Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey, May 27-29, 2006 (pp76-79)

GPCA Method for Fraud Detection in Mobile Communication Networks

WANG Da-zhen, Wan Fang
Computer Science Department
Hubei University of Technology, Wuhan 430068
P.R.China
http://www.newtelesoft.com

Abstract: - To improve the fraud detection accuracy by SVM(support vector machine), a feature extraction method named GPCA based on IG (information gain) and PCA (principal component analysis) is proposed. It analyzes the data on CDR(call detail record), customer information，paying and arrear information etc in mobile communication networks , and then the data can be used by the classifier SVM to build the fraud detection model and the user can predict the potential fraud customers. Despite of its simplicity, GPCA outperforms some of the most popular feature extraction methods such as BS (bivariate statistics), IG and PCA in predicting accuracy and training time. To get the higher predicting accuracy, a binary SVM using RBF (Radial Basis Function) kernel is used. The experiments show that the classifier with GPCA has fine predicting accuracy.

Key-Words: PCA; GPCA; SVM;

1. Introduction

Fraud in mobile communication networks has been a serious problem to the network operator. Statistics shows that the loss due to fraud in mobile communication networks was even 1 billion last year. Fraud may be defined as a dishonest or illegal use of services, with the intention to avoid service charges. With the aid of the fraud detection models, fraudulent activity in a mobile communications network may be revealed. This is beneficial to the network operator, who may lose several percent of revenue to fraud, since the service charges from the fraudulent activity remain uncollected.

The traditional technological method used in fraud detection in mobile communication networks in developed countries is classification in data mining, which mainly includes Decision Tree and Neural Networks. However, many Decision Tree algorithms are designed to solve large sample problem and Neural Networks has the disadvantage of excessive learning. To solve the arrear problem with small sample, Support Vector Machine is first chosen. SVM is initiated by Vapnik in 1995. It is designed to solve the two classes problems in pattern recognition. A non-linear SVM is used to classify the data. The input to the non-linear should be a low dimensional vector for reducing the non-linear functions to be calculated. According to this, a feature extraction step must be included prior to the classifier. Additionally, a low dimension input space can improve the generalization of the classifier [3].

The objective of this paper is to evaluate different feature extraction methods that are designed to improve the classifier's accuracy. Next is a general fraud detection model in mobile communication networks.

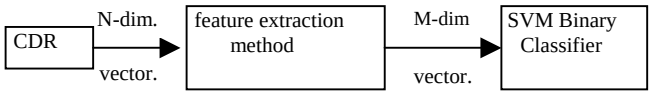

Fig.1 A general fraud detection model in mobile communication networks

2. Feature Extraction

In order to reduce the dimensionality of the raw feature space, several methods are proposed. Four procedures will be analyzed: BS(bivariate statistics), IG(information gain), PCA(principal component analysis) and a new method (named GPCA) mainly combining PCA and IG.

2.1 Bivariate Statistics

BS shows that the correlation between two variables reflects the degree to which the variables are related. When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from -1 to +1. A correlation of +1 means that there is a perfect positive linear relationship between variables.

2.2 Principal Component analysis

PCA is widely used in signal processing, statistics, and neural computing. The basic idea of PCA is to find the components $c_1, c_2, \cdots, c_n$, so that they explain the maximum amount of variance possible by linearly transformed components. PCA can be defined in an intuitive way using a recursive formulation. Define the direction of the first principal component, say w_1, by:

$$w_1 = \arg\max_{\|w\|=1} E\{(w^T x)^2\} \qquad (1)$$

where w_1 is of the same dimension m as the random data vector X. Thus the first principal component is the projection on the direction in which the variance of the projection is maximized. Having determined the first k-1 principal components, the k-th principal component is determined as the principal component of the residual:

$$w_1 = \arg\max_{\|w\|=1} E\{[(w^T (x - \sum_{i=1}^{k-1} w_i w_i^T X)^2]\} \quad (2)$$

The principal components are then given by $s_i = w_i^T X$. In practice, the computation of the w_1 can be simply accomplished using the (sample) covariance matrix $E\{XX^T\} = C$. The w_i are the eigenvectors of C that correspond to the n largest eigenvalues of C

2.3 Information Gain

Information Gain is a measure based on Entropy, given a finite set E of classified examples and a set $C=\{c_1,\ldots,c_n\}$ of possible classes. Let $|E_{c_i}|$ denote the number of examples of class c_i and let $p_i = |E_{c_i}|/|E|$, Then the entropy of E is defined as:

$$\text{entropy}(E) = -\sum_{i=1,\ldots,n} p_i * \log_2(p_i) \qquad (3)$$

The Entropy measures the impurity of a set of examples. It is lowest, if there is at most one class present, and it is highest, if the proportions of all present classes are equal. Given a partition $P=\{E_1,\ldots,E_n\}$ of , then the Information Gain is defined as:

$$\text{ig}(E,P) = \text{entropy}(E) - \sum_{i=1,\ldots,n} \text{entropy}(E_i) * |E_i|/|E| \quad (4)$$

Information Gain measures the decrease of the weighted average impurity of the attributes $E_1,\ldots,E_n$ compared with the impurity of the complete set of examples E [6]. The feature selection algorithm computes the information gain of every attribute $E_1,\ldots,E_n$. The attribute which has the higher information gain should be kept. Given a threshold, the first higher m attributes can be selected as the new features.

As it can be easily seen, these three methods are very powerful for finding a small set of features from a bigger one in order to have the highest amount of useful information to face up the classifier. However, each method has its own advantage and disadvantage. Both BS and IG make use of the labels, however, they have no consideration of the relationship between the attributes, while PCA has no consideration of the label.

3. SUPPORT VECTOR MACHINE

Recently, SVM has become a very effective method in statistical machine learning. SVM is a type of model that is optimized so that prediction error and model complexity are simultaneously minimized [2].

Consider a set of data points, $\{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}$, such that $x_i \in R^d$ is an input and y_i is a target output. A SVM is a model that is calculated as a weighted sum of kernel outputs. There are mainly three kinds of kernels that are currently studied. The first is polynomial kernel:

$$K(x, x_i) = [(x \cdot x_i) + 1]^q \qquad (5)$$

What can get is a q rank polynomial classifier. The second is RBF function:

$$K(x, x_i) = \exp\left[\frac{\|x - x_i\|^2}{\sigma^2}\right] \qquad (6)$$

The classifier is different from the traditional RBF in that its basic function center corresponds to a support vector whose output and weight are decided by the algorithm automatically. The third is the sigmoid function:

$$K(x, x_i) = \tanh(v(x \cdot x_i) + c) \qquad (7)$$

Any other function that obeys Mercer's condition can also serve as a kernel. Thus, the output of an SVM is either a linear function of the inputs, or a linear function of the kernel outputs.

Because of the generality of SVM, they can take forms that are identical to nonlinear regression, radial basis function networks, and multilayer perceptrons. The difference between SVM and other methods lies in the objective functions that they are optimized with respect to and the optimization procedures that one uses to minimize these objective functions. In the linear, noise-free case for classification, with $y_i \in \{-1, 1\}$, the output of an SVM is written as

$f(x,\omega) = x \cdot \omega + \omega_0$, and the optimization task is defined as:

$$\text{minimize} \quad \frac{1}{2}\|\omega\|^2$$

$$\text{subject to} \quad y_i(x \cdot \omega + \omega_0) \geq 1 \quad \forall i \qquad (8)$$

Intuitively, this objective function shows the notion that one should find the simplest model that explains the data. This basic SVM framework has been generalized to include slack variables for miss-classifications, nonlinear kernel, regression, as well as other extensions for other problem domains. Generally, SVM has three characteristics. Firstly, SVM is based on the statistical learning theory and minimizes structural risk rather than the empirical one. Secondly, the maximal margin algorithm is given, which is only concerned with the operation of dot product. Finally, the kernel method is applied to solve nonlinear problems. Then training SVM is converted to find the solution of Quadratic Programming [3].

4. GPCA

Please, follow our instructions faithfully, otherwise you have to resubmit your full paper. This will enable us to maintain uniformity in the conference proceedings as well as in the post-conference luxurious books by WSES Press. The better you look, the better we all look. Thank you for your cooperation and contribution. We are looking forward to seeing you at the Conference.

As a part of this family of feature extraction methods, based on IG and PCA, a simple and effective procedure named GPCA is proposed.

PCA deals well in finding the directions of the input space in which the components have the higher energy. These directions are orthogonal, and are obtained by the simple process of eigen-decomposition of the covariance matrix. However, it makes no use of the label. Our aim is now to find, by applying IG, the attributes which have the higher information with the label. (Fig.2). As a result, the algorithm training time can be saved, and the predicting accuracy can be improved.

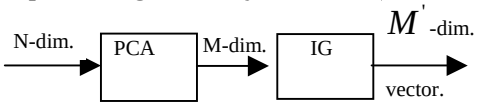

Fig.2 GPCA scheme based on PCA and IG

Next the GPCA algorithm is given:

Given $X=\{x_1,\ldots,x_n\}$

Step1: compute the mean $\overline{x_i}$ and variance δ_i of the train samples, and standardize the training samples, get the resulted samples $X' = (x_1', x_2', \ldots, x_n')$, where $x_i' = \frac{(x_i - \overline{x_i})}{\delta_i}$.

Step2: compute all the eigenvectors $p_i (i = 1,\ldots,M)$ and eigenvalue $\lambda_i (i = 1,\ldots,M)$

Step3: sort the eigenvalues, such that every eigenvector p_i keeps the largest λ_i of the residual variance

Step4: compute the contributing ratio of every eigenvector p_j with the formula $\lambda_j / \sum_{i=1}^{n} \lambda_i$, give a threshold R of the ratio and decide the vectors by R which should be kept.

Step5: compute the new vector's information gain by formula (4), give a threshold G of information gain and decide the final vectors by G which should be kept.

5. Experiments

The raw data in our experiment consist of 55 attributes most of CDR attributes including calling counts, normal (time) calling counts, international calling counts, total fee etc. The data samples can be divided into two classes, one for training and the other for testing. There are 4615 samples in training data where there are 882 positive samples in our data, and there are 2139 samples in testing data where there are 455 positive samples.

The SVM package used for the experiments is LIBSVM. This package is under active development and has several advantages over other packages. LIBSVM is developed by Chih-Chung Chang and Chih-Jen Lin [Chang and Lin, 2001] and its features include parameterized kernel, different SVM formulations (variable optimization algorithms), and multi-class classification.[3]

In order to evaluate the performance of these methods, an analytic criterion named accuracy was used. Accuracy is the ratio of the number of the right classified samples to the number of the wrong classified samples. However, when one class samples are fewer than another class samples, the accuracy would be doubtful, because it can't explain the exact samples the classifier can identify. If the negative samples are far more than the positive numbers, then sensitivity and specificity can be used as a measurement of accuracy. Sensitivity and specificity can be defined as follows:

$$sensitivity = \frac{t\_neg}{neg} \qquad (9)$$

$$specificity = \frac{t\_pos}{pos} \qquad (10)$$

where t_pos is the number of the true classified positive samples， pos is the number of the positive

samples， t_neg is the number of the true classified negative samples， neg is the number of the negative. Then accuracy can be defined by sensitivity and specificity [1] as follows:

$$accuracy = sensitivity \bullet \frac{\text{neg}}{(pos+neg)} + specificity \bullet \frac{\text{pos}}{(pos+neg)} \quad (11)$$

Table 1 is the comparison of predicted results for each feature extraction method with SVM classifier.

	BG & SVM	IG& SVM	PCA& SVM	GPCA & SVM	Total samples
predicted samples(0)	1309	1394	1498	1537	1684
predicted samples(1)	324	367	402	412	455

Table1 Comparison of predicted results for each feature extraction method with SVM

The predicted accuracy can be computed by formula (7), and Fig.3 gives a comparison of the predicted accuracy with four classifiers. Fig.3 shows that the classifier GPCA&SVM has higher accuracy comparing with other classifiers.

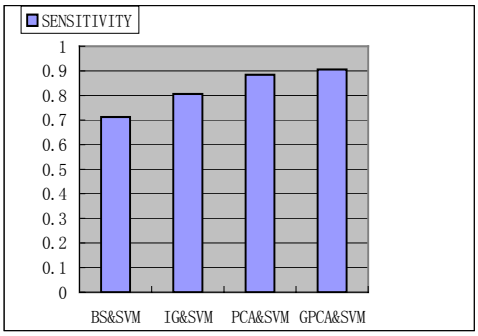

Fig.3 Comparison of the predicted accuracy with four classifiers

Also an experiment for the algorithm's training time is done. Fig.4 compares the training time with four classifiers. Fig.4 clearly shows that GPCA&SVM has faster training time.

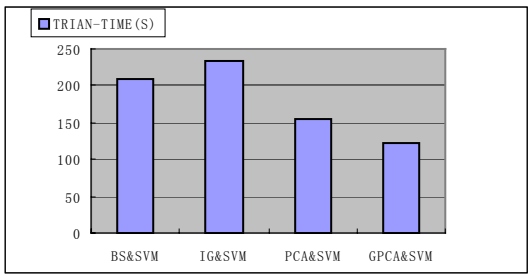

Fig.4 Comparison of the training time with four classifiers

6. Conclusions

Each feature extraction method such as BS, IG, and PCA has its own advantage and disadvantage. GPCA was proposed just by combining the advantage of IG and PCA. The experiments in fraud detection in mobile communication networks show that GPCA has fast training efficiency and fine predicted accuracy with the actual data. However, how to decide the threshold and how to optimize the parameters of SVM classifier are still questions, and there is still much work to do.

References:

[1] Jiawei Han Micheline Kamber Data Mining Concepts and Techniques 2001

[2] Vapnik.V. The Nature of Statistical Learning Theory [J]. Springer Verlag. 1995.

[3] CHANG Chih-Chung, LIN Chih-Jen.Libsvm: a library for support vector machines (version 2.3)[M]. 2001.

[4]WANG Jia-qi. Kernerl Projection Algorithm for Large-scale SVM Problems [J]. Comput.Sci &Technol .2002.

[5] TRAN Quang-anh, ZHANG Qian-li, LI Xing. SVM classification-based intrusion detection system [J]. Journal Of China Institute of Communications.2002

[6]http://kiew.cs.uni-dortmund.de:8001/mlnet/instances/81d91eaa-da13f5785e