An Optimized Cost-Sensitive SVM for Imbalanced Data Learning

Peng Cao[1,2], Dazhe Zhao[1] and Osmar Zaiane[2]

[1]Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, China; [2]University of Alberta, Canada
{cao.p, zhaodz}@neusoft.com, zaiane@ualberta.ca

Abstract. Class imbalance is one of the challenging problems for machine learning in many real-world applications. Cost-sensitive learning has attracted significant attention in recent years to solve the problem, but it is difficult to determine the precise misclassification costs in practice. There are also other factors that influence the performance of the classification including the input feature subset and the intrinsic parameters of the classifier. This paper presents an effective wrapper framework incorporating the evaluation measure (AUC and G-mean) into the objective function of cost sensitive SVM directly to improve the performance of classification by simultaneously optimizing the best pair of feature subset, intrinsic parameters and misclassification cost parameters. Experimental results on various standard benchmark datasets and real-world data with different ratios of imbalance show that the proposed method is effective in comparison with commonly used sampling techniques.

1 Introduction

Recently, the class imbalance problem has been recognized as a crucial problem in machine learning and data mining [1]. This problem occurs when the training data is not evenly distributed among classes. This problem is also especially critical in many real applications, such as credit card fraud detection when fraudulent cases are rare or medical diagnoses where normal cases are the majority. In these cases, standard classifiers generally perform poorly. Classifiers usually tend to be overwhelmed by the majority class and ignore the minority class examples. Most classifiers assume an even distribution of examples among classes and assume an equal misclassification cost. Moreover, classifiers are typically designed to maximize accuracy, which is not a good metric to evaluate effectiveness in the case of imbalanced training data. Therefore, we need to improve traditional algorithms so as to handle imbalanced data and choose other metrics to measure performance instead of accuracy. We focus our study on imbalanced datasets with binary classes.

Much work has been done in addressing the class imbalance problem. These methods can be grouped in two categories: the data perspective and the algorithm perspective [2]. The methods with the data perspective re-balance the class distribution by re-sampling the data space either randomly or deterministically. The main disadvantage of re-sampling techniques are that they may cause loss of important information or the model overfitting, since that they change the original data distribution.

A cost-sensitive classifier tries to learn more characteristics of samples with the minority class by setting a high cost to the misclassification of a minority class sample. It does not modify the data distribution. Weiss [3] left the questions "why doesn't the cost-sensitive learning algorithm perform better given the known drawbacks with sampling; and are there ways to improve the effectiveness of cost-sensitive learning algorithms." We need to improve the effectiveness of cost sensitive learning algorithms by optimizing factors which influence the performance of cost sensitive learning.

There are two challenges with respect to the training of cost sensitive classifier. The misclassification costs play a crucial role in the construction of a cost sensitive learning model for achieving expected classification results. However, in many contexts of imbalanced dataset, the misclassification costs cannot be determined. Beside the cost, the feature set and intrinsic parameters of some sophisticated classifiers also influence the classification performance. Moreover, these factors influence each other. This is the first challenge. The other is the gap between the measure of evaluation and the objective of training on the imbalanced data [4]. Indeed, for evaluating the performance of a cost-sensitive classifier on a skewed data set, the overall accuracy is irrelevant. It is common to employ other evaluation measures to monitor the balanced classification ability, such as G-mean [5] and AUC [6]. However, these cost-sensitive classifiers measured by imbalanced evaluation are not trained and updated with the objective of the imbalanced evaluation. To achieve good prediction performance, learning algorithms should train classifiers by optimizing the concerned performance measures [7].

In order to solve the challenges above, we design a novel framework for training a cost sensitive classifier driven by the imbalanced evaluation criteria. The training scheme can bridge the gap between the training and the evaluating of cost sensitive learning, and it can learn the optimal factors associated with the cost sensitive classifier automatically. The significance of the scheme has two questions to fix: how to optimize these factors simultaneously; and using what evaluation criteria for guiding their optimization. These two issues are our key steps for improving the cost sensitive learning in the context of the class imbalance problem without cost information. Our main contributions in this paper are centered around the questions above.

The contributions of this work can be listed as follows:

1) Optimizing the factors (ratio misclassification cost, feature set and intrinsic parameters of classifier) simultaneously for improving the performance of cost-sensitive SVM.

2) Imbalanced data classification is commonly evaluated by measures such as G-mean and AUC instead of accuracy. However, for many classifiers, the learning process is still largely driven by error based objective functions. We use the measure directly to train the classifier and discover the optimal parameter, ratio cost and feature subset based on different evaluation functions like the G-mean or AUC. Different metrics can reflect different aspect performance of classifiers.

2 Related Works

The common methods to solve data imbalance are data re-sampling perspective and algorithm perspective. Re-sampling methods are attractive under most imbalanced circumstances. This is because re-sampling adjusts only the original training dataset, instead of modifying the learning algorithm; therefore it provides a convenient and effective way to deal with imbalanced learning problems using standard classifiers by balancing the instances of the classes. Weiss and Provost observed that the naturally occurring distribution is not always optimal [8]. Therefore, one needs to modify the original data distribution. The idea of sampling is to purposefully manipulate the class distributions by under-sampling and over-sampling.

The methods with the algorithm perspective adapt existing common classifier learning algorithms to bias towards the small class, such as cost-sensitive learning. Cost-sensitive learning is one of the most important topics in machine learning and data mining, and attracted significant attention in recent years. Cost-sensitive learning methods consider the costs associated with misclassifying examples. The objective of cost-sensitive methods is to minimize the expected cost of misclassifications without changing the class distribution [9]. A closely related idea to cost-sensitive learners is shifting the bias of a machine to favor the minority class so as to obtain better recognition ability by adjusting the costs associated with misclassification rather than to seek the minimum of total misclassification cost [4, 10-12]. In the construction of cost sensitive learning, the parameter of misclassification cost plays an indispensable role.

There is another issue in the class imbalance problem. The importance of feature selection to class imbalance problems, in particular, was realized and has attracted increasing attention from machine learning and data mining communities. Wrappers and embedded methods are feature subset selection methods that consider feature interaction in the selection process. Some authors have conducted studies on using feature selection to combat the class imbalance problem [13, 14]. Zheng and Srihari [14] suggest that existing measures used for feature selection are not appropriate for imbalanced datasets. The wrapper feature selection seems a good approach.

3 Cost-Sensitive SVM

Support Vector Machines (SVM), which has strong mathematical foundations based on statistical learning theory, has been successfully adopted in various classification applications. SVM maximizes a margin in a hyperplane separating classes. However, it is overwhelmed by the majority class instances in the case of imbalanced datasets because the objective of regular SVM is to maximize the accuracy. In order to provide different costs associated with the two different kinds of errors, cost-sensitive SVM (CS-SVM) [15] is a good solution. CS-SVM is formulated as follows:

$$\begin{aligned}
&Min\ \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C_+\sum\nolimits_{i:y_i=+1}\xi_i + C_-\sum\nolimits_{j:y_j=-1}\xi_j \\
&s.t.\ \ y_i[(\boldsymbol{w}^T\boldsymbol{x}_i)+b] \ge 1-\xi_i \quad i=1,\cdots,n \\
&\qquad \xi_i \ge 0
\end{aligned} \tag{1}$$

where the C_+ is the higher misclassification cost of the positive class, which is the primary interest, while C_- is the lower misclassification cost of the negative class. Using the different error cost for the positive and negative classes, the hyperplane could be pushed away from the positive instances. In this paper, we fix $C_- = C$ and $C_+ = C \times C_{rf}$, where C and C_{rf} are respectively the regularization parameter and the ratio misclassification cost factor. In the construction of cost sensitive SVM, the misclassification cost parameter plays an indispensable role. For the cost information, Veropoulos et al. have not suggested any guidelines for deciding what the relative ratios of the positive to negative cost factors should be.

In general, the Radial Basis Function (RBF kernel) is a reasonable first choice for the classification of the nonlinear datasets, as it has fewer parameters (γ).

4 Optimized cost sensitive SVM by measure of imbalanced data

SVM tries to minimize the regularized hinge loss; it is driven by an error based objective function. However, the overall accuracy is not an appropriate evaluation measure for imbalanced data classification. As a result, there is an inevitable gap between the evaluation measure by which the classifier is to be evaluated and the objective function based on which the classifier is trained. The classifier for imbalanced data learning should be driven by more appropriate measures. We inject the appropriate measures into the objective function of the classifier in the training with PSO. The common evaluation for imbalanced data classification is G-mean and AUC. However, for many classifiers, the learning process is still driven by error based objective functions. In this paper we explicitly treat the measure itself as the objective function when training the cost sensitive learning. We designed a measure oriented training framework for dealing with imbalanced data classification issues. Chalwa et al. [6] propose a wrapper paradigm that discovers the amount of re-sampling for a dataset based on optimizing evaluation functions like the f-measure, and AUC. To date, there is no research about training the cost sensitive classifier with measure based objective functions. This is one important issue that hinders the performance of cost-sensitive learning.

Another important issue of applying the cost-sensitive learning algorithm to the imbalanced data is that the cost matrix is often unavailable for a problem domain. The misclassification cost, especially the ratio misclassification cost, plays a crucial role in the construction of a cost sensitive approach; the knowledge of misclassification costs is required for achieving expected classification result. However, the values of costs are commonly given by domain experts. They remain unknown in many domains where it is in fact difficult to specify the precise cost ratio information. It is not exact to set the cost ratio to the inverse of the imbalance ratio (the number of majority instances divided by the number of minority instances); especially it is not accurate for some classifier such as SVM. Some cost sensitive learning use a heuristic approach to search the optimal cost matrix, such as Genetic Algorithm [10] or grid search to find the optimal cost setup [12].

Apart from the ratio misclassification cost information, feature subset selection and the intrinsic parameters of the classifier have a significant bearing on the performance. Both factors are not only important for imbalanced data classification, but also for any classification. Feature selection is the technique of selecting a subset of discriminative features for building robust learning models by removing most irrelevant and redundant features from the data. Optimal feature selection can concurrently achieve good accuracy and dimensionality reduction. Unfortunately, the imbalanced data distributions are often accompanied by high dimensionality in real-world datasets such as text classification, bioinformatics, and computer aided detection. It is important to select features that can capture the high skew in the class distribution [1]. Moreover, proper intrinsic parameter setting of classifiers, such as regularization cost parameter and the kernel function parameter for SVM, can improve the classification performance. It is necessary to use the grid search to optimize the regulation parameter and kernel parameters. Moreover, these three factors influence each other. Therefore, obtaining the optimal ratio misclassification cost, feature subset and intrinsic parameters must occur simultaneously.

Based on the reasons above, our specific goal is to devise a strategy to automatically determine the optimal factors during training of the cost sensitive classifier oriented by the imbalanced evaluation criteria (G-mean and AUC).

In this paper, for the multivariable optimization, especially the hybrid multivariable, the best methods are swarm intelligence techniques. We choose the particle swarm optimization as our optimization method because it is mature and easy to implement. Particle swarm optimization (PSO) is a population-based global stochastic search method [16]. PSO optimizes an objective function by a population-based search. The population consists of potential solutions, named particles. These particles are randomly initialized and move across the multi-dimensional search space to find the best position according to an optimization function. During optimization, each particle adjusts its trajectory through the problem space based on the information about its previous best performance (personal best, *pbest*) and the best previous performance of its neighbors (global best, *gbest*). Eventually, all particles will gather around the point with the highest objective value.

The position of individual particles is updated as follows:

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{2}$$

With *v*, the velocity calculated as follows:

$$v_{id}^{t+1} = w \times v_{id}^t + c_1 \times r_1 \times (pbest_{id}^t - x_{id}^t) + c_2 \times r_2 \times (gbest^t - x_{id}^t) \tag{3}$$

Where v_i^t indicates velocity of particle *i* at iteration *t*; w indicates the inertia factor; C_1 and C_2 indicate the cognition and social learning rates, which determine the relative influence of the social and cognition components. r_1 and r_2 are uniformly distributed random numbers between 0 and 1, x_i^t is current position of particle *i* at iteration *t*, $pbest_i^t$ indicates best of particle *i* at iteration *t*, $gbest^t$ indicates the best of the group.

Evaluation measures play a crucial role in both assessing the classification performance and guiding the classifier modeling. The purpose of cost-sensitive learning is usually to build a model with total minimum misclassification costs. However, it should be based on the known cost matrix condition. The purpose of our cost sensitive learning is to get a best AUC or G-mean evaluation metric. We train the cost sensitive learning using performance measures as the objective functions directly. Through training the cost sensitive classifier with measure based objective functions, we can discover the best factors in terms of the different evaluation. The evaluation metrics value is taken as the fitness function to adjust the position of a particle. These two different evaluations reflect different aspect of the classifier. AUC affects the ranking ability and G-mean involves the accuracies of both classes at the same time.

For binary class classification, the cost parameter is only one parameter, which means the relative cost information, ratio misclassification cost factor C_{rf}. Since the RBF kernel is selected for the cost sensitive SVM, γ and C are the parameters to be optimized. We need to combine the discrete and continuous values in the solution representation since the costs and parameters we intend to optimize are continuous while the feature subset is discrete. Each feature is represented by a 1 or 0 for whether it is selected or not. The major difference between the discrete PSO [17] and the original version is that the velocities of the particles are rather defined in terms of probabilities that a bit will change to one. Using this definition a velocity must be restricted within the range [0, 1], to which all continuous values of velocity are mapped by a sigmoid function:

$$v'^{t}_{i} = sig(v^{t}_{i}) = \frac{1}{1+e^{-v^{t}_{i}}} \tag{4}$$

Equation 4 is used to update the velocity vector of the particle while the new position of the particle is obtained using Equation 5.

$$x^{t+1}_{i} = \begin{cases} 1 & if \quad r_i < v'^{t}_{i} \\ 0 & otherwise \end{cases} \tag{5}$$

Where r_i is a uniform random number in the range [0,1] .

Algorithm 1: MOCSSVM (optimized cost sensitive SVM by imbalanced data measure)

```
Input: Training set D; termination condition T; population size SN; metric E; NumFolds =5
Randomly initialize particle population positions and velocities (including cost matrix, intrinsic parameters, and feature subset)
repeat
  foreach particle i
     Construct the $D_i$ with the feature selected by the particle i
     for k=1 to NumFolds
        Separate $D_i$ randomly into $Trt^k_i$ (80%) for training and $Trv^k_i$ (20%) for validation
        Train CS-SVM with cost matrix and intrinsic parameters optimized by the particle i on the $Trt^k_i$
        Evaluate the cost sensitive classifier on the $Trv^k_i$ , and obtain the value $M^k_i$ based on E
     end for
     $M_i$=average($M^k_i$); Assign the fitness of particle i with $M_i$
     if fitness ($pbest_i$) <= fitness ($x_i$)
         then $pbest_i = x_i$
     end if
   end foreach
   set gbest as best pbest
   foreach particle i
      update $velocity_i$ and $position_i$ with Eq. 2 and 3.
   end foreach
until termination condition
output optimal parameters, cost ratio and feature subset of gbest
```

The solution (i.e. particle) includes three parts: the ratio misclassification cost, the intrinsic parameters of classifier, and the feature subsets. Figure 1 illustrates the mixed solution representation in the PSO.

Ratio cost	Intrinsic parameters		Feature subset					
C_{rf}	C	γ	f_1	f_2	…	f_{n-1}	f_n

Fig. 1 Solution representation

The detailed algorithm MOCSSVM to optimize cost sensitive SVM by imbalanced data measure is shown in Algorithm 1. It is a wrapper framework for empirically discovering the potential misclassification cost ratio, feature subset, and intrinsic parameters for CSL oriented by the imbalanced evaluation criteria (G-mean and AUC).

5 Experimental study

5.1 Dataset Description

To evaluate the classification performance of our proposed method in different classification tasks, and to compare with other methods specifically devised for imbalanced data, we tried several datasets from the UCI database. We used all available datasets from the combined sets used in [4]. This also ensures that we did not choose only the datasets on which our method performs better. The minority class label (+) is indicated in Table 1. The datasets chosen have diversity in the number of attributes and imbalance ratio. Moreover, the datasets used have both continuous and categorical attributes. All the experiments are conducted by 10-fold cross-validation.

Table 1. The data sets used for experimentation
The dataset name is appended with the label of the minority class (+)

Dataset (+)	Instances	Features	Class balance
Hepatitis (1)	155	19	1:4
Glass (7)	214	9	1:6
Segment (1)	2310	19	1:6
Anneal (5)	898	38	1:12
Soybean (12)	683	35	1:15
Sick (2)	3772	29	1:15
Car (3)	1728	6	1:24
Letter (26)	20000	16	1:26
Hypothyroid(3)	3772	29	1:39
Abalone (19)	4177	8	1:130

5.2 Experiment I

In this experiment, the comparison is conducted between our method and the intermediate method or basic method, such as basic SVM with and without the feature selection, cost sensitive SVM, cost sensitive SVM with grid search and our method MOCSSVM with/without the feature selection. For the basic SVM with feature selection, it is a common wrapper feature selection method with evaluation by classification performance. As for CSSVM, the misclassification cost ratio is searched iteratively to maximize the measure score within a range of cost value. CSSVM uses a grid search for optimization. We also need to treat this misclassification cost ratio as a hyperparameter, and locally optimize this parameter. However, it is not feasible to use a triple circulation for optimizing the best parameters, so we optimize the best parameter pair(C and γ) firstly, then locally optimize the cost ratio parameter based on the best parameter pair(C and γ). All SVM models in this experiment use the same kernel, RBF, and for basic SVM and CSSVM, the intrinsic parameters are fixed with default values (C=1 and γ =1).

For the PSO setting of our method MOCSSVM, the initial parameter values of it in our proposed method were set according to the conclusion drawn in [18]. The parameters were used: C_1=2.8, C_2=1.3, w=0.5. To empirically provide good performance while at the same time keeping the time complexity tractable, the particle number was set dynamically according to the amount of the variables optimized (=1.5×|variables to be optimized|), and the termination condition could be a certain number of iterations (500 cycles) or other convergence condition (no changes any more within 2× |variables to be optimized| cycles). Besides these parameters in PSO, the other parameters are the upper and lower of limit parameter of model to be optimized. For Grid-CSSVM and MOCSSVM, the ranges for C and γ are based on a grid search for SVM parameters as recommended in [19]. The range of C is (2^{-5}, 2^{15}), and the range of γ is (2^{-15}, 2^{3}). The range of ratio misclassification cost factor C_r was empirically set between 1 and 10×*ImbaRatio* (ratio between the instance amounts of two classes).

In this experiment, we assess the overall quality of classifiers with only the AUC evaluation metric. From the result in Table 2, we found that simultaneously optimizing the feature subset, parameter and cost ratio generally help the base classifiers learned on the different data sets, regardless of feature selecting or not.

Table 2. Experimental results between all the methods based on the SVM

Dataset	Basic SVM		CS-SVM	Grid-CSSVM	MOCSSVM	
	without *FS*	*FS*	without *FS*	without *FS*	without *FS*	*FS*
Hepatitis	0.632	0.714	0.707	0.801	**0.861**	0.855
Glass	0.952	0.957	0.953	0.955	0.994	**1**
Segment	1	1	1	1	1	1
Anneal	0.876	0.925	0.957	**1**	**1**	**1**
Soybean	1	1	1	1	1	1
Sick	0.728	0.761	0.788	0.848	0.908	**0.975**
Car	0.990	0.987	0.990	0.999	**1**	**1**
Letter	0.898	0.895	0.909	0.983	0.980	**0.999**
Hypothyrid	0.830	0.855	0.887	0.945	0.973	**0.988**
Abalone	0.638	0.712	0.722	0.839	0.867	0.893
Average	0.854	0.881	0.892	0.937	0.957	0.971

Under the condition where the feature selection is not carried out, we found that the simultaneous optimization for all the factors using PSO outperforms the optimization using grid search, which optimizes the intrinsic parameters first, then searches the optimal misclassification cost parameter based on the best intrinsic parameters. It lacks many potential parameter pairs not searched in the parameter space. Hence, it shows that the parameters need to be search at the same time. Moreover, in MOCSSVM, the use of feature selection was found to improve the AUC for each dataset except the Hepatitis dataset.

Although, we take some dynamic strategies for improving the efficiency of the PSO algorithm, the average running iterations for PSO-based approach is slightly inferior to that of the grid search algorithm. However, it significantly improves the classification accuracy and obtains fewer input features for the classifiers. Therefore, we can draw the conclusion that by simultaneously optimizing the intrinsic, misclassification cost parameter and feature selection with the imbalanced evaluation measure guiding improves the classification performance of the cost sensitive SVM on different datasets.

5.3 Experiment II

The comparison is conducted between our method and the other state-of-the-art imbalanced data classifiers, such as the random under-sampling (RUS), SMOTE [20], SMOTEBoost [21], and SMOTE combined with asymmetric cost classifier [5]. For the under-sampling algorithm, the SMOTE and SMOTEBoost, the re-sampling rate is unknown. In our experiments, in order to compare equally, no matter under-sampling or over-sampling method, we also use the evaluation measure as the optimization objective of the re-sampling method to search the optimal re-sampling level. The increment step and the decrement step are both set at 10%. This is a greedy search, which process repeats, greedily, until no performance gains are observed. The optimal re-sampling rate is decided in an iterative fashion according to the evaluation metrics. Thus, in each fold, the training set is separated into training subset and validating subset for searching the appropriate rate parameters. The evaluation metrics are also used with the G-mean and AUC. For the CS-SVM with SMOTE, for each re-sampling rate searched, the optimal misclassification cost ratio is determined by searching under the evaluation measure guiding under the current over-sampling level of SMOTE.

As shown in bold in Table 3, our MOCSSVM outperforms all the other approaches on the great majority of datasets. It did not get the best result only on the Glass dataset. From the results, we can see that the random under-sampling has the worst performance. This is because it is possible to remove certain significant examples and under-sampling the majority class causes larger angles between the ideal and learned hyperplane, and also reduces the total number of training instances which also contributes to increasing angles [5]. Both the SMOTE and SMOTEBoost improve the classification on the imbalanced data. The over-sampling algorithm that tries to improve on it inevitably sacrifices some specificity in order to improve the sensitivity; but the degree of sensitivity improved is larger than the lost specificity. However, they have a potential disadvantage of distorting the class distribution. SMOTE combined with a different cost classifier is better than only SMOTE over-sampling, and it is the method that shares most of the second best results. In the majority of cases, the G-mean value from the G-mean wrapper is higher than the one of the AUC wrapper, but in some cases, the G-mean value from the AUC wrapper is higher, such as Hepatitis and Abalone datasets for MOCSSVM and Glass. Even for MOCSSVM, the average G-mean from AUC optimization is better than the one from G-mean optimization. From this, we believe that by using AUC as the wrapper evaluation function we get better performances, which is the similar conclusion as in [6]. We believe that employing the AUC evaluation measure as optimization objective could lead to more generalized performances. Similarly, the two evaluation metrics wrapper optimizations for the same classifier result in different misclassification cost, feature subset and intrinsic parameters, since they optimize different properties of the classifier.

The feature selection is as important as the re-sampling in the imbalanced data classification, especially with high dimensional datasets. However, feature selection is often ignored. Our method does feature selection in the wrapper paradigm, hence improves the classification performance on the datasets which have higher dimensionality, such as Anneal, Sick and Hypothyroid.

We use the MOCSSVM method as a baseline and compare the other methods against it. Although all methods are optimized under the evaluation measure oriented, we can clearly see that MOCSSVM is almost always equal to, or better than other methods. What is most important is that our method does not change the data distribution, while the re-sampling may make the generalization not as good as the training, since that the data distribution are different between the training set and test set.

Table 3. Experimental comparison between MOCSSVM method and other imbalanced data methods

Dataset		**RUS**		**SMOTE**		**SMOTE Boost**		**SMOTE-CSSVM**		**MOCSSVM**	
		wrapper metric		**wrapper metric**		**wrapper metric**		**wrapper metric**		**wrapper metric**	
		AUC	**GM**	**AUC**	**GM**	**AUC**	**GM**	**AUC**	**GM**	**AUC**	**GM**
Hepatitis	**AUC**	0.663	0.528	0.754	0.721	0.788	0.759	0.813	0.783	**0.855**	0.823
	GM	0.598	0.487	0.672	0.667	0.558	0.592	0.628	0.729	**0.805**	0.801
	Fea.	19								7	8
Glass	**AUC**	0.955	0.948	0.988	0.986	0.981	0.978	0.992	0.975	**1**	0.995
	GM	0.817	0.803	0.844	0.858	0.874	0.862	0.965	**0.988**	0.986	0.971
	Fea.	9								5	4
Segment	**AUC**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	GM	0.993	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0.998	**1**
	Fea.	19								10	11
Anneal	**AUC**	0.882	0.866	0.912	0.876	0.891	0.889	0.957	0.934	**1**	**1**
	GM	0.616	0.535	0.758	0.821	0.761	0.784	0.819	0.835	0.999	**1**
	Fea.	38								14	12
Soybean	**AUC**	**1**	0.992	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	GM	0.876	0.953	0.947	0.965	0.992	0.997	**1**	0.997	**1**	**1**
	Fea.	35								12	12
Sick	**AUC**	0.784	0.742	0.822	0.799	0.841	0.824	0.931	0.874	**0.975**	0.954
	GM	0.206	0.141	0.452	0.528	0.508	0.512	0.811	0.825	0.893	**0.915**
	Fea.	29								9	7
Car	**AUC**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	GM	0.964	0.964	0.962	0.958	0.979	0.981	0.995	**0.998**	0.996	**0.998**
	Fea.	6								4	4
Letter	**AUC**	0.907	0.896	0.966	0.956	0.987	0.965	0.988	0.980	**0.999**	0.995
	GM	0.925	0.933	0.947	0.954	0.934	0.922	0.965	0.961	0.983	**0.985**
	Fea.	16								12	10
Hypothyroid	**AUC**	0.876	0.843	0.971	0.915	0.967	0.955	0.973	0.971	0.988	**0.989**
	GM	0.482	0.612	0.853	0.894	0.876	0.903	0.876	0.901	0.964	**0.968**
	Fea.	29								9	14
Abalone	**AUC**	0.781	0.613	0.822	0.754	0.799	0.780	0.846	0.812	**0.893**	0.855
	GM	0.618	0.687	0.712	0.814	0.645	0.744	0.698	0.817	**0.853**	0.785
	Fea.	8								4	5
Average	**AUC**	0.885	0.843	0.924	0.900	0.925	0.915	0.950	0.933	**0.971**	0.961
	GM	0.710	0.711	0.815	0.814	0.813	0.830	0.876	0.910	**0.948**	0.943
win/tie/lose	**AUC**	0/3/7	0/2/8	0/3/7	0/3/7	0/3/7	0/3/7	0/3/7	0/3/7	base	1/4/5
	GM	0/0/1	0/1/9	0/1/9	0/1/9	0/1/9	0/1/9	0/2/8	1/2/7	3/1/6	base

Many papers conclude that there is no consistent clear winner between the sampling approaches and the cost-sensitive technique. However, the conclusions were based on the default condition without sufficient search in the parameters space. In this paper, we have empirically shown that under the evaluation measure guiding, the performances of cost sensitive SVM with cost, feature subset and intrinsic parameter optimized are better than the re-sampling methods with sampling level optimized.

5.4 Experiment III

Computer aided detection provides a computer output in order to assist radiologists in the diagnosis of Lung Cancer on medical images. It can be divided into initial nodule identification step and false-positive reduction step. The purpose of false-positive reduction is to remove false positives (FPs) as much as possible while retaining a relatively high sensitivity. It is a typical class imbalance issue since the two classes are typically skewed and have unequal misclassification costs. Our database consists of 98 thin section CT scans with 106 solid nodules, obtained from Guangzhou hospital in China. We obtained the appropriate candidate nodule samples objectively using a candidate nodule detection algorithm, which identifies 95 true nodules as positive class and 592 non-nodules as negative class from the total CT scans; the class imbalance ratio is 1:6. The imbalance level is not extremely high, but the misclassification costs of each class are very different. The imbalance level is dependent on reliability and accuracy of the initial detection process. Our feature extraction process generated 43 features from multiple views. Using these features, we construct the input space for our classifiers. Our method outperforms the other common approach (Table 4). It means that our method can be applied on the nodule or other lesion detection. The measure optimization used is the AUC metric.

Table 4 Experiment result of candidate nodule classification

metric	**SVM**	**CSSVM**	**RUS**	**SMOTE**	**SMOTE-Boost**	**SMOTE-CSSVM**	**MO CSSVM**
AUC	0.681	0.785	0.603	0.948	0.948	0.956	**0.969**
GM	0.208	0.662	0.590	0.826	0.818	0.867	**0.937**

6 Conclusion

Learning with class imbalance is a challenging task. We propose a wrapper paradigm oriented by the evaluation measure of imbalanced dataset as objective function with respect to misclassification cost, feature subset and intrinsic parameters of SVM. Our measure oriented framework could wrap around an existing cost-sensitive classifier. The proposed method has been validated on some benchmark imbalanced data and real application. The experimental results presented in this study have demonstrated that the proposed framework provides a very competitive solution to other existing state-of-the-arts methods, in optimization of G-mean and AUC for combating imbalanced classification problems. These results confirm the advantages of our approach, showing the promising perspective and new understanding of cost sensitive learning. In the future research, we will extend the framework to the imbalanced multiclass data classification.

Reference

1. Chawla, N.V., Japkowicz, N.& Kolcz, A. (2004): Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets* 6 (1):1-6.
2. Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. (2006): Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*:25-36.
3. Weiss G., McCarthy K., Zabar B. (2007): Cost-sensitive learning vs. sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? IEEE ICDM, pp. 35–41.
4. Yuan, B. & Liu, W.H. (2011): A Measure Oriented Training Scheme for Imbalanced Classification Problems. Pacific-Asia Conference on Knowledge Discovery and Data Mining Workshop on Biologically Inspired Techniques for Data Mining. pp:293–303.
5. Akbani, R., Kwek, S. & Japkowicz, N. (2004): Applying support vector machines to imbalanced datasets. European conference on machine learning.
6. Chawla, N.V., Cieslak, D.A., Hall, L.O. & Joshi, A. (2008): Automatically countering imbalance and its empirical relationship to cost. Utility-Based Data Mining: A Special issue of the International Journal Data Mining and Knowledge Discovery.
7. Li, N., Tsang, I., Zhou, Z. (2012): Efficient Optimization of Performance Measures by Classifier Adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume: PP , Issue: 99, Page(s): 1.
8. Weiss, G. & Provost, F. (2003): Learning when training data are costly: the effect of class distribution on tree induction, J Artif Intel Res 19:315–354.
9. Zhou, Z.H. & Liu, X.Y. (2006): Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1): 63-77.
10. Sun, Y., Kamel, M.S. & Wang, Y. (2006): Boosting for Learning Multiple Classes with Imbalanced Class Distribution. Proc. Int'l Conf. Data Mining:592-602
11. Wang, B.X. & Japkowicz, N. (2008): Boosting support vector machines for imbalanced data sets, Journal of Knowledge and information Systems 4994, 38–47.
12. Thai-Nghe N. (2010): Cost-Sensitive Learning Methods for Imbalanced Data, *Intl. Joint Conf. on Neural Networks*.
13. Forman, G. (2003): An Extensive Empirical Study of Feature Selection Metrics for Text Classification. J. Machine Learning Research, vol. 3, pp. 1289-1305.
14. Zheng, Z., Wu, X. & Srihari, R. (2004): Feature selection for text categorization on imbalanced data. SIGKDD Explorations, 6(1):80-89.
15. Veropoulos, K., Campbell, C. & Cristianini, N. (1999): Controlling the sensitivity of support vector machines. International Joint Conference on AI, 55–60.
16. Kennedy, J., Eberhart, R.C. (1995): Particle swarm optimization, *IEEE Int. Conf. Neural Networks*, pp.1942–1948.
17. Khanesar, M.A., Teshnehlab, M. & Shoorehdeli, M.A. (2007): A novel binary particle swarm optimization. In Control & Automation. Mediterranean Conference on, pp. 1–6
18. Carlisle, A. & Dozier, G. (2001): An Off-The-Shelf PSO. *PSO Workshop*. pp. 1–6.
19. Hsu, C.W, Chang, C.C. & Lin, C.J. (2003): A Practical Guide to Support vector Classification, National Taiwan University Technical Report.
20. Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. (2002): SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 16:321–357.
21. Chawla N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W. (2003): SMOTEBoost: Improving Prediction of the Minority Class in Boosting. European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 107-119.