Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 3, March 2015, pg.693 – 701

RESEARCH ARTICLE

Interesting Pattern Generation in Smart Electricity Meter Data Using Subgroup Discovery Algorithm

Dr.V.Karpagam[1], Punitha.S[2], Sathiya.S[3], Suvetha.S[4]

[1]Associate Professor, Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India
[2]Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India
[3]Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India
[4]Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India
[2] Punithamani12@gmail.com; [3] sathiyasivakumar35@gmail.com; [4] suvethajan14@gmail.com

Abstract: ***At present, data mining and knowledge discovery in electricity meter data suffer from insufficient focus on intelligent data analysis of subgroups whose patterns vary significantly from aggregate patterns present in an entire dataset, lack of effort towards generating intuitively understandable and practically applicable knowledge for industrial practitioners and limited knowledge regarding the link between unusual consumption patterns and household consumers' socio demographic characteristics. The proposed work overcomes these problems by using pattern recognition which helps to identify unusual consumption patterns within the dataset. Pattern recognition can be accurately implemented using subgroup discovery algorithms. This project work addresses these practically important but technically challenging issues by applying subgroup discovery algorithms to smart electricity meter dataset. K-means clustering is used to partition the dataset instance into usual and unusual consumption patterns. Subgroup discovery is applied on the unusual consumption patterns into discovering interesting relationships between different objects in the set with respect to a property which is of interest to the user the target variable. Subgroups are patterns which are extracted normally represented in the form of rules. Subgroups whose patterns are unusual and whose sizes are large enough are discovered, and their descriptive and predictive models are generated. The performance is evaluated quality measure using accuracy as a measures the support and confidence of the generated rule within the subgroups. The methodologies and algorithms presented for electricity dataset here, generic and applicable to a wider range of data mining problems where predictive analytics.***

Keywords---- Data mining, knowledge discovery, pattern recognition, and search space reduction.

I. INTRODUCTION

The knowledge and experience gained from smart- metered electricity system will enhance analysing and detecting abnormalities in electricity consumption data. This will aid in informed decision making in day-to-day operations, pricing and tariff design and outage management, and offer valuable ways to implement energy savings and demand response management. In the existing system, general patterns aggregated from yearly consumption patterns may lead to significant errors when the consumption patterns are statistically different from the general ones, most of the techniques use regression analysis which is not suitable for non-linear modelling, and artificial neural networks is suitable for non-linear modelling but it is computationally expensive [1]. The proposed system makes use of clustering techniques and subgroup discovery algorithm to identify unusual consumption patterns in subsets of consumers and associate it with consumer behaviour will overcome the existing system limitations.

A. Literature Review

1) Clustering: Meter readings of households are separated by clustering methods into heterogeneous daily consumption profiles, with an emphasis on their seasonal and temporal features using k-means algorithm [2]. Given historic hourly electricity consumption data, incremental summarization identifies daily and weekly patterns and updates them upon receiving new data so that to new actionable information is generated [3]. For example, learning from historic 24-hourly power loads over a year, a clustering-based fuzzy wavelet neural network has been used to predict the maximum and minimum load in the next year [4]. These works extract general and aggregated daily and yearly consumption patterns from the complete records in a dataset. A practical concern here is that these general patterns might lead to significant errors when they are applied to those subsets of consumers whose consumption patterns are statistically different from the general ones.

2) Pattern Recognition: The unusual pattern collected from clustered data is given as a input to Cortana, an open source software package. Cortana is a subgroup discovery tool. Typically return a large set of subgroups that satisfy the users inputs of options and parameters.

3) Associating Consumption Data with Types of Consumers: Apriori algorithm is used to select frequent item set. To find support and confidence choose the frequent item set. For example, using annual data on power, temperature, age, household size and income as features, together with household annual consumption data [5]. Popular data mining approaches used for pattern discovery in electricity consumption data include: **regression analysis** which estimates relationships between target and features [6], [7]. Artificial neural networks are more suitable for nonlinear modelling and complex datasets with unknown factors [5], [8], but they cannot produce the

underlying structure of the relations, so it is difficult to interpret the results, and furthermore the technique is computationally expensive. **Incremental summarization** [3] identifies the cyclic nature of electricity consumption patterns. These techniques are capable of extracting general and aggregated, typically daily and yearly consumption patterns from the complete records in a data set[9]. To classify electricity consumers according to their consumption patterns usually involves two steps: 1) partitions the dataset and group typical consumption curves, and2) then classify consumers [10], [11].

II. SUBGROUP DISCOVERY

Subgroup discovery is a branch of data mining concerned with the exploratory analysis of large data. Specifically, subgroup discovery methods aim to find particular regions in the dataset (known as subgroups) where the data shows an unusual distribution. Before starting the analysis, one needs to identify a specific variable of interest, known as the target. Subgroup discovery will then identify subgroups where the target for example is surprisingly high (or low) compared to what one would expect considering the entire dataset. Alternatively, when the target variable is categorical (e.g., "yes" and "no"), one would expect subgroups to contain significantly many occurrences of one of the categories. Typically, subgroups are identified by a set of conditions on variables other than the target, such that one can often think of them as (probabilistic) if-then rules, where the subgroup forms the if-part and the target the then-part [12]. There are four important components in subgroup discovery.

1) Subgroup Description Language: A subgroup is expressed as a conjunction of conditions and can be thought of as an induced rule. A condition is a test on a variable (or in data mining terms, an attribute), for example "" or"." The number of conditions, indicating the complexity of a subgroup, is 1 in these examples. The conjunction "AND" has two conditions joined by AND. A selector is a test on a variable. Two features "Location" and "Age" are tested here and the complexity of the subgroups so defined is two.

2) Target: This is a variable whose patterns interest users and on which a quality measure assesses subgroups.

3) Search Strategy: The simple brute-force search method exhaustively tests all combinations of selectors. It is practically prohibitive because when the number of selectors increases, the search space increases exponentially. On the other hand, the beam search method [13], in any iteration, only expands a fixed number of the most promising selectors or conjunctions of selectors found so far, by appending additional selectors. This fixed number is called beam width.

4) Quality measure: A quality measure evaluates the quality of candidate subgroups. Most quality measures are designed to obtain a balance between the two objective values.

1. Coverage that indicates the size, or the number of samples, of a subgroup. A large coverage is always desirable.

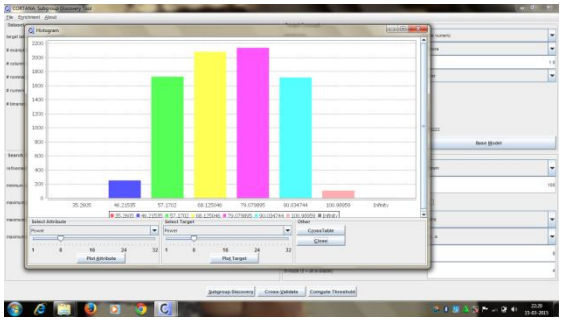

Fig. 1.Histogram

2. Unusualness that indicates whether the target patterns of a subgroup significantly differ from those of the entire dataset. Highly unusual patterns are desirable. These two objectives can be conflicting. When the coverage is large, it is often difficult to find unusual subgroups. When the coverage is small enough, it is often easy to find highly unusual subgroups. In most cases, the quality measures of subgroup discovery are trade-offs between unusualness and coverage. Those subgroups with a very low coverage would be in appropriately considered to be sufficiently "significant" or "representative". They are better treated as exceptional cases. A range of subgroup discovery methods exist for a variety of target types, including a single Boolean target [12], a multi-class target [14], and multiple targets [15]. For such targets, the following quality measures have been proposed [16].

1. Mean test: It is used to calculate average target value of a subgroup and is the average target value of the entire dataset, and is the coverage of a subgroup.

2. z-score: It is used to evaluate standard deviation of the target values of the dataset.

III. SMART METER DATA

A. Dataset

The smart electricity meter consumption data used in this paper were collected by the Energy Demand Research Project (EDRP) [17]. This dataset contains the electricity consumption in kWh of every one of nearly 8000 domestic consumer premises (households), with 30-min resolution, from April 2013 to April 2014. These consumption data were recorded by the installed smart meters at 24 time points per day: 0:00,1:00,….,24:00.

Table-I describes and explains these variables with examples. To scale this "socio-demographic enrichment" to practical population sizes is not technically difficult (various companies sell raw data at street and even address level), but serious issues of privacy and accuracy would need to be overcome. This research deals with these issues with data anonymization and cleansing steps. Practical projects will need to demonstrate the benefit of such analysis so it can be balanced against public concern.

Socio demographic Variables	**Description**	**Number of categories**	**Example(s)**
Power	Based on consumer usage	60-85	Consumers used 65.987
Temperature	Based on climate	19-25	e.g. 22
Household size	Total size of the house	4	4+bedrooms
Income	Consumers monthly income	25,000	25000-30000

TABLE-I

B. Consumption-Related Variables

Two consumption-related variables of wide industrial interest are chosen here to model consumption patterns:

1) Average Daily Load Factor ( ): The ratio between the averaged mean daily consumption and the averaged maximal daily consumption. Average daily load factor is shortened to load factor. A household with a large load factor suggests a constant consumption; conversely, a household with a small load factor suggests the maximum of the peaks is much higher than its daily average.

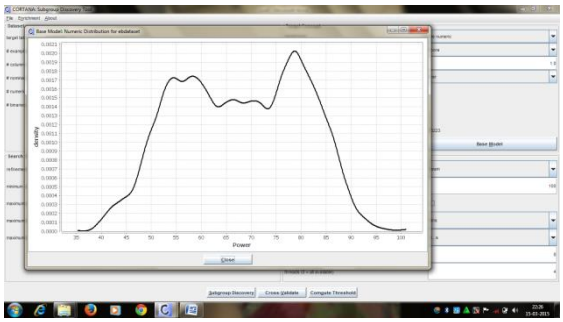

Fig. 2. Base model

2) Profile Error: It is widely used to evaluate the accuracy of profiles by calculating the error between profile estimate and actual consumption [4]. Given a household, these profiles return estimates in 30-minresolution over a year, taking into account environmental factors such as seasonal temperature changes, school and working hours and daylight hours. Actual readings for most households are manually recorded quarterly or half-yearly, making it impossible to accurately calculate profile errors.

The newly available dataset provides an opportunity to test the accuracy of the customer profiles currently in use. Percentage absolute profile error (PAPE) calculates the ratio of the profile error (the absolute difference between the actual measured consumption and the profile estimate) to the profile. The PAPE at a given half-hour indicates how household specific factors make impacts on their consumption. MeanPAPE averages the PAPE over a period of time.

Target concept **Target type**	**Example** **quality measures**
Single numeric	Meantest, t-test, z-score
Double regression	Slope difference
Double correlation	Coefficient, distance
Search condition	
Refinement depth	Maximal number of selectors
Search strategy	
Strategy type	Beam (width=100),best first, depth first
Numeric operation	<=,>=,=
Numeric strategy	All, bin, best
Number of bins	8

It gives a good indication of the profile error of a household. Volatility in smart meter data is due to multiple factors. They can be roughly divided into two classes: 1) more predictable variation across days (as factors such as temperature and daylight times vary); and 2) less predictable variation in customer behavioural patterns. Although the second class is more difficult to treat (i.e., to discover valid subgroups within) due to its randomness.

It is believed that because large sample sizes are used and many days worth of data are included in the calculations, there will be averaging effects such that subgroup discovery operates correctly.

IV. EXPERIMENTS

Cortana, an open source software package, is used for subgroup discovery in this paper. Subgroup discovery tools typically return a large set of subgroups that satisfy the users' inputs of options and parameters. Up to three resulting subgroups of each experiment are selected as examples to demonstrate the usefulness of this method.

There are various ways to perform significance tests on the statistical difference of a test statistic between the dataset and a subgroup, and to compare unusualness among subgroups.

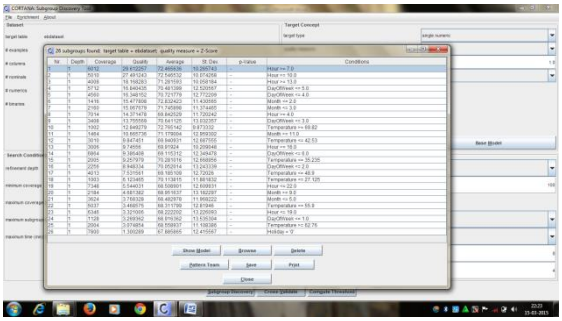

Fig. 3. Subgroup discovery

A) Regression Model as Target: The target of the experiment is a single variable. The concept of exceptional model mining expands subgroup discovery. So that a target can be a model with more than one variable[15], such as a linear regression model as demonstrated in the next experiment. Common sense tells us that property values are highly dependent on household incomes. Linear regression models are applied to model their relationship in the dataset.

B) Correlation Model as Target: The next experiment's target is a correlation model consisting of a socio-demographic variable, and meanPAPE which is a consumption-related variable.

V. COMPARATIVE STUDIES

The usefulness and effectiveness of subgroup discovery algorithms need to be empirically evaluated through their predictive power and classification accuracy, both being important to industrial practitioners.

A) Classification Accuracy: The classification accuracy [16], is compared with the dataset partitioned by subgroup discovery and by k-means clustering. Both subgroup discovery and clustering take socio-demographic variables. It was found that using subgroups to partition the dataset achieves higher classification accuracy. A subgroup's condition separates the dataset into one group satisfying this condition and its complement. It is not suitable for partitioning the dataset into more than two exclusive groups; for this purpose clustering is appropriate. A goal of this research was to search for unusual patterns, so it is worth comparing the patterns of meanPAPE in the

subgroups, in the two resulting exclusive clusters, and in the dataset. This comparison demonstrates the advantage of applying subgroup discovery over clustering, where the aim is to explore the data for subsets possessing distinct patterns which are not present or obscured in the dataset. Since clustering is a process of segmentation, it would require a larger number of clusters to produce a similar result. Another goal is to describe the features of households showing unusual patterns. As shown, subgroups describe this by induced rules. On the other hand, clustering only calculates the cluster centroids of variables.

VI. CONCLUSION

The proposed system of smart meters presents a considerable challenge to large-scale data analytics. This work explores subgroup discovery as a data mining method to analyse smart electricity meter data. The experiments found subgroups given various features and targets as inputs. This approach has constructed descriptive and interpretable models whose samples demonstrate statistically different patterns on target from the general patterns in the entire smart electricity meter data, and such models also cover statistically large number of samples. The comparative studies suggest that this approach outperforms more conventional data mining methods in terms of their predictive power and classification accuracy, while consuming similar computational resources.

The methodologies are generic and expected to be applicable to a wide range of problems for detecting unusual patterns, possibly including knowledge discovery in water and gas consumption data. It also hoped that the knowledge and experience gained will enhance analysing and detecting abnormalities in smart meter data, lead to informed decision making in day-to-day operation, pricing and tariff design and outage management, and offer valuable ways to implement energy savings and demand response management.

REFERENCES

[1] Nanlin Jin, Member, IEEE, Peter Flach, Tom Wilcox, Royston Sellman, Joshua Thumim, and Arno Knobbe, "Subgroup discovery in smart electricity meter data," IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 10, NO. 2, MAY 2014.

[2] A. M. Ferreira, C. A. Cavalcante, C. H. Fontes, and J. E. Marambio, "Anewmethod for pattern recognition in load profiles to support decision-making in the management of the electric sector," Int. J. Electr. Power Energy Syst., vol. 53, pp. 824–831, 2013.

[3] D. D. Silva, D. Yu, D. Alahakoon, and G. Holmes, "A data mining framework for electricity consumption analysis from meter data," IEEE Trans. Ind. Informat., vol. 7, no. 3, pp. 399–407, Aug. 2011.

[4] M. Amina, V. Kodogiannis, I. Petrounias, and D. Tomtsis, "A hybrid intelligent approach for the prediction of electricity consumption," Int. J. Electr.Power Energy Syst., vol. 43, no. 1, pp. 99–108, 2012.

[5] A. Azadeh and Z. Faiz, "A meta-heuristic framework for forecasting household electricity consumption," Appl. Soft Comput., vol. 11, no. 1, pp. 614–620, 2011.

[6] G. K. Tso and K. K. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, vol. 32, no. 9, pp. 1761–1768, 2007.

[7] A. Azadeh, O. Seraj, and M. Saberi, "An integrated fuzzy regression analysis of variance algorithm for improvement of electricity consumption estimation in uncertain environments," Int. J. Adv. Manuf. Technol., vol. 53, pp. 645–660, 2011.

[8] M. Khashei and M. Bijari, "An artificial neural network (p,d,q) model for time series forecasting," Expert Syst. Appl., vol. 37, no. 1, pp. 479–489, 2010.

[9] W. Labeeuw and G. Deconinck, "Residential electrical load model based on mixture model clustering and Markov models," IEEE Trans. Ind. Informat.,vol. 9, no. 3, pp. 1561–1569, Aug. 2013.

[10] M. Espinoza, C. Joye, R. Belmans, and B. DeMoor, "Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series," IEEE Trans. Power Syst., vol. 20, no. 3, pp. 1622–1630, Aug. 2005.

[11] N. Lavrac, B. Kavsek, P. Flach, and L. Todorovski, "Subgroup discovery with CN2-SD," J. Mach. Learn.Res., vol. 5, pp. 153–188, 2004.

[12] P. Clark and T. Niblett, "The CN2 induction algorithm," Mach. Learn.,vol. 3, no. 4, pp. 261–283, Mar. 1989.

[13] T. Abudawood and P. Flach, "Evaluation measures for multi-class subgroup discovery," in Eur. Conf. Mach. Learn.Princ.Pract.Knowl. Discovery Databases (ECML/PKDD'09), Sep. 2009, pp. 35–50.

[14] D. Leman, A. Feelders, and A. Knobbe, "Exceptional model mining," in Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases, Part II, ser. ECML PKDD '08, 2008, pp. 1–16.

[15] B. Pieters, A. Knobbe, and S. Dzeroski, "Subgroup discovery in ranked data, with an application to gene set enrichment," in Proc. Pref. Learn.Workshop (PL2010) at ECML PKDD, 2010.

[16] G. Raw and D. Ross, "Energy demand research project," Office of GasElect.Markets, Tech. Rep. 60163857, 2011 [Online]. Available: http://www.ofgem.gov.uk/Sustainability/EDRP/Pages/EDRP.aspx

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," SIGKDD Explor., vol. 11, no. 1, pp. 10–18, 2009.

