Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 118 (2015) 1189 – 1194

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Sustainable Design, Engineering, and Construction

Semi-Supervised Energy Modeling (SSEM) for building clusters using machine learning techniques

Hariharan Naganathan[a], Wai K. Chong[a*], Xue-wen Chen[b]

[a]*School of Sustainable Engineering and Built Environment, Arizona State University, Tempe, AZ, United States.*
[b]*Professor, Computer Science Department, Wayne State University, Detroit, Michigan, United States.*

Abstract

There are many data mining and machine learning techniques to manage large sets of complex energy supply and demand data for building, organization and city. As the amount of data continues to grow, new data analysis methods are needed to address the increasing complexity. Using data from the energy loss between the supply (energy production sources) and demand (buildings and cities consumption), this paper proposes a Semi-Supervised Energy Model (SSEM) to analyse different loss factors for a building cluster. This is done by deep machine learning by training machines to semi-supervise the learning, understanding and manage the process of energy losses. Semi-Supervised Energy Model (SSEM) aims at understanding the demand-supply characteristics of a building cluster and utilizes the confident unlabelled data (loss factors) using deep machine learning techniques. The research findings involves sample data from one of the university campuses and presents the output, which provides an estimate of losses that can be reduced. The paper also provides a list of loss factors that contributes to the total losses and suggests a threshold value for each loss factor, which is determined through real time experiments. The conclusion of this paper provides a proposed energy model that can provide accurate numbers on energy demand, which in turn helps the suppliers to adopt such a model to optimize their supply strategies.

Peer-review under responsibility of organizing committee of the International Conference on Sustainable Design, Engineering and Construction 2015

Keywords: Semi-supervised learning, Energy Modeling, Demand- supply analysis, Energy losses, Labelled and Unlabelled factors.

Corresponding author: E-mail address:ochong@asu.edu

Peer-review under responsibility of organizing committee of the International Conference on Sustainable Design, Engineering and Construction 2015
doi:10.1016/j.proeng.2015.08.462

1. **Introduction**

The liberalization on power sectors allows customers the freedom of choosing their suppliers from their choices [1]. This creates competition among the suppliers to reach their targets of consumers and most of the suppliers try to go by traditional and manual statistical methods to balance their demand-supply problems. Although equipment and physical systems have well-established energy models that define systems in real-time aspects, the pre-defined system models for *multi-tiers energy demand* (equipment, systems, building contents, building, building cluster, community, city and state) are still relatively new. The accuracy and usability of traditional energy demand forecasting methods are limited by the nature and extensiveness of the data and their analysis techniques. Energy inefficiencies occurs at every tiers, and the benefits of new energy saving technologies may not contribute to the reduction of energy supply. The energy saving technologies elevates the indoor energy quality to a greater extent. This cannot help the power suppliers to reduce their production as energy losses in various forms are not considered by most of the energy models.

Machine learning is becoming an extremely popular tool to assist engineers better manage energy production. This research focuses on developing a Semi-Supervised Energy Model (SSEM), a real-time energy demand and supply network model that would accurately estimate the energy consumption of building clusters. This is done by predicting the energy demand and supply for every cluster through extensive implementation of semi-supervised learning techniques that involves deep machine learning algorithms and techniques. SSEM trains the machine through a definite pattern of labeled data, and integrates the reliable unlabeled data (which is the percentage of loss factors in this research) in order to determine the energy loss values. Through the learning process, the machine can predict the energy loss percentage accurately by analyzing labeled and unlabeled factors that account for the energy loss. This helps the researchers to understand the characteristics of losses and how much energy is lost in real time. This model can be developed into a dynamic model that will be an important decision-making and marketing strategy tool for the practitioners and industrial pioneers. With the large volume of labeled and unlabeled data, the focus of the research is to propose a model to reduce the energy losses between demand and supply sources.

2. **Semi-Supervised learning techniques**

"Deep" machine learning is currently being explored as one of the more reliable techniques for semi-supervised machine learning. One of the advantages of deep learning over traditional neural networks is the ability to utilize unlabelled data for unsupervised pre-training of machine [2]. This technique discovers the feature of the data by itself and follows by a fine-tuning stage where labelled data carves the parameters further for discrimination and accuracy of outputs. Unsupervised pre-training is often time consuming process.

Recent research demonstrated that unlabelled data can be used differently to improve the reliability of data analysis [3], i.e. unlabelled data and labelled data can be learned simultaneously in a semi-supervised manner. Compared with other semi-supervised approaches, which are usually based on Support Vector Machine (SVM) methods, deep learning based approaches are expected to be more reliable. Deep learning has become the new state-of-the-art technique for many difficult artificial intelligence tasks. In addition, the learning process is less complicated since both labelled and unlabelled data can be learned simultaneously and hence it is less time consuming.

Using labelled data, the paper develops a machine learning technique to estimate energy loss between supply and demand sources. The paper also lays out a novel approach for a semi-supervised learning based on the deep learning framework. The approach carefully selects part of unlabelled data that has high confidence interval. The data and analyses are then verified so that the time frame and the methodology to integrate the unlabelled data with the supervised learning process can be optimized.

3. **Cluster Analysis and Its Limitations**

Data were collected from the Energy Information System (EIS) for 119 buildings from an unidentified university campus. The data includes electricity (generated by different sources), heating load, cooling load, watt/square feet, human counts, occupancy at different time of the day, heat index and outside temperature of the respective building.

The data are divided into fifteen minutes intervals and include energy supply and demand from the buildings. Each building is connected with one of the four substations on campus (labelled as north, south, west and central). The research team carried out an extensive analysis on each substation, and each substation supplies electricity to a cluster of buildings. Figure 1 below show four different building clusters (each supplied by a substation) and their supply-demand curve (for one month, 2013). The analyses for the building clusters highlights the supply-demand curve of which the energy loss can be determined. The energy loss varies widely with respect to each cluster.

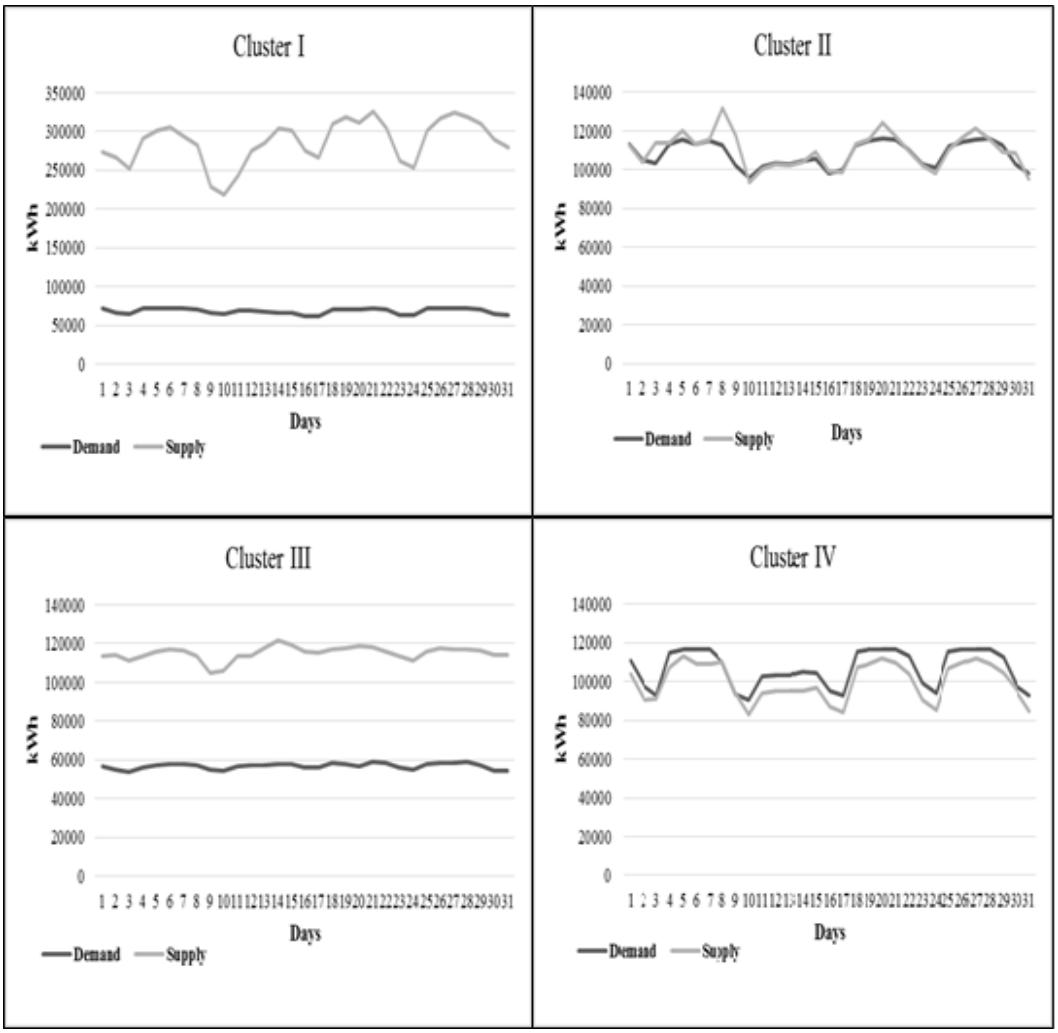

Fig. 1. Supply-Demand Curves for Clusters I-IV for one month (2013)

For example, the energy loss for Clusters I is around 10000 kWh which makes no sense as there cannot be such a huge loss from one cluster of buildings. The research team need to verify the data and look at various reasons that contribute to this loss. Similarly Cluster III has the biggest of losses between supply and demand. Thus Cluster I and III are much higher than Clusters II's and IV's – indicated by the large gaps between the demand and supply curves in the figures. The gaps are caused by unknown direct or indirect parameters due to both technical and non-technical energy losses, which are explained later in the paper. The analyses indicate the energy loss suffered by each cluster. Figure above explains the losses for only one month and involves regression and other statistical analysis to determine various other reasons for the losses. Since the reasons behind the loss are unknown (self-learning by machines are impossible to generate the reasons), the researchers have to conduct extensive studies to determine the

reasons. The reasons can swirls around data loss, distribution line loss, material damage, meter losses, theft or any other unidentified losses. This approach, however, is far less time consuming than traditional statistical methods which is highly labor intensive and can result in low accuracy due to human errors.

4. Semi-Supervised Energy Model

Techniques used to model building energy consumption can broadly be grouped into top-down approach, which includes econometric and technological approaches whereas bottom-up includes engineering and statistical approaches [4]. The cluster analysis above comes under the bottom-up approach since it involves statistics. The primary drawback is the calculation or simulation techniques of the bottom-up models can be complex [4]. The proposed model is a simpler and more effective method that semi-supervises the machine using deep learning approach where pseudo-labels of unlabeled data are calculated during every update based on current parameters [5]. And the pseudo-labels are treated as the real labels so that unlabeled data can be learned as if they are labeled data. The overall loss function is mentioned in the equation 1 below.

$$L = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f_i\right) + \alpha(t)\frac{1}{n'}\sum_{j=1}^{n'} L\left(y_j, f_j\right) \tag{1}$$

where the first term represents loss value between demand-supply of building clusters of "n" labeled data with y_i, (the desired output vector for energy supply) for sample x_i (identify each building cluster) and f_i (actual output for energy demand). The second term represents "n'" unlabeled data with y_j being the pseudo-label for sample x_j and f_j the actual output. The difference between the desired output and the actual output gives the loss value of demand-supply curve.

The second term includes various loss factors that contribute to the total losses. These loss factors include both technical and non-technical losses. Technical losses include losses through circuits, meters, transformers and distribution. Each factor has a threshold value which is the base loss percentage. The International electro technical commission suggests that the distribution loss percentage ranges from 7% -10% on an average *[6]*. So the threshold value for this distribution loss is assumed to be 10%, which is taken as actual output expected during distribution to the building clusters. Thus, the research decides on different threshold percentage for different loss factors. The non-technical loss generally includes time switch errors, theft, metering and recording errors and unmetered supplies. This paper focuses only on technical losses at this point since the data collection on non technical losses is still under process and requires more time to collect enough data to be included into the research. Lee (2013) treated unlabeled data as equal for the loss function, even though in reality, they are not. Treating them equally will result in labeling them with wrong pseudo-labels. In other words, those wrongly predicted unlabeled data (loss factors) might be playing a misleading role, and result in poor or unstable generalization performance.

The key difference (unique to this research) between the proposed method and the actual method proposed by Lee (2013) is that, instead of taking all the unlabeled data into the training process and gradually increasing the importance for each set of data, the research team selected the data base on their expected confidence interval and treated them as labeled data throughout the analysis. Since the proposed method relies on the prediction confidence, consequently, it is called confidence-based semi-supervised learning (CSL). CSL approach may have an issue with the threshold values that define the prediction confidence level of each data point. A reasonable choice of the threshold can overcome the issue by guaranteeing that majority of the evaluation samples are correctly classified.

4.1 Selection of Loss Factors

In order to select reliable unlabeled data, the differences between the desired and actual loss factor percentages are observed. The percentage is used to determine the confidence level of the data. Table 1 represents how the loss factors contribute to the total energy savings. Table explains the three difference instances of integrating the loss factors (unlabeled data) into the labeled data and how the machine learns.

The loss factors in the Table 1, L_1, L_2 and L_3 are the desired output percentage gathered from data collection. The actual output (the base threshold value) is selected based on various parameters. The threshold plays another role to

indicate the confidence level of selected unlabeled data. The difference between the outputs is observed to determine whether the percentage is positive or negative. The machine is trained to understand that if the desired output is lesser than the actual output (i.e. negative), it has to be omitted from the cumulative loss percentage. This process is repeated for all loss factors and the total loss percentage is calculated only from the positive values.

Table 1 Loss factors and its selection

Energy data and Loss factors		**Instance 1**		**Instance 2**		**Instance 3**	
Factors	Outputs	Loss (%)	Inference	Loss (%)	Inference	Loss (%)	Inference
	Desired	a_1		a_2		a_3	
L_1	Actual	b_1		b_2		b_3	
	Difference	$a_1 - b_1$	Positive	$a_2 - b_2$	Positive	$a_3 - b_3$	Positive
	Desired	p_1		p_1		p_3	
L_2	Actual	k_1		k_1		k_3	
	Difference	$p_1 - k_1$	Negative	$p_1 - k_1$	Positive	$p_3 - k_3$	Negative
	Desired	z_1		z_1		z_3	
L_3	Actual	c_1		c_1		c_3	
	Difference	$z_1 - c_1$	Positive	$z_1 - c_1$	Negative	$z_3 - c_3$	Negative
Total Loss %			**L_1+L_3**		**L_1+L_2**		**L_1**

The cumulative total loss percentage is then compared with the output from first term to understand how much losses could be reduced. Users of the model are required to take into consideration the timing where the unlabeled data are included in the model (referred to as the transition point). A recommendation is for the users to wait until the supervised learning on labeled data is convergent with the unlabeled data, so that the unlabeled data become clearer and thus more identifiable. Consequently, the proposed method minimizes the conditional entropy for unlabeled data to lower the density of class overlapping at the decision boundary, which explains why unlabeled data can help improve the classification performance.

5. Conclusion

The research team proposes a semi-supervised learning that predicts the classes of unlabeled data using labeled data in the first stage, and selects only those reliable sets of unlabeled data to be included in the semi-supervised learning stage. Instead of utilizing all the unlabeled data indiscriminately, the proposed method measured the confidence level of the data before using them. This helps to improve the accuracy of the output results on loss prevention. The proposed method also identifies the contributions of the positive loss factors toward energy savings. The proposed concept can be extended to incorporate with a different cluster and include identified and non-identified loss factors which improves the efficiency of the output. Such machine learning helps suppliers understand the underlying reasons behind the losses, by integrating the expertise of facility managers, engineers and architects.

The paper proposes a preliminary framework for Semi-Supervised Energy Model (SSEM). The reliability and accuracy of this model is under testing. It is known that complexity requires more time to implement such kind of the model and also requires extensive knowledge on machine learning and cybernetic concepts. Training the machine with algorithms are considered to be the high complex task of this research after which the machine learns the pattern and automate the model for greater accuracy. These complexities claims the drawbacks of this research at this time which eventually will be eliminated in later part of the research by confidence based semi supervised learning techniques.

The future scope involves almost half a million data with labelled and unlabeled data. Also the research will be directed towards utilizing the model for different factors like heating, cooling to understand their patterns and

automate their efficient strategies. This can ameliorate the energy savings and provide more insights to the decision makers on the important factors to achieve sustainability for the future.

6. References

[1] Figueiredo, V., Rodrigues, F., Vale, Z. and Gouveia, J.B., "An electric energy consumer characterization framework based on data mining techniques," *Power Systems,* vol. 20, no. 2, pp. 596-602, May 2005.

[2] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," ScienceMag, 2006.

[3] J. Weston, F. Ratle, H. Mobahi and R. Collobert, "Deep Learning via Semi-Supervised Embedding," in *The 25th International Conference on Machine learning*, Finland, 2008.

[4] L. G. Swan and I. Ugarsal, "Modeling of end-use energy consumption in the residential sector: A review of modelling techniques," *Renewable and Sustainable Energy Reviews,* vol. 13, no. 8, pp. 1819-1835, 30 September 2008.

[5] D. Lee, "Psuedo Label: The Simple and Effcient Semi-Supervised Learning Method for Deep neural Networks," Atlanta, 2013.

[6] IEC, "Energy," IEC, Geneva, 2010.

[7] H. Kim, A. Stumpf and W. Kim, "Analysis of an energy efficient building design through data mining approach," *Automation in construction,* vol. 20, no. 1, pp. 37-43, 3 May 2010.

[8] Sohn, "Electricity distribution losses," Sohn Associates, London, 2009.

[9] IEC, "Efficient electrical energy transmission and distribution," IEC, Geneva, 2010.