Journal of Modern Applied Statistical Methods

Volume 13 | Issue 1 Article 21

5-1-2014

Hierarchical Clustering with Simple Matching and Joint Entropy Dissimilarity Measure

A Mete Çilingtürk
Marmara University, Istanbul, Turkey, acilingi@marmara.edu.tr

Özlem Ergüt
Marmara University, Istanbul, Turkey, ozlem.ergut@marmara.edu.tr

DOI: 10.22237/jmasm/1398918000

Journal of Modern Applied Statistical Methods
May 2013, Vol. 13, No. 1, 329-338.

ISSN 1538 − 9472

Hierarchical Clustering with Simple Matching and Joint Entropy Dissimilarity Measure

A. Mete Çilingirtürk
Marmara University
Istanbul, Turkey

Özlem Ergüt
Marmara University
Istanbul, Turkey

Conventional clustering algorithms are restricted for use with data containing ratio or interval scale variables; hence, distances are used. As social studies require merely categorical data, the literature is enriched with more complicated clustering techniques and algorithms of categorical data. These techniques are based on similarity or dissimilarity matrices. The algorithms are using density based or pattern based approaches. A probabilistic nature to similarity structure is proposed. The entropy dissimilarity measure has comparable results with simple matching dissimilarity at hierarchical clustering. It overcomes dimension increase through binarization of the categorical data. This approach is also functional with the clustering methods, where a-priori cluster number information is available.

Keywords: Categorical data, clustering, dissimilarity, entropy

Introduction

Clustering analysis is a process used for classifying objects so that homogeneous subsets are built in heterogeneous groups. A variety of distance/similarity criteria are used when classifying objects in groups according to their similarity. One important criterion for choosing the distance or similarity measure, when classifying objects into groups, is the type of the data. In the literature it can be seen that most studies examine the clustering of continuous data. If the data set consists of continuous data, Euclid and Manhattan are the distance measures most widely used in applications. However, in a data set with categorical data it is not possible to use this type of distance measures. These variables are first transformed into binary data and then the analysis is applied, which increases the

A. Mete Çilingtürk is Professor of Statistics in the Department of Econometrics. Email him at acilingi@marmara.edu.tr. Özlem Ergüt is an Assistant in the Department of Economics. Email at ozlem.ergut@marmara.edu.tr.

number of dimensions when there are multinomial variables in the data. This procedure increases memory allocation.

There are different techniques and approaches for finding clusters with categorical data. One includes transformation of categorical variable into dummy variable, independently from calculation of distances. Applications of such hierarchical method algorithms are single linkage, complete linkage, average linkage, etc. (Chaturvedi et al., 2001).

Another approach uses the k-means algorithm for clustering of categorical data developed by Ralambondrainy (1995). In this approach multiple category attributes are turned into binary variables, which are assumed to be numeric variables and thus, the *k*-means algorithm is applied. The drawback of this approach is the increase in the number of binary variables when there are too many categories in variables. Further, cluster centres, given as 0 and 1, do not reflect the real characteristics of clusters (Huang, 1998). The basics of *K*-medoids algorithm is founded on finding *k* number of objects representative of several structural features of data. A Medoid is the most central point of the cluster with minimum average distance to other objects that are located in the same cluster (Kaufman and Rousseeuw, 2005; Xu and Wunsch, 2009). Due to the distance measure used in K-means algorithm, this method is not used in clustering categorical data. As the data set consists of categorical data, k-modes method, which is an extension of *k*-means model, is used for clustering categorical data, which was developed by Huang (1998). In this algorithm,

1. simple matching dissimilarity measure for categorical objects,
2. mod is used for clusters instead of mean,
3. frequency-based method is used for updating modes (Huang, 1998).

An extended–modes algorithm was proposed by Aranganayagi and Thangwell (2010), which uses a probability weighted single matching dissimilarity function.

Initially, expectation maximization algorithm assigns randomly different possibilities to each class or category. These probabilities are determined with consecutive iterations so as to maximize the similarity value of the data, which will also fit a pre-set number of clusters. The EM algorithm assumes that the model is suitable for a non-observable latent variable and that the stochastic model performs maximum likelihood estimations of the parameters (Agarwal et al., 2010). The optimization algorithm determines the convergence of the parameters.

ROCK (RObust Clustering using linKs) is an adaptation of the hierarchical clustering algorithm developed for clustering of categorical data. In this algorithm similarity value between two objects is calculated using Jaccard coefficient, then the threshold value (θ), defined between 0 and 1 by the researcher, is compared to decide adjacent points. In order that a given point q_i is adjacent to a point q_j for an i^{th} object in an m-dimensional space, similarity value has to exceed threshold value (θ) (Guha et al., 1999).

$$sim(q_i, q_j) \geq \theta$$

If this condition is met, it can be said that the points are neighbours. This algorithm classifies the objects into clusters according to their link ability. The link ability between two clusters gives the number of common adjacent points between q_i and q_j. The higher the linkability of q_i and q_j, the higher is the possibility of q_i and q_j being in the same cluster.

COOLCAT is proposed for categorical clustering analysis as an entropy-based algorithm (Barbara et al., 2002). The entropy-based algorithm consists of two steps, namely initialization and incremental steps. In the initialization step K most dissimilar records are selected from the sample. In the next step remaining records in the data set are assigned to appropriate clusters. The algorithm groups objects in the data set trying to minimize the expected entropy of the clusters. Similarly, He et al. (2005) maximized Ensemble algorithms with the average normalized mutual information [0,1] function based on entropy in separating of units with the purpose of categorical clustering.

Definitions and Notations

X and Y are two categorical objects defined by n and m attributes, the dissimilarity measure between X and Y is the sum of mismatches in relevant variable attributes of the two objects. The smaller the number of mismatches, the more similar are two objects. This measure is also a kind of generalized Hamming distance (Ng et al., 2007).

$$d(X,Y) = \sum_{k=1}^{m} \delta(x_k, y_k) \qquad (1)$$

$$\delta\left(x_k, y_k\right) = \begin{cases} 0\left(x_k = y_k\right) \\ 1\left(x_k \neq y_k\right) \end{cases} \tag{2}$$

As statistics is applied to physics, the development of statistical physics earned entropy new meanings with entropy, which is an indicator of the irregularity and uncertainty in a physical system. The increase in irregularity in the system is proportionate to the increase in entropy. The uncertainty of occurrence of x_i situation in system X, which is the entropy of x_i situation, is shown as $c\left(p_i\right) = -\log p\left(x_i\right)$, while the entropy of the system is expressed as (Roy, 2002; Müller, 2003)

$$H\left(X\right) = -\sum_{i=1}^{n} P\left(x\right) \log P\left(x\right). \tag{3}$$

As the logarithmic operations are performed, the entropy becomes an additive quantity for independent systems (Georgii, 2003).

For a given n, when $p\left(x_1\right) = p\left(x_n\right) = \cdots = p\left(x_n\right) = \frac{1}{n}$,

$$\begin{aligned} H_{max} &= -\sum \frac{1}{n} \log \frac{1}{n} = -n \frac{1}{n} \log \frac{1}{n} \\ H_{max} &= \log\left(n\right) \end{aligned} \tag{4}$$

is obtained. This means that H reaches its maximum value when it is equal to log(n). When a two-dimensional (X, Y) random variable is in question, P joint probability matrix becomes $P = \left\{p_{ij}\right\} = P\left(X = x_i, Y = y_j\right)$ and thus the entropy becomes

$$H\left(i, j\right) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log p_{ij}. \tag{5}$$

The uncertainty coefficient calculated asymmetrically and symmetrically based on entropy in cross-tables is more appropriate for use. The uncertainty coefficient for symmetric structures is calculated as

$$U\left(X,Y\right)=2\left[\frac{H\left(x_i\right)+H\left(y_j\right)-H\left(x_i,y_j\right)}{H\left(x_i\right)+H\left(y_j\right)}\right]. \tag{6}$$

Proposed Method

Taking observation units as variables, the proposed method ensures that calculation of combined entropy values remains on the same constant (log *m*) for *m* number of categorical attributes.

The ***S*** matrix, which shows that *n* number of objects take identical values, provides the basis of entropy dissimilarity measure approach, unlike the simple matching dissimilarity measures matrix. Each row/column in this matrix shows the number of similar objects for each *m* variables. Therefore each row/column of the matrix is the frequency distribution of its similarity with another observation. The uncertainty coefficient given in equation (6) aims that a single value is generated for a cross-table; thus, the formula has been organized with the help of the following equations with the purpose of measuring uncertainty based on entropy.

$$\begin{aligned} H\left(i.\right)&=p_{i.}\log\left(p_{i.}\right),\\ H\left(.j\right)&=p_{.j}\log\left(p_{.j}\right)\text{ and}\\ H\left(i,j\right)&=p_{ij}\log\left(p_{ij}\right). \end{aligned} \tag{7}$$

If *X* and *Y* are independent random variables, combined entropy is equal to the sum of the entropies of these two random variables

$$\begin{aligned} &H\left(i,j\right)=H\left(i.\right)+H\left(.j\right),\\ &U\left(i.j\right)=\begin{cases} 2\left[\dfrac{H(i)+H(j)-H(i,j)}{H(i)+H(j)}\right] & p_{ij}\neq 0\\ 2 & p_{ij}=0 \end{cases}. \end{aligned} \tag{8}$$

Equation (8) displays a symmetric dissimilarity matrix which does not consist of constant values: the reason for this is that the entropy of an object with itself depends on the frequency of encountering the characteristics in the total

distribution. The uncertainty of an object with frequently observable characteristics will be proportionately low. If two objects have no common features, $p_{ij} = 0$ and as logarithm is non-defined, maximum entropy dissimilarity value of −2 is used. However, as algorithm software used for clustering will accept a symmetric dissimilarity matrix with constant diagonal (0), $U(i,j)$ values are proportioned and corrected in 0-1 interval.

$$U^*(i,j) = \frac{U(i,j) - \text{diag}U(i,j)}{2 - \text{diag}U(i,j)} \tag{9}$$

The numerator of fraction brings the diagonal values, which are the smallest values of each row and column, to zero, whereas denominator proportions the dissimilarity of other values according to the maximum value and earns the value 1 for maximum dissimilarity.

Empirical Results

Dissimilarity matrices were formed based on simple matching dissimilarity measure and entropy in this article. The results obtained by using hierarchical methods in both dissimilarity matrices were compared with each other. The data used in the study was Teaching Assistant data obtained from UCI database (Loh, W. -Y. & Lim, T. -S., 1997). It was collected for evaluation of the performances of 151 research assistants at statistics department of Wisconsin-Madison University during three semesters and two summer schools. The scores were divided into 3 roughly equal-sized categories (low, medium, high) to form the class variable. The four variables chosen for determining the performance of 151 research assistants is:

1. Whether of not the TA is a native English speaker? (2 categories)
2. Course instructor (25 categories)
3. Course (26 categories)
4. Summer or regular semester (2 categories)

Within the scope of the study, Stata 11.0 program was used for application of hierarchical methods for entropy and simple matching dissimilarity measures. The results obtained from simple matching dissimilarity measure and hierarchical methods using single linkage, complete linkage and average linkage methods

were interpreted. In single linkage method, the two closest objects or clusters (minimum distance or biggest similarity) using distance/similarity values are combined. In complete linkage, the maximum of the distance between the new cluster formed after combining two clusters (objects) and the other cluster is taken. In the average linkage method, which is suggested as an alternative as it provides results between these two extreme techniques, the distance between two clusters is equal to the average values of the distances between observed couples located in two clusters.

One of the measures used in evaluating the success and quality of clustering results is F measure. This measure consists of a combination of precision and recall measures. F measure is basically the harmonic mean of precision and recall (Işık and Çamurcu, 2007). F measure, which is one of the measures that ensures (i) comparison of the classification which is known in advance and the clusters obtained as a result of clustering analysis (Loh and Shin, 1997) and (ii) evaluation of clustering, is calculated as follows for j.cluster and i.class.

$$F(i,j) = \frac{2 * r(i,j) * p(i,j)}{r(i,j) + p(i,j)}$$

where r means recall and p means precision.

$$r(i,j) = \frac{n_{ij}}{n_i} \; p(i,j) = \frac{n_{ij}}{n_j}$$

In n_{ij}, the number of observations in j.cluster and i.class, namely n_j and n_i, are respectively the magnitudes of j.cluster and i.class. Total F measure for a data set consisting of n number of observations is calculated as follows (Dalli, 2003):

$$F = \sum_i \frac{n_i}{n} max\left[F(i,j)\right]$$

If single linkage is used with simple matching dissimilarity measure, as there are considerable number of connections, observations are not classified into clusters and combined in a single cluster. In complete linkage method while observations are assigned to maximum three clusters; however, if average linkage method is preferred, observations can form maximum 34 clusters but the F

measure obtained in the case that there are three clusters is as, $F = 0{,}360$ with simple matching dissimilarity and $F = 0{,}387$ with entropy dissimilarity.

In single linkage, which is one of the three hierarchical methods, observations are classified into four clusters in the case that entropy dissimilarity is used. In the case that there are three clusters, 146 of the observations are assigned to the first cluster, four are assigned to the second cluster and one is assigned to the third cluster. In simple matching, dissimilarity observations cannot be classified into clusters, whereas clusters consisting of small number of observations occur in entropy dissimilarity. If the same measure is used in complete linkage method, as there are considerable number of connections, observations are not classified into clusters and combined in a single cluster. In average linkage method, however, observations are concentrated in the first cluster if there are three clusters.

Performance in the data set was evaluated in three categories namely good, mediocre and poor. The results obtained according to both dissimilarity measures and two were compared with these three categories, the level of concordance were determined. Accordingly,

In the average linkage method, 49 observations were correctly assigned (33 percent) if entropy dissimilarity measure was used.

In the average linkage method, 47 observations were correctly assigned (31 percent) if simple matching dissimilarity measure was used.

In the average linkage method, the F measure value obtained using simple matching dissimilarity, entropy measure were 0.36 and 0.38, respectively.

Conclusion

In categorical data, with the exception of data mining algorithms, clustering algorithms are applied with two-step clustering method and simple matching measure is used. Two-step clustering first digitalizes the categorical variables and then performs distance calculations. Parameter estimations require optimized solutions with iterations. The simple matching method however does not take the frequency of observing a certain characteristic in categorical variables and the possibility of a unit for having this unique characteristic into the consideration.

The selection of distance and/or similarity measure lies in the foundation of all clustering methods. The findings are based on the selection of both clustering

methods and distance measure. Therefore, this study offers an estimation of entropy matrix based on dissimilarity of categorical variables. The method also provides a solution to the problem of increase in the number of variables by using dummy variable in the case of existence of categorical variables. The study can also be used for developing a different clustering algorithm with a non-constant diagonal, which therefore will take into consideration the low level of uncertainty that is caused by having frequently encountered characteristics.

References

Agarwal, P., Alam, M. A., & Biswas, R. (2010). Analyzing the Agglomerative Hierarchical Clustering Algorithm for Categorical Attributes. *International Journal of Innovation, Management and Technology, 1*: 186-190.

Aranganayagi, S., & Thangavel, K. (2010). Extented K-modes with Probability Measure. *International Journal of Computer Theory and Engineering*, *2*: 431-435.

Barbara, D., Couto, J., & Li, Y. (2002). COOLCAT: An Entropy-based Algorithm for Categorical Data. In C. Nicholas (Chair). *Proceedings of the 11th International ACM Conference on Information and Knowledge Management*, McLean, VA, pp. 582-589.

Chaturvedi, A., Green, P. E. & Carroll, J. D. (2001). K-modes Clustering. *Journal of Classification, 18*: 35-55.

Dalli, A. (2003). Adaptation of the F-Measure to Cluster Based Lexion Quality Evalution. *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics,* Budapest, Hungary. Available at http://aclweb.org/anthology/W/W03/W03-2807.pdf

Georgii, H. (2003). Probabilistic Aspects of Entropy. In A. Greven, G. Keller & G. Warnecke (Eds.). *Entropy.* New Jersey: Princeton University Press, pp. 37-52.

Guha, S., Rastogi, R., & Shim, K. (1999). ROCK: A Robust Clustering Algorithm for Categorical Attributes. In M. Kitsuregawa, L. Maciaszek, M. Papazoglou & C. Pu (Eds.). *Proceedings of the 15th IEEE International Conference on Data Engineering*, Sydney, Australia. Available at theory.stanford.edu/~sudipto/mypapers/categorical.pdf.

He, Z., Xu, X., & Deng, S., (2005). A Cluster Ensemble Method for Clustering Categorical Data. *Information Fusion, 6*: 143-151.

Huang, Z. (1998). Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *304*: 283-304.

Işık, M., & Çamurcu, A. Y. (2007). K-Means, K-Medoids ve Bulanık C-Means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 6*: 31-45.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley and Sons.

Loh, W. -Y, & Shin, Y. -S (1997). Split Selection Methods for Classification Trees, *Statistica Sinica*, *7*: 815-840. Avilable at citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.7375[1].pdf

Loh, W. -Y. & Lim, T. -S. (1997). Teaching Assistant Evaluation Data Set. *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*. Irvine, CA: University of California, School of Information and Computer Science. Available at http://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation

Müller, I. (2003). Entropy: A Subtle Concept in Thermodynamics. In A. Greven, G. Keller & G. Warnecke (Eds.). *Entropy.* New Jersey: Princeton University Press, pp. 19-35.

Ng, M., Li, M. J., Huang, J. Z., & He, Z. (2007). On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm. *Pattern Analysis and Machine Intelligence, 29*: 503-507. Available at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.53.

Ralambondrainy, H. (1995). A conceptual version of the K-means algorithm. *Pattern Recognition Letters, 16*: 1147-1157.

Roy, B. N. (2002). *Fundamentals of Classical and Statistical Thermodynamics*. New York: John Wiley and Sons.

Xu, R., & Wunsch, D. (2009). *Clustering.* New York: John Wiley and Sons.