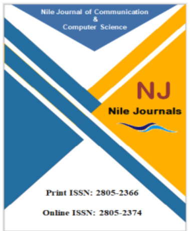
NJ
Nile Journals

Nile Journal of Communication & Computer Science

Volume 1, Number 1, August 2021

Journal Webpage: https://njccs.journals.ekb.eg



Nile Journal of Communication
&
Computer Science

NJ
Nile Journals

Print ISSN: 2805-2366

Online ISSN: 2805-2374

A Survey on Data Mining Techniques in Smart Grids (SGs)

Ahmed I. Saleh, Hesham A. Ali, and Asmaa H. Rabie
Computers and Systems Department, Faculty of Engineering, Mansoura University, Egypt

Abstract

Smart Grids (SGs) have already achieved wide adoption in information sensing, transmission, and processing. SGs are considered as an advanced digital 2-way power flow power system and are capable of self-healing, adaptive, resilient, and sustainable with foresight for prediction. Data mining play an effective role in SGs in which it enable SGs to be transformed from traditional grids to be an intelligent ones. In this paper, many classification methods which will affect performance of networks in the future are discussed. In fact, classification methods are used in SGs to provide accurate predictions such as electrical load prediction. Electrical load forecasting is a vital process for the electrical power system operation and planning. There are many methods used to improve the load forecasting accuracy in which these methods differ in the mathematical formulation and features used. The classical load forecasting techniques have more complex computational operations with low performance when compared to load forecasting methods based on data mining techniques. A review for feature selection and outlier rejection methods is presented as these processes are very important in data preprocessing phase that enable the prediction models to perform their tasks well.

***Keywords**: smart grids, load forecasting, data mining, classification, feature selection, outlier rejection.*

1. Introduction

SGs is a developed and intelligent electricity transmission and distribution network (grid) that employs modern information, communication (internetworking system), and control technologies to develop economy, efficiency, reliability, as well as security of the grid. Load forecasting is an essential process for the power system planning and operation to provide intelligence to energy management within smart grids. Electric load forecasting is categorized in terms of the planning horizon's duration: up to 1 day/week ahead for short-term, 1 day/week to 1 year ahead for medium-term, and more than 1 year ahead for long-term [1-6].Short-term forecasts are applied to schedule the generation of electricity. Medium-term forecasts are applied to schedule the fuel purchases. Long-term forecasts are applied for planning purposes of the power supply and delivery system.

Electric load forecasting techniques may be provided in two essential groups: widely used classical methods, and new soft computing techniques based on artificial intelligence [1-6]. Electric loads are affected by a variety of variables (features) such as time factors, weather conditions, class of customers, special events, population, economic indicators, trends in using new technologies, and electricity price [1-6].The main problem is that demand pattern is almost very complex due to the deregulation of energy markets. Therefore, finding an appropriate forecasting model for a specific electricity network is not an easy task. Thus, we are seeking to find appropriate forecasting model suitable for stable and mature utilities, as well as dynamic and normal or fast growing utilities with very high accuracy and speed. Consequently, the main objective of the development of forecasting models is to improve the forecasting accuracy and reduce it's time. Accurate forecasts lead to substantial savings in operating and maintenance costs, increased reliability of power supply and delivery system, and correct decisions for future development [1-6].

Data preprocessing is very important before implementing any prediction model as it remove any bad data from the input dataset to improve the performance of the model [1-6]. In preprocessing phase, there are two main processes which are performed called; feature selection and outlier rejection. While feature selection aims to select the most significant features, outlier rejection process aims to remove all an hasted data [1-6]. Feature selection methods are classified into two main types called filter and wrapper [3]. Outlier rejection methods are classified into three groups, namely; Statistical-Based Approaches (SBAs), Cluster-Based Approaches (CBAs), and Neighbor-Based Approaches (NBAs) [4].

In this paper, a review of different widely used classification methods applied as load forecasting techniques will be discussed. Additionally, feature selection and outlier rejection techniques will be discussed. The rest of the paper is organized as follows: Section 2 shows an overview of the widely used classification methods. Section 3 provides an overview about load prediction methods. Section 4 introduce a discussion about feature selection using data mining techniques. Section 5 discuss outlier rejection using data mining techniques. Finally, conclusions are discussed in Section 6.

2. Classification Techniques

Over the centuries, several researchers directed their biggest attention to improve classification techniques [1-9]. These techniques are extended from traditional to soft computing or data mining techniques. Some of classification techniques which are widely applied are called Fuzzy logic (FL), Support Vector Machines (SVMs), regression technique, and time series models. Recently, ensemble classification model has been used to predict the load. Ensemble model is a combined model that mixes several single algorithms together to improve the prediction accuracy [10].

2.1. Fuzzy Logic Techniques

FL is an artificial intelligence technology that has a large application in ELF [11]. It is applied to handle the information of input parameters after detailed analysis of data and knowledgebase (IF-THEN rules) [34]. FL provides fuzzy values rather than crisp values of loads in which the predicted load under several conditions had different memberships according to every category of the load levels. Hence, FL can address ELF perfectly because it is similar to human decision making. Additionally, it has the ability to produce accurate solutions from certain or approximate information [12]. Using membership function, FL can be applied to provide the highly non-linear relationship between input parameters and their effects on the electrical load [11].

2.2. SVM Techniques

SVM is considered one of the popular and efficient techniques that is able to solve problems related to classification and regression. Even if there is a limited amount of data, SVM has the ability to provide accurate results [10]. Non-linear mapping can be performed on the data by SVMs to represent it in a high dimensional space using kernel functions. Selecting appropriate kernel parameters is very important because these parameters indicate to the general performance of the SVM model. SVM is better than other popular data-driven techniques such as ANNs [10]. Indeed, ANNs try to define complex functions of the input feature space. On the other hand, SVMs are able to create linear decision boundaries in the new space by using simple linear functions. In fact, SVMs have a competitive forecasting performance, and its application is more flexible. Although SVMs have many benefits, the evaluation of their parameter is a hard but crucial process for providing accurate predictions [10].

2.3. Regression Technique

Regression is a more recent powerful method applied as statistical technique in which its accuracy depends on the data representation. Data representation means the appropriate representation of future conditions or factors based on historical data [8]. For ELF, regression methods have been applied to model the functional relationship of load consumption and other factors. For example on these factors are; customer class, day type, and weather. Regression is defined as a mathematical model by (1):

$$L(t)=Ln(t)+\sum ai\ xi(t)+e(t) \qquad (1)$$

Where the standard load at time t is denoted by Ln(t). The estimated slowly varying coefficients are denoted by ai, and the independent affected factors like weather effect are denoted by xi(t). Additionally, a white noise component is dented by e(t), and the number of observations is symbolized by n, in the most cases is 24 or 168 [8]. Although the simplicity of implementing regression technique, it has no ability to deal with non-linear problems [10]. Besides, it needs a huge amount of data to capture all possible scenarios.

2.4. Time Series Models

Time series models analyze the load evolution by using a mathematical tool [1,17]. Load analysis aims to find the dynamic characteristics of the load. Moreover, it aims to extrapolate these characteristics into the future. Thus, the load can be decomposed into components that can be treated separately. In time series forecasting, the future load is represented only as a function of the previous loads. In a time series plot, the modeling patterns can provide more reliable predictions [13]. Thus, time series is a prediction model that generates a mapping between input and output data [1]. The structure of time series models is simple and observations on the variable under study are completely sufficient. Although time series models have many benefits, it have drawbacks such as a cause-and-effect relationship cannot be described. Additionally, insights into why changes have taken place in the variable cannot be introduced by these models.

2.5. Ensemble Classification Method

Ensemble classification method is a promising approach that helps to improve machine learning results by integrating several models [14]. The idea of ensemble approach is to produce a better predictive performance compared to a single model. Surely, ensemble classification method is a data-driven technique built for prediction applications [9,10]. Basic of a typical ensemble classification is illustrated in figure 1. The framework of ensemble modeling consists of two essential steps which are;

Several sub-models: these sub-models either are called “base learners” for homogenous ensemble classification models or “base models” for heterogeneous ones [10].
Perform the comparison on the respective prediction results from these sub-models: the accuracy of these results are used to give a weight of them. Then, the best output of the ensemble classification model is produced by the integration of these results [10].

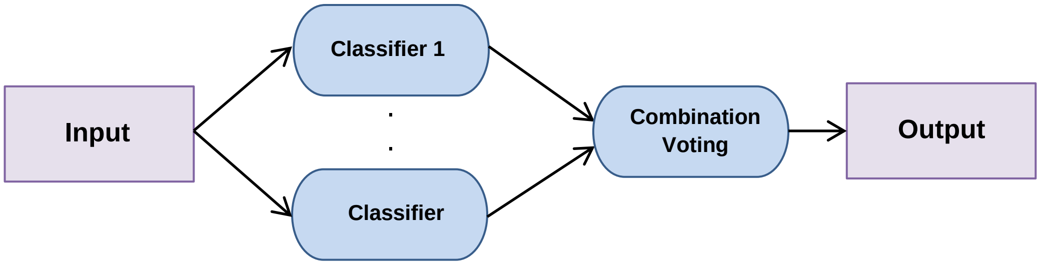


Figure 1: Basic of a typical ensemble classification.

Generally, there are two main approaches for the ensemble modeling process which are; homogenous modeling, and heterogeneous modeling [10,36]. Homogenous modeling creates different subsets from the original dataset. These subsets can be analyzed and processed by models of the same type, and then the results for each subset of data are produced. The prediction performances of the base learners are used for giving a weight to each base leaner. Then, these base leaners are combined into the ensemble classification model. On the other hand, heterogeneous modeling depends on using many different single prediction methods which are trained on the same dataset (in *step i*). Then, it is a vital process to give weights to the prediction results of the base models for providing the ensemble classification model (in *step ii*).

3. Load Prediction Techniques

Since 1990's, there are several efforts by researchers to solve ELF's problem. Many ELF techniques have been provided for improving the forecasting accuracy [6]. These techniques depend on a different mathematical formulation, and every formula also uses different features. Usually, there are three major groups used to classify the ELF techniques. These groups are; (i) Traditional Load Forecasting (TLF) techniques, (ii) Modified Traditional Forecasting (MTF) techniques, and (iii) Soft Computing Forecasting (SCF) techniques [6]. These techniques are illustrated in figure 2.

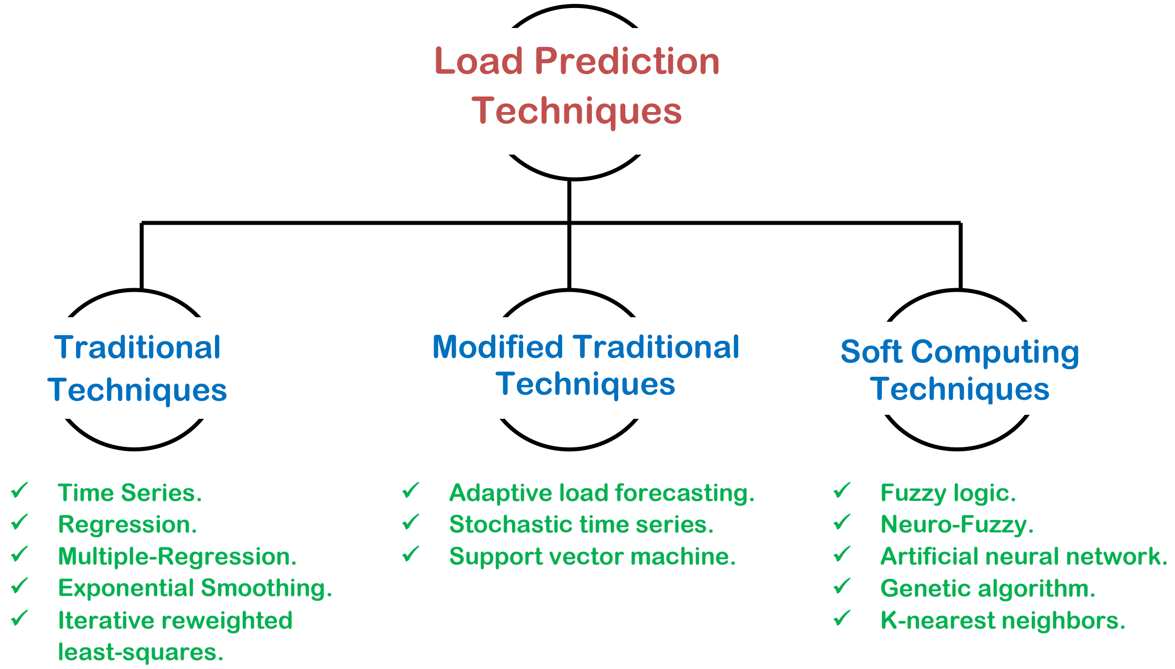

Load Prediction Techniques
Traditional Techniques
Modified Traditional Techniques
Soft Computing Techniques
✓ Time Series.
✓ Regression.
✓ Multiple-Regression.
✓ Exponential Smoothing.
✓ Iterative reweighted least-squares.
✓ Adaptive load forecasting.
✓ Stochastic time series.
✓ Support vector machine.
✓ Fuzzy logic.
✓ Neuro-Fuzzy.
✓ Artificial neural network.
✓ Genetic algorithm.
✓ K-nearest neighbors.

Figure 2: Recent load prediction techniques [6].

4. Feature Selection Methods

Feature selection process is a vital process in preprocessing phase to analyze the data. Feature selection aims to select the most informative data before starting to learn the perdition model. Wrapper and filter represents the main categories of feature selection methods. Filter approach depends on the information content of feature subsets to assess these subsets using correlation measures...etc. On the other hand, classifier is implemented on each subset of features considered by the search in the wrapper approach. A common taxonomy of feature selection techniques are given in table 1 [6].

Table 1: A taxonomy of feature selection techniques [6].

		Advantages	Disadvantages	Examples
Filter	Univariate	• Fast. • Scalable. • Independent of the classifier.	• Ignores feature dependencies. • Ignores interaction with the classifier.	• Chi-square. • Euclidean distance. • t-test. • Information gain, Gain ratio.
	Multivariate	• Models feature dependencies. • Independent of the classifier. • Better computation complexity than wrapper methods.	• Slower than univariate techniques. • Less scalable than univariate techniques. • Ignores interaction with the classifier.	• Correlation based feature selection (CBFS). • Markov blanket filter (MBF). • Fast correlation based feature selection (FCBF).
Wrapper	Deterministic	• Simple. • Interacts with the classifier. • Models feature dependencies. • Less computationally intensive than randomized methods.	• Risk of over fitting. • More prone than randomized algorithms to getting stuck in a local optimum (greedy search). • Classifier dependent selection.	• Sequential forward search (SFS). • Sequential backward elimination (SBE). • Plus *q* take-away *r*. • Beam search.
	Randomized	• Less prone to local optima. • Interacts with the classifier. • Models feature dependencies.	• Computationally intensive. • Classifier dependent selection. • Higher risk of overfitting than deterministic algorithms.	• Simulated annealing. • Randomized hill climbing. • Genetic algorithms. • Estimation of distribution algorithms.

5. Outlier Rejection Methods

Outlier rejection process is very important in the preprocessing phase. Outlier detection targets to find an observation that deflects too much from other observations. the identification of outliers can be the main target of the analysis or can help to discover the useful knowledge [4,6]. Thus, outlier rejection mainly aims to detect and then eliminate bad “noise” data. Outlier rejection techniques can broadly classified into two main categories called classic outlier approach and spatial outlier approach as illustrated in figure 3 [6]. As shown in figure 3, outlier rejection methods are classified into three groups, namely; Statistical-Based Approaches (SBAs), Cluster-Based Approaches (CBAs), and Neighbor-Based Approaches (NBAs) [4].

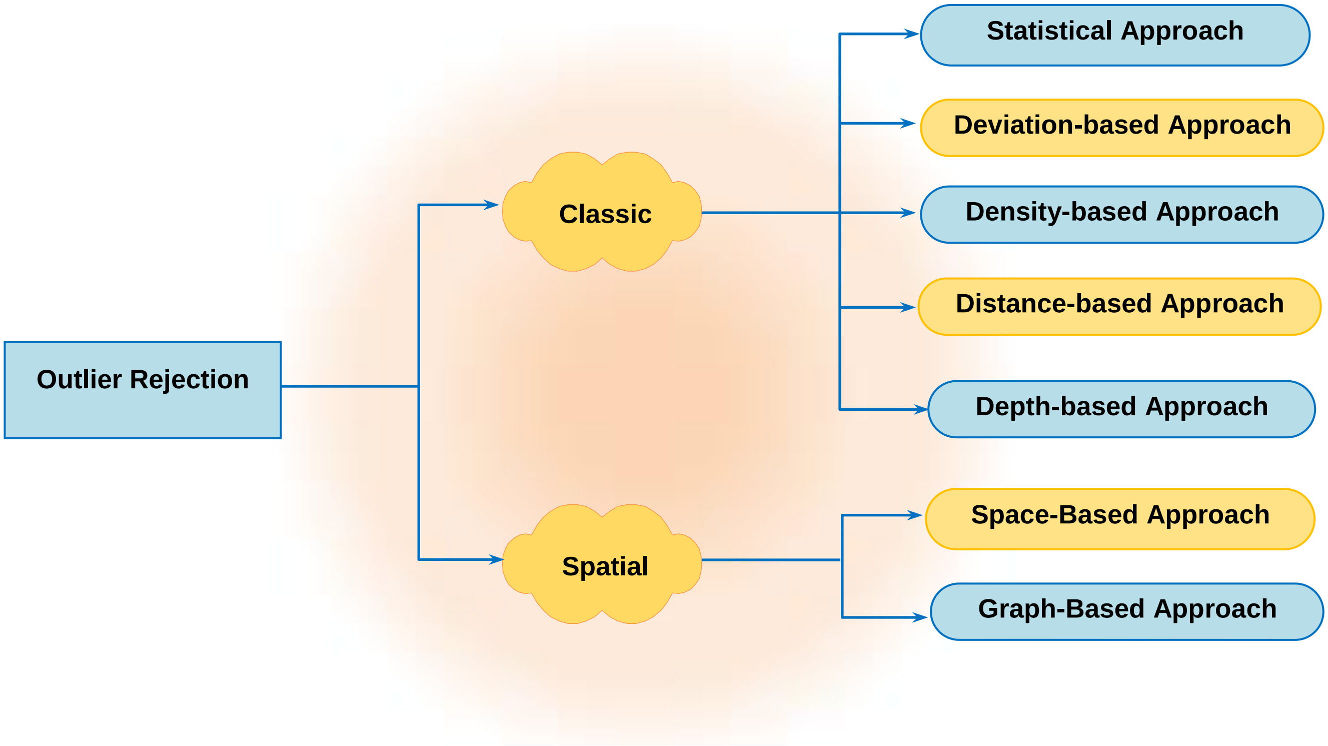


Figure 3: Outlier Rejection Techniques [6].

6. Conclusions

SGs are becoming a global trend nowadays which are capable of self-healing, adaptive, resilient, and sustainable with foresight for prediction. In fact, The usage of classical load prediction methods either wasted the time or degraded the prediction accuracy. Thus, it is conducted that using the modern classification methods based on data mining techniques as prediction methods in SGs are important to provide fast and accurate predictions. Classifications methods such as; FL, SVMs, Regression, Time series, and ensemble classification. Preprocessing of data is very important to enable the classifier or the prediction model to perform its tasks well. Many feature selection methods were provided and also many outlier rejection were discussed as these method are important in the preprocessing phase.

7. References

[1] A. Rabie, A. Saleh, and K. Abo-Al-Ez, *"A New Strategy of Load Forecasting Technique for Smart Grids,"* International Journal of Modern Trends in Engineering and Research (IJMTER), Volume 2, Issue 12, 2015, PP.332-341.

[2] A. Saleh, A. Rabie, and K. Abo-Al-Ezb, *"A data mining based load forecasting strategy for smart electrical grids,"* Advanced Engineering Informatics, Elsevier, Volume 30, Issue 3, 2016, PP.422- 448.

[3] A. Rabie, S. Ali, H. Ali, and A. Saleh*, "A fog based load forecasting strategy for smart grids using big electrical data,"* Cluster Computing, Springer, Volume 22, Issue 1, 2019, PP.241- 270.

[4] A. Rabie, S. Ali, A. Saleh, and H. Ali*, "A new outlier rejection methodology for supporting load forecasting in smart grids based on big data,"* Cluster Computing, Springer, Volume 23, 2020, PP. 509-535.

[5] A. Rabie, S. Ali, A. Saleh, and H. Ali*, "A fog based load forecasting strategy based on multi-ensemble classification for smart grids,"* Journal of Ambient Intelligence and Humanized Computing, Springer, Volume 11, Issue 1, 2020, PP. 209-236.

[6] A. Rabie, A. Saleh, and H. Ali, *"Smart electrical grids based on cloud, IoT, and big data technologies: state of the art,"* Journal of Ambient Intelligence and Humanized Computing, Springer, https://doi.org/10.1007/s12652-020-02685- 6, 2020, PP. 1-32.

[7] N. Mansour, A. Saleh, M. Badawy, and H. Ali, "*Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy,*" Journal of Ambient Intelligence and Humanized Computing, Springer, https://link.springer.com/article/10.1007/s12652-020-02883-2, 2021, PP. 1-33.

[8] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi,*"Big Data management in smart grid: concepts, requirements and implementation,"* Journal of Big Data, Springer, Volume 4, Issue 13, 2017, PP.1-9.

[9] C. Tu, X. He, Z. Shuai, and F. Jiang, *"Big data issues in smart grid – A review,"* Renewable and Sustainable Energy Reviews, Elsevier, Volume 79, 2017, PP. 1099-1107.

[10] M. Ansari, V. Vakili, and B. Bahrak, "*Evaluation of big data frameworks for analysis of smart grids,"* Journal of Big Data, Springer, Volume 6, Issue 109, 2019, PP.1-14.

[11] R. Moghaddass and J. Wang, *"A hierarchical framework for smart grid anomaly detection using large-scale smart meter data,"* IEEE Transactions on Smart Grid, Volume 9, Issue 6, 2018, PP. 5820-5830.

[12] K. Vimalkumar and N. Radhika, *"A big data framework for intrusion detection in smart grids using Apache Spark,"* Proceedings of the 2017 International conference on advances in computing, communications and informatics (ICACCI), IEEE, 2017, PP. 198-204.

[13] G. Sheng, H. Hou, X. Jiang, and Y. Chen, *"A Novel Association Rule Mining Method of Big Data for Power Transformers State Parameters Based on Probabilistic Graph Model,"* IEEE Transactions on Smart Grid, Volume 9, Issue 2, 2018, PP. 695-702.

[14] A. Kumari, S. Tanwar, S. Tyagi, et al., *"Fog data analytics: A taxonomy and process model,"* Journal of Network and Computer Applications, Elsevier, Volume 128, 2019, PP.90-104.