VOTCL and a Case Study of its Application*

Guang-tong Zhou, Yi-long Yin, Xin-jian Guo, Cai-ling Dong, Qing-yuan Wang
School of Computer Science and Technology, Shandong University, Jinan, 250101, China
ylyin@sdu.edu.cn

Abstract

In many real-world applications, the problem of class imbalance and cost-sensitive always arise simultaneously. To address this problem, we propose an effective solution named VOTCL: first, we generate several balanced training datasets by combining under-sampling and over-sampling techniques; then, they are trained to get base learners; at last, voting based on optimal threshold is proposed to ensemble those base learners for decision-making. Experiments on the cross-selling dataset provided by PAKDD2007 competition show the effectiveness of our solution with AUC 0.6037.

1. Introduction

In many real-world applications, such as fraud detection and medical diagnosis, the default assumption of balanced class[1] distribution and equal misclassification costs underlying conventional machine learning algorithms is most likely violated simultaneously [1]: the class distribution of the datasets is skewed (i.e. class imbalance) and the cost of misclassifying a minority class sample is substantially greater than the cost of misclassifying a majority class sample (i.e. cost-sensitive). The performance of learning algorithms on CICS[2] problem is limited if class imbalance and cost-sensitive are not considered, as they tend to be overwhelmed by the majority class and ignore the minority class [2, 3, 4]. Most research concentrates on the problems of class imbalance or cost-sensitive separately. Approaches for addressing the problem of class imbalance can be divided into two directions: resampling techniques and algorithm-based improvements. Resampling manipulates the datasets directly to generate a balanced class distribution by over-sampling minority class [5] or under-sampling majority class [6]. Algorithm-based improvements change the internal or external structure of learning algorithms. External structure changes includes thresholding [3, 4] and reweighting [7], and internal structure changes focuses on specific learning algorithms, such as support vector machine [8, 9] and decision tree [10], etc.. To address the problem of cost-sensitive, direct MEC can be easily applied [2, 3] if both misclassification costs and class membership probabilities are known. However, in many real-world applications, it is usually difficult and sometimes impossible to obtain the costs and probabilities, so both cost estimators and probability estimators should be learned [11] firstly, and then resampling or algorithm-based improvements can be applied to reduce overall misclassification cost.

Despite of various algorithms addressing the problem of class imbalance and cost-sensitive separately, there are few successful models for the problem of CICS. Theoretically, we can make use of the class imbalance rescale ratio and cost-sensitive rescale ratio to address this problem [12], but as the rescale ratios are difficult to obtain in real-world applications, the practice of the theoretical method is questionable.

A solution named VOTCL (Voting based on Optimal Threshold for Class imbalance and cost-sensitive Learning) is proposed to address the problem of CICS: first, we generate several balanced training datasets by combining under-sampling and over-sampling techniques; then, they are trained to get base learners; at last, voting based on optimal threshold is proposed to ensemble those base learners for decision-making. Training based on balanced datasets can effectively weaken the impact of class imbalance and cost-sensitive. Voting based on optimal threshold is applied to minimize the overall misclassification cost, since this method avoids misclassifying minority class samples as many as possible.

VOTCL relaxes the requirements for class imbalance rescale ratio and cost-sensitive rescale ratio. It is a wrapper-based solution that we can choose different base learners flexibly according to the applications. Furthermore, VOTCL may be helpful

* Corresponding author: Yi-long Yin, *ylyin@sdu.edu.cn*.

[1] In this paper, we only discuss binary class problems.

[2] In this paper, the term "CICS" is used to represent the phenomenon that serious class imbalance and cost-sensitive arising simultaneously.

978-0-7695-3304-9/08 $25.00 © 2008 IEEE
DOI

when dealing with large datasets. The experimental results also show the effectiveness of our solution.

This paper is organized as follows. Section 2 describes VOTCL. Section 3 illustrates the procedures of the experiments and presents some experimental results. Section 4 explains the decision-making principle of VOTCL. Section 5 concludes this paper.

2. VOTCL

VOTCL is composed of three main processes: (1) Resampling. (2) Training of base learners. (3) Voting based on optimal threshold. The framework of VOTCL is shown in Fig. 1, and the pseudo code of VOTCL is shown in Tab. 1.

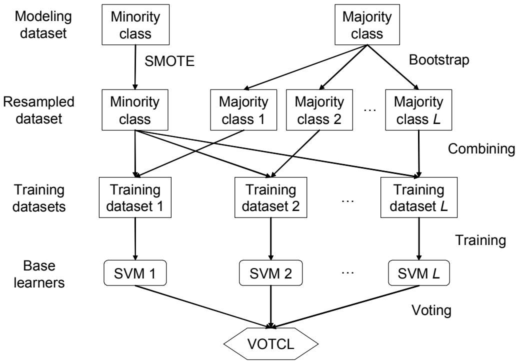


Fig. 1 Framework of VOTCL

Tab. 1 Process of VOTCL

```
Input:
$D$ : modeling dataset;
$P$ : prediction datasets;
$L$ : number of base learners (i.e. SVMs);
Process:
  $D^+ \leftarrow$ the positive subset of $D$;
  $D^- \leftarrow$ the negative subset of $D$;
  $D^+_{SMOTE} \leftarrow$ SMOTE( $D^+$ );
  for i = 1 to $L$ do
      $D^-_i \leftarrow$ Bootstrap( $D^-$ );
      $D^i \leftarrow D^-_i \cup D^+_{SMOTE}$;
      $SVM_i \leftarrow$ Training SVM on $D^i$;
  end for
Output $Label(x)$ & $R(x)$:
  $v(x) = \sum_{i=1}^{L} SVM_i(x)$;
  Select Optimal Threshold $K_{opt}$ by $F$ Measure;
  if $v(x) \geq K_{opt}$ then
      $Label(x) = 1$;
      $R(x) = \frac{1}{v(x)} \sum_{j=1}^{v(x)} prob_j^+(x)$
      (for $SVM_j(x) = 1$);
  else
      $Label(x) = 0$;
      $R(x) = \frac{1}{L - v(x)} \sum_{j=1}^{L-v(x)} prob_j^+(x)$
      (for $SVM_j(x) = 0$);
  end if
```

2.1 Resampling

Liu and Zhou [12] indicated that when dealing with a not-seriously imbalanced dataset and not-seriously unequal misclassification costs, learning based on natural class distribution will achieve a good performance; while if misclassification costs are seriously unequal or dataset is seriously imbalance, learning based on a balanced class distribution is more favorable. As the latter always presents in the datasets of real-world applications, we have to balance the class distribution by resampling first. However, both over-sampling and under-sampling have their own drawbacks when dealing with seriously imbalanced dataset [14], so we combine over-sampling and under-sampling to balance the dataset. It can be described as follows.

First, over-sample the minority class with SMOTE [5] to some extent; then under-sample the majority class with bootstrap [6] so that both sides are of the same size. The benefit of this process is that this approach inherits the strength of both SMOTE and bootstrap, and alleviates the over-fitting and information loss problems [14]. We bootstrap the majority class for L times (with the same size of majority class after SOMTE), and each under-sampled majority class is combined with the over-sampled minority class to form a training dataset.

Our resampling technique may be helpful when dealing with large datasets since it is adjustable of the size of the training datasets and the training of base learners can also be carried out in parallel.

2.2 Training of Base Learners

In this paper, we select support vector machine (SVM) [13] as the base learner, and radial base function (RBF) as kernel function. Base learners are trained on each balanced training datasets.

2.3 Voting based on Optimal Threshold

For a sample *x*, let $SVM_j(x)$ denotes the predicted label (0 for majority class and 1 for minority class) given by the *j*th SVM, $prob^+(x)$ denotes the corresponding posterior probability for minority class, *j*=1, 2,…, *L*. We propose voting based on optimal threshold to ensemble base learners, and the classification function *Label*(*x*) and regression function *R*(*x*) can be defined as follows.

(1) Classification function *Label*(*x*)

First, we calculate the voting number *v*(*x*) (i.e. the number of base learners which classify the given sample *x* to be minority class). Next, for each threshold *K* varies from 1 to *L*, generate a classification function as:

if $v(x) \geq K$, then *Label*(*x*) = 1;

else *Label*(*x*) = 0;

Then test each of those classification functions with *F* measure and the optimal threshold K_{opt} is estimated. At last, voting based on optimal threshold K_{opt} is applied to ensemble the base learners for decision-making.

On the criterion of *F* measure, increasing of *recall* means avoid misclassifying minority class samples as many as possible, and therefore, we prefer a high *recall* to a high *precision*, so the parameter β of *F* measure is set to be less than 1. The benefit of doing so is that the overall misclassification cost is minimized since the cost of misclassifying a minority class sample is substantially greater than the cost of misclassifying a majority class sample.

(2) Regression function *R*(*x*)

If VOTCL classifies a sample *x* to be minority class based on the final classification function, *R*(*x*) is obtained by averaging the posterior probabilities $prob^+(x)$ estimated by the SVMs which predict the sample *x* as minority class; otherwise, *R*(*x*) is obtained by averaging the posterior probabilities $prob^+(x)$ estimated by the SVMs which predict the sample *x* as majority class.

3. Experiments

3.1 Datasets

Experiments of VOTCL were carried out on the cross-selling problem of PAKDD competition 2007, and the real-world dataset was donated by a consumer finance company. The problem can be describes as follows.

The company currently has a customer base of credit card customers as well as a customer base of home loan customers. Both of these products have been on the market for many years, although for some reason the overlap between these two customer bases is currently very small. The company would like to make use of this opportunity to cross-sell home loans to its credit card customers, but the small size of the overlap presents a challenge when trying to develop an effective scoring model (a higher score indicates a greater possibility to become a home loan customer) to predict potential cross-sell take-ups.

Problem of CICS presents in the dataset of this cross-selling task.

(1) Class imbalance

In the modeling dataset given by the company, the ratio between the number of credit card customers (majority class) and the number of home loan customers (minority class) is 40,000:700, so seriously class imbalance exists in the modeling dataset.

(2) Cost-sensitive

From the potential interest of the consumer finance company, it may be anticipated that no potential home loan customers are scored low and the credit card customers are scored as low as possible, because a home loan customer may bring much more profit to the company than a credit card customer. It is obviously that producing a low score for a potential home loan customer is with a higher cost than producing a high score for a credit card customer, so seriously cost-sensitive exists in the modeling dataset.

3.2 Design and Analysis

3.2.1 Resampling

We over-sample the minority class with SMOTE to the amount of 2100. Then, we repeat bootstrap for 20 times (*L*=20) to generate 20 different majority class and each is combined with the over-sampled minority class to form a training dataset. There are 2100 samples in each bootstrapped majority class, and we get 20 different training datasets with the same size of 4200.

3.2.2 Training of Base Learners

Even if the datasets is balanced, we have to take the problem of cost-sensitive into account when training base learners. Given these circumstances, a model optimized on error rate alone may end up with building a useless model [2, 3, 11]. AUC [15] has proved itself a good evaluation criterion when dealing with class imbalance problem and cost-sensitive problem, so we choose AUC as the evaluation criterion to optimize parameters for base learners.

We test the two parameters *C* (cost parameter of SVM) and γ (parameter of RBF kernel function) of the

SVMs, and optimal parameters C and γ are estimated by AUC, for instance, C=21, γ=5 for SVM1.

3.2.3 Voting based on Optimal Threshold

By adjusting threshold K from 1 to 20 (L is 20 as mentioned above), we get 20 classification functions. As shown in Fig. 2, we test true positive fractions (TPF) and true negative fractions (TNF) of those classification functions. The corresponding F measures are shown in Fig. 3 with the parameter β=0.1. Fig. 3 reflects that the F measure achieves the highest value (0.99979) when K equals 15, so we set the optimal threshold K_{opt} to be 15 to ensemble the base learners for decision-making.

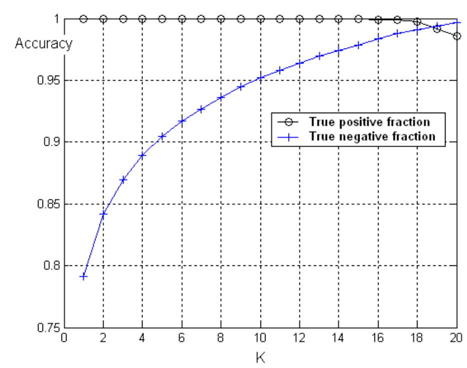


Fig. 2 TPF and TNF

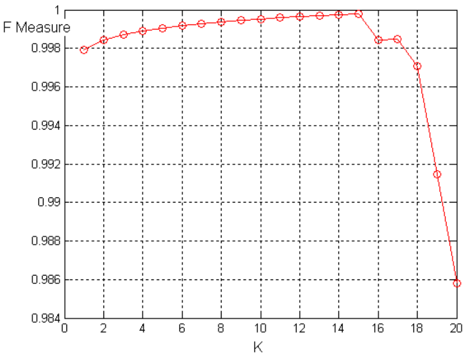


Fig. 3 F measure

3.2.4 Prediction and Scoring

The scores are given by the regression function $R(x)$, and Fig. 4 shows the distribution of scores for all the samples in modeling dataset. We find that most scores of the minority class samples (red points) are greater than 0.9, and the scores of the majority class samples (yellow points which are misclassified and blue points which are correctly classified) are less than 0.9 contrastively. Moreover, only 2.15% of minority class samples are misclassified. This indicates that the classification function is of high precision and the regression function is of good ranking.

Experiment on prediction dataset provided by PAKDD Competition 2007 shows the effectiveness of VOTCL with its AUC value 0.6037[3]. The AUC of the Grand Champion is 0.7001.

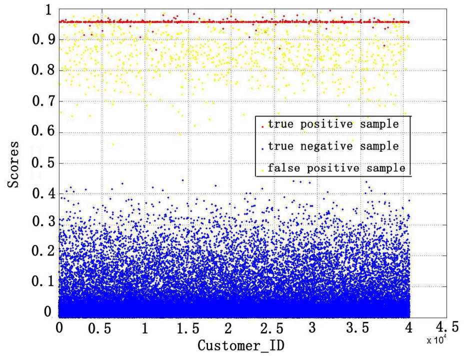


Fig. 4 Distribution of Scores

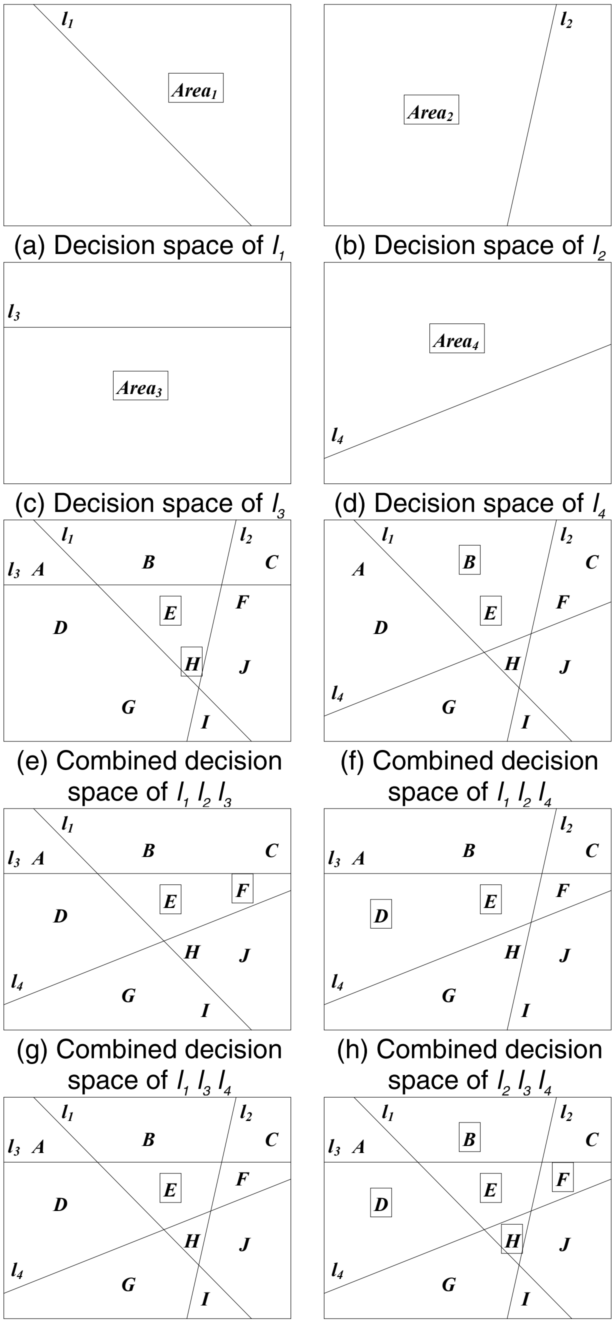
4. Decision-making Principle of VOTCL

The decision-making principle of VOTCL can be interpreted as follows. We can establish L hyper-planes since we train L different SVMs, and getting an ensemble of those SVMs with optimal threshold K_{opt} indicates that we combine at least K_{opt} hyper-planes to partition a subspace. K_{opt} or more hyper-planes can be combined to represent the decision space of minority class. By integrating various hyper-plan combinations, the decision space of VOTCL is dynamic when classifying a given sample. Hence, VOTCL is suitable for a more complex decision-making situation.

As shown in Fig. 5, there are four decision line l_1, l_2, l_3, l_4 with their corresponding decision space[4] $Area_1$, $Area_2$, $Area_3$, $Area_4$ (Fig.5 (a), (b), (c) and (d)). Three or more decision lines can be combined to represents the decision space of minority class if we have optimal threshold K_{opt} =3 (Fig.5 (e), (f), (g), (h) and (i)). The decision space of VOTCL is the union of all the decision spaces in Fig.5 (e), (f), (g), (h) and (i).

[3] The competition result can be seen on http://lamda.nju.edu.cn/conf/pakdd07/dmc07/results.htm, and the ID of our solution is P060.

[4] The letters with black border indicate the decision space of minority class.



(a) Decision space of l_1 (b) Decision space of l_2

(c) Decision space of l_3 (d) Decision space of l_4

(e) Combined decision space of l_1 l_2 l_3 (f) Combined decision space of l_1 l_2 l_4

(g) Combined decision space of l_1 l_3 l_4 (h) Combined decision space of l_2 l_3 l_4

(i) Combined decision space of l_1 l_2 l_3 l_4 (j) Decision space of VOTCL

Fig. 5 Decision-making Principle of VOTCL (Diagrammatic Sketch)

5. Conclusion and Future Work

A solution named VOTCL is proposed to address the problem of CICS: first, we generate several balanced training datasets by combining under-sampling and over-sampling techniques; then, they are trained to get base learners; at last, voting based on optimal threshold is proposed to ensemble those base learners for decision-making. Training based on balanced datasets can effectively weaken the impact of class imbalance and cost-sensitive, and the overall misclassification cost is minimized by voting based on optimal threshold.

VOTCL relaxes the requirements for class imbalance rescale ratio and cost-sensitive rescale ratio and it may be helpful when dealing with large datasets. It is a wrapper-based solution that we can choose different base learners flexibly according to the applications. Experiments on a real-world cross-selling problem also show the effectiveness of VOTCL. Furthermore, analysis of the decision-making principle indicates that our solution is suitable for a more complex decision-making situation by integrating various hyper-plane combinations.

We are working on the automatic optimization of the parameters used in training base learners. In future researches, we hope to test the performance of VOTCL when applied to other CICS problems.

6. References

[1] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification. 2nd Edition*, Wiley, New York, USA, 2000.

[2] S. Viaene and G. Dedene, "Cost-sensitive learning and decision making revisited", *European Journal of Operational Research*, 2005, 166(1), pp.212-220.

[3] C. Elkan, "The foundations of cost-sensitive learning", *In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, Morgan Kaufmann, Seattle, Washington, USA, 2001, pp. 973-978.

[4] G. Weiss, "Mining with rarity: A unifying framework", *SIGKDD Explorations*, 2004, 6(1), pp. 7-19.

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

[6] L. Breiman, "Bagging predictors", *Machine Learning*, 1996, 24(2), pp. 123-140.

[7] K.M. Ting, "An instance-weighting method to induce cost-sensitive trees", *IEEE Transactions on Knowledge and Data Engineering*, 2002, 14(3), pp. 659-665.

[8] S. Lessmann, "Solving imbalanced classification problems with support vector machines", *In: Proceedings of 2004 International Conference on Artificial Intelligence (IC-AI'04)*, CSREA Press, Las Vegas, Nevada, USA, 2004, pp. 214-220.

[9] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines", *In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)*, *Workshop ML3*, Stockholm, Morgan Kaufmann, Sweden, 1999, pp. 55-60.

[10] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data", *Technical Report 666*,

Statistics Department, University of California at Berkeley, 2004.

[11] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown", I*n: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, ACM, San Francisco, California, USA, 2001, pp. 204-213.

[12] X.Y. Liu and Z.H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study", *In: Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM'06)*, IEEE Computer Society, Hong Kong, China, 2006, pp. 970-974.

[13] V. Vapnik, "The nature of statistical learning theory", Springer, New York, USA, 1995.

[14] Y. Liu, A. An, and X. Huang, "Boosting prediction accuracy on imbalanced datasets with SVM ensembles", *In: Proceedings of the tenth Pacific-Asia Knowledge Discovery and Data Mining conference (PAKDD'06)*, *Lecture Notes in Computer Science (LNCS) 3918*, Springer, 2006, pp. 107-118.

[15] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, 1997, 30, pp. 1145-1159.