# 3. DATA MINING TECHNIQUES FOR UNMETERED POWER LOSSES ANALYSIS

## 3.1 Introduction

In this chapter, the main part of data mining techniques useful for unmetered power losses analysis are reviewed. Advanced methods for the fundamental functionalities involved have been developed and are summarized in the five main modules comprising the proposed unmetered power losses analysis framework presented in Figure 3.1. The sequence of major data mining tasks within the framework that are detailed here are designated as Data Pre-processing, Feature Selection [74], Detection of Techniques, Classification of Techniques, and Prediction of Techniques.

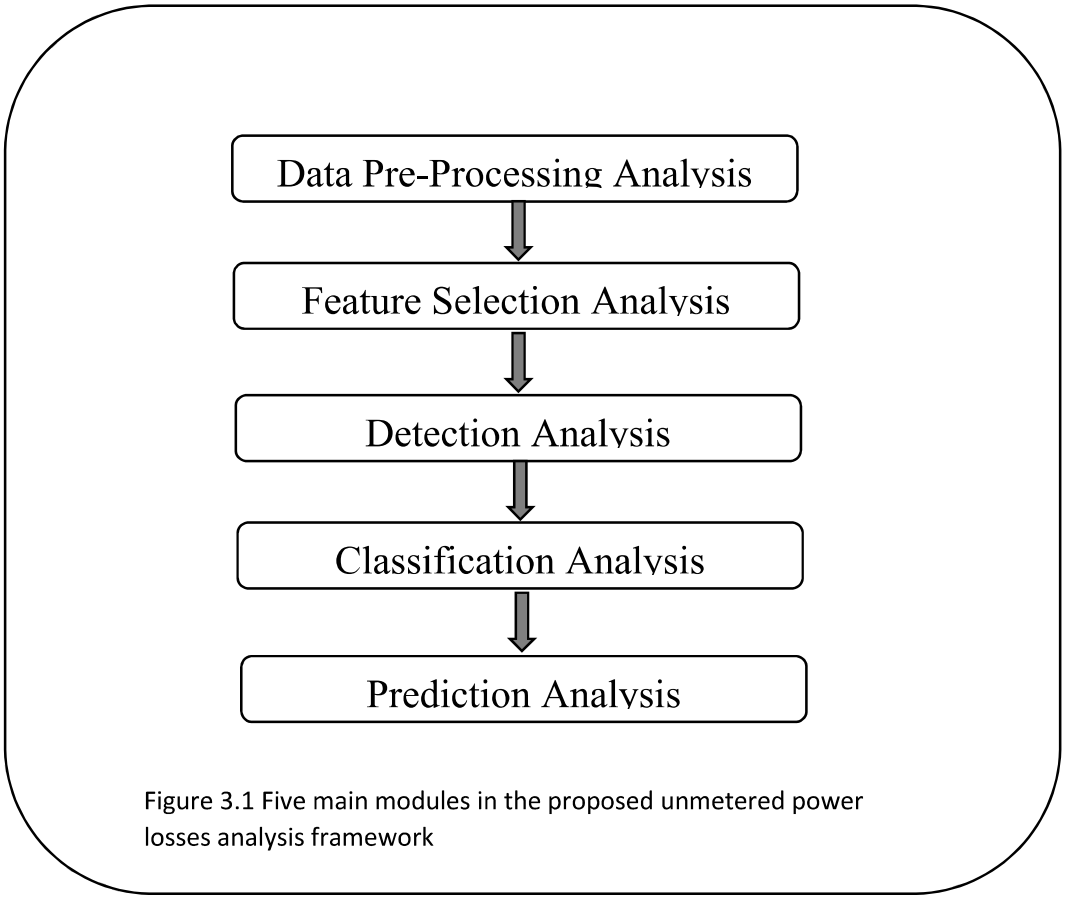


Figure 3.1 Five main modules in the proposed unmetered power losses analysis framework

More comprehensive analyses are to be undertaken in subsequent chapters as follows. Data pre-processing [15] and detection tasks and features election analysis in chapter 4, classification analysis in chapter 5 and prediction analysis in chapter 6.

## 3.2 Feature Selection

Feature selection plays an important role in the subsequent classification task. It contains the search process of attributing this space for an appropriate subset. To eliminate any irrelevant attributes is undertaken in order to upgrade the performance of a selected learning algorithm (LA). The objective of feature selection is that to determine the subset of features [77] and that will improve data mining (DM) performances compare to all the features. It is essential process that has been used in many applications, including text processing and in general expression arrays. The necessity for it comes in large database, it is impractical to use in all the available features. Moreover, the potential for the curse of dimensionality to occur must be recognized. In the ensuing discussion, a review of some key feature selection topics will be undertaken from three perspectives: 1) the search problem, 2) evaluation criteria, and 3) model selection.
The search problem arises when endeavouring to find an optimal subset and it includes two main considerations, namely search directions and search strategies. Four kinds of search directions are examined below.

- Sequential Forward Generation (SFG) – SFG, known simply as "forward selection", begins with an empty set of features and sequentially adds the features one at a time until all the required features accumulate to form a full set.

- Sequential Backward Generation (SBG) – SBG, known simply as "backward elimination", begins with a full set of features and sequentially removes features one at a time until all the features that remain constitute the required set.

- Bidirectional Generation (BG) – BG starts searching in both directions concurrently and ceases either when one of the searches has identified all the required features before reaching the middle of the search space or when both searches have reached the middle of the search space.

- Random Generation (RG) – RG starts searching in a random direction and adds and deletes features randomly. The intention is to avoid being trapped into local optima by not adhering to any fixed means of subset generation.

The selection of search directions influences the performance of feature selection and the selection of search strategies influences the search directions. The combination of search directions and the accompanying determination of search strategies can improve the feature selection technique performance to an extent that depends on the demands of the problem. Three categories of search strategies are included here: 1) Complete/Exhaustive Search, 2) Heuristic Search, and 3) Nondeterministic Search.

The second perspective to be reviewed concerns evaluation criteria. The goodness of a subset is always determined by certain criteria and these are categorized in two groups, criteria in form of independent and dependent, respectively. A special methods are used to evaluate the strength of a feature or subset of features [73] through exploiting the intrinsic behaviour of the training

set. Some popular independent methods are cited with the four types of measures involved in evaluation criteria considered here being 1) Distance Measures, 2) Information Measures, 3) Dependence Measures, and 4) Consistency Measures

- Distance Measures – The distance measures are known as separability, divergence or discrimination measures.
- Information Measures - The information measures typically determine the information gain from feature set.

- Dependence Measures – The dependence measures are also known as correlation measures or similarity measures. They measure the ability to predict the value of one variable form to the value of another.

- Consistency Measures –This measures utilizes the class information, biased with Min-Features for selecting feature subset.

The two different perspectives including search strategies, search directions and evaluation measures above are summarized into the following table 3-1:

Table 3.1: Search Strategies based on Search Directions and Evaluation Measures

| Search Direction | Evaluation Measures | Search strategies | | |
|---|---|---|---|---|
| | | Complete | Heuristic | Nondeterministic |
| SFG | Accuracy | Yes | Yes | No |
| | Consistency | Yes | Yes | No |
| | Information | Yes | Yes | No |
| | Distance | Yes | Yes | No |
| | Dependence | Yes | Yes | No |
| SBG | Accuracy | Yes | Yes | No |
| | Consistency | Yes | Yes | No |
| | Information | Yes | Yes | No |
| | Distance | Yes | Yes | No |
| | Dependence | Yes | Yes | No |
| BG | Accuracy | Yes | Yes | No |
| | Consistency | Yes | Yes | No |
| | Information | Yes | Yes | No |
| | Distance | Yes | Yes | No |
| | Dependence | Yes | Yes | No |
| RG | Accuracy | No | Yes | Yes |
| | Consistency | No | Yes | Yes |
| | Information | No | Yes | Yes |
| | Distance | No | Yes | Yes |
| | Dependence | No | Yes | Yes |

In the third perspective to be reviewed, the selection of models is in focus. The combination of the search strategies and search directions directly yields two models of particular relevance,

namely the wrapper model and the filter model. A hybrid model drawing from both these alternatives is also possible.

3.2.1 Wrapper model – The wrapper model requires one pre-determined data mining algorithm and uses its performance as the evaluation criterion. It generates a set of candidate features, runs the induction algorithm on the training data, and uses the accuracy of the resulting description to evaluate the feature set. It also searches for better features that are more suited to the algorithm concerned in that they improve the data-mining performance. This model will be computationally more expensive to run than the alternative filter model.

3.2.2 Filter model – The filter model relies on general characteristics of the data to select and evaluate feature subsets without involving any data mining algorithm. It filters out irrelevant attributes before the induction process occurs.

3.2.3 Hybrid model – The hybrid model takes advantage of both the filter and wrapper models by exploiting their different evaluation criteria in different search stages. Liu [111] and Zhao [112] provided comprehensive reviews of the procedures involved in feature analysis and list the most popular subset generated strategies available as Random Search, Greedy-stepwise Search, Complete Search, Best-first Search, Ranker and Genetic Search. The five search methods that are included within the WEKA data-mining software package to be used here are set out below.

3.2.4 Best-First Search – It uses its experimental strategy by searching forward, backward, or bi-directionally. Even though it is among the more expensive searches to run, it remains tractable in some domains. However, it often losses the global optimum as a consequence. It determines the space of greedy hill-climbing and employs a backtracking facility to restrict the local optima.

3.2.5 Exhaustive Search – The exhaustive search is claimed to be impractical as there exist 2a possible subsets of an attributes.

3.2.6 Genetic Search – It uses a simple Genetic Algorithm (GA) to locate the global optimum, its speed is comparable with the Best-first and Greedy stepwise options.

3.2.7 Greedy-stepwise – The Greedy-stepwise search selects a random point in the feature space that performs greedy forward and backward searches and it thereby achieves local optima. It generates a ranked list of attributes through traversing the space from one side to the other and preserving the order in which the included attributes are selected.

3.2.8 Random Search – The Random search, without involvement of any deterministic rule generates each candidate set. This strategy produces global optima with large percentage however, the process intractable.

3.2.9 Ranker – The Ranker is the fastest search strategy. It utilizes the correlation between the features however, it could be very unreliable. The features it selects may be ranked according to appropriate evaluation values.

## 3.3 Classification

Electricity customer load profiles are very rich in hidden information that has the potential to be used for making intelligent business decisions and for acquiring knowledge of customer behaviour [76]. These load profiles contain consumption readings every half an hour and are linked to other databases, including those used for billing and involving tariff rates. In the present research, a classification task known as "supervised learning" is used to extract models to describe customer behaviour classes and subsequently to predict future customer behaviours. A classification model is useful in the process of predicting customer consumption status in accordance with three categories: non-unmetered power losses, normal behaviour, and suspicious behaviour. This process takes as given their load profiles and other relevant factors such as weather, holiday banks, rates, and the nature of their businesses. This classification task comprises a very important module in the proposed unmetered power losses framework of analysis because the classification behaviour results are useful for monitoring suspicious behaviour. Moreover, the developed behaviour model can be applied to predicting the behaviour classes of new customers [23].

Classification of electricity customers is becoming increasingly important, particularly in competitive electricity markets where suppliers have been given freedom to formulate dedicated tariff options for different customer classes. However, the present purpose of such classification is to provide electricity utilities with the means of detecting customer behaviour changes, most especially where such changes involve shifts from normal to abnormal and suspicious consumption patterns.

Several alternative classification models of relevance here are outlined below.

3.3.1 Decision Tree - The most well-known decision tree [64] algorithms are ID3 and its descendent C4.5. These two algorithms employ a variety of pruning techniques and information gain as a heuristic. The original ID3 algorithm used a criterion called gain or attribute selection to select the attribute to be tested and was based on the information theory concept of entropy. Chang [113] and Pitt [114] used decision trees for clustering customers' load profiles, while Filho [115] used them to detect fraud within an electricity utility.

3.3.2 Naïve Bayes - The Naïve Bayes [77] classifier is a classification technique that is based on Bayesian statistical theory. Statistical approaches usually determine the class label by calculating the conditional probability of a class label being relevant for a given input vector. A Bayesian classifier, in particular, assigns a class label to an input vector in accordance with the largest posterior probability. Wang used the dynamic Bayes model to classify customers' operating styles.

3.3.3 Back Propagation Neural Network – Back propagation comprises a neural network learning algorithm that is also known as "connectionist learning". It consists of a set of connected input/output units, with each connection having an associated weight. Synaptic strengths, or weights, are changed in proportion to the error times associated with the input signal, a procedure that diminishes the errors in the direction of the gradient.

3.3.4 Support Vector Machine - SVM [47] has appeared as one of the most popular and useful techniques for data classification and it has been receiving increasing attention in many areas of research due to its remarkable generalization performance. SVM [116, 117] originates from Vapnik's statistical learning theory

and has been observed as being useful for robust outlier detection. The simplest form of SVM classification is the maximal margin classifier. It is used to solve the most basic classification problem, namely the case of a binary classification with linear separable training data. Most generally, the objective of SVM is to generate a model that determines the target value of data instances in the testing set in which only properties are described. The classification target in SVM is to separate the two classes by means of a function derived from available data and thereby design a classifier that will work well on further unseen data.

The most appropriate model from among these alternatives is to be selected on the basis of estimations of the predictive accuracy of the model and the fastest processing speed. For the present purpose, the predictive accuracy of a model is defined as "the percentage of test samples that are correctly classified by the model", while the processing speed is referred to in terms of the resulting "computation costs involved in generating and using the model."

## 3.4 Prediction

Prediction techniques applied in the context of electric power utilities involve the prediction of daily, weekly, monthly, and yearly data relating to system loads, peak loads, and system energy. This prediction task, commonly known as electricity load forecasting, is one of the important tasks in the planning and operational activities of power utilities. It benefits the utilities in many ways, including unit commitment scheduling, fuel allocation, dispatch, system security, and off-line network analysis. As such, load forecasting can be categorized into Short-Term Load Forecasting, Medium-Term Load Forecasting, and Long-Term Load Forecasting.
Short-Term Load Forecasting (STLF) [70] is used to forecast loads over short periods of time, such as in daily forecasts. It is also defined as predictions of system loads with forecasted intervals ranging daily to one week. Three principle objectives of STLF are cited as 1) formulating the basic generation scheduling function, 2) assessing the security of the power system at any point, and 3) providing timely dispatcher information. In addition, STLF has been used for on-line scheduling and providing security functions in an energy management system (EMS).

Medium-Term Load Forecasting (MTLF) is defined as encompassing one-day to one-year forecasting durations, while Long-Term Load Forecasting (LTLF) is defined as extending to forecasting durations ranging from one year to ten years. These two modes of forecasting are the basis for developing future generation capacity and the associated transmission and distribution facilities. They include forecasts for quantities of electricity, maximum loads, load rates, and representative and continuance daily load curves. A non-linear regression technique for long-range weekly forecasting was proposed for the Power System.

Generally, two approaches can be used in formulating STLF, MTLF, and LTLF, namely a deterministic approach and a probabilistic approach. A deterministic approach using a pattern recognition algorithm was applied to forecasting [68] hourly loads with lead times of 24 hours. Even though this method considers sensitive loads, special occasions, and economic changes, along with demographic and geographical factors, it has the particular drawback that it is only intended for use in small-area power systems. The effects of probabilistic inputs have been considered in a study of the effects of load forecasting accuracy. Electricity load formulation is known as a non-stationary process that is influenced by a number of factors, including weather factors, time factors, seasonal factors, calendar or holiday events, economic factors, and a range of other random effects. It is because of these factors that load forecasting has always been considered to be a challenging task and one concerning which research is ongoing.

The main challenge of load forecasting [66] is to achieve maximum possible accuracy. This is because the accuracy of load forecasting has significant effects on the efficacy of power system operations and on the associated operating costs. Most especially, as large forecasting errors will increase such operating costs, any reduction in average forecasting errors, however small, can save power utilities millions of dollars. With this as their motivation, several load forecasting studies have been conducted and empirically tested using various techniques, including statistical analysis and artificial neural networks [67].

## 3.5 Conclusion

In this chapter, three main data mining techniques have been reviewed, namely feature selection, classification, and prediction. Feature selection has been considered from the three perspectives of 1) search problem, 2) evaluation criteria, and 3) model selection. It comprises a process of searching the attributes space for the required subsets, with this being achieved by eliminating irrelevant attributes.

The classification task consists of extracting models to represent customer behaviour classes, as well as facilitating the prediction of future customer behaviours. Useful predictions concerning customer consumption status in the present context relate to three classification categories, namely non-unmetered power losses, normal behaviour, and suspicious behaviour. These classifications rely on given load profiles and on other relevant factors, such as weather, holiday banks, rates, and the nature of customers' businesses. This task forms a very important module within the proposed unmetered power losses framework of analysis because the classification behaviour results obtained are useful for monitoring any suspicious behaviour and because the developed behaviour model can be used to predict behaviour classes for new customers. Applying prediction techniques in electric power utilities involves the prediction of daily, weekly, monthly, and yearly of system loads, peak loads, and system energy. Such a prediction task is one of the most important among the planning and operational activities of these utilities that are designed to maximize the benefits accruing to them. Three approaches have been reviewed above, including the statistical approach, the time series approach, and the neural network approach.