Dissimilarity Computation for Objects of Different Variable Types

Zin Nyein Nyein Han, Ei Ei Moe Tun
Computer University, Magway
zinnyeinhan7@gmail.com, eieimoetun@gmail.com

Abstract

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the object in other cluster. Measuring the dissimilarity between data objects is one of the primary tasks for distance-based techniques in data mining and machine learning, e.g., distance-based clustering and distance-based classification. The quality of clustering can be accessed based on dissimilarity measures of objects which can be computed for various types of data. In this paper, we propose general framework for measuring a dissimilarity betweens various data analysis is proposed. The key idea is to consider the dissimilarity between two values of an attribute as a combination of dissimilarities between the conditional probability distributions of other attributes given these two values. In this system, the similarity is guessed by computing the dissimilarity measure between two objects. This can get the most similar values and the least similar values before clustering analysis.

Keywords- dissimilarity measure; cluster analysis; mixes types objects.

1. INTRODUCTION

Data mining, "knowledge discovery", or "machine learning" methods have many origins, drawing on insights from research on learning as it naturally occurs in humans (cognitive science), advances in computer science and algorithm design on how to
best detect automatically patterns in "unstructured" data, engineering and advances in machine learning (e.g., neural networks), to name a few [1].

Clustering of multidimensional data is one of the main tools in Knowledge Discovery from Data (KDD), a field that emerged from the need to extract useful information from the vast amount of data generated by simulations or measurements. The most often used measure of similarity is the Euclidean distance between the vectors representing the data features. Cluster analysis among various applications is the important issue in these days. Advantages of cluster-based process are that it is adaptable to change and helps single out useful features that distinguish different groups. The dissimilarity measure of various data objects can be used before cluster analysis [1].

Clustering is a data analysis technique in which a measure of similarity, or equivalently a measure of dissimilarity, is used to detect groups or patterns in data. Traditionally, these similarity and dissimilarity measures have been related linearly [2]. Measure of similarity can be computed for various types of data as dissimilarity measure between categorical data [3].

In this system, the dissimilarity measure based various data analysis is proposed for cluster analysis. In our paper, we used the various data types of objects such as interval-scaled variables, binary variables, nominal, ordinal, and ratio-scaled variables for measuring dissimilarity between these objects. By making this system, various data can be analysis before cluster analysis as preprocessing stage. The rest of the paper is organized as follows. In Section 2, we describe the background theory of our proposed system. Section 3 presents the overview of proposed system. In section 4 we can describe the proposed dissimilarity measures between various data objects and some evaluation analysis. In this section, we show sample computation with example heart data set. We conclude our proposed dissimilarity measure for cluster analysis in Section 5.

2. BACKGROUND THEORY

2.1 Cluster Analysis

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications [1].

Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, and market research.

Alternatively, it may serve as a preprocessing

step for other algorithms, such s characterization and classification, which would then operate on the detected clusters.

2.2 Similarity and Dissimilarity Measure

Similarity is expressed in terms of a distance function, which is typically metric and it is quite difficult to measure. Similarity is quantity that reflects the strength of relationship between two objects or two features. This quantity is usually having range of either -1 to +1 or normalized into 0 to 1. If the similarity between feature *i* and feature *j* is denoted by *d* (*i, j*), we can measure this quantity in several ways depending on the scale of measurement (or data type) that we have.

Dissimilarity measure the discrepancy between the two objects based on several features. Dissimilarity may also be viewed as measure of disorder between two objects. Each similarity or dissimilarity has its own characteristics [4].

2.3 Types of Data Cluster Analysis

In this section, the types of data that often occur in cluster analysis and how to preprocess them for a cluster analysis will describe. Suppose that the data set to be clustered contains *n* objects, which may represent persons, houses, documents, countries, and so on. Typically, usual clustering algorithms operate on either or the following two data structures [1].

1. Data matrix/two-mode matrix (or object-by-variable structure): This represents n objects, such as persons, with *p* variables (also called measurements or attributes), such as age, height, gender, race and so on. The structure is in the form of a relational table, or *n*-by-*p* matrix (*n* objects x *p* variables) as shown in Figure (1).

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Figure 1. N-by-P matrix

2. Dissimilarity matrix/one-mode matrix (or object-by-object structure): This stores a collection of proximities that are available for all pairs of *n* objects. It is often represented by n-by-n tables as shown in Figure (2):

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Figure 2. Dissimilarity matrix

Where *d (i,j)* is the measure difference or dissimilarity between objects *i* and *j* . In general, *d (i,j)* is a nonnegative number that is close to 0 when objects *i* and *j* are highly similar or "near" each other, and becomes larger the more they differ. In this paper, we used the dissimilarity matrix as our data structure for analyzing between two objects [1].

2.4 Data Types for Dissimilarity Measure

There are five data types that used in object dissimilarity computation as follows:

Interval-scales Variables: are continuous measurements of a roughly linear scale. These measures includes the *Euclidean, Manhattan, and Minkowski distances.* Examples like Weight and height, latitude and longitude coordinate weather temperature. Dissimilarity Distance Measure measures the similarity or dissimilarity between two data objects namely Euclidean as shown in formula (1):

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots\ldots\ldots + (x_{in} - x_{jn})^2} \quad (1)$$

Binary Variables: We often face variables that has only binary value such as Yes and No, or Agree and Disagree, True and False, Success and Failure, 0 and 1, Absence or Present, Positive and Negative, etc. For such binary variables, there are only two possible values, which can be represented as positive and negative. Similarity or dissimilarity (distance) of two objects that represented by binary variables can be measured in term of number of

occurrence (frequency) of positive and negative in each object[1].

Table 1. Binary values for two objects

Feature of Fruit	Sphere shape	Sweet	Sour	Crunchy
Object *i* =Apple	1	1	1	1
Object *j* =Banana	0	1	0	0

For example: Table (1) shows the objects Apple (*i*) and Banana (*j*) with related binary values. The coordinate of Apple is (1,1,1,1) and coordinate of Banana is (0,1,0,0). Because each object is represented by 4 variables, we say that these objects have 4 dimensions.

Let:

q = number of variables that positive for both objects

r = number of variables that positive for the i_{th} objects and negative for the j_{th} object

s = number of variables that negative for the i_{th} objects and positive for the j_{th} object

t = number of variables that negative for both objects

$p=t+q+r+s$ is total number of variables

Table 2. A contigency table for binary variables

		Object *j*	
		1	**0**
Object *i*	**1**	*q*	*r*
	0	*s*	*t*

The most common use of binary dissimilarity (distance) is described in formula (2).

$$d(i,j)=\frac{r+s}{q+r+s+t} \quad (2)$$

Nominal/Categorical Variable: In many cases, that cannot measure variable in quantitative way, but it is possible to measure in term of category. A nominal or categorical variable is used when number is only a symbol to represent something. For example, the fruits I like are categorized as follows:

1 = Apple, 2 = Banana and 3 = Orange.

Gender is a nominal variable with value of

1 = male and 2 = female

The dissimilarity between two objects *i* and *j* can be computed based on the ratio of mismatches as shown in formula (3).

$$d(i,j)=\frac{p-m}{p} \quad (3)$$

Where *m* is the number of matches (i.e the number of variables for which *i* and *j* are in the same state) and *p* is the total number of variables.

Ordinal Variables: Number usually has order. When there have sequence of number 1, 2, 3, these can be said that 3 is higher than 2 and 1, while 2 is higher than 1. Ordinal scale play very important role in behavioral survey because it is relatively easy to design, easy to answer by respondent.

Both ordering and rank are ordinal variables, though the labels are category. Nominal variable is best represented as existence of the choice, without order**.** Ordinal variable emphasize the sequence, or order of the choice. The dissimilarity computation with respect to *f* that means a variable from a set of ordinal variables describing n objects, involves the following steps:

1. The value of *f* for the i_{th} object is X*if* ,and f has M*f* ordered states of ranking 1,……,M*f*. Replace each X*if* by its corresponding rank , r*if* $\in$ {1,…..,M}.
2. Since each ordinal variable can have a different number of states, it is necessary to map the range of each variable onto [0.0, 1.0] so that each variable has equal weight .This can be achieved by replacing the rank r_{if} of the i_{th} object in the f_{th} variable by formula (4).

$$z_{if}=\frac{r_{if}-1}{M_f-1} \quad (4)$$

3. Then dissimilarity can be computed by using distance measure for interval-scale variables, using *zif* to represent the *f* value for the i_{th} object.

Ratio- scale Variables: **A** positive measurement on

a multiplicative scale, corresponding to exponential growth: such as *AeBt* or *Ae-Bt*, e.g., growth of a bacteria population. **Methods:** treat them like interval-scaled variables and *a*pply logarithmic transformation as formula (5).

$$y_{if} = \log(x_{if}) \quad (5)$$

2.5 Variables of Mixed Type

One approach to compute the dissimilarity between objects of mixed variable types is to group each kind of variable together, performing a separate cluster analysis for each variable type. The data set contains *p* variables of mixed type. The dissimilarity *d* (i, j) between object i and j is defined as in formula (6).

$$d(i, j) = \frac{\sum_{f=1}^{P} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{P} \delta_{ij}^{(f)}} \quad (6)$$

Here $\delta_{ij}^{(f)} = 0$ if (1) x_{if} or x_{jf} is missing,

(2) $x_{if} = x_{if} = 0$ and

$\delta_{ij}^{(f)} = 1$ Otherwise

The contribution of variable f to the dissimilarity between i and j, that is, computed dependent on its type:

-if *f* is interval –based: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$

where h runs overall no missing objects for variable *f*.

-if *f* is binary or categorical: = 0
if $x_{if} = x_{if}$ otherwise = 1.

-if *f* is ordinal: compute the ranks r_{if} and

$z_{if} = \frac{r_{if} - 1}{M_f - 1}$, treated z_{if} as interval scaled.

-if *f* is ratio scaled: either perform log arithmetic transformation and treats the transformed data as interval-scaled.

3. OVERVIEW OF PROPOSED SYSTEM

In this system, the dissimilarity measure for cluster analysis before clustering is proposed. By measuring the dissimilarity, user can guess which two objects are the most similar and other two objects are least similar for clustering different types of data objects. Figure 3 show the system flow diagram of proposed system.

3.1 System Flow Diagram of Proposed System

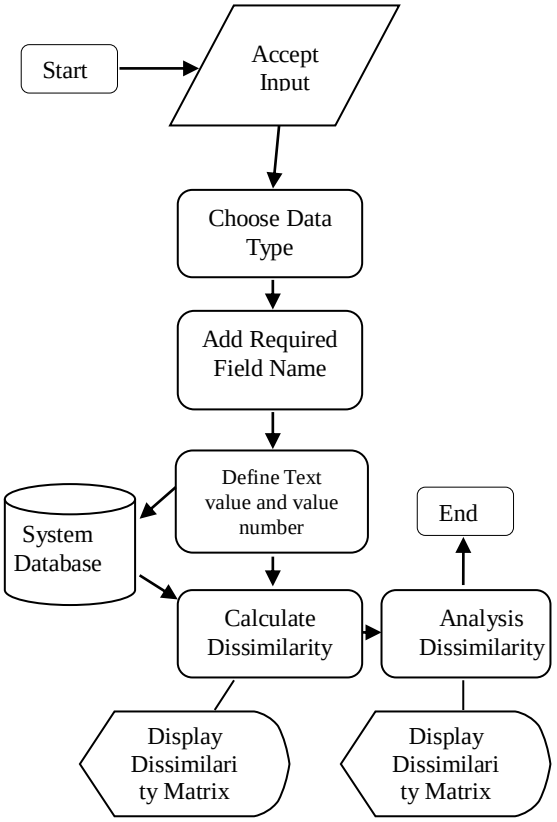

Figure 3. System flow diagram of proposed system

3.2 Algorithm for Computing Dissimilarity

In Figure (5), the algorithm for computing dissimilarity between various types of data objects is described.

Input: Data set *D*
Output: Dissimilarities between data objects
BEGIN
Step1. Choose two data objects *i* and *j*
Step2. Define the data values by data types
Step3. For any pair of values *i* and *j* of data objects, Compute dissimilarity *d (i, j)*
Step4. Analysis dissimilarity results
END

Figure 5. Algorithm for computing dissimilarities

4. EXMERIMENTAL ANALYSIS

In proposed system, we use the heart diseases data set as sample input data source. It includes the various attributes as shown in Table (3). Some record sets are shown in Table (4) for computing dissimilarity measure between data objects. According to the results, we can get exact results and can analysis which two are most similar and which are least similar before cluster analysis. We also compute the dissimilarity measure between mixed types of variables in this data set.

Table 3. Attributes of heart data set

Attribute Name	Data Type	Description
age	Ratio Scale	Age of patients
sex	Binary	Male, Female
Chest Pain	Categorical	Chest Pain Type (4 values)
Rest BP	Ratio Scale	Resting Blood Pressure
Cholesterol	Ratio Scale	serum cholesterol in mg/dl
Blood Sugar	Binary	fasting blood sugar > 120 mg/dl
ECG	Categorical	resting electrocardiography results (values 0,1,2)

Table 4. Sample heart (record) dataset

Old	Age	Sex	CP	RBP	Chop	BS	ECG
1	60	1	4	130	206	0	2
2	54	1	4	110	239	0	0
3	54	0	2	132	288	1	2
4	55	0	4	180	327	0	1
5	41	1	2	135	203	0	0

4.1 Defining Data Values by Data Types

1. Sex-> 1=Male, 0=Female
2. CP -> 1,2,3,4
3. BS -> 0=absent, 1=present
4. ECG-> 0, 1, 2

4.2 Example Calculation

For Chest Pain (Categorical Data)

$$d(2,1) = \frac{1-1}{1} = 0$$

……...

$$d(5,4) = \frac{1-0}{1} = 1$$

Output Dissimilarity Matrix of chest pain data

$$\begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 1 & 1 & 0 & & \\ 0 & 0 & 1 & 0 & \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$

We have calculated the all data types of 7 in sample heart data set and also computed the dissimilarity between variables of mixed types as follows:

$$d(2,1) = \frac{1(0)+1(0.33)+1(0.33)+.....+1(0)}{7} = 0.27$$

……..

$$d(5,4) = \frac{1(0)+1(0.33)+1(0.33)+.....+1(0)}{7} = 0.68$$

According to dissimilarity matrix of mixed

type's variables, we can guess the analysis results as follows:

d (2,1) is the lowest value for any pair of different objects, so objects 2 and 1 are the most similar

d (5,4) is the highest value for any pair of different objects, so objects 5 and 4 are the least similar

Output Dissimilarity Matrix of Mixed types data

$$\begin{pmatrix} 0 & & & & \\ 0.27 & 0 & & & \\ 0.59 & 0.52 & 0 & & \\ 0.48 & 0.45 & 0.67 & 0 & \\ 0.53 & 0.53 & 0.63 & 0.68 & 0 \end{pmatrix}$$

5. CONCLUSION

In this paper, we proposed a general framework for computing dissimilarity measure between objects of various types. By computing this dissimilarity, user can guess which objects are most similar and which ones are least similar. In this system, user can compute any data types or data sets according their desires. This system can enhance the clustering analysis before clustering among various objects and can get exact results in clustering process.

REFERENCES

[1] J.Han and M.Kamber. *Data Mining: Concepts and Techniques*, Simon *Fraser University*, 2008

[2] A. Socolovsky. "ADissimilarity Mesure for Clustering High –and Infinite Dimensional Data" Hampton University, Hampton, Virginia, 2002

[3] S.Q. Le and T .B .Ho, "An association-based dissimilarity measure for categorical data", *ELSEVIER, Pattern Recognition Letters 2549–255*, 26, 2005

[4] D. Malerba, F. Esposito, V. Gioviale and V. Tamma, "Comparing Dissimilarity Measures for Symbolic Data Analysis", 2003