Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 24 (2018) 166–172

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

4th International Conference on System-Integrated Intelligence

Improving Cluster-based Methods for Usage Anticipation by the Application of Data Transformations

Andres Auquilla[a,b,*], Yannick De Bock[a], Joost R. Duflou[a]

[a]Department of Mechanical Engineering KU Leuven - Flanders Make, B-3001, Heverlee, Belgium
[b]Department of Computer Science University of Cuenca, Cuenca, Ecuador

Abstract

The wide adoption of Internet of Things (IoT) infrastructure in recent years has allowed capturing data from systems that make intensive use of electrical power or consumables typically aiming to create predictive models to anticipate a system's demand and to optimize system control, assuring the service while minimizing the overall consumption. Several methods have been presented to perform usage anticipation; one promising approach involves a two step procedure: profiling, which discovers typical usage profiles; and, prediction that detects the most likely profile given the current information. However, depending on the problem at hand, the number of observations to characterize a profile can increase greatly, causing high dimensionality, thus complicating the profiling step as the amount of noise and correlated features increase. In addition, the profile detection uncertainty increases, as the cluster intra-variability becomes larger and the distances between the centroids become similar. To overcome the difficulties that a usage profile with high dimensionality poses, we developed a methodology that finds the intrinsic dimensionality of a dataset, containing binary historical usage data, by performing dimensionality reductions to improve the profiling step. Then, the profile detection step makes use of the transformed actual data to accurately detect the current profile. This paper describes the implementation details of the application of such techniques by the analysis of two use cases: (1) usage prediction of a laser cutter machine; and, (2) occupancy prediction in an office environment. We observed that the dataset dimensionality and the cluster intra-variability was greatly reduced, making the profile detection less prone to errors. In conclusion, the implementation of methodologies to enhance the separability of the original data by dimensionality transformations improves the profile discovery and the subsequent actual profile detection.

Keywords: usage profiling; cluster detection; dimensionality reduction; logisticSVD; logisticPCA;

[*] Corresponding author. Tel.: +32-16-32-01-73.
E-mail address: andres.auquilla@kuleuven.be

10.1016/j.promfg.2018.06.044

1. Introduction

In recent years, Internet of Things (IoT) technologies have been widely adpted in the industry, allowing to capture large amounts of data regarding systems that make intensive use of electrical power or consumables [17, 10]. The logged data are frequently used to create models which anticipate a system's demand to optimize its control such that the service is guaranteed while minimizing the overall consumption [5].

Several methods have been presented in literature to perform usage anticipation to automatically steer the system's control while maintaining functionality. Shu et al. (2017) [15] presented a non-exhaustive study of these methods and how they are related to smart systems, enumerating three types of methods: (1) pervasive computing; (2) usage profile/prediction; and, (3) intelligent control. Pervasive computing methods inform the user about the current system state, such that the user acts upon this information. Usage profile/prediction methods extend the functionality of pervasive computing by incorporating a two-step process: profiling and prediction. The former involves discovering typical usage profiles, while the latter involves the detection of the most likely profile for the current time [3]. In this way, the system automatically adapts itself to the user expectations. Intelligent control methods implement the usage predictions into a physical controller.

Usage prediction is typically performed by applying machine learning techniques. For instance, Honarmand et al. (2014) [7] modelled the driving patterns of different users by a Markov chain model. This information was then included in an optimization model to define charging/discharging schedules for plug-in electric vehicles. In addition, Basu et al. (2013)[2] studied different machine learning techniques to perform usage prediction in home appliances, such as artificial neural networks, Bayes networks, and decision trees. These methods are data-driven by nature, but they do not provide information about the system from a global and understandable perspective. Therefore, a method such as the one proposed in [3] involving profiling and prediction is more informative to take actions on the system, as it first discovers the typical usage profiles, i.e. ways how the system is typically used, which are then used for prediction. The profile information provided by this method is highly interpretable and the computation power required to perform predictions is very low, which represent a substancial benefit with respect to black box models. However, depending on the problem at hand, the number of observations to characterize a profile can increase greatly, causing high dimensionality, thus complicating the handling process in terms of storage capacity and computational time. In addition, the profiling step becomes less effective, as many of the dimensions are correlated or noisy. Furthermore, the intra-cluster variability increases proportionally to the noise, decreasing the separability between clusters, as the cluster centroids become closer to each other. The type of data that characterizes a profile, e.g. numeric or categorical, has a big impact on the selection of the methodology to discover the profiles since the similarity measurement selection depends heavily on this type of data [1, 14].

To overcome the difficulties that a usage profile with a high dimensionality poses, we developed a methodology that finds the intrinsic dimensionality of a dataset containing binary historical usage data, by performing dimensionality reductions to improve the profiling step. Then, the profile detection step makes use of the transformed and enhanced data to accurately detect the current profile. This paper describes the implementation details of the application of such techniques by the analysis of two use cases: (1) usage prediction of a laser cutter machine; and, (2) occupancy prediction in an office environment.

2. Dataset Description

Two datasets were used for the experiments in this work. One dataset contains information about the machine control mode of an industrial 5 kW CO_2 laser cutting machine for a period of 252 working days. During this period, the machine status was recorded every second. This machine has the following states: start-up, standby, three levels of cutting, and off. For the sake of simplicity, the dataset was altered to differentiate only two states: not-working (off, standby, and start-up) and working (cutting). Then, the dataset was re-sampled to a new resolution of 10 minutes. Therefore, each day contains 144 measurements.

The second dataset corresponds to a time series of occupancy information of a single-user academic office. The occupancy data were collected by a wireless sensor network (WSN), consisting of two motion sensors and a magnetic reed switch attached to the door. The office occupancy was derived from a decision tree [3]. The presence was logged for 239 working days at a time resolution of 15 minutes. In this case, every day is characterized by 96 measurements.

3. Methodology

One effective way to anticipate usage in a system is to identify its recurrent usage patterns, which are then used to estimate the most likely pattern given the current information. The most likely profile is used to predict the expected usage, as it contains probabilistic information in function of time and the variability of the usage patterns. Thus, this process can be divided into two main steps: (1) profiling, which is performed offline and aims to discover the recurrent usage patterns in the dataset; and, (2) profile detection, which is performed online by comparing the information of the current day to the different usage profiles using a similarity score. The usage profiles were segmented into days since human behaviour tends to be strongly correlated on a daily and weekly basis.

3.1. Discovering usage profiles (offline)

The main objective of this step is to discover recurrent usage patterns in the dataset, by applying clustering techniques. For the experiments, two cases were considered to discover profiles: (1) clustering the original dataset (full dimensionality); and, (2) clustering the dataset after transforming it into a dataset with lower dimensionality. For the first case, let $n = 1, \ldots, N$ be the day of each observation, $t = 1, \ldots, T$ the time of the day, $x_n = (x_{n,1}, \ldots, x_{n,T})$ the set of observations for the day n, where $x_{n,t} = \{0, 1\}$ ($x_{n,t} = 0$: no usage, $x_{n,t} = 1$: usage). The number of variables (T) required to characterize a profile depends on the time resolution at which the dataset was logged (144 for the first case and 96 for the second case). As a result of the clustering process, a set of profiles is discovered, i.e. clusters $k = 1, \ldots, K$ such that their cluster intra-variability is minimized and their separation maximized. The clustering process must also consider that not all clusters contain a similar number of days.

Since the observations $x_{n,t}$ are categorical, the selection of the clustering methods and similarity metrics must be oriented to this type of data [14, 13]. For instance, a concept such as the Euclidean distance between two categorical points is not valid, as their distance cannot be represented by such metric. Therefore, the similarity metric used in this experiment was the Jaccard score [8], which provides a valid interpretation of distances between categorical points. The following clustering techniques were used: Hierarchical Clustering, KMedians, and DBSCAN which automatically finds the number of clusters. For a detailed description of these clustering techniques, the reader is referred to [9]. In the case of Hierarchical Clustering and KMedians, the number of clusters is an algorithm parameter. A simple and intuitive method to find the number of clusters was presented by A. Fujita et al. (2013) [6]: the slope statistic is based on the silhouette score proposed by Rousseeuw (1987) [12], and aims to find a dataset partition that maximizes the overall silhouette score and that cannot be further partitioned. The slope statistic is defined as follows $\hat{k} = argmax_k - [s(k + 1) - s(k)]s(k)^p$ where $s(k)$ is the overall silhouette score for k clusters. The optimal cluster configuration is achieved when the configuration of k clusters provides a high silhouette score and a low score for a $k + 1$ configuration.

In the next experiment, the original dataset was transformed into a lower dimensional space by applying logistic principal component analysis (logPCA) [11] and logistic single value decomposition (logSVD) [4]. Logistic PCA is an extension of ordinary PCA oriented to binary data. In the case of logistic PCA, principal components are estimated by using maximum-likelihood by computing a sequence of singular value decompositions. These techniques are valid for the methodology, as the original dataset is binary and the traditional PCA and SVD techniques are unable to cope properly with this type of data. Let $x'_n = (x'_{n,1}, \ldots, x'_{n,T'}), T' << T$ be the observation of day n transformed to a lower space. This data transformation changes the observation type of data such that $x'_{n,t'} \in \mathbb{R}, t' \in T'$. The transformed data are then used in a clustering process with a Euclidean metric as a similarity score. For this experiment, a Hierarchical Clustering process was used due to its simplicity and ability to deal with unbalanced cluster memberships.

3.2. Detecting usage profiles (online)

The profile detection process is continuously performed once the usage profiles have been discovered; it compares the usage information of the current day to the usage profiles. The detection process aims to find the most likely usage profile given the already observed information of the current day. Let $x_{N+1} = (x_{N+1,1}, \ldots, x_{N+1,t^*}), t^* \in T$ be the partial observation of a new day $N + 1$ at time t^*. Let $\delta(x_{N+1})$ be a function that computes the similarity scores of day $N + 1$ with respect to the centroids of the usage profiles $c_k = (c_{k,1}, \ldots, c_{k,t^*}), k \in K$. Then, the most likely cluster is selected

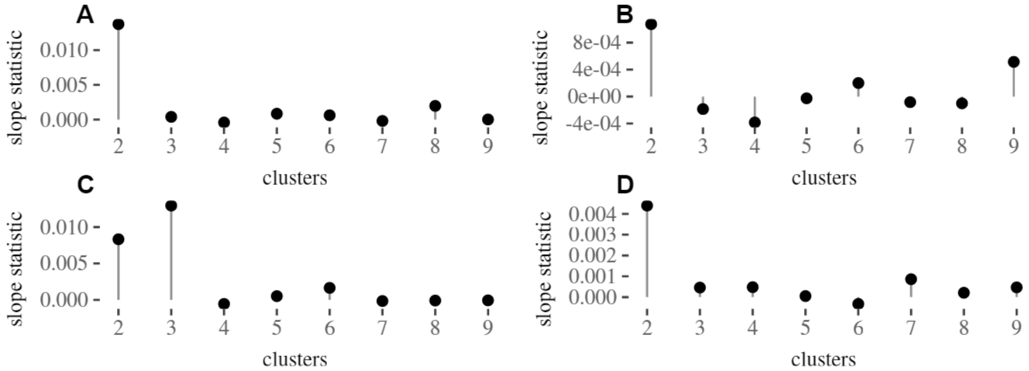

Fig. 1. Slope statistic results for the laser cutter machine usage (first row) and room occupancy dataset (second row). The first and second column depicts the results of using a Hierarchical Clustering and K-centroids algorithms respectively.

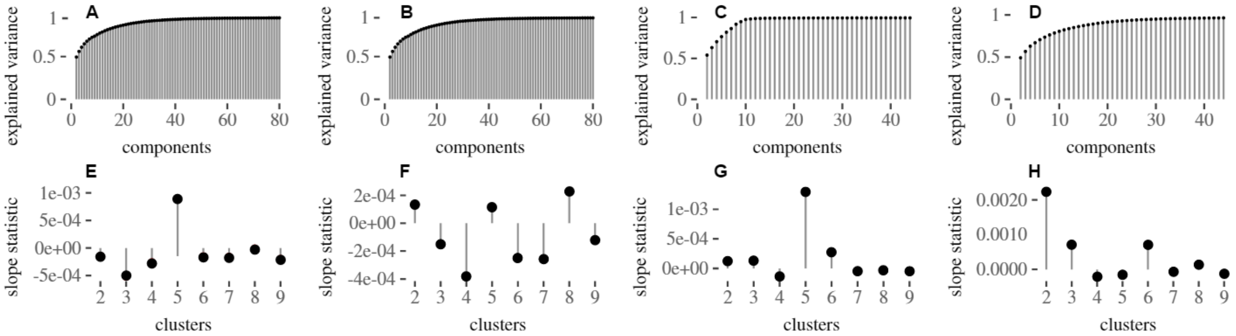

Fig. 2. Explained variance (first row) and slope statistic (second row) results for the laser cutter machine usage (first and second columns) and room occupancy dataset (third and forth columns). This figure depicts the results when using logSVD (first and third column) and when using logPCA (second and forth columns).

according to the similarity metric defined in $\delta(\bullet)$. For the experiments with the original dataset, this score is computed by the Jaccard distance.

For the experiments using the dataset with a lower dimensionality, the new observation x_{N+1} was transformed, and then compared with the cluster centroids of the transformed data $c'_k = (c'_{k,1}, \ldots, c'_{k,t'}), k \in K$. Likewise, the profile with the maximum similarity score was selected as the most likely profile for the current day: $k^* = argmax_k \delta(c'_k, x_{N+1})$. For the transformed dataset case, Euclidian distance was used a similarity metric ($\delta(\bullet)$), since it is simple to interpret, fast, and is a valid score for the transformed type of data.

The last 15 days of the dataset were separated as testing data. The profile detection accuracy was tested at different times on the testing days, i.e. from 8 am till 8 pm in intervals of two hours.

4. Usage Profiles Discovery

In this section, the results regarding the identification of profiles are depicted. Firstly, the profiles discovered by using all features in the dataset are shown; then, the profiles discovered with the dimensionality reduction methods, i.e. logSVD and logPCA, are analyzed.

When clustering using all the dataset features (Figure 1), there is a very strong presumption that the best choice is two clusters for both datasets, i.e. laser cutter machine (Figure 1A-B) and occupancy in an office (Figure 1C-D), regardless of the clustering algorithm. However, having only two clusters to perform usage anticipation is rudimentary as those clusters contain an average of the typical days: workday and holiday; this results in an uninformative dataset partitioning, supporting the idea that many of the variables of the original dataset are noisy and highly correlated to each other.

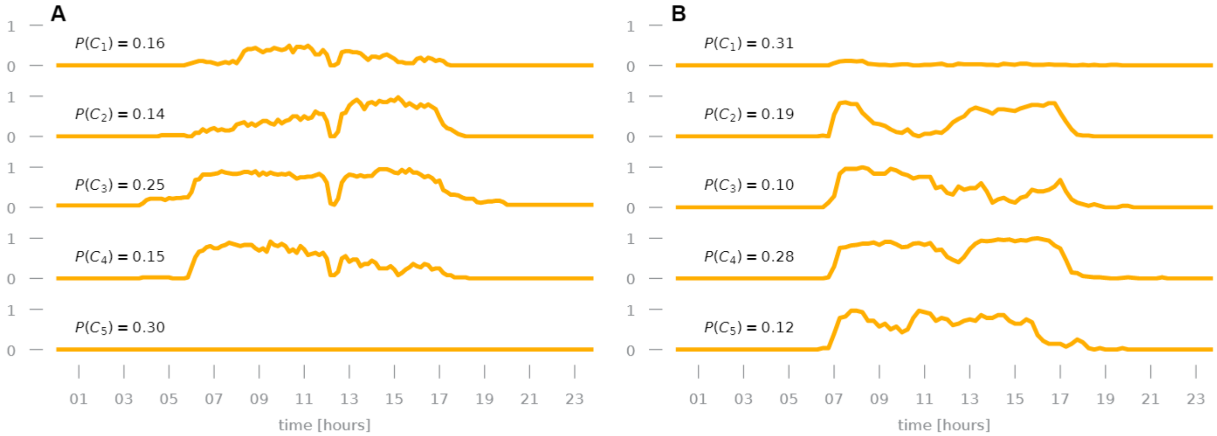

Fig. 3. Usage profiles of the laser cutter machine (left) and the room occupancy in an office environment (right). The profiles are shown with their cluster probability which is computed as the number of days that belongs to the cluster divided by the total number of days observed.

When the original dataset is transformed by applying logSVD and logPCA, the number of components required to characterize the original dataset decreases strongly. In the case of the laser cutter machine, 20 and 45 dimensions were required by the logSVD and logPCA methods (Figure 2A-B) to explain ≥ 93% of the original variance. In the case of the office environment, 15 and 25 dimensions were required to explain ≥ 93% of the original variance by the logSVD and logPCA methods (Figure 2C-D). In all cases, logSVD requires fewer features to explain larger levels of the original dataset variance. This is an interesting characteristic since only a small proportion of the original feature space is required to characterize the original dataset. The explained variance, in the case of logSVD, grows faster as the number of components increases compared to using logPCA. Therefore more components are required to explain the same variance when using this last method compared to logSVD.

The determination of the number of clusters is simpler for both datasets when using logSVD as a dimensionality reduction method since the slope statistic shows clear candidates (Figure 2E,G). When using logPCA, the cluster configurations are more complicated to select, since the slope statistic for both datasets shows a few possible candidates (Figure 2F, H). When there are several cluster configuration candidates, the selection is performed based on expert knowledge about the dataset. For instance, it is expected to have at least two clusters in both datasets, i.e. holidays and working days; however, this data partition is not informative enough to perform usage prediction. Therefore, cluster configurations of $k > 2$ are preferred. On the other hand, when k is too large, the cluster detection process becomes more complicated as the chance to select a wrong profile increase; in addition, there could be clusters with limited membership, thus indicating overfitting. Based on the results of Figure 2, logSVD provides the best results in terms of number of components to characterize the data and number of clusters determination for both datasets. The dimensionality reduction process proved to be successful as the suggested cluster configurations by the slope statistic are more informative, i.e. contain more clusters, compared to the case of using all features, which is of crucial importance for the usage prediction step. For both datasets, DBSCAN was also tested; this technique discovered two clusters for both datasets, regardless of the number of dimensions. In other words, DBSCAN was only able to discover the clusters representing holidays and working days, which is not an informative clustering configuration.

The results of the clustering exercises obtained by applying logSVD on the datasets are shown in figure 3. In the case of the laser cutter dataset (Figure 3A), the discovered clusters, except for C1, contain large regions where the usage probability is close to 1 or 0, indicating a high degree of certainty. In the case of the occupancy in an office environment (Figure 3B), the discovered clusters also have large regions with occupancy probability close to 0 or 1.

5. Usage Profile Detection

In this step, the discovered profiles were detected at different times during the day by comparing the known cluster centroids with the current information of the day. The results are compared in terms of accuracy, i.e. the ratio of correctly detected profiles, for both datasets when using all features and after performing a dimensionality reduction. These results are depicted in Table 1. The accuracy results of the columns with the label *eucl* were computed when using Euclidian distance as a similarity score. Likewise, the accuracy results of the columns with label *lda* were computed when using a linear discriminant analysis (LDA) model to classify the type of profile given the current usage information. This model was trained by using the training dataset and their cluster membership labels. In addition, the accuracy results of the columns with label *likelihood* were computed when using the similarity metric presented in [16].

Table 1. Accuracy results for the laser cutter machine and office occupancy datasets when detecting clusters by using all the features and after a dimensionality reduction by logSVD and logPCA.

Time	Laser cutter usage						Occupancy in an office					
	All features		logSVD		logPCA		All features		logSVD		logPCA	
	likelihood	eucl	eucl	lda	eucl	lda	likelihood	eucl	eucl	lda	eucl	lda
8	46.7	46.6	60.0	33.3	66.7	66.7	33.3	53.3	60.0	60.0	40.0	53.3
10	46.7	40.0	73.3	53.3	60.0	80.0	53.3	53.3	66.7	60.0	60.0	53.3
12	46.7	66.7	73.3	60.0	60.0	80.0	80.0	80.0	86.7	100.0	60.0	53.3
14	66.7	66.7	73.3	80.0	80.0	86.7	93.3	93.3	93.3	100.0	66.7	73.3
16	100.0	100.0	100.0	86.7	80.0	86.7	100.0	100.0	100.0	100.0	80.0	93.3
18	86.7	100.0	100.0	86.7	80.0	86.7	100.0	100.0	100.0	100.0	86.7	100.0
20	86.7	100.0	100.0	100.0	86.7	93.3	100.0	100.0	100.0	100.0	86.7	100.0

In all cases, the results achieved by the methods using transformed datasets outperformed the ones achieved by the methods using all features. At the beginning of the day, the information is scarce, however, the accuracies of the logSVD and logPCA methods are larger than the ones using all features. For both datasets, logSVD obtained better accuracy results: in the case of the laser cutter machine with a Euclidian distance score, and for the occupancy case with an *LDA* classification model.

6. Conclusions

The use of techniques to reduce the dimensionality in binary datasets, such as logistic SVD and logistic PCA, is beneficial as their representation becomes more concise and the noisy variables are discarded in the process. As a result, the intrinsic dimensionality of the dataset can be discovered, thus helping the clustering process to discover the most relevant profiles from the historical data. Dimensionality reduction techniques are also beneficial to the process of detecting the most likely cluster given the already available information of the current day. When using a transformed dataset, the detection accuracy increases and becomes perfect faster than when using the original dataset. As our results show, logSVD provides a better data representation for the datasets as the number of dimensions required to explain the original variance is lower than for logistic PCA. In addition, logSVD improves the separability between clusters, as shown in the detection results for both test sets. On the contrary, using all the features of a dataset can be problematic, since the noisy features can become dominant, reducing the chances of discovering the relevant profiles in the historical data and their subsequent detection.

References

[1] Ahmad, A., Dey, L., 2007. A k-mean clustering algorithm for mixed numeric and categorical data. Data and Knowledge Engineering 63, 503–527. doi:10.1016/j.datak.2007.03.016.

[2] Basu, K., Hawarah, L., Arghira, N., Joumaa, H., Ploix, S., 2013. A prediction system for home appliance usage. Energy and Buildings 67, 668–679. URL: http://linkinghub.elsevier.com/retrieve/pii/S0378778813000789, doi:10.1016/j.enbuild.2013.02.008.

[3] De Bock, Y., Auquilla, A., Kellens, K., Vandevenne, D., Nowé, A., Duflou, J.R., 2017. User-Adapting System Design for Improved Energy Efficiency During the Use Phase of Products: Case Study of an Occupancy-Driven, Self-Learning Thermostat. Springer Singapore, Singapore. pp. 883–898. URL: https://doi.org/10.1007/978-981-10-0471-1_60, doi:10.1007/978-981-10-0471-1_60.

[4] De Leeuw, J., 2006. Principal component analysis of binary data by iterated singular value decomposition. Computational Statistics and Data Analysis 50, 21–39. doi:10.1016/j.csda.2004.07.010.

[5] Duflou, J.R., Auquilla, A., De Bock, Y., Nowé, A., Kellens, K., 2016. Impact reduction potential by usage anticipation under comfort trade-off conditions. CIRP Annals - Manufacturing Technology 65, 33–36. doi:10.1016/j.cirp.2016.04.087.

[6] Fujita, A., Takahashi, D.Y., Patriota, A.G., 2014. A non-parametric method to estimate the number of clusters. Computational Statistics and Data Analysis 73, 27–39. URL: http://dx.doi.org/10.1016/j.csda.2013.11.012, doi:10.1016/j.csda.2013.11.012.

[7] Honarmand, M., Zakariazadeh, A., Jadid, S., 2014. Optimal scheduling of electric vehicles in an intelligent parking lot considering vehicle-to-grid concept and battery condition. Energy 65, 572–579. URL: http://linkinghub.elsevier.com/retrieve/pii/S0360544213010153, doi:10.1016/j.energy.2013.11.045.

[8] Jaccard, P., 1901. Etude de la distribution florale dans une portion des alpes et du jura 37, 547–579.

[9] Joshi, A., Kaur, R., 2013. A Review: Comparative Study of Various Clustering Techniques in Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering 3, 2277–128.

[10] Koseleva, N., Ropaite, G., 2017. Big Data in Building Energy Efficiency: Understanding of Big Data and Main Challenges. Procedia Engineering 172, 544–549. URL: http://dx.doi.org/10.1016/j.proeng.2017.02.064, doi:10.1016/j.proeng.2017.02.064.

[11] Landgraf, A.J., Lee, Y., 2015. Dimensionality Reduction for Binary Data through the Projection of Natural Parameters URL: http://arxiv.org/abs/1510.06112, arXiv:1510.06112.

[12] Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53 – 65. URL: http://www.sciencedirect.com/science/article/pii/0377042787901257, doi:https://doi.org/10.1016/0377-0427(87)90125-7.

[13] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.T., 2017. A review of clustering techniques and developments. Neurocomputing 267, 664–681. URL: http://dx.doi.org/10.1016/j.neucom.2017.06.053, doi:10.1016/j.neucom.2017.06.053.

[14] Seung-Seok, C., Sung-Hyuk, C., Tappert, C.C., 2010. A Survey of Binary Similarity and Distance Measures. Journal of Systemics, Cybernetics & Informatics 8, 43–48. URL: http://ezproxy.uthm.edu.my/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=59856128&site=ehost-live&scope=site, doi:10.1.1.352.6123.

[15] Shu, L.H., Duflou, J., Herrmann, C., Sakao, T., Shimomura, Y., De Bock, Y., Srivastava, J., 2017. Design for reduced resource consumption during the use phase of products. CIRP Annals - Manufacturing Technology 66, 635–658. doi:10.1016/j.cirp.2017.06.001.

[16] Truong, N.C., McInerney, J., Tran-Thanh, L., Costanza, E., Ramchurn, S.D., 2013. Forecasting multi-appliance usage for smart home energy management, in: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press. pp. 2908–2914. URL: http://dl.acm.org/citation.cfm?id=2540128.2540547.

[17] Zhou, K., Fu, C., Yang, S., 2016. Big data driven smart energy management: From big data to big insights. Renewable and Sustainable Energy Reviews 56, 215–225. doi:10.1016/j.rser.2015.11.050.