Granulating Data On Non-Scalar Attribute Values

Lawrence Mazlack
Sarah Coppock

Computer Science
University of Cincinnati
Cincinnati, Ohio 45220
{mazlack, coppocs}@uc.edu

Abstract

Data mining discouvers interesting information from a data set. Mining incorporates different methods and considers different kinds of information. Granulation is an important aspect of mining. The data sets can be extremely large with multiple kinds of data in high dimensionality. Without granulation, large data sets often are computationally infeasible; and, the generated results may be overly fine grained.

Most available algorithms work with quantitative data. However, many data sets contain a mixture of quantitative and qualitative data. Our goal is to group records containing multiple data varieties: quantitative (discrete, continuous) and qualitative (ordinal, nominal). Grouping based on different quantitative metrics can be difficult. Incorporating various qualitative elements is not simple. There are partially successful strategies as well as several differential geometries. We expect to use a mixture of scalar methods and soft computing methods (rough sets, fuzzy sets), as well as methods using other metrics.

To cluster whole records in a data set, it would be useful to have a general similarity metric or a set of integrated similarity metrics that would allow record to record similarity comparisons. There are methods to granulate data items belonging to a single attribute. Few methods exist that might meaningfully handle a combination of many data varieties in a single metric. This paper is an initial consideration of strategies for integrating multiple metrics in the task of granulating records.

GROUPING RECORDS TOGETHER

Granulation helps data mining accomplish: association rule discouvery, classification, partitioning, clustering, and sequence discouvery. Without granulation, large data sets often are computationally infeasible; and, the generated results may be overly fine grained.

Data mining a data set composed of varied data stored in records can focus on either: granulating individual attributes, data extracted from the records making up a data set, or whole records. To cluster whole records in a data set, it would be useful to have a general similarity metric that would allow record to record similarity comparison. There are methods to granulate data items belonging to a single attribute. Unfortunately, few methods exist that meaningfully account for a combination of many data varieties.

Clustering groups 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 data distribution; and, aid in the discouvery of similar objects described in the data set. A good characterization of the resulting clusters can also be a valuable data mining product.

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 largest. This is done repeatedly until all objects are in the same cluster. Conversely, divisive begins with all objects in the same cluster and does the reverse. Because most well understood approaches use a similarity metric, an similarity appropriate metric or a way to integrate diverse metrics is desirable. Any mix of data varieties without losing the meaning behind the metric's is necessary.

Another approach to grouping records is partitioning. 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 or fuzzy values.

A data set can have millions of records with hundreds of attributes. The attributes may have many disparate kinds of data. Some algorithms offer promise in handling multiple kinds of data. Unfortunately, they are not scalable as their complexity is geometric. They are only useful for small data sets. In addition, some approaches lose the meaning of the metric when trying minimize algorithmic complexity.

DATA VARIETIES

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*.

For measuring the similarity of two quantitative values, a function of the difference in magnitudes is used. For evaluating similarity between qualitative values, it is common to use simple matching; i.e., if the two values match, then the similarity is *1*; otherwise, the similarity is *0*.

Generalizing distance clustering algorithms to handle a mix of data varieties often loses the meaning of a metric in an effort to restrain complexity. An example of this is simply dividing the attributes into quantitative and qualitative. Then, apply a different metric to each. Finally, add everything together for the overall similarity measure. Unfortunately, the resulting metric loses the consistency found in individual measures of the same nature. It is open to question if this will be successful in obtaining a good picture of the overall similarity between records. For example, one quantitative attribute can contribute infinitely many possible values; while simple matching on a qualitative attribute may only contribute two possible values, *0* or *1*. Modification to make individual metrics consistent with each other would introduce significant computational complexity..

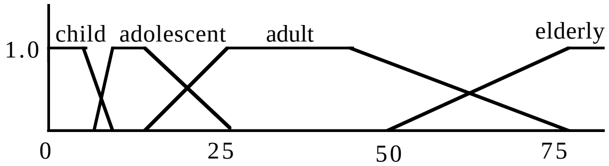
A consequence of the lack of scalar metric is the difficulty in quantitatively measuring the similarity between two values. Ordinal data provides more information than nominal. Ordinal data's progressively increasing values allow the creation of an index that ranks the ordinal data. The increasing index values supply some relative information. A temptation to be avoided is to use the ranked index as a scalar measure. An index built on nominal data provides no information useful for clustering purposes. Sometimes naive workers attempt to apply nominal indices as a scalar measure.

Many data sets contain multiple varieties of data. When clustering records, it is important that this is considered in evaluating record similarity (or conversely, the dissimilarity). Developed clustering methods work sufficiently well for quantitative data where the dissimilarity between two records could be a Euclidean distance measure[*]. An unanswered question is whether it would be useful to use a different geometry or a transformation into a different space (analogously to a dual transformation in operations research).

If the data to be clustered is ordinal, the values can be mapped into rank index values, e.g., *1..m*, according to their natural order. Some authors have tried to use the ranked index as a distance measure (Han, 2001, 344-5) (Li, 1998). The difficulty with mapping ordinal data is that an assumed, artificial scale is imposed on the data. For example, if the values are: {*large, medium, small*} then mapping them as: {*large: 1, medium: 2, small: 3*} has the inference that the similarity between *medium* and *small*, and *medium* and *large* are equal. The same problem occurs when mapping nominal values, except an artificial ordering is also imposed. We believe that these approaches are unsatisfactory as a general approach due to their inherent artificiality. An effort offering greater promise might be to express a non-equal increment scale using fuzzy distributions.

Data can vary in magnitude, whether it is quantitative or qualitative. Sometimes, data is normalized so attributes with different maximum magnitudes can be grouped together by relative size. For example, the range for human male height might be asserted as being up to 220 cm; while the range for grasshoppers might be asserted as being up to 10 cm. If we wanted to group critters as: tiny, small, medium, or large, we might first normalize the data (to a maximum of a convenient value, say 1.0). If the distinction between sizes is linear (it usually is not), then grouping by size is relatively straightforward. If not, then other techniques have to be used.

Often the range of values to be included in a particular group is uneven and the discrimination point between them unclear. For example, if we are grouping humans into: {child, adolescent, adult, elderly} the dividing points both depend on perception. Even potential definition items vary with location and situation; such as: school entry date, required age before leaving school, legal age to vote, age at which it is first possible to retire with full benefits, required age of retirement legal age to have sex[*], legal age to drive.



Sometimes data is expressed across a distribution such as time. How to best normalize data for magnitude, uniformity, and distribution is an active research question in several areas, including genomics. Normalization of ordinal data is more

[*] A variety of other distance metrics are possible; e.g., the Minkowski metric. The Minkowski metric is defined as: $[\Sigma(|x_i-y_i| ^p)]^{1/p}$. Euclidean geometry is the case where p=2.

[*] Age of consent varies widely. Worldwide, the range is none to 21. In the USA, depending on state, the range is 14 to 18. Reference: http://www.ageofconsent.com/ageofconsent.htm

problematic as the ordinal labels may not be uniform or complete.

Data is not always represented linearly. For example, it is common to show data on a logarithmic scale. Clustering logarithmically scaled data is often done. The decision to represent/cluster data logarithmically is usually done by a human. For the purposes of granulation, little work has been done to computationally decide when data should be represented non-linearly (e.g., logarithmically).

There are many differential geometries that may be applied to the data besides Euclidean. Mathematicians have developed many geometries that are primarily theoretical play things. However, non-Euclidean differential geometries are commonly used in physics, astronomy, chemistry, and biology; for example, Riemann space. Various theories are dependent on the non-Euclidean differential geometries; e.g., string theory and space-time. There are transformations that can be used to change Euclidean values to non-Euclidean; e.g., Lorentz transformation. Some of the geometries have higher dimensionality; e.g., Minkowski space unifies Euclidean 3-space with time in Einstein's theory of special relativity. There does not appear to be any work done using differential geometries to partition records.

SIMILARITY/DISSIMILARITY IN MIXED DATA

Many current similarity metrics use pair-wise comparisons in the measurement of the 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

$$\text{sim}(a1,a2) \oplus \text{sim}(b1,b2) \oplus \text{sim}(c1,c2)$$

where $\oplus$ indicates some combination operator and sim(x,y) is a measurement of similarity between the attribute values x and y. We suggest that a more useful inter-record similarity measure would be

$$w_{a(1,2)}*\text{sim}(a1,a2) \oplus w_{b(1,2)}*\text{sim}(b1,b2) \oplus w_{c(1,2)}*\text{sim}(c1,c2)$$

where $w_{Ä}$ represents weights for each attribute pair. How to determine these weights, especially if done computationally, is unclear.

Although the similarity measure between records can be either quantitative or qualitative, many current metrics attempt to derive a quantitative, scalar measure. This is desirable in the clustering task because many clustering methods utilize scalar distances.

There are 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). It is important to note the difference between the *simple matching coefficient* and *simple matching*. *Simple matching* results in a match or no match result (*0* or *1*) as two qualitative values are different or the same. The *simple matching coefficient* (Sneath, 1973) is the proportion of number of matching values to the total number of values compared. When qualitative values are mapped into a form appropriate for metrics suitable for quantitative data, the utility of the measure is lost. For example, starting with a data set containing three records r_i *(fruit*$_i$ *, color*$_i$ *, bag size*$_i$*)*:

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

Let 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*}.
In these cases, the values *1, 2, etc.* are arbitrarily assigned numeric integers. In this example, *bag size* is already a numeric character. Whether it is a quantitative value or an arbitrary qualitative value (ordinal or nominal) is not described. Assume we are using Euclidean distance, defined as $([x_i\text{-}y_i]^2)^{1/2}$, where X and Y are the records being compared and x_i and y_i are the values for the i^th attribute of X and Y. With the mapping π_1, record r_2 has equal distance to both r_1 and r_3. With the mapping π_2, this is not the case. Let d(x,y) represent the distance between records x and y,

$d(r_1, r_2) > d(r_1,r_3)$ with π_1

but

$d(r_1, r_2) < d(r_1,r_3)$ with π_2.

The difficulty is that Euclidean distance is defined for quantitative values, but we are imposing an ordering and a scale not be reflected in the real world. Consequently, it is unlikely that the result is useful. This example demonstrates that we can construct an arbitrary mapping, but we cannot be sure about the utility of the resulting measure.

The difficulty of applying metrics commonly used with quantitative data is not limited to extensions to qualitative data. Guha (2000) provides an example of a problem that can occur when using distance metrics, such as Euclidean distance, on binary data with centroid-based clustering. It is possible to reach a situation where a given record is calculated as closer to one cluster's mean, when it is really not. If all dimensions are Boolean (i.e. there are no quantitative attributes), then a distance computation, such as Euclidean, does not discriminate. A record could be considered close to a mean, when it actually has no values in common with the mean (or less values in common than it has with another record). This occurs regardless of whether the Boolean values were developed from qualitative match/no-match testing or from other sources.

Conversely to the situation of trying to apply quantitative methods to qualitative data, using the metrics developed for qualitative data causes loss of information when applied to quantitative data (Li, 1998). For example, if using simple matching between quantitative values, it is possible to obtain the same pair-wise similarity between values. For example, consider the values 3.4, 3.5, and 4.2. They will have the same pair-wise similarity between them (*0*). The fact that the value 4.2 is more dissimilar to 3.4 than to 3.5 is lost. Even if simple matching were used on quantitative values that have been mapped to discrete intervals such as [3.0,3.5], information loss is still likely. In this case, much like the last, the loss is the ordering of objects imposed by similarity. Considering the same values with 3.1 in addition, 3.1, 3.4, and 3.5 will now be pair-wise considered to be the same (i.e., their pair-wise similarity will be *1*). In both cases, the ordering imposed by a similarity measure will be lost.

Some qualitative clustering approaches cluster values extracted from the records rather than the records themselves (Gibson, 2000) (Han, 1997) (Zhang, 2000). Most qualitative clustering methods simply group items together. They do measure of closeness for any two particular values. For example, if it is discovered that *a*, *b* & *c* belong to the same cluster, how can we decide whether *a* closer to *b* or to *c*? If this could be derived, then it could be used in the clustering of whole records.

Gibson (2000) and Zhang (2000) use a dynamic system to propagate weights from an initial value. The weight is propagated according to frequency and co-occurrences. The weights are then used to consider two clusters, one containing the initial value and the other its complement. It is not clear whether the resulting weights can be used to compare pair-wise similarity. Han's (1995) approach uses a hypergraph to represent the frequencies and weights the graph edges according to the co-occurrences. The metrics used in these approaches are based on the frequency of the values and their occurrences together in the records. This is directed towards the problem discussed by Kanal (1993) when he discussed Watanabe's (1969, 1985) "Theorem Of The Ugly Duckling;" namely the need to have weighted memberships. One way of achieving weighted memberships is to use soft computing tools.

Wang (1999) and Ganti (1999) also use the value's frequency to find the clustering. Wang uses the metric in a function to evaluate clustering goodness. Values are termed *large* if they are present within a cluster above a threshold (human specified). If values are not found to be *large* in a cluster, they are termed *small*. By determining two sets for each cluster, one for *large* values and the other *small* values, an evaluation function can be defined. The goal is to minimize this function over any possible clustering. The evaluation function is a sum of the inter-cluster measure and intra-cluster metrics. A weight is provided to allow for more emphasis on either the inter or intra-cluster similarity. The intra-cluster similarity measure is the count of small values over all of the clusters. The inter-cluster similarity measure is the number of values that are large in multiple clusters (i.e. overlap between clusters).

An approach that does not use frequency for the metric is the extension to the k-means algorithm by Huang (1997) (1999) and Ralambondrainy (1995). Originally, the distance between records is a quantitative metric such as Euclidean distance. They modify the distance based k-means algorithm to handle categorical data by combining two measures, one for the quantitative values and one for the qualitative values. The qualitative measure is the sum of the reverse of simple matching for categorical attributes (*1* or *0*) indicating non-matching and matching respectively. The distance between two records is then a weighted sum of the quantitative value's measure and the categorical value's measure. In (Huang, 1999), similarity is computed as the sum of square differences for the numerical attributes added to a weighted summation of matches for the categorical attributes.

This approach does not attempt to fit one type of metric to all kinds of data present. It attempts to find a meaningful way to combine multiple metrics to obtain the overall distance. The similarity for two records is a combination of two metrics, one for quantitative and one for categorical. Huang weights only the qualitative measure, while Ralambondrainy (1995) weights both measures. Ralambondrainy offers a possible way to weight attributes.

In both cases (Huang, Ralambondrainy), the discovery of an appropriate weighting adds complexity to the algorithm. With a meaningful weighting, the numerical values maintain and contribute their magnitude; while the categorical values, with no magnitude to contribute, can contribute proportionally to the measure. Without a good weight parameter, a difficulty arises because the metric loses the meaning of "similar." This point was previously made by Goodall (1966). The following example shows the difficulty with the added weight(s) parameter. If the data set is:

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

The respective dissimilarity using Huang's approach with weight of 1 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 $d(t_1,t_2)=1$ and $d(t_1,t_3)=9$ where $d(t_i,t_j)$ is the distance (dissimilarity) for objects t_i and t_j. Clearly, this is a case of the quantitative measure dominating the qualitative measure. An algorithmic question is: How should the mismatching qualitative values between t_1 and t_2 contribute to the measure while the quantitative difference between t_1 and t_3 contributes according to its magnitude? One option is to find a quantitative metric consistent across all of the attributes being considered in the measure. This follows from the idea of normalizing the data before the computation of similarity (Everitt, 1993). Another option finding appropriate attribute weights. Huang (1997) suggests that the weight be selected according to the distribution of the quantitative attributes. In the example above, using the standard deviation as the weight gives minimal change to the distances in the example.

Li (1998) developed their method using the Goodall similarity metric (Goodall, 1966). This metric measures the amount of weight that a categorical value contributes to the overall similarity measure. For example, if two records have the same value for a qualitative attribute k, then the similarity is not necessarily the values returned from match/non-match (*0,1*) that most qualitative similarity metrics assign. The given value assigned for a match between qualitative values is therefore a real number between zero and one. This similarity value that is assigned is decided according to the frequency of the value within the data. Let $s(a,b,c,d)$ represent the fact that c is more similar to d than a is to b. Then when computing the measure for two quantitative values, say x and y, the probability of other pairs of values, s and t, such that $s(x,y,s,t)$ is used. This use of the frequency distribution of values allows for a more meaningful measure. The difference in magnitude for quantitative values is used in the deciding of $s(a,b,c,d)$. The chi-squared (χ^2) statistic is used for computing the measure between records. When using this statistic, there is an assumption of independence among the attributes.

Often independence among attributes cannot be guaranteed. Many methods have the underlying requirement of attribute independence. However, often there is dependence. How to handle dependent attributes is a significant, unsolved question.

The distributions of the quantitative values offer added information in the computation of similarity. It is not uncommon that data sets are biased. That is, they are not representative of the population as a whole; e.g., a data set of diabetic patients. In this case, if two records have matching qualitative values, if the value is commonly found with diabetics, then it should not be as significant as matching values that are not common.

The above discussion includes two metrics, one based on dissimilarity (or distance), Huang's metric, and the other based on similarity, Li's metric. We cannot directly compare the two metrics, but we can attempt to compare them in an indirect way. To begin with, if we think about dissimilarity as the complement of similarity, then it is quite possible that the two metrics are not consistent with each other. It is difficult to discover whether the metrics actually correspond to each other. If we are given one metric, say similarity, we cannot easily compute its complement (dissimilarity). Richter (1992) refers to some properties common to similarity and dissimilarity metrics.

EPILOGUE

Our goal is to group records containing multiple data varieties: quantitative (discrete, continuous) and qualitative (ordinal, nominal). This paper is an initial consideration of strategies for integrating multiple metrics in the task of granulating records.

Data mining a data set composed of varied data stored in records can focus on either: granulating individual attributes, data extracted from the records making up a data set, or whole records. To cluster whole records in a data set, it would be useful to have a general similarity metric that would allow record to record similarity comparison. There are methods to granulate data items belonging to a single attribute. Few methods exist that meaningfully account for a combination of many data varieties in a single metric.

Most available algorithms work with quantitative data. However, many data sets contain a mixture of quantitative and qualitative data. Our goal is to group records containing multiple data varieties: quantitative (discrete, continuous) and qualitative (ordinal, nominal). This is a difficult task. Even grouping based on different quantitative metrics can be difficult. There are several known partially successful strategies. Incorporating qualitative elements is not simple. We expect to use a mixture of scalar methods, soft computing (rough sets, fuzzy sets), as well as methods using other metrics. Potentially, there are also several possible differential geometries that might be applied.

For a metric to cluster records in a data set, it would be useful to have a single 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.

It appears that modification to make individual metrics consistent with each other may introduce significant computational complexity. How to best do this is an open question.

Finding a way to integrate or combine a different metrics offers more promise than developing a metric general enough to use on all types of data. The lack of magnitude and scale in nominal data creates a difficulty in discovering the weighting needed to develop a useful metric. So far, methods to develop generalized distance clustering algorithms to handle a mix of data varieties often lose the meaning of a metric in an effort to restrain complexity.

To cluster whole records in a data set, it would be useful to have a general similarity metric or a set of integrated similarity metrics that would allow record to record similarity comparisons. Our research seeks to accomplish this.

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, p 9-13

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*, p 21-34

A. Jain, R. Dubes (1988) **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*

H. Ralambondrainy (1995) "A Conceptual Version of the K-Means Algorithm," *Pattern Recognition Letters 16*, p 1147-1157

M.M. Richter (1992) "Classification and Learning of Similarity Measures," *Proceedings der Jahrestagung der Gesellschaft fur Klassifikation*, Studies in Classification, *Data Analysis and Knowledge Organisation*, Springer Verlag

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