Multi-Modal Data Fusion: A Description

Sarah Coppock and Lawrence J. Mazlack

ECECS Department
University of Cincinnati
Cincinnati, Ohio 45221-0030
USA
{coppocs,mazlack}@uc.edu

Abstract. Clustering groups records that are similar to each other into the same group, and those that are less similar into different groups. Clustering data of mixed types is difficult due to different data characteristics. Extending Gower's metric for nominal and ordinal data is incorporated into an agglomerative hierarchical clustering algorithm to cluster mixed type data. This paper describes the extensions and algorithm.

1 Introduction

Enormous amounts of data are collected daily for a variety of reasons. Potential information can be hidden within the collected data. Clustering records provides information such as the discovery of the overall distribution, the discovery of groups that are similar, and the discovery of outliers. For example, a private business such as a grocery store may want to categorize its customers.

Clustering groups records that are more similar to each other in the same group, placing records that are less similar to each other into different groups [1], [2], [3], [4]. Deciding which records should be grouped together is subjective and can depend on both the domain and the distribution of the data set.

Clustering is an unsupervised process; no semantic knowledge of the data is explicitly used. Clustering is used in many fields and is a well studied topic. Most clustering algorithms can handle all quantitative attributes. Recently, research has developed clustering algorithms to handle all qualitative attributes [5], [6], [7], [8].

How best to cluster the records containing mixed data such as those in Table 1 is a difficult problem). Some of the difficulty is due to the different characteristics of data and how similarity (or dissimilarity) is defined for records of mixed data types. Without assuming semantic knowledge about the data, the similarity, dissimilarity or distance metric[1] typically is defined based on inherent characteristics of data.

[1] Unless specified differently, the word *metric* refers to the standard from which a measure is taken, rather than the mathematical definition.

Table 1. Example mixed data set

	—— a_1	a_2	a_3	a_4
x_1	Coke	C	18	Cincinnati
x_2	Pepsi	A	30	Cincinnati
x_3	Pepsi	B	12	St. Petersburg
x_4	Budweiser	D	26	St. Petersburg
x_5	Budweiser	F	23	Hamilton
x_6	Guinness	B	35	Tampa
x_7	Guinness	D	21	Tampa
x_8	Budweiser	C	40	Hamilton
x_9	Pepsi	A	32	Tampa
x_{10}	Pepsi	A	31	Cincinnati

A *data set* consists of records; a *record* has a fixed number of attribute-value pairs. Let $D = \{x_1, x_2, \ldots, x_n\}$ denote a data set where x_i is the i^{th} record in the set, and x_{ik} denote the value for the k^{th} attribute in record x_i.

Different types of data exist [1], [2]. Two inherent characteristics of data may be present: *order* and *scale*. While data may be categorized in multiple ways [1], [4], in this paper, the following categories will be used:

- *quantitative* data which has both scale and order. This includes both discrete, e.g. *number of children*, and continuous, e.g. *height* and *age*. Quantitative can also be referred to as *numeric* or *scalar* data;
- *ordinal*, a subcategory of the more general qualitative. Ordinal data has order, but is without scale. Some example include *level of education*, or *academic letter grade*;
- and *nominal*, which is also qualitative and has no order or scale associated with it. Examples of nominal data include: *gender*, *race*, *area of study*.

Although there exist multiple approaches to clustering[2], [4], we selected a hierarchical agglomerative approach to cluster mixed data. Prior work has focused on extending the popular k-means algorithm [13]. Although the k-means has better complexity, a good selection for the parameter k needs to be known prior to clustering in order to obtain good results.

The algorithm utilizes an extension of a similarity measure developed by Gower [9] that handles both quantitative and qualitative attributes. A previous approach by [10] also uses an agglomerative hierarchical approach. It is based on Goodall's similarity measure [11] and requires the discovery of probability distributions.

Hierarchical clustering builds a hierarchical structure according to a given similarity or distance measure. The bottom level of the structure has singleton clusters while the top level contains one cluster with all records. *Divisive* algorithms construct the hierarchy in a top down fashion by splitting clusters, while *agglomerative* algorithms build the structure from the bottom up by merging selected clusters.

In an agglomerative hierarchical algorithm, during each iteration, two clusters are heuristically selected. These selected clusters are then merged to form a new cluster. The selection process depends on how similar or distant two clusters. There may be a stopping condition; when the condition is met, the algorithm stops. There may be a heuristic in determining the "best" level, or clustering. This is the level determined to be better than all other possible levels. This can be based on the used similarity function such as in [12] or may be related to the stopping condition.

2 Similarity

Similarity is domain dependent. It can be defined in many ways. In the area of taxonomy, a similarity metric was developed by Gower [9]. This metric is a function of the similarity between pair-wise attribute-values of each attribute. It allows for missing values and for individual attribute weighting when measuring similarity. It may be the case that one or more attributes are "more important" than other attributes; these attributes would have a larger weighting than the others.

$$S_{ij} = \frac{\sum\limits_{k} s_{ijk} w_k}{\sum\limits_{k} \delta_{ijk} w_k} \tag{1}$$

gives the weighted equation for Gower's metric [9], where s_{ijk} is the similarity between the k^{th} attribute values for records i and j; w_k is the k^{th} attribute weighting; δ_{ijk} is a binary parameter indicating whether the similarity is defined for the k^{th} attribute values for records i and j. s_{ijk} is defined as:

$$s_{ijk} = \begin{cases} \begin{cases} 1, \text{ if } x_{ik} = x_{jk} \\ 0, \text{ if } x_{ik} \neq x_{jk} \end{cases}, & k \text{ qualitative} \\ 1 - \dfrac{|x_{ik} - x_{jk}|}{R_k}, & k \text{ quantitative} \end{cases} \tag{2}$$

Gower's metric does not include information provided by ordinal qualitative data. For example, we would consider A more similar to B than to C. Also, any information provided by a concept hierarchy is ignored. For example, we would consider *Pepsi* to be more similar to *Coke* than it is to *Budweiser* because both *Pepsi* and *Coke* are brands of colas and *Budweiser* is a brand of beer.

2.1 Extending Gower's Measure

Previous metrics used to measure similarity between records of mixed data do not allow individual attribute weighting and are limited to simple matching techniques in handling categorical and ordinal data [13]. Gower's metric is extended to incorporate the information provided by order and concept hierarchy into the similarity measurement. Ordinal values are mapped as in [2]; nominal values are defined as in [14]. The general equation, equation 1, is the same, but the attribute value similarity function, s_{ijk}, is extended. The extension of the attribute value similarity function is given in equation 3.

$$s_{ijk} = \begin{cases} 1 - \dfrac{|x_{ik} - x_{jk}|}{R_k}, & k \text{ quantitative} \\ 1 - \dfrac{|\pi(x_{ik}) - \pi(x_{jk})|}{\pi_{\max} - \pi_{\min}}, & k \text{ ordinal} \\ \dfrac{2 * depth(LCA(x_{ik}, x_{jk}))}{depth(x_{ik}) + depth(x_{jk})}, & k \text{ nominal with hierarchy} \\ \begin{cases} 1, \text{ if } x_{ik} = x_{jk} \\ 0, \text{ if } x_{ik} \neq x_{jk} \end{cases}, & k \text{ nominal without hierarchy} \end{cases} \tag{3}$$

x_{ik} is the k^{th} value for record i; $w_k \geq 0$ is the weighting for attribute k. The initial weighting for all attributes is 1. δ_{ijk} is a binary variable indicating any missing values. R_k is the range for the the k^{th} attribute. $\pi(xik)$ is the mapping for the k^{th} value of the i^{th} record. depth(x) is the number of edges between x and the root of the associated tree. $LCA(x_{ik}, x_{jk})$ is the common ancestor of the nodes, which has maximum depth in the tree, of the nodes corresponding to x_{ik} and x_{jk}. The remaining variables are the same as defined in equations 1 and 2.

Note that for nominal attributes, the concept hierarchy structure is limited to tree structures. With this definition, network concept hierarchy structures cannot be appropriately handled. The challenge with network structures is how to decide which root to use in determining similarity. The structure can be either autonomously discovered or user-supplied. Also, note that we are imposing a fixed scale to ordinal data as an attempt at satisficing.

Using Gower's metric allows for the quantitative similarity measure to incorporate similarity from each dimension and also allows for individual attribute weighting without using a single weighting factor between all categorical and all quantitative. In addition, the metric allows for missing attribute values.

3 The Algorithm

The clustering algorithm is an agglomerative hierarchical one. With agglomerative clustering, the structure is constructed from the bottom up. The structure used is implemented as a binary forest. Without any given stopping conditions such as a number of clusters or a similarity threshold, the final structure is a binary tree.

Each node has exactly two children corresponding to the clusters with are merged to create the new cluster associated with the node. Each node structure consists of the following:

- a sequence number: the iteration at which the node was created;
- a similarity value: the similarity of the clusters that were merged to create the node;

– a cluster: the cluster associated with the node.

During each iteration, the two best nodes, n_{i1} and n_{i2}, are selected. A new node, n_i is created as the parent node of n_{i1} and n_{i2}. Figure 1 is a visual example of the first two iterations for six records denoted by a, b, c, d, e and f. Initially, there are six tree roots for each of the items (Figure 1 I.). Suppose e and f have maximal similarity, denoted by μ_{ef}, according to the measure used. Then e and f are selected and merged to create the updated structure to be that of Figure 1 II. The similarity for the new node containing e and f is μ_{ef}. The sequence number for the new node is *1*.

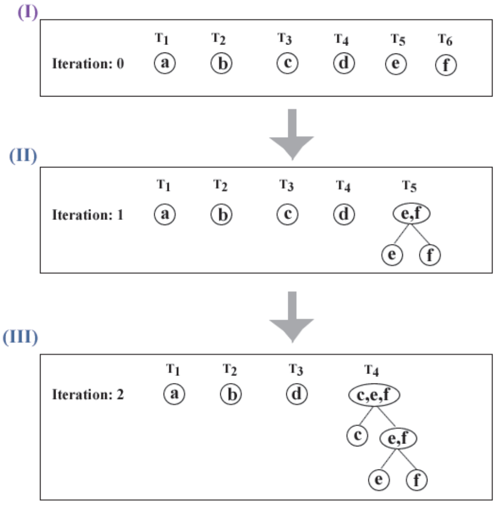

Fig. 1. Example structure in initial iterations

The structure is initialized with the one root node for each of the records (Figure 1 I). The sequence number for each of the nodes in the initial structure is zero. The similarity for each node in the initial structure is one since there is only one record in the cluster. For each iteration, two roots are selected heuristically and the trees merged into one.

3.1 Selection Of Clusters To Merge

During an iteration, heuristics are employed to select the clusters to merge. Some possible heuristics are inter-cluster average similarities, maximum inter-cluster similarity, and minimum inter-cluster similarity between records from each cluster. Similar to partitional clustering approaches, the decision of which cluster to merge depends on the similarity between cluster representations. The two clusters that have maximal similarity are selected. Before continuing on, the cluster representation is discussed.

Table 2. Example cluster from Table 1

	a_1	a_2	a_3	a_4
x_1	Coke	C	18	Cincinnati
x_2	Pepsi	A	30	Cincinnati
x_3	Pepsi	B	12	St. Petersburg
x_9	Pepsi	A	32	Tampa
x_{10}	Pepsi	A	31	Cincinnati

Table 3. Example cluster from Table 1 showing multiple possible modes for nominal attribute a_4

	a_1	a_2	a_3	a_4
x_2	Pepsi	A	30	Cincinnati
x_3	Pepsi	B	12	St. Petersburg

3.2 Cluster Representation

The cluster representation used is equivalent to a center. There is an attribute-value pair for each attribute, similar to the mode representation used in partition clustering approaches such as [13]. For a quantitative attribute, the mathematical median is used; for an ordinal attribute, the median; and for a nominal attribute, the mode. The median of a set of values is defined as the value which half of the values in the set are less than or equal and the other half are greater than or equal to. The median is a more appropriate representation in that it is unknown, but likely, that the quantitative attributes are skewed. The mode of a set of values is defined as the value that has a largest number present in the set of values.

An example would make this clearer. Suppose a cluster contains some of the records from Table 1 given in Table 2. The representation would be computed as: {*Pepsi*, *A*, *30*, *Cincinnati*}.

When computing the mode for a nominal attribute, there may be a tie between one or more possible values. For example, suppose we have the cluster $\{x_2, x_3\}$ from Table 1 given in Table 3. Both *St. Petersburg* and *Cincinnati* are possible modes for a_4; both values have a count of one in the cluster. The value selected to represent the attribute is the value that has the smallest distribution in the data set. Since the overall distribution for *Cincinnati*, which has a count of three, is greater than that of *St. Petersburg*, which has a count of two, *St. Petersburg* is used in the representation. The reason for this selection is matching a value that is least common would be more important. It is still possible to have a tie. If this is the case, a random one is selected.

4 Conclusion

It is important to note that the given hierarchies are not the only possible structures. A future goal of this work is to discover how sensitive the algorithm, specifically the measure, is to the hierarchies supplied. It is believed that learned hierarchies will give better results; although, how to evaluate results from clustering mixed data is yet to be addressed.

References

1. Everitt, B.: Cluster Analysis. 3rd ed. Hodder & Stoughton London (1993)
2. Han, J. and Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufman, San Diego. (2001)
3. Jain, A. K. and Dubes, R. C.: Algorithms for Clustering Data. Prentice Hall, New Jersey. (1988)
4. Sneath, P. and Sokal, R.: Numerical Taxonomy. Freeman and Company, San Fransisco. (1973)
5. Ganti, V., Gehrke, J., and Ramakrishnan, R.: CACTUS: Clustering Categorical Data Using Summaries. in Knowledge Discovery and Data Mining. (1999) 73–83
6. Gibson, D., Kleinberg, J., and Raghavan, P.: Clustering Categorical Data: an Approach Based on Dynamical Systems. in Proceedings of the 24th VLDB Conference. **8**:3/4 (2000) 222–236
7. Guha, S., Rastogi, R., and Shim, K.,: ROCK: A Robust Clustering Algorithm for Categorical Attributes, in Information Systems, **25**:5 (2000) 345–366
8. Han, E., Karypis, G., Kumar, V., and Mobasher, B.: Clustering Based on Association Rule Hypergraphs. in Proceedings of SIGMOD '97 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'97) (May 1997)
9. Gower, J. C.: A General Coefficient of Similarity and Some of Its Properties. Biometrics. **27**:4 (1971) 857–871
10. Li, C. and Biswas, G.: Conceptual Clustering with Numeric-and-Nominal Mixed Data: A New Similarity Based System. IEEE Transcript on KCE 1998. (1998)
11. Goodall, D.: A New Similarity Index Based On Probability. in Biometrics. **22**:4 (1966) 882–907
12. Salvador, S. and Chen, P.: Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. Florida Institute of Technology (2003) 8 pages
13. Huang, Z. and Ng, M.K.: A Fuzzy k-Modes Algorithm for Clustering Categorical Data. IEEE Transactions on Fuzzy Systems **7**:4 (1999) 446–452
14. Ganesan, P., Garcia-Molina, H. and Widom, J.: Exploiting Hierarchical Domain Structure to Compute Similarity. ACM Transactions on Information Systems, **21**:1 (2003) 64–93