Unsupervised Clustering based on Feature-value / Instance Transposition Selection

Akira Kusaba
Research Institute for Applied Mechanics
Kyushu University
Fukuoka, Japan
kusaba@riam.kyushu-u.ac.jp

Takako Hashimoto
Commerce and Economics
Chiba University of Commerce)
Chiba, Japan
takako@cuc.ac.jp

Kilho Shin
Computer Centre
Gakushuin University
Tokyo, Japan
yoshihiro.shin@gakushuin.ac.jp

David Lawrence Shepard
Evidation Health
Los Angeles, CA, USA
shepard.david@gmail.com

Tetsuji Kuboyama
Computer Centre
Gakushuin University
Tokyo, Japan
ori-tencon2020@tk.cc.gakushuin.ac.jp

***Abstract*—This paper presents FITS, or Feature-value / Instance Transposition Selection, a method for unsupervised clustering. FITS is a tractable, explicable clustering method, which leverages the unsupervised feature value selection algorithm known as UFVS in the literature. FITS combines repeated rounds of UFVS with alternating steps of matrix transposition to produce a set of homogenous clusters that describe data well. By repeatedly swapping the role of feature and instance and applying the same selection process to them, FITS leverages UFVS's speed and can perform clustering in our experiments in tens milliseconds for datasets of thousands of features and thousands of instances.**

We performed feature selection-based clustering on two real-world data sets. One is aimed at topic extraction from Twitter data, and the other is aimed at gaining awareness of energy conservation from time-series power consumption data. This study also proposes a novel method based on iterative feature extraction and transposition. The effectiveness of this method is shown in an application of Twitter data analysis. On the other hand, a more straightforward use of feature selection is adopted in the application of time series power consumption data analysis.

***Index Terms*—feature selection, clustering, twitter data, time-series data**

I. Introduction

Unsupervised clustering is one of the hardest problems in machine learning. The difficulty arises from two fundamental challenges, tractability and explainability. Clustering problems can be *intractable* because a dataset can be described by a union of feature sets, each of which consists of unrelated observations of the same instances, which may have n potential combinations, each of which may fall into m clusters, yielding a theoretical maximum of m^n potential clusterings.

The number of instances can be significantly less than the sufficient number for effective clustering since the exponential function m^n rapidly exceeds the size of practical datasets as n increases.

Clustering problems can also be *inexplainable* because the emergent behavior of some algorithms can make identifying the features that determine a particular clustering difficult, as when a set of features is substantially different in particular columns, and the clustering grows to encompass apparently-unrelated instances. In supervised learning, these problems matter less because class labels establish a relationship between features and instances, which guarantees explicability. The guarantee of explicability in turn places a limit on the possibility of completely independent feature sets, which guarantees the problems will be tractable [6]. In unsupervised learning, these problems affect every approach.

Preemptive feature selection, on the other hand, can solve both tractability and explainability problems. An effective feature selection algorithm can pare a feature set down to a tractable subset, which reduces the work of the clsutering algorithm. A reduced feature subset is usually more explainable. Moreover, a *feature value selection algorithm* can yield better results because feature values explain clusters better than features [6].

This paper presents FITS (Feature-value/Instance Transposition Selection), a novel method that solves the tractability and explainability problems by incorporating a new idea into the process of feature selection.

The idea originates from feature/instance transposition, or swapping the role of "feature" and "instance" to refine the dataset. It is common to view a dataset as a matrix, with columns corresponding to features and rows as instances. Transposing a matrix, however, swaps the role of "feature" and "instance." Hence, applying feature selection to the transposed matrix (dataset) results in selection of instances.

This method, however, will not work effectively because of the lack of symmetry between features and instances: For a single feature, the values across instances vary, but within the same scale; For a single instance, by contrast, values across features may belong to unrelated scales.

One way to avoid this problem is to convert features into feature values by one-hot encoding. This binarization converts all values to numbers in $\mathbb{F}_2 = \{0, 1\}$. Turning values into a dimensionless quantity establishes a symmetry between feature values and instances.

By iterating feature value selection and matrix transposes, we refine the feature values selected into values that categorize the instances into consistent, explainable clusters.

Because our method requires many iterations of feature value selection, we use UFVS [6], which requires only a few tens of milliseconds to process datasets with thousands of features and instances on a typical laptop PC.

This paper demonstrates the effectiveness of our method by applying it to two real-world problems. First, we extract topics from Twitter data. Twitter is characterized by very short texts in an informal style, and existing topic modeling methods such as LDA and matrix decomposition often perform poorly. Our method, on the other hand, extracts high quality topics with meaningful features. The second example is discovering patterns from classroom power consumption sensor data on a university campus. In this paper, we present these two cases as instances in which FITS extracts meaningful information from real-world datasets.

This paper is organized as follows. Section II introduces related work on feature (value) selection and real world data clustering. Section III describes our proposed method. Section IV demonstrates experimental results of our method. Finally, Section V offers directions for future research.

II. Related Work

A. Feature Selection and Feature Value Selection

A number of practical feature selection algorithms for supervised learning have been proposed in the literature [3]–[5], [7].

In many of them, feature selection is described as a process of iteratively eliminating features irrelevant to class labels and features mutually redundant [4] until reaching a sufficiently small set of features. In [7], it is claimed that interacting relevant features, which are individually irrelevant but relevant to class labels as a group, should be incorporated into selection. Under the combined definition, we note that, in supervised learning, class labels play a crucial role in guiding feature selection.

Unsupervised learning, in contrast, is performed without class labels. Therefore, for unsupervised feature selection, we need an alternative principle that replaces class labels. Some unsupervised feature selection algorithms in the literature leverage pseudo-labels as a substitute for class labels [11]–[13], which are determined prior to feature selection using some known clustering algorithms. Some other methods select features to preserve intrinsic structures of data such as manifold structures [1], [10], [14] and data-specific structures [8], [9]. One of the problems of these approaches is low time efficiency. In fact, generating pseudo-labels and finding underlying structures require heavy computation such as decomposing huge matrices.

UFVS [6], on the other hand, selects feature values rather than features under the constraint that the percentage of instances that can be explained by the selected feature values must not be lower than a predetermined value specified as the coverage parameter ξ. In other words, UFVS selects minimal feature value subsets guaranteeing the minimum *explainability* determined by ξ. The most important advantage of UFVS is its significantly high time efficiency. In fact, it processes datasets with thousands of features and instances in only a few tens of milliseconds. In this paper, we use UFVS because can be repeated many times very quickly.

B. Real World Data Clustering

This paper explores two real-world applications of FITS: Twitter topic extraction, and discovering patterns in energy consumption.

Our first example is Twitter topic extraction. Topic models, such as Latent Dirichlet Allocation (LDA) [23], are widely used for topic extraction. However, Twitter is characterized by very short texts and informal styles, it is not easy to apply a topic model that assumes a certain length of texts. In order to overcome this problem, aggregating all the tweets of one user as a single document has been proposed [15], [16]. However, a single tweet is usually about a single topic, there is a risk of losing each topic information in aggregated tweets. [17] proposed adding labels (hashtags) to tweets in advance, and combining Tweets into a single document. But this has the disadvantage of making extracted topics label-dependent. [18] uses hashtags, authors, and a network of followers to create a single document from Tweets. [19] proposed the Twitter-LDA model that assumes Twitter has a fixed number of topics and each of which is represented by a word distribution. These methods require additional information for topic extraction.

Alternatives to LDA exist for topic extraction, such as dimensionality reduction using non-negative matrix factorization (NMF) [20]. However, NMF loses the original features, or words, which makes extracting meaning difficult.

Our second example is analysis of time series electricity consumption data. To promote saving energy, research into analysis of electricity consumption data monitored by smart meters has been ongoing [21], [22], [24], [27]–[29]. Gajowniczek et al. [25] analyzed data from smart meters to detect household characteristics for contributing to higher energy awareness. They used some machine learning, data mining and visualization techniques: k-means clustering, multidimensional scaling, grade correspondence analysis and over-representation map. University campuses are an especially good case study for energy conservation as they contain large numbers of buildings that consume significant energy: reducing consumption is both good for the environment and can reduce costs. The University of Tokyo is promoting the "The Green University of Tokyo Project" [30], which utilizes IT/ICT tools and encourages awareness of energy conservation by visualizing energy consumption in real time. In this paper, as one example of the application, we focus on clustering after feature selection on time series electricity consumption data of a large number of classrooms in university campus.

III. Clustering Based on Feature-value/Instance Transposition Selection (FITS)

We first provide an overview of UFVS and then explain the process of FITS-based clustering.

A. UFVS [6]

As mentioned, we deploy UFVS [6] as a feature value selection engine because of its high time efficiency. In this section, a brief description of UFVS is given.

When an unlabeled dataset D is inputted into UFVS [6], D is first equivalently transformed into D^b by one-hot encoding so that all the features of D^b take binary values 0 or 1. When $\mathcal{F}$ denotes the entire feature set that describe D, we let $\mathcal{F}^b$ denote the entire feature set of D^b.

UFVS takes two parameters: the coverage parameter ξ and the threshold parameter t. ξ determines the minimum

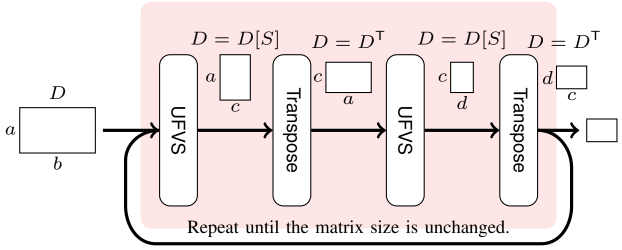

Fig. 1. Feature-value/Instance Transposition Selection: An input D is a matrix with a rows and b columns over $\mathbb{F}_2$ by applying one-hot encoding, if necessary.

percentage of instances to be explained by the selected feature values. t is a minimum number of instances that a feature value must occur in to be considerd for selection. By varying the parameter values, UFVS can produce a wide range of different sets of feature values, which are, in other words, different local solutions in the search range.

B. FITS-based Clustering

Algorithm 1 describes FITS with UFVS as a feature value selection engine, and how FITS can be combined with clustering. We should note that the framework of FITS is independent of any specific feature value selection engine, as far as it is fast enough to perform many iterations.

Algorithm 1 FITS and FITS-based clustering

```
Require: An unlabeled dataset $D$ described by binary features $\mathcal{F}$; a coverage $\xi \in \left[\frac{1}{2}, 1\right]$; a threshold generator $t(\cdot)$.
Ensure: A clustering of instances of $D$.
 1: Let $S_{\text{old}} = \mathcal{F}$;
 2: Let $S_{\text{new}} = \text{UFVS}(D, \xi, t(\xi))$;
 3: Let $D = D[S_{\text{new}}]$;
 4: while $S_{\text{new}} \subsetneq S_{\text{old}}$ do
 5:     Let $D = D^{\mathsf{T}}$;
 6:     Let $D = D[\text{UFVS}(D, \xi, t(\xi))]$;
 7:     Let $D = D^{\mathsf{T}}$;
 8:     Let $S_{\text{old}} = S_{\text{new}}$;
 9:     Let $S_{\text{new}} = \text{UFVS}(D, \xi, t(\xi))$;
10:     Let $D = D[S_{\text{new}}]$;
11: end while
12: Run a clustering algorithm on $D$.
```

Algorithm 1 describes our clustering algorithm: For simplicity, we assume $D = D^b$ and $\mathcal{F} = \mathcal{F}^b$, that is, all the features of D are binary; In other words, D is viewed as a matrix over the binary field of $\mathbb{F}_2 = \{0, 1\}$; The threshold generator function $t(\xi)$ generates a value of the threshold parameter t; Although the range of possible threshold values depends on ξ, it always includes zero; An example of $t(\cdot)$ is to always select $t = 0$; The other extreme exmple is to always select the possible maximum; A setting so that $t(\cdot)$ selects a threshold following a predetermined probability distribution is possible; D^{T} denotes the transpose of D; Also, when S is a subset of the column set of D, $D[S]$ denotes the sub-matrix of D that consists of the column vectors in S; UFVS is understood as a function $\text{UFVS}(D, \xi, t)$ that takes a dataset D, a coverage ξ and a threshold t as arguments; At the last step (Step 12), an arbitrary clustering algorithm is applied to the relevant D and returns clusters of instances; The instances that are in the original D but not in the last shrunken D form a single cluster in the output.

IV. Applications

A. Twitter Topic Extraction

Data gathering was a two-step process. First, we searched the twitter API for all tweets that mentioned a series of coronavirus-related keywords. We collected about a week's worth of tweets for each keyword, between roughly March 13-March 21. Then, we extracted a list of all users in our dataset. This yielded around 600,000 users. We selected a sample of around 100,000 users. For each of these users, we gathered all tweets the user had sent between February 10, 2020 and the beginning of April.

From the above data, we extracted a dataset of of 24,142 tweets posted from 9:00 pm to 9:30 pm on March 1, 2020. In early March, when the corona epidemic was gradually spreading in Japan, so 1717 Tweets with the word "corona" can be found in the target tweets. Morphological analysis (MeCab) was used to stem the Japanese text, and extract keywords from each Tweet. This produced a matrix of 24,142 tweets (instances) x 49,342 unique words (features). Then we applied FITS to form clusters. As described in III, we applied UFVS several times by transposing the matrix. Figure 2 shows the process for applying FITS to the target Twitter data. For simplicity, in Figure 2 we assume $D = D^b$ and all the features of D are binary.

First, we set the coverage parameter $\xi = 0.9$ and the threshold generator generated the threshold value $t = 20$. The first round of UFVS produced a matrix of 24,142 tweets (instances) by 2,132 words (features). The first matrix transpose produced a matrix of 2,132 words (instances) by 24,142 tweets (features). The second UFVS produced a matrix of 2,132 words and 9,827 tweets with the coverage parameter $\xi = 0.9$ and the threshold parameter $t = 20$. The second transpose produced a matrix of 9,827 tweets (instances) and 2,132 words. For the third round of UFVS, we changed the coverage parameter to $\xi = 0.9$ and the threshold generator generated the threshold value $t = 0$, which produced a matrix of 9,827 tweets (instances) by 1,853 words (features). We then stopped the process because we confirmed that the 1,853 words (features) were included in the 2,132 words (features) of the previous matrix. Furthermore, the dimension of the word in the matrix obtained was reduced to two dimensions for visualization by using t-SNE [26], a popular dimensionality reduction method, with Hamming distance and a perplexity parameter of 30. We then applied the DBSCAN clustering algorithm [2], setting the maximum distance between samples (eps) to 2 and number of samples in a neighborhood (min_samples) to 10. This produced 309 clusters. Figure 3 shows the clustering results.

Figure 4 shows the word clouds from the clusters derived from FITS. In this figure, five clusters from A to E are given as examples of extracted topics. The word clouds show that each topic has the following contents.

1) Cluster A: The third anniversary of a game's release
2) Cluster B: A popular animated TV program that aired on March 1

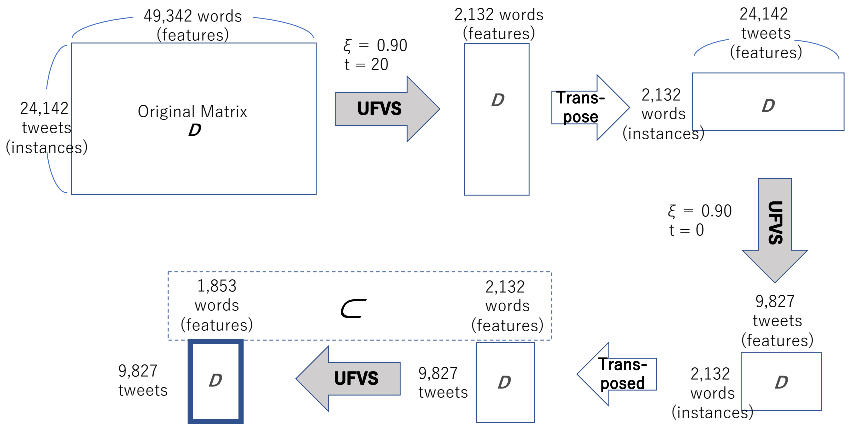

Fig. 2. Applying FITS to Twitter Data

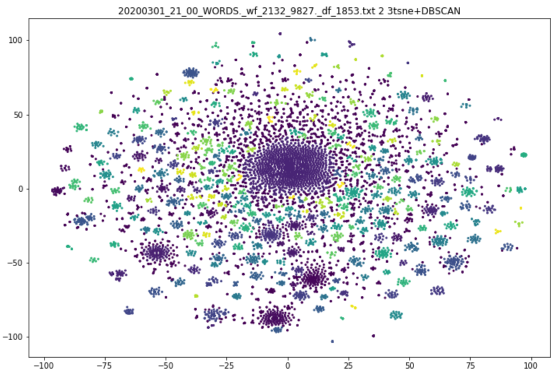

Fig. 3. Clustering Results for the Target Data

3) Cluster C: A popular drama that aired on March 1
4) Cluster D: COVID-19
5) Cluster E: A YouTube Live concert on March 1

Of these topics, four are about entertainment topics, and only one is related to COVID-19. The COVID-19 topic includes words such as "spread of infection" and "masks," which shows people starting to worry about the pandemic. The meaning of each topic is readily apparent from the keywords. While existing clustering methods tend to form clusters consisting of retweets, our method can extract semantically-close and clearer topics by extracting words as features and further transposing them to extract important tweets. Thus, FITS-based clustering can extract quality topics.

B. Electricity Consumption

Chiba University of Commerce was looking for ways to reduce electricity consumption in a large number of classrooms. This example shows how we performed dimension reduction visualization after feature selection.

The electricity consumption of each classroom was recorded in 30-minute time slots when classes were in session (9:00 to 18:30) during a summer semester (16 weeks). An instance is one classroom on one day, with a feature (column) for each time slot, which yields a dataset of 3782 instances and 19 features. For simplicity, we have treated the electricity consumption data as on-off binary data. Figure 5 shows this matrix by heat map.

First, we perform feature selection. Here, we used UFVS and the parameters are set to $\xi = 0.95$ and $t = 2000$. As a result, the time slots 9:00, 11:00, 11:30, 15:30, and 16:30 were selected as features and the number of features was reduced from 19 to 5. Next, we performed a dimensionality reduction visualization using t-SNE with Manhattan distance.

Figures 6 and 7 show t-SNE visualizations using a perplexity parameter of 5, with and without the feature selection, respectively. In both cases, a large cluster of classrooms consuming electricity all day long (Continuously On) and a large cluster of classrooms keeping electricity consumption zero all day long (Continuously Off) are found. On the other hand, in Fig. 6, classrooms with no electricity consumption only in the early morning (Early Morning Off) form a moderate-size cluster, whereas in Fig. 7, no such a cluster is formed. Changing the perplexity parameter up to 30 in the t-SNE visualization without feature selection, a small cluster of Early Morning Off are formed, as shown in Fig. 8. In this case, clusters of Continuously On and Off are not formed and such classrooms are spread broadly.

Figures 9 and 10 are heat maps of clusters of Early Morning Off in Figs. 6 and 8, respectively. After applying UFVS, although some classrooms with no electricity consumption on the dropped features (time slots) are incorporated to the cluster, a larger number of classrooms with similar electricity consumption patterns are gathered into the cluster. On the other hand, without UFVS, only classrooms with precisely-equal consumption pattern cluster together.

By adjusting the UFVS parameters appropriately, feature selection can be performed with a slight reduction of information while retaining sufficient information from the original data. As a result, it is possible to find larger clusters with coarse-graining, which can be useful when a small variance in the cluster is not important, as in this case.

V. Conclusion

Our work has both shown that FITS solves the problems of intractability and inexplicability in two very different

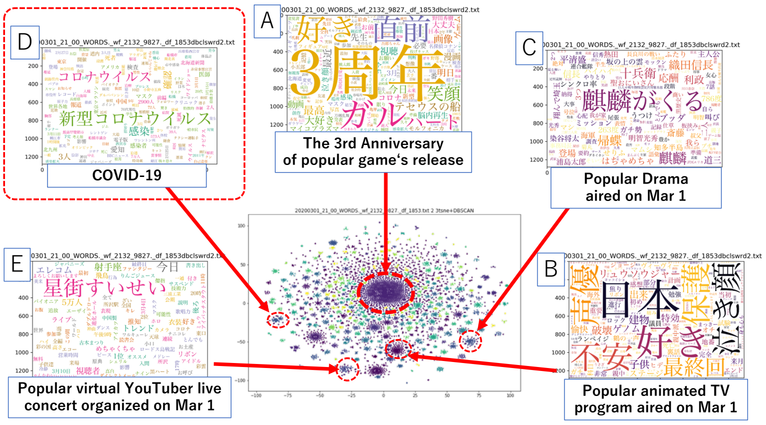

Fig. 4. Word Clouds of Clusters

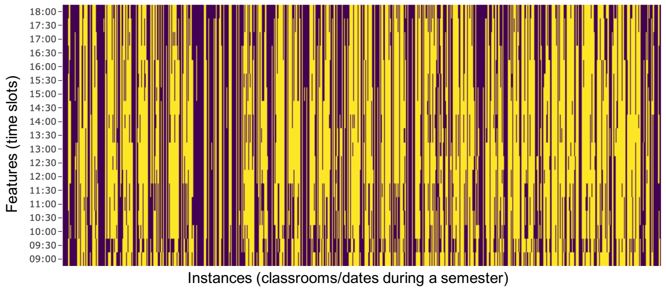

Fig. 5. Heat map of Data Matrix (yellow is on, and blue is off.)

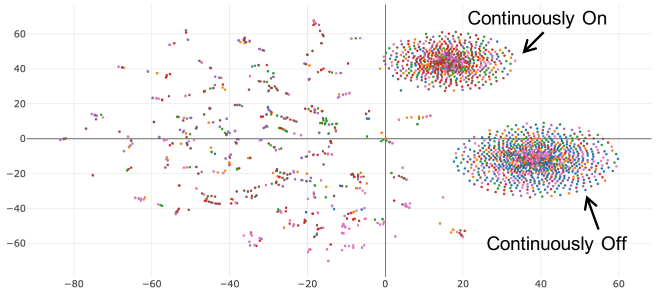

Fig. 7. t-SNE Visualization (Perplexity: 5) without UFVS

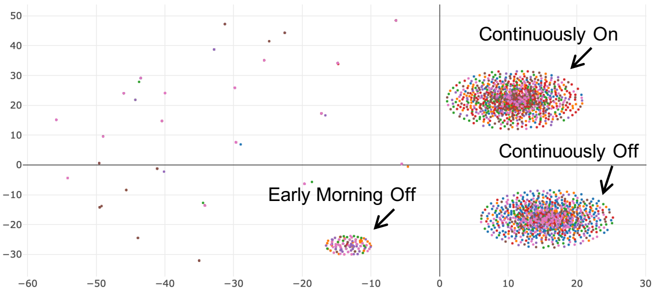

Fig. 6. t-SNE Visualization (Perplexity: 5) with UFVS

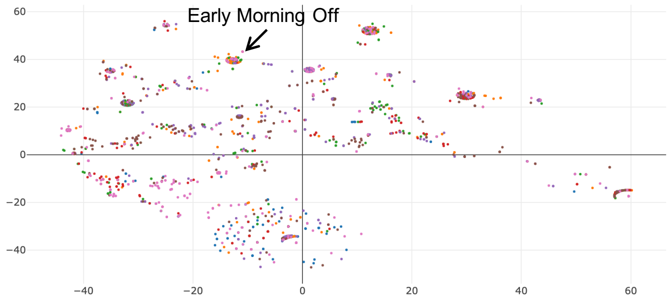

Fig. 8. t-SNE Visualization (Perplexity: 30) without UFVS

applications. These applications demonstrate our algorithm's generality and the range of possibilities it offers. Our future work will evaluate the performance of FITS in relation to other clustering approaches, and examine its application in a wider variety of problems.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Numbers 18K11443, 19K12125, 19H01133, and 17H00762.

References

[1] Cai, D., Zhang, C., He, X.: In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010). pp. 333–342 (2010)

[2] Ester, Martin, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 96(34), 226–231, 1996.

[3] Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: ICML 2000 (2000)

[4] Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. IEEE Transaction on Pattern Analysis and Machine Intelligence **27**(8) (August 2005)

[5] Shin, K., Kuboyama, T., Hashimoto, T., Shepard, D.: sCWC/sLCC: Highly scalable feature selection algorithms. Information **8**(4) (2017)

[6] Shin, K., Okumoto, K., Shepard, D., Kuboyama, T., Hashimoto, T., Ohshima, H.: A fast algorithm for unsupervised feature value selection. pp. 203–213 (01 2020).

[7] Zhao, Z., Liu, H.: Searching for interacting features. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2007). pp. 1156 – 1161 (2007)

[8] Wei, X., Cao, B., Yu, P.S.: Multi-view unsupervised feature selection by cross-diffused matrix alignment. In: Proceedings of 2017 International Joint Conference on Neural Networks (IJCNN 2017). pp. 494–501 (2017)

[9] Wei, X., Cao, B., Yu, P.S.: Unsupervised feature selection on networks: A generative view. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2016). pp. 2215–2221 (2016)

[10] Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: Proceedings of the 24th International Conference on Machine Learning (ICML 2007). pp. 1151–1157 (2007)

[11] Qian, M., Zhai, C.: Robust unsupervised feature selection. In: Proceed-

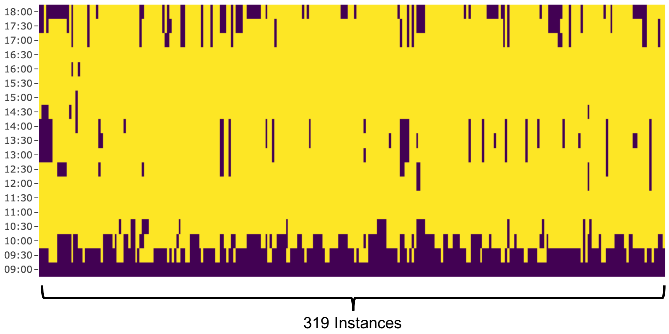

Fig. 9. Heat map of Early Morning Off Cluster with UFVS

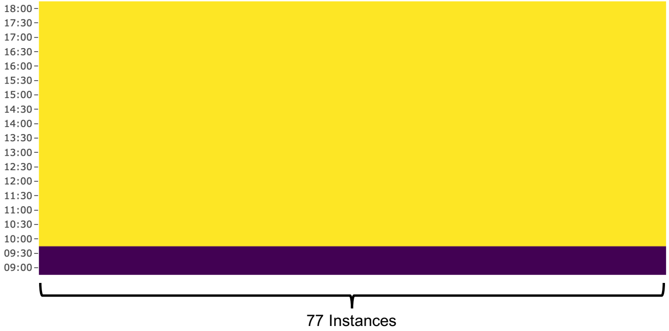

Fig. 10. Heat map of Early Morning Off Cluster without UFVS

ings of 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013). pp. 1621–1627 (2013)

[12] LI, Z., Liu, J., Yang, Y., Zhou, X., Liu, H.: Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Transactions on Knowledge Data Engineering **26**(9), 2138–2150 (2014)

[13] Liu, H., Shao, M., Fu, Y.: Consensus guided unsupervised feature selection. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2016). pp. 1874–1880 (2016)

[14] He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Advances in Neural Information Processing Systems (NIPS 2005). pp. 507–514 (2005)

[15] Weng, J., Lim, E. P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining pp. 261-270 (2010)

[16] Hong, L., Davison, B. D.: Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics, pp. 80-88 (2010)

[17] Ramage, D., Dumais, S., Liebling, D.: Characterizing microblogs with topic models. In Fourth international AAAI conference on weblogs and social media (2010)

[18] Lim, K. W., Chen, C., Buntine, W.: Twitter-network topic model: A full Bayesian treatment for social network and text modeling. arXiv preprint arXiv:1609.06791(2016)

[19] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In European conference on information retrieval, pp. 338-349, Springer, Berlin, Heidelberg (2011)

[20] Casalino, G., Castiello, C., Del Buono, N., Mencar, C.: Intelligent Twitter data analysis based on nonnegative matrix factorizations. In International Conference on Computational Science and Its Applications, pp. 188-202, Springer, Cham (201)

[21] AS Ahmad, MY Hassan, MP Abdullah, HA Rahman, F Hussin, H Abdullah, and R Saidur. A review on applications of ann and svm for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, 33:102–109, 2014.

[22] Adrian Albert and Ram Rajagopal. Smart meter driven segmentation: What your consumption says about you. *IEEE Transactions on power systems*, 28(4):4019–4030, 2013.

[23] David M. Blei, Andrew Y. Ng, Michael I Jordan. Latent Dirichlet Allocation *Journal of Machine Learning Research*, 3 (4–5): 993–1022, 2003.

[24] Christoph Flath, David Nicolay, Tobias Conte, Clemens van Dinther, and Lilia Filipova-Neumann. Cluster analysis of smart metering data. *Business & Information Systems Engineering*, 4(1):31–39, 2012.

[25] Krzysztof Gajowniczek and Tomasz Zabkowski. Data mining techniques for detecting household characteristics based on smart meter data. *Energies*, 8(7):7407–7427, 2015.

[26] Laurens van der Maaten, and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research* 9, 2579–2605, 2008.

[27] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied energy*, 141:190–199, 2015.

[28] M Santamouris, G Mihalakakou, P Patargias, N Gaitani, K Sfakianaki, M Papaglastra, C Pavlou, P Doukas, E Primikiri, V Geros, et al. Using intelligent clustering techniques to classify the energy performance of school buildings. *Energy and buildings*, 39(1):45–51, 2007.

[29] Zhun Yu, Benjamin CM Fung, Fariborz Haghighat, Hiroshi Yoshino, and Edward Morofsky. A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and buildings*, 43(6):1409–1417, 2011.

[30] Hideya Ochiai. Power data management on the internet space: Green ict projects in japan. In *2012 IEEE Colombian Communications Conference (COLCOM)*, pages 1–2. IEEE, 2012.