying visualization technique on the dataset.

In another case attributes related cluster disjunct oversampling can also be performed to improve the skewed dataset. In any case depending on the amount of minority examples generated, the final "strong set" can or cannot be balanced i;e number of majority instances and minority instances in the strong set will or will not be equal.

The presented MAJOR_CD algorithm is summarized as below. The datasets is partitioned into majority and minority subsets. As we are concentrating over sampling, we will take minority data subset for further visualization analysis to identify cluster disjuncts.

11 b) Improve cluster disjunct by removing noisy and borderline instances

Minority subset can be further analyzed to find the noisy or borderline instances so that we can eliminate those. For finding the weak instances one of the ways is that find most influencing attributes or features and then remove ranges of the noisy or weak attributes relating to that feature.

How to choose the noisy instances relating to that cluster disjunct from the dataset set? We can find a range where the number of samples are less can give you a simple hint that those instances coming in that range or very rare or noise. We will intelligently detect and remove those instances which are in narrow ranges of that particular cluster disjunct. This process can be applied on all the cluster disjuncts identified for each dataset.

12 c) Forming the strong dataset

The minority subset and majority subset is combined to form a strong and balance dataset, which is used for learning of a base algorithm. In this case we have used C4.5 or Naïve Bayes as the base algorithm.

V.

13 Evaluation Metrics

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:

The Area under Curve (AUC) measure is computed by equation (**??**), (1) The Precision measure is computed by equation(2),

The F-measure Value is computed by equation(3),

14 Experimental Framework

In this study MAJOR_CD are applied to twelve binary data sets from the UCI repository [40] with different imbalance ratio (IR). Table **??** summarizes the data selected in this study and shows, for each data set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority) and IR. In order to estimate different measure (AUC, precision, Fmeasure, TP rate and TN rate) we use a tenfold cross validation approach, that is ten partitions for training and test sets, 90% for training and 10% for testing, where the ten test partitions form the whole set. For each data set we consider the average results of the ten partitions.

15 Table 3 : Summary of benchmark imbalanced datasets

To validate the proposed MAJOR_CD algorithm, we compared it with the traditional Support Vector Machines (SVM), C4.5, Functional Trees (FT), SMOTE (Synthetic Minority Oversampling TEchnique) and CART algorithm.

16 VII.

17 Results

For all experiments, we use existing prototype's present in Weka [41]. We compare the following domain adaptation methods: The True Negative Rate measure is computed by equation (**??**),

The True Positive Rate measure is computed by equation (**??** We compared proposed method MAJOR_CD with the SVM, C4. **??** [42], FT, SMOTE [43] and CART state-of -the-art learning algorithms. In all the datasets using proposed MAJOR_CD learning algorithm. Second, we compare the classification performance of our proposed MAJOR_CD algorithm with the traditional and class imbalance learning methods based on all datasets.

Following, we analyze the performance of the method considering the entire original algorithms, without pre-processing, data sets for SVM, C4.5, FT and CART. we also analyze a pre-processing method SMOTE for performance evaluation of MAJOR_CD. The complete table of results for all the algorithms used in this study is shown in Table 4 to 9, where the reader can observe the full test results, of performance of each approach with their associated standard deviation. We Table 4, 5, 6, 7, 8 and 9 reports the results of AUC, Precision, F-measure, TP Rate, TN Rate and 9 provide both the numerical average performance (Mean) and the standard deviation (SD) results. If the proposed technique is better than the compared technique then '?' symbol appears in the column. If the proposed technique is not better than the compared technique then '?' symbol appears in the column. The mean performances were significantly different according to the T-test at the 95% confidence level. The results in the tables show that MAJOR_CD has given a good improvement on all the measures of class imbalance learning. This level of analysis is enough for overall projection of advantages and disadvantages of MAJOR_CD. A two-tailed corrected resampled paired t test is used in this paper to determine whether the results of the cross-validation show that there is a difference between the two algorithms is significant or not. Difference in accuracy is considered significant when the p-value is less. algorithm. The method achieves competitive or better results compared to state-of-the-art baselines. We emphasize that our approach is learnerindependent: visualization can be used in conjunction with many of the existing algorithms in the literature. Furthermore, the fact that we select samples in the model space, as opposed to the feature space, is novel and sets it apart from many previous approaches to transfer learning (for both classification and ranking). This allows us to capture the "functional change" assumption and incorporate labeled information in the transfer learning process.

Finally, we can say that MAJOR_CD are one of the best alternatives to handle class imbalance problems effectively. This experimental study supports the conclusion that a cluster disjunct approach for cluster detections and elimination can improve the class imbalance learning behavior when dealing with imbalanced data-sets, as it has helped the MAJOR_CD method to be the best performing algorithms when compared with four classical and well-known algorithms: SVM, C4.5, FT and CART and a wellestablished pre-processing technique SMOTE.

18 VIII.

19 Conclusion

Class imbalance problem have given a scope for a new paradigm of algorithms in data mining. The traditional and benchmark algorithms are worthwhile for discovering hidden knowledge from the data sources, meanwhile class imbalance learning methods can improve the results which are very much critical in real world applications. In this paper we present the class imbalance problem paradigm, which exploits the cluster disjunct concept in the supervised learning research area, and implement it with C4.5 as its base learners. Experimental results show

that MAJOR_CD have performed well in the case of multi class imbalance datasets. Furthermore, MAJOR_CD is much less volatile than C4.5.

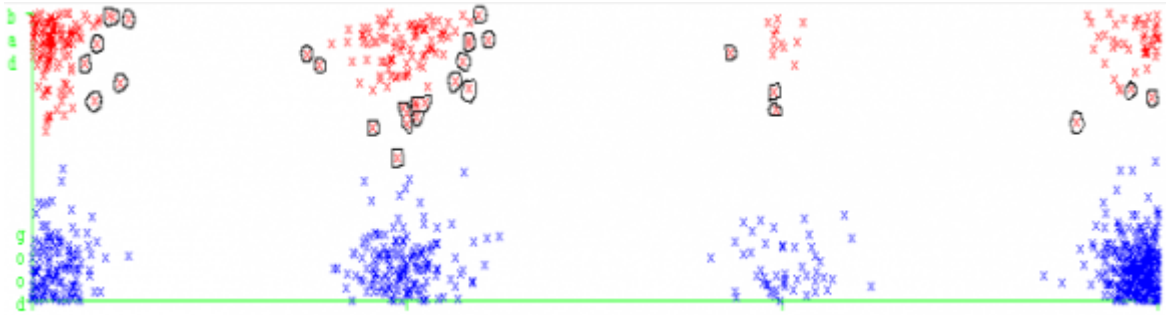
In our future work, we will apply MAJOR_CD to more learning tasks, especially high dimensional feature learning tasks. Another variation of our approach in future work is to analyze the influence of different base classifier effect on the quality of synthetic minority instances generated.

20 Global Journal of Computer Science and Technology

Volume XIV Issue VII Version I [1] [2]

1

Figure 1: Figure 1 :

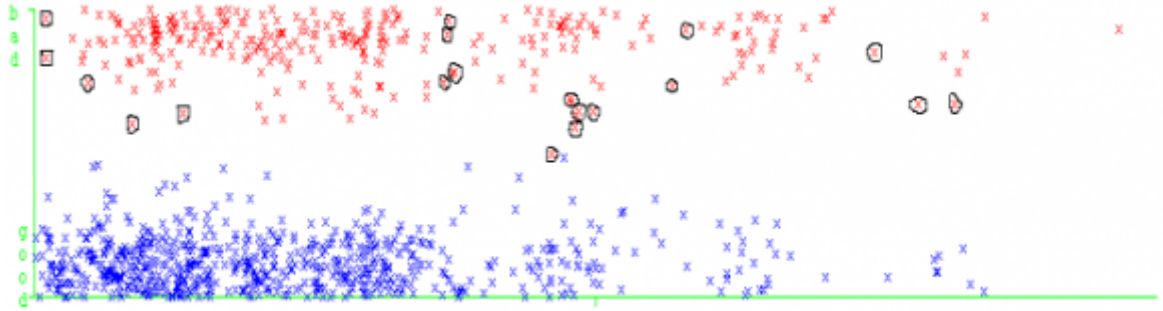


2

Figure 2: Figure 2 :

[1] © 2014 Global Journals Inc. (US)An under-Sampled Approach for Handling Skewed Data Distribution using Cluster Disjuncts

[2] © 2014 Global Journals Inc. (US)



1

Figure 3: Algorithm 1 :

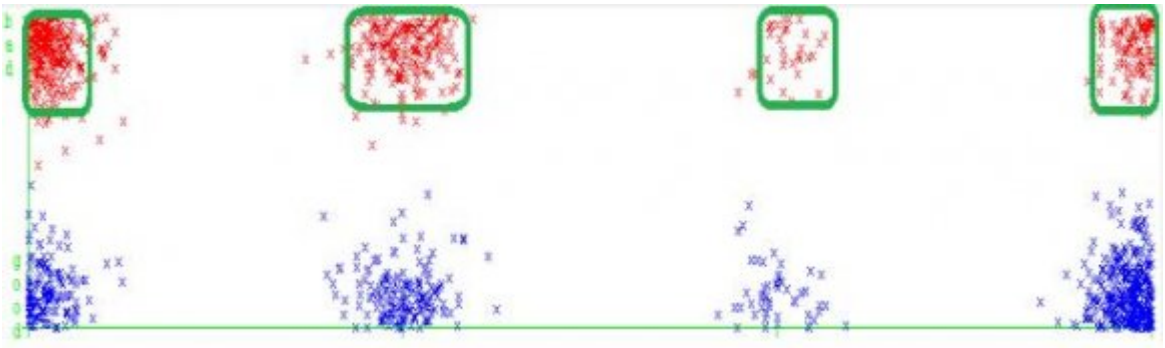


Figure 4:

1

? SAMPLING METHODS
? BASIC SAMPLING METHODS
? Under-Sampling
? Over-Sampling
? ADVANCED SAMPLING METHODS
? Tomek Link
? The SMOTE approach
? Borderline-SMOTE
? One-Sided Selection OSS
? Neighbourhood Cleaning Rule (NCL)
? Bootstrap-based Over-sampling (BootOS)
? ENSEMBLE LEARNING METHODS
? BAGGING
? Asymmetric bagging, SMOTE Bagging
? Over Bagging, Under Bagging
? Roughly balanced bagging
? Lazy Bagging
? Random features selection
? BOOSTING
? Adaboost
? SMOTEBoost
? DataBoost-IM
? RANDOM FORESTS
? Balanced Random Forest BRF
? Weighted Random Forest WRF
? COST-SENSITIVE LEARNING
? Direct cost-sensitive learning methods
? Methods for cost-sensitive meta-learning
? Cost-sensitive meta-learning
? Thresholding methods

Figure 5: Table 1 :

2

ALGORITHM	DESCRIPTION
DCEID	Combining ensemble learning with cost-sensitive learning. [27]

Figure 6: Table 2 :

4

Datasets	SVM	C4.5	FT	SMOTE	CART	MAJOR_CD

Figure 7: Table 4 :

5

An under-Sampled Approach for Handling Skewed Data Distribution using Cluster Disjuncts

experiments we estimate AUC, Precision, F-measure, TP rate and TN rate using 10-fold cross-validation. We experimented with 12 standard datasets for UCI repository; these datasets are standard benchmarks used in the context of high-dimensional imbalance learning. Experiments on these datasets have 2 goals. First, we study the class imbalance properties of the

Breast	67.27±7.28?	74.28±6.05?	6

Datasets	Breast_w	Colic	Credit-g	Diabetes	Hepatitis	Ionosphere	Kv-rs-kp	Labor	Mushroom	Sick	Sonar	___
Colic	0.802±0.073?	0.843±0.070?	0									
Credit-g	0.650±0.075?	0.647±0.062?	0									
Diabetes	0.793±0.072?	0.751±0.070	0									
Hepatitis	0.757±0.195?	0.668±0.184?	0									
Ionosphere	0.900±0.060?	0.891±0.060?	0									
Kr-vs-kp	0.996±0.005?	0.998±0.003	0									
Labor	0.971±0.075?	0.726±0.224?	0									

[Note:

Figure 8: Table 5 :

6

Datasets	SVM	C4.5	FT	SMOTE	CART	MAJOR_CD

Figure 9: Table 6 :

7

Datasets	SVM	C4.5	FT	SMOTE	CART	MAJOR_CD

Figure 10: Table 7 :

8

	1.000±0.00	1.000±0.00
Sick	0.990±0.014?	0.952±0.040?
Sonar		

Breast	0.745±0.051?	0.753±0.042
Breast_w 0.988±0.019? Colic 0.845±0.060? Credit-g 0.776±0.033? Diabetes 0.793±0.037?		0.965±0.026
		0.851±0.055
		0.767±0.025
		0.797±0.045

Year 2014

8 Hepatitis Ionosphere 0.906±0.080? 0.895±0.084 0.604±0.271? 0.510±0.371?

Volume XIV Issue VII Version I

Kv-rs-kp Labor Mushroom 1.000±0.000 0.991±0.008? 0.915±0.197? Sick 0.997±0.003? Sonar 0.764±0.11

D D D D) c (Global Journal of Computer Science and Technology

Colic	0.833±0.055?	0.888±0.044

Datasets Credit-g Diabetes Hepatitis Ionosphere 0.787±0.098? SVM 0.802±0.027 0.778±0.037? 0.469±0.

Hepatitis	0.448±0.273?	0.374±0.256
Ionosphere 0.689±0.131?		0.821±0.107
Kv-rs-kp	0.916±0.021?	0.995±0.005
Labor	0.845±0.243?	0.640±0.349
Mushroom 1.000±0.000		1.000±0.000
Sick	0.984±0.006?	0.995±0.004

Figure 11: Table 8 :

9

Datasets			
Breast	0.26[illegible]	[illegible]	
Breast_w	0.93[illegible]	[illegible]	
Colic	0.717±0.119?	0.734±0.118?	
Credit-g	0.398±0.085?	0.469±0.098?	
Diabetes	0.60[illegible]	[illegible]	
Hepatitis	0.90[illegible]	[illegible]	

Ionosphere 0.940±0.055? 0.949±0.046? Kv-rs-kp 0.993±0.007? 0.990±0.009? Labor 0.865±0.197? 0.945±0.1

Sick Sonar

than 0.05 (confidence level is greater than 95%). In discussion of results, if one algorithm is stated to be bet

complete competitive set of methods and an improvement of results is expected in the

benchmark algorithms i;e SVM, C4.5, FT and CART. However, they are not able to outperform MAJOR_C

Figure 12: Table 9 :

[Information Processing and Management ()] , *Information Processing and Management* 2011. 47 p. .

[Japkowicz ()] '\Balancing Machine Learning Training Data'. N Japkowicz . *Proc. AAAI Workshop Learn. Imbalanced Data Sets*, (AAAI Workshop Learn. Imbalanced Data Sets) 2000. 2004.1. 6 p. .

[Li et al. ()] 'A learning method for the class imbalance problem with medical data sets'. Der-Chiang Li , Chiao-Wenliu , Susanc , Hu . *Computers in Biology and Medicine* 2010. 40 p. .

[Celebi et al. ()] 'A methodological approach to the classification of dermoscopy images'. M E Celebi , H A Kingravi , B Uddin , H Iyatomi , Y A Aslandogan , W V Stoecker , R H Moss . *Comput.Med. Imag. Grap* 2007. 31 (6) p. .

[Batista et al.] *A Study of the Behavior of Several Methods for*, G E A P A Batista , R C Prati , M C Monard .

[Kubat and Matwin ()] 'Addressing the Curse of Imbalanced Training Sets: One-Sided Selection'. M Kubat , S Matwin . *Proc. 14th Int'l Conf. Machine Learning*, (14th Int'l Conf. Machine Learning) 1997. p. .

[María Dolores Pérez-Godoy et al. ()] 'Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets'. Alberto María Dolores Pérez-Godoy , Antonio Jesús Fernández , María José Rivera , Jesus Del . *Pattern Recognition Letters* 2010. 31 p. .

[Hsu et al. ()] 'Bayesian decision theory for support vector machines: Imbalance measurement and feature optimization'. Che-Chang Hsu , Kuo-Shong Wang , Shih-Hsing Chang . *Expert Systems with Applications* 2011. 38 p. .

[Chawla et al. ()] N V Chawla , N Japkowicz , A Kotcz . *Proc. ICML Workshop Learn. Imbalanced Data Sets*, (ICML Workshop Learn. Imbalanced Data Sets) 2003.

[Chawla et al. ()] N V Chawla , N Japkowicz , A Kolcz . *Special Issue Learning Imbalanced Datasets*, 2004. 6.

[Chi et al. ()] Z Chi , H Yan , T Pham . *Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition*, 1996. World Scientific.

[Prati et al. ()] 'Class Imbalances versus Class Overlapping: An Analysis of a Learning 22. System Behavior'. R C Prati , G E A P A Batista , M C Monard . *Proc. Mexican Int'l Conf. Artificial Intelligence*, (Mexican Int'l Conf. Artificial Intelligence) 2004. p. .

[Jo and Japkowicz ()] 'Class Imbalances versus Small Disjuncts'. T Jo , N Japkowicz . *ACM SIGKDD Explorations Newsletter* 2004. 6 (1) p. .

[Japkowicz ()] 'Class Imbalances: Are We Focusing on the Right Issue?'. N Japkowicz . *Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II*, (Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II) 2003.

[Cieslak et al. ()] 'Combating imbalance in network intrusion datasets'. D Cieslak , N Chawla , A Striegel . *IEEE Int. Conf. Granular Comput*, 2006. p. .

[Freitas et al. ()] 'Comparison of different strategies of utilizing fuzzy clustering in structure identification'. A Freitas , A Costa-Pereira , P Brazdil . *Data Warehousing Knowl. Discov*, Lecture Notes Series in Computer Science I Song, J Eder, T Nguyen, . K Eds, I B Kilic¸, Tu¨rksen (ed.) 2007. 177 p. . (Costsensitive decision trees applied to medical data)

[Witten and Frank ()] *Data Mining: Practical machine learning tools and techniques*, I H Witten , E Frank . 2005. San Francisco. (2nd edition Morgan Kaufmann)

[Phridviraj and Gururao ; Chidchanok Lursinsap ()] 'Data miningpast, present and future -a typical survey on data Streams'. M S B Phridviraj , C V Gururao ; Chidchanok Lursinsap . *Chumphol Bunkhumpornpat, Krung Sinapiromsaran*, 2014. 2012. 12 p. . (Appl Intell)

[Sha et al. ()] 'EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining'. Hongzhou Sha , Tingwen Liu , Peng Qin , Yong Sun , Qingyun Liu . *Procedia Computer Science* 2013. 17 p. .

[Huang et al. ()] 'Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem'. Y.-M Huang , C.-M Hung , H C Jiau . *Nonlinear Anal. R. World Appl* 2006. 7 (4) p. .

[Che et al. ()] 'Exploiting probabilistic topic models to improve text categorization under class imbalance'. Enhong Che , Yanggang Lin , Hui Xiong , Qiming Luo , Haiping Ma . *C4.5: Programs for Machine Learning*, . J R Quinlan (ed.) (San Mateo, CA) 1993. Morgan Kaufmann Publishers. (1st ed)

[Wu et al. (2008)] 'Fast asymmetric learning for cascade face detection'. J Wu , S C Brubaker , M D Mullin , J M Rehg . *IEEE Trans. Pattern Anal. Mach. Intell* Mar. 2008. 30 (3) p. .

[Batuwita and Vasile Palade ()] 'FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning'. Rukshan Batuwita , Vasile Palade . *IEEE TRANSACTIONS ON FUZZY SYSTEMS* 2010. JUNE 2010. 18 (3) p. .

[Lu and Wang ()] 'Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme'. W.-Z Lu , D Wang . *Sci. Total. Enviro* 2008. 395 (2-3) p. .

[Burez and Van Den Poel ()] 'Handling class imbalance in customer churn prediction'. J Burez , D Van Den Poel . *Expert Systems with Applications* 2009. 36 p. .

[Ishibuchi et al. ()] 'Hybridization of fuzzy GBML approaches for pattern classification problems'. H Ishibuchi , T Yamamoto , T Nakashima . *IEEE Transactions on System, Man and Cybernetics B* 2005. 35 (2) p. .

[Quinlan ()] 'Induction of Decision Trees'. J R Quinlan . *Machine Learning*, 1986. 1 p. .

[Weiss ()] 'Mining with Rarity: A Unifying Framework'. G M Weiss . *Procedia -Social and Behavioral Sciences* 2004. 2013. 6 (1) p. . (ACM SIGKDD Explorations Newsletter)

[Weiss (2004)] 'Mining with rarity: A unifying framework'. G M Weiss . *ACM SIGKDD Explor. Newslett* Jun. 2004. 6 (1) p. .

[Martino et al. ()] 'Novel classifier scheme for imbalanced problems'. Matías Di Martino , Alicia Fernández , Pablo Iturralde , Federico Lecumberry . *Pattern Recognition Letters* 2013. 34 p. .

[Fernández et al. ()] 'On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets'. Alberto Fernández , María José Del Jesus , Francisco Herrera . *Information Sciences* 2010. 180 p. .

[Garcia et al. ()] 'On the effectiveness of preprocessing methods when dealing with different levels of class imbalance'. V Garcia , J S Sanchez , R A Mollineda . *Knowledge-Based Systems* 2012. 25 p. .

[Fernández et al. ()] 'On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets'. Alberto Fernández , María José Del Jesus , Francisco Herrera . *Expert Systems with Applications* 2009. 36 p. .

[Peng and King ()] 'Robust BMPM training based on second-order cone programming and its application in medical diagnosis'. X Peng , I King . *Neural Netw* 2008. 2007. Springer. 21 (2-3) p. .

[Chawla et al. ()] 'SMOTE: Synthetic minority over-sampling technique'. N Chawla , K Bowyer , P Kegelmeyer . *J. Artif. Intell. Res* 2002. 16 p. .

[Japkowicz and Stephen ()] 'The Class Imbalance Problem: A Systematic Study'. N Japkowicz , S Stephen . *Intelligent Data Analysis* 2002. 6 p. .

[Malof et al. ()] 'The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support'. Jordan M Malof , Maciej A Mazurowski , Georgia D Tourassi . *Neural Networks* 2012. 25 p. .

[Mazurowski et al. ()] 'Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance'. M A Mazurowski , P A Habas , J M Zurada , J Y Lo , J A Baker , G D Tourassi . *Neural Netw* 2008. 21 (2-3) p. .

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