International Journal of Pure and Applied Mathematics
Volume 119 No. 14 2018, 1445-1451
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

ADAPTIVE HIGH-ORDER POSSIBILISTIC C-MEANS ALGORITHM USING HADOOP FRAMEWORK

M. Suresh[1], Dr. M. S. Anbarasi[2], A.Sarathkumar[3]
[1]Assistant Professor, IT, SMVEC, Pondicherry University, India.
[2] Assistant Professor, IT, PEC, Pondicherry, India.
[3] UG Students, IT, SMVEC, Pondicherry University, India.

ABSTRACT: Electricity is one of the major need for our daily lives. This project proposes to find an electricity consumption behavior pattern of customers using possibilistic c-means algorithm (PCM) which has been widely used in knowledge discovery. Initially PCM does not produce good results for clustering, especially for heterogeneous datasets, because it initially works with smaller datasets. To overcome this drawback a Higher order possibilistic c-means clustering based on map reduce algorithm is proposed in this project. This proposed method works with large amount of heterogeneous data sets and protect the private data without any special encryption schemes. By clustering the electrical consumption behavior will give out the overall electricity consumption for a year in different places and the predict the amount of electric power need for the future is also calculated. . Experimental results indicate that proposed method can effectively cluster a large number of heterogeneous data using map reduce.

INTRODUCTION

Data mining techniques are used to analyze a power consumption model in a city 's regional level and to capture knowledge about the use of electricity in relation to the weather temperature .Its geographical geographical features include river, agriculture, land and road. Based on the power consumption K-means clustering algorithm is used to create different groups of temperature and consumers. Association governance analysis creates associations rules regarding power consumption to describe the effect of physical geographical objects and physical distance in different areas. To analyze fuel consumption, the best analytical process is data mining, which is exploration analysis and preprocessing, often sample mining and data mining, such as associations, clustering and outlining deduction.

Electricity is a special object. Production and consumption should be equal each time. Availability of power consumption and quantity of production charts is essential. Electricity Consumption is based on Forecast Time Series analysis method. There is a definite future for different programs such as capital budgeting, sales forecasting, market research, financial planning, and inventory planning and control. Time series analysis techniques are useful for classification and assessment of time-based processes.

Larger data sets have a variety of objects, texts, images and audio, resulting in structured data and structured data forming high diversity. However, various objects have different information, but they are associated with each other. For example, a portion of the Sport video with Meta-Information uses a large number of further images to display a workout and uses some meta-communications, such as annotation and surrounding lessons to show additional information.

Power consumption depends on some trends such as temperature trends, dependence on daylight. Different types of data are collected from many users and they cluster separately as a group. The higher order clustering algorithm for large data is by using a tensor vector space to model correlations in various ways. Cluster big data efficiently, especially different data is harder. The problematic C-means algorithm is an important part of blur clustering. This reflects the distinction of each object to different clusters and can prevent corruption of the noise in the clustering process.

1.LITERATURE SURVEY

Load profiling, which refers consumers' power consumption behaviors to a particular period of time, eg, helps LSE understand how to use power for different users and how to use customer loading profiles or load samples. [4], nodal or customer scale load assessment, [5], the goal of demand response and fuel efficiency [6] and non-technical loss (NTL) 7 in tariff design of load profiling (to) Load profiling can be

classified into two groups: direct clustering and indirect clustering [8]. Direct clustering means that clustering methods are directly applicable for loading data. There are a number of clustering methods that are widely studied, including K-means [9], mess k-means [10], sequential clustering [11], self-management maps (SOM) [12], support vector, clustering [13], sub-clustering [14], ant colony clustering [15] and so on. Each clustering technique uses different criteria by assessing and computing the performance, Clustering Output Index (CDI), Scatter Index (SI), Davis-Boldin Index (DBI) and Average Index Integrity (MIA). [16].
Impacts of power consumption with a wider and high-frequency collection of electric current introduces great challenges for data storage, communication and analysis. In this case, the reduction methods are used effectively to reduce pre-load data before closing, defined as indirect clustering. Such clustering can be classified into two sub-categories, feature extraction-based clustering and time series based clustering. The feature extraction [17] is often used to reduce the level of input data to convert data into a low-volume space in high-volume space. Main component analysis (PCA) [18] [19] is a frequently used linear reduction method. It tries to keep most of the covariance of data attributes with the lowest artificial variables. Some vague size reduction techniques, such as Sammone maps, correlulator spatial analysis (CCA) [20] and deep practice [21], also apply to power consumption data. Moreover, power consumption data is particularly time series. Discrete Fourier transform (DFT) [22] [23], discrete wavelet transformations (DWT) [24], symbolic aggregate approximation (SAX) [25] and hidden markov model (HMM) 26] are discussed in the literature.

In addition to the Markov model, this work attempts to address the "data deluge" issue in three other ways. First, SAX is used to convert loading curves into a formal string, reduce storage space, and simplify communication traffic in smart meters and communications centers. Secondly, the recently reported effective clustering technique reported by Fast Search and Density Peaks (CFSFDP) uses the power consumption behaviors as a profile, which has less time-intensive complexity and weakness for sound points [28]. As measured by Kulbak-Leibler (K-L) distance [29], the differences between the two approaches are described in the dynamics of power consumption. Thirdly, to overcome the challenges of large and scattered data, CFSFDP technology is connected to a partitioning and concealing mechanism to further enhance the efficiency of data processing, where custom K is applicable for obtaining representative customers on local sites and the CFSFDP method performed on global sites. This technique can be further applied to larger data applications.

Machine Learning (ML) came with new data emerging as the Big Data emerged as a way to find value from data. ML platforms for Big Data began with disk-based systems such as disco-mute [7], which contained disk orientation from the underlying Hadoop architecture. As disk access is slow, new memory-based processes have been developed. Examples of Apache Spark and Oxdata H2O memory-driven platforms, and the Meehout Machine learning algorithms have been transformed into these platforms. Zhang et al. [15] Interviewed Big Data Management and Processing. They identified two types of in-memory systems: batch-based systems like spark and H2O, real time or drain processing systems such as storm. Systems of Energy Conservation Assessment are mainly of the Batch category.

Al-jarra et al. [6] Reviewed energy-efficient machine learning practices and new approaches with improved memory requirements. They saw local training as one of the key mechanisms of learning machine with Big Data because of the ability to reduce computational cost. As a promise to provide representative learning to complex problems, they considered a significant technology as a deep practice.

2. CLUSTERING OF ELECTRICITY CONSUMPTION BEHAVIOR

2.1 Fuzzy c-means algorithm

The algorithm works by allocating each data point to each cluster center based on distance between the cluster center and the data point. The algorithm provides the best result for the overlapping data set and by then the k-is the algorithm better. The data point is the result of each cluster center. The FCM program applies to various types of geostatistics data analysis problems. The program produces shading partitions and patterns for any data set. Fuzzy clustering is referred to as soft clustering.

Steps

The Fuzzy c-means algorithm is explained by the following steps,

a. Choose a number of clusters.
b. Assign coefficients randomly to each data point for being in the clusters.

2.2 Implementation

Stage 1

A user/application can submit a job to the Hadoop for required process by specifying the following items:
1. The location of the input and output files in the distributed file system.
2. The java classes in the form of jar file containing the implementation of map and reduce functions.
3. The job configuration by setting different parameters specific to the job.

Stage 2

The Hadoop Job Client then provides Job Tracker job (jar / executable etc) and formatting slaves, scheduling tasks and monitoring information, providing job-customer status and diagnostic information.

Stage 3

The Task Trackers on different nodes execute the task as per Map Reduce implementation and output of the reduce function is stored into the output files on the file system.

2.3 Higher order C-means clustering:

The most important steps of higher order C-means clustering is to calculate the membership matrix and the clustering centers. Therefore, we use the Map function to calculate the membership matrix and use the Reduce function to calculate the clustering centers. For updating the membership matrix, only the object xi and clustering centers V = {v1; v2; : : : ; vc} are required for calculating the membership values of the object xi towards each clustering center. Therefore, to reduce the storage of each computing node and the communication, we partition the membership matrix into p sub-matrices U = {U1;U2; : : : ; Up} by columns. The dataset X is also partitioned into p subsets X = {X1;X2; : : : ;Xp} accordingly. In the Map phrase, we dispatch each sub-matrix, the corresponding subset and all the clustering centers to one computing node for updating the membership matrix.

3. SYSTEM ARCHITECTURE

In this section, architecture of proposed system is explained. Fig.1 demonstrates the working of proposed system. To reduce the scale of large electricity consumption dataset using Symbolic Aggregate Approximation method. Clustering the electricity consumption behavior using k-means clustering. To obtain the typical dynamics of consumption behavior, with the difference between any two consumption patterns

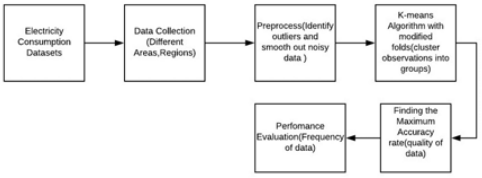


Fig.1: Proposed System Architecture

Step 1: The data's are collected for individual customers are handled separately. Divide the big data set into k parts, each marked as **Li**. Note that the data on one distributed site can be further partitioned to make the size of the data sets on each site more even.

Step 2: An adaptive k-means method is performed for each individual part to obtain a certain number of cluster centers. Each cluster center can represent all the objects belonging to this cluster with a small error. All these cluster centers of **Li** are selected as the representative objects **Mi**, which are defined as a local model.

Step 3: A modified Energy Consumption method is applied to all the representative objects (local models) that are centralized and gathered to classify them into several groups **R,** which are defined as a global model. Then, according to the final clustering result, the cluster label of each local site would be updated.

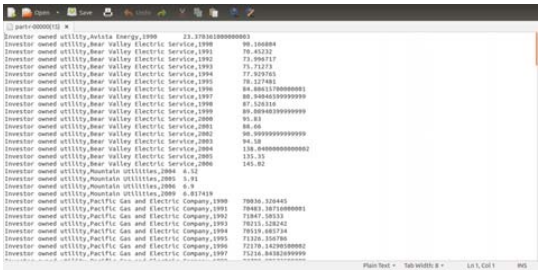
4. PERFOMANCE EVALUATION

Performance measurement can be used for all types of learning. This determines the type of task that it automatically learns and calculates the most common criteria for that type. For better performance calculation, use the above operators. Use performance (user-based) operator to allow performance actions if none of them are needed are.

a. Accuracy
b. Precision
c. Recall

4.1 Results

The Electricity Consumption analysis is carried out through Hadoop frame work and clustering of similar customers is carried out through the r- programming for visualization of similar customers.



The above values indicate the electricity consumption behavior pattern identified for every year by taking account with different customers.

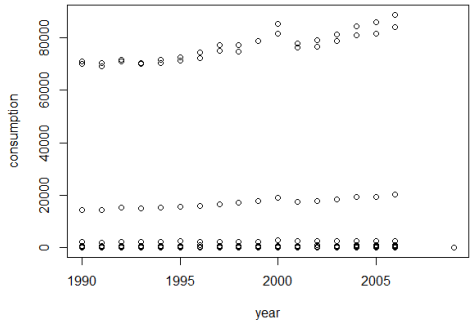


Fig 2.a Electricity Consumption Values

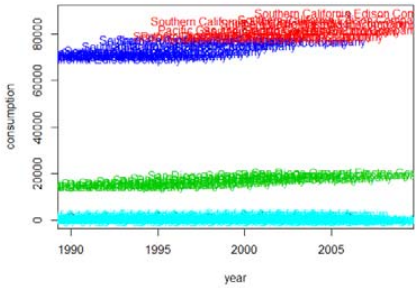


Fig 2.b Clustering of consumption values

In the Figure 2.a, the Graph plotted by the x-axis as Year, Y-axis as Consumption, the consumption values are imported over the year 1990-2005 and processed by k-means clustering method. In the Figure 2.b ,the figure represents the Overall Consumption value of 8000 Watts which is used in the regions of Southern California and San Diego are been shown in the graph. From the above presented graph it's evident that southern California and Bear Valley electric customers are consuming more than 80000 watts per year.

5. CONCLUSION

In this work, a novel approach has been proposed for dynamics clustering in the use of electrical power towards large data sets. A mess-c-clustering technique is performed to detect distinct dynamics in different groups of power consumption and segment consumer. Finally, each individual is a result of dynamic clustering to determine the potential for each user's consumer response to the consumer per year and the total energy consumption. The impact of external factors such as temperature, day type and economy in electricity consumption is limited by data sets and is not considered deep in the work. Future works focus on feature extraction and data mining techniques, combining power consumption with external factors.

REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data Mining with Big Data," IEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.
[2] B. Ermis, E. Acar, and A. T. Cemgil, "Link Prediction in Heterogeneous Data via

Generalized Coupled Tensor Factorization," Data Mining and Knowledge Discovery, vol. 29, no. 1, pp. 203-236, 2015.

[3] Q. Zhang, L. T. Yang, and Z. Chen, "Deep Computation Model for Unsupervised Feature Learning on Big Data," IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 161-171, Jan. 2016.

[4] N. Soni and A. Ganatra, "MOiD (Multiple Objects Incremental DBSCAN) - A Paradigm Shift in Incremental DBSCAN," International Journal of Computer Science and Information Security, vol. 14, no. 4, pp. 316-346, 2016.

[5] Z. Xie, S. Wang, and F. L. Chung, "An Enhanced Possibilistic c-Means Clustering Algorithm EPCM," Soft Computing, vol. 12, no. 6, pp. 593-611, 2008.

[6] Q. Zhang, C. Zhu, L. T. Yang, Z. Chen, L. Zhao, and P. Li, "An Incremental CFS Algorithm for Clustering Large Data in Industrial Internet of Things," IEEE Transactions on Industrial Informatics, 2015. DOI: 10.1109/TII.2017.2684807.

[7] X. Zhang, "Convex Discriminative Multitask Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 1, pp. 28-40, Jan. 2015.

[8] B. Gao, T. Liu, T. Qin, X. Zheng, Q. Cheng, and W. Ma, "Web Image Clustering by Consistent Utilization of Visual Features and Surrounding Texts," in Proceedings of the 13th Annual ACM International Conference on Multimedia, 2005, 112-121.

[9] Y. Chen, L. Wang, and M. Dong, "Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1459-1474, Oct. 2010.

[10] L. Meng, A. Tan, and D. Xu, "Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2293-2306, Aug. 2014.

[11] Q. Zhang, L. T. Yang, Z. Chen, and Feng Xia, "A High-Order Possibilistic-Means Algorithm for Clustering Incomplete Multimedia Data," IEEE Systems Journal, 2015, DOI: 10.1109/JSYST.2015.2423499.

[12] R. Krishnapuram and J. M. Keller, "A Possibilistic Approach to Clustering," IEEE Transactions on Fuzzy Systems, vol. 1, no. 2, pp. 98-110, May 1993.

[13] R. Krishnapuram and J. M. Keller, "The Possibilistic c-Means Algorithm: Insights and Recommendations," IEEE Transactions on Fuzzy Systems, vol. 4, no. 3, pp. 385-393, Aug. 1996.

[14] Q. Zhang and Z. Chen, "A Weighted Kernel Possibilistic c-Means Algorithm Based on Cloud Computing For Clustering Big Data," International Journal of Communication Systems, vol. 27, no. 9, pp. 1378-1391, 2014.

[15] Q. Zhang, L. T. Yang, and Z. Chen, "Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning," IEEE Transactions on Computers, vol. 65, no. 5, pp. 1351-1362, May 2016.

[16] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A Possibilistic Fuzzy c-Means Clustering Algorithm," IEEE Transactions on Fuzzy Systems, vol. 13, no. 4, pp. 517-530, Aug. 2005.

[17] M. Yang and C. Lai, "A Robust Automatic Merging Possibilistic Clustering Method," IEEE Transactions on Fuzzy Systems, vol. 19, no. 1, pp. 26-41, Feb. 2011.

[18] M. Filippone, F. Masulli, and S. Rovette, "Applying the Possibilistic c-Means Algorithm in Kernel-Induced Spaces," IEEE Transactions on FuzzySystems, vol. 18, no. 3, pp. 572-584, Jun. 2010.

A. Schneider, "Weighted Possibilistic c-Means Clustering Algorithms," in Proceedings of the 9th IEEE International Conference on Fuzzy Systems, 2000, pp. 176-180.

B. Liu, S. Xia, Y. Zhou, and X. Han, "A Sample-Weighted Possibilistic Fuzzy Clustering Algorithm," Acta Electronica Sinica, vol. 30, no. 2, pp. 371-375, 2012.

[19] R. Zhao and W. Grosky, "Narrowing the Semantic Gap Improved Text- Based Web Document Retrieval Using Visual Features," IEEE Transactions on Multimedia, vol. 4, no. 2, pp. 189-200, Jun. 2002.

[20] T. Jiang and A.-H. Tan, "Learning Image-Text Associations," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 2, pp. 161-177, Feb. 2009.

[21] M. Rege, M. Dong, and J. Hua, "Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering," in Proceedings of the 17th international conference onWorld Wide Web, 2008, pp. 317-326.

[22] X. Cai, F. Nie, H. Huang, and F. Kamangar, ”Heterogeneous Image Feature Integration via Multi-Modal Spectral Clustering,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1977-1984.

[23] B. Long, X. Wu, Z. Zhang, and P. Yu, ”Spectral Clustering for Multi-Type Relational Data,” in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 585-592.

[24] Q. Gu and J. Zhou, ”Co-Clustering on Manifolds,” in Proceedings ofthe 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 2009, pp. 359-367.

[25] R. Bekkerman, M. Sahami, and E. Learned-Miller, ”Combinatorial Markov Random Fields,” in Proceedings of the 17th European Conference on Machine Learning, 2006, pp. 30-41.

[26] L. Kuang, F. Hao, L. T. Yang, M. Lin, C. Luo, and G. Min, ”A Tensor-based Approach for Big Data Representation and Dimensionality Reduction,” IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 280-291, Sept. 2014.

[27] Y. Liu, Y. Liu, and K. Chan, ”Tensor Distance based Multilinear Locality- preserved Maximum Information Embedding,” IEEE Transactions on Neural Network, vol. 21, no. 11, pp. 1848-1854, Nov. 2010.

[28] J. Dean and S. Ghemawat, ”MapReduce: Simplified Data Processing on Large Clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

[29] J. Yuan and S. Yu, ”Privacy Preserving Back-propagation Neural Network Learning Made Practical with Cloud Computing,” IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 1, pp. 212-221, Jan. 2014.

[30] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, ”NUSWIDE: a Real-World Web Image Database from National University of Singapore,” in Proc. of A of ACM International Conference on Image and Video Retrieval, 2009, pp. 1-9.