International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 2, Issue 3, May – June 2013 **ISSN 2278-6856**

An Exhaustive Literature Review on Class Imbalance Problem

K.P.N.V.SATYASREE[1], Dr.J.V.R.MURTHY[2]

[1]Assistant Professor, Department of Computer Science and Engineering , Vignan's Nirula Institute of Technology in Science for WOMEN, Guntur, A.P, India,
[2]Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Kakinada, Kakinada, A.P, India

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. Unbalanced dataset learning is a new paradigm of machine learning which has applicability in real time, since all the datasets of real time are of unbalanced in nature. This paper presents an updated literature survey of current methods for constructing classification models for imbalanced datasets. The paper suggests a unified framework for recent developments and describes the benchmark datasets and methodologies.*

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 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].

2.1. 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.

2.2. Decision Trees

A decision tree is a tree data structure with the following properties:

- Each leaf is labeled with the name of a class;
- The root and each internal node (called a decision node) are labeled with the name of an attribute;
- Every internal node has a set of at least two children, where the branches to the children are labeled with disjoint values or sets of values of that node's attribute such that the union of these constitutes the set of all possible values for that attribute. Thus, the labels on the arcs leaving a parent node form a partition of the set of legal values for the parent's attribute.

A decision tree can be used to classify a case by starting at the root of the tree and moving through it until a leaf is reached [7]. At each decision node, the case's outcome for the test at the node is determined and attention shifts to the root of the sub-tree corresponding to this outcome. When this process finally (and inevitably) leads to a leaf, the class of the case is predicted to be that labeled at the leaf.

2.3. Building Decision Trees

Every successful decision tree algorithm (e.g. CART [8], ID3 [9], C4.5 [7]) is an elegantly simple greedy algorithm:

i. Pick as the root of the tree the attribute whose values best separate the training set into subsets (the best partition is one where all elements in each subset belong to the same class);
ii. Repeat step (i) recursively for each child node until a stopping criterion is met.

Examples of stopping criteria are:

- Every subset contains training examples of only one class;
- The depth of the tree reaches a predefined threshold;
- Lower bound on the number of elements that must be in a set in order for that set to be partitioned further is reached (CART [8], C4.5 [7]).

The dominating operation in building decision trees is the gathering of histograms on attribute values. As mentioned earlier, all paths from a parent to its children partition the relation horizontally into disjoint subsets. Histograms have to be built for each subset, on each attribute, and for each class individually.

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. In this paper, we are laying more stress to propose an external CIL method for solving the class imbalance problem.

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, undersampling is performed on majority classes either randomly or from areas far away from the decision boundaries. We note that random undersampling 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 undersampling of majority classes [26].

Re-sampling techniques can be categorized into three groups. Undersampling 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 data-set 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.

- *Random undersampling*: It is a non-heuristic method that aims to balance class distribution through the random elimination of majority class examples. Its major drawback is that it can discard potentially useful data, which could be important for the induction process.
- *Random oversampling*: In the same way as random oversampling, it tries to balance class distribution, but in this case, randomly replicating minority class instances. Several authors agree that this method can increase the likelihood of occurring over fitting, since it makes exact copies of existing instances.
- *Hybrid Methods:* In this hybrid method both undersampling and oversampling will be applied for the datasets so as to make it a balance dataset.

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]. The following two sections briefly discuss the external-imbalance and internal-imbalance learning methods.

The external methods are independent from the learning algorithm being used, and they involve preprocessing of the training datasets to balance them before training the classifiers. Different re-sampling methods, such as random and focused oversampling and undersampling, fall into to this category. In random undersampling, the majority-class examples are removed randomly, until a particular class ratio is met [28]. In random oversampling, the minority-class examples are randomly duplicated, until a particular class ratio is met [27]. 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].

5. Evaluation Criteria's for Class Imbalance Learning

In this section, we follow a design decomposition approach to systematically analyze the different unbalanced domains.

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 numbers 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,

$$TruePositi veRate = \frac{TP}{(TP) + (FN)}$$

The True Negative Rate measure is computed by,

$$TrueNegati veRate = \frac{TN}{(TN) + (FP)}$$

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 1
SUMMARY OF BENCHMARK IMBALANCED DATASETS

S.no	Datasets	#Ex.	#Atts.	Class (-,+)	IR
1.	Breast	268	9	(recurrence; no-recurrence)	2.37
2.	Breast_w	699	9	(benign; malignant)	1.90
3.	Colic	368	22	(yes;no)	1.71
4.	Credit-g	1000	21	(good; bad)	2.33
5.	Diabetes	768	8	(tested-potv; tested-negtv)	1.87
6.	Heart-c	303	14	(<50,>50_1)	1.19
7.	Heart-h	294	14	(<50,>50_1)	1.77
8.	Heart-stat	270	14	(absent, present)	1.25
9.	Hepatitis	155	19	(die; live)	3.85
10.	Ionosphere	351	34	(b;g)	1.79
11.	Kr-vs-kp	3196	37	(won; nowin)	1.09
12.	Labor	56	16	(bad ; good)	1.85
13.	Mushroom	8124	23	(e ; p)	1.08
14.	Sick	3772	29	(negative , sick)	15.32
15.	Sonar	208	60	(rock ; mine)	1.15

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.

6. Recent Advances on Class Imbalance Learning

Currently, the research in class imbalance learning mainly focuses on the integration of imbalance class learning with other AI techniques. How to integrate the class imbalance learning with other new techniques is one of the hottest topics in class imbalance learning research. There are some of the recent research directions for class imbalance learning as follows:
A comprehensive review of different CIL methods can be found in [32]. The following two sections briefly discuss the external-imbalance and internal-imbalance learning methods. The external methods are independent from the learning algorithm being used, and they involve preprocessing of the training datasets to balance them before training the classifiers. Different re-sampling methods, such as random and focused oversampling and under-sampling, fall into to this category. In random under-sampling, the majority-class examples are removed randomly, until a particular class ratio is met [33]. In random oversampling, the minority-class examples are randomly duplicated, until a particular class ratio is met [32]. Synthetic minority oversampling technique (SMOTE) [34] 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 [35].

In [36] authors have proposed a minority sample method based on k-means clustering and genetic algorithm. The proposed algorithm works in two stages, in the first stage k-means clustering is used to find clusters in the minority set and in the second stage genetic algorithm is used to choose the best minority samples for re-sampling. Classifier ensemble techniques can also be efficiently used for handling the problem of class imbalance learning. In [37] authors have proposed another variant of clustering-based sampling method for handling class imbalance problem. In [38] authors have proposed a pure genetic algorithm based sampling method for handling class imbalance problem. Evolutionary algorithms are also of great use for handling the problem of class imbalance. In [39] authors have proposed evolutionary method of nested generalized exemplar family which uses Euclidean n-space to store objects when computing distances to the nearest generalized exemplar. This method uses evolutionary algorithm for selection of most suitable generalized exemplars for re-sampling. In [40] authors have proposed an evolutionary cooperative competitive strategy for the design of radial-basis function networks CO2RBFN by using evolutionary cooperative competitive technique with radial-basis function on imbalanced datasets. In CO2RBFN framework where an individual of population represents only a part of the solution, competing to survival and at the same time cooperating in order to build the whole RBFN, which achieves good generalization for new patterns by representing the complete knowledge about the problem space.

In [41] authors have proposed a dynamic classifier ensemble method (DCEID) by ensemble technique with cost sensitive learning. A new cost-sensitive measure is proposed to combine the output of ensemble classifiers. A comparative analysis of external and internal methods for handling class imbalance learning is conducted [42]. The results of analysis are to study deeply about data intrinsic properties for proposal of data shifting and data overlapping. In [43] authors have given suggestion for applying gradient boosting and random forest classifier for better performance on class imbalanced datasets. In [44] authors have introduced a new hybrid sampling/boosting algorithm, called RUSBoost from its individual component AdaBoost and SMOTEBoost, which is another algorithm that combines boosting and data sampling for learning from skewed training data. In [45] authors have proposed a max-min technique for extracting maximum negative features and minimum positive features for handling the problem of class imbalance datasets for binary classification. The proposed two models which can do the max-min extraction of features have been implemented simultaneously thereby producing effective results. The application of fuzzy rule based technique for handling class imbalance datasets is proposed as Fuzzy Rule Based Classification Systems (FRBCSs) [46] and Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML) algorithm [47]. These fuzzy based algorithmic approaches had also shown a good performance in terms of metric learning.

In [48] authors have proposed to deal the imbalanced datasets by using preprocessing step with fuzzy rule based classification systems by application of an adaptive inference system with parametric conjunction operators. In [49] authors have proposed the applicability of K-Nearest Neighbor (k-NN) classifier to investigate the performance on class imbalanced datasets. In [50] authors have conducted a systematic study to investigate the effects of different classifiers with different re-sampling strategies on imbalanced datasets with different imbalance ratios.

Gangs Wuet al. [51] have proposed a kernel-boundary-alignment algorithm for handling class distribution imbalance problem. The proposed model improved on prediction accuracy of SVMs which is sensitivity to class boundaries. Zhao Zhang *et al.* [52] have proposed two novel multimodal nonlinear methods known as semi-supervised Laplacian Eigenmaps (S2LAE) and semi-supervised Locally Linear Embedding (S2LLE) for marginal manifold visualization. The pair wise constraints of cannot-link and must-link are used to induce neighborhood graph. Zhi-Hua Zhou*et al.* [53] have proposed two novel methods known as hard-ensemble (hard voting) and soft-ensemble (soft voting) using the concept of threshold moving. Threshold moving tries to move the output threshold towards inexpensive

classes so that examples with lower costs will be easy to misclassify and examples with higher cost will be difficult to misclassify. One of the limitations of this approach is that the cost of all the examples has to be provided explicitly. Jing Li *et al.* [54] have proposed a naïve approach for general optimal router query which can be efficiently applied for large datasets. This approach is motivated to avoid repeated computation for general query generation by using two different methodologies using backward search and forward search. They also used their proposed methods for answering a variant of optimal route queries.

Hanjiang Lai *et al.* [55] have proposed a sparse learning to rank problem using an fenchel and primal duality perspective. The proposed learning strategy efficiently addresses the optimization problem for information retrieval. One of the limitation of the proposed model is it can be only efficiently implemented on the ranking models which are constrained to be with few nonzero constraints. Le Xu *et al.*[56] have proposed a power distribution systems using fuzzy classification algorithm introduced by Ishibuchi *et al.*[57] to alleviate the effect of imbalanced data constitution, is applied to Duke Energy outage data for distribution fault cause identification. Steven Euijong Whang *et al.* [58] has proposed an Entity Resolution method pay-as-you-go for finding the matching records by using hints. One of the advantage of the proposed method is it uses limited amount of work for identifying matching records. The proposed learning strategy also solves the problem of tolerating any entity resolution processing for large amount of times. MdNasir *et al.* [59] has proposed a class imbalance algorithm which works by resorting to stochastic gradient algorithm. One of the strength of the proposed algorithm is the applicability of statistical techniques for handling the uneven class distribution. Xuan Li *et al.* [60] have proposed a graph-based ranking method for retrieving the similar type of documents by using salience of the sentences in the current documents. The proposed approach handled the problem of proper document summarization by using the manifold and large-margin constraint ranking. Due to the limitation of being an NP-hard problem only approximation methods had been proposed with polynomial time complexity. Xu-Ying Liu*et al.* [61] has proposed two ensemble techniques known as Easy Ensemble and Balance Cascade. In the first proposed approach several subsets samples from the majority class are prepared and trains a learner using each of them, and combines the outputs of those learners. The second proposed approach trains the learners sequentially, where in each step, the majority class examples that are correctly classified by the current trained learners are removed from further consideration. One of the limitations of the both proposed approach are the increase in the overall complexity.

Kelvin Sim *et al.* [62] have proposed a clustering framework CAT seeker which works on 3d subspaces from proper clustering of biological datasets using domain specific knowledge of datasets with parameter insensitivity and singular value decomposition. The proposed approach solved the problem of wrong threshold setting and reduced clustering quality. Haiying Cao *et al.* [63] have proposed adual-input nonlinear model based on a real-valued Volterra series for compensation of the nonlinear frequency-dependent I/Q imbalance. Le Xu*et al.* [64] have proposed a supervised algorithm known as artificial immune recognition system (AIRS). The problem of data imbalance is handled in duke energy outage dataset using three major causes (tree, animal, and lightning) as prototypes. Shan-Hung Wu *et al.* [65] have developed an extension of SVM known as asymmetric SVM (ASVM) for reducing cost of false predictions from different classes. The main objective of the proposed ASVM is to allow user tolerance on false-positive rate as specified by the user. One of the strength of ASVM is that it can increase the accuracy of the overall model with improved training time. David P. Williams *et al.* [66] have proposed an infinitely imbalanced logistic regression (IILR) algorithm for handling the problem of class imbalance specifically for remote-sensing classification problems. One of the advantages of the proposed algorithm is the formulation strategy used for explicitly handling the class imbalance problem efficiently.

Jisu Oh *et al.* [67] have developed a predictive as well as reactive method for data-intensive real-time applications. The model is predictive in terms of having efficient capability for providing data-services for real-time applications and reactive in terms of adjusting load bounds for dynamically varying work load. One of the limitation of this model is it trades off reactivity for predictive and vice versa. Chris Seiffert *et al.* [68] have presented comprehensive study different data-mining techniques for handling class imbalance problem specifically for software–quality assurance. The efficient applicability of data sampling and boosting algorithms for high-assurance systems is one of the strengths of the study. Y. S. Ihn*et al.* [69] have developed an adaptive dynamic rotor balance scheme, which is of great worth. The use of least mean squares algorithm for the formulation of the control law has improved the robustness of the imbalance correction and improved manufacturing yields. Yangqiu Song *et al.* [70] have proposed hidden Markov random field (HMRF) using exception maximization (EM) algorithm for optimization of the algorithm. The proposed algorithms had couple of variants using automatic construct document constraint and automatic construct word document constraint for unsupervised learning.

Junfeng Pan *et al.* [71] have proposed a staged framework for data preprocessing to support data mining. The proposed framework pushes the cost sensitivity and data imbalance of customer retention data into the data preprocessing itself. One of the strength of the framework is the applicability in the field of customer churning or

attrition for the industries. Chris Seiffert *et al.* [72] have developed a couple of new hybrid algorithms known as RUSBoost and SMOTEBoost. The first algorithm uses data sampling techniques to minimize the problem of class imbalance. The second algorithm combines both boosting and data sampling techniques to solve the problem of class imbalance. Rukshan Batuwita *et al.* [73] have a present FuzzySVMs (FSVMs) is a variant of the SVM algorithm, which has been proposed to handle the problem of outliers and noise. In FSVMs, training examples are assigned different fuzzy-membership values based on their importance, and these membership values are incorporated into the SVM learning algorithm to make it less sensitive to outliers and noise.

In [74] Taghi M. Khoshgoftaar *et al.* This study presents a comprehensive empirical investigation using neural network algorithms to learn from imbalanced data with labeling errors. In particular, the first component of our study investigates the impact of class noise and class imbalance on two common neural network learning algorithms, while the second component considers the ability of data sampling (which is commonly used to address the issue of class imbalance) to improve their performances. In [75], Taghi M. Khoshgoftaar *et al.* authors presented algorithms include SMOTEBoost, RUSBoost, Exactly Balanced Bagging, and Roughly Balanced Bagging, combine boosting or bagging with data sampling to make them more effective when data are imbalanced.
Mikel Galar*et al.* [76] have given a review of the state of the art on ensemble techniques in the framework of imbalanced data-sets, with focus on two-class problems. They also presented propose a taxonomy for ensemble-based methods to address the class imbalance where each proposal can be categorized depending on the inner ensemble methodology in which it is based. Shuo Wang*et al.* [77] have given a brief overview of ensemble techniques used for class imbalance learning. The study also focuses on class imbalance of single and multi-class. Shuo Wang*et al.* [78] have studied the challenges posed by the multiclass imbalance problems and investigate the generalization ability of some ensemble solutions. The authors also proposed a novel algorithm AdaBoost.NC, with the aim of handling multiclass imbalance problem. The proposed algorithm uses multi-majority and multi-minority concepts on basic re-sampling techniques.

Urvesh Bhowan*et al.* [79] have proposed ensemble of genetic program classifiers using Multi-objective Genetic Programming (MOGP) approach to handle class imbalance problem. The evolved ensembles comprise of non-dominated solutions in the population where individual members vote on class membership. One of the limitations of this model is to develop an effective fitness evaluation strategy in the underlying MOGP algorithm to evolve good ensemble members. Nicolás García-Pedraja *et al.* [80] have a presented a new approach to dealing with the class-imbalance problem that is scalable to data sets with many millions of instances and hundreds of features. This proposal is based on the divide-and-conquer principle combined with application of the selection process to balanced subsets of the whole data set. This divide-and-conquer principle allows the execution of the algorithm in linear time. Furthermore, the proposed method is easy to implement using a parallel environment and can work without loading the whole data set into memory.

Hualong Yu *et al.* [81] have presented a skewed gene selection algorithm that introduces a weighted metric into the gene selection procedure. The extracted genes are paired as decision rules to distinguish both classes, with these decision rules then integrated into an ensemble learning framework by majority voting to recognize test examples; thus avoiding tedious data normalization and classifier construction. Urvesh Bhowan *et al.* [82] has presents a novel method that first converts the imbalanced binary-class data into balanced multiclass data and then builds a defect predictor on the multiclass data with a specific coding scheme.

Obviously, there are many other algorithms which are not included in this literature. A profound comparison of the above algorithms and many others can be gathered from the references list.

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.

7. Conclusion

In this paper, the state of the art methodologies to deal with class imbalance problem has been reviewed. This issue hinders the performance of standard classifier learning algorithms that assume relatively balanced class distributions, and classic ensemble learning algorithms are not an exception. 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. However, there was a lack of framework where each one of them could be classified; for this reason, a taxonomy where they can be placed has been taken as our future work. Finally, we have concluded that prominent based algorithms are the need of the hour for improving the results that are obtained by the usage of data preprocessing techniques and training a single classifier.

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 MichelineKamber. 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] 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.

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

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

[35] 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.

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

[37] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 40–49, 2004.

[38] S. Zou, Y. Huang, Y. Wang, J. Wang, and C. Zhou, "SVM learning from imbalanced data by GA sampling for protein domain prediction," in *Proc. 9th Int. Conf. Young Comput. Sci.*, Hunan, China, 2008, pp. 982– 987.

[39] Salvador Garcı´a, Joaquı´nDerrac, Isaac Triguero, Cristobal J. Carmona, Francisco Herrera, "Evolutionary-based selection of generalized instances for imbalanced classification", Knowledge-Based Systems 25 (2012) 3–12.

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

[41] Jin Xiao, Ling Xie, Changzheng He, Xiaoyi Jiang," Dynamic classifier ensemble model for customer classification with imbalanced class distribution", Expert Systems with Applications 39 (2012) 3668–3675.

[42] Z. Chi, H. Yan, T. Pham, Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition, World Scientific, 1996.

[43] V. Garcia, J.S. Sanchez , R.A. Mollineda," On the effectiveness of preprocessing methods when dealing with different levels of class imbalance", Knowledge-Based Systems 25 (2012) 13–21.

[44] Alberto Fernández, María José del Jesus, Francisco Herrera," On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets", Information Sciences 180 (2010) 1268–1291.

[45] Jinguha Wang, JaneYou ,QinLi, YongXu," Extract minimum positive and maximum negative features for imbalanced binary classification", Pattern Recognition 45 (2012) 1136–1145.

[46] Alberto Fernández, María José del Jesus, Francisco Herrera," On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets", Expert Systems with Applications 36 (2009) 9805–9812.

[47] Victoria López, Alberto Fernández, Jose G. Moreno-Torres, Francisco Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics", Expert Systems with Applications 39 (2012) 6585–6608.

[48] Alberto Fernández, María José del Jesus, Francisco Herrera," On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets", Expert Systems with Applications 36 (2009) 9805–9812.

[49] Jordan M. Malof, Maciej A. Mazurowski, Georgia D. Tourassi," The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support", Neural Networks 25 (2012) 141–145.

[50] V. Garcia, J.S. Sanchez , R.A. Mollineda," On the effectiveness of preprocessing methods when dealing with different levels of class imbalance", Knowledge-Based Systems 25 (2012) 13–21.

[51] Gang Wu, Edward Y. Chang, "KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2005, pp.no: 786-795.

[52] Zhi-Hua Zhou and Yuan Jiang, NeC4.5: Neural Ensemble Based C4.5, *IEEE Transactions on Knowledge and Data Engineering*, VOL. 16, NO. 6, page no: 770-773, June 2004..

[53] Zhi-Hua Zhou, Xu-Ying Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 1, JANUARY 2006, pp.no: 63-77.

[54] Le Xu, Mo-Yuen Chow, Leroy S. Taylor " Power Distribution Fault Cause Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification E-Algorithm", IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 22, NO. 1, FEBRUARY 2007, pp.no: 164 – 171.

[55] Jordan M. Malof, Maciej A. Mazurowski, Georgia D. Tourassi," The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support", Neural Networks 25 (2012) 141–145.

[56] Le Xu, Mo-Yuen Chow, Leroy S. Taylor " e Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification E-Algorithm", IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 22, NO. 1, FEBRUARY 2007, pp.no: 164 – 171.

[57] H. Ishibuchi and T. Yamamoto, "Comparison of heuristic rule weight specification methods," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2002, pp. 908–913.

[58] Liang Wang, Uyen T.V. Nguyen, James C. Bezdek, Christopher A. Leckie, and Kotagiri Ramamohanarao. iVATand aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment,*M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI* 6118, pp. 16–27, 2010.

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

[60] Liang Wang, Uyen T.V. Nguyen, James C. Bezdek, Christopher A. Leckie, and KotagiriRamamohanarao. iVATand aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment,*M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI* 6118, pp. 16–27, 2010.

[61] Xu-Ying Liu, Jianxin Wu, Zhi-Hua Zhou" Exploratory Undersampling for Class-Imbalance Learning", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 39, NO. 2, APRIL 2009, pp.no:539 – 550.

[62] Jordan M. Malof, Maciej A. Mazurowski, Georgia D. Tourassi," The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support", Neural Networks 25 (2012) 141–145.

[63] Jin Xiao, Ling Xie, Changzheng He, Xiaoyi Jiang," Dynamic classifier ensemble model for customer classification with imbalanced class distribution", Expert Systems with Applications 39 (2012) 3668–3675.

[64] Le Xu, Mo-Yuen Chow, Leroy S. Taylor " Power Distribution Outage Cause Identification With Imbalanced Data Using Artificial Immune Recognition System (AIRS) Algorithm", IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 22, NO. 1, FEBRUARY 2007, pp.no: 164 – 171.

[65] Shuo Wang, Xin Yao "Relationships between Diversity of Classification Ensembles and Single-Class Performance Measures", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013, pp.no: 206 – 219.

[66] David P. Williams, Vincent Myers, Miranda SchattenSilvious "Mine Classification With Imbalanced Data", IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 6, NO. 3, JULY 2009, pp.no: 528 – 532.

[67] Chris Seiffert, Taghi M. Khoshgoftaar, *Member, IEEE*, and Jason Van Hulse, *Member, IEEE "* Improving Software-Quality Predictions With Data Sampling and Boosting", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 39, NO. 6, NOVEMBER 2009, pp.no: 1283 – 1294.

[68] Haiying Cao, Ali SoltaniTehrani, Christian Fager, Thomas Eriksson, and Herbert Zirath," I/Q Imbalance Compensation Using a Nonlinear Modeling Approach", IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES, VOL. 57, NO. 3, MARCH 2009, pp.no:513 – 518.

[69] Junfeng Pan and Qiang Yang, Yiming Yang and Lei Li, Frances Tianyi Li and George Wenmin Li "Cost-Sensitive-Data Preprocessing for Mining Customer Relationship Management Databases", JANUARY/FEBRUARY 2007, A Technical Report.

[70] Claudia Diamantini, DomenicoPotena "Bayes Vector Quantizer for Class-Imbalance Problem" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009, pp.no:638 – 651.

[71] Y. S. Ihn, J. K. Lee, D.H.Oh, H. S. Lee, J. C. Koo " Active Correction of Dynamic Mass Imbalance for a Precise Rotor", IEEE TRANSACTIONS ON MAGNETICS, VOL. 45, NO. 11, NOVEMBER 2009, pp.no: 5088 – 5093.

[72] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, Amri Napolitano" RUSBoost: A Hybrid Approach to Alleviating Class Imbalance", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 40, NO. 1, JANUARY 2010. pp.no: 185 – 197.

[73] RukshanBatuwita, Vasile Palade "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.

[74] Taghi M. Khoshgoftaar, Jason Van Hulse, Amri Napolitano "Supervised Neural Network Modeling: An Empirical Investigation Into Learning From Imbalanced Data With Labeling Errors", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 21, NO. 5, MAY 2010, pp.no: 813 – 830.

[75] Taghi M. Khoshgoftaar, Jason Van Hulse, Amri Napolitano "Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 41, NO. 3, MAY 2011, pp.no: 552 – 568.

[76] MikelGalar, Alberto Fern´andez, EdurneBarrenechea, HumbertoBustince, Francisco Herrera " A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 42, NO. 4, JULY 2012, pp.no: 463 - 484.

[77] Shuo Wang, Xin Yao "Relationships between Diversity of Classification Ensembles and Single-Class Performance Measures", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013, pp.no: 206 – 219.

[78] Shuo Wang, XinYao"Multiclass Imbalance Problems: Analysis and Potential Solutions", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 42, NO. 4, AUGUST 2012, pp.no:1119 – 1130.

[79] UrveshBhowan, Mark Johnston, Mengjie Zhang and Xin Yao "Evolving Diverse Ensembles usingGenetic Programming for Classification with Unbalanced Data" This article has been accepted for publication in a future issue of this IEEE journal, Copyright (c).2012 IEEE.

[80] NicolásGarcía-Pedrajas, Javier Pérez-Rodríguez, Aida de Haro-García "OligoIS: Scalable Instance Selection for Class-Imbalanced Data Sets", IEEE TRANSACTIONS ON CYBERNETICS, VOL. 43, NO. 1, FEBRUARY 2013, pp.no: 332 – 346.

[81] Hualong Yu; Jun Ni, Yuanyuan Dan, SenXu "Mining and Integrating Reliable Decision Rules for Imbalanced Cancer Gene Expression Data Sets", Tsinghua Science and Technology, December 2012, 17(6): 666-673.

[82] Zhongbin Sun, Qinbao Song, and Xiaoyan Zhu "Using Coding-Based Ensemble Learning to Improve Software Defect Prediction", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 42, NO. 6, NOVEMBER 2012, pp.no"1806 – 1817.

[83] Yubin Park, JoydeepGhosh "Ensembles of α-Trees for Imbalanced Classification Problems", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 6, NO. 1, JANUARY 2007, Digital Object Indentifier 10.1109/TKDE.2012.255.

[84] Alberto Cano, Amelia Zafra, Sebastián Ventura "Weighted Data Gravitation Classification for Standard and Imbalanced Data" IEEE TRANSACTIONS ON CYBERNETICS, Digital Object Identifier 10.1109/TSMCB.2012.2227470.