The Calculation of Similarity and Its Application in Data Mining

Shaohua Teng, Junlei Li, Rigui Li, and Wei Zhang

School of computer science and technology, Guangdong University of Technology, Guangzhou Guangdong 510006, China
{Shteng,weizhang}@gdut.edu.cn, {609376502,836386635}@qq.com

Abstract. The Similarity is a measure, which is used to measure the strength of the relationship between two objects and their closely degree. According to different object types, similarity calculation method is also different. Similarity calculation is widely used in classifing data, it is the basis of object classification. In this paper, the data objects were divided into three kinds: numerical type, non numeric type and mixed type. And these similarity calculation methods of different types are discussed. Finally, we illustrated the application of similarity in the data classification and data cluster.

Keywords: similarity, object, data mining, data type.

1 Introduction

With the development of data mining technology, its application is more and more broad. The similarity calculation of the object has become a very important topic. The Similarity measure is an important means to measure the strength of the relationship between variables and their closely degree.

The tasks of data mining mainly involves: data classification and prediction [1-5], data clustering [1-6], association rules [1-5], sequential patterns [1-7], the dependent relations or the dependent model [1-7], abnormal and trend [1-5] and so on. They all deeply depend on the similarity calculation. The Measurement and calculation of the similarity plays a decisive role in the application of data mining, involving almost the majority of data mining algorithms, especially to study the same degree between objects and their relations, the similarity calculation is an important functionality of data preprocessing, the measure to calculate the objects' similarity directly influences the final mining result. Therefore, different measures and different methods are used to calculate the similarity of different data types, which depend on their data types and their practical applications. The study of the measures and calcualation methods about the similarity is very important.

The concept of similarity is firstly presented in the paper, and then the data objects are divided into three types: numeric type, non-numerical type and mixed type. These different similarity calculation formulas for different data objects are proposed. Finally the applications of the similarity calculation are given through some examples.

Q. Zu et al. (Eds.): ICPCA/SWS 2013, LNCS 8351, pp. 563–574, 2014.

2 The Similarity

On data mining [1-5], machine learning [3-4, 6-8], pattern recognition [6-8], statistical analysis [6-8], and other computer applications, the similarity is used to present a measure of similarity between two objects. Informally, the similarity between two objects is numerical measure of the degree to which the two objects are alike. The more similar between them, the higher is the similarity meseasure. Similarities are usually nonnegative, which are between 0 (no similarity) and 1 (complete similarity) [2].

The dissimilarity between two objects is a numerical measure of the degree to which the two objects are different [2]. The lower the dissimilarity between objects, the more similar is the pairs of objects. Frequently, the term distance is used as a synonym for dissimilarity, although it is often used to refer to a special class of dissimilarities.

Most data mining algorithms are related to the similarity measure between objects, but the similarity definition and its calculation are different with different data types. For example, the similarity measure of two numerical objects is often converted to the dissimilarity, the latter is usually an Euclidean distance. It is used to present the degree of diversity between two objects; and for the nominal data, the similarity between two objects is related to the number of same values which are from their corresponding attributes.

According to the data types of attribute values, we divide them into three data types: numeric type, non numeric type and mixed type [1-2].

(1) Numeric: also known as the continuous or quantitative variables, namely there are infinite values between two different values of the specific numerical attributes. Frequently, natural numbers or the units of measurement are used to measure similarities directly, such as temperature, height etc. Numeric variables can be divided into interval-scaled variables and scale variables, while the interval-scaled variables are a linear scale variable, and the scale variables are generally nonlinear.

(2) Non-numerical: the attribute values are non-quantitative, but qualitative data, such as, a person's gender or the excellent grades of achievements etc. Usually this kind of attribute values is a finite number of states (letter or ordinal number). Non-numeric attributes can be changed into nominal attributes, binary attributes and ordinal attributes. Nominal attributes are unordered, while ordinal are ordered.

(3) Mixed type: usually refers to mixed by numeric and non-numeric type.

3 The Similarity Calculation

3.1 Similarities between Numerical Objects

The numerical data can be divided into the interval-scaled variables and the similarity measure is the proportion of numerical variables.

3.1.1 The Interval-Scaled Variables

The interval-scaled variables: it is roughly a linear scaling of the continuous variables, interval value of the property is its order. It can be positive or negative, also zero. The similarity of interval attribute is usually converted to dissimilarity. Processing method is used to standardize the variables commonly. The dissimilarity between objects is calculated based on the distance between them. Here is the common calculation method