Proceedings of ICTAMI
International Conference on Theory and Applications in Mathematics and Informatics
17^{th}–20^{th} of September, 2015, Alba-Iulia, Romania
Computer Science Section

pp. 147-153

PARTITIONAL CLUSTERING METHODS WITH ORDINAL DATA

A.M. Coroiu

Abstract. One of the most important unsupervised technique of data mining is represented by cluster analysis. Cluster analysis means grouping similar objects and separating the dissimilar ones. Each object in the data set is assigned to a class in the clustering process using a distance measure. Partitioning method clustering is one of the major approaches in cluster analysis and the efficient partitioning of data sets is an important problem in this field of study.

Alike, in cluster analysis, one issue is constituted by using ordinal data sets. An ordinal attribute of a data is an attribute with possible values that have a meaningful order or ranking among them, but the value of the measure between successive values is not known. Ordinal attributes are useful for registering subjective assessments of qualities that cannot be measured objectively, therefore ordinal attributes are often used in studies for evaluations such as ratings. The challenge in using ordinal data sets is due to the fact that algorithms have to work with arbitrary similarity and with some measures of a distance which are suitable for this particular type of data. This paper presents the results gathered from performing some partitioning cluster algorithms to these ordinal variable and their efficiency in terms of clustering external evaluation.

2010 *Mathematics Subject Classification*: 68T10, 62H30

Keywords: partitioning cluster algorithms, cluster analysis, ordinal variables.

1. Introduction

The world in which we are living is full of data. Every day, people must have the ability to deal with a large amount of information and store or represent it as data, for further analysis and management. One of the vital means of dealing with these data is to classify or group them into a set of categories or clusters. Actually, as one of the most primitive activities of human beings, classification plays an important and indispensable role in the long history of human development.

In consideration of learning a new object or understanding the apparition of a new phenomenon, people always try to discover the features that can describe it, and after that to compare it with other known objects or phenomena, based on the similarity or dissimilarity, generalized as proximity, according to some certain standards or rules. In other words, classification systems are either supervised or unsupervised, depending on whether they assign new inputs to one of a finite number of discrete supervised classes or unsupervised categories, respectively [3]. Data mining is new technology/process of finding novel, hidden, interesting, and useful information, or knowledge from the large volumes of raw data [8]. This useful information or knowledge can be used to predict or to tell us something new.

2. Distance functions used in cluster analysis

The data represent an essential entity, but only if we know how to retrieve or extract useful data from the large volumes of raw data. Data mining technique helps us in accomplishing this. The most important technique of data mining is represented of clustering. This is a technique of grouping similar data objects together, so that the objects in each group (called clusters) share the same pattern of information. The domains in which the technique called cluster is used are countless: financial data classification, spatial data processing, satellite photo analysis, engineering and medical figure auto-detection or social network analysis.

There are two types of clustering techniques - partitioning and hierarchical clustering technique. In unsupervised classification, called clustering or exploratory data analysis, no labeled data are available. As pointed out Backer and Jain [1], in cluster analysis a group of objects is split up into a number of more or less homogeneous subgroups on the basis of an often subjectively chosen measure of similarity (chosen subjectively based on its ability to create interesting clusters), such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups.

Clustering algorithms partition data into a certain number of clusters (groups, subsets, or categories). The researchers describe a cluster considering the internal homogeneity and the external separation, i.e., patterns in the same cluster should be similar to each other, while patterns in different clusters should not.

Distance functions

Without claiming that we will present all of the similarity distance existed in literature, in the next section are briefly presented some distance that can be used when we have to deal with mixed (including ordinal) data set [2], [6]:

Cosine similarity this is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0 is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90 have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in a positive space, where the outcome is neatly bounded in [0, 1].

Jaccard similarity function this measure, Jaccard coefficient is one of the metrics used to compare the similarity and diversity of sample sets. It uses the ratio of the intersecting set to the union set as the measure of similarity. Thus it equals to zero if there are no intersecting elements and equals to one if all elements intersect [7].

Gower distance - this measure compute all the pairwise distances between observations in the data set. The original variables may be of mixed types. In this case, or whenever metric = Gower is set, a generalization of the Gower's formula is used [9].

3. Computational experiment

In the following sections of the paper are presented results achieved from performing some algorithms of cluster analysis on the ordinal data set. In terms of interpretation the comparisons for our results, we have used measures for internal and external evaluation of a clustering.

For external evaluation we have choose some methods of evaluation: pair counting measures, Bcubed based measures, editing distance measures [5]. In the following section, we will present the results achieved from our experiments within these evaluation methods.

For Pair counting measures we have gathered the following values indexes: Jaccard, Recall, Rand and Fowlkes Mallows. Jaccard Index is a similarity coefficient of two clustering of the same data set under the assumption that the two clustering are independent. For two clustering of the same data set, this function calculates the Jaccard similarity coefficient of the clustering from the co-memberships of the observations. The co-membership is defined as the pairs of observations that are clustered together; The Jaccard index has values between 0 - independent clustering and 1 identical clustering. Rand index is an index which correctly classified pairs of elements, its value between 0 and 1 reveals the fact that two partitions agree perfectly for 1 value or not at all for 0 value. FMI is an external evaluation method that is used to determine the similarity between two clustering.

The second evaluation methods on which we gathered values of indexes is BCubed-based measures - BCubed metrics decompose the evaluation process estimating the

precision and recall associated to each item in the distribution. The item precision represents how many items in the same cluster belong to its category. In the same time, the recall associated to one item represents how many items from its category appear in its cluster. In our case, our experiments outline a higher value of precision which means that the cluster achieved is more precision-oriented. A higher value of precision is achieved in our experiment

Finally the set matching based measures and editing distance measures this metric share feature of assuming a one to one mapping between clusters and categories. In our experiments, the indexes gathered from this measures method are purity which is one of primary validation measure to determine the cluster quality. A greater value of purity indicates good clustering. Purity serve out the noise in a cluster, but it does not reward grouping items from the same category together; if we simply make one cluster per item, we reach trivially a maximum purity value. In our experiments, the results shows a lower value of inverse purity and a higher value of purity, which means that the clustering methods return good results.

For internal evaluation of a clustering, the analyzed and achieved value is the silhouette index which validates clustering performance based on the pairwise difference of between and within cluster distances. In our experiments, the achieved values are within 0.50 and 0.70 and these values of silhouette index outline that reasonable structure has been found.

As stages in cluster analysis, we can count the next:

1. Decide on the clustering variables
2. Decide on the clustering procedure
3. Hierarchical methods
4. Partitioning methods
5. Select a measure of similarity or dissimilarity
6. Choose a clustering algorithm
7. Validate and interpret the cluster solution

For this experiment, data set used was Dermatology data set - a data set with 366 instances and 33 attributes and the attribute characteristics are categorical. The original data set can be access here: https://archive.ics.uci.edu/ml/datasets/Dermatology. In this data set, every feature (clinical and histopathological) was given a degree in the range of 0 to 3. 0 value indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

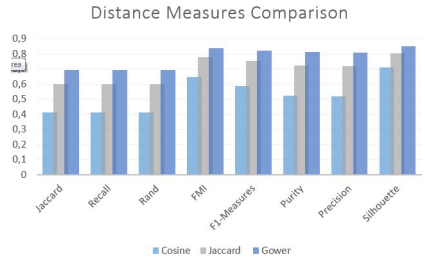
Table 1: Distance measures comparisons

Index \Distance	Cosine	Jaccard	Gower
Jaccard	0,41	0,60	0,69
Recall	0,41	0,60	0,69
Rand	0,41	0,60	0,69
FMI	0,65	0,78	0,84
F1-Measures	0,59	0,75	0,82
Purity	0,52	0,72	0,81
Precision	0,52	0,72	0,81
Silhouette	0,71	0,81	0,85

We have considered some clustering algorithms that have to its ability to handle the mixed data set, including ordinal data. For our experiments, we have used an open source data mining software written in Java called ELKI on which focus is research in algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection [4]. Nowadays for researchers, clustering validation represents essential issues in clustering application. And in the same time, using ordinal data set in different cluster analysis is a challenge. In this paper we have perform clustering algorithms on ordinal data set. The purpose of this was to establish which is the proper distance to be applied when we are working with ordinal data set.

Results interpretation. The final goal of clustering is to provide meaningful insights from the original data, so that they can effectively solve the problems encountered. After applying a clustering method on a data set, we want to assess how good the resulting clusters are. A number of measures can be used. Some methods measure how well the clusters fit the data set, while others measure how well the clusters match the ground truth, if such truth is available. There are also measures that score clustering and thus can compare two sets of clustering results on the same data set. The results achieved are in Table 1 and are highlighted in Figure 1

In terms of interpretation the comparisons for our results, we have used measures for internal and external evaluation of a clustering. In this case we have gather values for the following indexes: Jaccard, Recall, Rand, Fowlkes Mallows, F1-Measures, Purity and Precision for external evaluation of the clustering and Silhouette index for internal evaluation of a clustering.



We have used three distance measures (all of them can be suitable to be applied when we are using ordinal data set, and the results reveals that if we choose Jaccard similarity function we achieve good values of indexed used in cluster validity (evaluation).

4. Conclusions

The paper presented a timely concept, as far as we know. Cluster analysis still represents a field of interest for researchers from different domains and, furthermore, due to their characteristics and their manner of achieving, the ordinal data sets are also a particular importance.

The paper is a work in progress, so, as a future work directions we can enumerate the other distance to achieve more accurate results and some new approaches in order to find out more relevant information from our data.

References

[1] E. Backer, A. Jain, *A clustering performance measure based on fuzzy set decomposition*, IEEE Trans. Pattern Anal. Mach. Intell., vol.PAMI-3, no. 1, (1981), 6675.

[2] C. Charu, *Data Classification: Algorithms and Applications*, CRC Press, (2014).

[3] V. Cherkassky F. Mulier, *Learning From Data*, Concepts, Theory,and Methods, Wiley, (1998) 23-26.

[4] ELKI: Environment for Developing KDD-Applications Supported by Index-Structures, http://elki.dbs.ifi.lmu.de, (2014).

[5] C. Goutte, E. Gaussier, *A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation*, Xerox Research Centre Europe 6, (2004).

[6] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[7] D. Lin, *An information-theoretic definition of similarity*, ICML Vol. 98, (1998), 296-304.

[8] M. Muntean, H. Vlean, I. Ileana, C. Rotar, *Improving Classification with Support Vector Machine for Unbalanced Data*, Proceedings of 2010 IEEE International Conference on Automation, Quality and Testing, Robotics, (2010), 234-239.

[9] J. Podani, *Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions*, Journal of Vegetation Science, 16.5, (2005), 497-510.

Adriana Mihaela Coroiu
Department of Computer Science, Faculty of of Mathematics and Computer Science,
Babes-Bolyai Univeristy,
Cluj-Napoca, Romania
email: *adrianac@cs.ubbcluj.ro*