ABNORMALITIES AND FRAUD ELECTRIC METER DETECTION USING HYBRID SUPPORT VECTOR MACHINE AND MODIFIED GENETIC ALGORITHM

Hooi Loong POK
Keem Siah YAP
Izham Z.ABIDIN
COE, UNITEN, Malaysia
jason@uniten.edu.my
yapkeem@uniten.edu.my
izham@uniten.edu.my

Amir Hisham HASHIM
Zahrul Faizi HUSSIEN
PEC, UNITEN, Malaysia
amir@uniten.edu.my
zahrul@uniten.edu.my

Abdul Malik MOHAMAD
TNBR, Malaysia
m.malik@tnrd.com.my

Abstract: This paper presents a new intelligent system to detect abnormalities and fraud electric meter using hybrid Support Vector Machine (SVM) and Modified Genetic Algorithm (MGA). The main motivation for this study is to reduce the distribution loss (technical and non-technical), estimated around 15% at present in Sabah State, Malaysia. The hybrid algorithm is able to pre-select suspected customers to be inspected on-site for abnormalities or potential fraud according to their consumption patterns and other characteristics. SVM is relatively a novel classification technique and it has shown higher performance than traditional learning methods in many applications. A practical difficulty of using SVM is the selection of parameters such as C and kernel parameter, σ in Gaussian RBF kernel. The purpose of choosing parameters is to get the best generalization performance. Modified Genetic Algorithm (MGA) is used to search for the best parameter of SVM classification by using combination of random populated genomes and genomes from Pre Populated Database (PPD). In MGA, Dynamic Fitness-Based Crossover (DFBC) operator is used for a better evolutionary approach to the optimization problem that has been an issue in SVM implementation. Training and testing of the algorithm have been carried out by using actual customer consumption data from Sabah Electicity Sdn Bhd. The representation approach has been implemented via a computer program in order to achieve optimized performance.

Key words: Support Vector Machine, Modified Genetic Algorithm, Dynamic Fitness-Based Crossover, Pre Populated Database, Dual Lagrangian Optimization

INTRODUCTION

One of the main initiatives in Sabah Electricity Sdn. Bhd. (SESB) is to reduce the distribution loss (technical and non-technical), estimated around 15% at present. This research will only focus on reducing non-technical losses (NTL). NTL can be defined as losses due to meter malfunction, billing error and fraud. In order to combat NTL, inspections are carried out randomly without any specific criterion or direction due to unavailability of a system that can shortlist the abnormalities and possible suspects from the Customer Information Billing System (CIBS). Therefore, this research proposed an intelligent system which uses hybrid SVM-MGA to solve these problems.

The main objective of this intelligent system is to assist the utility to increase the effectiveness of their inspections. The developed system will provide the tools for non-technical loss classification and detection based on the available customer data and complements the already existing on-going actions. SVM is a new pattern recognition technique developed by Vapnik et al. [1-2]. The basic idea of SVM is to design a linear classifier with maximal classification margin, while minimizing the training error. Maximizing the margin plays the role of capacity control so that the learning machine will not only have small empirical risk but also hold good generalization ability [1-2]. SVM have been successfully applied to a number of applications ranging from particle identification, face identification and text categorization to engine-knock detection, bioinformatics, and control.

A practical difficulty of using SVM is the selection of parameters such as tradeoff parameter, C and kernel parameter, σ in Gaussian RBF kernel. The purpose of choosing parameters is to get the best generalization performance. Tuning these parameters is usually done by minimizing the estimated generalization error [3-5] or using gradient descent methods [6] to do iterative searching.

For any classification task, performance of the classification will deteriorate if hyper-parameters are not well chosen. Therefore, picking the best optimized values for parameters is very important. In this research, MGA is used to perform the optimization searching of the parameters. MGA is a hybrid algorithm modified from typical Genetic algorithm (GA). Two significant modifications in MGA are the use of DFBC and PPD. GA is a stochastic search techniques based on the mechanism of natural selection and natural genetics. GA with its proven characteristics of high efficiency and global optimization are widely applied in many areas [7].

PARAMETERS OF SVM

This section briefly introduces the concept of SVM. Given training data extracted from SESB, $(\mathbf{x_1}, y_1)\ldots(\mathbf{x_i}, y_i)$, where

$\mathbf{x_i}$ are input vectors and y_i are the associated output values of $\mathbf{x_i}$ which indicate good or fraud. Please refer Methodology section for details.
The optimal hyperplane is obtained by maximizing the margin between the two classes (refer Figure 1) and the hyperplane is subjected to the constraint:

$$y_i [\vec{w} \cdot \vec{x}_i + b] \geq 1, \text{ for all i} \quad (2.1)$$

$\vec{w}$ is termed as the weight vector and b termed as the threshold or bias

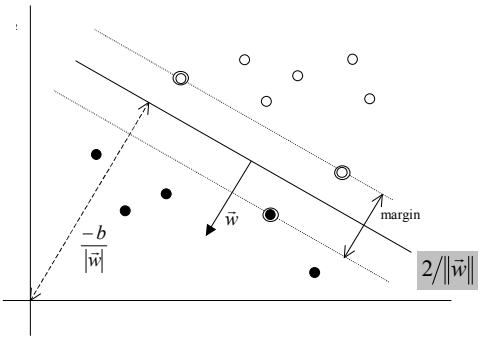

Figure 1 Maximum margin concept

Because $\|\vec{w}\|^2$ must be minimized to maximize the margin of the hyperplane, dual variables Lagrangian, L is introduced by imposing the Karush-Kuhn-Tucker (KKT) conditions and by introducing the Lagrange multiplier α_i

$$L = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (2.2)$$

Subject to constraints:

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0 \quad (2.3)$$

$$0 \leq \alpha_i \leq C, \; i = 1,2,3,...,\ell \quad (2.4)$$

Kernel function is used in equation 2.2 to map the dot product of input vector $\mathbf{x_i}$ into a high dimensional feature space. This can avoid the mathematical complexity of dot product calculation. L under the constraints of (2.3) and (2.4) will obtained the Optimal Separating Hyperplane(OSH). Input data that can satisfy $\alpha_i > 0$ are called Support Vectors (SVs). The optimal decision function is as follows:

$$f(x) = sign\left(\sum_{i=1}^{\ell} y_i \alpha_i K(x, x_i) \right) \quad (2.5)$$

Cristiani et al. [6] proposed the Kernel-Adatron (KA) Algorithm which can automatically select models without testing on a validation data. Unfortunately, this algorithm is ineffective if the data has a flat ellipsoid distribution [2]. This might be often happened in the real world case. Unlike the KA Algorithm, this study developed a new method named SVM-MGA to optimize the all parameters of SVM simultaneously. The real-valued genetic algorithm was adapted to determine the optimal values of SVM parameters. In the proposed SVM-MGA model, the SVM parameters are dynamically optimized via MGA evolution process and the SVM model then performs the prediction task using these optimal values. The process of SVM-MGA is illustrated in Fig. 2. The MGA tries to search for the optimal values to enable SVM to fit various datasets. The GA-SVM was developed and coded in the Microsoft Visual 6 C++ environment.

GENETIC ALGORITHM OPTIMISATION

Lagrange parameter selection for SVM is complex in nature and quite hard to solve by conventional optimization techniques. Among three evolutionary algorithms: genetic algorithm (GA), evolutionary programming (EP), and evolutionary strategy (ES), GA is perhaps the most widely known type of evolutionary algorithms today. GA has been widely used in a variety of fields such as Traveling Salesman Problem (TSP) and Network Optimization [8].

Chromosome Representations

Unlike traditional GA which uses binary coding for chromosome, all corresponding real-valued parameters are directly coded to form a chromosome. Hence, the representation of the chromosome is straightforward in MGA. All alpha, sigma and C, of SVM were directly coded to form the chromosome. Consequently, the chromosome $\boldsymbol{X}$ was represented as $\boldsymbol{X_n} = \{\alpha_1, \ldots \alpha_i, p_1, p_2\}$, where i is the number of features, p_1 and p_2 denote the C and σ (the parameters of the kernel function), respectively.

Genetic Operators

Three important operators in a typical GA are selection, crossover, and mutation operators. They are used to generate and evaluate the offspring of the existing population. The proposed GA-SVM model incorporated a well-known selection method which is called Roulette Wheel method. In this method, each chromosome is given a slice of the circular roulette wheel. The area of the slice is proportional to the chromosome fitness ratio R_f , and it is calculated by the formula,

$$R_f = \frac{f(i)}{\sum_{i=1}^{n} f(i)} \times 100\% \quad (3.1)$$

where $f(i)$ = fitness of the i^{th} chromosome

The bigger the fitness ratio is, the larger the area of the slice is. This will increase the probability of fitter chromosomes

to be crossover. Once a pair of chromosomes has been selected for crossover, one or more randomly selected positions are assigned into the to-be-crossed chromosomes. The newly-crossed chromosomes then combine with the rest of the chromosomes to generate a new population. In this experiment, uniform mutation was used in the presented model.

Dynamic Fitness-Based Crossover

In this paper, a new alternative crossover operator named the Dynamic Fitness-Based Crossover (DFBC) was developed, and its performance evaluated for a variety of input sizes. The results indicate that the use of the proposed operator has a marked influence on the time necessary to converge on the best combination of alpha to maximize the objective function.

By using the newly proposed DFBC operator, an earlier checking has been executed at the beginning of each population before crossover. Initially, the pivot point for crossover, which is dynamically located, will be determined. Then, the components at each side of the pivot point will be randomly shuffled only in their own group. Similar process applies to the other chromosome before the crossover takes place. Crossover rate will be dynamically adjusted based on best L generated. If the best L consistently stays at same value for 10 times then cross over rate will be reduced by 0.1. This process helps to avoid local optimum and increase the convergence speed [7].

Pre Populated Database

The main objective of Pre Populated Database (PPD) is to reduce the searching space of MGA and to get a better gene for the same set of problem. Generally, training of the SVM to obtain optimal solution is time consuming. The proposed database will store 10 best genomes from the initial training process. During the subsequent training process, 10 best genomes will be mixed with the random populated genomes. All genomes are evaluated by MGA and the 10 fittest genomes will be restored to the PPD. This method is relatively easy to implement and a significant speedup can be expected. More convincingly, the fitness produced via this approach in every instance, out-performed the other standard evolutionary algorithm mechanism.

METHODOLOGY

Raw data obtained from SESB is in text file format and it describes the monthly reading of each customer. The meter reading obtained range from 3 to 6 years. For this experiment, a total of 190 SESB customers were inspected. 96 customers are identified as good customer (where inspection team of SESB has confirmed no fraud on the customer meter) and 94 remaining customers are confirmed to have fraud meter. On average, 85 months of electrical consumption were used. The raw data is then preprocessed to obtained monthly consumption of each customer. Then, consumption of each customer is arranged in row and indication of good or bad customer is shown in the last column. +1 indicates good customer and -1 indicate bad customer.

The processed data is then stored as a text file in the same folder of the SVM-MGA program. The program will take 90% of the data for training and the remaining 10% for validation. The flow chart of the SVM-MGA program is shown in figure below. 10 fold cross validation is used to ensure that the accuracy of the result does not overfit. Overfitting is a condition where the model fits to closely to the training data (including noise) and cannot generalize other unseen dataset.

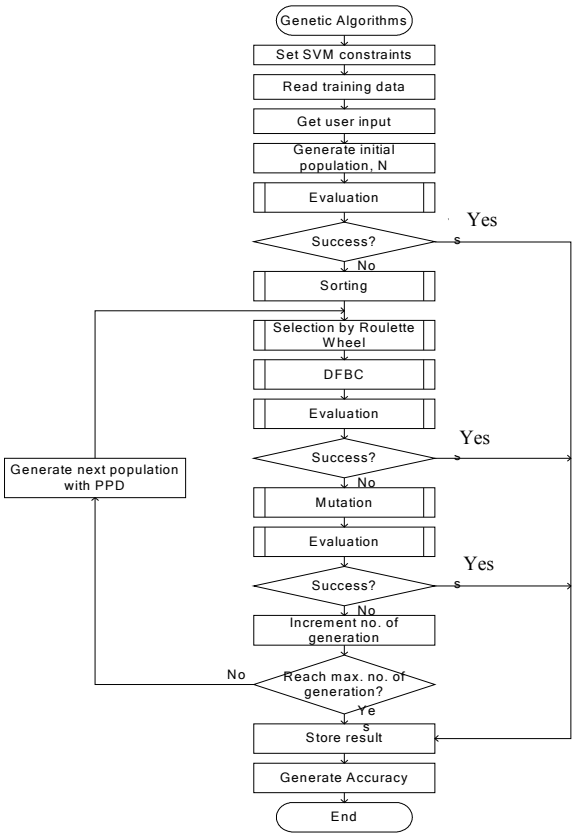

Figure 2 SVM-MGA Program Flow Chart

In this research, DFBC is used to minimize randomness in high fitness chromosomes. The default crossover rate is set between 0.5 and 1. The mutation process is also employed to avoid local optimum problem, the mutation date is empirically set between 0.01 and 0.08 [7].

DATA ANALYSIS & RESULTS

This section will discuss on the analysis and results of the

developed model using the SESB data.

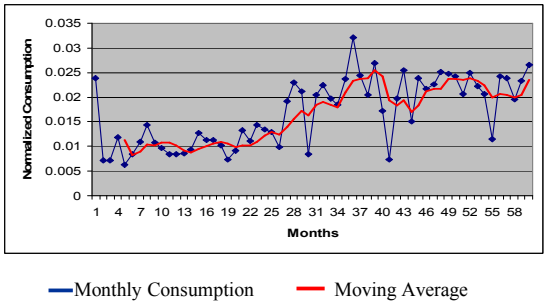

Figure 3 Monthly electrical consumption of a typical fraud meter

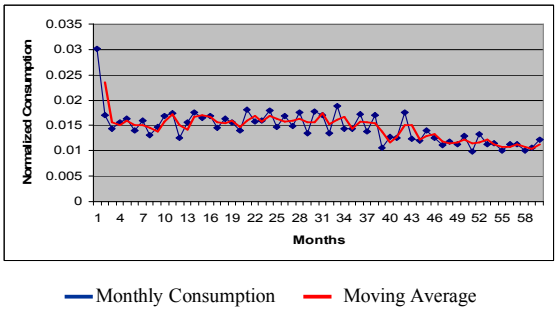

Figure 4 Monthly electrical consumption of a typical good meter

Figure 3 and 4 showed that the monthly consumption of fraud meter tends to deviate more from the moving average if compare to the good meter reading. This is important to show the correlation between the monthly consumption pattern and classification of the data. This is also the rationale of choosing monthly consumption as the feature of the model.

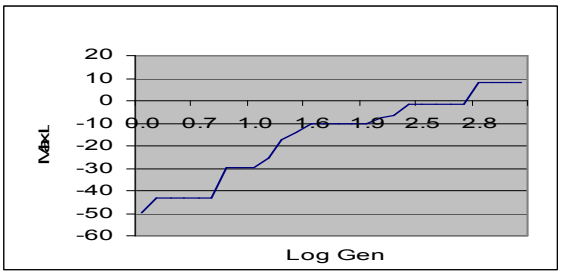

Figure 5 Graph of Maximum L VS Log10 of Generation used

With referring to Figure 7, the highest 10 fold cross validation accuracy was 94%. It is noted accuracy of detection decreases when it has reach saturation point. This is mainly because the model generated tends to be over generalizing the data given. The maximum L obtained is not the local optimum because the MGA has reached the stopping criterion (1000 Generations) with a constant 10 FCV accuracy.

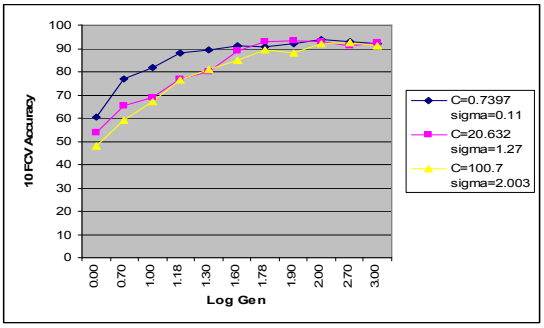

Figure 6 Graph of 10 Fold Cross Validation Accuracy VS Log10 of Generation used

CONCLUSIONS

The intelligent system using hybrid SVM-MGA was developed and it will assist SESB inspection teams to increase their effectiveness to reduce NTL. . Hybrid SVM-MGA has also shown to be an alternative to search for the best parameters for Dual Lagrangian optimization. With this hybrid algorithm, weakness of KA can be avoided.

REFERENCES

[1] Vapnik V N. "*Statistical Learning Theory*", New-York, John Wiley & Sons, 1998.

[2] X. Zhang, "Introduction to statistical learning theory and support vector machines", 2000.

[3] Chapelle O. and Vapnik V., "*Advances in Neural Information Processing Systems*", pp.230-237, Cambridge, Mass: MIT Press, June 2000.

[4] Joachims T., "Estimating the generalization performance of a SVM efficiently", *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, Kaufman, pp.431-438, 2000.

[5] Vapnik V. and Chapelle O., Bounds on Error Expectation for SVM, *Advances in Large Margin Classifiers*, Cambridge, MA: MIT Press, pp.261-280, 2000.

[6] Chapelle O., Vapnik V., Bousquet O. and Mukherjee S., "Choosing multiple parameters for support vector machines", *Machine Learning*, Vol. 46, no.1-3, pp.131-159, Jan.-March 2002.

[7] Grefenstette, J.J. "Optimization of Control Parameters for Genetic Algorithms," *IEEE Trans. Systems, Man, and Cybernetics*, Vol. SMC-16, No. 1, Jan./Feb. 1986, pp. 122-128.

[8] Liepins, G. E., Hilliards, M.R., Palmer, M. and Morrow, M., Greedy Genetics: Genetic Algorithms and Their Applications, *Proceedings of the Second International Conference on Genetic Algorithms*, Cambridge, MA, 1987.