INTELLIGENT SYSTEM FOR DETECTION OF ABNORMALITIES AND PROBABLE FRAUD BY METERED CUSTOMERS

Abdul Rahim Ahmad,
Fariq Izwan Ismail, Fadhilah Abdul Razak
Universiti Tenaga Nasional, Malaysia
abdrahim@uniten.edu.my

Abdul Malik Mohamad
Tenaga Nasional Research, Malaysia
m.malik@tnrd.com.my

ABSTRACT

This paper describes an on going project in Tenaga Nasional Berhad, (TNB) Malaysia to create an intelligent system to detect fraud by customers of SESB, a subsidiary of TNB in Sabah, East Malaysia. It presents a methodology to obtain a list of abnormal users from the customer database using a popular and recent intelligent method, the support vector machine (SVM). Instead of spending a lot of money on inspection campaign on all customers, a list of likely abnormal (fraud) customers will be generated and checked. First, an SVM model is constructed using samples of verified customer list. Then, the model is used for the detection of fraud customers from the 400,000 customers.

INTRODUCTION

Power companies loose revenue from both technical and non-technical failures. Non-technical losses are normally caused by the consumers and it can be circumvented by remote monitoring of the devices at customers sites for malfunction or tampering. This is possible for cases of new installations. However, most existing installations do not have remote monitoring feature. Monitoring customers are then done by using their consumption data to categorize them into normal or abnormal.

This paper presents a methodology to obtain a list of abnormal users from the customer database using a popular and recent intelligent method. The aim of the detection of abnormal customers can be for detecting fraud cases or just for customer consultancy purposes. It may also be used for detecting and correcting equipment errors such as meter malfunction. Instead of spending large amount of money on inspection campaign on all customers, a reduced group of likely abnormal customers can be generated.

This is achieved by modelling the consumption profiles of each customer using Support Vector Machine (SVM). A list of verified customer of normal and abnormal categories is used to construct the model which will be used for the overall detection. The specific application of this methodology is under implementation at Sabah Electricity Sendirian Berhad (SESB), a subsidiary company of the Malaysian power distributor, Tenaga Nasional Berhad (TNB). Readers can refer to other earlier similar works which are based on non-supervised ANN [1], decision tree and rough set [2].

This paper is organized as follows: the following section describes about fraud detection in general. This is followed by the description of SVM and its uses in the third section. Usage of SVM in fraud detection for SESB is described in the fourth section. Results are then discussed in Results section. The papers wraps up with a suggestion and conclusion section.

FRAUD DETECTION

Overview

Fraud is defined by The Concise Oxford Dictionary as ‘criminal deception; the use of false representations to gain an unjust advantage'. Fraud can either be prevented or detected. In prevention, steps are taken to discourage fraud, while in detection; measures are not taken to prevent but only to detect it for taking action against the perpetrator [3]. Fraud occurs in all fields of everyday life, to name a few; education, banking, electricity and telecommunication businesses. In electricity business, fraud mainly involves tempering with meters by consumers in order to reduce the consumption bill.

An electric meter - a kilowatt-hour meter record the reading which is used to generate an invoice of the electricity consumption. Many meters in use are electromechanical induction meters which can easily be tampered. Newer solid state meters uses remote current-carrying conductors that can also record many other load parameters such as maximum demand, power factor and reactive power used etc. Most importantly, they are more difficult to tamper.

The newer automatic electricity meter reading (AMR) technology allows automatic collection of data from electricity metering devices and transferring that data to a central database for billing. This allows billing on actual consumption rather than on an estimate. It also eliminates the need for meters to be visually read. In Europe, solid state meters and AMR meters are popular. However, in Malaysia, electromechanical induction meters are still largely in use which causes a concern in term of tampering and fraud.

Fraud in Power consumption

Some electrical power consumers temper their meters to cause it to under-register and thus, pay less. This is fraud, and illegal in most countries. The meter is normally sealed so that it cannot be tampered-with without breaking the seal. Given the tamper resistance in meters, some consumers attempt fraud by bypassing the meter, wholly or in part, to use the power without it being recorded at the meter.

Power companies also can investigate discrepancies between the total billed and the total generated. These investigations are effective in discovering tampering. Anti tamper techniques are also well-known in the industry, but were not widely applied in developed countries because tampering was rare. AMR meters often have sensors that can detect tempering. These features are not available on all meters, though, and it could be catastrophic for grid operators and utilities if information about these weaknesses would be wide-spread, since the objective with AMR meters is that no visit at the meter is required, which would allow permanent tempering to meters to not be noticed.

In this project, we are mainly concern about tampering in developing countries, in particular Malaysia where latest meter technology is not widely applied and tampering can be widespread. In this case fraud detection relies fully on detection of abnormal patterns in meter readings and consumption profiles of consumers. As such it is a pattern recognition problem and techniques in pattern recognition can be applied similarly.

SUPPORT VECTOR MACHINE

Introduction

Support Vector Machine (SVM) [4] is a method in pattern recognition and classification. It is a classifier to predict or classify patterns into categories; in the case of electricity consumption, between good or fraudulent customers. As any artificial intelligence tool, it has to be trained to obtain a learned model. It has been used in many classical pattern recognition problems such as text categorization, image, objects, face or speech recognition and in power load prediction etc.

SVM have been introduced initially as a two-class classifier which can be an alternative to neural network. As opposed to neural network, SVM theoretical framework makes use of training data and structural behaviour together to give a good classifier. SVM uses the principle of structural risk minimization (SRM) which aims to maximize the margin of class separation. Thus, SVM classifier is also known as a large margin classifier.

Basic SVM formulation is meant for linearly separable problems. However, many real life problems such as our fraud problem are non-linear in nature. In this case, with a small modification, SVM can be used by using kernel functions to indirectly map the non-linear input space to a linear feature space where the maximum margin decision function is approximated. The problem of outliers is handled through soft-margin SVM formulation. The general SVM formulation is non-linear soft-margin SVM in which linear and hard-margin problem are special cases.

SVM Theory

A quick and brief idea of SVM follows: We have a set of N training data pair: $\{(x_1, y_1), (x_2,y_2),\ldots, (x_n, y_n)$ where $x_i \in R^d$ and $y_i \in \{\pm 1\}$. The x's are training data and y's the classes. We want to learn a classifier: $f(x) = \text{sgn}(w.x + b)$ which have the maximum separating margin with respect to the two classes. Specifically, we want to find a plane: H: $y = w.x + b = 0$ and two hyper planes parallel to it and with equal distances to it, H_1: $y = w.x + b = +1$ and H_2: $y = w.x + b = -1$ with the condition that there are no data points between H_1 and H_2, and the distance between H_1 and H_2 is maximized. See figure 1.

For any separating plane H and the corresponding H_1 and H_2, we can always normalize the coefficients vector w so that: H_1 be $y = w.x + b = +1$, and H_2 be $y = w.x + b = -1$. We want to maximize the distance between H_1 and H_2. So there will be some positive examples on H_1 and some negative examples on H_2. These examples are called support vectors because only they participate in the definition of the separating hyper plane, and other examples can be removed and/or moved around as long as they do not cross the planes H_1 and H_2.

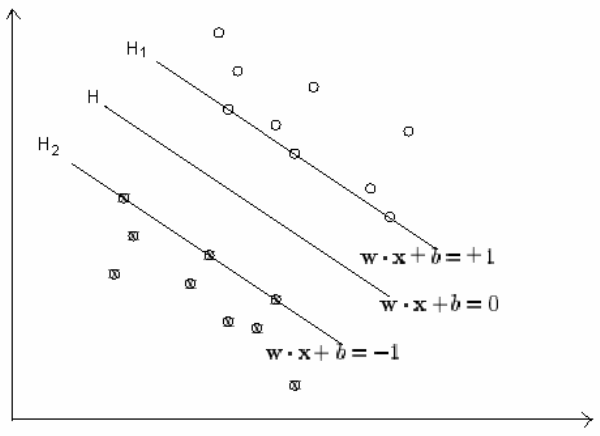

Figure 1 SVM as Maximal margin classifier

The distance between H_1 to H is $\frac{|wx+b|}{\|x\|} = \frac{1}{\|w\|}$ and thus

between H_1 and H_2 is $\frac{2}{\|w\|}$. Therefore to maximize the margin, we need to minimize $|w| = w^T w$ with the condition that no data points between H_1 and H_2 satisfy:

$w.x + b \geq +1$ for positive examples $y_i = +1$,

$w.x + b \geq +1$ for negative examples $y_i = -1$.

The two conditions can be combined into $y_i(w.x + b) \geq 1$. So, our problem can be formulated as $\min_{w,b} \frac{1}{2} w^T w$ subject to $y_i(w.x + b) \geq 1$. This is a convex, quadratic programming problem (in *w* and *b*), in a convex set which can be solved by introducing Lagrange multipliers $\alpha_1, \alpha_2, \ldots, \alpha_N \geq 0$. Thus we have the following Lagrangian:

$$L(w,b,\alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i y_y (w.x_i + b) + \sum_{i=1}^{N} \alpha_i \qquad (1)$$

We can now maximize $L(w,b,\alpha)$ with respect to α, subject to the constraint that the gradient of $L(w,b,\alpha)$ with respect to the primal variables w and b vanish: $\frac{\partial L}{\partial w} = 0$ and $\frac{\partial L}{\partial b} = 0$ and that $\alpha \geq 0$. We then have

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

Substitute them into $L(w,b,\alpha)$, we have

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \qquad (2)$$

in which the primal variables w and b are eliminated. Using the available data and its known classes, we solve for α_i,. Then, substituting into $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ we get w, and our decision function is:

$$f(x) = sgn(w.x + b)$$
$$= sgn((\sum_{i=1}^{N} \alpha_i y_i x_i).x + b) \qquad (3)$$

In prediction, a newly unseen data from unknown class is used together with the α_is and the corresponding support vectors to calculate f(x).

As mentioned earlier, in the case of non-linearly separable input space, data inputs can be mapped to another high dimensional feature space that the data points will be linearly separable. If the mapping function is *Φ(.),* we just solve:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(x_i) \Phi(x_j) \qquad (4)$$

Generally, if the dot product $\Phi(x_i).\ \Phi(x_j)$ is equivalent to a kernel $k(x_i, x_j)$, the mapping need not be done explicitly. Thus, the equation above can be replaced by:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \qquad (5)$$

Using the kernel in input space is equivalent to performing the map into feature space and applying dot product in that space. There are many kernels that can be used that way. Any kernel that satisfies Mercer's condition can be used.

There are a few possible kernels that can be chosen :

(a) Polynomial kernels: $K(x,y) = (x.y + 1)^d$,

(b) Radial basis function (RBF) kernel:

$$K(x\ y) = e^{\frac{-|x-y|^2}{2\sigma^2}}$$

(c) Hyperbolic tangent kernel: $K(x,y) = tanh(ax.y - b)$.

In the case of imperfectly separable input space, a penalty C is enforced if data points cross the boundaries. Using similar formulation as in linear case, we obtained the same dual Lagrangian but with a different constraint for α_i,, which is $0 < \alpha_i < C$ where C is the penalty. The value of σ obtained is constraint to be positive for perfectly separable case and between 0 and C in the case of non-linearly separable data. The value C is the penalty term and needs to be chosen prior to training the SVM. RBF kernel is the most popular kernel for SVM. Choosing C and σ for the kernel, constitute parameter selection which can be handled by various methods. Pok et. al. [5] discusses a method to choose these parameters using genetic algorithm (GA).

Practical SVM Usage

There are two tasks in the use of SVM; training and prediction. Training involves solving the convex quadratic programming problem. The final solution solves for nonzero parameters α in the formulation and extracts a subset of training data corresponding to the parameter, as support vectors (SV). These α's and the SVs constitute the SVM model that will used during prediction.

A number of methods of SVM training have been developed over the years. SVM have been made popular by the availability of stable implementation packages. There are a few implementation packages available publicly and have been popularly used as reported by many researchers. Among them are LIBSVM, SVMTorch and SVMLight. We have adopted the LIBSVM package [6] in our project.

METHODOLOGY

Data Collection

SESB currently have close to 0.4 million customers. The estimated overall distribution loss (technical and fraudulent) is about 15%. About 7.5% of the total customers have been checked by visiting selected customer premises in the last 2 years at some cost. Of the total premises checked an estimate of only about 6% have been confirmed to be fraudulent. With the use of the proposed intelligent system, we hope to generate a possible fraud list to be checked that gives a high strike rate which is much better than the current manual method of 6%.

In developing an intelligent system, verified data is needed. Thus, we have made use of the verified data from the 7.5% customers that have been checked. During inspection exercise, premises had been randomly chosen. Upon inspection, customers were categorized into good or fraud categories. A summary of the categorization grouped by the device number is submitted to the head office in hard copy forms for record and analysis. We have used these forms and as a result, save a lot of effort in data collection for our SVM modelling. However, since the data provided by the inspection does not contain the history of meter readings they need to be extracted from the SAP customer database, maintained at the head office. Since we know the good and the bad customers by their device numbers, we only need to provide the information to the customer management centre for them to generate text files containing the monthly readings for the customers we listed. We have extracted a few sets of customers data based on a few criteria: (a) random, no specific criteria (b) according to consumer category (commercial or domestic) and (c) by location/region. For this paper we only reported the results based on the data collected without specific criteria.

Features Extracted

Using the generated output from the SAP system, electricity consumptions for the last few years can be calculated for each customer. In order to obtain the features for SVM training, we have used the electricity consumption for the last 85 months. In the case of less than 85, we have decided to use a number of methods to project the missing data: (a) by filling with zeros (b) by using the average values for the customer and (c) by using the average profile of all customers. For SVM parameter selection, we have used cross validation and GA. The result using GA is reported in another paper [5]. In validating our models, we have used a few different data sets for training and validation, depending on the initial availability and also based on sizes. We hereby, give the validation results based on cross validation.

RESULTS AND DISCUSSION

Table 1, gives the summary of results obtained in the course of training and validation of the SVM model. The table gives the percentage of recognition using 10-fold cross validation for the different sample data that was used. Training and validation was done using the LIBSVM package. All programs are coded in C++. The initial dataset of 332 customers was manually collected independently from the set of checked and verified customers. The other three dataset were taken from the verified list of customers.

Table 1

Training Data	Validation accuracy using 10-fold cross validation
332 customers initial set (158 + /174 -)	76.51 %
190 customers verified set (94 + /96-)	93.12%
2000 customers verified set (1200 + / 800 -)	73.4 %
13000 customers verified set (12200 + / 800-)	68.56 %

CONCLUSION

In this paper we have described the background in customer fraud prediction using support vector machine. The steps taken in implementing a predictor for classifying fraud customers in electricity power business were shown. The result of > 60% indicates that prediction accuracy of the system is very promising which can give much saving in cost of inspecting customer premises.

ACKNOWLEDGMENTS

The authors would like to thank Tenaga Nasional Berhad and Tenaga Nasional Research for providing the funds for this project. The authors are also indebted to following colleagues for their valuable assistance in the project: Ir. Dr Zahrul Faizi Hussien, Dr Izham Zainal Abidin, Yap Keem Siah, Jason Pok Hooi Loong and Prof. Dr Wan Ahmad Tajuddin.

REFERENCES

[1] J.R. Galván, A. Elices, A. Muñoz, T. Czernichow, M.A. Sanz-Bobi; System for Detection of Abnormalities and Fraud in Customer Consumption; 12th Conference on the Electric Power Supply Industry; November 2-6, 1998, Pattaya, Thailand

[2] J. E. Cabral, E.M. Gontijo, Fraud detection in electrical energy consumers using rough sets, 2004 IEEE International Conference on Systems, Man and Cybernetics, 10-13 Oct. 2004.

[3] Richard J. Bolton, David J. Hand, Statistical Fraud Detection: A Review, January 2002.

[4] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Kluwer Academic Publishers, Boston, 1998.

[5] H. L. Pok, A. H. Hashim, A. M. Mohamad, K. S. Yap, Z. F. Hussein, I.Z.Abidin, Abnormalities And Fraud Electric Meter Detection Using Hybrid Support Vector Machine And Modified Genetic Algorithm, CIRED 2007, Vienna, Austria, May 2007

[6] Chang, C. C., Lin, C. J. (2001). "LIBSVM: a library for support vector machines. ." from http://www.csie.ntu.edu.tw/~cjlin/libsvm.