# A Hybrid Re-sampling Method for SVM Learning from Imbalanced Data Sets

Peng Li
*College of Computer Science and Technology, Harbin University of Science and Technology*
*E-mail:pli@insun.hit.edu.cn*
Pei-Li Qiao
*College of Computer Science and Technology, Harbin University of Science and Technology*
Yuan-Chao Liu
*School of Computer Science and Technology, Harbin Institute of Technology*

## Abstract

*Support Vector Machine (SVM) has been widely studied and shown success in many application fields. However, the performance of SVM drops significantly when it is applied to the problem of learning from imbalanced data sets in which negative instances greatly outnumber the positive instances. This paper analyzes the intrinsic factors behind this failure and proposes a suitable re-sampling method. We re-sample the imbalance data by using variable SOM clustering so as to overcome the flaws of the traditional re-sampling methods, such as serious randomness, subjective interference and information loss. Then we prune the training set by means of K-NN rule to solve the problem of data confusion, which improves the generalization ability of SVM. Experiment results show that our method obviously improves the performance of the SVM on imbalanced data sets.*

## 1. Introduction

Imbalanced data sets (IDS) are usually objective data of observation in many real-world applications field, for a two-class classification problem, in which most of the instances belong to a larger class (majority) and far fewer instances belong to the other class (minority), and yet people usually focus on the occurrence of minority cases. (In the remainder of the paper negative is always taken to be the majority class and positive is the minority class) Example applications include vision recognition [1], bioinformatics[2], credit card fraud detection[3], the detection of oil spills[4] and so on. These studies and many others show that imbalanced data sets may result in poor performance of standard classification algorithms (e.g. decision tree, nearest neighbor and naïve bayes methods). The problem of IDS classification has been considered as a challenging direction by researchers in computer science. In recent years, there are four high-level international conference makes IDS as a special issue [5].

Support Vector Machine (SVM) is a powerful learning algorithm and promising results have been obtained in many fields, such as text categorization [6], vision detection and handwriting recognition [7] and so on. However, the success of SVM is very limited when it is applied to the problem of learning from imbalanced datasets. In this paper, we specifically chose SVM as learning algorithm to solve the problem of IDS because it has strong theoretical foundations and some empirical results show that it performs well with moderately imbalanced datasets even without any modification. The reason is that SVM only takes into account those instances that are close to the boundary, this principle means that SVM is unaffected by non-noisy negative instances far away from the boundary even if they are huge in number. Therefore, SVM may be the most suitable learning algorithm for IDS.

Resampling methods are commonly used for dealing with the class-imbalance problem since it straightforwardly balances the data at the stage of pre-processing. Their advantage over other methods is that they are external and thus, easily transportable. Two re-sampling strategies are popular used to adapt machine learning algorithm to IDS: under-sampling or down-sampling and over-sampling or up-sampling. Many researchers, suach as Batista, G [8], Estabrooks [9] and Japkowicz [10], proposes various approach about two re-sampling strategies. However, none of the approaches consistently outpreforms the other and it is also difficult to determine an ultimate conclusion in under-sampling or over-sampling strategy which consistently leads to the best results. We adopt under-sampling for our method because we agree to similar conclusions that effectively eliminate noisy instances can improve classifier performance. Meanwhile, under-sampling can also be used as a means to obviously reduce training set size. The major drawback from under-sampling is that it disregards possible useful information. Therefore, the goal of our method is how to eliminate huge noisy instances and try our best to reduce the loss of valuable information in the process.

The rest of the article is organized as follows. We analyze the influencing factors of SVM learning from

978-0-7695-3305-6/08 $25.00 © 2008 IEEE
DOI 10.1109/FSKD.2008.407

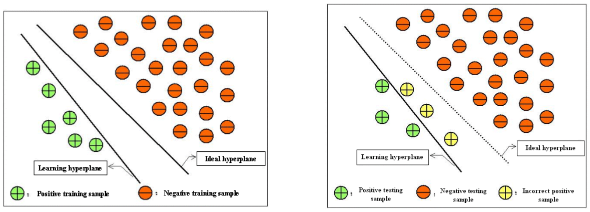
IEEE computer society

imbalanced datasets in the next section. Then we put forward an under-sampling method based on variable SOM clustering in section 3. K-NN sample-pruning algorithm is introduced in section 4. Finally, experiments and conclusions are described in section 5 and 6.

## 2. Influencing factors of SVM learning from imbalanced datasets

Imbalanced datasets have two inner factors, namely, imbalance ratio (IR) and lack of information (LI). Imbalance ratio is the value of *Number of Majority/ Number of Minority* and LI is the lack of information for the monority class. For a data set consisting of 100:10 majority : minority examples the imbalance factor IR is the same as in a data set of 1000:100, but the intuition implicate us there are several defference in them. In the first case the minority class is poorly respresented and suffers more from the LI factor than in the second case. Both the above inherent factors are present in every IDS learning problem, in combination with other external factors, such as overlap, complexity, size of the data and high dimension etc.

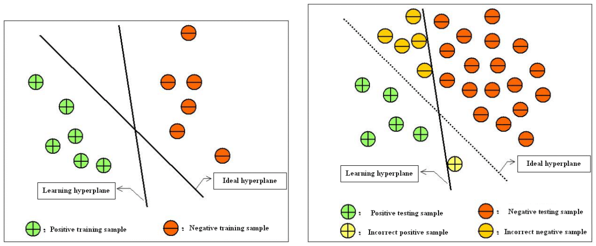
We theoretically analyze Influencing factors by means of linear separable imbalanced datasets. In figure 1(a), SVM learning from an imbanlanced data set, the result show that the learning hyperplane has almost the same orientation as the ideal hyperplane, but the distance of the learning hyperplane is far away from the ideal hyperplane. Furthermore, learning hyperplane is too close to the positive support vectors. In figure 1(b), the learning hyperplane will lean toward the negative instances and mis-identify the positive instances to negative ones in the process of testing. The phenomemon is data-whelming, when the training data gets more imbalanced, the ratio between the positive and negative support vectors also becomes more skewed. They dicide the learning hyperplane far off the ideal hyperplane, the neighborhood of a testing instances close to the boundary is more likely to be dominated by negative support vectors and hence the decision function is more likely to classify a boundary point negative and led to the majority whelm the minority.



(a) (b)

Figure 1 The phenomemon of data-whelming

If we randomly under-sampe the majority instances of imbalance training data, until their numbers are equal to the minority instances in gross. In figure 2(a), we can see that the learning hyperplane is close to ideal hyperplane but the orientation of the learning hyperplane is no longer accurate. In figure 2(b), the learning hyperplane will lead to severely wrong classification result in the process of testing. This phenomemon is information loss. The reason is that mass negative instances are cut down randomly lead to many valuable information is lossing, the remainder negative instances can no longer give good cues about the orientation of the hyperplane and there is a greater degree of freedom for the orientation to vary.



(a) (b)

Figure 2 The phenomemon of information loss

From above analysis, we get the guiding principle to our re-sampling method that is to find a strategy to filter large number of majority instances, which are far away from the target boundary, without losing too many minority instances. This allows us to concentrate on distinguishing the more difficult boundary instances and reduce the imbablance ratio which makes the learning task more tractable.

## 3. A under-sampling method based on variable SOM clustering

Under-sampling methods are commonly used for dealing with the class-imbalance problem and the key is to reduce mass majority instances with little information loss. We adopt the variable SOM clustering to solve the paradox problem. Unlike various under-sampling techniques, clustering will split imbalanced training data based on their distribution into meaningful clusters. If instances of a cluster are all negative instances, then we will eliminate the cluster.

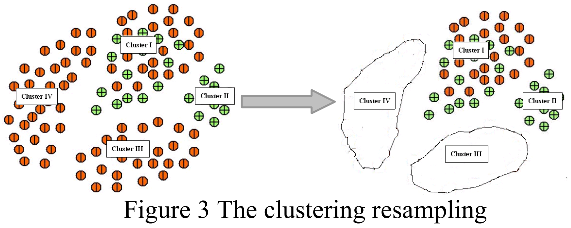


Figure 3 The clustering resampling

The self-organizing map (SOM) introduced by Kohonen is an unsupervised neural network method which has both clustering and visualization properties. . It can be considered as an algorithm that maps a high dimensional data space, to a lower dimension, generally 2, and which is called a map. Hence, SOM is very suitable to deal with the high dimensional data in real world. This projection enables the input data to be partitioned into "similar" clusters while preserving their topology. In this paper, we adopt variable SOM is to avoid the phenomenon of under-utilization due to the neuron expansion and overcome the problem of boundary effect because of rectangular or other structure [11]. The adjustment function of neuron threshold is shown in following:

$$n_j(t+1) = n_j(t) + \varphi(t)\bullet r_j(t)\bullet dis(x_i, n_j(t)) \quad (1)$$

$$dis(x_i, n_j(t)) = 1 - sim(x_i, n_j(t)) \quad (2)$$

$n_j(t+1)$ and $n_j(t)$ respectively denote the weight vector of neuron $n_j$ after and before adjustment. $\varphi(t)$ is the learning rate function and $r_j(t)$ is the neighborhood function, both are gradual decrement in the process of training. $dis(x_i, n_j(t))$ means the distance between sample vector $x_i$ and neuron vector $n_j(t)$ and we can get the value by means of similarity computation. In general, similarity is calculated by cosine formula.

$$sim(x,n) = \frac{\sum_{i=1}^{l} W_{x_i} W_{n_i}}{\sqrt{\sum_{i=1}^{l} W_{x_i}^2}\sqrt{\sum_{i=1}^{l} W_{n_i}^2}} \quad (3)$$

In formula (3), $l$ denotes the dimension of vector, $W_{x_i}$ means the weight value of sample vector $x$ in the $i$ th dimension, $W_{n_i}$ is the weight value of neuron vector $n$ in the $i$ th dimension.

We adopt $R^2$ clustering criterion coefficient as judgment guideline , and seek the balance between under-utilization and over-utilization. We assume $m_i$ is the corresponding vector of neuron $N_i$ , then the sum of deviations squares of mapping sample of $N_i$ is

$$S_i = \sum_{x_j \to N_i} dis(x_j, m_i) \quad (4)$$

In the time of $t$ , we hypothesize there are $c$ neurons in output layer, and then define $P_c = \sum_{k=1}^{c} S_k$ . T is used to denote the general sum of deviations squares of all samples. Then,

$$T = \sum_{i=1}^{|D|} dis(x_i, \bar{x}) \quad (5)$$

$\bar{x} = \frac{1}{|D|}\sum_{i=1}^{|D|} x_i$ is the mean vector of all training samples and $|D|$ is the total number of input sample. Then,

$$R^2 = 1 - \frac{P_c}{T} \quad (6)$$

the range of clustering criterion coefficient $R^2$ is 0 to 1 and the value is monotonically increase according to the increasing of net scale. Therefore, we set threshold $\mu$ to terminate the increasing of net in order to avoid the under-utilization.

## 4. A sample-pruning algorithm based on K-NN rule

In section 2, the imbalance data set, which is used to analyze influencing factors, is linear separable imbalanced datasets. However, these datasets are not existing in real world. The phenomenons of overlap and complexity are always existing in the field of real application as shown in figure 4. If we want to solve the problem of application, these actual cases must be taken into account. Therefore, we propose a sample-pruning algorithm to improve these environment.

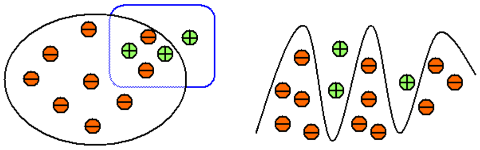


Figure 4 The overlap and complexity of data

SVM constructs an optimal hyperplane from a small set of samples near the boundary. If the training data have the phenomenons of overlap and complexity, those samples intermixed in another class are seriously will result in the number of support vector will increase greatly and the performance of training and classification will become worse, instead they may greatly increase the burden of computation and their existence may lead to overlearning and decrease the generalization ability.

Li Hong-Lian et,al. proposed a NN-SVM method to solve above problem and gain a good result [12]. It first prunes the training set, reserves or deletes a sample according to whether its nearest neighbor has same class label with itself or not, then trains the new set with SVM to obtain a classifier. In imbalance data sets, however, it is more serious for negative instances to overlap in positive instances. In this way, there is an attendant phenomenon in noisy data. As shown in figure 5, two negative instances associated existing and the method of nearest neighbor is invalidation in this case.

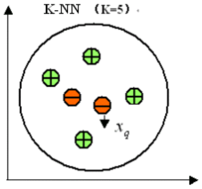


Figure 5 The attendant phenomenon of noisy data

We present a sample-pruning algorithm based on K-NN to solve this difficulty. The kernel thinking is to exam every sample and its K nearest neighbors, then calculate the predominant value of current inquiry sample and judge the computational attributive is same to self-attributive or not. In real application, we adopt different threshold to negative and positive due to the positive instance is sparser than negative instance, so the positive information is more valuable. In imbalance data sets, moreover, the number of negative instances intermixes in positive is more than the number of positive instances intermixes in negative. Hence, the thinking of different control-threshold is apt to delete the negative intermixed instances and possibly reserve the positive information.

We assume all sample correspond to the node of $n$ dimension space and adopt standard Euclidean distance as the distance of two vectors.

$$\langle \alpha_1(x),\ \alpha_2(x),\ \dots,\alpha_n(x)\rangle \qquad (7)$$

where $\alpha_k(x)$ denotes the $k$ th attribute of instances $x$. Then the distance between $x_i$ and $x_j$ is defined

$$d(x_i,x_j)=\sqrt{\sum_{k=1}^{n}(\alpha_k(x_i)-\alpha_k(x_j))^2} \qquad (8)$$

Assuming the value of class attribute is $f(x_i)\in\{1,-1\}$, predominant attribute value of inquiry instance $\psi(x_q)$ is calculated by

$$\psi(x_q)=\frac{\sum_{\mathrm{i=1}}^{k} f(x_i)}{k} \qquad (9)$$

## 5. Experiment

People usually focus on the performance of minority classification in real IDS application. We apply three evaluation criteria, which are precision, recall and F-measure of minority, to evaluate classifiers on imbalanced datasets. The functions are given below.

$$\Pr ecision=\frac{correct\ positive\ inst.}{correct\ positive\ inst.+incorrect\ negative\ inst.} \qquad (10)$$

$$\operatorname{Re}call=\frac{correct\ positive\ inst.}{correct\ positive\ inst.+incorrect\ positive\ inst.} \qquad (11)$$

$$F-measure=\frac{2\times\Pr ecision\times\operatorname{Re}call}{\Pr ecision+\operatorname{Re}call} \qquad (12)$$

We used 3 different UCI datasets and a MUC dataset with varying degrees of class imbalance in table 1. Each datasets was randomly split into train and test sets in the ratio 5 to 5, while sampling them in a stratified manner to ensure each of them had the same negative to positive ratio.

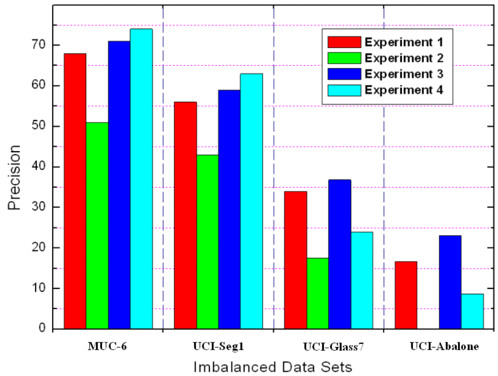
Table 1 Data sets of experiment

| Datasets | Negative Insts. | Positive Insts. | Imbalance Ratio |
|---|---|---|---|
| MUC-6 | 159815 | 11266 | 14.2:1 |
| UCI-Seg1 | 1980 | 330 | 6:1 |
| UCI-Glass7 | 185 | 29 | 6.4:1 |
| UCI-Abalone | 4145 | 32 | 130:1 |

In our experiments, we compared the performance of our classifier with regular RBF kernel SVM and design four experiments to exam the performance of different strategies:

- Experiment 1: Classification without resampling
- Experiment 2: Classification with random resampling
- Experiment 3: Classification with resampling based on variable SOM clustering
- Experiment 4: Classification with resampling based on variable SOM clustering and K-NN pruning

The results of our experiments are given below.



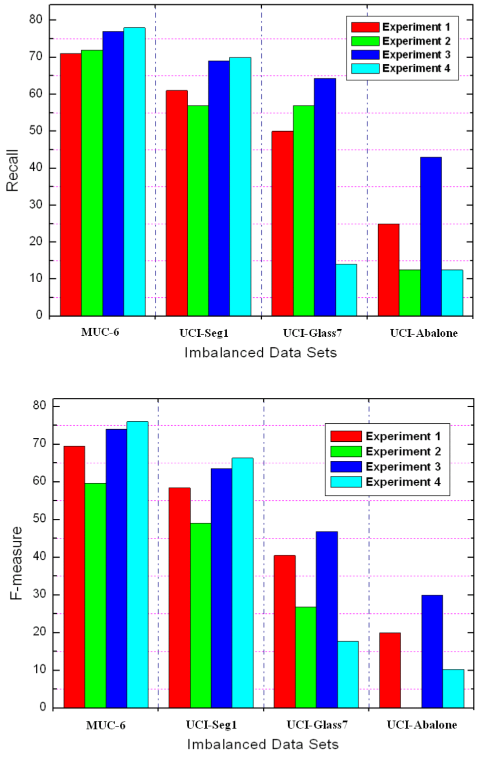


Figure 6 The results of four different experiments

The results show that our method obviously improves the performance of classification. However, our sample-pruning algorithm is invalidation in the data sets of UCI-Glass7 and UCI-Abalone. In particular, the imbalance ratio of UCI-Seg1 is same to UCI-Glass7, but the effect is differential to them. We analyze the reason is because of the lack of information. Therefore, the lack of information is also a very important factor in the problem of IDS classification.

## 6. Conclusion

In this paper, we present a hybrid method for SVM classifier to learn from Imbalanced data sets. We analyze the intrinsic factors behind this failure and propose a suitable re-sampling method. We re-sample the imbalance data by using variable SOM clustering so as to overcome the flaws of the traditional re-sampling methods, such as serious randomness, subjective interference and information loss. Then we prune the training set by means of K-NN rule to solve the problem of data confusion, which improves the generalization ability of SVM. The results of experiment show that our method obviously improves the performance of IDS classification.

There are several possible areas for future work. It is potential for us to improved performance through more up-to-date machine learning methods and sophisticated use of NLP techniques. In particular, we will explore a effective way to solve the influence of the lack of information.

## 7. References

[1] Maloof, M. Learning when data sets are imbalanced and when costs are unequal and unknown, *Proceedings of the Workshop on Learning from Imbalanced Data Sets.* 2003

[2] Liu Guo-ping, Yao Li-xiu and Yang Jie, Solving the Problem of Imbalanced Dataset in the Prediction of Membrane Protein Types Based on Weighted SVM, *Journal of Shanghai Jiaotong University*, 2005, pp. 1676-1684.

[3] Chan, P. and Stolfo, S. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection, *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining,* 1998. pp. 164-168.

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

[5] Sofia Visa, Anca Ralescu, Issues in Mining Imbalanced Data Sets - A Review Paper, *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference,* MAICS-2005, Dayton, April 16-17, 2005, pp. 67-73.

[6] Joachims, T. . Text Categorization with SVM: Learning with Many Relevant Features. *Proceedings of ECML-98, 10th European Conference on Machine Learning.* 1998

[7] Liu Han, Guo Yong, et al. Edge detection based on least squares support vector machines, *Acta Electronica Sinica,* 2007, 34(7) pp. 1275-1279.

[8] Batista, G.; Prati, M., and Monard, M. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations* 2004.6(1) pp.20–29.

[9] Estabrooks A, Jo TH, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence,* 2004,20(1). pp.18-36.

[10] Japkowicz, N. Learning from imbalanced data sets: A comparison of various strategies. *In Proceedings of Learning from Imbalanced Data,* 2000. pp.10–15.

[11] LIU Yuan-Chao. Research of Text Clustering Based on Dynamic Self-Organizing Maps Model. *Ph.D thesis*. 2006.

[12] LI Hong-Lian, WANG Chun-Hua and YUAN Bao-Zong. An Improved SVM: NN-SVM. *Chinese Journal of Computer.* 2003,26(8) , pp.1015-1020.