Cost-Sensitive learning on classification

Qin Yang[1]*, Changyao Zhou[2]

[1]*School of Business, Sichuan Agricultural University, Dujiangyan 611830, Sichuan, China*

[2]*School of Resources and environment, Sichuan Agricultural University, Wenjiang 611130, Sichuan, China*

Received 1 March 2014, www.cmnt.lv

Abstract

Real-world predictive data mining (classification or regression) problems are often cost sensitive, meaning that different types of prediction errors are not equally costly. In this paper we propose a new algorithm for cost-sensitive classification in a multiple time series prediction problems. The fitness function of the genetic algorithm is the average cost of classification when using the decision tree, including both the costs of tests (features, measurements) and the costs of classification errors. The proposed model is evaluated in a real world application based on a network of satellite network map distributed in land spatial pattern evolution in Chengdu Plain. These satellite networks generate multiple time series data representing land spatial pattern. This study presents a new algorithm for cost-sensitive classification that deal with class imbalance using both recompiling and CSL. The method combines and compares several sampling methods with CSL using support vector machines (SVM). We build our cost-benefit model for the prediction process as a function of satellite network in a distributed land spatial and measured the optimum number of satellite network that will balance the expenses of the system with the prediction accuracy.

Keywords: cost sensitive learning, prediction, distributed satellite network, cost benefit analysis

1 Introduction

Machine learning and data mining reply heavily on a large amount of data to build learning models and make predictions, and thus, the quality of data is ultimately important. To evaluate in a real world application based on a network of satellite network distributed in land spatial pattern evolution in a region. We use real-world data-sets in land spatial pattern evolution from 2008 to 2013. Cost/Benefit Analysis is a relatively simple and widely used technique for deciding whether to make a change [1]. Profit and costs drive the utility of every land resource use decision. As land resource use decision making, from strategic to operational planning, is based upon future realizations of the decision parameters [2]. The real trick to doing a cost benefit analysis well is making sure you include all the costs and all the benefits and properly quantify them. Cost-benefit analysis has already attracted much attention from the machine learning and data mining communities [3]. Thus, cost-sensitive learning algorithms should make use of only known values. Under cost-sensitive learning, we impute values of data, and the learning algorithms make use of values to minimize the total cost of tests and classifications.

In the literature of cost sensitive analysis for data mining applications, most emphasis is given to classification problems. Recent research and related publications show that cost sensitive analysis is not deeper analysed, modelled, and applied to prediction problems. Conversely, data mining methods for regression and time series analysis generally disregard economic utility and apply simple accuracy measures [2]. Only some theoretical approaches exist for specific data mining methods such as neural networks and support vector machines.

In this research, we propose a new approach and develop a methodology to apply the cost benefit analysis to real world prediction problems. Our application is prediction of land spatial pattern. Data for our research is collected from a patented distributed satellite network of land spatial pattern. In this paper, we have proposed a model for cost-benefit analysis in a multiple time series prediction problems. We have identified three distinct areas of the cost function and analysed the behaviour in each region. We have experimentally found out the threshold values corresponding to these regions. We presents a new algorithm for cost-sensitive classification that deal with class imbalance using both resembling and CSL. The method combines and compares several sampling methods with CSL using support vector machines (SVM). Results are compared and analysed. We have discussed how this methodology can be utilized in similar distributed systems.

Following a brief introduction to previous work in section 2. Section 3 introduces multiple time series prediction problems and its formalization for prediction of land spatial pattern based on daily satellite network data. We report our recent research efforts in introducing cost sensitive analysis for this real world prediction application in Section 4. Conclusions are given in Section 5.

Corresponding author e-mail: yangqin@sicau.edu.cn

2 Related work

In recent years data mining community has attempted incorporating cost benefit analysis into classification problems, where the "Cost" could be interpreted as misclassification cost, training cost, test cost, or others [5]. Among all different types of costs, the misclassification cost is the most popular one. In general, misclassification cost is described the cost of predicting that an example belongs to class *i* when in fact it belongs to class *j* [6]. Hollmen et al., and Elkan introduce a cost model that inland resource uses the specific properties of objects to be classified. Instead of a fixed misclassification cost matrix, they utilize a more general matrix of cost functions. These functions operate on the data to be classified and are re-calculated for each data point separately [7, 8].

Most of the existing cost sensitive classifiers assume that datasets are either noise free or noise in the datasets is less significant. However, real-world data is never perfect and suffer from noise that may impact models created from data. Zhu and Wu have addressed the problems of class noise, which means the errors introduced in the class labels. They have studied the noise impacts on cost sensitive learning and proposed a cost guided noise handling approach for effective learning [6]. The class imbalance problem has been recognized as a crucial problem in machine learning and data mining. Such a problem is encountered in a large number of ranges, and it can lead to poor performance of the learning method [9]. It has been indicated that cost-sensitive learning is a good solution to the class imbalance problems and Zhou and Liu have studied methods that address the class imbalance problems applied to cost-sensitive neural networks [3].

Similarly, for predictive data mining problems of regression and time series analysis the costs arising from invalid point prediction, costs of under prediction versus over prediction, etc. are also analysed in the literature. Crone et al. have analysed the efficiency of a linear asymmetric cost function in inventory management decisions, training multilayer perceptions to find a cost efficient stock-level for a set of seasonal time series directly from the data [2]. A similar work has been proposed by Christoffersen and Diebold by introducing a new technique for solving prediction problems under asymmetric loss using piecewise-linear approximations to the loss function [10].

Wang and Stockton have investigated how the constraints imposed by changing export market affect the identification of "cost estimating relationships" and investigated their relative benefits and limitations in terms of their effects on the overall cost model development process. Neural network architecture has been used and a series of experiments have been undertaken to select an appropriate network [11].

Cost estimation generally involves predicting labour, material, utilities or other costs over time given a small subset of factual data on "cost drivers." Alice has examined the use of regression and SVM models in terms of performance, stability and ease of use to build cost estimating relationships. The results show that SVM have performed well when dealing with data, which there is little prior knowledge about the cost estimating relationship to select for regression modelling. However, regression models have shown significant improvements in terms of accuracy in cases where an appropriate cost estimating function can be identified [12].

3 Prediction of total power production

This paper presents our research results in analysis of distributed land spatial pattern in a region, and development of prediction model based on data collected from multiple distributed satellite network related to specific land spatial pattern. These land spatial pattern operate data throughout the year continuously.

Each land spatial keeps record of real-time land spatial pattern, current land spatial output and real time changes and variations in land spatial usage and supply availability. These data are taken at specific time intervals and they vary from a day to a week or a month. The sampling frequency depends on the type of data that are collected at the land spatial.

The data for our analysis comprises readings of satellite network at 417 distributed land spatial pattern, and the cumulative variable that correspond to the total land spatial pattern of those 417 land spatial pattern. The data set that we are using is collected daily and we use a repository of three years from year 2008 to year 2013. It has 365 data for each land spatial in years 2011, 2012 and 2013. These historic data can be used to build and improve relative models used for short and long term land spatial pattern forecasting. They are currently used by land trading and marketing firms and federal regulatory agencies including Homeland Security.

Prediction systems have to obtain these data sets at a cost. Getting data from more land spatial pattern to make prediction means increase in expenses. Our goal is to predict the total land spatial pattern with reduced number of satellite network, where we can compromise between the expenses for collecting data and the quality of prediction accuracy. Based on the analysis of different multiple time series prediction methods we have selected SVM as the best candidates for our study in building cost sensitive prediction model [13-15]. Our approach shows that it is possible to measure the total land spatial pattern using reduced number of satellite network stations. The modelling results and general approach may be used in other systems to determine the required number of satellite network to be used for data collection.

3.1 DATA COLLECTION AND PRE-PROCESSING

The data for analysis is collected daily for the years from year 2008 to year 2013. Most of the machine learning methods including SVM requires that all data sets to be normalized. We use MATLAB to normalize data between

[-1,+1] for 417 land spatial pattern. In the raw data set there are some entries with 0 values. They correspond to days in which a satellite network does not record a value for a given day. In our problems, all those 0 values are taken as actually recorded values after proving with the authorities [16].

We have used year 2008 data as training data set and year 2012 and year 2013 data as testing data. Main goals of our research are a) To measure and compare the quality of different prediction methodologies, and b) To measure optimum number of satellite network, which will give the best results balancing expenses of the system. We used several sampling methods with CSL using support vector machines (SVM). Accuracy of this model is measured, by computing the correlation coefficient, between actual values and predicted values in the testing data set. Prediction accuracy is lie on the amount of satellite network we use, i.e. number of power land spatial inputs we use in the prediction. To prove this we have built training and testing data sets by varying the number of satellite network from which the data are collected. In our study we considered data from 16, 64, 128, 512 and 1024 land spatial inputs. An overview of these steps is presented in Table 1. In the following sections, we elaborate on these steps.

TABLE 1 Classify land resources

Step: Assessment of land resources:
1. Land resources classification:
–identify the land resources role in the organization
2. Define land resources policy goals:
–assign weights to land resources policy goals (C,I,A)
3. Classify land resources:
–enumerate resources available on the given system
–determine the resource importance for each land resources policy goal
–compute the overall resource weight for each land resources policy goal

3.2 EXPERIMENTAL RESULTS

In SVM, a selected number of satellite network data (time series) are used in prediction. We have 417 satellite networks as data sources, and there are various ways of selecting the subset. Selection of a subset of satellite network (in our case 10, 20, 30, 40 tc.) will cause combinatorial explosion. We made one heuristic approximation in this step to eliminate computational complexity. Each of 417 satellite network time series is compared with the total land spatial pattern by computing the correlation factor. We wanted to measure how much each of satellite network measurements are correlated to the output. Then, we sorted satellite network based on the correlation factor, and selected subset is a portion of top ranked satellite network.

We used a feed forward neural network with back propagation learning using only one hidden layer. The algorithm was implemented in MATLAB ver6.5. Experiments showed that equal number of input and hidden nodes give the best results in prediction. Inputs to the network are the data columns corresponding to satellite network' recordings at each land spatial throughout the year and the single output represents predicted value for the land spatial pattern in the region. We have experimented the SVM model with different combinations of the parameters and determined that the values 0.001 for accuracy parameter and 0.04 for learning rate with Tangent-Sigmoid activation function give the best prediction results. Figure 1 shows the results for prediction using the SVM with 40, 70, and 100 satellite network with the correlation coefficients (r) 0.84, 0.93, and 0.95 respectively. A toolbox called LIBSVM (A Library for Support Vector Machines) was used for SVM methodology [17].

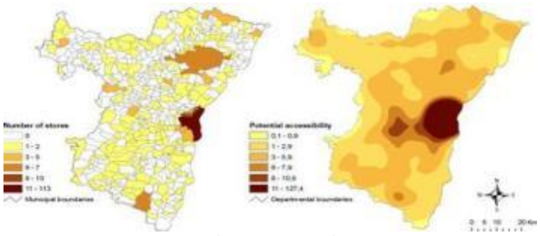

FIGURE 1 Graph for solution of all hyper dataset.

We use support vector machines (SVM) as the base classifier. Grid search is used to determine the best hyper parameters for SVM and the resembling techniques.

(a) Method 1: Combination of Sampling techniques with CSL, called S-CSL.

Algorithm 1: *S-CSL*
Begin
while *NFC < MAX_{NFC}* ***dob***
for *i=1 to ps* ***do***
Update velocity of the ith location point according to (1);
Update position of the ith location point according to (2);
Calculate the fitness value of the ith location point;
NFC++;
end for
Update the pbest and gbest, if needed.
for *i=1 to ps* ***do***
if *rand(0,1) < p_o* ***then***
Generate a mutant location point according to (3);
Calculate the fitness values of the mutant location point;
NFC++;
Select the fitter one between the current location point and the mutant location point as the new current location point;
end if
end for
Update the pbest and gbest, if needed.
end while
End

(b) Method 2: Using CSL by Optimizing Cost Ratio Locally, called CSL-OCRL.

Algorithm 2: *CSL-OCRL*
Begin
while *NFC < MAX_{NFC}* ***do***
for *i=1 to ps* ***do***
Update velocity of the ith location point according to (1);
Update position of the ith location point according to (2);
Calculate the fitness value of the ith location point;
NFC++;
end for
Update the pbest and gbest, if needed;
Calculate the boundaries of current search space;
Create a random location point Y according to (3);
Create a trail location point X according to (5);*

Calculate the fitness value of X;*
NFC++;
If** X* is better than X_w **then
Replace X_w with X;*
end if
Update the pbest and gbest, if needed.
end while
End

Standard facilities of matrix algebra in MATLAB are used for MR analysis. The Figure 1 shows the variations of quality of prediction results obtained with SVM 128 input satellite network, r = 100
(c) Using SVM with 16 input satellite network, r = 10
(d) Using SVM with 1024 input satellite network, r =510
Input satellite network for three methods MR for different number of input time series.

Detailed comparison of these methods and discussion about optimal number of satellite network for different models are given in our previous article. The hypothesis that optimal number of satellite network should be the trade off between accuracy of prediction and costs of the monitoring system is the initiating point of the current research [16].

4 Cost benefit analysis in multiple time series prediction

Should we collect data from all areas? If not what would be the optimum number of areas to use for prediction? What are economic benefits from the prediction system? These questions are essential part in analysis of a solution for our prediction system based on distributed satellite network system which monitors land spatial pattern. A cost benefit analysis for multiple time series prediction, we are developing and applying in this research, will give some of the answers.

$$V(t+1)=wV(t)+c_1r_1\left(pbest_i(t)-X_i\right)+ c_2r_2\left(gbest_i(t)-X_i(t)\right), \tag{1}$$

$$X_i(t+1)=X_i(t)+V_i(t+1), \tag{2}$$

$$X_i^*=gbest=a_1\left(gbest-X_i\right)+\left(1-a_1\right)\left(X_{i1}-X_{i2}\right). \tag{3}$$

In order to verify the performance of the proposed approach, ten famous benchmark functions are selected in our experiments [12]. According to their properties, they are divided into two types: functions f_1 to f_7 are uni-modal functions and f_8 to f_{10} are multimodal functions. All the functions are minimized problems. The specific definitions, dimensions, and the global optimum are listed as follows:

$$f_1=\sum_{i=1}^{D}x_i^2,$$

where $x_i\in[-100,100]$, D=30, and the global optimum is 0.

$$f_2=\sum_{i=1}^{D}|x_i|+\prod_{i=1}^{D}x_i,$$

where $x_i\in[-10,10]$, D=30, and the global optimum is 0.

$$f_3=\sum_{i=1}^{D}\left(\sum_{j=1}^{i}x_j\right)^2,$$

where $x_i\in[-100,100]$, D=30, and the global optimum is 0.

$$f_4=\max_i\left(|x_i|,1\le i\le D\right),$$

where $x_i\in[-100,100]$, D=30, and the global optimum is 0.

$$f_5=\sum_{i=1}^{D-1}\left[100\left(x_{i+1}-x_i^2\right)^2+\left(x_i-1\right)^2\right],$$

where $x_i\in[-30,30]$, D=30, and the global optimum is 0.

$$f_6=\sum_{i=1}^{D}\left(\lfloor x_i+0.5\rfloor\right)^2,$$

where $x_i\in[-100,100]$, D=30, and the global optimum is 0.

$$f_7=\sum_{i=1}^{D}ix_i^4+rand[0,1),$$

where $x_i\in[-1.28,1.28]$, D=30, and the global optimum is 0.

$$f_8=\sum_{i=1}^{D}-x_i\sin\left(\sqrt{|x_i|}\right),$$

where $x_i\in[-500,500]$, D=30, and the global optimum is -12569.5.

$$f_9=\sum_{i=1}^{D}\left[x_i^2-10\cos\left(2\pi x_i\right)+10\right],$$

where $x_i\in[-5.12,5.12]$, D=30, and the global optimum is 0.

$$f_{10}=-20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right)- \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos\left(2\pi x_i\right)\right)+20+e,$$

where $x_i\in[-32,32]$, D=30, and the global optimum is 0.

There are several costs involved in building our prediction system. They can be of two types: fixed cost and variable cost. In optimization problems, fixed cost only shifts the cost curve into a higher or lower level. So, we only considered variable cost in our analysis. In building and maintaining the satellite network system, these expenses correspond to hardware installation and maintenance, data collection, data processing and data analysis.

These expenses would include, procedure OPTIMIZECOSTRATIO(DTrain;_; _)

Input: DTrain, SVM parameters _, step length _
Outputs: the best cost ratio for GMean
2: (DLocalTrain;DHoldout) DTrain . split for 5-fold CV
3: ImbaRatio jMajorj
jMinorj . imbalance ratio of DTrain
4: maxRatio ImbaRatio _ 1:5
5: currentRatio 1:0
6: bestGMean 0
7: while currentRatio <= maxRatio do
8: buildLocalModel(DLocalTrain; _)
9: currentGMean testLocalModel(DHoldout)
10: if (currentGMean > bestGMean) then
11: bestGMean currentGMean
12: bestCostRatio currentRatio
13: end if
14: currentRatio currentRatio + _
15: end while
16: return bestCostRatio
17: end procedure

Cost may also include a risk factors that would compensate increase of prices, depreciation etc. For these reasons it should be determined very carefully by experts in the range.

Accuracy of prediction is a nonlinear function of the number of satellite network [16]. We selected the polynomial model with relative small error and at the same time enough simple. We experimentally confirmed that polynomial function of third order makes a good approximation of the prediction non-linearity:

1: procedure HYPERSEARCH(DTrain;)
returns the best hyperparameters _ for eval. metric E
2: (DLocalTrain;DHoldout) DTrain //split for 5-fold CV
//Raw search:
3: bestC; besty ←0
4: for i ←15,……10 do
5: for j ←15,……0 do
6: y←2j ; C←2i
7: buildLocalSVM(DLocalTrain; ;C)
8: TestLocalModel(DHoldout) //using metric E
9: Update bestC; best
10: end for
11: end for //Smooth search:
12: for i← bestC-1,…… bestC + 1; step r do
13: for j← best -0.1,…… best + 0:1; step q do
14: y←j; C←i
15: buildLocalSVM(DLocalTrain; ;C)
16: TestLocalModel(DHoldout) //using metric E
17: _ C; //Update the best parameter values
18: end for
19: end for
20 : return _
21: end procedure

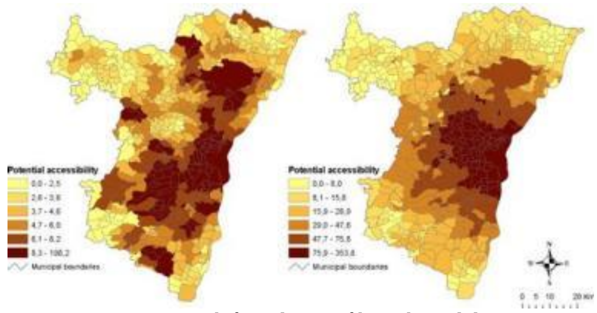

FIGURE 2 Graph for solution of hypothyroid dataset

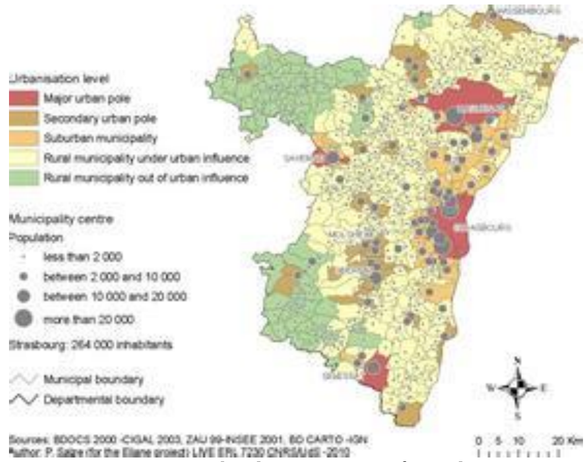

FIGURE 3 Graphical interpretation of Ann dataset

4.1 COST BENEFIT MODEL

We experimentally determined parameters in relative formulae *y*(*n*) for the SVM (SVM *y*) methodologies. We can find the relative relationship between the accuracy and the number of inputs for each of the methods applying a curve fitting procedure and assuming that the function is a polynomial third order. The approximated equations fits the graphs of Figure 2:

These three functions represent our cost-benefit models for prediction. Each model is based on two input parameters:' C -ratio of cost constants, and n - number of satellite network.

4.2 INTERPRETATION OF EXPERIMENTAL RESULTS

We can analyse how the actual cost function changes with values of constant 1 2 C C or ' C, and we will try to find a minimum of these functions for different ' C values using experimental results.

4.2.1 Large benefits

Figure 3 gives a plot for cost function C for very small value of ' C 0.0001based on our experimental results. Three graphs represent the three applied prediction methodologies. Small ' C means that the value of 2 C is very high, relative to the value of 1 C i.e. the users measure that the benefit of prediction accuracy is very high and its overweight any cost 1 C for satellite network installation and maintenance. As it is expected, the cost function C is continuously decreasing function, where its minimum is with maximum number of satellite network n (in our case 417). That means, we accept all expenses for installation of all 417 satellite network and there is no prediction system. Output will be just calculated as a sum of all satellite network values. Even if the satellite network are expensive, it is not a sufficient reason to influence the cost function, which is minimum for maximum *n*.

4.2.2 Large costs

Figure 3 shows the variation of total cost as a function of a number of input satellite networks for cost 64. It gives its minimum value with minimum number of input areas close to 0. If the installation of satellite network determined by cost CSL is very costly, and benefits (profit) determined by Cost CSL is relatively low, the function will be continuously increasing. The Figure 3. Graph for solution may found at extremely small number of satellite network (close to 0). Again, we do not need prediction system because the expenses are so high, that the system is not economically feasible.

4.2.3 Costs and benefits are balanced

The most interesting case is when the cost ratio is increasing. We believe that it is the common case in real world applications. Minimum value for total cost , will determine necessary number of satellite network to obtain maximum benefits from the prediction system. Based on the number of satellite network we can measure the quality of prediction for the recommended configuration.

It shows the case where the minimum of cost, occurs between the minimum and the maximum number of areas. As expected, simple SVM model shows the minimum cost where the required number of input satellite network is around 128. Both simple SVM and CSL using support vector machines (SVM). It shows a minimum at around 100 input satellite networks for this value. As shows the variation the total cost with the different values, and the number of satellite network. In addition, data shows the accuracy of prediction at the optimum number of satellite network for each case. As we discussed so far, it is clear that the value plays a major role in selecting the minimum value for the cost function. In general cost will determine, if it is useful to build the prediction system or not. Based on the experimental results, we can see that the behaviour of the cost function defines three regions. The cost function behaves as an asymptotically increasing function, where the minimum cost can be found at minimum number of satellite network in all three methods. Accuracy at this situation is at its minimum values. In both previous cases we do not need a prediction system. The cost function clearly defines a minimum value for the number of satellite network at a point inside the range.

Accuracy and the number of satellite network used depend on prediction technique. This is a good engineering design heuristics, obtained from cost benefit analysis, to determine whether prediction analysis is economically feasible for our system.

5 Conclusions

In this paper we describe a new methodology for the cost-benefit analysis in multiple time series prediction. It is applied in a real world distributed satellite network. Satellite network generate time series data representing land spatial pattern evolution in Chengdu Plain. For establishing the prediction model, we use three common prediction methods SVM. Training and testing data were available for year’s 2008 to 2013 land spatial pattern evolution in Chengdu Plain. Our results show that CSL using support vector machines (SVM). Technique gives the best prediction model.

Acknowledgments

This work was supported by Sichuan Province Science and technology support program (2013GZ0024).

References

[1] Li P, Qiao P L, Liu Y C 2008 A hybrid re-sampling method for SVM learning from imbalanced data sets *Proceedings of fifth Int. Conf. on Fuzzy Systems and Knowledge Discovery* Washington DC USA 65-9

[2] Zhang Y, Zhou, Z H 2008 Cost-sensitive face recognition *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Anchorage AK 416-422

[3] Yang Q, Wu X 2006 10 challenging problems in data mining research *International Journal of Information Technology and Decision Making* **5**(4) 597-604

[4] Chawla N V, Bowyer K, Hall L, Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *Journal of AI Research* **16** 321-57

[5] Chawla N V, Japkowicz N Kotcz A 2004 Special issue on learning from imbalanced data sets *SIGKDD Explorations* **6** 1-6

[6] Nickerson A, Japkowicz N, Millos E 2001 Using unsupervised learning to guide resampling in imbalanced data sets *The Eighth International Workshop on AI and Statitsics* 261-5

[7] Raskutti B, Kowalczyk A 2004 Extreme re-balancing for SVMs: acase study *SIGKDD Explorations* **6**(1) 60-9

[8] Chen X, Gerlach B, Casasent D 2005 Pruning support vectors for imbalanced data classification *IEEE International Joint Conference on Neural Networks* **3** 1883-8

[9] Lessmann S 2004 Solving imbalanced classification problems with support vector machines *Proceedings of International Conference on Artificial Intelligence* 214-20

[10] Liu X Y, Wu J, Zhou Z H 2009 Exploratory under-sampling for class-imbalance learning *IEEE Transactions on Systems, Man and Cyberernetics Part B* 539-50

[11] Li X, Wang L, Sung E 2008 AdaBoost with SVM-based component classifiers *Engineering Applications of Artificial Intelligence* **21**(5) 785-95

[12] Sheng S, Ling C X, Ni A, Zhang S 2006 Cost-sensitive test strategies *The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference* **1** 323-31

[13] Ting K M 2002 A study on the effect of class distribution using costsensitive learning *Discovery Science* 98-112

[14] Akbani R, Kwek S, Japkowicz N 2004 Applying support vector machines to imbalanced datasets *Proceedings of the 15th ECML* 2004 39-50

[15] Batista G E A P A, Prati R C, Monard M C 2004 A study of the behavior of several methods for balancing machine learning training Data *SIGKDD Explorations* **6**(1) 20-6

[16] Van Hulse J, Khoshgoftaar T M, Napolitano A 2007 Experimental perspectives on learning from imbalanced data *The 24th ICML* ACM 935-42

[17] Platt J C 1999 Probabilistic outputs for SVM and comparisons to regularized likelihood methods A*dvances in Large Margin Classifiers* MIT Press 61-74

[18] Domingos P 1999 Metacost: A general method for making classifiers costsensitive *The 5th International conference on Knowledge Discovery and Data mining KDD1999* 155-164

[19] Margineantu D D 2000 When does imbalanced data require more than cost-sensitive learning? *Papers from the AAAI Workshop Technical Report WS-00-05* 47-50

Authors

Qin Yang, born in March, 1977, China

Current position, grades: researcher at Sichuan Agricultural University, China.
University studies: MS degree in Computer Software and Theory, University of Electronic Science and Technology of China.
Scientific interests: Computational intelligence: location point swarm optimization, ant colony optimization, genetic algorithm, differential evolution.

Changyao Zhou, born in March, 1989, China

Current position, grades: Master Student in Sichuan Agricultural University, China.
Scientific interests: Computational intelligence: location point swarm optimization, sustainable use of land resource.