Outlier Identification for Symbolic Data with the Application of the DBSCAN Algorithm

Marcin Pełka

Abstract Outliers have a significant negative impact on the data quality, data analysis results. If a large dataset contains only few outliers it is essential to identify them and then remove them (e.g., for credit card transactions only some of them will be fraudulent) or not. The paper presents an application of ensemble learning for symbolic data as a tool for outlier detection, where the DBSCAN (density-based SCAN) algorithm is applied. In the empirical part ensemble learning and single DBSCAN algorithm is used to detect outliers in an unbalanced data sets. The results show that ensemble approach can be efficiently used to detect outliers in symbolic data sets.

Keywords Symbolic data analysis · Outliers · DBSCAN

1 Introduction

Outliers have a significant negative impact on the data quality. According to Hawkins (1980), an outlier is an observation that deviates so much from the other observations in the dataset as to arouse suspicious that it was generated by some different mechanism than other data entities.

Detection of outliers is a fundamental issue in data analysis, its main goal is to detect and remove anomalous objects from the data. Because the technology changes rapidly, number of databases and their size grows over time.

There are many different methods for outlier detection than can be used, starting from the simplest ones to more complex ones like feature bagging, subsampling, rotated bagging, isolation forests, outrank method, the approach proposed by Nguyen et al. (2010) (see, for example, Aggrawal 2013, 2015; Aggrawal and Sathe 2017; Nguyen et. al. 2010).

M. Pełka (✉)
Wroclaw University of Economics and Business, Wrocław, Poland
e-mail: marcin.pelka@ue.wroc.pl

K. Jajuga et al. (eds.), *Modern Classification and Data Analysis*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-10190-8_5

Usually outlier detection algorithms are seen as statistical models of data that allow identify objects that do not fit the model, whereas the aim of the distance-based approaches is to measure the distance between data. In such case, outliers are the data for which the distance is greater than given threshold (see, for example, Aggrawal 2013, 2015; Zhang 2013). One of such distance-based algorithms is the DBSCAN (density-based spatial clustering algorithm for applications with noise) algorithm.

Ester et al. (1996) have proposed a density-based algorithm for discovering clusters for classical data. This algorithm groups together points that have many neighbors and also detects outliers that are left in low-density subregions. In 1998, Sander et al. have proposed a generalized version of DBSCAN—the GDBSCAN. The generalized algorithm can cluster objects as well as spatially extend objects according to both their spatial and non-spatial attributes (Sander et al. 1998). Compello et al. (2013) have proposed a hierarchical version of DBSCAN. In 1999, Ankrest et al. proposed the OPTICS algorithm that extends the ideas of DBSCAN (Ankrest et al. 1999). Other DBSCAN extensions are SUBCLU and PreDeCon (Jahirabadkar, Kulkarni 2013; Kailing et al. 2004). Both use similar subspace clustering ideas that are similar to those of DBSCAN. WaveCluster (Sheikholeslami et al. 1998) is another modification of DBSCAN. It uses the wavelet transform to the dimension space and unfortunately is applicable only to low-dimensional datasets. DenClue (Hinnburg and Keim 1998) is another efficient algorithm that uses information about density to cluster objects.

The paper presents an application of ensemble learning for symbolic data as a tool for outlier detection, where the DBSCAN algorithm is applied. Also it analyzes how DBSCAN's initial parameters impact the number of detected outliers and the clustering quality itself. It is also the first paper that deals the problem of outlier detection with application of DBCSAN for symbolic unbalanced data sets.

In the empirical part of this paper, ensemble learning and single DBSCAN algorithm with different distance measures is used to detect outliers in an unbalanced data sets. The paper presents also an impact of the parameters that are essential for DBSCAN on the outlier detection and partition quality (in terms of Silhouette index).

2 DBSCAN Ensemble for Symbolic Data

In the classical data, each object is described by a set of single-valued variables. Such situation allows to describe objects as a vector of quantitative or qualitative measurements where each column represents a variable. However, the classical approach may be too restrictive to represent more complex data. In order to take into consideration uncertainty and variability of the data, we must assume sets of categories or intervals with frequencies or weights. This kind of data representation has been studied in the Symbolic Data Analysis (SDA).

In symbolic data analysis, each symbolic object can be described by following variables (Bock and Diday 2000; Billard and Diday 2006; Diday and Noirhomme-Fraiture 2008; Noirhomme-Fraiture and Brito 2011):

1. Quantitative (numerical) variables:
    - numerical single-valued,
    - numerical multi-valued,
    - interval variables,
    - histogram variables.
2. Qualitative (categorical) variables:
    - categorical single-valued,
    - categorical multi-valued,
    - categorical modal.

Examples of symbolic variables with their realizations are presented in Table 1.

As the objects in the symbolic data analysis are described by non-classical variables, we can describe any type of phenomena in a more detailed way. However, symbolic data representation requires to apply special distance measures, methods, and algorithms that can deal with complex data.

DBSCAN algorithm that will be used in the empirical part of this paper has several advantages over traditional partitioning techniques, like possibility to detect non-spherical shapes of clusters, groups of different size, and robustness against outliers. But sometimes the DBSCAN algorithm can lead to large number of clusters and the interpretation of such results can be difficult. Some authors suggest to use clustering visualization methods as the support for DBSCAN (Nowak-Brzezińska and Xięski 2014).

Table 1 Examples of symbolic variables with realizations

Symbolic variable	Realizations	Variable type
1	*2*	*3*
Price of a new car (in PLN)	<27,000, 38,000>; <35,000, 50,000> <20,000, 30,000>; <25,000, 37,000>	Interval-valued (non-disjoint intervals)
Engine's capacity (in ccm)	<1000, 1200>; <1300, 1400> <1500, 1800>; <1900, 2200>	Interval-valued (disjoint intervals)
Chosen car color	{red, black, green, blue}	Categorical multi-valued

Preferred car	{Toyota (0.3); Volvo (0.7)}
	{Audi (0.6), VW (0.4), Skoda (0.05)}
		Categorical modal

Distance traveled	<10, 20> (0.65); <21, 30> (0.35)	Histogram
Sex	{M; F}	Nominal
Number of customers	(0, 1, 2, 3, …)	Ratio

Source Own elaboration

The DBSCAN algorithm for symbolic data requires to select two initial parameters: ε and *minPts*. The ε parameter controls how similar are the objects in the same group. The *mintPts* is the minimal number of objects that are needed to form a cluster. The *minPts* value can be derived from the number of dimensions in the data set (D) as *minPts* $\geq D + 1$. If *minPts* $= 1$, every data point will be a cluster. When *minPts* ≤ 2, the results will be the same as for hierarchical clustering with the single metric, with the dendrogram cut at height of ε. So *minPts* should be at least 3. Larger minPts values are useful for data sets with noise and larger data sets. In general, minPts should be equal or greater than data dimensionality (number of variables). Sander et al. (1998) suggest to use *minPts* that is twice bigger than number of variables.

The ε can be found by using a k-distance graph (see Sander et al. 1998 for further details) and plotting the distance to the $k =$ *minPts*. If ε are too small, a large part of the data set will not be clustered. Too large values will merge almost all data points.

There are also other proposals in the literature that deal with the problem of parameter selection for DBSCAN. Some of them use differential evolutions, some propose to detect sharp distance increases generated by a function which computes a distance between each element of a data set and its kth nearest neighbor others propose to use some other clustering algorithm as the initial tool (see, for example, Starczewski et al. 2020; Karami and Johansson 2014; Chen et al. 2019).

For the DBSCAN algorithm also a suitable distance measure for symbolic data is needed. In the paper, three of them will be compared (Gatnar and Walesiak 2011):

1. Normalized Ichino-Yaguchi distance (unweighted):

$$d(A_i, A_k) = \sqrt[q]{\sum_{j=1}^{m} \psi\left(v_{ij}, v_{kj}\right)^q}, \tag{1}$$

   where $\psi\left(v_{ij}, v_{kj}\right) = \frac{\varphi\left(v_{ij}, v_{kj}\right)}{|V_j|}$ with $\varphi\left(v_{ij}, v_{kj}\right) = \left|v_{ij} \oplus v_{kj}\right| - \left|v_{ij} \otimes v_{kj}\right| + \gamma\left(2 \cdot \left|v_{ij} \oplus v_{kj}\right| - \left|v_{ij}\right| - \left|v_{kj}\right|\right)$, v_{ij}, v_{kj}—symbolic variables, $\oplus$—Cartesian sum, $\otimes$—Cartesian product, $||$—length of a symbolic interval-valued variable or number of elements in a symbolic categorical multi-valued variable, V_j—domain of a symbolic variable, $\gamma \in \left[0, \frac{1}{2}\right]$.
2. Normalized de Carvalho distance based on description potential:

$$d(A_i, A_k) = \frac{\left[\pi|A_i \oplus A_k| - \pi|A_i \otimes A_k| + \gamma(2\pi|A_i \oplus A_k| - \pi|A_i| - \pi|A_k|)\right]}{\pi\left(A^E\right)}, \tag{2}$$

   where: A^E—maximum object according to the description potential, $\pi(A_i)$—description potential of a symbolic object, other elements like in Eq. 1.
3. Second normalized de Carvalho distance that is based on description potential:

$$d(A_i, A_k) = \frac{\left[\pi|A_i \oplus A_k| - \pi|A_i \otimes A_k| + \gamma(2\pi|A_i \oplus A_k| - \pi|A_i| - \pi|A_k|)\right]}{\pi(A_i \oplus A_k)}, \tag{3}$$

where all elements as in Eq. 3.

Other distances for symbolic data are described in Bock and Diday (2000), Gatnar and Walesiak (2011).

The ensemble learning, in general, means aggregation of results of many different models into one model that reaches better results. Such an idea was successfully applied both in supervised and unsupervised approaches for classical and symbolic data. However, this idea can be also used to detect outliers (see, e.g., Aggrawal and Sathe 2017). For all distance measure computations the R package symbolicDA will be applied (see: Walesiak et. al. 2018).

In this paper, two different symbolic distance measures, *minPts*' and ε*,* are used to obtain one ensemble model that allows to detect outliers in more precise way. The results of the ensemble model will be compared to single models.

3 Simulations and Their Results

To check if the DBSCAN for symbolic data can be a suitable tool for outlier detection in real data sets, unbalanced symbolic data sets were prepared for experiments with application of `cluster.Gen` function from `clusterSim` algorithm (Walesiak and Dudek 2020):

1. Data set I that contains 100 symbolic objects in three elongated clusters of equal size and 10 outliers in two dimensions. The observations in each cluster are independently drawn from bivariate normal distribution with means $(0, 0)$, $(1.5, 7)$, $(3, 14)$ and covariance matrix $\sum(\sigma_{jj} = 1, \sigma_{jl} = -0.9)$.
2. Data set II that contains 190 symbolic objects and 20 outliers in four clusters of following sizes $(70, 40, 30, 30)$ that are described by three variables. The observations are drawn from multivariate normal distribution $(-4, 5, -4)$, $(4, 14, 5)$, $(14, 5, 14)$, $(5, -4, 5)$ and identity variance matrix Σ, where $\sigma_{jj} = 1(1 \le j \le 3)$ and $\sigma_{jl} = 0(1 \le j \ne l \le 3)$.
3. Data set III that contains 180 observations five clusters of following sizes $(20, 30, 40, 50, 50)$ that are not well separated and 20 outliers in two dimensions. The observations are independently drawn from bivariate normal distribution with means $(5, 5)$, $(-3, 3)$, $(3, -3)$, $(0, 0)$, $(-5, -5)$ and identity covariate matrix $\sum(\sigma_{jj} = 1, \sigma_{jl} = 0.9)$.
4. For the DBSCAN algorithm, the following initial parameters have been assumed:
5. As we have two or three variables in the data sets then *minPts* will be equal to 3, 5, and 7.
6. The ε was selected by using a k-distance graph (see Sander et. al. 1998) and plotting the distance to the $k = minPts$.

7. Table 2 presents the results for the *minPts* values for all distances that were taken into consideration in this research.

The larger the number of *minPts* parameter in the model, the greater the number of detected outliers and better clustering quality. Similar results were reached for classical data by Nowak-Brzezińska and Xsięski (2017), p. 65.

Table 2 Results of simulations—*minPts* parameter

Dataset	minPts	3	5	7
Normalized Ichino-Yaguchi distance				
I	No. of clusters	2	3	3
	No. of outliers	out of 10	8 out of 10	9 out of 10
	Clustering quality*	0.3564	0.4776	0.5662
II	No. of clusters	3	4	4
	No. of outliers	11 out of 20	17 out of 20	17 out of 20
	Clustering quality*	0.3112	0.4432	0.5201
III	No. of clusters	4	5	6
	No. of outliers	5 out of 15	10 out of 15	11 out of 15
	Clustering quality*	0.2212	0.3555	0.4110
Normalized de Carvalho distance based on description potential				
I	No. of clusters	2	3	3
	No. of outliers	6 out of 10	9 out of 10	9 out of 10
	Clustering quality*	0.4356	0.5680	0.5680
II	No. of clusters	3	4	4
	No. of outliers	12 out of 20	18 out of 20	18 out of 20
	Clustering quality*	0.3523	0.4767	0.4767
III	No. of clusters	4	6	6
	No. of outliers	7 out of 15	10 out of 15	12 out of 15
	Clustering quality*	0.4405	0.4703	0.5309
Second version of the normalized de Carvalho distance based on description potential				
I	No. of clusters	2	3	3
	No. of outliers	6 out of 10	9 out of 10	9 out of 10
	clustering quality*	0.4457	0.5774	0.6102
II	No. of clusters	3	4	4
	No. of outliers	11 out of 20	18 out of 20	18 out of 20
	Clustering quality*	0.3106	0.4817	0.4817
III	No. of clusters	4	6	6
	No. of outliers	7 out of 15	10 out of 15	12 out of 15
	Clustering quality*	0.4612	0.4745	0.5509

*—measured by Silhouette index
Source Own elaboration

In the case of the *minPts* parameter normalized de Carvalho distances perform usually better, in terms of clustering quality, than normalized Ichino-Yaguchi distance.

Table 3 presents the results for the ε parameter for all distances.

If the parameter ε is small then larger number of outliers is being detected and better clustering quality is archived in general. The choice of a distance measure is quite important and usually normalized de Carvalho distances perform better than normalized Ichino-Yaguchi distance.

Table 4 presents the results for aggregated model.

When considering an ensemble model for each datasets, *minPts* and ε values, we can see that aggregated models allow to detect more existing outliers and their quality is also better.

4 Final Remarks

The DBSCAN algorithm can be easily applied for the symbolic data case. The only thing that differs it from the classical version of the DBSCAN is the distance measure for symbolic data.

When looking at the DBSCAN's parameters—*minPts* (minimal number of data points to form a cluster) and ε (maximum distance for objects in a cluster), both have significant impact on the clustering results—both in terms of clustering quality and number of outliers that have been detected. Higher *minPts* values lead to larger number of outliers in the data and also to higher clustering quality (that was measured by Silhouette index). However, higher ε values lead to lower number of clusters and less outliers and usually worse clustering quality. So selection of initial parameters can lead to different clustering results. Similar results were obtained by Nowak-Brzezińska and Xsięski (2014) for classical data sets with outliers.

Ensemble approach that uses information from different models (three different distance measures, three different *minPts* values, and three different ε values) allows to save some time on initial parameters tuning and as the most important fact it leads to better clustering results in terms of clustering quality and number of detected outliers in unbalanced datasets.

Table 3 Results of simulations—ε parameter

Dataset	ε	0.2873	0.6073	0.9021
Normalized Ichino-Yaguchi distance				
I	No. of clusters	3	3	3
	No. of outliers	6 out of 10	4 out of 10	4 out of 10
	Clustering quality*	0.5456	0.4243	0.4375
II	No. of clusters	4	4	3
	No. of outliers	15 out of 20	11 out of 20	10 out of 20
	clustering quality*	0.5082	0.4553	0.4023
III	No. of clusters	5	4	4
	No. of outliers	12 out of 15	10 out of 15	9 out of 15
	Clustering quality*	0.5210	0.4521	0.3532
	ε	0.4032	0.6443	0.9827
Normalized de Carvalho distance based on description potential				
I	No. of clusters	3	3	3
	No. of outliers	7 out of 10	5 out of 10	4 out of 10
	Clustering quality*	0.6532	0.5102	0.4874
II	No. of clusters	4	4	3
	No. of outliers	16 out of 20	13 out of 20	11 out of 20
	Clustering quality*	0.6081	0.4652	0.4001
III	No. of clusters	5	4	4
	No. of outliers	13 out of 15	11 out of 15	9 out of 15
	Clustering quality*	0.5287	0.4987	0.3493
	ε	0.3002	0.6758	0.9928
Second version of the normalized de Carvalho distance based on description potential				
I	No. of clusters	3	3	3
	No. of outliers	8 out of 10	6 out of 10	5 out of 10
	Clustering quality*	0.6232	0.5457	0.5108
II	No. of clusters	4	4	3
	No. of outliers	15 out of 20	13 out of 20	10 out of 20
	Clustering quality*	0.5083	0.4655	0.4027
III	No. of clusters	5	4	4
	No. of outliers	13 out of 15	12 out of 15	9 out of 15
	clustering quality*	0.5297	0.5211	0.3490

*— measured by Silhouette index
Source Own elaboration

Table 4 Results of simulations—distances and all models

Aggregated model	I	II	III
No. of clusters	3	4	6
No. of outliers	9 out of 10	17 out of 20	13 out of 15
Clustering quality*	0.6732	0.5476	0.6927

*—measured by Silhouette index
Source Own elaboration

References

Aggrawal CC (2013) Outlier analysis. Springer

Aggrawal CC (2015) Data mining—the textbook. Springer

Aggrawal CC, Sathe S (2017) Outlier ensembles—an introduction. Springer

Ankrest M, Breunig M, Kriegel H-P, Sander J (1999) OPTICS: ordering points to identify the clustering structure. In: ACM SIGMOD international conference on management of data, pp 49–60

Bock H-H, Diday E (eds) (2000) Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data. Springer, Berlin-Heidelberg

Billard L, Diday E (2006) Symbolic data analysis. conceptual statistics and data mining. Wiley, Chichester

Compello R, Moulavi D, Sander J (2013) Density-based clustering based on hierarchical density estimates. Adv Knowl Discov Data Min 160–172

Chen S, Liu X, Ma J, Zhao S, Hou X (2019) Parameter selection algorithm of DBSCAN based on K-means two classification algorithm. J Eng 23:8676–8679

Diday E, Noirhomme-Fraiture M (2008) Symbolic data analysis and the SODAS software. John Wiley & Sons, Wiley, Chichester

Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd ACM international conference on knowledge discovery and data mining, Portland, pp 226–231

Gatnar E, Walesiak M (red.) (2011) Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R. C.H. Beck, Warszawa

Hawkins D (1980) Identification of outliers. Chapman and Hall

Hinneburg A, Keim D (1998) An efficient approach to clustering in large multimedia databases with noise. In: Proceedings of the 4th international conference on knowledge discovery and data mining, pp 58–65

Jahirabadkar S, Kulkarni P (2013) Clustering for high dimensional data: density based subspace clustering algorithms. Int J Comput Appl 63(20):29–35

Kailing K, Kriegel H-P, Kröger P (2004) Density-connected subspace clustering for high-dimensional data. In: Proceedings of SIAM internatinal conference on data mining, pp 246–257

Karami A, Johansson R (2014) Choosing DBSCAN parameters automatically using differential evolution. Int J Comput Appl 91(7):1–11

Noirhomme-Fraiture M, Brito P (2011) Far beyond the classical data models: symbolic data analysis. Stat Ana D Min 4(2):157–170

Nguyen H, Ang H, Gopalakrishnan V (2010) Mining ensembles of heterogeneous detectors on random subspaces. In: International conference on database systems for advanced applications, pp 368–383. Springer, Berlin, Heidelberg

Nowak-Brzezińka A, Xięski T (2014) Exploratory clustering and visualization. Procedia Comput Sci 35C:1082–1091

Sander J, Ester M, Kriegel H-P, Xu X (1998) Density-based clustering in spatial databases: the algorithm gdbscan and its applications. Data Min Knowl Disc 2(2):169–194

Sheikholeslami G, Chatterjee S, Zhang A (1998) Wavecluster: a multi-resolution clustering approach for very large spatial databases. In: Proceedings of the 24th VLDB conference, pp 428–439

Starczewski A, Goetzen P, Er MJ (2020) A new method for automatic determining of the DBSCAN parameters. J Artif Intell Soft Comput Res 10(3):209–221

Walesiak M, Dudek A (2020) The clusterSim package for R software. www.r-project.org

Walesiak M, Dudek A, Pełka M (2018) The symbolicDA package for R software. www.r-project.org

Zhang J (2013) Advances of outlier detection: a survey. ICST Trans Scalable Inf Syst 13(1):1–26