

Wavelet Based Feature Extraction and Multiple Classifiers For Electricity Fraud Detection

Rong Jiang, Harry Tagaris and Andrei Lachsz

Abstract—Electricity consumer dishonesty is a serious problem faced by all utilities. Finding efficient measurements for detecting fraudulent energy usage has been an active research area. The most effective way is to use intelligent/smart electronic meters that make fraudulent activities more difficult and easily detectable. In this paper we propose a new automatic feature analysis method using wavelet techniques and combining multiple classifiers to identify fraud in electricity distribution networks. Based on the assumption that meter-reading data present abnormalities when fraud events occur, the feature extraction scheme is carried out in both time and wavelet domains and the combination of multiple classifiers is applied through a cross identification and a voting scheme. Simulation results prove the proposed method to be effective in electricity fraud identification. For a relatively small amount of data, the classification accuracy reaches 78% on the training dataset and 70% on the testing dataset.

Index Terms—Combination of multiple classifiers, feature extraction, fraud detection, modeling and simulation, wavelets.

I. NOMENCLATURE

Minerva—subsystem of the convergent automation technology system developed by InovaTech Limited, Australia, that consists of network management, customer web and customer care, power quality, data and systems management, tariff builder, communication, fraud analysis, external integration, data analysis and reporting.

II. INTRODUCTION

Electricity utilities lose large sums of money each year due to electricity consumer fraud. Electricity fraud can be defined as a dishonest or illegal use of electricity equipment or service with the intention to avoid billing charge. It is difficult to distinguish between honest and fraudulent customers. Realistically, a utility will never be able to eliminate fraud. It is possible, however, to take measures to detect, prevent and control fraud. An essential part of this process is an efficient detection mechanism.

Neural network techniques have gained acceptance as useful problem solving tools in the power industry, with

feature selection and extraction being critical to achieving good learning capabilities and generalization performance [6], [17], [11]. Neural network techniques are also widely applied as powerful tools in commercial fraud-detection system fields such as financial transactions, telemarketing, health care, insurance, e-business etc [5], [13].

Electricity fraud detection is becoming more and more of a necessity. Finding the efficient measurements preventing fraudulent energy usage has been an active research area. The most effective way to reduce commercial losses is to use intelligent/smart electronic meters that make fraudulent activities more difficult and easily detectable [15]. However, the patterns of electricity fraud are nonstationary, noisy, time varying and have multiscale properties: fraudulent activities appear randomly initiated and terminated, and fraudulent methods, magnitude and frequency change over time. Therefore, it is very difficult to analyse such data using conventional techniques. Currently, there are no reported approaches for electricity fraud detection.

Recently, wavelets have been successfully introduced as an efficient tool for various time series analyses [1], [14]. Due to the properties of localization and multiresolution analysis, wavelets are suited not only for waveforms that are smooth but also for those with abrupt changes, transients and other irregularities. Wavelets can be used to determine the time of a change, its type (change in the first or the second derivative) and its amplitude. The reason for choosing wavelets over conventional methods is their ability to capture localized features, which means that a more accurate model can be expected. Different predictors can be fitted to the wavelet related subseries and they can be effectively combined.

There are several different types of fraud that can occur, but our research concentrates on the scenario where the abrupt or gradual changes appear in the meter readings, that possibly indicate fraudulent activities. They could be step or gradual waveform changes in time, frequency or in amplitude.

In this paper a new automatic feature extraction/analysis method is proposed using the wavelet techniques and combination of multiple classifiers is developed to identify fraud in electricity distribution networks. Based on the assumption that meter-reading data present abnormalities when a fraud event occurs, the feature extraction scheme is carried out in both time and wavelet domains. The combination of multiple classifiers is implemented on cross identification and a voting scheme.

Concepts of wavelet transforms and the proposed approach

Rong Jiang is with InovaTech Limited, St. Leonards, NSW 2065, Australia (e-mail: junejiang@inovatech-inc.com).

Harry Tagaris is with InovaTech Limited, St. Leonards, NSW 2065, Australia (e-mail: harrytagaris@inovatech-inc.com).

Andrei Lachsz is with InovaTech Limited, St. Leonards, NSW 2065, Australia (e-mail: andreilachsz@inovatech-inc.com).

are introduced in Section III. Tests are simulated and discussed in Sections IV. and V. respectively. Conclusions are drawn in Section VI.

III. PRINCIPLE

In this section, the introduction algorithms and the principles are briefly described.

A. Brief review of wavelets

In $L^2(\mathbb{R})$, given certain conditions [7], a function f in the signal domain can be transformed into the wavelet domain by applying a discrete wavelet transform (DWT)

$$\begin{aligned} W_{\psi} f_{j,k} &= (W_{\psi} f)\left(\frac{k}{2^j}, \frac{1}{2^j}\right) \\ &= \int_{-\infty}^{\infty} f(t) \left[2^{j/2} \psi(2^j x - k) \right] dx \\ &= \left\langle f, \psi_{j,k} \right\rangle \end{aligned} \quad (1)$$

Where ψ is the mother wavelet and its derived forms are:

$$\left\{ \psi_{j,k}(x) \right\} = 2^{j/2} \psi(2^j \cdot x - k), \quad j \in Z, k \in Z, \quad (2)$$

where j is the dilation and k is the translation. $W_{\psi} f_{j,k}$ contains the information about the function f near the time point $k/2^j$ and near the frequency proportional to 2^j . At each resolution, the signal is analyzed with both wavelets and scaling function: the wavelets encode the details while the scaling function encodes an image of a signal at half resolution, taking one sample out of two. This process is repeated until nothing, or virtually nothing, is left.

The stationary wavelet transform (SWT) is popular because of its property of shift invariance: the new sequences have the same length as the original sequence. The slight change in the derived forms $\psi_{jk}(x)$ is:

$$\left\{ \psi_{jk}(x) \right\} = 2^{j/2} \psi \left[2^j (x - k) \right], \quad j \in Z, k \in Z. \quad (3)$$

That is, for all integers j and k , the wavelet decomposed coefficients at all resolution levels j appear at all positions k . The location of stationary wavelets does not depend on the resolution level. This is in contrast to the discrete wavelets that only appear at the locations $k/2^j$ — a dyadic grid.

Referring to Nason [9] and Sachs [10], the local wavelet spectrum can be defined as,

$$P_w(j, x) = \frac{1}{c_{\psi} 2^j} \left| W_{\psi}^s f_{j,k} \right|^2, \quad j \in Z, k \in Z, x \in Z. \quad (4)$$

Where the $W_{\psi}^s f_{j,k}$ is stationary wavelet coefficients at the j^{th} resolution level and the coefficient c_{ψ} is

$$c_{\psi} = \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega. \quad (5)$$

The local wavelet spectrum measures the contribution to the total energy from the vicinity of point x at the resolution level j , and unveils the frequency components of the subseries at that level. This is useful for detecting the spectral changes in the wavelet domain.

B. Wavelet shrinkage

Because the data used is collected for bookkeeping rather than for fraud detection, the noise level is high. In our research, the data is normalized and suppressed through wavelet shrinkage. Wavelet shrinkage is processed [8], [3], [2] as:

$$\delta_{\lambda_1 \lambda_2}^{HS}(w) = \begin{cases} 0, & \text{if } |w| \leq \lambda_1 \\ \text{sgn}(w) \frac{\lambda_2 (|w| - \lambda_1)}{\lambda_2 - \lambda_1}, & \text{if } \lambda_1 < |w| \leq \lambda_2 \\ w, & \text{if } |w| > \lambda_2 \end{cases} \quad (6)$$

where the parameters λ_1 and λ_2 are the lower and upper threshold respectively. The representation in the equation above is intermediate between hard and soft shrinkages. Hard shrinkage is obtained by setting $\lambda_1 = \lambda_2$ and soft shrinkage is obtained by setting $\lambda_1 = \infty$.

According to the IEEE standard [16], in wavelet analyses, statistical concepts are applied in the selection of thresholds such as in amplitude reduction, variance filtering, spectral shift and noise suppression. For brevity, the methods of selecting the wavelet thresholds are not discussed here, but a broad overview of the algorithms can be found in Donoho [2][3], Bruce [8] and others.

C. Proposed method for feature extraction

In general, a classification problem can be described as follows: given a set of training data, find a classification method to correctly label any data from the same source. As shown in Fig. 1, the procedure consists of three steps: pre-processing, feature selection and classification. However,



Fig. 1. General pattern recognition

sometime it is impossible to find a suitable feature set.

Wavelet analysis has been recognised as a powerful tool in solving problems in power systems [4], [12]. However, the particular difficulty in electricity fraud identification is that the background of many observations is not known. Our proposed schema of feature analysis, shown in the Fig. 2, is aimed at meeting this challenge.

In this approach, several techniques have been applied for feature analysis. Wavelet analysis plays a major role in the implementation of waveform recognition, pattern identification, pattern change location, and spectral distribution analysis in wavelet domains.

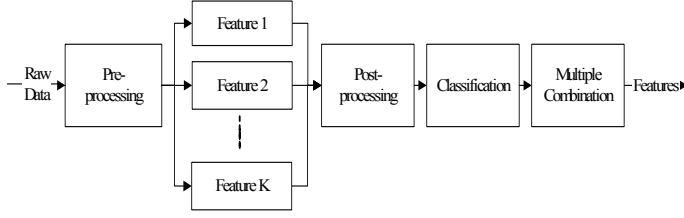


Fig. 2. Feature extraction and pattern recognition using wavelet technique

The principle of applying wavelet technique to feature analyses is shown in Fig. 3, where S_o is a signal, J is the total number of decomposition levels, w'_j are wavelet decomposition coefficients in the j^{th} resolution level, w''_j are the processed wavelet coefficients, $S_l (1 \leq l \leq L)$ are the features for a dataset, $p_j (1 \leq j \leq J)$ are the process, $A_i (1 \leq i \leq I)$ are the time domain analyses, $A_k (1 \leq k \leq K)$ are the wavelet domain analyses, $C_k (1 \leq k \leq K)$ are the analyzed results from coefficients, \oplus is multiple combination, Σf is the assembling

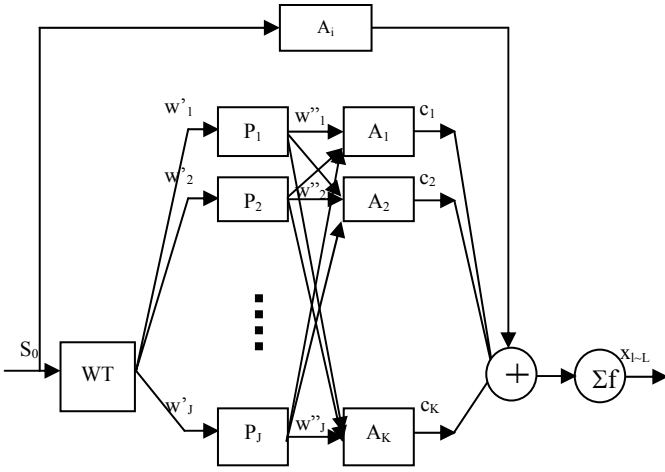


Fig. 3. Framework of applying wavelet technique to feature analyses.

and $X_l (1 \leq l \leq L)$ are the extracted features that are either in the time domain or in the wavelet domain.

D. Proposed method for classifier combination

As mentioned in the previous section, normally it is impossible to find the solution to the real-world classification problems by using one feature only. There is seldom one unique feature that is highly superior to any other for representing all the input samples.

Assuming that there exists a probabilistic relationship between the raw data profile (a time series of power demand from a customer) and a set of features, a soft competition algorithm can be applied to select optimal feature vectors from all extracted features. Suppose there are M classes C_1, \dots, C_M and N classifiers CL_1, \dots, CL_N . The probability that data sample D belongs to the m^{th} class can be presented as:

$$P(x_l) = P(D \in C_m | O_l = 1), \quad (7)$$

where $P(O_l = 1)$ is the probability that the l^{th} feature vector $\{X_j\}$ is optimal.

Considering a feature matrix X_{IL} with I records and L features, in terms of the input feature vector $x_{i,l} (1 \leq i \leq I; 1 \leq l \leq L)$, the probability that the i^{th} sample belongs to class C_m classified by the classifier CL_n is:

$$p_m^{(n)}(x_{i_n,l}) = \begin{bmatrix} p_{11}(x_{i_n,l}), \dots, p_{1M}(x_{i_n,l}) \\ \dots \\ p_{n1}(x_{i_n,l}), \dots, p_{nM}(x_{i_n,l}) \\ \dots \\ p_{N1}(x_{i_n,l}), \dots, p_{NM}(x_{i_n,l}) \end{bmatrix}^T, \quad (8a)$$

where

$$p_{nm}(x_{i_n,l}) \geq 0 \text{ and } \sum_{m=1}^M p_{nm}(x_{i_n,l}) = 1. \quad (8b)$$

If denote g_n as the weight function for the n^{th} classifier, then the probability of the i^{th} set data belonging to the m^{th} class is:

$$p_i = \sum_{n=1}^N g_n p_m^{(n)}(x_{i_n,l}). \quad (9)$$

The principle diagram of the combination of multiple classifiers is represented in Fig. 4. Each of the N classifiers is trained by a set of L features ($L < L_{all}$) where schemes are used to produce the combination weights. Based on the extracted features, the linear combination is implemented so that only four dominant features are used in classification.

E. Model generator

In the proposed method, model training and testing are processed in the model generator. The inputs include the feature table, customer labels and customer fraud history (optional, it can assist labeling, refer to Fig. 5). Techniques such as Bayes network and multiple layer perceptrons are applied and logical function gates are used to control the processing loop. In order to ensure generation performance of

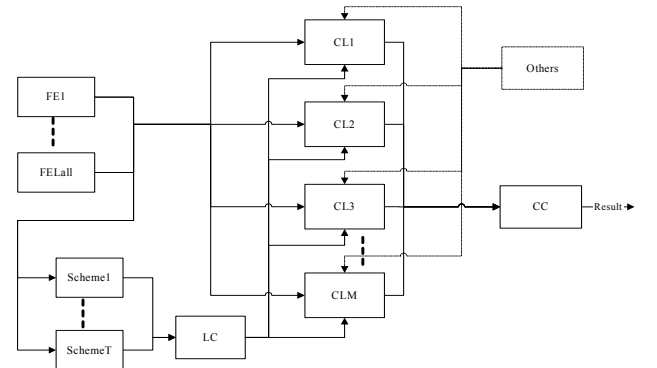


Fig. 4. Diagram of classifier combination (FE: feature extracted, $l=1 \sim L_{all}$; CL: classifier, $n=1 \sim N$; LC: linear combination; CC: cross combination; others are other input components for classification, refer to the next section)

classification, the generator only outputs models if the tested results are deemed to be satisfactory by model testing. These models are specific to particular groups of customers such as industry, commerce or individual residences.

The model generator can automatically create up-to-date classifying models routinely based on the latest input data. Thus the model generator can fill two tasks: generating or upgrading models periodically as scheduled by users.

F. System frame and procedure

As shown in Fig. 5, the complete automatic fraud detection consists of three parts: feature extractor, model generator and fraud detector. The feature extractor deals with data pre-processing, data analyses (spectrum, variance, the 1st and 2nd derivatives, periodogram and statistics), feature extraction and feature post-processing. The model generator deals with model generating, testing and upgrading. The fraud detector outputs fraud reporting. Inputs for whole system are data profiles, customer fraud history and customer labels, while outputs are fraud reports, models and accuracy rate. Features are inner variables.

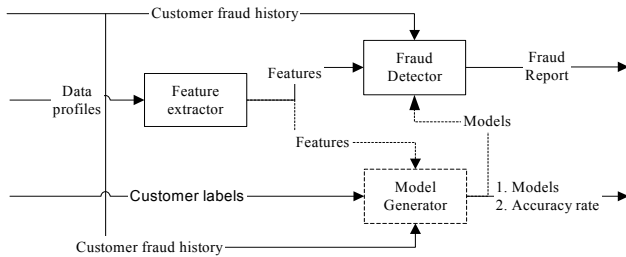


Fig. 5. Diagram of fraud detection

The connection of the fraud detection to Minerva is shown in Fig. 6. The inputs are the most recent meter readings, customer labels and customer fraud history record, while the outputs are fraud reports indicating the suspect customers with corresponding estimated probabilities. Also the used models and the corresponding accuracy rate are saved for data retrieving purpose.

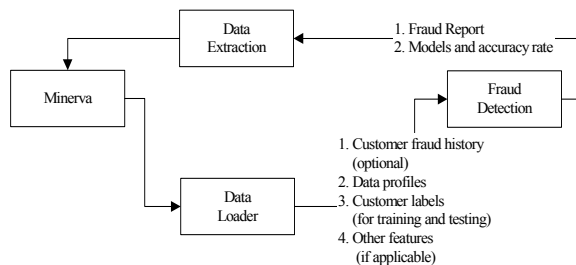


Fig. 6. Relation between Minerva and fraud detection application.

In the Minerva system, data access, feature extraction, model generating/upgrading, fraud detection and fraud reporting are scheduled processes running in the background without operator intervention.

IV. EXPERIMENTS

The data used in our simulation are from measurements made with energy demand meters installed at customers' premises.

A. Input data

The input data consists of average power measurements, recorded at fifteen-minute intervals during the year 2000. Fig. 7 shows two typical data sets. Data from customers known to be honest or fraudulent is labeled as "honest" or "fraud" respectively. They were mixed randomly in model training/testing and cross validation. The data profiles for the fraud scheme can be meter readings alone or meter readings plus other data such as customer educational level, marital status, and income etc. The additional features can be pre-processed in the appropriate domain and appended to the feature table directly.

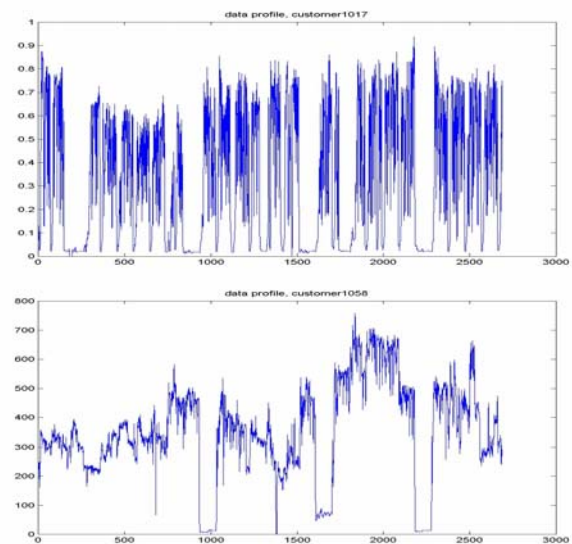


Fig. 7. Two typical data profiles (x-axis: time in 15-min interval; y-axis: energy demand (kW))

B. Wavelet base and neural network architecture

Either a "Daub 6" or a "Sym 8" filter is used for detecting abrupt or gradual changes in wavelet decomposition. In each case a total of eight levels of wavelet decomposition were employed. The extracted features were post-processed and then combined, thus the number of attributes reduced to four. One hidden layer neural networks are applied. The number of nodes for the input, hidden and output layers for the classifiers is 4, 32 and 1 respectively.

C. Evaluation criteria

If there are N classification labels, the number of error types equals $N^2 - N$. By assigning e_{pq} to the cost of an error

made by wrongly classifying p into q where $p \neq q$ and c_{pq} the corresponding type of error, the total error cost can be expressed as

$$C = \sum_{p=1}^P \sum_{q=1}^Q e_{pq} c_{pq}. \quad (10)$$

Errors do not equally impact performance. The question is which error is more significant? In this application, false positive identification will wrongly identify honest customers and may cause unnecessary investigation and the loss of honest customers. False negative identification ignores genuine fraudulent activities and loss of revenue for the utilities. There is a problem of balance.

In our experiments, acceptance accuracy rate is defined as the ratio of the number of the correctly reported instances of fraud to the total number of reported cases of fraud given a dataset X with total I meter readings and L analyzed features:

$$R(\text{correct} | X_{IL}, \text{report}) = \frac{N(\text{correct} \cap \text{report})}{N(\text{report})}, \quad (11)$$

where $N(\text{report})$ is the total number of cases of reported fraud, and $N(\text{correct} \cap \text{report})$ is the total number of cases correctly identified as fraud.

In the next section we give some quantitative experimental results based on the feature analysis scheme with decisions made using the indicator variables η_1 and η_2 from classification by applying the following rules:

- Pattern_i is accepted—if $\eta_1 \geq p_m^{(n)}(x_{i_n, l})$ and $\eta_2 \geq R(\text{correct} | x_{IL}, \text{report})$;
- Pattern_i is rejected—if $\eta_1 < p_m^{(n)}(x_{i_n, l})$ or $\eta_2 < R(\text{correct} | x_{IL}, \text{report})$.

V. DISCUSSION

Table 1 shows the relationship between $R(\text{correct} | x_{IL}, \text{report})$ and the number of data profiles given the agreement probability threshold $p_i = 0.55$ (refer to (9)). As expected, the larger the number of data profiles used, the more accurately fraud is detected. However, as the number of data profiles increases, the feature extraction and training of the classifiers become more time consuming.

It can be noticed (refer to table1) that the simulation results are acceptable only if 300 or more profiles are used when the accuracy rate is set to 70%. This is because a smaller number of samples cover a smaller space of the system operation, yielding poor generalization capability.

The relationship between the acceptance accuracy rate and the agreement probability threshold is shown in Fig. 8 with the number of data profiles being 600. Meter reading data are 15 minutes interval, 30 days period, averaged by 10 rounds with 95% confidence, and the y-axis represents accuracy rate $R(\text{correct} | x_{IL}, \text{report})$ in percent and the x-axis represents the probability threshold p_i . We can see that a higher acceptance accuracy rate can be obtained by setting a higher probability

threshold. For example, at p_i values 0.3, 0.4, and 0.5, $R(\text{correct} | x_{IL}, \text{report})$ is 71.44%, 72.13%, and 74.79% respectively.

Table 1: Relationship between acceptance accuracy rate and the total number of data profiles (averaged by 10 rounds with 95% confidence)

Number of data profiles	Accuracy rate
120	57.0875
180	60.5633
245	67.5000
300	72.4809
620	76.7827
900	78.2222
1240	79.0282
1600	79.4870

However, when the threshold equals or becomes larger than 0.6, the accuracy rate drops dramatically because classifiers fail to recognise genuine fraud. On the other hand, if the threshold is too low, the misidentification of “honest” is unavoidable and thus the performance is poor. In the latter case, the model testing fails in our simulations.

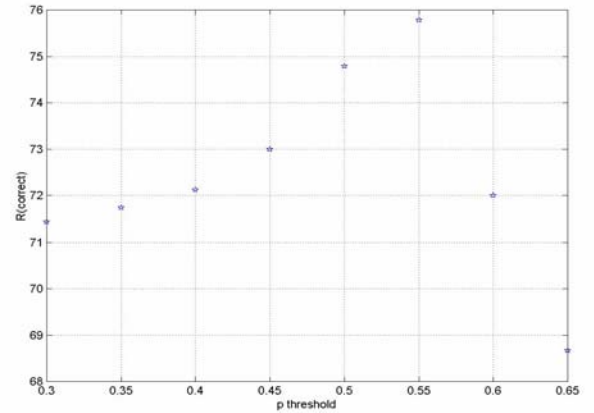


Fig. 8. Relationship between the acceptance accuracy rate.

VI. CONCLUSIONS

A major motivation for the study was to solve the automatic fraud recognition problem in electricity distribution systems. In this paper, we outlined a possible framework for electricity fraud detection. We proposed the use of new wavelet techniques to represent the multiple characteristics of meter readings and to build a new model that incorporates evidence from several basic models. This mechanism makes it feasible to add new components in the detection system without significant re-design.

Our results show that the proposed method of extracting features through wavelets and combining classifiers can be used for reliable detection of abnormal patterns. This method proved to be sensitive to local pattern changes, to step and gradual amplitude variation. Using a relatively small amount of data, the classification accuracy reached 78% on the

training dataset and 70% on the testing dataset. When compared with research in credit card fraud detection where 65% acceptance accuracy rate is claimed, the proposed method is promising [13].

Some useful features of the proposed method are listed below:

- It is effective in detecting the signal dynamics as time varies.
- The feature extraction approach has both frequency and time information presented simultaneously.
- The approach is sensitive to the abnormal changes of the amplitude and frequency in meter readings.
- The combination of classifiers can be very broad in terms of definition.

However, it is noted that in wavelet domain analyses, wavelet transforms tend to be somewhat insensitive to these slowly varying changes and tend to produce very small coefficients at all scales.

Future work may include applying post-filtering techniques to the processed datasets and modules according to the observed behavior in the data. It may also devise practical procedures to translate our automatic feature extracting and classifying rules into models for real-time use. It is probable that better classification performance can be achieved by using additional information with an appropriate selection method such as Monte Carlo, thus only the most relevant features being included in modeling.

VII. ACKNOWLEDGMENT

The first author gratefully acknowledge the contribution of Prof. Marwan Jabri for his valuable comments on the original version of this document.

VIII. REFERENCES

Periodicals:

- [1] Abry, P. and Veitch D. (1998), "Wavelet analysis of long-range-dependent traffic", *IEEE Trans. on Info Theory*, 44(1), pp2-15.
- [2] Donoho, D. (1995), "Denosing by soft-thresholding", *IEEE Trans. on Info Theory*, vol.41, No.3, pp613-627.
- [3] Donoho, D. and Johnstone, I. (1995), "Wavelet shrinkage: asymptopia?", *Royal statistical society*, B57: pp. 301-369.
- [4] Galli, A and Nielsen, O. "Wavelet analysis for power system transients", *IEEE, Computer application in power*, Jan 1999 pp16-25.
- [5] He, H., Wan J., Graco W. and Hawkins, S. (1997), "Application of neural networks to detection of medical fraud", *Expert systems with applications*, 13 (4), 329-336.

Books:

- [6] Bishop, Christopher (1995), <Neural networks for pattern recognition>, Oxford University Press.
- [7] Chui, Charlesk (1995), <An introduction to wavelets>, Academic Press.

Technical Reports:

- [8] Bruce, A. and Gao, H. Y., "WaveShrink: Shrinkage function and thresholds", 1995.
- [9] Nason, G P and Silerman, B W, "The stationary wavelet transform and some statistical application", Dept of Mathematics, University of Bristol, University Walk, Bristol, 1995 (BS8 1TW).
- [10] Sachs, R. V., Nason, G and Kroisandt, G, "Adaptive estimation of the evolutionary wavelet spectrum", Statistics Dept. Stanford University, 1997 (516).

Papers from Conference Proceedings (Published):

- [11] El-Sharkawi, M. A., "Neural network and its ancillary techniques as applied to power systems", IEE, Savoy Place, London WC2R OBL, UK, 1995.
- [12] Lee, C. H., Wang, Y. J. and Huang W. L., "A literature survey of wavelets in power engineering applications (invited review paper)", *Proc. Natl. Sci. Coun. ROC(A)*, Vol. 24, No. 4, 2000. pp. 249-258
- [13] Lee, W and Stolfo, S. (1998), "Mining in a dataflow environment: experience in network intrusion detection", *proceeding of the fourth international conference on knowledge discovery and data mining (KDD)*, New York, NY.
- [14] Meyer, Y. (editor 1989), <Wavelets and applications>: *proc. of Intern conf.*, France.
- [15] Rao, M V Krishna and Miller S H, "Revenue improvement from intelligent metering systems", *Conference of metering and tariffs for energy supply, IEE 1999.*

Standards:

- [16] "IEEE recommended practice for monitoring electric power quality", *IEE standard coordinating committee 22 on power quality*, SH94324, 1995.

Lecture notes:

- [17] Neal, R. (1994), <Bayesian learning for neural networks>, *Lecture notes in statistics*, Springer Press.

IX. BIOGRAPHIES

Rong Jiang, MSc/MEng, was born in Chengdu, China. She graduated from Sichuan University, Institute of Optics and Electronics, Chinese Academy of Sciences, and University of Sydney, Australia, majoring in computer engineering. Her employment experience includes Chengdu TV Equipment Company and Sichuan University, China and University of Sydney, Australia. Her special fields of interest include time series analysis and prediction, wavelet techniques and applications, machine learning and signal processing.



Harry Tagaris, MSc, was born in Melbourne, Australia. He graduated from University of Melbourne and University of NSW (ADFA), majoring in electrical engineering. His employment experience includes Boeing Australia, CEA Technologies Pty Ltd, Auspace Ltd, Australian CSIRO and Australian National University. His special fields of interest include in system engineering methodology, electronic and DSP design techniques dissemination.



Andrei Lachsz, MSc, was born in Bucharest, Romania. He graduated from the Polytechnic University, Bucharest, majoring in electrical machines. His special fields of interest include power and industrial electronics.

