USING SOFT COMPUTING TECHNIQUES TO INTEGRATE MULTIPLE KINDS OF ATTRIBUTES IN DATA MINING

SARAH COPPOCK AND LAWRENCE MAZLACK

Computer Science, University of Cincinnati, Cincinnati, Ohio 45220 USA

E-mail: {mazlack,coppock}@uc.edu

Data mining discouvers interesting information from a data set. Mining incorporates different methods and considers different kinds of information. Granulation is an important task of mining. Mining methods include: association rule discouvery, classification, partitioning, clustering, and sequence discouvery. The data sets can be extremely large with multiple kinds of data in high dimensionality.

Most current clustering algorithms deal with either quantitative or qualitative data, but not both. However, many data sets contain a mixture of quantitative and qualitative data. We are considering how to best group records containing multiple kinds of data. It is difficult to do this. Even grouping based on different quantitative metrics has its difficulties. There are many partially successful strategies as well as several different possible differential geometries. Adding in various qualitative elements is exceedingly difficult. We expect to use a mixture of scalar methods, soft computing (rough sets, fuzzy sets), as well as methods using other metrics.

To cluster records in a data set, it would be useful to have a similarity measure. Unfortunately, few exist that account meaningfully for any combination of kinds of data. The more meaningful metrics known are restrictive to a particular area or science. The method of combining magnitude difference and simple matching so that it is general enough for mining is a topic that is yet to be reasonably solved. We will present several strategies for integrating multiple metrics.

1 Grouping Records Together

Clustering groups of objects into clusters so that the similarity among objects within the same cluster (intra-cluster similarity) is maximized and the similarity between objects in different clusters (inter-cluster similarity) is minimized. Clustering increases granule size and is useful in data mining. Clustering can discouver the general distribution of the data, which is also useful in data mining. It allows discouvery of similar objects described in the data set. A good characterization of the resulting clusters can also be a valuable data mining product.

A data set can have millions of records over hundreds of attributes. The attributes have many disparate kinds of data. Some algorithms offer promise in handling multiple kinds of data. Unfortunately, their complexity is geometric and thus not scalable. They are useful for small data sets but not suitable for large data sets.

There are two types of hierarchical approaches to clustering: agglomerative and divisive. Agglomerative begins with all objects in their own cluster and combines clusters together for which the similarity is the greatest. This is done repeatedly until all objects are in the same cluster. Divisive begins with all objects in the same

cluster and does the reverse. Because these approaches are based on similarity, it is desirable that an appropriate metric would be available for measuring the similarity between records containing any mix of kinds of data.

Another approach to grouping records is partitioning. An initial clustering is formed and items are iteratively moved to other clusters to improve the quality of the clustering. Sometimes, the term *partitioning* is used as if synonymous with *clustering*. However, partitioning can also be approached as a purification process (Coppersmith, 1999) where partitions progressively become more pure. Increasing granule of small partitions is then a matter of relaxing partition boundaries through either rough sets of the incorporation of fuzzy values.

2 Kinds Of Data

Data can be classified by scale and kind (i.e., qualitative or quantitative). Most current clustering algorithms deal with quantitative data. It includes continuous values, such as a person's height, and discrete values, such as the number of cars sold. Qualitative data are simply symbols or names with no natural scale between the values. This includes nominal data such as the color of a car and ordinal data such as the doneness of a burger, rare, medium, well. A consequence of the lack of a fixed scale is difficulty in quantitatively measuring the similarity between two qualitative values. It is common to use simple Boolean matching; e.g., 1 if two values match and 0 if two values do not match. For two quantitative values, the difference in magnitude is suitable.

Many data sets contain a mixture of kinds of data. When clustering the records, it is important that this is taken into account when evaluating:

- the similarity (or its complement dissimilarity) between records,
- when evaluating the clustering, and
- when finding a good representation of a cluster (i.e., in centroid-based clustering).

3 Similarity/Dissimilarity In A Mixture Of Data

Most current similarity metrics use pair-wise comparisons in the measurement of the overall similarity between two records. For example, if the two records are:

a_1	b_1	c_1
a_2	b_2	c_2

Then the similarity between the two records would be defined as sim(a1,a2) $\oplus$ sim(b1,b2) $\oplus$ sim(c1,c2) where $\oplus$ indicates some combination operator, usually +.

Although the measure itself can be either kind of value, quantitative or qualitative, most current metrics attempt to derive a quantitative measure.

Developed methods work sufficiently well for quantitative data where the dissimilarity between two records could be a function such as an Minkowski metric (or L_p metric) (L_p: $[\sum(|x_i-y_i|^p)\ ]^{1/p}$). If the data to be clustered is ordinal such as *tall*, *medium* and *short*, it has been suggested that the values can be mapped into numerical values in a meaningful way. And, used in a distance measure for quantitative data (Han, 2001, 344-5) (Li, 1998). But, this is not guaranteed to be meaningful. For example, if the ordinal values are *tall*, *medium* and *short*, the difference between *tall* and *medium* may be smaller (or larger) than the difference between *medium* and *short*. In other words, we are assuming that the three values have a fixed and artificial scale. This becomes a difficulty because the metric for quantitative data uses the scale and magnitude in the computation of the measure.

There are also metrics such as those based on simple matching (e.g., the Jaccard coefficient) that work well for when all attributes are categorical (Sneath, 1973) (Wang, 1999). Unfortunately, these values cannot be sensibly mapped into a form appropriate for typical distance measures. For example let D be:

fruit	*color*	*bag size*
apple	red	5
orange	orange	3
apple	green	5

and one arbitrary mapping π_1 be

fruit={orange: 1, apple: 2}

color={red: 1, orange: 2, green: 3}

and another arbitrary mapping π_2 be

fruit={orange: 2, apple: 1}

color={red: 0, orange: 1, green: 6}

When applying Euclidean distance, the distances between the mapping target values *1&2* and *2&3* are equal (*1*); while the distances between *1&2* and *1&3* are not equal (*1* vs *2*). The difficulty is that the metric is defined on quantitative values, but we are imposing an ordering and a scale that may not have meaning in the real world; nor, is it necessarily useful. Therefore, we can construct an arbitrary mapping, but we cannot be sure about the utility of the resulting measure.

Guha (2000) gives an example of a problem that occurs when using numerical distance metrics, specifically Euclidean distance, on binary data that has been converted from market basket data. The difficulty arises when using the measure in centroid based clustering, i.e., one centroid or mean representation for a cluster. It is possible to reach a point where an item is determined closer to a particular cluster's

mean and consequently the item is added to that cluster, when it is actually closer to another cluster's mean.

Using the metrics developed for nominal data causes loss of information when applied to quantitative data (Li, 1998). For example, if using simple matching between quantitative values, 3.4, 3.5, and 4.2, will all have the same similarity between them. The fact that 4.2 is more dissimilar to 3.4 than 3.5 is lost. Even if matching were to be used on quantitative ranges such as [3.0,3.5], information will still be lost. For example, if three values are being compared, say 3.0, 2.9 and 3.4, then 3.0 will be considered as distant from 2.9 as 3.4 is from 2.9.

Some clustering approaches for categorical values only discover a clustering of the categorical values themselves (Gibson, 2000) (Han, 1997) (Zhang, 2000). This may provide useful information and could possibly supplement in the clustering of the records. In the current methods though, it is unclear how to derive a specific measure of closeness for any two particular values. For example, if it is discovered that *a*, *b* & *c* belong in the same cluster, how can we decide the closeness of each pair? Is *a* closer to *b* than to *c*? If this could be derived, then it could be used in the clustering of whole records. It is also not clear how these methods adjust for values that are found in the domain of multiple attributes. These metrics are based on the frequency and occurrences of values together as does the metric used by Ganti (1999). This is essentially the problem discussed by Kanal (1993) when he discussed Watanabe's (1969, 1985) "Theorem Of The Ugly Duckling;" namely the needed to have weighted memberships. One way of achieving weighted memberships is to use soft computing tools.

Other clustering methods find the clustering of the records by determining specific values' occurrence frequency. Wang (1999) developed a partitioning method using an evaluation function on the clusters. While it does not generalize to a mix of data kinds, it is an interesting approach. Wang identifies items that are present in a cluster above a support threshold. By determining two sets for each cluster, one for items occurring above the support threshold and the other below, a cost function is defined using these sets. The portion of the cost function that represents the intra-cluster similarity is a count of items in the cluster that are below the threshold.

Huang (1997) (1999) extends the distance-based method k-means algorithm to handle categorical data. By using an integer value, 1 or 0, to indicate non-matching and matching respectively, a categorical attribute is incorporated into the distance metric. In (Huang, 1997), similarity is computed as the sum of square differences for the numerical attributes simply added to a weighted summation of matches for the categorical attributes. Simply stated, the similarity is a combination of two metrics, one for quantitative and one for categorical. The numerical values maintain the characteristic of magnitude while categorical data have no magnitude or ordering. The

magnitudes of the quantitative attributes contribute to the measurement differently; a point previously made by Goodall (1966). For example, let D be:

t_1	1	a	2
t_2	1	b	2
t_3	4	a	2
t_4	2	a	1
t_5	1	b	4
t_6	3	a	3

For this example, the respective similarity using Huang's approach with weight of 1 (with the suggested weight of 1.27) is displayed in the following matrix:

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	0	1	9	2	5	5
t_2	1	0	10	3	4	6
t_3	9	10	0	5	14	2
t_4	2	3	5	0	11	5
t_5	5	4	14	11	0	6
t_6	5	6	2	5	6	0

Notice that $d(t_1,t_2)=1$ and $d(t_1,t_3)=9$ where $d(t_i,t_j)$ is the distance (dissimilarity) for objects t_i and t_j. Should the magnitude associated with a numerical attribute give considerably more to the (dis)similarity measure between records? Note that t_1 has two values in common with both t_2 and t_3. In this case, it makes more sense to find a quantitative metric consistent across all of the attributes being considered in the measure. This follows from the idea of standardizing the data before the computation of similarity (Everitt, 1993).

Li (1998) developed their method using the Goodall similarity metric (Goodall, 1966). Their metric measures the amount of weight that a categorical value contributes to the overall similarity measure. If two records have the same value for an attribute *k*, then the similarity is not necessarily 0 or 1 (match/no-match). Most previous metrics allow only match/no-match. Li's value for a match is between 0 and 1. The following table gives the distance using Li's metric for the previous data set:

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	0.16	0.31	0.82	0.29	0.67	0.66
t_2	0.31	0.03	0.71	0.31	0.10	0.93
t_3	0.82	0.82	0.41	0.58	0.99	0.44
t_4	0.29	0.31	0.58	0.82	0.92	0.52
t_5	0.67	0.10	0.99	0.92	0.10	0.71
t_6	0.66	0.93	0.44	0.52	0.71	0.82

The metric allocates a value proportional to the frequency as compared with other values. This weight allows for a more meaningful measure. This measure takes the distribution of values into account along with the magnitude. The chi-squared ($\chi 2$) statistic is used for computing the measure between records. Because this statistic is used, there is the assumption of independence among the attributes, which cannot be guaranteed. It also may not be the best idea to have the distribution of values affect the qualitative values' contribution to the measure. For example, in the same data set, if *apples* are a rare value for an attribute while *oranges* are frequent, if two records match on the value *apples*, then should *apples* contribute more to the measure than if the two records match on *oranges*?

The above example gives two distance measures. Even though both measures represent dissimilarity, we cannot compare the metrics directly. This is because Li's measure takes on real values 0 and 1, where Huang's measure takes on values between 0 and infinity. Yet, if we think about dissimilarity as imposing an ordering with respect to a record, we can show that we cannot decide whether either measure is useful. Let $t_i < t_j$ stand for t_i is closer to the record, t_k, than t_j is to t_k. Then with respect to t_1, we have $t_1 < t_2 < t_4 < t_5 = t_6 < t_3$ with Huang's metric, where we have $t_1 < t_4 < t_2 < t_6 < t_5 < t_3$ with Li's metric. This shows that the utility of the metric cannot be determined. If two records can be farther or closer depending on the metric used, then how can it be decided which metric is useful?

Epilogue

For a metric to cluster records having attributes with different kinds of data; e.g., scalar, categorical, etc., it would be useful to have a uniform similarity measure. Unfortunately, very few exist that can handle combinations of different kinds of data. The meaningful multi-modal metrics are so far restricted to particular scientific domains. A general method of combining differences in magnitude and kind so suitable

for data mining is a topic that has yet to be satisfactorily resolved. There are several open questions that wait for a persistent investigator.

Bibliography

G. Biswas, J. Weinberg, D. Fisher (1998) "ITERATE: A Conceptual Clustering Algorithm For Data Mining," *IEEE Transactions on Systems, Man And Cybernetics-Part C: Applications and Reviews*, v 28, n 2, May, p 219-230

D. Coppersmith, S.J. Hong, J. Hosking (1999) "Partitioning Nominal Attributes In Decision Trees," *Data Mining And Knowledge Discovery*, v 3 n 2, p 197-217

B. Everitt (1993) **Cluster Analysis**, 3rd ed. Hodder & Stoughton, London

V. Ganti, J. Gehrke, R. Ramakrishnan (1999) "CACTUS: Clustering Categorical Data Using Summaries" *Knowledge Discovery and Data Mining*, p 73-83

D. Gibson, J. Kleinberg, P. Raghavan (2000) "Clustering Categorical Data: An Approach Based On Dynamical Systems" *Proceedings of the 24th VLDB Conference*, v 8, n 3/4, p 222-236

D. Goodall (1996) "A New Similarity Index Based On Probability" *Biometrics*, v 22, n 4, p 882-907

S. Guha, R. Rastogi, K. Shim (2000) "ROCK: A Robust Clustering Algorithm For Categorical Attributes," *Information Systems*, v 25, n 5, p 345-366

E. Han, G. Karypis, V. Kumar, B. Mobasher (1997) "Clustering Based On Association Rule Hypergraphs," *Proceedings of SIGMOD '97 Workshop on Research Issues in Data Mining and Knowledge Discouvery (DMKD'97)*, May

J. Han, M. Kamber (2001) **Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, San Francisco

Z. Huang, M. Ng (1999) "A Fuzzy k-Modes Algorithm For Clustering Categorical Data," *IEEE Transactions on Fuzzy Systems*, v 7, n 4, August, p 446-452

Z. Huang (1997) "Clustering Large Data Sets With Mixed Numeric And Categorical Values," *Proceedings Of 1st Pacific-Asia Conference on Knowledge Discouvery And Data Mining*

A. Jain, R. Dubes (1998) **Algorithms For Clustering Data**, Prentice Hall, New Jersey

L. Kanal (1993) "On Pattern, Categories, And Alternate Realities," *Pattern Recognition Letters 14, p 241-255*

C. Li, G. Biswas (1998) “Conceptual Clustering With Numeric-And-Nominal Mixed Data-A New Similarity Based System," *IEEE Transactions on KCE*

P. Sneath, R. Sokal (1973) **Numerical Taxonomy**, Freeman and Company, San Francisco

K. Wang, C. Xu, B. Liu (1999) “Clustering Transactions Using Large Items," CIKM, p 483-490

S. Watanabe (1960) **Knowing And Guessing**, Wiley, New York

S. Watanabe (1985) **Pattern Recognition - Human And Mechanical**, Wiley, New York

Y. Zhang, A. Wai-chee Fu, C. Cai, P. Heng (2000) “Clustering Categorical Data," *16th International* Conference on Data Engineering