Similarity Measure Selection for Categorical Data Clustering

Guilherme Alves, Miguel Couceiro, Amedeo Napoli

Similarity Measure Selection for Categorical Data Clustering

Guilherme Alves, Miguel Couceiro, Amedeo Napoli

Université de Lorraine, CNRS, Inria Nancy G.E., LORIA
{guilherme.alves-da-silva, miguel.couceiro}@inria.fr, amedeo.napoli@loria.fr

Abstract. Data clustering is a well-known task in data mining and it often relies on distances or, in some cases, similarity measures. The latter is indeed the case for real world datasets that comprise categorical attributes. Several similarity measures have been proposed in the literature, however, their choice depends on the context and the dataset at hand. In this paper, we address the following question: *given a set of measures, which one is best suited for clustering a particular dataset?* We propose an approach to automate this choice, and we present an empirical study based on categorical datasets, on which we evaluate our proposed approach.

1 Introduction

Real-world problems that are related to people generate a tremendous amount of data, e.g., analysis of customer data, people analytics, health applications. Most of the datasets are of mixed-type, which means that they comprise different types of attributes, e.g., numerical and categorical (Šulc and Řezanková, 2019). To help in solving some of these problems, data mining techniques are often employed to obtain useful information. Clustering is one of such well-known data mining techniques, which aims to group data in order to find patterns, to summarize information, and to arrange it (Barioni et al., 2014).

The clustering process often relies on distances or, in some cases, similarity measures. Due to the key role of these measures, different similarity functions for categorical data have been proposed (Boriah et al., 2008). However, there is no measure that performs best in all cases. The choice of a suitable similarity measure depends on the context and the dataset characteristics.

Several approaches for automating the choice of algorithms have been proposed (Pimentel and de Carvalho, 2019) (Abdulrahman et al., 2018). Meta-learning has been employed to build models based on dataset characteristics and algorithm effectiveness/efficiency. The ideia is to use the system's past experiences to automatically select an algorithms suitable to a given dataset (Brazdil et al., 2008). However, to our knowledge, no research work has presented a strategy to automate the choice of similarity measures.

In this paper, we thus address the following question: *given a set of measures, which one is best suited for clustering a particular dataset?* We propose a strategy to automatically select similarity measures for clustering categorical data.

Organization of the Paper. This paper is organised as follows. In Section 2 we recall the main concepts needed throughout this work. In Section 3 we propose an approach and in

Section 4 we present the experimental setup and discuss our empirical results. Conclusions and perspectives of future work are given in Section 5.

2 Background and Literature Review

Meta-learning A meta-learning system exploits knowledge obtained from previous experiences (Brazdil et al., 2008). In the context of selection of algorithms, the algorithm effectiveness (based on a specific evaluation metric or a set of metrics) measured over a dataset can be seen as a previous experience. In order to represent the previous experiences, it is necessary to build a dataset named *meta-dataset*. Each instance of this set is a *meta-instance*, which is composed of features extracted from the original data. The extracted features for a meta-dataset are called *meta-features* (or meta-attributes). We assign to each meta-instance one class that represents the target, reffered here as *meta-target*. In our scenario, the meta-target is the similarity measure that provides the satisfactory results (according to the evaluation metric).

One practical example of the employment of meta-learning is the clustering algorithm selection (or recommendation). To recommend clustering algorithms, Pimentel and de Carvalho (2019) proposed an approach to extract meta-features by combining the distances and the Spearman's rank correlation coefficient among pairs of (numerical data only) instances. The authors build histograms from pairs of instances and use them to train a classifier, referred here as *meta-learner*. Then, the obtained model is able to predict which algorithm should be employed when a new dataset arrives.

Categorical Data Clustering. Categorical data clustering, or clustering of nonnumerical data, is in concern with a special case of the problem of partitioning a set of instances into groups where instances are defined over categorical attributes. In this case, there is a lack of metric space and there is no single ordering for the categorical values (Andritsos and Tsaparas, 2017). One way to employ traditional clustering algorithms over categorical data is by using binary transformation. However, this leads to loss of information and may affect the interpretability and explainibility of the final outcome. To cope with this issue, various research works have proposed different strategies without the necessity of transforming categorical data into numerical values.

The traditional clustering algorithm *k-means* has been extended to deal with categorical attributes. *K-modes* is the *k-means* extension algorithm proposed in (Ahmad and Dey, 2007) for clustering mixed-data. The authors introduced a distance measure for mixed-data and changed the cluster center description to cope with the numeric data only limitation of *k-means* algorithm. For clustering purely categorical data, the algorithm ROCK is proposed in (Guha et al., 2000) using an agglomerative hierarchical approach based on links. Recently, Ahmad and Khan (2019) has presented an extensive survey of state-of-the-art of mixed-data clustering algorithms.

Similarity Measures for Categorical Data. Similarity measures are used to quantify the similarity between two data instances. Several measures have been proposed in the literature and employed in different tasks. For example, Boriah et al. (2008) compare different measures in the context of outlier detection. Some of them are adopted in this work. Alamuri et al. (2014) propose a taxonomy of several similarity measures and a taxonomy for clustering approaches for categorical data. dos Santos and Zárate (2015) analysed the performance of different similarity measures (most of them used herein) with TaxMap clustering algorithm in

a pairwise approach to be able to identify a measure that contains characteristics which offers more stability and also provides satisfactory results.

Recently, Nguyen et al. (2019) presented two new similarity measures based on the variability of the data, the Variable Entropy (VE) and Variable Mutability (VM), and compared them against eleven measures for clustering data using hierarchical approach. In both research works, the authors argue that the performance of algorithms that use the similarity measures depend on the dataset characteristics.

In order to describe the measures used in this work, let $X = [x_i]_{i=1}^{n}$ be the data matrix, where n is the number of rows (data instances). Each instance x_i is described by m attributes, such that, x_{ik} denotes the value of the k-th attribute, A_k. Let $S(x_i, x_j)$ be the similarity between the data instances x_i and x_j. To compute $S(x_i, x_j)$, we first compute the similarity for each variable $S_k(x_{ik}, x_{jk})$ based on the matching $(x_{ik} = x_{jk})$ or mismatching $(x_{ik} \neq x_{jk})$. We then aggregate them using the formulas:

$$S'(x_i, x_j) = \frac{1}{m} \sum_{k=1}^{m} S_k(x_{ik}, x_{jk}) \tag{1}$$

and

$$S''(x_i, x_j) = \frac{\sum_{k=1}^{m} S_k(x_{ik}, x_{jk})}{\sum_{k=1}^{m} \log p(x_{ik}) + \log p(x_{jk})}. \tag{2}$$

The fourth column of Table 1 indicates which function is used for aggregating the similarities of all attributes.

Before providing a computed matrix for a clustering algorithm, we need to convert the similarity $S(x_i, x_j)$ into the dissimilarity measure $D(x_i, x_j)$, to be able to use it in a clustering algorithm for example. We thus employ the following formulas:

$$D'(x_i, x_j) = 1 - S(x_i, x_j) \tag{3}$$

and

$$D''(x_i, x_j) = \frac{1}{S(x_i, x_j)} - 1. \tag{4}$$

The fifth column of Table 1 indicates which function is used for converting each similarity measure. The combinations shown in the fourth and fifth columns in Table 1 are predefined by each similarity measure. These combinations depend on the range of each per-attribute measure, which oblige distinct functions to aggregate and convert them in a dissimilarity measure. Eq. 1 gives the same weight for all per-attribute similarity and Eq. 2 uses probability estimation for assigning different weight. Eq. 3 is applied for similarity measures that take values from zero to one and Eq. 4 is applied for the measures that exceed the value one.

We also need the following notation from (Boriah et al., 2008) to define similarity measures: $f(x)$ is the number of times that the value x appears in the attribute A_k (k-th attribute), $p_k(x)$ is the probability of attribute A_k takes the value x and it is calculated according to the formula $p_k(x) = \frac{f(x)}{n}$, $p_k^2(x)$ is another form for computing the probability of attribute A_k to take the value x and it is calculated as $p_k^2(x) = \frac{f(x)(f(x)-1)}{n(n-1)}$, and n_k is the number of distinct values taken by the k-th attribute, A_k.

	$S_k(x_{ik}, x_{jk})$			
Measure	$x_{ik} = x_{jk}$	$x_{ik} \neq x_{jk}$	$S(x_i, x_j)$	$D(x_i, xj)$
Overlap	1	0	Eq. 1	Eq. 3
ES	1	$\frac{n_k^2}{n_k^2+2}$	Eq. 1	Eq. 4
G1	$1 - \sum_{q \in Q} p_k^2(q)$	0	Eq. 1	Eq. 3
G2	$1 - \sum_{q \in Q} p_k^2(q)$	0	Eq. 1	Eq. 3
G3	$1 - p_k^2(q)$	0	Eq. 1	Eq. 3
G4	$p_k^2(q)$	0	Eq. 1	Eq. 3
IOF	1	$\frac{1}{1+\log f(x_{ik}) \log f(x_{jk})}$	Eq. 1	Eq. 4
LIN	$2 \log p(x_{ik})$	$2 \log(p(x_{ik}) + p(x_{jk}))$	Eq. 2	Eq. 4
LIN1	$\sum_{q \in Q} \log p(q)$	$2 \log \sum_{q \in Q} \log p(q)$	Eq. 2	Eq. 4
OF	1	$\frac{1}{1+\frac{n}{\log f(x_{ic})} \cdot \frac{n}{\log f(x_{jc}}}$	Eq. 1	Eq. 4

TAB. 1: Similarity measures. (Boriah et al., 2008)(Šulc and Řezanková, 2019)

Several dissimilarity measures have been proposed in different research works (Boriah et al., 2008) (Šulc and Řezanková, 2019). We focus on some of these measures. Table 1 shows the equations used to compute each similarity measure adopted herein. We briefly describe each measure here, but due to the lack of space, please refer to the references for more details.

G1, G2, G3 and G4 (Boriah et al., 2008) are similarity measures extended from the original Goodall's measure, introduced in (Goodall, 1966). The measures G3, G4 and LIN (Lin et al., 1998) consider the relative frequencies of observed categories. G1, G2, and LIN1 (an extension of LIN proposed in (Boriah et al., 2008)) are based on relative frequencies of selected categories. Hence, there are different definitions for the set Q in these measures. For LIN1, $\{Q \subseteq A_k : \forall q \in Q, p_k(X_k) \leq p_k(q) \leq p_k(Y_k)\}$, assuming $p_k(X_k) \leq p_k(Y_k)$. For measure G1, $\{Q \subseteq A_k : \forall q \in Q, p_k(q) \leq p_k(X_k)\}$, and for G2, $\{Q \subseteq A_k : \forall q \in Q, p_k(q) \geq p_k(X_k)\}$. The Inverse Occurrence Frequency (IOF) and the Occurence Frequency (OF) measures use absolute frequencies of the observed categories.

3 Proposed Approach

In this section, we present our approach to build a meta-learning system for automatically selecting similarity measures for clustering categorical data.

The goal is to exploit the previous cases of the employment of similarity measures for clustering categorical data to build a predictive model (meta-model). Thus, the main task is to build a meta-dataset, which is the input for a meta-learner, and then to build a predictive model. The main task leads us to define: (1) the meta-features that describe datasets, and (2) the evaluation of the previous experiences to determine a meta-target.

Meta-features. In order to describe the datasets to be used by a meta-learner, we adopted as meta-features the following classes of datasets characteristics: (i) characteristics that describe attributes, and (ii) characteristics that describe the nature of attributes. These meta-features are commonly used in the algorithm selection literature (Kalousis, 2002) (Brazdil et al., 2008) (Pimentel and de Carvalho, 2019). Related works conclude that there is a dependency between these characteristics and the task effectiveness when the measures are em-

Histogram	Intervals
Entropy of attributes	[0,0.2],(0.2,0.4],(0.4,0.6],(0.6,0.8],(0.8,1]
Skewness	$(-\infty, -1], (-1, -0.5], (-0.5, 0.5], (0.5, 1], (1, +\infty)$
Kurtosis	$(-\infty, -1], (-1, -0.5], (-0.5, 0.5], (0.5, 1], (1, +\infty)$

TAB. 2: Schema of meta-features represented by histograms.

ployed (Šulc and Řezanková, 2019). Therefore, we focus on the same characteristics to build our meta-dataset. It is also important to make sure that the meta-features can be computed quickly (Brazdil et al., 2008). We compute the following meta-features for each dataset:

— number of attributes,
— number of instances,
— dimensionality of the dataset: the proportion of attributes and the number of instances,
— histogram of attributes entropy,
— histogram of skewness of attributes,
— histogram of kurtosis of attributes.

Some meta-features are represented in the form of histograms with fixed number of bars, instead of being represented as real numbers. Ferrari and De Castro (2015) state that a histogram gathers most of the information needed to characterize a dataset. All histograms have five bars which encode different intervals. Table 2 shows the scheme for these meta-features, i.e., the interval for each bar of each histogram.

Meta-target. The meta-dataset construction is completed by including a meta-target value for each meta-instance. The idea is to map the dataset characteristics to the effectiveness of the similarity measures in the clustering process. In this step, we use the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) as an external criteria to decide whether a similarity measure provides the best match between the dataset class and the obtained clusters. For each dataset, we perform the clustering algorithm using all similarity measures shown in Table 1, since they are complementary and belong to different classes of measures. After, we evaluate the obtained clusters using ARI and then we take the measure that provided the highest ARI. Figure 1 shows the workflow for building one meta-instance of the meta-dataset.

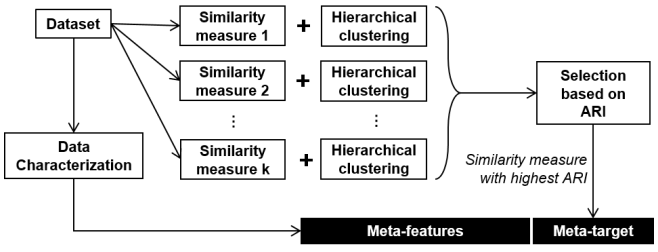

FIG. 1: Workflow for building one meta-instance.

Parameter	Value, set or interval
Number of attributes	{4,6,8,10}
Range of the size of clusters	[25,100]
Size of attribute domains	{2,3,4}, {2,3,5,6}, {6,8,10}
Number of noisy attributes	0
Number of outliers	0

TAB. 3: Parameters to generate the datasets employed in the experiments.

4 Experiments and Results

Experimental Setup. In order to evaluate our approach, the experiments were conducted using 60 datasets [1]. The datasets were generated by the *genRandomClust* function in the *clusterGeneration* R package. We varied the parameters function in order to generate datasets with different number of attributes, cluster structures, and attributes domain size. After the generation of the datasets, the numeric attributes were transformed into categorical attributes. The details concerning the parameters for generating the datasets are summarized in Table 3. The size of datasets ranges from 5 to 14 attributes and from 139 to 354 data instances. Together, all these datasets have labeled instances, which allow us to evaluate our experiments using ARI in order to assess the quality of the obtained clusters. As these are synthetic data, we can focus on how close the obtained clusters are to the predetermined labels. We noticed that there is no single similarity measure that performs better than all others in all datasets. The distribution of classes is presented here: G1=28, G3=9, ESKIN=7, IOF=6, OVERLAP=5, G2=4, OF=1.

We adopted the Hierarchical Clustering Algorithm (HCA) with average linkage for clustering data. Each dataset was clustered using HCA with all similarity measures mentioned in Table 1. We used the default parameters of HCA implementation available in Scikit-learn library (Pedregosa et al., 2011). Also, all experiments were implemented within the same platform using the Python programming language [2]. The experiments were run on an Intel Core i7 8th generation (2.11 GHz) with 16 GB of RAM, on Windows 7 x64.

During the execution of the experiments, we provided the actual number of clusters and we followed the leave-one-out protocol. Random Forest (RF), SVM, kNN (k=5), AdaBoost, Naive Bayes and Decision Tree (DT) were employed as meta-learners (Breiman, 2001), following the research works in algorithm selection literature (Brazdil et al., 2008) (Pimentel and de Carvalho, 2019). We evaluate the meta-learners performance using the accuracy over the predicted measures for each dataset.

Experimental Results. The main goal of these experiments was to assess the effectiveness of our proposed approach in comparison with: (1) the random choice of the similarity measure, (2) the ground truth, i.e., in order to have an upper bound, we compute the performance of an oracle that would always pick the measure with highest ARI, and (3) the similarity measure Overlap, which constitutes the simplest similarity between pairs of attributes (see Table 1). Table 4 shows the assessment in terms of accuracy of the built meta-model using different classifiers against the random strategy and the Overlap measure. We observe that the meta-learners trained over the meta-dataset outperform the random strategy and the Overlap measure. We

1. The datasets can be downloaded at: http://guilhermealves.eti.br/datasets
2. The source code is available at: http://github.com/asilvaguilherme/cat-sim-measure-selection

can also notice that Random Forest outperforms all the other supervised strategies, mainly because of the number of trees (n_estimators=50).

Concerning the meta-rules extracted from the Decision Tree from our meta-dataset, we noticed that the meta-features which represent the initial and final bars of the histograms appeared in the first nodes of the tree. This is expected because they show whether a dataset has several attributes with heavy-tailed or light-tailed relative to a normal distribution, maximum or minimum entropy, and distribution with positive or negative skew. These characteristics are relevant for several measures employed herein since they depend on the distribution of the domain of each attribute.

We also compared the quality of obtained clusters (see the second line of Table 4) by employing the predicted measure against the clusters obtained by using the best measure (Oracle), the randomly chosen measure (Random), and the simplest measure (Overlap). We can also notice that the predicted measures led us to get better clusters: using the meta-dataset to predict which similarity measure should be employed outperforms the Random approach and Overlap measure in terms of ARI. Therefore, the better quality of obtained clusters from the meta-model corroborates our research question.

TAB. 4: Prediction assessment (accuracy) and partition quality assessment (ARI).

	Random	**Overlap**	**Naive**	**DT**	**SVM**	**5NN**	**AdaBoost**	**RF**	**Oracle**
Accuracy	0.100	0.083	0.300	0.450	0.333	0.333	0.333	0.467	-
ARI	0.444	0.484	0.539	0.558	0.600	0.585	0.563	0.597	0.653

5 Conclusion

In this paper, we propose an approach for selecting similarity measures automatically for clustering categorical data. We demonstrate that our approach builds a meta-model that make better predictions and improve the obtained clusters.

This work opens avenues for future research. We can use the strategy of choosing a similarity measure to build an ensemble approach for clustering categorical data. It could also be extend to consider only internal criteria to compute the meta-target, therefore we would not need external information for evaluating the previous experiences. Finally, we should evaluate our approach using real datasets and also extend it to be able to work in a online setting.

References

Abdulrahman, S. M., P. Brazdil, J. N. van Rijn, and J. Vanschoren (2018). Speeding up algorithm selection using average ranking and active testing by introducing runtime. *Machine learning 107*(1), 79–108.

Ahmad, A. and L. Dey (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering 63*(2), 503–527.

Ahmad, A. and S. S. Khan (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access 7*, 31883–31902.

Alamuri, M., B. R. Surampudi, and A. Negi (2014). A survey of distance/similarity measures for categorical data. In *2014 International joint conference on neural networks (IJCNN)*, pp. 1907–1914. IEEE.

Andritsos, P. and P. Tsaparas (2017). Categorical data clustering. *Encyclopedia of Machine Learning and Data Mining*, 188–193.

Barioni, M. C. N., H. Razente, A. M. Marcelino, A. J. Traina, and C. Traina Jr (2014). Open issues for partitioning clustering methods: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4*(3), 161–177.

Boriah, S., V. Chandola, and V. Kumar (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*, pp. 243–254. SIAM.

Brazdil, P., C. G. Carrier, C. Soares, and R. Vilalta (2008). *Metalearning: Applications to data mining*. Springer Science & Business Media.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

dos Santos, T. R. and L. E. Zárate (2015). Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications 42*(3), 1247–1260.

Ferrari, D. G. and L. N. De Castro (2015). Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. *Information Sciences 301*, 181–194.

Goodall, D. W. (1966). A new similarity index based on probability. *Biometrics*, 882–907.

Guha, S., R. Rastogi, and K. Shim (2000). Rock: A robust clustering algorithm for categorical attributes. *Information systems 25*(5), 345–366.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of classification 2*(1), 193–218.

Kalousis, A. (2002). *Algorithm selection via meta-learning*. Ph. D. thesis, University of Geneva.

Lin, D. et al. (1998). An information-theoretic definition of similarity. In *Icml*, Volume 98, pp. 296–304. Citeseer.

Nguyen, T.-H. T., D.-T. Dinh, S. Sriboonchitta, and V.-N. Huynh (2019). A method for k-means-like clustering of categorical data. *Journal of Ambient Intelligence and Humanized Computing*, 1–11.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Pimentel, B. A. and A. C. de Carvalho (2019). A new data characterization for selecting clustering algorithms using meta-learning. *Information Sciences 477*, 203–219.

Šulc, Z. and H. Řezanková (2019). Comparison of similarity measures for categorical data in hierarchical clustering. *Journal of Classification 36*(1), 58–72.

Résumé

Le partitionnement de données est une tâche bien connue dans l'exploration de données et il repose souvent sur des distances ou, dans certains cas, des mesures de similarité. C'est notamment le cas pour les ensembles de données du monde réel qui comprennent des attributs catégoriels. Plusieurs mesures de similarité ont été proposées dans la littérature, mais leur choix dépend du contexte et de l'ensemble de données en cause. Dans ce travail, nous nous posons la question suivante : compte tenu d'un ensemble de mesures, laquelle est la mieux adaptée pour partitionner un jeu de données particulier ? Dans ce papier, nous proposons une approche pour automatiser ce choix, et nous présentons une étude empirique basée sur des ensembles de données catégorielles, sur lesquels nous évaluons l'approche proposée.