CLUSTERING OF HETEROGENEOUSLY TYPED DATA WITH SOFT COMPUTING

Angel Kuri-Morales[1], Luis Enrique Cortes-Berrueco[2] and Daniel Trejo-Baños[2]
[1]*Instituto Tecnológico Autónomo de México, Río Hondo No. 1 México D.F., Mexico*
[2]*Universidad Nacional Autónoma de México, Apartado Postal 70-600, Ciudad Universitaria, México D.F., Mexico*

Keywords: Clustering, Categorical variables, Soft computing, Data mining.

Abstract: The problem of finding clusters in arbitrary sets of data has been attempted using different approaches. In most cases, the use of metrics in order to determine the adequateness of the said clusters is assumed. That is, the criteria yielding a measure of quality of the clusters depends on the distance between the elements of each cluster. Typically, one considers a cluster to be adequately characterized if the elements within a cluster are close to one another while, simultaneously, they appear to be far from those of different clusters. This intuitive approach fails if the variables of the elements of a cluster are not amenable to distance measurements, i.e., if the vectors of such elements cannot be quantified. This case arises frequently in real world applications where several variables correspond to categories. The usual tendency is to assign arbitrary numbers to every category: to encode the categories. This, however, may result in spurious patterns: relationships between the variables which are not really there at the offset. It is evident that there is no truly valid assignment which may ensure a universally valid numerical value to this kind of variables. But there is a strategy which guarantees that the encoding will, in general, not bias the results. In this paper we explore such strategy. We discuss the theoretical foundations of our approach and prove that this is the best strategy in terms of the statistical behaviour of the sampled data. We also show that, when applied to a complex real world problem, it allows us to generalize soft computing methods to find the number and characteristics of a set of clusters.

1 INTRODUCTION

Clustering can be considered the most important unsupervised learning problem. As every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. In this particular case it is of relevance because we attempt to characterize sets of data trying not to start from preconceived measures of what makes a set of characteristics relevant.

When the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case, as will be discussed, geometrical distance). This is called distance-based clustering.

An important component of a clustering algorithm is the distance measure between data points.

Regardless of the distance we select it is clear that it implies handling of exclusively numerical vectors. To illustrate this fact we mention four of the most used metric clustering algorithms:

- K-means
- Fuzzy C-means
- Hierarchical clustering
- Neural Networks (Self Organizing Maps)

K-means is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is obvious and, lastly, SOMs are based on the connectionist paradigm.

The mentioned methods are conceptually different. These methods are all, however, metric. It is clear, that all of the metric algorithms are rendered useless when one or more of the variables is non-numeric. And this underlines the importance of being able to encode categorical variables. As stated before, we focus on finding an adequate encoding. A natural alternative, of course, is to abandon metric algorithms and there have been many attempts to do so (Shyam, 2008).

Kuri-Morales A., Cortes-Berrueco L. and Trejo-Baños D..
CLUSTERING OF HETEROGENEOUSLY TYPED DATA WITH SOFT COMPUTING.
DOI: 10.5220/0003690304910494
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 491-494
ISBN: 978-989-8425-79-9
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

2 UNBIASED ENCODING OF CATEGORICAL VARIABLES

We introduce an alternative which allows the generalization of numerical algorithms to encompass categorical variables. Our concern is that such encoding:

a) Does not induce spurious patterns

b) Preserves legal patters, i.e. those present in the original data.

By "spurious" patterns we mean those which may arise by the artificial distance induced by any encoding. On the other hand, we do not wish to filter out those patterns which are present in the categories. If there is an association pattern in the original data, we want to preserve this association and, furthermore, we wish to preserve it in the same way as it presents itself in the original data. The basic idea is simple: "Find the encoding which best preserves a measure of similarity between all numerical and categorical variables".

In order to do this we start by selecting Pearson's correlation as a measure of linear dependence between two variables. Higher order dependencies will be hopefully found by the clustering algorithms. This is one of several possible alternatives. Its advantage is that it offers a simple way to detect simple linear relations between two variables. Its calculation yields "r", Pearson's correlation, as follows:

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}} \tag{1}$$

Where variables X and Y are analyzed to search their correlation, i.e. the way in which one of the variables changes (linearly) with relation to the other. The values of "r" in (1) satisfy $-1 \le r \le +1$. What we shall do is to search for an encoding for categorical variable “A” so that the correlation calculated from such encoding does not yield a significant difference with any of the possible encodings of all other categorical or numerical variables.

2.1 The Algorithm

We define the *i-th instance* of a categorical variable V_X as one possible value of variable X. We denote the number of variables in the data as V. Further, we denote with r_{ik} Pearson's correlation between variables i and k. We would like to a) Find the mean μ of the correlation's probability distribution for all categorical variables by analyzing all possible combinations of codes assignable to the categorical variables plus the original (numerical) values of all non-categorical variables. b) Select the codes for the categorical variables which yield the closest value to μ. The rationale is that the absolute typical value of μ is the one devoid of spurious patterns *and* the one preserving the legal patterns. In the algorithm to be discussed next the following notation applies:

```
N← number of elements in the data
V← number of categorical variables
V[i]← the i-th variable
Ni  ← number of instances of V[i]
$r_j$ ← the mean of the j-th sample
S← sample size of a mean
$\mu_{\bar{r}}$← mean of the correlation's
distribution of means
$\sigma_{\bar{r}}$← standard deviation of the
correlation's distribution of means
```

Algorithm A1. Optimal Code Assignment for Categorical Variables.

```
01  for i=1 to V
02    j ← 0
03    do while $\overline{r}_j$ is not distri-
            buted normally
04       for k=1 to S
05        Assign a code for variable
          V[i]
06         Store this code
07        ℓ ← integer random number
          (1≤ℓ≤ V; ℓ≠i)
08        if variable V[ℓ] is cate-
          gorical
09            Assign a code for
                variable V[ℓ]
10         endif
11         $r_k = \frac{N\sum XY - \sum X\sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$
12       endfor
13       Calculate  $\overline{r_j} \leftarrow \frac{1}{S}\sum_{k=1}^{S} r_k$
14       j ← j+1
15    enddo
16 $\mu = \mu_{\bar{r}}$ ;   the   mean   of   the
correlation's distribution
17  $\sigma = \sqrt{ss}\cdot\sigma_{\bar{r}}$ ;  the std. dev. of the
correlation's distribution
18  Select the code for V[i] which
yields the $r_k$ closest to $\mu$
19  endfor
```

For simplicity, in the formula of line (11), X stands for variable V[i] and Y stands for variable V[ℓ]. Of

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

course it is impossible to consider all codes, let alone all possible combinations of such codes. Therefore, in algorithm A1 we set a more modest goal and adopt the convention that to *Assign a Code* [as in lines (05) and (09)] means that we restrict ourselves to the combinations of integers between 1 and Ni (recall that *Ni* is the number different values of variable *i* in the data). Still, there are Ni! possible ways to assign a code to categorical variable *i* and Ni! x Nj! possible encodings of two categorical variables *i* and *j*. An exhaustive search is, in general, out of the question. Instead, we take advantage of the fact that, regardless of the way a random variable distributes (here the value of the random encoding of variables *i* and *j* results in correlation r_{ij} which is a random variable itself) the *means* of sufficiently large samples very closely approach a normal distribution(Feller, 1966). Furthermore, the mean value of a sample of means $\mu_{\bar{r}}$ and its standard deviation $\sigma_{\bar{r}}$ are related to the mean μ and standard deviation σ of the original distribution by $\mu = \mu_{\bar{r}}$ and $\sigma = \sqrt{SS} \cdot \sigma_{\bar{r}}$. What a sufficiently large sample means is a matter of convention and here we made S=25 which is a reasonable choice. Therefore, the loop between lines (03) and (15) is guaranteed to end. In our implementation we split the area under the normal curve in deciles and then used a goodness-of-fit test with p=0.05 to determine that normality has been achieved. This approach is directed to avoid arbitrary assumptions regarding the correlation's distribution and, therefore, not selecting a sample size to establish the reliability of our results. Rather, the algorithm determines at what point the proper value of μ has been reached. Furthermore, from Chebyshev's theorem, we know that:

$$P(\mu - k\sigma \le X \le \mu + k\sigma) \ge 1 - \frac{1}{k^2} \qquad (2)$$

If we make k=3 and assume a symmetrical distribution, the probability of being within three σ's of the mean is roughly 0.95.

Three other issues remain to be clarified.

1) To Assign a code to V[i] means that we generate a sequence of numbers between 1 and Ni and then randomly assign a one of these numbers to every different instance of V[i].

2) To Store the code [as in line (06)] means NOT that we store the assigned code (for this would imply storing a large set of sequences). Rather, we store the value of the calculated correlation along with the root of the pseudo random number generator from which the assignment was derived.

3) Thereafter, selecting the best code (i.e. the one yielding a correlation whose value is closest to μ) as in line (18) is a simple matter of recovering the root of the pseudo random number generator and regenerating the original random sequence from it.

3 CASE STUDY: PROFILE OF CAUSES OF DEATH IN A LARGE HUMAN POPULATION

In order to illustrate our method we analyzed a data base corresponding to the life span and cause of death of 50,000 individuals between the years of 1900 and 2007. The confidentiality of the data has been preserved by changing the locations and regions involved. Otherwise data are a faithful replica of the original.

The database contains 50,000 tuples consisting of 11 fields: BirthYear, LivingIn, DeathPlace, DeathYear, DeathMonth, DeathCause, Region, Sex, AgeGroup, AilmentGroup and InterestGroup. Therefore, our working data base has 10 dimensions since the last variable (InterestGroup) corresponds to interest groups identified by human healthcare experts in this particular case and was not considered. This last field corresponds to a heuristic clustering of the data and could be used for the final comparative analysis of resulting clusters as the comparative analysis between the expert's clusters and the resulting clusters. We will explore this line in future works.

Once the data were encoded, we proceeded to use an unsupervised learning method for the clustering process. First we needed to determine the number of clusters in which we would group our sample.

We applied the fuzzy c-means algorithm to our coded sample. To determine the number of clusters we experimented with 17 different possibilities (assuming from 2 to 18 clusters). In each step we calculated the partition coefficient and classification entropy of the clustered data.

We applied the fuzzy c-means algorithm to our coded sample. To determine the number of clusters we experimented with 17 different possibilities (assuming from 2 to 18 clusters). In each step we calculated the partition coefficient (pc) and classification entropy (pe) of the clustered data, see (Lee, 2005); (Shannon, 1949); (Vinh, 2009). Plotting

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

the values of pc and pe against the number of clusters we got a graph.

To determine the number of clusters, we used the elbow criterion, see (Ganti, 1999), which indicates that we should use the value where the change in tendency is the most notable, in this case this occurred with 3 and 4 clusters. For clarity, we picked 3 clusters for this experiment.

Then we proceeded to apply Kohonen´s SOM to the data. As all data is now in numerical form, the algorithm was applicable without alterations over the set of variables. Therefore, after running the algorithm our data was classified in three clusters depending on the neuron which was closer to a specific data field. We interpreted the results according to the values of the mean of each variable on each cluster. We rounded the said values for BirthYear and DeathCause and obtained the following decoded values:

For cluster 1 the decoded values of the mean for BirthYear and DeathCause correspond to “1960” and cancer.

In cluster 2 the values are “1919” and Pneumonia.

In cluster 3 the values are “1923” and Heart stroke.

These logical backward results attest to their validity which, we emphasize, are obtained from an unbiased numerical encoding. Therefore we can infer that legal patterns are preserved and, furthermore, that such patterns (adequately encoded) allow a numerical method to find the patterns in spite of the fact that no unique valid metric does, in general, exist.

4 CONCLUSIONS

We have shown that we are able to find meaningful results by applying numerically oriented non-supervised clustering algorithms to categorical data by properly encoding the instances of the categories. We were able to determine the number of clusters arising from the data encoding according to our algorithm and, furthermore, to interpret the clusters in a meaningful way. Rather than a priori accepting essential limitations in numerical methods when applied to sets of categorical variables, we have adopted the point of view that machine learning techniques allow a broader scope of interpretation which are not marred by limitations of processing capabilities.

At any rate, the proposed encoding does allow us to tackle complex problems without limitation due to the non-numerical characteristics of the data. It is also a scalable method and independent of the set of data. Much work remains to be done, but we are confident that these is the first of a series of significant applications.

REFERENCES

V. Ganti, J. Gehrke, and R. Ramakrishnan. Cactus-*Clustering categorical data using summaries*. In KDD '99: *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 73-83, New York, NY, USA, 1999. ACM.

Lee, Y., and Choi, S., Minimum entropy, k-means, spectral clustering, Neural Networks, 2004. *Proceedings IEEE International Joint Conference on*, volume 1, 2005.

Shannon, C. E., and Weaver, W., The Mathematical Theory of Communication, *Scientific American*, July 1949.

Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *SDM*, pages 243-254, 2008.

Vinh, N. X., Epps, J., and Bailey, J., Information theoretic measures for clusterings comparison: is a correction for chance necessary?

Feller, William, An introduction to probability theory and its applications. Vol. II., Oxford, *England: Wiley*. (1966).

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS