IMBALANCED DATA LEARNING APPROACHES REVIEW

Mohamed Bekkar[1] and Dr. Taklit Akrouf Alitouche [2]

[1]ENSSEA, National School of Statistics and Applied Economics, Algiers, Algeria
moh.bekkar@gmail.com
taklitalitouche@yahoo.fr

ABSTRACT

The present work deals with a well-known problem in machine learning, that classes have generally skewed prior probabilities distribution. This situation of imbalanced data is a handicap when trying to identify the minority classes , usually more interesting one In real world applications. This paper is an attempt to list the different approachs proposed in scientific research to deal with the imbalanced data learning, as well a comparison between various applications cases performed on this subject.

KEYWORDS

imbalanced data, over-sampling, under-sampling, Bagging, Boosting, smote, Tlink, Random forests, cost-sensitive learing, offset entropy.

1. INTRODUCTION

Imbalances data learning problem has acquired in recent years a special interest from academics, industries, and research teams. considered as one of the top 10 Challenging problems in Data Mining [119],With great influx of attention devoted in scientific publication [117], due to the fact that this problem is faced in different applications areas, such as social sciences [116], credit card fraud detection [120] , taxes payment [92], customer retention[81], customer churn prediction[115], segmentation[99], medical diagnostic imaging[64], detection of oil spills from satellite images[121], environmental studies[70], bioinformatics [118], and more recently in improving mammography examinations for cancer detection [110].

When a model is trained on an imbalanced data set, it tends to show a strong bias to the majority class, since classic learning algorithms intend to maximize the overall prediction accuracy. Inductive classifiers are designed to minimize errors over the training instances, while Learning algorithms, can ignore classes containing few instances [8]. several methods was proposed to handle this kind of situation, from basic ones as sampling adjustment, to more complex like Algorithm modification.

we review in this article the proposed methods till date with comparison and assessment, starting by sampling adjustment, basic in section2 and advanced in section3.in subsequent cost-sensitive learning methods are detailed in section4, while section 5 describe the Features selection approaches, and final category about algorithm modification is analyzed in section6. finally, section 7 makes some comparison of applications among previous research and conclusion.

Table 1. Imbalanced Data learning Approaches.

<table>
<tr><th>SAMPLING METHODS</th><th>ENSEMBLE LEARNING METHODS</th></tr>
<tr><td>➢ 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</td><td>➢ 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</td></tr>
<tr><th>COST-SENSITIVE LEARNING</th><th>FEATURE SELECTION METHODS</th></tr>
<tr><td>• Direct cost-sensitive learning methods
• Methods for cost-sensitive meta-learning
• Cost-sensitive meta-learning thresholding methods MetCost
• Cost-sensitive meta-learning sampling methods</td><td>• Warpper
• PREE (Prediction Risk based feature selection for Easy Ensemble)</td></tr>
<tr><th colspan="2">ALGORITHMS MODIFICATION</th></tr>
<tr><td colspan="2">• Proposal for new splitting criteria DKM
• Adjusting the distribution reference in the tree
• Offset Entropy</td></tr>
</table>

2. BASIC SAMPLING METHODS

A common approach to deal with the imbalanced data is sample handling. The key idea is to pre-process the training set to minimize any differences between the classes. In other words, sampling methods alter the priors distribution of minority and majority class in the training set to obtain a more balanced number of instances in each class.

2.1. Under-Sampling and Over-Sampling

Two sampling methods are commonly used under-sampling (or down-sampling), and the oversampling (or up-sampling).Under Sampling is a non-heuristic method that removes instances of the majority class in order to balance the distribution of classes. The logic behind this is to try to balance the data set in order to overcome the idiosyncrasies of algorithms.

Japkowicz [4] suggest to distinguish between two different types of under-sampling; Random Under-Sampling RUS, that exclude randomly observations from majority class; and Focused Under-Sampling FUS, that exclude the majority class observations present on the borders between the two classes.

The over-sampling is an approach that increases the proportion of minority class by duplicating observations of this class. We distinguish, Random Over-Sampling which is based on a random selection of observation in duplication process, and Focused Over-Sampling that duplicate observations on the borders between the majority and minority class.

2.2. Assessment of under-Sampling and over-Sampling methods

The methods of under-sampling and over-sampling have been extensively studied in different research, particularly in learning with decision tree algorithm [3][10][5] [7] [1] .The findings of these studies were similar: the under-sampling leads to better results, while over-sampling produces little or no change in performance. However, no approach outperforms always the other, and it is difficult to determine a specific optimal rate of under-sampling or over-sampling which always leads to better results. Some studies have combined the two approaches, [11], use the over-sampling to improve the accuracy of classification, and under-sampling to reduce the size of the training set.

The main disadvantage of under-sampling is that it may exclude potentially useful data [9], which could be important for the model training process, and engender low performance of the classifier. While the over-sampling increases the size of the training set, in consequence the required time to build models. Even worse the addition of formal copies of instances can lead to a situation of over-fitting; in an extreme case, the classifier rules will be generated to cover one example duplicated several times[3]. As well the over-sampling does not introduce new observation, so it does not present a solution to the fundamental issue of lack of data; This explains why some studies have simply considered the over-sampling useless in improving model learning [3], and under-sampling seems to have an advantage in comparison with the over-sampling [2]

On the other hand, a more complex problem may appear with this two approach: knowing that the goal of learning in the statistical theory is to estimate the distribution of a statistic within the target population, we try to perform this through a representative random sample of the target population, the under-sampling and over-sampling modify the distribution within the sample, than will not be longer be considered random [8].

However, the disadvantages of these two approaches can be countered by more intelligent sampling strategies, or by the use of weights as an alternative; In the case of under-sampling a lowest weight is assigned to observation of the majority class; while in case of over-sampling, highest weight is given to the observations of the minority class, these alternatives was experimented in some studies as [16][12][14][13][15].

3. ADVANCED SAMPLING METHODS

3.1. Using Tomek Link

Tomek link abbreviated TLink was proposed by Ivan Tomek in 1976 [17] as a method of enhancing the Nearest-Neighbor Rule; Tlink algorithm is running as following :

- Having two observations x and y from different classes,
- The distance between these two observations is denoted $d(x, y)$,
- The pair (x, y) is called TLink if there is not an observation z as $d(x,z)<d(x,y)$ or $d(y,z)<d(x,y)$.

If two observations are a Tlink, so either one is a noise, or both are class boundaries.
The TLink can be used as a guide for under-sampling, or as a method of data cleaning, in the first case, the observations from to the majority class are removed, as shown in the Figure 1, while in the second both observation are discarded.

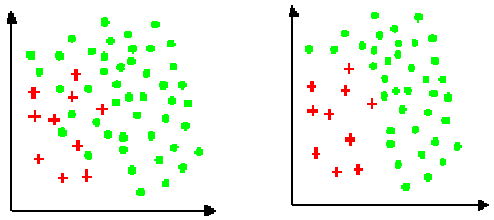


Figure 1. TLink application.

This procedure was tested in scientific research, Kubat and Matwin[21] who used TLink as a method of under-sampling by removing the observations of the majority class forming a TLink, since observations away from the border are more secure for learning, and less sensitive to noise. The TLink still relevant, this procedure was used in more recent research [22][20][23].

Another innovative approach based on TLink was proposed by Batista and Monard[18], using under-sampling in order to minimize the amount of potentially useful observations; elements of majority class are then classified, using TLink again as "safe "," borders "and" noise, they keep for learning only items classified as safe as the whole minority class.

3.2. The SMOTE approach

The SMOTE (Synthetic Minority Oversampling Technique) method is an advanced method of over-sampling introduced by Chawla & all [19]; essentially it aims to make the decision borders of minority class more general, and thus turned the issue over-fitting with basic over-sampling as detailed above.

The principle of this method is to generate new observations in the minority class by interpolating the existing ones. The algorithm is as following Figure 2:

- For each observation *x* of the minority class, identify its k-nearest neighbor,
- Select randomly a few neighbours (the number depends on the rate of over-sampling),
- Artificial observations are spread along the line joining the original observation *x* to its nearest neighbour.

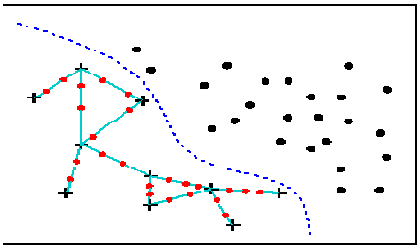
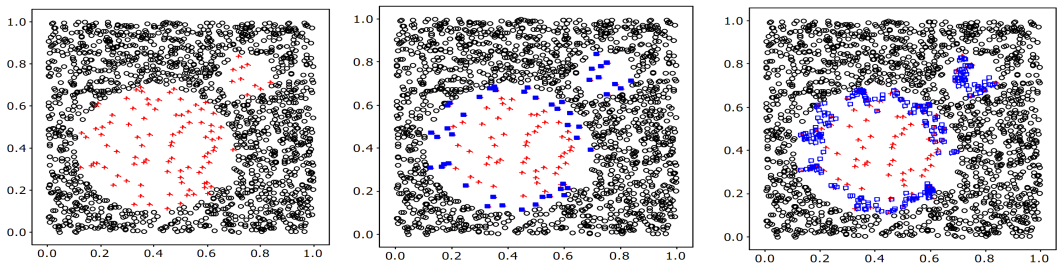


Figure 2. SMOTE application.

The effectiveness of this method was tested in Chawla [1], even some authors enhance the original concept like Han et al[25], who proposed borderline-SMOTE in which only the minority individuals close to the borders that are over-sampled.Figure3 illustrates the principle of applying SMOTE Borderline detailed as following:

(a) is the representation of the original data set, the black dots are the observations of the majority class, while red represents the minority class.(b) Identification of minority observations that are on the border with the majority class and are encircled in blue.(c) the final data set after applying SMOTE only to observations circled in blue.



a b c

Figure 3. SMOTE Borderline application.

The SMOTE Borderline produce better results than the original SMOTE, since observation located on the borders are the most likely ones to be misclassified.

Batista propose an approach combining SMOTE and TLink figure 4 detailed as following: (a) the initial imbalanced data set, (b) random over-sampling of the minority class using the SMOTE, (c) using TLink we detect the noise elements that appear on the majority class, (d) elimination of noise. This approach provid acceptable results; however, we observe that it expand significantly the boundaries of the minority class in detriment of the majority class.

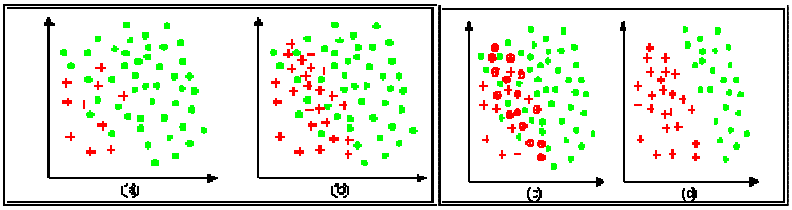


Figure 4. SMOTE and TLink combining approach.

3.3. One-Sided Selection OSS

The One-Sided Selection is an under-sampling approach proposed by Kubat and Matwin[21], in which the redundant observations, noise, and boundaries are identified and eliminated from the majority class; firstly we run TLink to locate noise and limits observations, then, closest nearest neighbor CNN to identify redundant observations, both of them are eliminated, remaining observations majority as well as minority class are used to reconstruct the training set.

The OSS was experienced in [25][29] , it is an efficient algorithm especially in the case of high imbalanced data, but it requires significant execution time and processing resources [28].

3.4. Neighbourhood Cleaning Rule NCL

Neighbourhood Cleaning Rule NCL [26] is under-sampling approach that use the Wilson's Edited Nearest Neighbor Rule ENN[30] to remove some observations from the majority class.

For reminder, the Wilson's Edited Nearest Neighbor Rule ENN identify the three nearest neighbours of each observation, then it eliminates all observations whose class labels differs from the 2/3 nearest neighbor.

The ENN algorithm removes noisy observations as it gathers the boundary points closely to the decision boundary, it avoids the over-fitting.

3.5. Bootstrap-based Over-sampling BootOS

The On BootOS is an over-sampling approach proposed by Zhu and Hovy[29].

As illustrated by Jo Japkowicz [28] the classic over-sampling, although it may reduce the imbalance between classes, it increases per cons the imbalance within classes due to the random duplication of minority observations. Bootstrap set avoids exact duplication of observations in the minority class, and secondly, it can provide a smoothing of distribution on the training samples, which can impair the problem of imbalance in the class generated by a basic over-sampling.

4. ENSEMBLE LEARNING METHODS

The use of ensemble learning methods was popular for a long time to boost classifier performance on a data set; these methods are based primarily on the work of Breiman[34], they was adapted to the context of unbalanced data across different research.

4.1. Bagging

The principle behind baaging-Bootstrap AGGregatING- is to combine classifiers by altering their inputs (learning observations) using iteration of bootstrap sampling technique on the training set [34] or by assigning weights to observations; at the end the prediction is given by majority voting process, which ensure that the errors will be ignored. The main contribution of baaging is to reduce the variance of the MSE (mean squared error), therefore, this method shows a significant improvement if it is associated to relatively unstable and inputs sensitive algorithms such as decision tree[49].

However, knowing that the bootstrap sampling is performed on all data regardless their class labels (majority or minority), the unbalanced distribution will be hold in each sub-sample pulled; that's the main failure associated with basic version of bagging. To be more useful in the context of imbalanced data, several variants were developed from the original one, which we quote the most recent:

- Asymmetric bagging [56]: in this method, in each bootstrap iteration, the full minority class is maintained, then a partition of equal size is derived from the majority class
- SMOTEBagging [58]: combination of SMOTE and bagging, this approach uses initially SMOTE to generate synthetic observations in the minority class, and then apply the bagging to the majority class in second stage.
- Over Bagging [58]: apply a random over-sampling on the minority class on each bootstrap iteration.
- Under Bagging [46]: apply a random under-sampling on the majority class on each bootstrap iteration.
- Roughly balanced bagging [43]: within this method weights are assigned to observations in order to ensure the balancing between classes on each bootstrap iteration. This variant was also experienced by [38] and returned suitable results.
- Lazy Bagging [60]: apply bagging only on the k nearest points using identified using the nearest neighbor algorithm.
- Random features selection [53]: random features selection combined with random sub-sample selection. On [53][48] demonstrate that combination of random space in the SVM can be an effective method for learning highly unbalanced financial data.

4.2. Boosting

The boosting finds its origin in the PAC (Probably Approximately Correct learning) which was proposed by [57][44] were first asked the question whether a low learning algorithm that performs just slightly better than random can be "boosted" in the PAC model to become a strong arbitrarily reliable algorithm. The idea remained so obscure until development Adabost, one of the first fully functional methods boosting implementation.

4.2.1 Adaboost

The Adaboost (Adaptive Boosting) is a method of iterative boosting introduced by Freund and Schapire[40][41]. It allocates variants weight to the observations during training. So, after each iteration, the weight of misclassified observations increases, while that of correctly classified decreases. Contained weights correction imposes on the learning process to focus more on misclassified in subsequent iterations. Given that in case of imbalanced data, most often the minority class is incorrectly classified, the boosting will therefore improve the accuracy of the obtained results [9].

The boosting, although it is as effective technique, easy to implement, it shows risk of over-training on outliers, which are often positioned at class boundaries limits, most probably incorrectly classified in the learning. This point was qualified in [38], which deducts following an application of bagging and boosting on the same training set, the bagging guarantee better performance, while boosting solution is a highly profitable and high-risk at once. We find a similar comparison in [51] applying the bagging and boosting learning in combination with decision trees.

Other applications was made through Adaboost or the general concept of boosting in various area such as fraud detection [55], text recognition [50], we also find [47] comparing the performance of AdaBoost, and SMOTE associated with SVM algorithm, conclude the superiority of AdaBoost in some cases. With [54][59] have built-in a cost component to the weighting phase, which emphasized the importance of the minority basis, and improves the accuracy rate. Noting that boosting may increase risk of over-training of minority class, Chawla [36] proposed SMOTEBoost algorithm adding artificial individuals by SMOTE method instead of simply increasing the weight of minority class observation.

The DataBoost-IM is also another algorithm developed by Guo et al, it combines data generation and boosting to improve the predictive accuracy within two classes, without focusing on minority at the expense of the majority class. Several other variants of AdaBoost have had proposed, including LP-Boost [37], AdaBoostReg, LPReg / QPReg-AdaBoost [25] and Nonlinear Boost [35].

However, Banfield and al [31] have demonstrated, from experiments performed on 57 data sets, that improvement of accuracy with boosting is limited to cases decision trees use, only when the training set is quite broad.

On the other hand it is often face with a debate between the boosting by re-sampling and boosting by re-weighting; Breiman [32] conclude at this point that boosting by re-sampling increases accuracy in the case of not pruned decision trees.

4.3. Random Forests

Random Forests (RF), proposed by Breiman [33], is a generalization of standard decision trees, based on bagging from a single training set of random not pruned decision trees.

Two sources of randomization are used sequentially in the algorithm:

- A random sample derived with replacement of the whole training set (bagging)
- Only a random subset of exogenous variables is used in the splits of each node during the construction phase.

The classification of a given observation is made by majority vote out of all trees results.

During the random bootstrap phases, about 1/3 of training set observations are not used in building decision tree [62]; these observations are called out-of-bag (OOB). For each tree, the OOB are used as test set, which allows generating an unbiased estimator of the error rate. Consequently random forests do not need a set of additional tests or cross-validation to evaluate its results.

This approach of random forests is more relevant and effective for highly multidimensional data sets, when randomization coupled with the multiplication of trees allows better exploration of the representation space [100]. it was used in case of imbalanced data as in predicting customer profitability and retention[81], segmentation using imbalanced data[99], ecological study[70], model learning in medical imaging[100][64]. Random forests have a significant performance improvement compared to standard decision trees such as C&RT, C5.0, and adaboost.

Chen et al [2] proposed two alternatives that are better suited to highly imbalanced data situation.

4.3.1 Balanced Random Forest BRF

in the case of using Random forest in severely imbalanced data, there is a strong probability that a bootstrap sample contains little or no observation of the minority class, which causes a decision tree with poor prediction performance on the minority class. The stratified bootstrap, which is the source of innovation provided by the balanced random forests, is a solution to this problem. The BRF algorithm is detailed as following:

- For each iteration in the random forest, take a bootstrap sample of minority class,
- Extract the same number of observations of the majority class; the sample is than balanced.
- Produce a tree from each bootstrap sample using a number of variables randomly selected.
- Aggregate predictions all using majority voting

4.3.1 Weighted Random Forest WRF

Another approach to make more appropriate random forests for highly skewed data learning is to include classes' weights; so we attribute an important penalty to misclassified minority cases. The weights are incorporated in two locations: in the tree induction process the weight are used for balances the Gini criteria used in the split; and in the leaf nodes of each tree, the weight considered again. The assignment of class to each leaf node is determined by "weighted majority vote."

5. COST-SENSITIVE LEARNING

The techniques listed until now acting on the distribution of classes in the training set to ensure a better balance; However, in several imbalanced data contexts such as fraud detection, intrusion prevention, medical diagnosis, or risk management, it is not only the distribution that is asymmetric but also the costs of misclassification, whereas most conventional learning algorithms

assume that misclassification within the same training set have identical costs. From these findings, the incorporation of costs in learning sequences has proven be a practical and effective solution to the unbalanced data issue

Based on actual cases, in medical diagnosis example, the cost of a false positive (false alarm) is limited to additional medical tests that the patient will suffer, while the cost of a false negative (misdiagnosis) will be fatal as potentially affected patient will be considered healthy. Likewise in fraud detection in banking transactions, false positives induce further infertile investigation, or the false negative result in exorbitant fraudulent transactions. finally, in customer relationship management, the false positive results in additional direct marketing costs (customer call, letter, visit to shop ...), while a false negative for the company is a loss of a potential customer, so less revenue, in such situations, it is important to accurately classify the minority class in order to reduce the overall cost.

5.1. Cost-Sensitive Learning Algorithm

To illustrate this method, we consider the following confusion matrix M associated to a cost matrix C

Table 2. Confusion and Cost Matrix

	Predicted Positive '1'	Predicted Negative '0'
Actual Positive '1'	TP (True positive)	FN (False Negative)
Actual Negative '0'	FP (False Positive)	TN (true Négative)

	Predicted Positive '1'	Predicted Negative '0'
Actual Positive '1'	C (1,1) TP	C (0,1) FN
Actual Negative '0'	C (1,0) FP	C (0,0) TN

Note that C (i, i) combined with TP or TN is usually considered an advantage or gain (more precisely denied cost) as the observation is correctly predicted in both cases.

Usually, the minority or rare class is considered as positive class. It is often more expensive misclassified a real positive as negative (FN), to classify a real negative as positive example (FP). In other words, the value C (0.1) assigned to FN is generally greater than that of C (1.0) associated to FP, and this is what we deduce from the above examples (medical diagnosis, bank fraud, customer relationship management).

Given the cost matrix, an example should be classified in the class with the minimum expected cost. The expected cost R (i | x) to classify an observation *x* in class *i* can be expressed as follows:

$$R(i \mid x) = \sum_{j} P(i \mid x) * C(i, j) \qquad (1)$$

Where $P(i \mid x)$ is the estimation of priori probability that observation *x* belong to a class i.

An observation *x* is predicted positive if and only if:

$$P(0|x)\, C(1,0) + P(1|x)\, C(1,1) \leq P(0|x)\, C(0,0) + P(1|x)\, C(0.1) \qquad (2)$$

Which is equivalent to $P(1|x)\,(C(1,0) - C(0,0)) \leq P(0|x)\,(C(0,1) - C(1,1))$ (3)

The initial cost of the matrix can be converted to a simpler, subtracting $C(1,1)$ of the first column, and $C(0,0)$ of the second column

Table 3. Simplified Cost Matrix

	Predicted Positive '1'	Predicted Negative '0'
Actual Positive '1'	0	C(0,1)- C(0,0)
Actual Negative '0'	C(1,0)- C(1,1)	0

According to the simplified costs matrix, the classifier predicts an observation *x* as positive if and only if P (0 | x) C (1,0) ≤ P (1 | x) C (0.1)

By projecting this matrix to the case of customer retention:

• The cost of a false positive *C(FP) = C (1,0) - C (1,1)* = Cost of a gift given to retain a customer
• The cost of a false negative *C(FN) = C (1,0) - C (1,1)* = loss of a customer

The total cost of misclassification = *FP*C (FP) + FN*C (FN)*

Since *P (0| x) = 1 - P (1| x),* we can get threshold *p** for classifying an observation *x* positive if *P(1|x) ≥ p*,* with the development:

$$P(0 \mid x)\, C(1,0) \le P(1 \mid x)\, C(0,1)..$$
$$(1 - P(1 \mid x))\, C(1,0) \le P(1 \mid x)\, C(0,1)..$$
$$C(1,0) \le P(1 \mid x)\, (C(0,1) + C(1,0))...$$
$$\frac{C(1,0)}{C(1,0) + C(0,1)} \le P(1 \mid x)...$$

In conclusion: $$p^* = \frac{C(1,0)}{C(1,0) + C(0,1)} = \frac{FP}{FP + FN} \qquad (4)$$

So if a cost-insensitive classifier may produce a posterior probability estimation *P(1| x)* for observations *xi*, we can make cost sensitive by selecting the classification threshold in terms of(1), and classify any observation as positive when *P (1|x) ≥ p**. This is the principle on which meta-learning costs sensitive algorithms are based such as "Relabeling". For other algorithms that do not offer the option to include costs directly (such as regression algorithms: Generalized linear, logistic, PLS ..), Elkan [74] specifies that we can make cost sensitive through a re-sampling performed as follows:

- Maintain all observations of the minority class
- Made a sub-sampling the majority class with the multiplier *C(1,0)/C(0,1) = FP/FN*

Knowing that generally *C(1.0) <C(0.1),* the multiplier is less than 1.

"Proportional sampling" is another alternative that consists to create a sample include minority and majority classes observations in accordance with the proportion

*P(1)*FN: P(0)*FP* (5)

Where *P(1)* and *P(0)* are the prior probabilities of positive and negative observations, in the initial data set.

In cost-sensitive learning, usually costs are not precisely known, we tend to use approximations or ratios of proportionality; on the other hand, as stated Domingos [72], the cost is not necessarily monetary value, it can be a waste of time or even the severity of the disease in some cases; Turney [101] provides a comprehensive overview of different types of costs, it grouped into ten categories, in addition to the cost of misclassification, there is the cost of data acquisition (for observations and attribute), calculation costs, human-machine interaction costs, testing costs, and so on. However, the cost of misclassification is most considered in literature [90], although some methods were developed to account for these varieties of costs, like [71] who use operating costs of coupling misclassification and test costs.

Ling and Sheng [85] propose to consider two families of applying the cost sensitive learning:

5.2. Direct cost-sensitive learning methods

That introduce and use direct costs in the learning algorithm, several experiments were carried out according to this approach, particularly by associating decision trees as [73][86][98][104], furthermore, some research, have analyzed the behaviour of decision trees under the cost-sensitive learning, in order to understand the interaction between costs and imbalance data such as [102][88][105].

5.3. Methods for cost-sensitive meta-learning

Methods of cost-sensitive meta-learning convert cost insensitive classifiers to cost-sensitive one. They operate as intermediate component that pre-processes training data, or post-processes output. These methods can be classified into two main categories: thresholding methods and sampling methods, based respectively on equations (4) and (5) mentioned above.

5.3.1 Cost-sensitive meta-learning thresholding methods

MetCost [72] is the most known algorithm in this family; the idea is that we affect each observation to the class that minimizes the final global cost, running as following:

- A set of models is generated on different bootstrap samples.
- The probability of belonging to each class is estimated for each individual using vote.
- Then each individual is assigned to the class that minimizes the total cost.
- The final result is obtained on the re-labelled data set.

Other algorithms were developed under this category, including [97][66][103].

5.3.1 Cost-sensitive meta-learning sampling methods

The sampling methods alter in the first class distribution of training data in terms of (5) and apply directly costs insensitive classifiers to the sampled data. Two main methods are positioned under this category: the Costing [113] and Weighting [61].

Other than these categories of costs sensitive learning, other emerging approaches was developed, which consist of:

- Combine costs with boosting: as in the case of AdaCl [54] or AdaCost[75]; the first one introduce the costs within the exponent of Adaboost weights update formula; while the second instead applying the cost elements directly, it uses a costs adjustment function that

increases aggressively the weight of costly errors misclassified observations, and decreases at the same time the weight correctly classified observations.

- Use advanced sampling methods in meta-learning cases: by "smart" re-sampling mainly through TLink and SMOTE as recently performed by Thai-Naghe and all [103].

6. FEATURE SELECTION METHODS

Is a relevant approach for large data sets exploration, particularly adopted with sets of high-dimensional data [77]. In the context of unbalanced data, the Feature selection was accommodated to select attributes that lead to greater separability between classes [67].

Warpper method proposed by Kohavi [79] is one of the first concrete feature selection applications in unbalanced data. As describe in figure () learning algorithm is executed so recurring over a separate part from the dataset, using different subsets of attributes; the attributes subset with the best performance evaluation is used as a final set to build final classifier over all learning set .

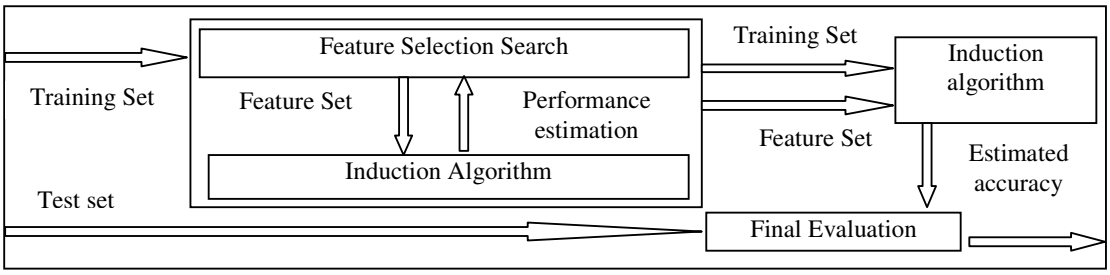


Figure 5. Warpper algorithm

In the same perspective, Zheng [111] propose an improved framework for features selection considering a selection for positive and negative classes independently, then combining them explicitly.

Feature selection was combined in some experiments with other methods, including the Ensemble learning methods, particularly in risk prediction through the PREE method [84] (Prediction Risk based feature selection for Easy Ensemble) we also find the combination with Tomek Link[112], or even thresholding [106]. Castillo and Serrano [68] present another innovative approach, they do not focus on the feature selection specially, but it fits into their work environment, they use a multi-classifier system strategy to build several classifiers, each classifier do its own feature selection based on genetic algorithm.

7. ALGORITHMS MODIFICATION

The aim of modifying algorithm is to provide adjustments on the learning algorithm (decision tree, regression, factor analysis...) in order to make them more relevant and appropriate to imbalance data situations. This approach is used mainly with Decision Tree and SVM; however few studies were done through this approach, since the options and opportunity within are limited compared to those detailed so far.

As a reminder, the decision trees are based on information gain criteria to split each node parent in the tree; the Gini index, Kh-2, and Shannon entropy are the gain criteria usually used with trees C&RT, CHAID and C5.0 respectively. We can list three methods classified under this approach.

7.1. Proposal for new splitting criteria

A more recent splitting criteria was proposed by Dietterich and al [69] known as DKM, which is more sensitive to the asymmetry than classical entropy. Various authors such as [73][76][109] have experienced as the DKM in decision tree, and aquire a performance improvement in most of imbalanced data cases.

Following same approach, Cieslak [38] proposes to use the Hellinger distance as splitting criteria, he develop a decision tree called HDDT, which presents the best performance out of conventional algorithms, and even demonstrated superiority over the DKM.

7.2. Adjusting the distribution reference in the tree

The adjustment of the distribution reference - implicitly assumed to be uniform - was proposed in the literature through the involvement index developed by Gras and al [78] as a measure of quality of association rules. For a given rule, it is defined from the cons-examples; in the case of decision trees, it is in each leaf node the number of cases that are not matching the assigned category; thus, instead of measuring deviations from uniform distribution to assign node classes, it is measured against this initial training set distribution. This technique was tested in [95][93].

7.3. Offset Entropy

Introduced per Lallich et al [80] and Marcellin et al [91], Off Center Entropy (OCE) idea is to consider the prior distribution of classes in the partitioning criteria.

This approach comprises moving the standard maximum entropy to the point where it takes its maximum according to the classes' distribution, allowing the user to determine the point of maximum uncertainty. The effect of OCE is exposed in Figure 2, where at left we have the classical probability distribution of entropy uncertainty, and at right offset entropy modified to fit with a distribution of two classes (90 %, 10%) in the training set.

This approach was experienced mainly in the work of Lenca et al [82] and Zighed et al [114].

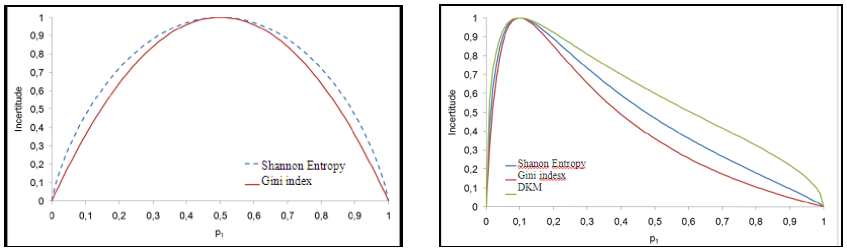


Figure 6. Off Set Entropy illustration

7. DISCUSSION AND CONCLUSION

The multitude of methods described so far shows a part of the richness of the subject, and secondly the difference and disparity logical resolution adopted by researchers, this gives rise to a more complex question: What is the best approach to use?.

Experiments done on different methods conclude with ambiguous results: while Anand et al[63], certified by Li et al[15] opt for sampling methods as optimal solution, we observe on the other front McCarthy et al [65]in agreement with Liu et al[88] on the superiority of the cost sensitive learning; while Quinlan[94] and Thomas [100] are approving ensemble learning methods; on the other hand Cieslak[38] and Marcellin[90] defend the algorithm modification approaches.

Looking back over the families of methods distinctively, we find that the sampling methods, which are based on remove or duplicate some observations, are facing firstly to the difficulty of distinguishing between minority and noise observations as specified by Kotasiantis et al[8], although this point is trying to be solved partly in advanced sampling methods (such as Tlink or SMOTE), it remains difficult to identify objectively the over-sampling (or sub-sampling) rate; as a matter of principle is revealed: at what level is it acceptable to delete, duplicate or generate observations in the learning sample?, noting that these samples are, in some cases, a partial representation of phenomenon mainly in Social Science or customer behaviour analysis with influencing factors that cannot be exhaustively surveyed. However, these critics do not exclude the main advantage of sampling methods that are easily transportable and can be associated with the majority of statistical learning algorithms. counter to costs sensitive learning, that even it is based on more robust thinking sense, it operate exclusively with limited progressive learning method such as decision trees, neural networks, or some regression model; and similarly, the of accurate costs determination is another major drawback associated with the cost-sensitive learning, since in proportion to the importance given to costs difference in learning, it may cause a significant volatility results by different user on the same training set.

The ensemble learning methods have the advantage to require less of setting, and modest user interaction, as they occur in successive iterations, which makes them effectively with large volumes of data. However, they are limited to use with decision trees which constitute the basis of ensemble learning methods.

Finally, the algorithms modification methods, despite being effective even with small sample sizes as demonstrated by Lallich et al[80], they are suffering from the development complexity and limited options available in this category.

To summarize, each family of methods offers advantages and show disadvantages which varied depending on the context and scope of training data; well as in some cases these different approach are aligned as conclude Maloof [89] by observing that the sampling methods, cost-sensitive learning and Off Center Entropy have similar effects. In conclusion, the comparison of different concepts reminder us the famous theory of Wolpert's[107][108] “No free Lunch Theorems” that assume “the learning algorithms cannot be universally superior”, therefore, in imbalanced data learning, the unique optimal solution does not exist, will the best solution depend on the context of learned data.

REFERENCES

[1] Chawla N.V. (2003) “C4.5 and Imbalanced Data Sets: Investigating the Effect of Sampling Method, Probabilistic Estimate, and Decision Tree Structure”, Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II.

[2] Chen, C., Liaw, A., Breiman, L. (2004) “Using Random Forest to Learn Imbalanced Data”, Tech. Rep. 666, University of California, Berkeley.

[3] Drummond, C., Holte, R.C. (2003) “C4.5, class imbalance, and cost sensitivity: Why undersampling beats over-sampling”, In Workshop on Learning from Imbalanced Datasets II, held in conjunction with ICML 2003.

[4] Japkowicz, N, (2000) “Learning from imbalanced data sets: a comparison of various strategies”, AAAI Tech Report WS-00-05. AAAI.

[5] Japkowicz N, Stephen S, (2002) “The class imbalance problem: A systematic study”, Journal Intelligent Data Analysis Volume 6 Issue 5, Pages 429 – 449

[6] Jonathan Burez, Dirk Van den Poel, (2009) “Handling class imbalance in customer churn prediction”, Expert Syst. Appl. 36(3): 4626-4636.

[7] Joshi M, V., Kumar, V., Agarwal R, C. (2001) “Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction”, ACM SIGMOD May 21-24.

[8] Kotsiantis S., Kanellopoulos D., Pintelas P., (2006), “Handling imbalanced datasets: A review”, GESTS International Transactions on Computer Science and Engineering, Vol.30 (1), pp. 25-36.

[9] Weiss G M, (2004)"Mining with rarity: A unifying framework", SIGKDD Explorations, 6:7-9
[10] Weiss G M, (2003) "The Effect of Small Disjuncts and Class Distribution on Decision Tree Learning", PhD Dissertation, Department of Computer Science, Rutgers University.
[11] Zhang J, Mani I, (2003) "kNN approach to unbalanced data distributions: A case study involving information extraction", In Proceedings of the ICML'2003.
[12] Alejo R., Garcia V., Sotoca J, M., Mollineda R, A., Senchez J. S, (2007),"Improving the performance of the RBF neural networks trained with imbalanced samples", Lecture Notes in Computer Science, Springer-Verlag Berlin.
[13] Anand A., Pugalenthi G., Fogel G, B., Suganthan, P, N., (2010), "An approach for classification of highly imbalanced data using weighting and under-sampling", School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.
[14] Fu X., Wang L., Chua K, S., & Chu, F (2002), "Training RBF neural networks on unbalanced data", In Neural Information Processing, Inst of High Performance Computer, Singapore.
[15] Li Q., Wang Y., Bryant S H., (2009),"A novel method for mining highly imbalanced high-throughput screening data in PubChem", Bioinformatics 25 (24): 3310-3316.
[16] Nguyen G, H., Bouzerdoum, A., & Phung S L, (2008) "A Supervised Learning Approach for Imbalanced Data Sets"; 19th International Conference on Pattern Recognition (ICPR 2008), IEEE, Dec 8-11, Florida, USA..
[17] Tomek Ivan, (1976) "An Experiment with the Edited Nearest-Neighbor Rule", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 6, No. 6, pp. 448-452.
[18] Batista, G., Carvalho, A., Monard, M, C. (2000),"Applying Onesided Selection to Unbalanced Datasets"; In Proceedings of MICAI, 315–325. Springer Verlag.
[19] Chawla N, V., Bowyer K, W., Hall L, O., Kegelmeyer W, P., (2002) "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research 16, P 321–357.
[20] Gu J, Zhou Y and Zuo X, (2007), "Making Class Bias Useful: A Strategy of Learning from Imbalanced Data", Lecture Notes in Computer Science, Intelligent Data Engineering and Automated Learning - IDEAL.
[21] Kubat M, Matwin S, (1997),"Addressing the curse of imbalanced training sets: One-sided selection"; In Douglas H. Fisher, editor, ICML, pages 179–186. Morgan Kaufmann.
[22] Thai-Nghe ,N., Do T, N., Schmidt-Thieme, L.,(2000),"Learning Optimal Threshold on Resampling Data to Deal with Class Imbalance", Proc. of the 8th IEEE International Conference on Computing and Communication Technologies (RIVF)
[23] Yueai Z and Junjie C, (2009), "Application of Unbalanced Data Approach to Network Intrusion Detection", 1st Inter Workshop on Database Technology and Applications, DBTA, IEEE.
[24] Batista, G., Ronaldo, C., Monard, M, C., (2004), "A study of the behavior of several methods for balancing machine learning training data", ACM SIGKDD Explorations -Special issue on learning from imbalanced datasets, Vol 6 Issue 1.
[25] Han, H., Wang, W,Y., Mao, B, H., (2005), "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning" ; Proc. Int'l Conf. Intelligent Computing, 878-887.
[26] Laurikkala J, (2001), "Improving Identification of Difficult Small Classes by Balancing Class Distribution", AIME, LNAI 2101, pp. 63–66, Springer-Verlag Berlin Heidelberg.
[27] Suman Sanjeev, Laddhad Kamlesh, Deshmukh Unmesh, (2005), "Methods for Handling Highly Skewed Datasets", Indian Institute of Technology Bombay.
[28] Taeho Jo and Japkowicz N, (2004), "Class imbalances versus small disjuncts"; ACM SIGKDD Explorations -Special issue on learning from imbalanced datasets, Vol 6 Issue 1.
[29] Zhu Jingbo, Hovy Eduard, (2007) "Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem"; Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 783–790, Prague.
[30] Wilson D L (1972), "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data", IEEE Transactions on Systems, Man, and Communications 2, 3, 408-421.
[31] Banfield, R., Hall, L, O., Bowyer, K, W., Kegelmeyer, W, P., (2007), "A Comparison of Decision Tree Ensemble Creation Techniques"; IEEE Trans. on Pattern Analysis and Machine Intelligence, 29(1):832–844.
[32] Breiman L (1998),"Rejoinder to the paper 'Arcing Classifiers by Leo Breiman",Annals of Statistics, 26(2):841–849.
[33] Breiman L, (2001), "Random forest". Machine Learning, 45, 5–32.
[34] Breiman L, (1996), "Bagging Predictors", Machine Learning, Kluwer Academic Publishers 24(2):123–140.

[35] Chen C, S., Tsai C, M., Chen J, H., Chen C, P., (2004),"Nonlinear Boost", Tech. Rep. No.TR-IIS-04-001, Taipei, Taiwan: Institute of Information Science, Academia Sinica.
[36] Chawla N,V., Lazarevic A., Hall L, O., Bowyer K, W., (2003),"Smoteboost: Improving prediction of the minority class in boosting", In 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, pages 107–119, Dubrovnik, Croatia.
[37] Demiriz A., Bennett K, P., Shawe-Taylor J., (2002),"Linear Programming Boosting via Column Generation", In Machine Learning , Vol 46, pp. 225–254; Hingham, MA, USA
[38] Cieslak David A, (2009), "finding problems in, proposing solutions to, and performing analysis on imbalanced data", Phd Dissertation, University of Notre Dame, Indiana.
[39] Efron B (1979), "Bootstrap Methods: Another Look at the Jackknife", In The Annals of Statistics, vol. 7, no 1, p. 1-26.
[40] Freund Yoav, Schapire Robert E, (1999), "A Short Introduction to Boosting"; Journal of Japanese Society for Artificial Intelligence, 14(5):771-780.
[41] Freund Y, Schapire R,(1996), "Experiments with a new Boosting Algorithm"; In Proc. 13th International Conference on Machine Learning 148–146. San Francisco: Morgan Kaufmann.
[42] Guo H, Viktor H, (2004),"Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach", Conference SIGKDD Explorations ,ACM,6(1).
[43] Hido S., Kashima H., (2008), "Roughly Balanced Bagging for Imbalanced Data", In SIAM International Conference on Data Mining (SDM), pp 143–152.
[44] Kearns M., Valiant L, G., (1994), "Cryptographic limitations on learning Boolean formulae and finite automata"; Journal of the Association for Computing Machinery, 41(1):67–95.
[45] Kearns M., Valiant L, G.,(1988), "Learning Boolean formulae or finite automata is as hard as factoring"; Technical Report TR-14-88, Harvard University Aiken Computation Laboratory.
[46] Liu Y., Chawla N, V., Harper M., Shriberg E., Stolcke A.,(2006), "A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech"; Computer Speech and Language, 20, 468-494.
[47] Li X., Wang L., Sung E., (2008), "AdaBoost with SVM-based component classifiers", Engineering Applications of Artificial Intelligence, 21(5):785 – 795.
[48] Machon Gonzalez I, Lopez-Garcia H,(2008), "Using Multiple SVM Models for Unbalanced Credit Scoring Data Sets"; In Artificial Neural Networks, p 642–651.
[49] Mennicke Jörg (2008),"Classifier Learning for Imbalanced Data", Vdm Verlag.
[50] Pio N., Sebastiani F., Sperduti A., (2003),"Discretizing Continuous Attributes in AdaBoost for Text Categorization" ; Proc. of ECIR-03, 25th , Pisa, 320-334.
[51] Quinlan J R; "Bagging, Boosting, and C4.5". In AAAI 06: Proceedings of the 13th National Conference on Artificial Intelligence (Vol. 2, pp. 725–730). Portland, OR ; 2006.
[52] Rätsch, G., Onoda, T., Müller, K, R.,(2001),"Soft Margins for AdaBoost", Machine Learning,42(3), 287–320.
[53] Schebesch K, Stecking R, (2008) ,"Using Multiple SVM Models for Unbalanced Credit Scoring Data Sets"; In Classification, Data Analysis, and Knowledge Organization, p515–522.
[54] Sun Y., Kamel M., Wong A., Wang Y., (2007),"Cost-Sensitive Boosting for Classification of Imbalanced Data", Pattern Recognition, 40(12):3358–3378.
[55] Stijn V., Derrig R, A., Dedene G., (2002), "Boosting Naive Bayes for Claim Fraud Diagnosis"; in Lecture Notes in Computer Science 2454, Berlin: Springer.
[56] Tao D., Tang X., Li X., Wu X., (2006), "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(7):1088–1099.
[57] Valiant L G, (1984),"A theory of the learnable", Communication of the ACM, 27(11).
[58] Wang S & Yao X (2009), "Diversity Analysis on Imbalanced Data Sets by using Ensemble Models", In Proc. of The IEEE Symposium on Computational Intelligence and Data Mining.
[59] Wang B X, Japkowicz N, (2008), "Boosting Support Vector Machines for Imbalanced Data Sets", In Foundations of Intelligent Systems, p38–47.
[60] Zhu X & Yang Y, (2008),"A Lazy Bagging Approach to Classification", Pattern Recognition, 41(10):2980 – 2992.
[61] Alejo, R., Gracia, V., Sotoca, J., Mollineda, R., Sanchez, J., (2007), "Improving the Performance of the RBF Neural Networks Trained with Imbalanced Samples", In Proceedings of Computational and Ambient Intelligence, pp 162-169, San Sebastian, Spain,.
[62] Archer K J and Kimes R V,(2008),"Empirical characterization of Random Forest variable importance measures"; Computational Statistics and Data Analysis, 52, 2249-2260.

[63] Anand, A., Pugalenthi, G., Fogel, GB, Suganthan, P,N., (2010), "An approach for classification of highly imbalanced data using weighting and under-sampling", Amino Acids, Vol. 39(5).
[64] Bosch, A., Zisserman, X., Muñoz (2007), "Image Classification Using Random Forests and Ferns", IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil.
[65] McCarthy, K., Zabar, B., Weiss, G, M.,(2005),"Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes?", Proc. Int'l Workshop Utility-Based Data Mining, pp 69-77
[66] Chai, X., Deng, L., Yang, Q., Ling, C, X.,(2004),"Test-Cost Sensitive Naïve Bayesian Classification",In Proceedings of the Fourth IEEE International Conference on Data Mining,UK.
[67] Chawla N, V., Japkowicz N., (2004),« Editorial: Special Issue on Learning from Imbalanced Data Sets", SIGKDD Explorations, Volume 6, Issue 1 - pp. 1.
[68] Castillo M and Serrano J ,(2004), "A multistrategy approach for digital text categorization from imbalanced documents", SIGKDD Explorations, 6(1):70-79.
[69] Dietterich, T., Kearns, M., Mansour, Y.,(1996),"Applying the weak learning framework to understand and improve C4.5", In Proc 13th International Conference on Machine Learning, pp 96–104. Morgan Kaufmann,.
[70] Cutler D R et al, (2007),"Random Forests For Classification In Ecology", Ecology Review, 88(11), pp. 2783–2792; Ecological Society of America.
[71] Davis Jason V et al, (2006), "Cost-sensitive decision tree learning for forensic classification", ECML, volume 4212 of Lecture Notes in Computer Science, pp. 622–629, Springer.
[72] Domingos P (1999),"MetaCost: A general method for making classifiers cost-sensitive",In Proceed of the 5th International Conference on Knowledge Discovery and Data Mining, 155-164.
[73] Drummond C, Holte R, (2000), "Exploiting the cost (in)sensitivity of decision tree splitting Criteria", In Proceedings of the 17th International Conference on Machine Learning, 239-246
[74] Elkan C, (2001), "The foundations of cost-senstive learning", 17th International Joint Conference on Artificial Intelligence, pp. 973–978.
[75] Fan, W., Stolfo, S, J., Zhang, J., Chan, P, K., (1999), "AdaCost: Misclassification Cost-Sensitive Boosting" , Proc Int'l Conf. Machine Learning, pp. 97-105.
[76] Flach P. A,(2003),"The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics", In International Conference on Machine Learning (ICML).
[77] Guyon I and Elissee A,(2003),"An introduction to variable and feature selection", Journal of Machine Learning Research, 3:1157-1182.
[78] Gras R, Couturier R, Blanchard J, Briand H, Kuntz P, Peter P, (2004), "Quelques critères pour une mesure de qualité de règles d'association", Revue des nouvelles technologies de l'information RNTI E-1, 3–30.
[79] Kohavi, R., John, G., (1998),"The wrapper approach", In Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers, pp.33-50.
[80] Lallich, S., Lenca, P., Vaillant, B., (2007),"Construction d'une entropie décentrée pour l'apprentissage supervisé", 3ème Atelier QDC-EGC 07, Namur, Belgique, pp 45–54.
[81] Larivière B and Van den Poel D, (2005),"Predicting customer retention and profitability by using random forests and regression forests techniques", Expert Systems With Applications, 29.
[82] Lenca, P., Lallich, S., Do, T., Pham, N, K., (2008), "A comparison of different off-centered entropies to deal with class imbalance for decision trees", In Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD, Osaka, Japan, pp. 634–643.
[83] Liu Tian-Yu, (2009),"EasyEnsemble and Feature Selection for Imbalance Data Sets", International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing.
[84] Liu, Tian-yu., Li, Guo-zheng., You, Ming-yu., (2009),"Feature Selection for Imbalanced Fault Diagnosis", Journal of Chinese Computer Systems, 05.
[85] Ling Charles X, Sheng Victor S, (2008),"Cost-Sensitive Learning and the Class Imbalance Problem", Encyclopedia of Machine Learning. C. Sammut (Ed.),Springer.
[86] Ling, C, X., Yang, Q., Wang, J., Zhang, S.,(2004), "Decision Trees with Minimal Costs", In Proceedings of 2004 International Conference on Machine Learning (ICML'2004).
[88] Liu, X, Y., Zhou Z, H, (2006),"The influence of class imbalance on cost sensitive learning: An empirical study", in Proceedings of the 6th ICDM. Washington, DC, USA, pp 970–974.
[89] Maloof, M, (2003),"Learning when data sets are imbalanced and when costs are unequal and unknown", In ICML Workshop on Learning from Imbalanced Data Sets II.
[90] Marcellin Simon, (2008), "Arbres de décision en situation d'asymétrie", Phd Thesis informatique, Université Lumière Lyon II, France.

[91] Marcellin S, Zighed D A, Ritschard G, (2006), "Detection of breast cancer using an asymmetric entropy measure", In COMPSTAT-Proced. in Computational Statistics, pp.975–982. Springer.
[92] Miguel Pironet San-Bento Almeida, (2009),"Classification for Fraud Detection with Social Network Analysis", Masters Degree Dissertation, Engenharia Informática e de Computadores; Instituto Superior Téchnico, Universidade Técnica de Lisboa, Portugal.
[93] Pisetta, V., Ritschard, G., Zighed, D, A., (2007), "Choix des conclusions et validation des règles issues d'arbres de classification", Extraction et Gestion des Connaissances (EGC 2007), Volume E-9 of Revue des nouvelles technologies de l'information RNTI, pp. 485–496. Cépaduès,.
[94] Quinlan J,R., (2006),"Bagging, Boosting, and C4.5", In AAAI 06: Proceedings of the 13th National Conference on Artificial Intelligence (Vol. 2, pp. 725–730). Portland.
[95] Ritschard, G. (2005) ," De l'usage de la statistique implicative dans les arbres de classification", Actes des 3eme Rencontres Internationale ASI Analyse Statistique Implicative, Palermo, N.15
[96] Ritschard, G., Zighed, D, A., Marcellin S., (2007), "Données déséquilibrées, entropie décentrée et indice d'implication", 4èmes Rencontres Inter Analyse Statistique Implicative ,España
[97] Sheng, V, S., Ling, C, X., (2006), "Thresholding for Making Classifiers Cost-sensitive", In Proceedings of the 21st National Conference on Artificial Intelligence, 476-481, Boston.
[98] Sheng, S., Ling, C, X., Yang, Q.,(2005),"Simple test strategies for cost sensitive decision trees"; in ECML,LNAI 3720, pp. 365 – 376, Springer
[99] Schroff , F., Criminisi, A., Zisserman, A.,(2008),"Object Class Segmentation using Random Forests", Dept. of Engineering Science, University of Oxford.
[100] Thomas J,(2009), « Apprentissage supervisé de données déséquilibrées par forêt aléatoire » ; Phd Thesis Informatique, Université Lumière Lyon 2, France
[101] Turney P D, (2000),"Types of cost in inductive concept learning", In Proceed of the Workshop on Cost-Sensitive Learning at the 7th Inter Conference on Machine Learning, California.
[102] Ting K M, (2002), "A study on the effect of class distribution using cost sensitive learning", in Discovery Science, Lecture Notes in Computer Science Volume 2534, pp 98-112, Springer.
[103] Thai-Nghe, Nguyen., Gantner, Zeno., Schmidt-Thieme, Lars.,(2010), "Cost-Sensitive Learning Methods for Imbalanced Data", Inf. Syst. & Machine Learning Lab, Univ. Hildesheim, Germany
[104] Van Hulse, J., Khoshgoftaar T, M., Napolitano, A., (2007), "Experimental perspectives on learning from imbalanced data", in Proceedings of 24th ICML. ACM, pp. 935–942.
[105] Weiss, G, M., Provost, F.,(2003),"Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction", Journal of Artificial Intelligence Research 19, 315-354.
[106] Wasikowski, Mike., Chen, Xue-wen., (2010),"Combating the Small Sample Class Imbalance Problem Using Feature Selection", IEEE Transactions on Knowledge and Data Engineering, pp.1388-1400,vol. 22 no. 10.
[107] Wolpert David, (1996),"The Lack of A Priori Distinctions between Learning Algorithms", Neural Computation, pp. 1341-1390.
[108] Wolpert, D., Macready, W, G., (1997),"No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67.
[109] Zadrozny Bianca, Elkan Charles,(2001), "Learning and making decisions when costs and probabilities are both unknown"; Proceedings of 7th ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, USA.
[110] Zadrozny, B., Langford, J., Abe, N., (2003), "Cost-sensitive learning by Cost-Proportionate instance Weighting", In Proceedings of the 3th International Conference on Data Mining.
[111] Zheng, Z., Wu, X., Srihari, R., (2004),"Feature selection for text categorization on imbalanced data", SIGKDD Explorations, 6(1):80-89.
[112] Zhu, Quanyin., Cao, Suqun., (2009), "A Novel Classifier-Independent Feature Selection Algorithm for Imbalanced Datasets", 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing.
[113] Zadrozny B and Elkan C, (2001), "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers", In Proc 18th International Conf. on Machine Learning, pp 609–616, Morgan Kaufmann, San Francisco, CA,.
[114] Zighed, D, A., Marcellin, S., Ritschard G., (2007), "Mesure d'entropie asymétrique et consistante" , In EGC 2007, Volume E-9 of RNTI, pp.81–86.
[115] Bekkar M, (2009),"Développement d'un modèle de prédiction du churn clientèle en télécommunication", Master 2 theis stat et économétrie, Unive Toulouse 1 Sciences Sociales.
[116] Bressoux P, (2008),"Modélisation statistique appliquée aux sciences sociales", De Boeck, Bruxelle , p326.

[117] Haibo He and Edwardo A. Garcia, (2009),"Learning from Imbalanced Data", IEEE transactions on knowledge and data engineering, vol. 1, no 9.
[118] Meng Y A, Yu Y, Cupples L A et al, (2009),"Performance of random forest when SNPs are in linkage disequilibrium", BMC Bioinformatics,10, 78.
[119] Qiang Yang, Xindong Wu,(2006),"10 Challenging Problems in Data Mining Research", International Journal of Information Technology & Decision Making, Vol. 5, No. 4 597–604.
[120] Shen A, Tong R, Deng Y,(2007),"Application of Classification Models on Credit Card Fraud Detection", International Conference on Service Systems and Service Management, IEEE.
[121] Kubat M., Holte R C., Matwin S., Kohavi R., Provost F., (1998),"Machine learning for the detection of oil spills in satellite radar images", in Machine Learning, pp. 195–215.

Authors

Mr. Mohamed Bekkar is a Phd student in ENSSEA, Algiers. He holds a Msc in Economy from ENSSEA, Msc in statistics form Toulouse I university, France; and Msc in Entreprise Administration from IAE Paris, France. Currently he is working as Predictive Analytics Expert with a major telecom operator in Middle East. His current research interests include Imbalanced data learning, social network analysis, Social media influencer ranking and Big data integration within industry

Pr. Taklit Akrouf Alitouche holds a PhD in Statistics from Plekhanov Russian University of Economics, currently she is a Professor of statistics in the ENSSEA (Ecole Nationale Supérieure de Statistique et d'Economie Appliquée),Algiers. She supervised several Msc and doctoral thesis. His research interests include statistical processing of survey basis, and survey methods enhancement.