International Journal of Digital Information and Wireless Communications (IJDIWC) 7(2): 118-123
The Society of Digital Information and Wireless Communications, 2017 ISSN: 2225-658X (Online); ISSN 2412-6551 (Print)

Data Mining and Analysis for off Grid Solar PV Power System

Mohamed D. Almadhoun
Department of IT, University College of Applied Sciences
Gaza, Palestine
mdmadhoun@ucas.edu.ps

ABSTRACT

Solar power generation systems production changes with respect to geographical location. By this paper we studied how solar systems behave in Gaza which is located in the Middle East. Log readings of off grid solar PV power supply system included PV module, array combiner, off grid inverter, battery and load. Data entered many techniques of data mining to discover the hidden and meaningful patterns, and to analyze results.
Classification was applied to discover rules which can help in supporting solar power system functionality of alerting about expected power generation and needed costs. Association rules were applied to find relationships between variables and plot events of power changing. Clustering were used to categorize power production values and study properties of each.
The scope of study extended to discover effects of solar and moon months, input voltage and charging current, daytime hours, daytime temperatures, and effects of week days on power generation.
Resulting knowledge were novel, actionable, understandable, and valid. This was stated by tracking the confidence rates and applying evaluation by cross validation.

KEYWORDS

Renewable energy, Solar systems, PV power generation, data mining, Knowledge Discovery in Database.

1 INTRODUCTION

Data mining is an approach to extract patterns from large data sets and deduce knowledge insights from patterns [1]. Data mining can find unsuspected relationships and summarize the data in novel ways that are both understandable and useful to the data owner [2].

Photovoltaic (PV) is a method of generating electrical power by converting solar radiation into direct current electricity using semiconductors that exhibit the photovoltaic effect. PV power generation employs solar panels composed of a number of solar cells containing a photovoltaic material. Manufacturing of solar cells and photovoltaic arrays has advanced considerably in recent years because of growing demand for renewable energy sources [3].

Operating costs for the electricity system may increase because of the unpredictability of the sun and unexpected variations of a PV output which leads to potential threats to the reliability of electricity supply. Predicting changes is a main concern to schedule the reserve capacity and to administer the grid operations [4]. PV power is affected by the weather and other natural factors dramatically. To reduce the energy dissipation and maintain the security of power grid we should give priority to predicting PV energy accurately [3].

In the last years data mining and related applications were used in predicting solar radiation and solar energy system design to give an overview of such predictive data mining techniques [5]. Due to the strong increase of solar power generation, the predictions of incoming solar energy are too important and necessary to predict the amount of energy which will be produced, up to 72 h before, and deviations of energy production are strongly penalized [5][6].

In this study data mining is used to find relationships and rules, and the development trends of PV solar power generation management systems [7][5].

This paper follows the process of knowledge discovery with all steps starting from data gathering, followed by data cleaning, and aggregation to make data ready to be utilized for data visualization and data mining, reaching to the evaluation and knowledge representation. Data mining phase will be shown in details and knowledge represented from each data mining model.

Three data mining models used: first was classification by decision tree, second was

association rules by FP-Growth approach, and finally clustering by k-Means approach.

2 RELATED WORKS

Qazi et. al. [5] proposed an overview of predictive data mining techniques and highlighted the importance of solar energy systems in terms of clean environment. Their paper concluded that ANN-based prediction offers greater accuracy, and hence it's much more dependable and demanding in the domain of renewable energy resource predication. Their study demonstrated that ANN models predict solar radiation more accurately than statistical, conventional, linear, nonlinear and fuzzy logic models. As a future work they aim to carry out in depth research on acceptance of solar systems in Malaysia.

Hossain et. al. [4] presented an architectural framework for the construction of hybrid intelligent predictor for solar power. Their research investigated the applicability of heterogeneous regression algorithms for 6 hour ahead solar power availability forecasting using historical data from Rockhampton, Australia. Real life solar radiation data was collected across six years with hourly resolution from 2005 to 2010. They observed that the hybrid prediction method is suitable for a reliable smart grid energy management. Prediction reliability of the proposed hybrid prediction method was carried out in terms of prediction error performance based on statistical and graphical methods. The experimental results show that the proposed hybrid method achieved acceptable prediction accuracy.

Wang et. al. [3] proposed a big data system (Solar Photovoltaic Power Forecasting System, called SPPFS) to calculate and predict the power under real-time conditions. In their system, they utilized the distributed mixed database to speed up the rate of collecting, storing and analysis the meteorological data. In order to improve the accuracy of power prediction, the given neural network algorithm has been imported into SPPFS. By adopting abundant experiments, they show that the framework can provide higher forecast accuracy-error rate less than 15% and obtain low latency of computing by deploying the mixed distributed database architecture for solar-generated electricity.

3 DATA MINING FOR UCAS SOLAR ENERGY SYSTM

A. Data resource

Dataset was created by EAST solar inverter of model GF6000. This system is equipped in the University College of Applied Science (UCAS) which is located in Gaza via Palestine; Dataset is for device readings for 18 months, with 276,657 instances.

Dataset elements are numerous, general like Date time, AC input voltage, AC output voltage, AC output frequency, load level, temperature, PV input voltage, PV charging current, battery voltage, battery capacity level, PV power, daily generated energy, and total generated energy. Some of these elements were not used in data mining techniques because they are not related to input/output operations. In addition, four new attributes where calculated and extracted from date/time field, first of them was the period of solar month (such as Jan, Feb, or Dec), second was period of moon month (lunar month such, includes months such as Muḥarram, Safar, or Ramadan), third was week day, and fourth was time hour period.

B. Data mining process and tasks

Figure 1 lists data mining process steps. First phase of data mining process was business understanding of how PV solar system works and what benefits of this system to UCAS. This information was gotten from specialists in the computer center department in UCAS.

Second phase data understanding achieved by collecting information about each data element, understanding what these values mean and what are effects of it.

By the aid of MS excel and rapidminer software third phase data preparation started. MS Excel operations helped in generating new attributes from date/time field. Solar month period was the first which includes one of three values: 1, 2, or 3, value of 1 means that this record was added in a date between 1 and 10 in solar month. Second was Moon month period which includes values of 1,2, or 3, and this was calculated on basis of Hijri calendar. Thirdly, week day was calculated. and finally time period number which implements day time hours, that means if value of time period num attribute was

10 then this reading sample was added to database between 10 AM to 11 AM.

After then on rapid miner processes, interesting elements were selected (Temperature, PV input voltage, PV charging current, PV power, solar month, moon month, time period number, and week day). Instances with PV power value of zero value were filtered out, so number of instances became 125,765. Because of lots of values and to make value meaningful, discretize by binning was applied to each of PVInputVoltage, PVPower, and Temprature with number of bins equals five.

Fourth phase modeling was applied through three different tasks, each of which informs about different knowledge. They were applied using rapidminer software.

First task was classification using decision tree technique which is a tree-structured plan of a set of attributes having several possible alternative branches of values in order to predict the class label. Classification can be useful in the field of PV power generation systems in expecting how much power can be generated in case we have some information about atmosphere and calendar. It also can give us an indication on quantities of fuel needed to power generation station in case of cities that use hybrid power sources. Figure 2 shows a branch of resulting decision tree.

Decision tree model was rich of knowledge, pruning was used to get rid of useless branches to avoid overfitting. Its criterion was information gain, and maximal depth of 4. In case of selecting attributes: moon month period, solar month period, temperature, PVpower (class label), and PVChargingCurrent, some knowledge inferred from decision tree was such as: most range values of PVPower matched the same range value of PVChargingCurrent attribute which means positive relationship, solar month period appeared in the tree but moon month period did not and that means solar month period has more effect on PVChargingCurrent than moon, also we find in two cases that third part of solar month causes a degradation in generated power despite of high charging current. Decision tree model algorithm seeks to build tree with least levels and this depends on calculating information gain that increases with the average purity of subsets that an element produces [8], so PVChargingcurrent attribute was selected as root element in the tree, and that confirms that the most affecting input of generated PV power is PVChargingcurrent.

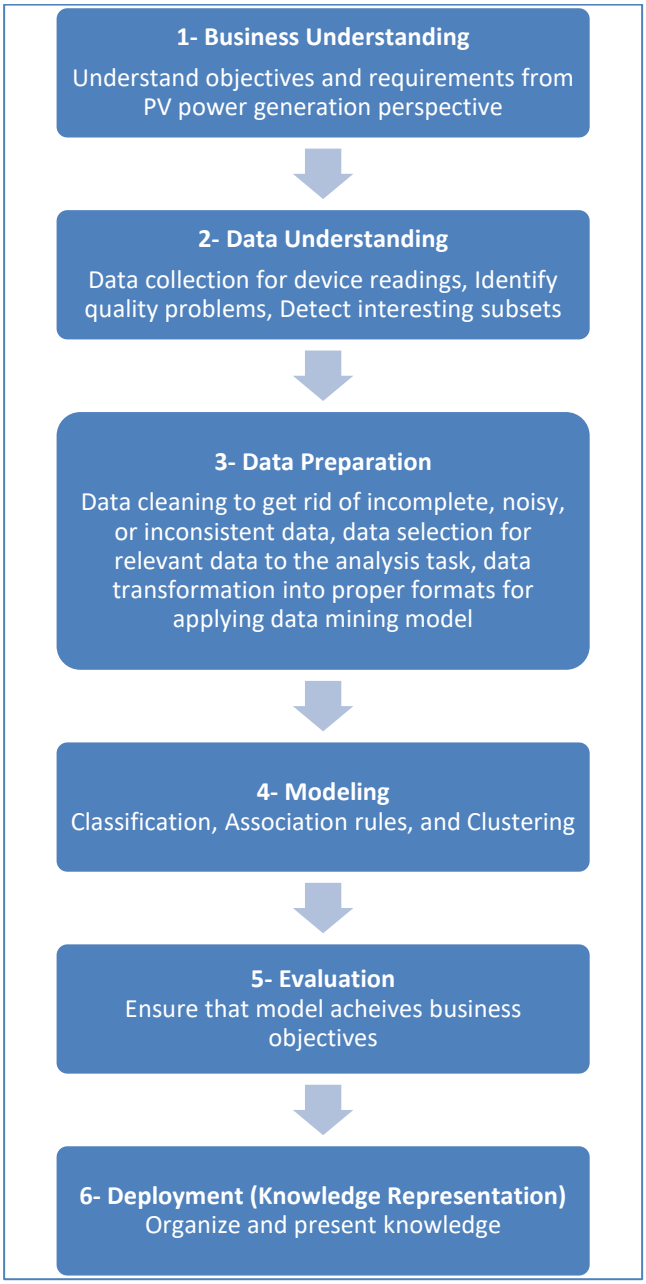

Figure 1 Data mining process steps and tasks

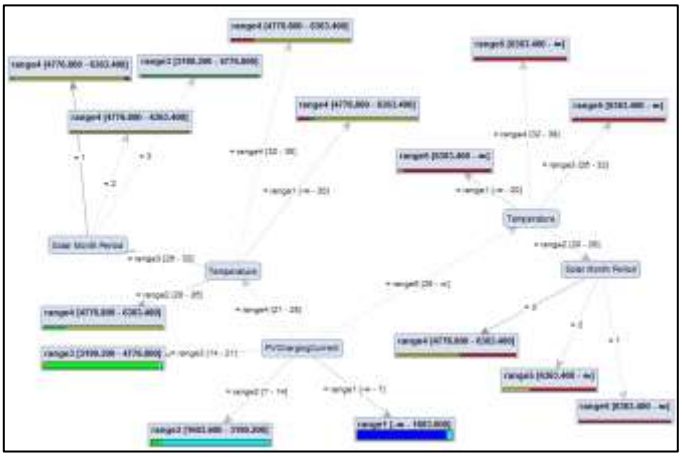

Figure 2 PVChargingCurrent Decision Tree

In case of selected attributes: moon month period, solar month period, temperature, PVpower (class label), and PVInputVoltage, we will get Figure 3 tree in which the PVInputVoltage is the root. 79.3% of range3 PVInputVoltage samples produced a PVpower value in range1 which means that voltage of solar cells current does not ensure high power like PVChargingCurrent. Moon month has more effect on PVInputVoltage where it appeared in tree.

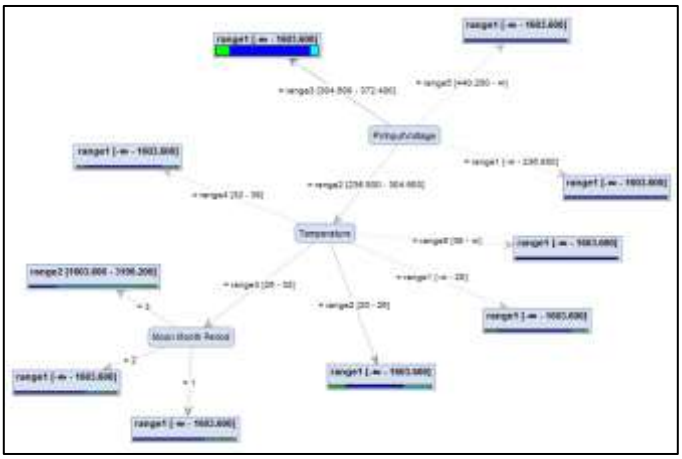

Figure 3 PVInputVoltage Decision Tree

Second task of data mining modeling phase was creating association rules by FP-Growth operator. This model is used to discover interesting relationships between disjoint items of dataset. Items can be defined as different values of elements. Relationship can be presented in a pattern of the form x -> y where x and y are item sets [8]. Association rules can be useful in the field of PV power generation systems by informing about those relations between elements and give useful knowledge about changeability relations between elements as a guidance of how to get more generated power. Figure 4 shows some resulting rules by association rules model with their evaluation factors support, confidence, and lift. Selected attributes for this model were Moon Month Period, PVInputVoltage, PVPower, Solar Month Period, Temperature, Time Period Num, and Week day.

Association rules were numerous, knowledge were extracted carefully according to background in that field and good consideration for data mining factors which are support, confidence, and lift. So association rules added that most likely high produced PVInputVoltage values can be in the time period between 12:00 and 13:00 in Gaza, despite of high PVInputVoltage the PVpower values will be low in case of temperature between 32 and 38, it's more likely to get lower PVpower values when it's a day between 11 and 20 of solar month (solar month period 2), Wednesday is strongly related to low PVpower values, and probability of low PVpower values can increase when a day fall in solar month period 2 and in moon month period 3 or a day falls in solar month period 3 and in moon month period 1.

Figure 4 Some Resulting Rules from Association Rules Model

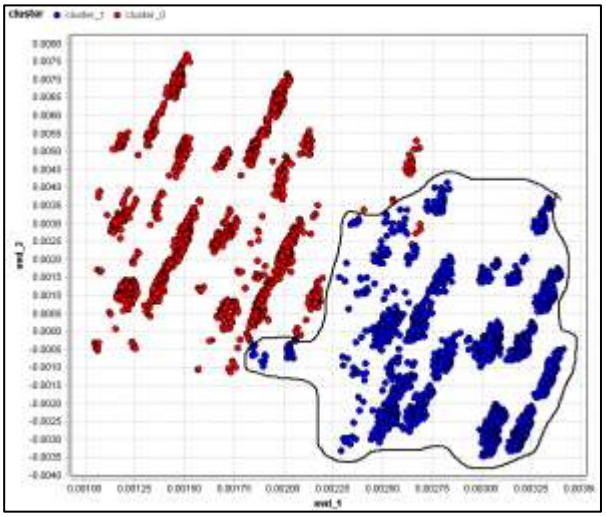

Figure 5 Resulting Clusters

Third task was clustering by k-Means operator. This model is used to group data instances with similar characteristics or features together into a separate cluster [8]. Clustering helps to categorize affecting situations on generating PV power so this gives an indication on how to deal with that situation. Figure 5 shows how points are distributed between clusters.

Clusters were distributed up to their counts as shown in Figure 6, It's obvious that cluster_1 is the majority (94308 items=75% of samples), and by studying centroid table of clustering model it was noticed that the cluster_0 has moderate and high values of PVpower and PVChargingCurrent, but cluster_1 has low values of PVpower and PVChargingCurrent.

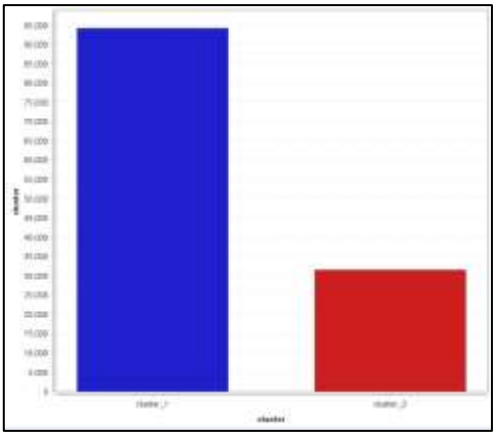

Figure 6 Clusters Counts

C. Evaluation

In order to estimate the statistical performance of a learning operator such as decision tree model, we use cross validation as an evaluation technique with ten validations. Cross validation defines a training set and testing set and runs classification model many times with changing testing set each time to reach to an unbiased success rate.

Accuracy of constructed decision tree model was 81.74%, Figure 7 shows performance vector.

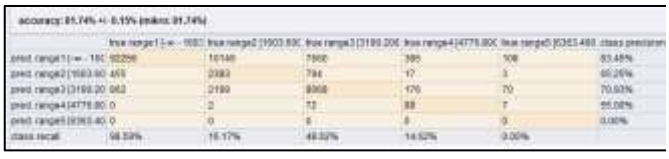

Figure 7. Performance Vector of Decision Tree Classification Model

4 CONCLUSION

PV solar power generation management systems can be exploited well if a good data mining process was applied on it with the guidance of specialists. This paper followed all data mining process steps to produce trusted results.

Main goal of data mining is to announce novel, actionable, understandable, and valid patterns. By applying association rules process and decision tree model a set of novel, actionable, and understandable knowledge was gotten.

Most important knowledge informed about that solar month periods affects PV charging current, but moon month periods affects PV input voltage. In addition, input voltage coming from solar cells does not ensure high power but PV charging current does. Also, highest produced PV input voltage values can be in the time period between 12:00 and 13:00 in Gaza. Furthermore, high temperatures produces low PV power values. Mostly PV power values decreases in the day of Wednesday, but can increase when day fall in solar month period 2 and in moon month period 3.

Future work can be considered if another datasets were available to re-apply models and get more accurate and actionable patterns, and to consider months as new attribute given a dataset of many years. Recommendation to PV management software developers is to add the functionalities of alerting about expected PV power generation to make decision makers take actions in the right way.

REFERENCES

1. M. Kraft, K. Desouza, and I. Androwich, "Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population," Proceedings of the 36th Hawaii International Conference on System Sciences, 2002.
2. A. Gosain, and A. Kumar, "Analysis of Health Care Data Using Different Data Mining Techniques," Proceedings of International Conference on Intelligent Agent & Multi-Agent Systems. IAMA 2009.
3. J. Wang, Y. Chen, R. Hua, P. Wang, and J. Fu, "A distributed big data storage and data mining framework for solar-generated electricity quantity forecasting," Proceedings of the SPIE, Volume 8333, article id. 83330S, pp. 7, 2011.
4. M. Hossain, A. Oo, and A. Ali, "Hybrid Prediction Method for Solar Power Using Different Computational Intelligence Algorithms," Smart Grid and Renewable Energy, vol. 4, pp. 76-87, 2013.
5. A. Qazi, H. Fayaz, and R. Raj, "Discourse on Data Mining Applications to Design Renewable Energy

Systems," Proceedings of International Conference on Advances in Engineering and Technology (ICAET'2014), Singapore, March 2014.

6. L. Martín, L. Zarzalejo, J. Polo, A. Navarro, R. Marchante, and M. Cony, "Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning," Solar Energy, Volume 84, Issue 10, October 2010, Pages 1772-1781, ISSN 0038-092X.
7. M. Almadhoun and A. El-Halees, "Different Mining Techniques for Health Care Data Case Study of Urine Analysis Test," Proceedings of Fifth International Conference on Intelligent Computing and Information Systems (ICICIS), Egypt, 2011.
8. J. Han, and M. Kamber, "Data Mining: Concepts and Techniques," the Morgan Kaufmann, 2001.
9. M. Lloyd-Williams, "Case Studies in the Data Mining Approach to Health Information Analysis," IEE Colloquium on Knowledge Discovery and Data Mining (1998/434), 1998.