Clustering Numerical and Categorical Data

Ricardo Linden

Lecturer – Faculdade Salesiana Maria Auxiliadora
R Monte Elíseo S/Nº
27943-180 Macaé , RJ – Brazil
rlinden@pobox.com

Abstract. Clustering is an important technique for data mining which allows us to discover unknown relationships in our data sets. Clustering algorithms that use metrics based on the natural ordering of numbers cannot be applied to categorical (non-numerical) data. In this tutorial we will review the main methods for numerical data clustering (K-Means, Hierarchical Clustering and Fuzzy C-Means) and then study two methods for categorical data clustering: CLICK (based on graphs) and STIRR (based on dynamical systems).

1 Introduction

In the last decades we have seen an exponential growth in the amount of information stored in data bases in many different areas such as finance and biology. Most data bases have doubled every two years, e.g., storing information on clients, transactions performed, DNA sequences and many others.

These huge data sets have no value if we cannot extract useful information and understand the hidden meaning in the data. Therefore, we need to extract information that can support business decisions or even to understand the rules that have generated those data. There can be hidden patterns and trends that, if uncovered, can be used in many different areas, ranging from marketing optimization to protein studies [1].

This process is called data mining and has been the focus of many important studies. Data mining encompasses several techniques, and one of the most important of them is clustering. Clustering algorithms intend to separate the elements into groups such that every element is more similar to the ones in its group than to any element in any other group, according to a specific criterion.

2 Clustering Numerical Data

Clustering numerical data relies on a metric that determines the distance of data pairs (how similar each pair is). The main metrics used are Euclidean distance, the Canberra metric, the correlation coefficient and the Mahalanobis distance [6].

Using these metrics we can define two different approaches: hierarchical and non hierarchical ones. A non-hierarchical approach to forming good clusters is to specify a desired number of clusters, say, k, then assign each case (object) to one of k clusters so as to minimize a measure of dispersion within the clusters [7].

One of the most common of these non-hierarchical algorithms is the k-means algorithm. It starts with an initial partition of the cases into k clusters and iterates in order to find the best cluster (the one that minimizes the distances intra-clusters). This is a very fast algorithm that has the drawback on relying on the pre-definition of the number of clusters by the user. Sometimes, knowledge of the ideal number of clusters in not available.

In hierarchical clustering the data are not partitioned into a pre-defined set of clusters but are linked to the nearest group forming a single cluster containing all objects (agglomerative methods) or divided in up to n clusters, each containing a single object (divisive methods) [6]. The groups are joined in a dendrogram that shows the similarity structure of the data and the number of groups generated depends on a cut parameter based on the user's analysis of the dendrogram generated.

The Fuzzy c-Means Algorithm is a fuzzy clustering algorithm used to establish an optimal classification for the data. It is a generalization of the K-Means algorithm that makes the class membership to become a relative one, allowing an object to belong to several classes at the same time with different degrees.

3 Clustering Numerical Data

Traditionally, clustering techniques are not directly applicable to categorical data. Preprocessing is commonly used to obtain the numeric features from categorical data for clustering. For example, in information retrieval, the vector model is applied: the frequency of occurrence of a word in document is used

As a numerical feature for clustering. However, there are also many datasets containing categorical data, which cannot be transformed to numerical features appropriately, so that special categorical clustering is needed [2]. In this tutorial we will concentrate on two important types of algorithms: graph-based and dynamical systems-based categorical clustering.

CLICK is an algorithm which finds clusters in categorical datasets based on a search method for k-partite maximal cliques. CLICK is able to detect subspace clusters and scales very well for high dimensional datasets.

STIRR is an approach based on an iterative method for assigning and propagating weights on the categorical values in a table that can be studied analytically in terms of certain types of non-linear dynamical systems [4]. The algorithm represents each attribute value as a weighted vertex in a graph. Starting with the initial conditions, the system is iterated until a "fixed point" is reached. When the fixed point is reached, the weights in one or more of the "basins" isolate two groups of attribute values on each attribute. STIRR has a problem dealing with real valued attributes that we have addressed using fuzzy rounding [8], which will also be described thoroughly.

4 Conclusion

Other methods, both numerical and categorical [1,3,5] are available, but will not be discussed due to time constraints. This tutorial should help attendees to develop a good notion of the difficulties concerning the problem of clustering as well as introducing several ideas on how to cope with them.

5 References

[1] Aggarwal, C. C., Magdalena, C., and Yu, P. S. Finding localized associations in market basket data. IEEE Trans. Knowledge and Data Eng. 14, 1 (2002), 51–62.

[2] Chen, K. and Liu L. – "Towards Finding Optimal Partitions of Categorical Datasets", Technical Report, 2003

[3] Cristofor, D. and Simovici, D. A. – "An information theoretical approach to clustering categorical databases using genetic algorithms", USA, 2000

[4] Gibson, D., Kleinberg, J. and Raghavan, P. Clustering categorical data: An approach based on dynamical systems. Proc. of VLDB 8, 3–4 (2000), 222–236.

[5] Guha, S., Rastogi, R., And Shim, K. Rock: A robust clustering algorithm for categorical attributes. Proc. Of IEEE Intl. Conf. on Data Eng. (ICDE) (1999).

[6] Gordon, A. D. – "Classification", 1st Edition, Chapman and Hall Ed., USA, 1981

[7] Hair, J F Jr. ; Anderson, R. E. et al – "Multivariate Data Analysis", 4th Edition, Prentice Hall, USA, 1995

[8] Linden, R. and Bhaya, A. – "Evolving Fuzzy Rules for Classification: An Application in Medical Diagnosis", submitted, Brazil, 2004.