Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008

PARALLEL CLASSIFIERS ENSEMBLE WITH HIERARCHICAL MACHINE LEARNING FOR IMBALANCED CLASSES

YUN ZHANG, BING LUO

Faculty of Automation, Guangdong University of Technology, Guangzhou 510006, China
E-MAIL: yz@gdut.edu.cn, luobing8888@163.com

Abstract:

Imbalanced distributions and mis-classified costs of two classes made conventional classification methods suffered. This paper proposed a new fast parallel classification method for imbalanced classes. Considering imbalanced distributions, the approach adopted a fast simple classifier with less features input working parallel with a complicated one. Most samples would be correctly recognized by the first classifier, and the second relatively slower classifier could be ended. The second one was only trained and worked for less difficult samples. Experimental results in machine vision quality inspection showed that the approach could effectively improve classification speed and decrease total risk for imbalanced classes' classification.

Keywords:

Pattern recognition; Imbalanced classes; Hierarchical machine learning; Parallel processing; ROC

1. Introduction

In bi-classification pattern recognition, samples obey iid. $F(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ are samples features for input and y is the label of a sample with $y \in \{-1, +1\}$.

The target of the classification is to find best function $f(\boldsymbol{x}, \boldsymbol{\alpha})$ with parameters $\boldsymbol{\alpha}$ that make the classification risk minimized:

$$R(f, \boldsymbol{\alpha}) = \int L\left(y, f(\boldsymbol{x}, \boldsymbol{\alpha})\right) dF(\boldsymbol{x}, y) \qquad (1)$$

Here $L\left(y, f(\boldsymbol{x}, \boldsymbol{\alpha})\right)$ is the classification error cost function defined as:

$$L\left(y, f(\boldsymbol{x}, \boldsymbol{\alpha})\right) = \begin{cases} 0, if\ (\boldsymbol{x}, \boldsymbol{\alpha}) = y \\ 1, if\ y = -1, f(\boldsymbol{x}, \boldsymbol{\alpha}) = +1 \\ n, if\ y = +1, f(\boldsymbol{x}, \boldsymbol{\alpha}) = -1 \end{cases} \qquad (2)$$

In conventional classification research, it was suggested that two classes were balanced with n=1 in expression (2) above. However, in quality inspection, diseases diagnosis, data mining and knowledge discovery etc., classes are imbalanced, and n>>1.

Imbalanced classes referred to two classes for classification with two appears simultaneously [1].

(1) Distributions of two classes were imbalanced with over ten times difference at least:

$P_- >> P_+$, where $P_- = \int F(\boldsymbol{x}, y = -1)\, d\boldsymbol{x}$ and $P_+ = \int F(\boldsymbol{x}, y = +1)\, d\boldsymbol{x}$.

(2) Error costs were imbalanced:

$L(y = +1, f(x) = -1) >> L(y = -1, f(x) = +1)$.

That was n>>1 in expression (2).

The class with minority distribution and more error cost was called positive class and the other was called negative one [2].

Conventional classification methods made accuracy as machine learning target which had ignored error cost imbalance, so positive samples would be classified with more errors and total risk would increase [3].

Drawbacks of conventional classification methods with imbalanced classes aroused interests of researchers. Many proposals have been appeared in literatures to dealing with the problem. These proposals can be divided into three sorts: moving decision thresholds [2, 3], adjusting cost or weights [3, 4] and re-sampling [5-7].

The first method biased the minority by moving decision threshold, and similarly the second one improved minority class accuracy by enhancing the minority class error cost or samples weights. The third one pursued to decrease imbalance by down-sampling majority samples or up-sampling minority samples. However the first two methods adjusted these parameters mainly by trial and error. Hence, it is hard to build direct connections between the parameters and the biased classification accuracy quantitatively. Therefore, they cannot rigorously handle imbalanced data. For the sampling method, the problem is that up-sampling may introduce excessive weight on the

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

noise data and lead to overfitting, while down-sampling probably may lose some critical information. To solve this problem, Chawla et al. proposed synthetic minority over sampling technique (SMOTE) to introduce minority data points and remove redundant majority points "intelligently" [7]. Huang et al. proposed to use biased minimax probability machine that minimized the max misclassified probability of minority class with permitted majority error rate [8]. These methods are considered as one of the state-of-the-art approach for imbalanced-classes pattern recognition. Here, these researches mainly concerned the bi-classification, where the majority class was defined as negative class and the minority one as positive class.

On the other hand, many classification applications are required high speed for real time processing. Inspired by human's way of recognizing things, we proposed fast parallel classification method to deal with imbalanced classes' classification and improve classifying speed.

The rest of the paper was organized as below. In part 2, suffering of conventional methods caused by imbalanced classes were described and in part 3, classifiers performance evaluation for imbalanced classes was discussed. In part 4, we described fats parallel classification method and analyzed the method in part 5. Then experiments of machine vision quality inspection were taken to test performance of our approach in part 6. Finally we conclude the paper in part 7.

2. Imbalanced classes made conventional methods suffered

Conventional classification methods were based on balanced sample classes and made total classification accuracy *A* as evaluation index and machine learning target [9]:

$$A = \frac{N_T}{N} = \Pr\{y = f(\boldsymbol{x})\}, \tag{3}$$

Where N_T is the number of the samples which correctly classified, *N* is total samples number, *f* is classifying function and Pr stands for probability.

For balanced classes, error costs of two classes are the same. So maximizing total accuracy *A* is the same as minimizing the total risk *R*. However, for imbalanced classes, this would lead to bias for negative class and make important positive class more errors. Much higher error cost of positive class made conventional classifiers total error risk increased.

For an example, in automatic quality inspection, distribution of normal products and defect ones were: $P_- = 0.99$, $P_+ = 0.01$. Most products were normal. Classifier A simply classified all products as normal, and then its accuracy was $A_A = 99\%$. Classifier B could correctly classified out 50% defect products but with 5% normal ones mis-classified, that $A_{B-} = 95\%$ and $A_{B+} = 50\%$. Then its total accuracy was:

$$\begin{aligned} A_B &= P_- \cdot A_{B-} + P_+ \cdot A_{B+} \\ &= 0.99 \times 95\% + 0.01 \times 50\% = 94.55\% \end{aligned}$$

So machine learning in conventional classifiers designing would tend to classifier A which was meaningless with no defect product checked out. The total risks of two classifiers were:

$$\begin{aligned} R_A &= P_- \cdot (1 - A_{A-}) \cdot L_- + P_+ \cdot (1 - A_{A+}) \cdot L_+ \\ &= 0.01 \times 1 \times 20 = 0.2 \\ R_B &= P_- \cdot (1 - A_{B-}) \cdot L_- + P_+ \cdot (1 - A_{B+}) \cdot L_+ \\ &= 0.1495 \end{aligned}$$

Obviously the risk of classifier *A* was higher then that of *B*, which means conventional method suffered.

3. Performance Evaluation on Classifiers for Imbalanced Classes

The distribution function in expression (1) was usually unknown and was difficult to determine. So some practicable indexes must be transformed from expression (1) for machine learning and evaluation.

Four criteria were proposed in literatures, they were: (1) the minimum cost (*MC*); (2) the maximum geometry mean (*MGM*) of the accuracies on the majority class and the minority class; (3) the maximum sum (*MS*) of the accuracies on the majority class and the minority class; and (4) the receiver operating characteristic (*ROC*) analysis. We review these criteria as follows.

To bias minority class accuracy, Buses et al. proposed the criteria of maximum sum (*MS*) of the accuracies on the majority class and the minority class [9]. This criterion is also widely used in other fields as graph detection, especially line detection and arc detection, where it is called vector recovery index [10, 11]. Similarly, kubat et al. proposed to use the geometric mean instead of the sum of the accuracies [12]. However, compared to maximizing the sum, this criterion has a nonlinear form, which is not easy to be optimized. Besides, when the cost of misclassification is known, *MC* measure defined in (4) should be used [13]:

$$\cos t = F_p C_{F_p} + F_n C_{F_n}. \tag{4}$$

where F_p is the number of misclassified minority-class

(positive-class) samples, C_{Fp} is the cost of misclassifying a positive sample, F_n is the number of misclassified majority-class (negative-class) samples and C_{Fn} is the cost of one of this error.

However, because the cost of misclassification is generally unknown in real cases, the usage of this measure is somewhat restricted.

Considering this point, some researchers introduced the ROC analysis. [11, 14] This criterion plots a so-called ROC curve to visualize the tradeoff between the false-positive rate and the true-positive rate and leaves the task of the selection of a specific tradeoff to the practitioners.

The ROC curve is a graphical plot of the true positive rate (along the vertical axis) against the false positive rate (along the horizontal axis). It has been suggested that the area under an ROC curve (*AUC*) can be used as a measure of accuracy in many applications [11, 14]. Thus, a better classifier should have a larger *AUC*.

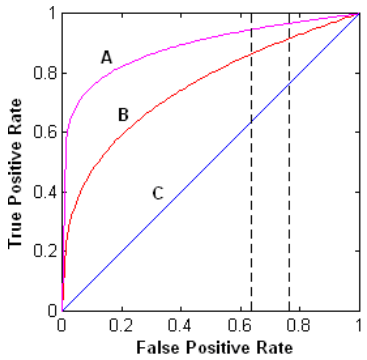

Figure1. ROC curve of classifiers

Fig. 1 shows ROC curves of three classifiers. It shows that classifier *A* is better than *B* and curve *C* is the worst with a random guessed result. Sometimes partial *AUC* shown in fig. 1 between two dash lines under curve was used to evaluate performance for some indexes were particularly concerned or some could not be reached.

In this paper, we use *ROC* analysis, *MC* and *MS* to evaluate performance of classifiers for imbalanced multiple classes.

4. Fast parallel classification method for imbalanced classes

Consider how human to classify a group of imbalanced samples. At first we will separate most common majority-class samples that are easily distinguished, and then we will use more information to do more careful separation on the rest. In the same way, we propose to do classification hierarchically to deal with imbalanced classes' problem and improve classification speed.

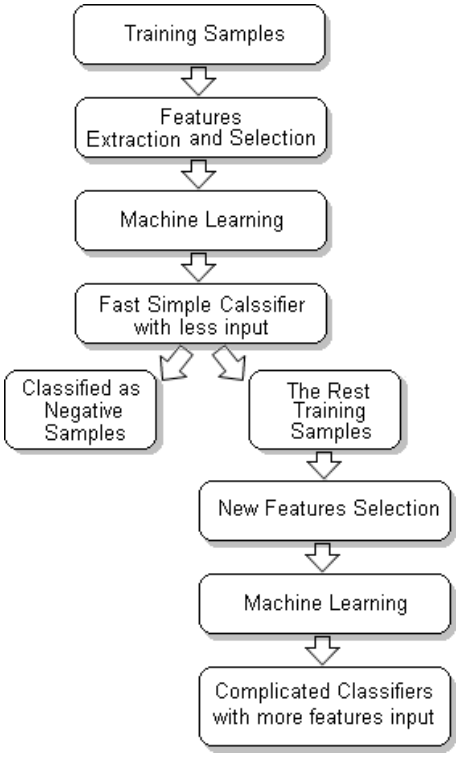

Figure2. Designing and Training for Parallel Classifiers

In fast parallel classification, two classifiers were working parallel. These two classifiers were designed and trained as fig.2. At first a simple classifier with less feature input was designed and trained to classified out most negative samples quickly, and maintain the positive ones error lower than allowed rate. And some negative samples which could be considered as difficult ones were allowed to be mis-classified and left for the other classifiers dealing.

Then another complicated classifier was designed and trained by the samples that the simple classifier classified as non-negative class. Classes distributions of these samples were almost tend to balanced and conventional methods could work and the target of this classifier was to get highest accuracy.

After designing and machine learning, two classifiers work parallel in classification as shown in fig.3. If a sample was classified into negative class by the fast simple classifier, the complicated classifier could be stopped immediately. For imbalanced classes, most samples were negative class, so most samples could be classified only by the fast classifier.

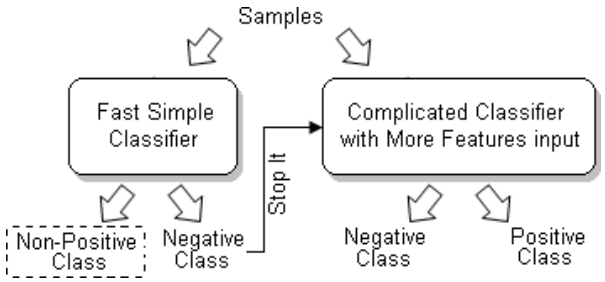

Figure3. Designing and Training for Parallel Classifiers

Only those difficult samples were classified by the complicated classifier with more features and high accuracy.

By this kind of parallel cooperation, classification speed was improved meanwhile error risk kept lower.

5. Performances analysis for fast parallel classification method

5.1. Machine learning effect

Conventional method would not biased important positive class. In our suggested fast parallel approach, learning of the fast simple classifier was to solve a optimal problem with restrict:

$$\max_{\alpha} A_n, \qquad \text{s.t. } A_p \geq A_\alpha . \tag{5}$$

Where α is the parameters of the classifier.

With less input features, simple structure and less parameters, this classifier had quick convergence in machine learning.

For the complicated classifier, imbalanced distribution in training set with rest samples which were classified as non-negative class by the simple classifier was largely relaxed. Negative samples were decreased and positive ones maintained, total samples number were decreased also. Now the machine learning would bias important positive class and implied conventional matured methods. Decreased samples number would reduce training computation.

5.2. Error risk

Set accuracy of a conventional single classifier for positive class as A_{S+}, accuracy of fast parallel classification as A_{P+}, positive class samples number as N_+, number of positive class samples error classified by the fast simple classifier as *a*, accuracy of the classifier for positive class as A_{1+}, number of positive class samples error classified by the complicated classifier as *b*, accuracy of the classifier for positive class as A_{2+}. Then we can get

$$A_{1+} = \frac{N_+ - a}{N_+}, \tag{6}$$

$$A_{2+} = \frac{N_+ - a - b}{N_+ - a} = 1 - \frac{b}{N_+ - a}, \tag{7}$$

$$\begin{aligned} A_{H+} &= \frac{N_+ - a - b}{N_+} = A_{1+} - \frac{b}{N_+} \\ &> A_{1+} - \frac{b}{N_+ - a} = A_{1+} + A_{2+} - 1 \end{aligned} . \tag{8}$$

Because imbalanced distribution were largely relaxed for the complicated classifier machine learning, the accuracy of the classifier would be increased: $A_{2+} > A_{S+}$. For the simple classifier, we can set least A_α as $A_\alpha = 1 - (A_{2+} - A_{S+})$, and it satisfied:

$A_{1+} \geq A_\alpha = 1 - (A_{2+} - A_{S+})$. Then

$$A_{P+} > A_{S+} . \tag{9}$$

So our approach can increase accuracy for important positive class.

For negative class, mis-classification in the simple classifier would not influence the accuracy of negative class, because the error classified negative samples would be left for the complicated classifier dealing. And the complicated classifier would have a normal accuracy for negative class. So total classification error risk would be decreased.

5.3. Classification speed

Suppose the time that a conventional classifier dealing a sample was t_S, then the total time for *N* samples was $T_S = N \cdot t_S$.

Suppose the time that the fast simple classifier dealing a sample was t_1, the complicated was nearly the same as conventional ones, the simple classifier had correctly classified N_{1-} negative samples, then the total time for *N* samples by our approach was:

$$\begin{aligned} T_H &= N_{1-} \cdot t_1 + (N - N_{1-}) \cdot t_s \\ &= N \cdot t_s - N_{1-} \cdot (t_s - t_1) \end{aligned} . \tag{10}$$

Because $t_1 < t_S$, we got $T_H < T_S$ and

$$T_S - T_H = N_{1-} \cdot (t_s - t_1), \quad (11)$$

$$\frac{T_S - T_H}{T_S} = \frac{N_{1-}}{N} \cdot \frac{t_s - t_1}{t_s}. \quad (12)$$

That means our approach can save total classification time.

5.4. Generalization ability

Generalization ability is an important performance of a classifier. Usually simple structure and less training samples would have relative better generalization ability [9].

In our approach, the fast simple classifier has a relative simple structure and less feature input, the complicated one has less training samples. So total generalization ability would be improved better than conventional method.

6. Experiments

To test our approach, we carried machine vision quality inspection for printed circuit board surface mounted technology assembly. In PCB SMT assembly line, default rate was always lower than 10%. The misclassification costs were imbalanced and the inspection speed should catch the assembly speed.

We took 1000 images of the same part from a PCB SMT assembly line product. 500 images selected randomly composed the training set and the others for testing set. After features selection, 5-10 features that correlation coefficients between inspected images conventional features with template images features were selected for inspection [15].

For comparison, we also inspected these samples in other three methods: Biased MPM [8], re-sampling [7] and directly ANN AdaBoost ensembles [16]. The cost of misclassification of normal class, the cost of defaults confusion and cost of misclassifying default-class to normal were assumed as 2:1:20 according to SMT assembly practice. Inspection time is counted for 500 testing samples.

The experimental results were listed in table 1 and their ROC curves were drawn in fig.4. The experimental results showed that our fast parallel classification approach had effectively improved the classification performance and accelerated inspection speed.

Table 1Comparison of Experimental Results

Methods	Ave Accuracy (%)	Cost	Time (s)
BMPM	85.65	68	14.17
Resample	89.57	61	14.38
AdaBoost	95.78	21	19.55
Our approach	97.83	9	9.62

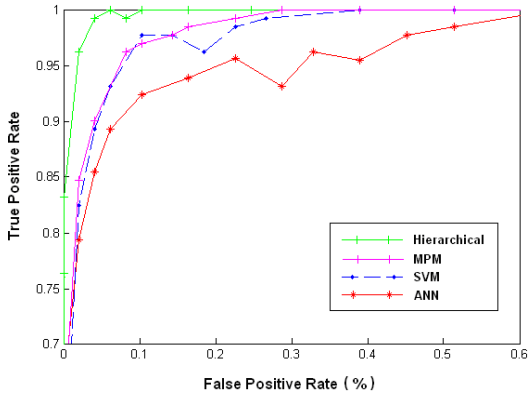

Figure4. Roc Curves of 4 Classifiers for AOI

7. Conclusions

Imbalanced classes with imbalanced distributions and error costs made conventional classification methods suffered. This paper proposed a fast parallel classification approach. Two classifiers with a simple one and a complicated one were trained serially but worked parallel. The fast simple classifier was designed with a simple structure and less features input. It's to classify out most negative samples. The complicated one was trained by left samples that the first one classified as non-negative class.

In parallel classification, for most negative samples, the fast simple classifier can quickly get negative class result and can stop the other classifier. Only those difficult negative samples and positive samples were classified by the complicated classifier. So classification speed and error risk can be improved.

Experimental results for quality inspection showed that our approach had effectively improved performance and accelerated speed.

Detail classifiers designing for their best cooperation and quantities analysis of our approach are needed further research.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China under projects U0735003, 60604006 and Research Fund for the Doctoral Program of

Higher Education (20070562005).

References

[1] Japkowicz N., Stephen S., "the Class Imbalance Problem: A Systematic Study", Intelligent Data Analysis, Vol 6, No. 5, pp. 429-450, 2002.

[2] Provost F., Fawcett T., "Robust Classification for Imprecise Environments", Machine Learning, Vol 42, No. 3, pp. 203-231, 2001.

[3] Provost F., "Machine Learning from Imbalanced Data Sets", Proc. of 17th Nat. Conf. AAAI, Workshop on Imbalanced Data Sets, Austin, TX, pp. 43-45, 2000.

[4] Sun Y., Kamel M., Wang Y., "Boosting for Learning Multiple Classes with Imbalanced Class Distribution", Proc. of 6th Int. Conf. on Data Mining, pp. 592-602, 2006.

[5] Maloof M.A., Langley P., Binford T.O. et al., "Improved Rooftop Detection in Aerial Images with Machine Learning", Machine Learning, Vol 53, No. 1, pp. 157-191, 2003.

[6] Cardie C., Howe N., "Improving Minority Class Prediction Using Case Specific Feature Weights", Proc. of 14th Int. Conf. on Machine Learning, Nashville, TN, pp. 57-65, 1997.

[7] Chawla N., Bowyer K., Hall L. et al., "Smote: Synthetic Minority Over-Sampling Technique", Artificial Intelligence Research, Vol 16, pp. 321-357, 2002.

[8] Huang K., Yang H. and King I. et al., "Imbalanced Learning with a Biased Minimax Probability Machine", IEEE Trans. on systems, man and cybernetics, Vol 36, No. 4, pp. 913-923, 2006.

[9] Vapnik, V.N., "Statistics Learning Theory", Beijing: Electronics Industry Press, 2004.

[10] Provost F., Fawcett T., "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions", Proc. of 3rd KDD, Newport Beach, CA, pp. 43-48, 1997.

[11] Bradley A, "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithm", Pattern Recognition, Vol 30, No. 7, pp. 1145-1159, 1997.

[12] Swets J., "Measuring the Accuracy of Diagnostic Systems", Science, Vol 240, No. 4857, pp. 1285-1293, 1988.

[13] Webb A.R., "Statistical Pattern Recognition", Beijing: Electronics Industry Press, 2004.

[14] Swets J., Pickett R., "Evaluation of Diagnostic Systems: Methods from Signal Detection Theory", New York: Springer-Verlag, 1982.

[15] Bing L., Yun Z., "ANN Ensembles Based Machine Vision Inspection for Solder Joints", Proc. of 6th IEEE Int. Conf. On Control and Automation, pp. 626-630, 2007.

[16] Sun Y., Kamel M.S., Wang Y., "Boosting for Learning Multiple Classes with Imbalanced Class Distribution", Proc. of 6th Int. Conf. on Data Mining, pp. 592-602, 2006.