Discovering an Effective Measure in Data Mining

Takao Ito
Ube National College of Technology, Japan

INTRODUCTION

One of the most important issues in data mining is to discover an implicit relationship between words in a large corpus and labels in a large database. The relationship between words and labels often is expressed as a function of distance measures. An effective measure would be useful not only for getting the high precision of data mining, but also for time saving of the operation in data mining. In previous research, many measures for calculating the one-to-many relationship have been proposed, such as the complementary similarity measure, the mutual information, and the phi coefficient. Some research showed that the complementary similarity measure is the most effective. The author reviewed previous research related to the measures in one-to-many relationships and proposed a new idea to get an effective one, based on the heuristic approach in this article.

BACKGROUND

Generally, the knowledge discover in databases (KDD) process consists of six stages: data selection, cleaning, enrichment, coding, data mining, and reporting (Adriaans & Zantinge, 1996). Needless to say, data mining is the most important part in the KDD. There are various techniques, such as statistical techniques, association rules, and query tools in a database, for different purposes in data mining. (Agrawal, Mannila, Srikant, Toivonen & Verkamo, 1996; Berland & Charniak, 1999; Caraballo, 1999; Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996; Han & Kamber, 2001).

When two words or labels in a large database have some implicit relationship with each other, one of the different purposes is to find out the two relative words or labels effectively. In order to find out relationships between words or labels in a large database, the author found the existence of at least six distance measures after reviewing previously conducted research.

The first one is the mutual information proposed by Church and Hanks (1990). The second one is the confidence proposed by Agrawal and Srikant (1995). The third one is the complementary similarity measure (CSM) presented by Hagita and Sawaki (1995). The fourth one is the dice coefficient. The fifth one is the Phi coefficient. The last two are both mentioned by Manning and Schutze (1999). The sixth one is the proposal measure (PM) suggested by Ishiduka, Yamamoto, and Umemura (2003). It is one of the several new measures developed by them in their paper.

In order to evaluate these distance measures, formulas are required. Yamamoto and Umemura (2002) analyzed these measures and expressed them in four parameters of a, b, c, and d (Table 1).

Suppose that there are two words or labels, x and y, and they are associated together in a large database. The meanings of these parameters in these formulas are as follows:

a. The number of documents/records that have x and y both.
b. The number of documents/records that have x but not y.
c. The number of documents/records that do not have x but do have y.
d. The number of documents/records that do not have either x or y.
n. The total number of parameters a, b, c, and d.

Umemura (2002) pointed out the following in his paper: "Occurrence patterns of words in documents can be expressed as binary. When two vectors are similar, the two words corresponding to the vectors may have some implicit relationship with each other." Yamamoto and Umemura (2002) completed their experiment to test the validity of these indexes under Umemura's con-

Table 1. Kind of distance measures and their formulas

No	Kind of Distance Measures	Formula
1	the mutual information	$I(x_1; y_1) = \log \frac{an}{(a+b)(a+c)}$
2	the confidence	$conf(Y \mid X) = \frac{a}{a+c}$
3	the complementary similarity measure	$S_c(\vec{F}, \vec{T}) = \frac{ad - bc}{\sqrt{(a+c)(b+d)}}$
4	the dice coefficient	$S_d(F, T) = \frac{2a}{(a+b)+(a+c)}$
5	the Phi coefficient	$\varphi_{DE} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
6	the proposal measure	$S(\vec{F}, \vec{T}) = \frac{a^2 b}{1+c}$

cept. The result of the experiment of distance measures without noisy pattern from their experiment can be seen in Figure 1 (Yamamoto & Umemura, 2002).

The experiment by Yamamoto and Umemura (2002) showed that the most effective measure is the CSM. They indicated in their paper as follows: "All graphs showed that the most effective measure is the complementary similarity measure, and the next is the confidence and the third is asymmetrical average mutual information. And the least is the average mutual information" (Yamamoto and Umemura, 2002). They also completed their experiments with noisy pattern and found the same result (Yamamoto & Umemura, 2002).

MAIN THRUST

How to select a distance measure is a very important issue, because it has a great influence on the result of data mining (Fayyad, Piatetsky-Shapiro & Smyth, 1996; Glymour, Madigan, Pregibon & Smyth, 1997) The author completed the following three kinds of experiments, based upon the heuristic approach, in order to discover an effective measure in this article (Aho, Kernighan & Weinberger, 1995).

RESULT OF THE THREE KINDS OF EXPERIMENTS

All of these three kinds of experiments are executed under the following conditions. In order to discover an effective measure, the author selected actual data of a place's name, such as the name of prefecture and the name of a city in Japan from the articles of a nationally circulated newspaper, the *Yomiuri*. The reasons for choosing a place's name are as follows: first, there are one-to-many relationships between the name of a prefecture and the name of a city; second, the one-to-many relationship can be checked easily from the maps and telephone directory. Generally speaking, the first name, such as the name of a prefecture, consists of another name, such as the name of a city. For instance, Fukuoka City is geographically located in Fukuoka Prefecture, and Kitakyushu City also is included in Fukuoka Prefecture, so there are one-to-many relationships between the name of the prefecture and the name of the city.

The distance measure would be calculated with a large database in the experiments. The experiments were executed as follows: