Comparison of Subspace Projection Method with Traditional Clustering Algorithms for Clustering Electricity Consumption Data

Minghao Piao[1], Hyeon-Ah Park[1], Kyung-Ah Kim[2], Keun Ho Ryu[1]
[1]Database/Bio informatics Laboratory, Chungbuk National University, Cheongju, South Korea
[1]{bluemhp, hapark, khryu}@dblab.chungbuk.ac.kr
[2]Department of Biomedical Engineering, Chungbuk National University, Cheongju, South Korea
[2]{kimka}@chungbuk.ac.kr

***Abstract*—There are many studies about using traditional clustering algorithms like K-means, SOM and Two-Step algorithms to cluster electricity consumption data for definition of representative consumption patterns or for further classification and prediction work. However, these approaches are lack of scalability with high dimensions. Nevertheless, they are widely used, because algorithms for clustering high dimensional data sets are difficult to implement and it is hard to find open sources. In this paper, we adopt several subspace and projected clustering algorithms (subspace projection method) and apply them to the electricity consumption data. Our goal is to find the strength and weakness of these approaches by comparing the clustering results. We have found that traditional clustering algorithms are better to be used for load profiling by considering global properties and subspace or projected methods are better to be used for defining load shape factors by analyzing local properties without prior knowledge.**

Keywords-subspace projection; traditional clustering; K-menas; SOMs; Two-Step; local property; global propert.

I. INTRODUCTION

The knowledge of how and when consumers use electricity is essential to the competitive retail companies. This kind of knowledge can be found in historical data of the consumers collected in load research projects developed in many countries. One of the important tools defined in these projects are different consumers' classes are represented by its load profiles. Load profiling or classification to assign different consumers to the existing classes has been a matter of research during last years. Traditional clustering algorithms like K-means, SOM and Two-Step algorithms are widely used for load profiling [1-10]. However, these approaches just considered the global properties of consumption patterns and ignored local properties during the process even there are many algorithms [11-30] can be used to analyze local properties of such high dimensional data sets.

Traditional clustering algorithms consider all of the dimensions of an input dataset in order to learn as much as possible. In high dimensional data, however, many of the dimensions are often irrelevant. These irrelevant dimensions can confuse clustering algorithms by hiding clusters in noisy data. In very high dimensions, it is common for all of the objects in a dataset to be nearly equidistant from each other.

However, in practice, the data points could be drawn from multiple subspaces, and the membership of the data points to the subspaces might be unknown. For example, trajectory sequences could contain several moving objects, and different subspaces might describe the motion of different objects in the different routing and place. Therefore, there is a need to simultaneously cluster the data into multiple subspaces and find a low-dimensional subspace fitting each group of points. Subspace and projected clustering algorithms localize their search and are able to uncover clusters that exist in multiple, possibly overlapping subspaces.

Another reason that many clustering algorithms struggle with high dimensional data is the curse of dimensionality. As the number of dimensions in a dataset increases, distance measures become increasingly meaningless. Additional dimensions spread out the points until, in very high dimensions, they are almost equidistant from each other.

The final goal of our study is trying to compare the strength and weakness between subspace projection method and traditional clustering algorithms for its application to cluster electricity consumption data.

Remaining paper is organized as following: section 2 lists several studies about clustering application to electricity data; section 3 introduces several subspace projection clustering methods; section 4 describes our experimental result and we made a conclusion on section 5.

II. RELATED WORKS

In [1], an electricity consumer characterization framework based on a KDD (Knowledge Discovery from Data) process is presented. The framework is a data mining model composed of load profiling module and the classification module. The load profiling module's goal is the partition of the initial data sample in a set of classes defined according to the load shape of the representative load diagrams of each consumer. This is made by assigning to the same class consumers with the most similar behavior, and to different classes consumers with dissimilar behavior. The first step of the module development was the selection of the most suitable attributes to be used by the clustering model and the best results are obtained by SOM [33] method. In the second step, the K-means algorithm [34] is used to group the weight vectors of the SOM's units and the final clusters are obtained. The load profiles for each class are obtained by averaging the representative load diagrams of the consumers assigned to the same cluster. In the classification module, load shape indexes are derived from each representative load profiles and discretized by using interval equalization method, C5.0 is used to build the tree based and rule based classification models.

In [2, 3], the approach is focusing on identifying typical usage profiles for households and clustering them into a few archetypical profiles with similar kinds of customers grouped together. The work tests the applicability of applying the framework to UK specific data and identifies possible enhancements or modifications to the framework in order to better fit the UK situation. Clustering algorithms are used to derive domestic load profiles that have been successfully used in Portugal and applied it to UK data. The paper found that Self Organizing Maps in Portuguese work is not appropriate for the UK data.

In [4], a novel clustering model is presented, tailored for mining patterns from imprecise electric load time series that are represented by interval numbers. The model consists of three components. First, to guarantee the correctness when comparing two load time series, normalization techniques like Z-score [36] and Max-Min linear normalization [36] are used to handle the differences of baselines and scales. Second, it adopts a similarity metric that uses Interval Semantic Separation based measurement. Third, the similarity metric is used with the k-means clustering method to handle imprecise time series clustering. The model gives a unified way to solve imprecise time series clustering problem and it is applied in a real world application, to find similar consumption patterns in the electricity industry. Experimental results have demonstrated the applicability and correctness of the proposed model.

In [5], Clustering is used to generate groupings of data from a large dataset with the intention of representing the behavior of a system as accurately as possible. To be precise, two clustering techniques K-means and Expectation Maximization (EM) have been utilized for the analysis of the prices curve, demonstrating that the application of these techniques is effective so to split the whole year into different groups of days, according to their prices conduct. Silhouette function is used for selecting number of clusters for K-means and cross validation is used for selecting number of clusters for EM. K-means has been confirmed to be the algorithm more suitable for daily prices classification.

Gabaldón et al. [6] proposed proposed a methodology in order to obtain a better support management decisions in terms of planning of bids and energy offers in real-time energy markets. Specifically, Self-Organizing Maps (SOM) and Statistical Ward’s linkage to cluster electricity market prices into different groups. SOM and Ward’s clustering provide a similar clustering for the price series in this study.

In [7], a SOM development is presented to achieve the segmentation and demand patterns classification for electrical customers on the basis of database measurements. The objective of this paper is to review the capacity of some of them and specifically to test the ability of Self-Organizing Maps (SOMs) to filter, classify, and extract patterns from distributor, commercializer or customer electrical demand databases. Before the clustering, demand data from time domain was transformed into frequency domain and it shows an improvement in clustering performance.

In [8], various unsupervised clustering algorithms (modified follow-the-leader, hierarchical clustering, K-means, fuzzy K-means) and the Self-Organizing Maps was tested and compared to group customers with similar electrical behavior into one. Furthermore, this paper discussed and compared various techniques like Sammon map [37], principal component analysis (PCA) [38], and curvilinear component analysis (CCA) [40] which are able to reduce the size of the clustering input data set, in order to allow for storing a relatively small amount of data in the database of the distribution service provider for customer classification purposes. The results of the clustering validity assessment performed in this paper show that the modified follow-the-leader and the hierarchical clustering run with the average distance linkage criterion emerged as the most effective ones. Both algorithms are able to provide a highly detailed separation of the clusters, isolating load patterns with uncommon behavior and creating large groups containing the remaining load patterns. The other algorithms tend to distribute the load patterns among some groups formed during the clustering process and, as such, are less effective.

In [9, 10], for deriving the load profiles, they have used K-means algorithm. The data was measured using Automatic Meter Reading (AMR) senses by Korea Electric Power Research Institute. Before the load profiling, the data was normalized into the range of 0 to 1 and the optimal number of clusters is determined by using reproducibility evaluation method.

From the above previous studies, we can see that the tasks of clustering multiple time series stream or many individual time series (i.e., electricity consumption data) have received significant attention, and most papers are using traditional clustering algorithms like K-means, SOM and Two-Step algorithms. However, these methods are not suitable for cluster analysis of time series like electricity data since these approaches are lack of scalability with high dimensions. Nevertheless, they are widely used, because algorithms for clustering high dimensional data sets are difficult to implement and it is hard to find open sources.

III. SUBSAPCE PROJECTION CLUSTERING METHODS

Subspace projection (Subspace clustering or projected clustering) is extension of traditional clustering that seeks to find clusters in different subspaces within a dataset. Subspace clustering algorithms might report several clusters for the same object in different subspace projections, while projected clustering algorithms are restricted to disjoint sets of objects in different subspace. These subspace projections also can be identified into three major paradigms characterized by the underlying cluster definition and parametrization of the resulting clustering [30].

Cell-based approaches search for sets of fixed or variable grid cells containing more than a certain number of objects. Subspaces are considered as restrictions of a cell in a subset of the dimensions. Cell-based approaches rely on counting objects in cells and with their discretization of data. This is similar to frequent itemset mining approaches. The first algorithm for cell-based clustering was introduced by CLIQUE [11]. CLIQUE defines a cluster as a connection of grid cells. Grid cells are defined by a fixed grid splitting each dimension in equal width cells. It consists of three steps: (1) Identification of subspaces that contain clusters, (2) Identification of clusters, (3) Generation of minimal description for the clusters. Base on similarity between mining frequent itemsets and discovering relevant subspace for a given cluster, MINECLUS [12] algorithm adapted FP-growth and employed branch-and-bound techniques to reduce the search space. The quality of the results was further improved by (1) pruning small clusters of low quality, (2) merging clusters close to each other with similar subspaces, and (3) assigning points close to some cluster, else considered as outliers. SCHISM [13] uses the notions of support and Chernoff-Hoeffding bounds to prune the space, and use a depth-first search with backtracking to find maximal interesting subspaces.

Density-based clustering paradigm defines clusters as dense regions separated by sparse regions. As density computation is based on the distances between objects, distances are computed by taking only the relevant dimensions into account in subspace clustering. They can be parametrized by specifying which objects should be grouped together according to their similarities or distances. Density-based approaches are based on the traditional clustering paradigm proposed in DBSCAN [14]. The first approach in this area was an extension of the DBSCAN named SUBCLU [15]. It works in greedy manner by restricting the density computation to only the relevant dimensions. Using a monotonicity property, SUBCLU reduces the search space by pruning higher dimensional projections in a bottom up way, but it results in an inefficient computation. A more efficient solution is proposed by FIRES [16]. It is based on efficient filter-refinement architecture and consists of three steps: first step is the pre-clustering, all 1-D clusters called base clusters are computed; second step is the generation of subspace cluster approximations. The base clusters are merged to find maximal dimensional subspace cluster approximations; third step is post-processing of subspace clusters, refines the cluster approximations retrieved after second step. INSCY [17] is extension of SUBCLU, which eliminates redundant low dimensional clusters which detected already in higher dimensional projections. INSCY is depth-first approach, processes subspaces recursively and prunes low dimensional redundant subspace clusters. Thus, it achieves an efficient computation of density-based subspace clusters. As the maximal high dimensional projection is evaluated first, immediate pruning of all its redundant low dimensional projections leads to major efficiency gains. DOC [18] is a density based optimal projective clustering. It requires each cluster to be dense only in its corresponding subspace. Density conditions refer only to the number of points that project inside an interval of given length, and do not make any assumption on the distribution of points.

Clustering-oriented approaches do not give a cluster definition like the previous paradigms. In contrast to the previous paradigms, clustering-oriented approaches focus on the clustering result by directly specifying objective functions like the number of clusters to be detected or the average dimensionality of the clusters. PROCLUS [19] partitions the data into *k* clusters with average dimensionality l, extending K-Medoids approach, which is called CLARANS [20]. The general approach is to find the best set of medoids by a hill climbing process, but generalized to deal with projected clustering. The algorithm proceeds in three phases: an initialization phase, an iterative phase, and a cluster refinement phase. The purpose of the initialization phase is to reduce the set of points and trying to select representative points from each cluster in this set. The second phase represents the hill climbing process that in order to find a good set of medoids. Also, it computes a set of dimensions corresponding to each medoid so that the points assigned to the medoid optimally form a cluster in the subspace determined by those dimensions. Finally, cluster refinement phase, using one pass over the data in order to improve the quality of the clustering. More statistically oriented, P3C [21] is comprised of several steps. First, regions corresponding to projections of clusters onto single attributes are computed. Second, cluster cores are identified by spatial areas that (1) are described by a combination of the detected regions and (2) contain an unexpectedly large number of points. Third, cluster cores are refined into projected clusters, outliers are identified, and the relevant attributes for each cluster are determined. P3C does not need the number of projected clusters as input. It can discover the true number of projected clusters under very general conditions. It is effective in detecting very low-dimensional projected clusters embedded in high dimensional spaces. P3C is the first projected clustering algorithm for both numerical and categorical data. Defining a statistically significant density, STATPC aims at choosing the best non-redundant clustering [22]. It consists of following steps: Detecting relevant attributes around data points; refining candidate subspaces; detecting a locally optimal subspace cluster; constructing the good candidate subspace; greedy optimization, solving the optimization problem on candidate subsapce by testing all possible subsets is still computationally too expensive in general.

There are also other subspace projection methods like DiSH [23], SUBCAD [24], CLICKS [25], PreDeCon [26], MAFIA [27], DUSC [28], FINDIT [29].

IV. Experiments Results and Discussion

The given test data set is the customers' power consumption data obtained from Korea Electric Power Research Institute (KEPRI). It has 165 instances with 25 attributes (one attribute is class label-contract types), and it is restructured as daily representative vectors:

$$V^{(c)} = \{V_0^C, ..., V_h^C, ..., V_H^C\}, \text{c = customer}, 0 < \text{h} < 24, \text{H} = 24 \qquad (1)$$

Where for each customer *c*, let $V^{(c)}$ denotes the total daily power usage of *c* in one day for 24 hours. The power usage is measured per 1 hour; therefore, each total daily power usages data has 24 dimensions, and dimension names are noted as H1, H2, H3,…, H14, H15,…, H23, H24, e.g., H14 means the measured time 14:00 PM.

Subspace projection algorithms are used with default parameter settings [10] except *Proclus*, the number of clusters of which is set to 11, since there were 11 different contract types in the test data set. From Tables 1 to 3, we can see that when using K-means, SOM and Two-Step to cluster the given data, SOM has achieved better result than K-means and Two-Step since K-means and Two-Step have assigned most number of instances into one cluster. At least, traditional clustering approaches, they have assigned all instances into the clusters. In contrast, for example, FIRES is used with default parameter settings and it resulted in 13 clusters where 92 un-clustered instances are. It means traditional algorithms are able to be used for load profiling to find representative consumption patterns as previous studies while subspace and projected clustering algorithms are not. The reason is that traditional clustering algorithms consider all dimensions which present the global properties of customers' consumption patterns. The given contract types are determined by consumer's consumption activities which characterize the global shapes (global property) of these consumption patterns. Since we have used these contract types as given class information, and subspace projection methods use subspace or projections of whole dimensions instead of all, the result of traditional approaches are better than subspace and projection methods.

TABLE I. CONFUSION MATRIX OF SOM

	Found Clusters										
Contract Types	1	2	3	4	5	6	7	8	9	10	11
1	0	1	0	0	0	0	0	3	0	0	0
2	4	3	1	22	2	0	0	56	2	0	0
3	0	0	0	0	0	0	0	7	0	0	0
4	0	0	0	0	0	0	0	0	0	0	2
5	0	1	2	5	1	0	1	6	0	3	0
6	0	0	1	0	0	1	0	1	0	0	0
7	0	0	0	0	1	0	1	3	2	1	1
8	0	1	0	1	0	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0	0
10	5	1	1	3	0	0	0	0	0	0	0
11	0	0	1	0	0	0	0	17	0	0	0

TABLE II. CONFUSION MATRIX OF K-MEANS

	Found Clusters										
Contract Types	1	2	3	4	5	6	7	8	9	10	11
1	3	0	1	0	0	0	0	0	0	0	0
2	88	0	2	0	0	0	0	0	0	0	0
3	7	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	1	0	0	0	0	0	0
5	12	1	3	0	0	1	1	0	0	0	1
6	1	0	1	0	0	0	0	0	0	1	0
7	6	0	0	0	0	1	0	0	1	0	1
8	1	0	1	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	0
10	9	0	0	0	0	0	0	1	0	0	0
11	17	0	0	0	0	0	0	1	0	0	0

TABLE III. CONFUSION MATRIX OF TWO-STEP

	Found Clusters										
Contract Types	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	0	0	0	0	0	0	3	1
2	1	0	0	0	0	0	0	0	0	88	1
3	0	0	0	0	0	0	0	0	0	7	0
4	0	0	0	0	0	0	0	1	1	0	0
5	2	0	0	2	1	1	0	0	0	11	2
6	0	1	0	0	0	0	0	0	0	1	1
7	0	0	0	0	1	1	0	0	1	6	0
8	0	0	0	0	0	0	0	0	0	1	1
9	0	0	0	0	0	0	0	0	0	1	0
10	0	0	1	0	0	0	0	0	0	9	0
11	0	0	1	0	0	0	0	0	0	17	0

Table 4 shows the dimensions relevant to definition of clusters. These relevant dimensions are describing the local properties of electricity consumption data and it is useful to define load shape factors [31, 32, 33]: for defining load shape factors, we have to find time intervals first which shows big difference of electricity usage during the time intervals. In previous studies, load shape factors are defined by experts according to their experience. From Table 4, we can see dimensions in each cluster which maximize the difference between itself with others. Definition of load shape factors can be done by considering time intervals of these neighboring relevant dimensions. For example, load shape factor for cluster-0 should consider time interval of 01:00 AM ~ 09:00 AM, and for cluster-11, have to consider time interval of 18:00 PM ~ 19:00 PM and 22:00 PM~23:00 PM. There will be some load shape factors overlaps in same time intervals. This problem will not cause serious problems since we can assign weight for each shape factors to cover it.

TABLE IV. DIMENSIONS RELATED TO DEFINITION OF CLUSTERS (DM:DIMENSIONS)

DM	**Cluster Numbers**												
	0	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1											
2	1		1								1		
3	1	1											
4	1		1										
5	1		1										
6	1		1		1								
7	1				1						1		
8	1					1							
9	1				1								
10				1		1	1						
11				1		1	1						1
12				1			1						
13				1		1	1						
14				1		1	1	1	1		1		
15				1		1	1		1		1		
16				1		1	1		1				
17				1			1		1		1		
18				1			1		1		1	1	
19				1			1	1			1	1	1
20				1			1						1
21				1			1	1					
22				1		1	1	1	1			1	
23				1				1		1		1	
24								1			1		1

From Figure 1, we can see some dimensions which is mostly related to the local property of the load usage patterns: often appeared dimensions tend to have most significant discriminating power to differentiate clusters. Therefore, suppose the given count threshold is 4 then dimensions {*11*, *14*, *15*, *16*, *17*, *18*, *19*, *22*, *23*} will be the most useful dimensions to define load shape factors. Combinations of these dimensions also can be most useful time intervals to derive load shape factor if and only the appearance of them are high enough and they are neighboring. Furthermore, we can use some relevance evaluation techniques to evaluate the defined load shape factors.

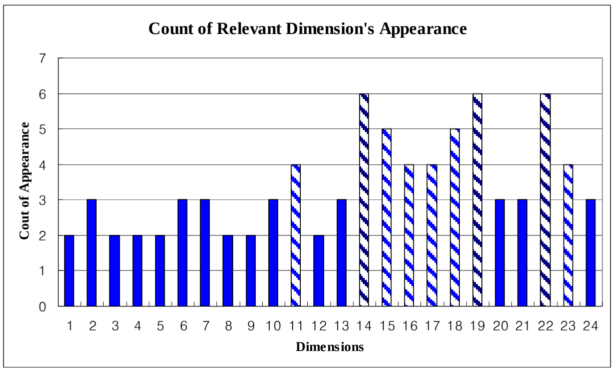

Figure 1. Count of relevant dimension’s appearance from table 3.

V. CONCLUSION

In this study, we have used small data sets consists of 165 instances with 24 dimensions. The main objective of clustering is to find high quality clusters within a reasonable time. However, we found that several subspace and projected clustering algorithms like P3C, CLIQUE and DOC took too much time to get in the result.

Also, we found that traditional clustering algorithms like K-means, SOM and Two-Step are better than subspace and projected clustering approaches when applied to define representative consumption patterns even they are lack of scalability with high dimensions - some instances are not assigned to the clusters when using subspace and projected clustering algorithms. However, relevant dimensions used to define clusters in subspace and projected approaches are able to be used for extracting load shape factors for classifications works by analyzing local and global properties of electricity consumption data set without prior knowledge.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0000478).

REFERENCES

[1] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, “An electric energy consumer characterization framework based on data mining techniques,” IEEE Transactions on Power Systems, vol. 20, May 2005, pp. 596-602, doi:10.1109/TPWRS.2005.846234.

[2] I. Dent, U. Aickelin, and T. Rodden, “The Application of a Data Mining Framework to Energy Usage Profiling in Domestic Residences using UK data,” Proc. Research Student Conference on “Buildings Do Not Use Energy, People Do?”, June 2011.

[3] I. Dent, U. Aickelin, and T. Rodden, “Application of a clustering framework to UK domestic electricity data,” Proc. The 11th Annual Workshop on Computational Intelligence, 2011, pp. 161-166.

[4] Q. D. Li, S. S. Liao, and D. D. Li, “A Clustering Model for Mining Consumption Patterns from Imprecise Electric Load Time Series Data,” Proc. Fuzzy Systems and Knowledge Discovery, Lecture Notes in Computer Science, vol. 4223, 2006, pp. 1217-1220, doi: 10.1007/11881599_152.

[5] F. Martínez-Álvarez, A. Troncoso, J.C. Riquelme, and J.M. Riquelme, “Partitioning-clustering techniques applied to the electricity price time series,” Proc. The 8th international conference on Intelligent data engineering and automated learning, 2007, pp. 990-999, 2007, doi: 10.1007/978-3-540-77226-2_99.

[6] A. Gabaldón, A. Guillamón, M.C. Ruiz, S. Valero, C. Álvarez, M. Ortiz, and C. Senabre, “Development of a methodology for clustering electricity-price series to improve customer response initiatives,” IET Generation, Transmission & Distribution, vol. 4, June 2010, pp. 706-715, doi:10.1049/iet-gtd.2009.0112.

[7] S. V. Verdu, M. O. Garcia, C. Senabre, A. G. Marin, and F. J. Garcia Franco, “Classification, filtering and identification of electrical customer load patterns through the use of SOM maps,” IEEE Transactions on Power Systems, vol. 21, no. 4, Nov. 2006, pp. 1672-1682, doi:10.1109/TPWRS.2006.881133.

[8] G. Chicco, R. Napoli, and F. Piglione, “Comparisons among Clustering Techniques for Electricity Customer Classification,” IEEE Transactions on Power Systems, vol. 21, no. 2, May 2006, pp. 933-940, doi:10.1109/TPWRS.2006.873122.

[9] M. H. Piao, H. G. Lee, J. H. Park, and K. H. Ryu, “Application of Classification Methods for Forecasting Mid-Term Power Load Patterns,” Proc. Communications in Computer and Information Science, vol. 15, 2008, pp. 47-54, doi:10.1007/978-3-540-85930-7_7.

[10] J. H. Shin, B. J. Yi, Y. I. Kim, H. G. Lee, and K. H. Ryu, “Spatiotemporal Load-Analysis Model for Electric Power Distribution Facilities Using Consumer Meter-Reading Data,” IEEE Transactions on Power Delivery, vol. 26, no. 2, April 2011, pp. 736-743, doi: 10.1109/TPWRD.2010.2091973.

[11] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications,” Proc. ACM SIGMOD international conference on Management of data, vol. 27, no. 2, June 1998, pp. 94-105, doi: 10.1145/276304.276314.

[12] M. L. Yiu and N. Mamoulis, “Frequent-pattern based iterative projected clustering,” IEEE International Conference on Data Mining, Nov. 2003, pp. 689-692, doi:10.1109/ICDM.2003.1251009.

[13] K. Sequeira and M. Zaki, “SCHISM: A new approach for interesting subspace mining,” International Journal of Business Intelligence and Data Mining, Dec. 2005, pp. 137-160, doi: 10.1504/IJBIDM.2005.008360.

[14] M. Ester, H. P. Kriegel, J. Sander, and X. Xu., “A density-based algorithm for discovering clusters in large spatial databases,” Proc. The 2nd International Conference on Knowledge Discovery and Data Mining, 1996. pp. 226-231.

[15] K. Kailing, H.-P. Kriegel, and P. Kroger, “Density-connected subspace clustering for high-dimensional data,” Proc. The 4th SIAM Conference on Data Mining, April 2004, pp. 246-257.

[16] H. P. Kriegel, P. Kroger, M. Renz, and S. Wurst, “A generic framework for efficient subspace clustering of high-dimensional data,” Proc. The 5th International Conference on Data Mining, Nov. 2005, pp. 250-257, doi:10.1109/ICDM.2005.5.

[17] I. Assent, R. Krieger, E. Muller, and T. Seidl, “INSCY: Indexing subspace clusters with in-process-removal of redundancy,” Proc. IEEE International Conference on Data Mining, Dec. 2008, pp. 719-724, doi:10.1109/ICDM.2008.46.

[18] C. M. Procopiuc, M. Jones, P. K. Agarwal, and T. M. Murali, “A Monte Carlo algorithm for fast projective clustering,” Proc. ACM SIGMOD International Conference on Management of Data, 2002, pp. 418-427, doi: 10.1145/564691.564739.

[19] C. Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park, “Fast algorithms for projected clustering,” Proc. ACM SIGMOD international conference on Management of data, June 1999, pp. 61-72, doi: 10.1145/304182.304188.

[20] R. T. Ng and J. W. Han, “Efficient and Effective Clustering Methods for Spatial Data Mining,” Proc. The 20th International Conference on Very Large Data Bases, 1994, pp. 144-155.

[21] G. Moise, J. Sander, and M. Ester, “P3C: A robust projected clustering algorithm,” Proc. IEEE International Conference on Data Mining, Dec. 2006, pp. 414-425, doi:10.1109/ICDM.2006.123.

[22] G. Moise and J. Sander, “Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering,” Proc. The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 533-541, doi:10.1145/1401890.1401956.

[23] E. Achtert, C. BÄohm, H. P. Kriegel, P. KrÄoger, I. MÄuller-Gorman, and A. Zimek, “Detection and visualization of subspace clusters hierarchies,” Proc. The 12th International Conference on Database systems for advanced applications, 2007, pp. 152-163, doi: 10.1007/978-3-540-71703-4_15.

[24] G. Gan and J. Wu, “Subspace clustering for high dimensional categorical data,” ACM SIGKDD Explorations Newsletter, vol. 6, Dec. 2004, pp. 87-94, doi:10.1145/1046456.1046468.

[25] M. Zaki, M. Peters, I. Assent, and T. Seidl, “CLICKS: an effective algorithm for mining subspace clusters in categorical datasets,” Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005, pp. 733-742, doi: 10.1145/1081870.1081965.

[26] C. Bohm, K. Kailing, H.-P. Kriegel, and P. Kroger, “Density Connected Clustering with Local Subspace Preferences,” Proc. IEEE International Conference on Data Mining, Nov. 2004, pp. 27-34, doi: 10.1109/ICDM.2004.10087.

[27] H. Nagesh, S. Goil, and A. Choudhary, “Adaptive grids for clustering massive data sets,” Proc. The 1st SIAM International Conference on Data Mining, 2001, pp. 1-17.

[28] I. Assent, M. Krieger, E.Muller, and T. Seidl, “DUSC: Dimensionality Unbiased Subspace Clustering,” Proc. IEEE International Conference on Data Mining, 2007, pp. 409-414, doi: 10.1109/ICDM.2007.49.

[29] K. G. Woo, J. H. Lee, M. H. Kim, and Y. J. Lee, “FINDIT: a Fast and Intelligent Subspace Clustering Algorithm using Dimension Voting,” Information and Software Technology, vol. 46, no. 4, 20045, pp.255-271, doi:10.1016/j.infsof.2003.07.003.

[30] E. Müller, S. Günnemann, I. Assent, and T. Seidl, “Evaluating Clustering in Subspace Projections of High Dimensional Data,” Proc. The 35th International Conference on Very Large Data Bases (VLDB 2009), vol. 2, Aug. 2009, pp. 1270-1281.

[31] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, “Electric energy customer characterization for developing dedicate market strategies”, Proc. IEEE Porto Power Tech conference, Sep. 2001, doi:10.1109/PTC.2001.964627.

[32] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, “Customer characterisation options for improving the tariff offer”, IEEE Transactions on Power Systems, vol. 18, no. 1, Feb. 2003, pp. 381-387, doi:10.1109/TPWRS.2002.807085.

[33] J. B. Lee, M. H. Piao, and K. H. Ryu, “Incremental Emerging Patterns Mining for Identifying Safe and Non-safe Power Load Lines”, Proc. IEEE International Conference on Computer and Information Technology, 2010, pp. 1424-1429, doi: 10.1109/CIT.2010.255.

[34] T. Kohonen, T. S. Huang, and M. R. Schroeder, Self-Organizing Maps, Springer-Verlag, December 2000.

[35] J. B. MacQueen, “Some Methods for classification and Analysis of Multivariate Observations,” Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297.

[36] J. W. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques. Third Edition, The Morgan Kaufmann Series in Data Management Systems, July 6, 2011.

[37] J. W. Sammon, “A nonlinear mapping for data structure analysis,”IEEE Transactions on Computers, vol. C-18, no. 5, May 1969, pp. 401-409, doi:10.1109/T-C.1969.222678.

[38] J. E. Jackson, A User’s Guide to Principal Components. New York: Wiley, 1991, pp. 1-25.

[39] P. Demartines and J. Herault, “Curvilinear component analysis: A selforganizing neural network for nonlinear mapping of data sets,” IEEE Transactions on Neural Networks, vol. 8, no. 1, Jan. 1997, pp. 148-154, doi:10.1109/72.554199.