МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2020
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2020
Proceedings of the Forty-ninth Spring Conference
of the Union of Bulgarian Mathematicians
2020

USING ENTROPY OF CATEGORICAL DATA FOR CLUSTERING*

Chavdar Dangalchev

One of the most important tasks of cluster analysis is determining the number of clusters. In this article we suggest a method for it, using entropy of categorical data. Examples, illustrating the method, are given.

1. Introduction. Algorithms for clustering have been studied for a long time and there are many different algorithms [3]. Most of them are using only numerical data. If there is a categorical data, the well-known approach is to transfer it to numerical: if the values are in the same category then the distance is 0, otherwise it is 1. After that, the distance is multiplied by a weight, corresponding to the importance of the categorical data. There are several ways to do clustering only on categorical data. Some of them use entropy (e.g. [1]). Here we suggest another way to use categorical data – calculating the entropy of a category (or categories) in cluster sets with different number of clusters and selecting the set with minimal entropy. There are other algorithms for clustering of numerical data using entropy (e.g. [2]), but not for determining the number of clusters.

2. Algorithm. By using hierarchical clustering algorithm, we can change the threshold of the distance, in order to receive different sets of clusters. Another way is to have limit on the number of clusters or use elbow method [4] to determine the number of clusters. Similar considerations can be done if we use other algorithms, which produce different sets of clusters (e.g. DBSCAN) in which the number of clusters to be picked can be determined. Here we use the entropy of one or more categorical features:

– For every cluster we calculate the entropy as:

$$E = \sum_i \frac{N_i}{N} \log\left(\frac{N}{N_i}\right),$$

where N is the total number of points in the cluster and N_1, N_2, ... are numbers in each category.

– We take the sum of the entropy of the clusters, after weighting it by the number of points.

The problem of this simple approach is that when the threshold is 0 (every point is a separate cluster) then the entropy is also 0. We would like to select a set of minimal number of clusters while also keeping the entropy low. The number of clusters as a function of the threshold decreases; the entropy as a function of threshold increases.

*2020 **Mathematics Subject Classification:** Primary: 62H30, Secondary: 68T10.
Key words: clustering, entropy, number of clusters.

Both functions have different unit measures. To combine them we need to normalize them. We put a weight (W) for the number of clusters and calculate the score by adding the entropy. Here we suggest one possible way to select the weight of the number of clusters: the weight is calculated by assuming equal scores when clusters are N and 1:

– N clusters (when the threshold is equal to 0) the entropy is 0 and the score is $N * W$.

– One cluster with entropy E (when the threshold is the maximal one): the score is $W + E$.

– From $NW = W + E$ we receive:

$$W = \frac{E}{N-1}$$

and therefore when the clusters are N and 1 the scores S are equal:

$$S = \frac{NE}{N-1} = NW = W + E.$$

– We calculate the score $S(n)$ for every cluster set S_n with n clusters:

$$S(n) = nW + E(S_n)$$

and select the number of clusters n with the minimal score.

3. Examples. Let us have 12 points with coordinates on two numerical features (see Fig. 1):

$P_1 : (0, 2)$, $P_2 : (0, 3)$, $P_3 : (1, 2)$, $P_4 : (1, 3)$, $P_5 : (1, 0)$, $P_6 : (2, 0)$,
$P_7 : (3, 0)$, $P_8 : (4, 1)$, $P_9 : (4, 2)$, $P_{10} : (5, 2)$, $P_{11} : (5, 3)$, $P_{12} : (6, 3)$.

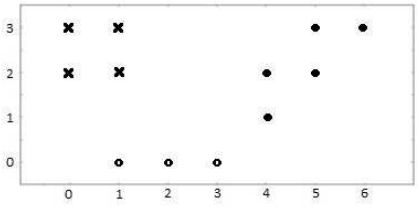

Fig. 1. Clusters count is 3

There are 4 different possible thresholds with Euclidean distance, respectively: 0, 1, 1.5 ($> \sqrt{2}$) and 2.

There are 12 clusters – every point is a cluster with threshold 0.

There are 3 clusters with threshold 1 (shown on Fig. 1):

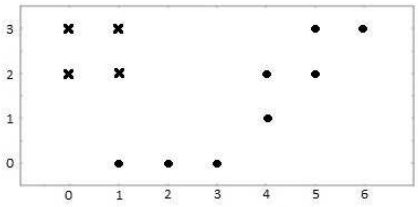

Fig. 2. Clusters count is 2

$C_1 = \{P_1, P_2, P_3, P_4\}$; $C_2 = \{P_5, P_6, P_7\}$; $C_3 = \{P_8, P_9, P_{10}, P_{11}, P_{12}\}$.

There are two clusters: cluster 2 and cluster 3 from Fig. 1 merge into one cluster with threshold 1.5.

There is only one cluster with threshold 2. These are all possible cluster sets.

Now let us use the entropy. Let there be a category "Size": Cluster 1 (from Fig. 1) has Size = "Small", Cluster 2 has Size = "Medium", Cluster 3 has Size = "Big".

Using logarithms with base 2, the entropy with threshold 2 (only 1 cluster) is:

$$E = \frac{4}{12}\log\left(\frac{12}{4}\right) + \frac{3}{12}\log\left(\frac{12}{3}\right) + \frac{5}{12}\log\left(\frac{12}{5}\right) = 1.554585.$$

The weight for the number of clusters is:

$$W = \frac{E}{12-1} = 0.141326$$

The scores with thresholds 0 and 2 are:

$$S(12) = 12 * 0.141326 = 1.696; \quad S(1) = E + 0.141326 = 1.696.$$

Calculating the other two scores (thresholds 1 and 1.5), we receive:

$$S(3) = 3 * 0.141326 = 0.424 \text{ (the entropy in every cluster is 0).}$$

$$S(2) = 2 * 0.141326 + \frac{3}{8}\log\left(\frac{8}{3}\right) + \frac{5}{8}\log\left(\frac{8}{5}\right) = 1.237.$$

By selecting the minimal score, we select the threshold 1 with 3 clusters.

We will receive the same result (3 clusters) even when the category "Size" does not strictly follow the 3 clusters distribution: e.g. P_3 is of size "Medium" and P_7 is of size "Big".

Let us create a new category "Age" from the old category "Size". Category "Age" has two values: "Kids" and "Adults". The "Kids" corresponds to size "Small". The value "Adults" combines sizes "Medium" and "Big". The calculations for clusters' scores are changing:

$$E = \frac{4}{12}\log\left(\frac{12}{4}\right) + \frac{8}{12}\log\left(\frac{12}{8}\right) = 0.918.$$

$$W = \frac{E}{12-1} = 0.08348.$$

$$S(12) = 0.08348 * 12 = 1.002; \; S(1) = 0.918 + 0.08348 = 1.002.$$

$$S(3) = 3 * 0.08348 = 0.250; S(2) = 2 * 0.08348 = 0.167.$$

The minimal score is when the threshold is 1.5 and has 2 clusters. Again, the same result is true if there are *small* changes of the categorical values.

4. Conclusion. Entropy of categorical data can be very useful in automation: for every data set we could have output cluster set, which follows the topology of the data and the category distribution. The number of clusters should not be given in advance. Yes, we can use different techniques to determine the cluster set (e.g. the elbow method), but this one is the method which can ensure synchronization with categorical data.

Acknowledgement. This work was done while the author was contracting for Microsoft (Azure BI) through Pactera Technology International Ltd. The idea and its verification occurred at the beginning of April 2016, patent documents (for Microsoft)

were submitted on 09.01.2017. At the end of March 2017, it was approved by Microsoft for patent application, but after the author's contract with Microsoft was finished the patent process was canceled and the permission for publishing the idea was granted by Microsoft lawyers. The permission to use real data has not been granted, and hence we use examples for demonstration.

The author would like to thank his colleagues from Microsoft Adeel Siddiqui and Preet Rihan and from Pactera Avinash Kamath for their comments and support of this work.

REFERENCES

[1] D. Barbara, J. Couto, Y. Li. COOLCAT: an entropy-based algorithm for categorical clustering. Proceedings of the Eleventh ACM CIKM Conference, 2002, 582–589.
[2] E. Aldana-Bobadilla, A. Kuri-Morales. A clustering method based on the maximum entropy principle. *Entropy*, **17**, (2015), No 1, 151–180.
[3] A. Jain, R. Dubes. Algorithms for Clustering Data. Prentice Hall, 1988.
[4] R. L. Thorndike. Who belongs in the family? *Psychometrika*, **18**, 4 (1953), 267–276.

Chavdar Dangalchev
18, Veliko Tarnovo
4000 Plovdiv, Bulgaria
e-mail: dangalchev@hotmail.com

ИЗПОЛЗВАНЕ НА ЕНТРОПИЯ НА КАТЕГОРИЙНИ ДАННИ ЗА КЛЪСТЪРИЗАЦИЯ

Чавдар Дангалчев

Една от най-важните задачи при клъстърния анализ е определяне на броя на клъстърите. В тази статия предлагаме метод, който използва ентропия на категорийни данни. Прилагат се примери, илюстриращи този метод.
2020 Mathematics Subject Classification: Основен: 62H30, Вторичен: 68T10.
Ключови думи: клъстърен анализ, ентропия, брой на клъстърите.