Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 175 (2020) 677–682

Procedia Computer Science

www.elsevier.com/locate/procedia

The 7th International Symposium on Emerging Inter-networks, Communication and Mobility (EICM)
August 9-12, 2020, Leuven, Belgium

Unsupervised learning algorithms applied to grouping problems

Amelec Viloria [a,*], Nelson Alberto Lizardo Zelaya[b], Noel Varela[c]

[a,c] *Universidad de la Costa, Barranquilla, Colombia.*
[b]*Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras*

Abstract

One of the tasks of great interest within process mining is the discovery of business process models, which consists of using an event log as input and producing a business process model by analyzing the data contained in the log and applying a process mining method, task and/or technique. The discovery allows the identification of the behaviors contained in the cases of the event log in order to detect possible deviations and/or validate that the business process is executed according to the business requirements. This paper presents an approach based on unsupervised learning techniques for the grouping of traces to generate simpler and more understandable models. The algorithms implemented for clustering are K-means, hierarchical agglomerative and density-based spatial clustering of applications with noise (DBSCAN).

Keywords: Trace grouping; Data mining; Unsupervised learning techniques.

1. Introduction

Process mining aims to discover, monitor and improve process models through knowledge extraction from data contained in event records [2]. These objectives have been addressed in different areas, for example, healthcare [1], industry [3], education [4], etc. Among the studies reported in the state of the art on the grouping of traces by means

* Corresponding author. Tel.: +57-3046238313.
E-mail address: aviloria7@cuc.edu.co

10.1016/j.procs.2020.07.099

of unsupervised learning techniques is the research conducted by [5] where 3 unsupervised learning algorithms are used, K-means, Hierarchical Agglomerative and the algorithm based on Self-Organization Map (SOM) neural networks. The tests carried out in this study included the execution of the unsupervised learning algorithms with 2 raw profiles and with the profiles processed by the techniques of random projection dimensionality reduction, principal components analysis (PCA) and singular value decomposition (SVD). In the paper of [6], trace grouping is applied to discover similar behaviors in the event log, being very difficult to find them manually in large event logs. On the other hand, in the paper of [7], a hierarchical grouping is applied to discover the behavior in a collaborative environment where a set of event records from different systems can be kept.

Therefore, this research presents an approach based on unsupervised learning techniques for trace grouping. The algorithms implemented for clustering are K-means, hierarchical agglomerative and DBSCAN. The proposal consists of performing the tuning or selection of the best parameters for each unsupervised learning algorithm using the Silhouette metric, which allows to measure the quality of the clusters performed, facilitating the acceleration of the selection of the best parameters in the clustering of the traces with an average aptitude close to 0.80, with which simple process models can be discovered.

2. Methods

2.1 Trace grouping

Unsupervised learning techniques group elements based on a similarity given by a type of distance (e.g. Euclidean or Hamming distances [8]), which indicates how similar a trace is compared to other traces contained in the event log. Also, it indicates that the information of the traces contained in the event log should be represented in a numerical vector space to determine the similarity. In this sense, the papers [9] and [10] propose a series of transformations called "trace profiles" which is a representation of the traces in a numerical vector space.

The first step in creating these profiles is to identify the unique events in the event log and their source [11]:

- Transition profile. For any combination of two events (A, B), this profile contains an element that records the occurrence of an event A being followed directly by another event B.
- Activity profile. For any event A, this profile contains an element that records the appearance of event A on the trace.
- Origin profile. For any combination of an event A and a user X, this profile contains an element that measures how often an event A has been performed by user X.

2.2 Unsupervised Learning Techniques

In techniques based on hierarchical agglomerative algorithms [12], where there are two variants (see Figure 1):

- All the elements are grouped in a single group and this is separated until there are as many groups as elements.
- It starts with as many groups as there are elements in the dataset, and in each iteration pairs of more similar sets or elements are grouped together. This algorithm is based on the use of 3 types of linkage criteria for element formation or separation (minimum distance, maximum distance, and average distance).

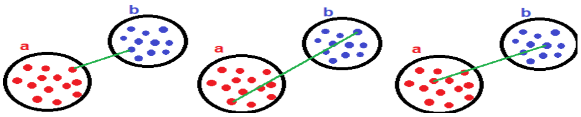


Fig. 1. Distances for the grouping of elements. a) Minimum distance b) Maximum distance c) Average distance.

2.3 Optimization of the Parameters of the Unsupervised Learning Techniques

One of the main disadvantages of clustering algorithms is to define the number of groups that will be formed and that are related to the parameters they receive. The K-means algorithm receives the K parameter which is the number of groups that will be formed. In the hierarchical agglomerative algorithm, the number of groups will always be one or as many groups as elements exist in the data set, provided that the number of groups formed is not limited.

In our proposal, the Silhouette Coefficient [13] is used to measure the quality of the groups formed and thus select the best parameters for the clustering algorithms. The Silhouette Coefficient refers to a method of interpretation and consistency validation within NN data sets. The value of this measure identifies how similar an object is to its own group (cohesion $aa(xx)$) compared to other groups (separation bb). The metric varies from -1 to +1, if the value is close to -1 it means a bad grouping, if the value is close to 0 the grouping is indifferent and if the value is close to 1 it means a good grouping. To calculate the Silhouette metric $ss(xx)$ for a group, the equation 1 is used and the Silhouette coefficient $CCCC$; for all clustering, the equation 2 [14] is used.

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \tag{1}$$

$$CS = \frac{1}{N}\sum_{i=1}^{N} s(x) \tag{2}$$

The following is a list of the steps followed by the methodology [1][3][5]:

- The event log is represented using a trace profile (of activities and transitions).
- The second stage consists of the execution of the grouping algorithms with a range of values between 2 and 50 as parameters, as well as the profiles generated in the previous task. The value "50" was selected as the maximum value, which is arbitrary because the best number of groups for the dataset is unknown. For each execution, the groups formed are evaluated using the Silhouette metric. It should be noted that during these runs it is not necessary to discover the process models.
- In the third stage, for each clustering algorithm, the parameters that obtained the best Silhouette metric are selected.
- The parameters selected in the third stage are then used to run the unsupervised learning algorithms. The trace groups formed by these algorithms are used to create new event records and discover the simplest process models. The PROM 6.8 Framework is used for the discovery process.

3. Results

3.1 Case Study

The event log generated by the EDS hospital information system [8], [11] is used in the evaluation of the implementation of the proposed unsupervised learning algorithms. The record comes from a hospital billing system, where each event refers to a service provided to a patient between 2015 and 2019. The event log is composed of 1,254,321 different event names, 2548 cases and a total of 954,325 events.

3.2 Test Scenario 1: Selecting the Best Parameters

The K-means algorithm was executed varying the value of K between values 2 to 50. Figure 2 shows the behavior of the Silhouette metric for this range. Where K = 10 has the best Silhouette metric equal to 0.569, using the activity profile and using the transition profile with a Silhouette metric equal to 0.582.

The hierarchical agglomerative algorithm was executed varying the total of groups formed between 2 and 50, this same configuration was executed with the 3 types of link criteria minimum, maximum and average distance. The average distance with the best Silhouette measure is 0.73, considering the formation of 21 groups and using the activity profile. The maximum distance with the best Silhouette measure is 0.57. On the other hand, using the profile of transitions the best Silhouette measure is 0.85, using the average distance considering 6 groupings. The maximum distance with the Silhouette measure is 0.84.

Figure 3 shows the behavior of the Silhouette metric for the different groups formed using the average distance bonding criterion. For the DBSCAN algorithm, the Eps and minPts parameters were calculated using the distances of the nearest neighbors (k-nearest) [12][14] [15]. The average of the distances of each trace to its K-nearest is calculated. The value of K used is equal to 4 and corresponds to the minPts parameter. The K distances are then plotted in ascending order.

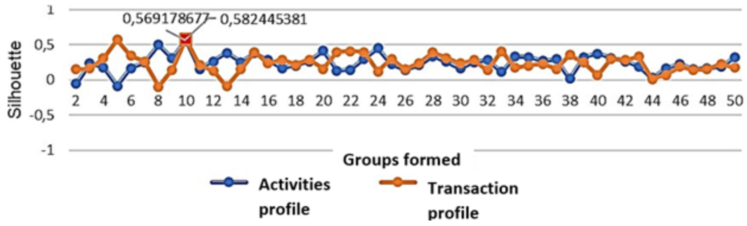


Fig. 2. Silhouette metrics for different K-values in K-means algorithm.

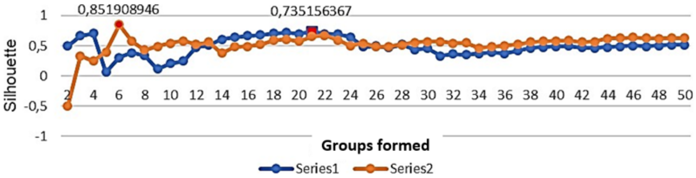


Fig. 3. Silhouette metric for the groups formed by the Hierarchical algorithm.

The objective is to determine the "knee" (drastic change), corresponding to the optimal Eps parameter. In this sense, a knee corresponds to a threshold where a sudden change occurs along the K-distance curve. Figure 4a shows the graph of the distances in the data set for the activity profile. The optimal value of the parameter Eps is between values 2.0 and 7.0. The algorithm was executed with all the values between these ranges, adding 0.1, producing the value of Eps equal to 3.4 with the best Silhouette measure equal to 0.59, considering the formation of 5 groups. For the transitions profile (Figure 4b), the optimal value of the Eps parameter is between values 2.0 to 4.5. Likewise, the activity profile, the DBSCAN algorithm was executed with all the possible values between this range obtaining the value of Eps equal to 3.7 with the Silhouette metric equal to 0.52, considering the formation of 3 groups. Figure 5 shows the behavior of the Silhouette metric for the different values of the Eps parameter.

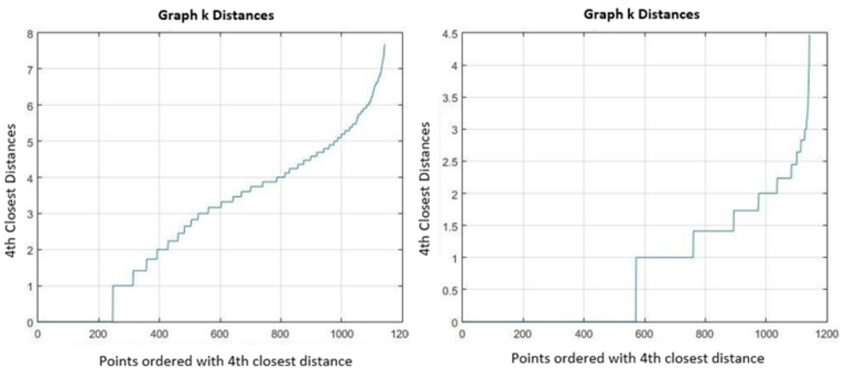
3.2 Test Scenario 2: Evaluation of the Processes Discovered

A means of validating that groups are well formed is discovering the process models and obtaining the average proficiency measure. Therefore, the aptitude of each group is measured and the average of the whole grouping is calculated, which is done for all the groupings made by the grouping algorithms considering the previously selected parameters. The result of the grouping of traces with a combination that shows the highest mean aptitude value is considered the best combination of grouping algorithm and event log representation.

Table 1 shows the results of the proficiency measurement for each of the groups generated by the K-means algorithm using the transition profile.

The results of the suitability of the groups formed by the hierarchical algorithm using the average distance linkage criterion are shown in Table 2 for the transition profile.

Similarly, the results of the DBSCAN algorithm are shown in Table 3. The "Group" column indicates the number of the group, the "#Traces" column indicates the total number of cases or traces within the group, the "#Events" column indicates the total number of events within the group and the "Aptitude" column indicates the aptitude of each group. Also, at the end of each table is the total number of traces, the total number of events and the average fitness of the entire group.



a) Activity profile b) Transition profile

Fig. 4. Selection of appropriate values for the DBSCAN algorithm.

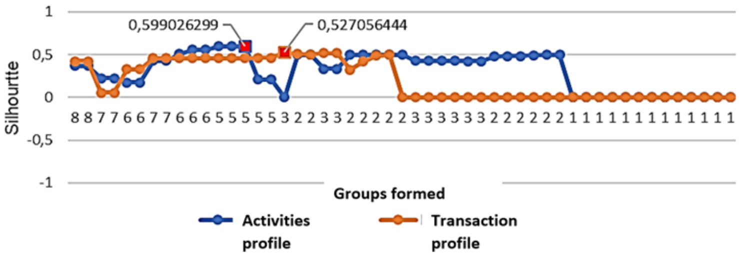


Fig. 5. Silhouette metrics for different numbers of groups formed by DBSCAN.

Table 1. Results of the K-means algorithm using the transition profile

Group	# Traces	#Events #	Aptitude	Group	# Traces	#Events #	Aptitude
1	14.254	7.524	0.3241	8	452	122.258	0.6752
2	18.524	320.145	0.6478	5	451	352.147	0.8014
3	10.458	314.258	0.9412	7	85	140.258	0.7954
4	13.654	104.250	0.6895	10	75	251.254	0.8452
5	1.254	208.148	0.8147	9	7	88.408	0.4574
						954.325	**Average: 0.6945**

Table 2. Results of the hierarchical algorithm using the transition profile.

Group	# Traces	#Events #	Aptitude
1	20	542	0.6547
2	15.254	415.254	0.8958
3	10	2.524	0.7845
4	15.478	245.247	0.4541
5	58	102.254	0.7478
6	12	188.504	0.7841
		954.325	**Average: 0.7885**

Table 3. Results of the DBSCAN algorithm using the transition profile

Group	**#Tracks #**	**#Events #**	**Aptitude**
1	512	548.254	0.7014
2	544	258.254	0.1548
3	141	147.817	0.7854
		954.325	**Average: 0.5412**

4. Conclusions

In this research, a study of the application of unsupervised learning techniques for the grouping of traces in event records was presented. The characteristic of these techniques is the generation of a spaghetti-type process model. The proposed methodology involves the tuning or selection of the appropriate parameters of the clustering algorithms. According to the results obtained in the experimentation, it is concluded that the use of the Silhouette metric allows to speed up the clustering of traces and the discovery of simple process models to select the adequate parameters of the clustering algorithms, with an aptitude mean of 0.7885, which is the result obtained from the clustering of the hierarchical algorithm using the activity profile, considered with the best performance of the 3 algorithms analyzed in this study..

References

[1] Celebi, M. E., & Aydin, K. (Eds.). (2016). Unsupervised learning algorithms (Vol. 9, p. 103). Springer.

[2] Albert, S., Teletin, M., & Czibula, G. (2018). Analysing protein data using unsupervised learning techniques. Int. J. Innovative Comput. Inf. Control, 14(3), 861-880.

[3] Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. Technological Forecasting and Social Change, 115, 131-142.

[4] Banerjee, N., Giannetsos, T., Panaousis, E., & Took, C. C. (2018, July). Unsupervised learning for trustworthy IoT. In 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-8). IEEE.

[5] Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data mining and analytics in the process industry: The role of machine learning. Ieee Access, 5, 20590-20616.

[6] Chauhan, R., Kaur, H., & Puri, R. (2017). An Empirical Analysis of Unsupervised Learning Approach on Medical Databases. In Emerging Trends in Electrical, Communications and Information Technologies (pp. 63-70). Springer, Singapore.

[7] Srinivas, C., & Rao, C. G. (2019, June). A novel approach for unsupervised learning of software components. In Proceedings of the 5th International Conference on Engineering and MIS (1-6).

[8] Fu, W., & Menzies, T. (2017, August). Revisiting unsupervised learning for defect prediction. In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (pp. 72-83).

[9] Packianather, M. S., Davies, A., Harraden, S., Soman, S., & White, J. (2017). Data mining techniques applied to a manufacturing SME. Procedia CIRP, 62, 123-128.

[10] Bokhari, S. M. A., & Khan, S. A. (2016). Applying Supervised and Unsupervised Learning Techniques on Dental Patients' Records. In Emerging Trends and Advanced Technologies for Computational Intelligence (pp. 83-102). Springer, Cham.

[11] Unnisa, M., Ameen, A., Raziuddin, S. (2016). Opinion mining on twitter data using unsupervised learning technique. International Journal of Computer Applications, 148(12), 975-8887.

[12] Henkel, J., Lahiri, S. K., Liblit, B., & Reps, T. (2019). Enabling Open-World Specification Mining via Unsupervised Learning. arXiv preprint arXiv:1904.12098.

[13] Viloria, A., Guerrero, I. M., Caraballo, H. M., Llinas, N. O., Valero, L., Palma, H. H., … Lezama, O. B. P. (2019). Effect on the demand and stock returns: Cross-sectional of big data and time-series analysis. In Communications in Computer and Information Science (Vol. 1122 CCIS, pp. 211–220). Springer. https://doi.org/10.1007/978-981-15-1301-5_17.

[14] Tax, N., Sidorova, N., Haakma, R., & van der Aalst, W. M. (2016, September). Event abstraction for process mining using supervised learning techniques. In Proceedings of SAI Intelligent Systems Conference (pp. 251-269). Springer, Cham.

[15] Viloria, A., Angulo, M. G., Kamatkar, S. J., de la Hoz – Hernandez, J., Guiliany, J. G., Bilbao, O. R., & Hernandez-P, H. (2020). Prediction Rules in E-Learning Systems Using Genetic Programming. In Smart Innovation, Systems and Technologies (Vol. 164, pp. 55–63). Springer. https://doi.org/10.1007/978-981-32-9889-7_5