New Oversampling Approaches Based on Polynomial Fitting for Imbalanced Data Sets

Sami Gazzah
Sami_gazzah@yahoo.fr

Najoua Essoukri Ben Amara
National Engineering School of Sousse-Tunisia
Najoua.benamara@eniso.rnu.tn

Abstract

In classification tasks, class-modular strategy has been widely used. It has outperformed classical strategy for pattern classification task in many applications [1]. However, in some modular architecture, such as one against all in support vector machines classifier, the training dataset for one class risks to heavily outnumber the other classes. In this challenging situation, the trained classifier will accurately classify the majority class; nevertheless, it marginalizes the minority class. As a result, True Negatives rate (TNr) will be very high while the True Positives rate (TPr) will be low. The main goal of this work is to improve TPr without much sacrifice in TNr. In this paper, we propose oversampling the minority class using polynomial fitting functions. Four new approaches were proposed: star topology, bus topology, polynomial curve topology and mesh topology. Star and mesh topologies approach had led to the best performances.

1. Introduction

During recent years, classification systems have achieved an advanced degree of maturity and great success in practical applications. Various classification methods and architectures have been proposed. A steadily growing interest in improving these classification models and their implementation in numerous areas has received more attention. In this context, class-modular architectures concept has gained a widely success against classical architectures for pattern classification task. In fact, it has been shown that the modular architecture offers superiority in terms of convergence and recognition capability over non-modular network [1, 2, 3, 4]. Furthermore, new multiclass conceptual learning systems have been developed using modularity such as one against all and one against one in SVM-support vector machine classifier. In those strategies, a complex single "*i*-classification" task is decomposed into "*i*" two-classification subtasks.

However, in some modular architecture (SVM-one against all for example) the training dataset for one class risks to heavily outnumber the other. In this case, the performance of the classification systems drops significantly and classical overall performance measures will not be significant if they don't take into consideration the relative distribution of each class. To overcome such limitations, a rebalance method seems to be a promising solution. This step aims to reinforce the training system by emphasizing the minority instance according to two ways: acting in a preprocessing stage to rebalance the training data or acting in the training stage at algorithmic level.

In this paper, we investigate the effects of imbalance training data on an earlier writer identification system developed using off-line Arabic handwriting in multiclass SVM one-against-all system [5]. The proposed earlier system is composed mainly of four modules: text image acquisition, image pre-processing, feature extraction and writer style identification (figure1).

The identification system is based on the combination of global and structural features. A one-against-all SVM classifier has been implemented; a complex single 60-classification task is devoted to 60 experts instead of only one. Each classifier is responsible for only 2-classification subtasks. During the phase of training, each classifier is trained with both text samples of the considered writer (authorship) and all 59 others (non-authorship). In this case, each expert will be trained by a dataset having a ratio of negative instances to positive ones about 59:1; hence, each one of the sixty classes is initially almost balanced. In such situation, average results have shown that the test samples belonging to the authorship class

978-0-7695-3337-7/08 $25.00 © 2008 IEEE
DOI 10.1109/DAS.2008.74

have been badly classified (56%) and then those belonging to the non-authorship class have been almost correctly classified (99%). The overall accuracy has been overwhelmed by non-authorship feature vectors correctly classified (98%).

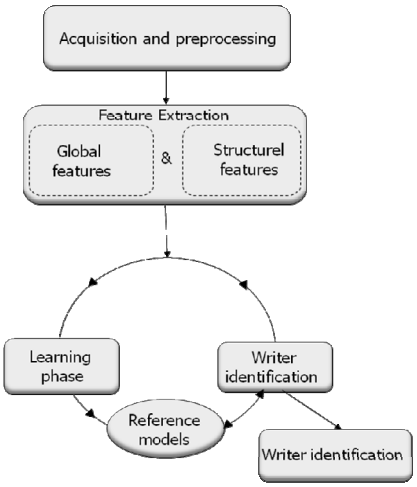

Figure 1. Flow chart of the proposed writer identification system in its first version [5]

In this work, we propose new oversampling approaches based on polynomial fitting function in order to improve the authorship instances correctly classified whilst preserving as much as possible the non-authorship instances correctly classified. In the following, we refer to "the minority class" when we consider positive class which encloses authorship data samples and to "the majority class" as a negative class for non-authorship feature vectors.

This paper is organized as follows: section 2 gives an overview for related works on resampling methods; section 3 introduces some common performance measure tools; section 4 presents detail about our proposed approaches; and section 5 presents experimental results for different architectures. Concluding remarks and some perspectives are addressed in the last section.

2. Literature survey

The synthesis of some previous works (cf table 2) has shown that the use of imbalanced data for the training of various classification systems such as SVM, LDA, C4.5, KNN, neural networks, etc, deteriorates considerably the classifier performances. In that situation, the minority instances are quite often neglected and the overall accuracy is dominated by the majority instances correctly classified [6]. A vast number of resampling methods have been proposed to solve this problem at two levels: algorithmic and data levels. Algorithmic approaches act mainly on classifier level; whereas, data level approaches act in feature space in order to oversample the minority class or to undersample the majority one (figure 2).

2.1. Algorithmic methods

Several research studies have manipulated the classifier architecture or parameters for resampling not equally represented data. This approach aims to enhance the classifier to pay more attention to the minority instances. In this paper we limit our study on the researches conducted using SVM classifier.

The literature [6, 7] has shown that the SVM classifier can be inefficient in determining the class boundary when the training instances are skewed. Several approaches were proposed to adjust the boundary of the SVM classifier. In [8] the authors use a general parameter optimization framework for one-class SVM classifier. They use only the minority instances for training the classifier, and then they use both majority and minority instances for test. Wu *et al.* [7] suggest adjusting the class boundary either by transforming the kernel functions when the training data can be represented in a vector space, or by modifying the kernel matrix when the data doesn't have vector space representation. Veropoulos *et al.* [9] have shown that we can control the balance between false positive and false negative rates with SVM classifier by using different error costs for each class.

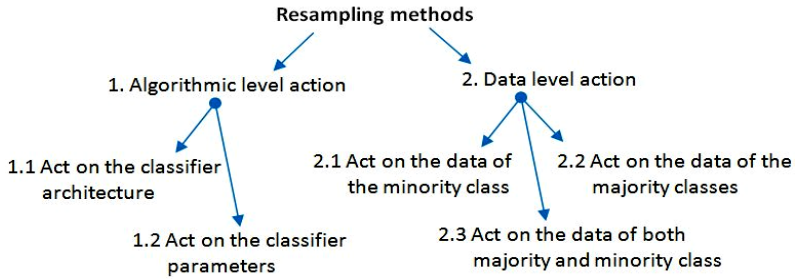

Figure 2. Resampling basic methods

2.2. Data level action

Data level action includes several techniques of rebalancing training data by adding synthetic examples to minority class or/and removing examples from majority class.

2.2.1. Undersampling methods. These methods seek to rebalance the training data by reducing the number of majority instances. This includes several methods such as random undersampling [10] or non-random undersampling [11]. The main advantage of these methods is that they reduce the training time. However, using these methods can lead to lose relevant information; for instance, when using SVM classifier, removing non-support vector instances has no effect but removing support vector instances may negatively affect the accuracy of the learned hyperplane [6]. Furthermore, if the majority class is composed of many classes, undersampling generates a within-class imbalance when skewing the intra-class distribution.

Chawla *et al.* [11] have tested a random undersampling method. The majority class is undersampled by randomly removing instances from the majority class population until the minority class reaches a specified percentage of the majority class. Kubat *et al.* [12] have proposed the SHRINK algorithm which consists of scanning the mixed regions (containing both positive and negative instance) and determining the best positive region to be labeled as positive.

2.2.2. Oversampling methods. These methods seek to rebalance the training data by adding a desired number of synthetic instances to minority data. This includes several methods such as oversampling by replicating the minority class [10] or by adding some generated synthetic samples until the desired class ratios are attained. Compared to the undersampling methods, this approach has the advantage to keep all training instances.

For each minority class sample, the SMOTE (Synthetic Minority Oversampling Technique) algorithm for continuous features [13] first computes the difference between each minority vector and the k nearest neighbors to it. Then it multiplies this difference by a random number between 0 and 1. Finally, it adds this difference to the feature value of the original feature vector to form a new synthetic vector. We have found [11] a comparison between the SMOTE algorithm and the random oversampling method that consists in replicating each minority instance. Lui *et al*., through a pilot study, conclude that replicating the minority class sample to equate classes has increased training time without any gain in classification performance [14].

3. Performance measures

In the case of an unbalanced training set, classical overall accuracy measures won't be significant if they don't take into consideration the relative distribution of each class. E.g., if we use a dataset consisted of 47 positive instances and 2767 negative ones, the overall classification rate will be 98% even though the classifier will fail to classify any of the positive instances. Balanced Classification Rate (BCR) is used to replace the overall performance when the distribution of data in training set is imbalanced [15]. The *BCR* is high when both *TPr* and *FNr* are high. The accuracy of a binary classifier is often described by confusion matrix (table 1) in which TP, FN, FP and TN are respectively true positive, false negative, false positive and true negative. Next, we will find the most common used evaluation metrics used in table2**.**

Table 1. Confusion matrix

	Expected positive	Expected negative
Actual positive	TP	FP
Actual negative	FN	TN

$$overall\ accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

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

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

$$BCR = \frac{TPr + FNr}{2}$$

$$F - value = \frac{(1 + \beta^2) * Recall * Precision}{\beta^2 * Recall + precision}$$

$$\text{Gmeans} = \sqrt{\text{sensitivity} * \text{precision}}$$

The *ROC* (Receiver Operating Characteristic) curves are the trade-off between *TP* and *FP* rates. The *AUC* area under the *ROC* curve is used when a general measure of productiveness is desired [16].

Table 2. Summary of some pervious works

Reference	Metric evaluation	Approach	Class number	Classifier
[17]	AUC, F, G_{means}, K/S, Precision, VP rate, ANOVA	Data level : BSM, CBOS, NONE, OSS, ROS, RUS, SM et WE	two	C4.5, MLP, RBF, 2NN , 5NN ,SVM, Naïve Bayes RIPPER, RF
[10]	AUC	Data level : Random Oversampling, Random undersampling, Tomak links and SMOTE	two	LDA
[18]	Mean precision	• Algorithmic level :One class SVM • Data level : Oversampling and undersampling	two	SVM
[19]	F, G_{means}, ROC curve	Data level : Oversampling and undersampling	two	SVM
[8]	F, G_{means}	• Algorithmic level : One-class SVM classifier	two	One-class SVM
[6]	G_{means}	Hybrid method : SMOTE with different error costs for the two classes	two	SVM
[11]	AUC	Data level : -2 oversampling methods by replicating each minority class example and SMOTE - random undersampling	two	C4.5 decision trees
[7]	G_{means} Sensitivity et specificity	• Algorithmic level : Adjust the class-boundary-alignment	two	SVM
[9]	ROC	Algorithmic level : used different error costs for the two classes	two	SVM
[20]	G_{means}	Data level : SHRINK approach undersampling	two	C4.5, 1-NN

4. Proposed approaches

To solve the imbalance problem, we propose to add a pre-processing module (Figure 3) for minority features to the system described in figure 1 to supply the learning system by a balanced training set.

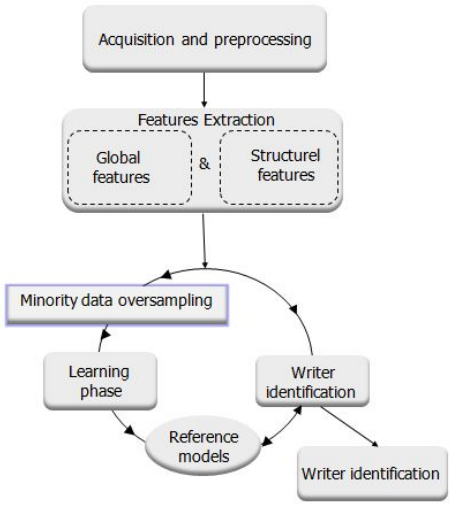

Figure 3. Flow chart of the modified writer identification system

Since the non-author class is composed of 59 subclasses, we have opted to oversampling the minority class rather than undersampling the majority one to avoid within-class imbalance problem. This problem can occur if the reduction of the number of majority class affects inequitably the subclasses.

We propose four approaches to rebalance the class distribution by adding synthetic instances into training data. Through these methods, we operate in-feature space to fill the minority feature subspace by adding an appropriate number of synthetic instances in different ways according to the oversampling rate. The synthetic instances have been generated using Curve Fitting Methods to find the coefficients of a polynomial *p(x)* of degree *n* that fits the minority instances.

Following, we will describe the four oversampling methods. Their names are inspired from network topologies.

Star topology: In this approach, we generate a synthetic instance arranged as a star silhouette (figure 4a). First, we consider a $n \times m$ minority class matrix in which $line \times column = (n\ feature\ values \times m\ instances)$. Second, we define a group of linear functions $f_i(x)$ for connecting each feature value to the mean of all ones: $\begin{cases} f_1(x) = a_1 x + b_1 \\ \dots \\ f_n(x) = a_n x + b_n \end{cases}$

Then, we identify the coefficients a_i and $b_i$in a way that each function $f_i(x)$ fits the mean value and the data. After that, we generate k linearly-spaced value x_k $(k \in [-1,1])$ according to the oversampling rate. Finally, the new synthetic minority samples are created by evaluating the value of the function $f_i(x)$ at x_k.

Polynomial curve topology: As shown in figure 4b, for each feature of the minority class matrix, we pass a curve through a set of minority instances in such a way that the estimated curve shows the best trend in these

instances. To define the "trend curve", we compute the coefficients of a polynomial p(x) of "n" degree that fits the features using the method of least square error:
$p(x) = p_1x^n + p_2x^{n-1} + p_3x^{n-2} + \cdots + p_nx^n + p_{n+1}$
Next, we generate k linearly-spaced value x_k ($k \in [-1,1]$) according to the oversampling rate. Then, the new synthetic minority samples are created by evaluating the value of the polynomial *p(x)* at x_k. The degree *n* has been determined empirically according to the best *TP* rate.

Bus topology: This approach is based on interpolation (figure 4c). First, for each feature of the minority class matrix, we plot a straight line connecting one minority data to the next. The single line connecting two succeeding data is described generally by the following linear function: $f_i(x) = ax + b$.
Then, we identify the coefficients "a" and "b" in such a way that f(x) fits the data. Next, we generate *k* linearly-spaced value x_k ($k \in [-1,1]$) according to the oversampling rate. Finally, the new synthetic minority samples are created between each two succeeding data by evaluating the value of the function $f_i(x)$ at x_k.

Mesh topology: This approach based on interpolation is a process for which each feature of the minority class matrix we plot a set of straight lines connecting one minority data to all others (figure 4d). For each single line connecting two data we proceed as stated in Bus topology concerning the synthetic data addition.

5. Experimentation and results

In this section, we first describe our database which has been designed for the study of a writer identification task. Then, we present the extracted features. Finally, we compare the results achieved using the four proposed topologies.

5.1. Database description

The database on which our experiments have been conducted contains handwriting samples of an Arabic letter source document written by 60 persons. Each person has been required to copy the same letter three times: two samples have been used for the training stage and the other for the tests which make a total of 180 A4-format sample pages. In our experiments, all handwritten text images have been initially digitized in greyscale at a resolution of 300 dpi. Then via the pre-processing stage we have removed salt and pepper noise by applying the median filter, extracted the text lines by horizontal projection profile method, and determined the bounding box of each text line. Structural and global features have been extracted from line text.

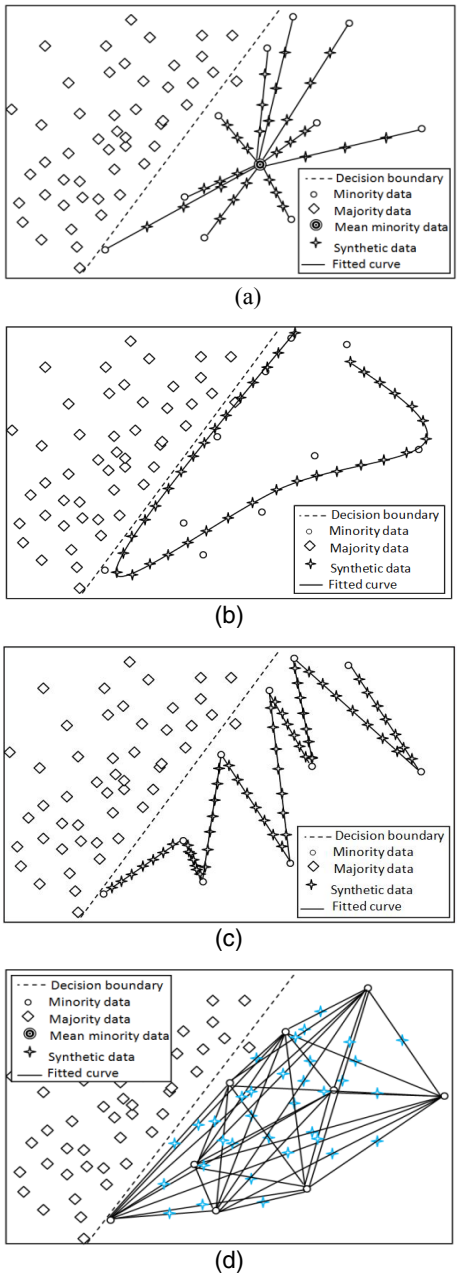

Figure 4. Proposed oversampling topologies using: (a) Star topology, (b) polynomial Curved-bus topology, (c) Bus topology, and (d) Mesh topology

5.2. Feature extraction

Features are extracted from handwriting samples at three levels: from ascenders, sub-words and lines. Two types of features have been considered: structural and textural resulting from the wavelets.

5.2.1. Structural features.

Line height: We fixed spaces between the lines to avoid the overlaps between two succeeding lines. Under these conditions, the extraction of the lines of writings becomes easier by the method of the horizontal projection profile. However, diacritical signs are ignored by the extraction procedure. The height of the line is measured directly on the histogram.

Spaces between sub-words: An algorithm allows to cross the baseline and to calculate the average of white spaces between sub-words.

Inclination of the ascenders: First, we extract the zone delimiting the ascenders. Then the segmentation of the ascenders is operated with connected-component neighbourhoods. Finally, we measure the angle of inclination on each ascender.

Dot boldness and shape: Diacritic dot shapes characterize the writer's style in Arabic handwriting. Our visual system detects the dot boldness by the density of black surface and their "ellipsoid" shape by comparing two constants, conventionally denoting major and minor axis.

Two features are extracted from isolated diacritic dots. First, the height and the width of each diacritic dot (isolated dots) are measured on each point's bounding box. Second, the ratio of the sum of black pixels is computed.

5.2.2. Holistic features.

Wavelet transforms: We have used Daubechies wavelet transforms to describe cursive Arabic word. The 2D wavelet decomposition is applied until level three. Four sub-band images have been yielded at this level: approximation image, horizontal details, vertical details and diagonal details at level three. Therefore, we compute the mean and the standard deviation from approximation matrix subband and the standard deviation from each detail matrix subbands.

5.3. Classifier description

We used an SVM classifier with radial base function core. However, SVMs are basically binary classifiers and the classification problem of the sixty writers requires a multiclass method. Therefore, we have adopted the one-against-all approach. Each SVM classifier learns the positive class examples as well as the negative ones trained independently. The SVM parameters C and γ of each classifier are optimized using cross validation method, and the ones with the best cross validation accuracy are kept.

For each proposed topology, twenty sets of experiments have been conducted. We have trained each classifier with: the positive instances only, the negative instances only, the original set, and the seventeen oversampled datasets. For the latter, we add a new positive generated synthetic data as part of a minority class to all the majority instances with a positive to negative instance ratio of 0.1:1 to 1.7:1 with a 0.1 step-value.

5.4. Results

Our results are presented in five parts: training with original data set and using the four proposed topologies. The purpose of these experiments is twofold: (1) to evaluate the necessity to use a preprocessing step to supply the classifier by balanced training set, and (2) to compare results provided by the four sampling approaches in order to select the best one. We begin by examining the classification behavior when the negative instances significantly outnumber the positive ones in the original training set (about 1.7% positive instances). The results show that the majority instances correctly classified perform very well ranging from 95% to 100% (figure 5), the TPr fluctuates between 0 (result provided by writer number 25 SVM-classifier) to 100%, and the Overall accuracy is almost equal to TNr. The overall average result shows that the test samples belonging to the authorship class have been poorly identified whereas those belonging to the non-authorship class have been almost correctly identified. The overall accuracy has been prevailed by non-authorship samples correctly identified. The impact of the different imbalance ratio using the four proposed algorithms is shown in figure 6. The proposed algorithms seek an acceptable TPr by adding new synthetic examples to the minority class, while preserving as much TNr as possible.

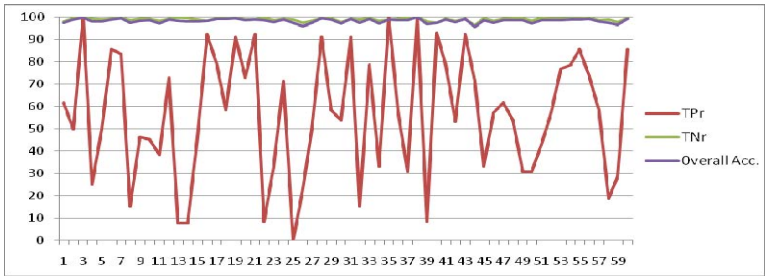

Figure 5. TPr, TNr, and overall accuracy provided with the original highly skewed class distribution

The results show that the performances of the classifier have been significantly improved in the four reported cases: the rate of true positive instances correctly classified increases and meets the TNr. However, in Bus topology only 0.4:1 positive to negative instance ratio is good enough to equalize TPr and TNr which matches earlier compared to others topologies (figure 6b). Nevertheless, The TPr decreases starting at rate of 1:1 even though if it was trained with only positive instances and the TPr and TNr do not exceed 85% when equalized. This set of experiments shows the importance of using a preprocessing step in which the minority instances will be oversampled in order to improve the classification rate of the positive instances while maintaining the performance of negative ones as much as possible.

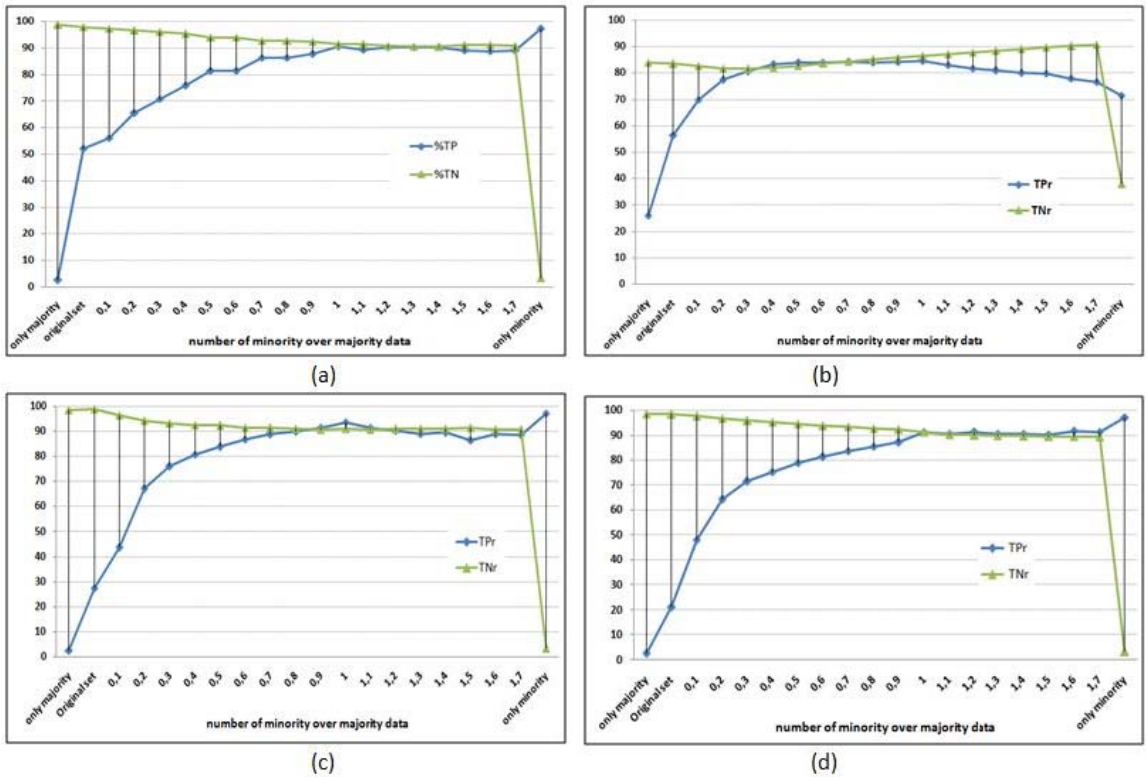

Figure 6. TPr and FNr with different ratio using (a) Star topology, (b) Bus topology, (c) polynomial curved-Bus topology, and (d) Mesh topology

Figure 7 compares the four proposed approaches. It shows that there are no significant differences between Star and Mesh topologies. The BCR performs better in the imbalance of non-authorship/authorship ratios between 0.8 and 1.2, but its performance drops when less than 0.8. The approach based on polynomial function topology provides better results at the oversampling ratio of 1.

6. Conclusions and future works

In this paper, we have addressed the problem of imbalance inducted by using the one-against-all SVM classifier when the one-class training dataset risks to heavily outnumber the other. Experiments have been applied on writer identification system using Arabic handwriting texts and one against all multiclass SVM classifier. During the phase of training, each i^{th} SVM classifier is trained with all the features of the i^{th} author and all the other fifty-nine authors. In this situation the negative instances significantly outnumber the positive ones (imbalance ratio of 1/59) and this problem will be worsened proportionally with the increase of the class number. To tackle this problematic situation, an oversampling minority data using polynomial fitting has been investigated intensively in several case studies. We have tested four new polynomial function topologies: Star, Polynomial curved-bus, Bus and Mesh topologies. Training with imbalanced data provides a poor true positives' rate. Our methods manifest the significance of sampling in improving notably the classification of the minority instances while maintaining the performance of majority ones. With these adequate oversampled data, the authorship identification is improved effectively.

Experimental results have confirmed that the Mesh and Star topologies outperform the other proposed ones. The results indicate that this research is quite promising and proves to be laudable for further investigation on the relationship between class overlap and class imbalance problems.

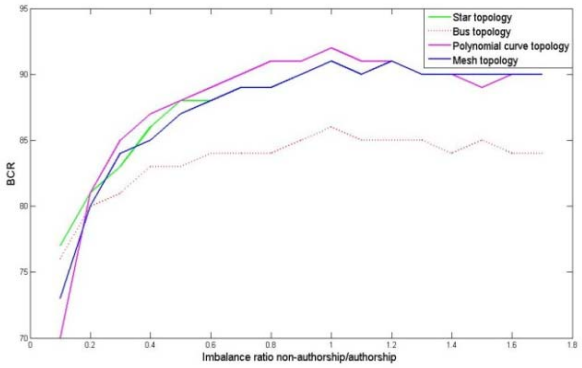

Figure 7. Overall Balanced Classification Rate for different considered imbalance ratio

7. Bibliography

[1] Il-S. Oh, and Ch.Y. Suen,"A class-modular feedforward neural network for handwriting recognition", *Pattern Recognition,* vol. 35, 2002, pp. 229-244.

[2] N. Essoukri Ben Amara, and S. Gazzah, "Une approche d'identification des fontes arabes", *Colloque International Francophone sur l'Ecrit et le Document*, La Rochelle/ France, 2004.

[3] M.N. Kapp, C.Freitas, J. Nievola, and R.Sabourin, "Evaluating the conventional and class-modular architectures feedforward neural network for handwritten word recogntion", *Brazilian Symposium on Computer Graphics and Image Processing*, Sao Carlos/Brazil, 2003, pp 315-319.

[4] K.Takahashi, and D.Nishiwaki, "A class-modular GLVQ ensemble with outlier learning for handwritten digit recognition", 7th *International Conference on Document Analysis and Recognition,* Edinburgh/Scotland, 2003, pp.268-272.

[5] S. Gazzah, and N. Essoukri Ben Amara, "Neural networks and support vector machines classifier for writer identification Using Arabic Script", *The International Arab Journal of Information Technology*, 2008*,* Vol. 5, No. 1, pp.93- 102.

[6] R. Akbani, S. Kwek, and N. Japkowics, "Applying support vector machines to imbalanced datasets", *European Conference on Machine Learning*, 2004, pp. 39-50.

[7] G. Wu, and E.Y. Chang, "Aligning boundary in kernel space for learning imbalanced dataset", *4th IEEE Int. Conf. Data Mining*, Brighton/U.K., 2004, pp. 265–272.

[8] L. Zhuang, and H. Dai, "Parameter optimization of kernel-based one class classifier on imbalance learning", *Journal of Computers,* Academy publisher, Filand,2006, Vol.1, No.7, pp.32-40.

[9] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines", *International Joint Conf. Artificial Intelligence*, Stockholm /Sweden, 1999, pp. 55-60.

[10] J. Xie, and Z. Qiu, "The effect of imbalanced data sets on LDA: A theoretical and empirical analysis", *Pattern recognition*, 2007, vol. 40, pp. 557-662.

[11] N. Chawla, "C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure", *Workshop on Learning from Imbalanced Datasets,* Washington DC, 2003.

[12] M. Kubat, R.C. Holte, and S.Matwin, "Machine learning for the detection of oil spills in satellite radar images", *Machine learning*, 1998, Vol. 30, pp.195-215.

[13] N. Chawla, K.W.Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique", *Journal of Artificial Intelligence Research*, 2002, Vol.16, pp. 321-357.

[14] Y. Lui, N.V. Chawla, M.P Harper, E Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech", *Computer Speech and Language*, 2006, vol. 20, pp. 468-494.

[15] P. Li, K.L. Chan, and W. Fang, "Hybrid kernel machine ensemble for imbalanced data sets", *18th International Conference on Pattern Recognition* (ICPR'06), 2006.

[16] T. Fawcett, "ROC graphs: notes and practical considerations for research", *Technical Rapport, HP Laboratories, Polo Alto*, USA, 2004.

[17] J.V. Hulse, T.M. Khoshgoftaar, and A. Napolitano, "Experimental perspective on learning from imbalanced data", *International Conference on Machine Learning*, Corvalis/Oregon USA, 2007, pp. 935-942.

[18] H.J. Lee, and S. Cho, "The novelty detection approach for different degrees of class imbalance", *International Conference on Neural Information Processing*, Hong-Kong/ China, Vol. 02, LNCS 4233, 2006, pp. 21-30.

[19] Y. Liu, A. An and X. Huang , "Boosting prediction accuracy on imbalanced datasets with SVM ensembles", *10th Pasific Asia Conference on Knowledge Discovery and Data Mining*, Singapore, 2006, pp. 107-118.

[20] M. Kubat, R. Holte, and S. Matwin, "Learning when Negative Examples Abound", *Proceeding of European Conference on Machine Learning,* 1997, pp. 146-153.