社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－ICS－138 (12)
2004／12／4

A Mixed Similarity Measure for Data with Numeric, Symbolic and Ordinal Attributes

Nguyen Ngoc Binh, Than Van Cuong, Nguyen Thanh Phuong
Hanoi University of Technology, Vietnam
{binhnn, phuongnt}@it-hut.edu.vn

Ho Tu Bao
Japan Advanced Institute of Science and Technology, Japan
bao@jaist.ac.jp

Abstract: Many methods of knowledge discovery in databases are distance-based, such as instance-based learning or clustering where similarity measures between objects plays an essential role. Besides, it is known that most our real-word data not only contain numeric, symbolic, and ordinal attributes individually but also carry all of them in mixed way. Therefore, a Mixed Similarity Measure (MSM) for numeric and symbolic attributes is not enough for various data process. Moreover, the high cost of $O(n^2 \log n^2)$ and $O(n^2)$ for time and complexities of the existing algorithms do not allow the MSM to be applied to large datasets in KDD. As a result, we have proposed a fast algorithm to compute the Goodall's MSM for numeric and symbolic attributes in a linear complexity. In this paper, as an extension of the MSM, we consider an MSM* for numeric, symbolic and ordinal attributes and describe a fast algorithm for MSM* with a linear complexity as well. The experimental results show that the proposed MSM* is also better than MSM and C4.5/See5.0 for the classification problem.

1. Introduction

Recently, some authors have proposed mixed similarity measures (MSMs) [1] for treating mixed symbolic and numeric data, for example those of C. Li and Biswas in [2]. These measures calculate the similarity between objects without discretizing numeric data and they seem promising as reported.

Particularly, the authors in [2] introduced a computing method for an MSM proposed by Goodall [1] for biological taxonomy. This MSM has been shown to be contributed greatly to clustering or classification as reported by Li and Biswas [2]. Their computation method has also been shown to be more efficient than the original computing method (with direct implementation of the MSM), but there are still some considerable limitations when applying it to KDD, where databases are always large or very large. Their method is with high costs of computation with the time complexity of $O(n^2 \log n^2)$ and the space of $O(n^2)$, where n is the maximum number of unique values meet in the database for any numeric attributes. As it will be shown in our experiments, such an MSM is not applicable even for moderately sized datasets.

In this paper, we propose a new method for calculating the same mixed similarity measure of Goodall but with lower cost of the $O(n)$ time and $O(n)$ space. Importantly, we show a fast algorithm to compute data with not only symbolic and numeric attribute but also ordinal attribute. The rest of the paper is organized as follows: Section 2 summaries Goodall's MSM; Section 3 describes our proposed computing methods with new and fast algorithms; Section 4 gives some experimental results to show the efficiency of the proposed algorithms and the comparison our solution with See5.0 in classification problem; and Section 5 summaries our work and gives some directions of applying our method to KDD.

2. MSM

The mixed similarity measure by Goodall in [1] is based on considering first the similarity in respect of single attributes, then combining the similarities in respect of different attributes. The problem is how to measure the similarity or dissimilarity between two

instances in the context of the given databases. We here describe Goodall's MSM for both numeric and symbolic attributes.

2.1. Symbolic attributes

The first type of attributes to be considered is symbolic, as purely qualitative, where the different values are notcapable of being measured. Here, pairs of differing values are all regarded as equally dissimilar, but pairs of values which agree are ordered according to the rule: agreement between two instances in possessing an uncommon value of the attribute is considered as indicating closer similarity than agreement in possessing a commoner value. The possible pairs of values having been ordered in this way, their probabilities are then calculated, and the measure of similarity for any pair is then the complement of the sum of the probabilities for all pairs of values equal or greater in similarity.

2.2. Numeric attributes

The numeric values as metric data: they have the order and the difference between any two values. Pairs with identical values are more similar than those which differ, those with a small difference are more similar than those with a larger difference. As between pairs of values differing by the same amount, however, it appears desirable to take into account the size of the groups encompassed by the two values. Again, once these ordering relations have been established, the degree of similarity is expressed by the complement of the probability of the observed pair or any more similar.

2.3. Mixed attributes

In fact that, objects consist both symbolic and numeric attributes. So, to compute the similarity measure (the distance), firstly we have to compute the similarity measure (the distance) for each attribute, then average (by weight) these measures, as follow:

$$s(x,y)=\frac{\sum_{j=1}^{m_i} w_j s_j(x,y)}{\sum_{j=1}^{m_i} w_j}$$

Where, s(x, y) is similarity measure between two objects x, y at the attribute jth and wis the weight.

2.4. Combination of similarities in respect of different attributes

Goodall, Li and Biswas have also described a method to combine similarities in respect of different attributes by using Fisher's transformation and Lancaster's transformation as follow:

Denote Og and O_h are two objects that is needed to compute the similarity. P_{gh}^{k} is the probability of a random pair of value at the attribute k will be as similar as, or more similar than, the pair $\left(v_g^k, v_h^k\right)$. Then, to reduce the number of the operations we can approximate the similarity S through Fisher's (for symbolic attributes) and Lancaster's (for numeric attributes) transformations as follow:

Fisher's transformation:

$$\mathrm{P}=e^{-\chi^2/2}\sum_{i=0}^{t}\frac{\left(1/2\,\chi^2\right)^i}{i!}$$

where $t=t_d+t_c-1$ and $\chi^2=\chi_d^2+\chi_c^2$, with

$$\chi_c^2=-2\sum_{k=1}^{t_c}\ln P^k$$

Lancaster's transformation:

$$\chi_d^2=2\sum_{k=1}^{t_d}\left\{1-\frac{P^k\ln P^k-P^{k'}\ln P^{k'}}{P_k-P^{k'}}\right\}$$

3. New computations – Fast algorithms

3.1. New computation for numeric attributes

To compute the similarity for a symbolic attributes, we have to consider to the probability of this attribute. Then, the computational complexity in this case is $\Theta(n)$, with n is the number of attribute's values.

On the other hand, in case of numeric attributes, if we compute directly by MSM of C. Li and G.. Biswas [2], the computational complexity in the worst case is $\Theta(n^2)$ complexity. In fact that, we have proposed another indirect solution which has the same results but is more simple with $\Theta(n)$ complexity as mentioned in [3, 4, 5].

The main idea is that we do not compute the measure by this formula:

$$P_{jk} = \sum_{i \in Q} \left\{ p_i^2 + 2 \sum_{t \in T_i} p_i p_t \right\}$$

where T is a set of indices t ($k < t \le n$) such that either

$$|V_t - V_i| < |V_k - V_j| \quad \text{or} \quad \left(|V_t - V_i| = |V_k - V_j|\right) \wedge \left(\sum_{u=j}^{k} p_u \ge \sum_{u=i}^{t} p_u \right)$$

In the contrast way, we compute indirectly through $S_{jk} = 1 - P_{jk}$ formula. We are only interested in the case of $j \ne k$, because when we have j = k then

$$P_{jk} = \sum_{i \in Q} p_i^2, \text{ với } Q = \{ i : (p_i < p_k) \}$$

Then compute Pjk with a linear complexity.

Easily to see that,

$$S_{jk} = 1 - P_{jk} = 2 \sum_{i=1}^{n} \left(p_i \sum_{t \in \overline{T}_i} p_t \right), \qquad j \ne k$$

where $\overline{T}_i$ satisfies

$$\overline{T}_i = \left\{ t : \left[\left(|V_t - V_i| > |V_k - V_j| \right) \vee \left[\left(|V_t - V_i| = |V_k - V_j| \right) \wedge \left(\sum_{u=j}^{k} p_u < \sum_{u=i}^{t} p_u \right) \right] \right] \wedge [i < t \le n] \right\}$$

We can also rewrite $\overline{T}_i$ in the simpler form

$$\overline{T}_i = \{ t : t_i \le t \le n \} \cup T_i^0$$

where t_i is the minimum value that satisfies $|V_t - V_i| > |V_k - V_j|$ and T_i^0 is the set of indices t satisfying

$$\left(|V_t - V_i| = |V_k - V_j| \right) \wedge \left(\sum_{u=j}^{k} p_u < \sum_{u=i}^{t} p_u \right).$$

T_i^0 is empty or has only one element.

We can transform the formula of S_{jk} as follow

$$S_{jk} = 2 \sum_{i=1}^{n} p_i (1 - s_i), \qquad j \ne k$$

where $s_i = \sum_{t=1}^{t_i - \left| T_i^0 \right| - 1} p_t$

The following figure (Figure 1) is the algorithm with the linear complexity.

Some notations in algorithm

Sij: the similarity of the pair of value (Vi, Vj). sumij: sum of frequency from i to j.
dij:absolute value of Vi and Vj. beforek: sum of frequency from 1 to k – 1.
sumt: sum of frequency from 1 to t. f[k]: frequency of k^{th} value.
n: number of attribute value.

Algorithm (in case of $i \ne j$)

```
t = 1; beforek = 0.0; sumt = 0.0; sij = 0.0; sumij = 0.0;

dij = abs(v[j] - v[i]);
for k = i to j do
sumij = sumij + f[k];
for k = 1 to n do {
if (v[k] + dij <= v[n]) {
    if (k > 1)
            beforek = beforek + f[k-1];
    while (t <= n and abs(v[t] - v[k] <= dij)) {
            sumt = sumt + f[t];
            t = t + 1;
    }
    sij = sij + 2 * f[k] * (1 - sumt);
    if (abs(v[t - 1] - v[k]) = dij and sumt - beforek <= dij)
            sij = sij + 2 * fk * f[t - 1];
}
}
```

Figure 1: Fast algorithm for a numeric attribute.

3.2. New computation for ordinal attributes

We applied similarly the way of numeric way for ordinal attributes with linear complexity.

Firstly, we rewrite the expression computing Pij as follows:

$$\begin{aligned} P_{ij} &= \sum_{k=1}^{n} p_k^2 + 2\sum_{k=1}^{n-1}\sum_{l=k+1}^{\tau} p_k p_l \\ &= p_n^2 + \sum_{k=1}^{n-1} p_k \left(p_k + 2\sum_{l=k+1}^{\tau} p_l \right) \\ &= p_n^2 + \sum_{k=1}^{n-1} p_k \left(2\sum_{l=k}^{\tau} p_l - p_k \right) \end{aligned}$$

where τ satisfies $(k+1 \le \tau) \wedge \left(\sum_{t=k}^{\tau} p_t \le \sum_{t=i}^{j} p_t < \sum_{t=k}^{\tau+1} p_t \right)$.

Then

$$P_{ij} = \sum_{k=1}^{n} p_k \left(2\sum_{l=k}^{\tau} p_l - p_k \right)$$

where τ multiplies k.

```
t = 1; beforek = 0.0; sumt = 0.0;
sumij = 0.0; pij = 0.0;

for k = i to j do
sumij = sumij + f[k];

for k = 1 to n do {
if (k > 1)
   beforek = beforek + f[k - 1];

BOOL stop = FALSE;
while (t <= n and not stop) {
   if (sumt - beforek <= sumij and
         sumt + f[t] - beforek > sumij)
         stop = TRUE;
   else {
         sumt = sumt + f[t];
         t = t + 1;
   }
}

pij = pij + f[k] * (2 * (sumt - beforek)
      - f[k]);
}
```

Figure 2: The fast algorithm for an ordinal attribute.

Using the notations shown in Figure 1, the fast algorithm for the ordinal attributes is described in Figure 2. For combining the mixed measure, the ordinal attributes are treated by the Lancaster's transformation.

4. Experiment and results

This section will present the experimental results by MSM measures (for numeric and symbolic attributes), MSM* (for numeric, ordinal and symbolic attributes) applied in classification problem using Nearest Neighbor Rules, k-NNR. The results are compared with decision tree technique through C4.5, See5.0 software of J.R. Quinlan in [6, 7]. This software is considered as a good software to classify. The datasets which are tried in experiments are standard data from UCI-KDD repository described detail in [8]. Each dataset is divided (equally and randomly) to 10 subsets of data and the experimental results are averages of results archiving from this subsets.

A dataset consists of 3 files with *.names*, *.data* and *.test* extensions. The file with *.names* extension is the structural file involved information of class name, attribute name with its type (symbolic, and/or numeric, and/or ordinal), and attribute value (discrete and/or ordinal). See 5.0 is free for training and education purpose with number of limit training records in *.data* file or test records in *.test* file - no more 200 records. Therefore, with each of dataset we only use 200 records after eliminating records with missing values to ensure the equality of the experimental condition.

There are 28 following datasets used in experiment with NAME (number of record; number of records; number of numeric attributes; number of symbolic attributes; number of ordinal attributes): **ATT(681; 2; 1; 4; 4)**, BAN(251; 2; 19; 11; 0), BCW(614; 2; 9; 0; 0), BIO(174; 2; 5; 0; 0), BLD(260; 2; 6; 0; 0), BOS(369; 3; 10; 6; 0), BPR(260; 2; 6; 0; 0), **CMC(1127; 3; 2; 4; 3)**, CRX(587; 2; 6; 9; 0),

DER(322; 6; 1; 32; 1), ECH(97; 2; 5; 1; 0), HAB(219; 2; 3; 0; 0), HCO(35; 2; 5; 14; 0); **HEA(267; 2; 5; 7; 1)**, HEP(72; 2; 6; 13; 0), **HIN(535; 3; 0; 1; 5)**, HUR(168; 2; 6; 0; 0), HYP(1800; 2; 6; 9; 0), IMP(177; 5; 13; 9; 0), **INF(191; 6; 0; 16; 2)**, LBW(182; 2; 2; 6; 0), PID(353; 2; 8; 0; 0), SEG(200; 7; 11; 0; 0), SMO(2088; 3; 8; 0; 8), TAE(125; 3; 1; 4; 0), USN(141; 3; 26; 1; 0), VEH(621; 4; 18; 0; 0), VOT(200; 2; 0; 16; 0). Among these sets, there are 6 bolded sets because of having ordinal attributes. The detail is described in [8].

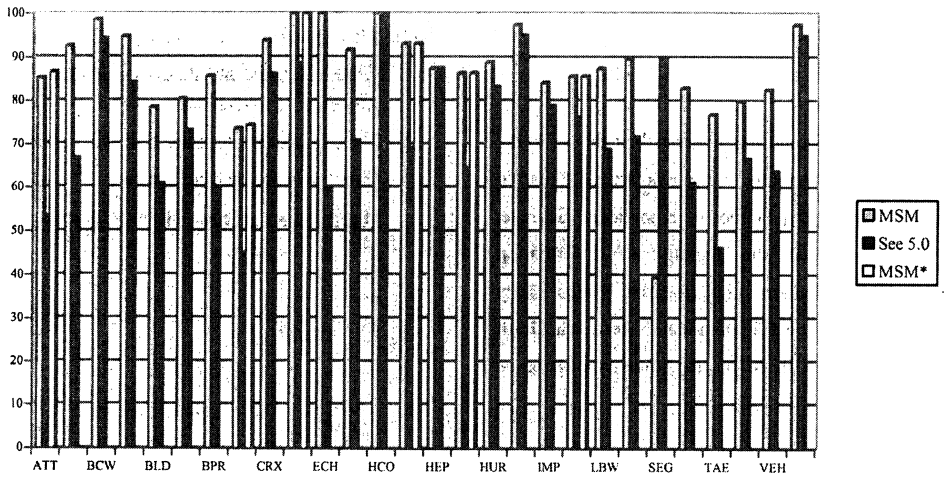

Figure 3: The accuracy (%) when classifying records by MSM, MSM* and See 5.0.

5. Conclusion and Future work

This paper presented a mixed similarity measure. The early measures often work at only one attribute. This problem leads the requirement of developing a new measure which is suitable with both symbolic and numeric attributes, especially ordinal attributes. The experimental results express that the MSM based on the early model of D.W.Goodall, work well with mixed database (included symbolic, numeric, ordinal attribute).

The main content of this paper is presenting the fast algorithm for computing the similarity of numeric and ordinal attributes. The fast algorithm for numeric attributes is discussed in [4]. The key idea is the indirect computation of the similarity measure, contrary to the direct formula. We also use the same way of [4] for ordinal and symbolic attributes in this paper.

The implementation of the MSM is carried out in MS Visual C++. The experiments on the datasets by MSM provide better result than by See 5.0. These results are also more potential in case that we consider the order of the symbolic attributes. As a result, MSM và MSM* are suitable for real-word data with hetegenenous attributes.

However, experimental data used in this paper are still small; it means that the run time is nearly immediate. Some other experiments with large database were also tried in [4], but did not include ordinal attributes. Studying

the MSM and implementing it with large database with all type of attributes: symbolic, numeric, ordinal is some of our future works. We also expect apply MSM* to distance-based algorithms. Another challenge is how we update incremental value of MSM, MSM* when there are changes of value in a huge database to save computing time and memory.

Acknowledgements

The authors would like express their thanks to the Researchers and the Students at the Knowledge Creating Methodology Laboratory, School of Knowledge Science, JAIST, Japan, and at the Faculty of Information Technology, HUT, Vietnam. This research is also partly supported by the National Fundamental Research Grant under Contract No.KHCB220102, the Ministry of Science and Technology, Vietnam.

Reference

[1]. D.W. Goodall: *A New Similarity Index Based on Probability.* Biometrics, 1966.

[2]. C. Li, G. Biswas: *Conceptual Clustering with Numeric and Nominal Mixed Data – A New Similarity Based System.* KDD: Techniques and Applications, World Scientific, 1997. (Also in IEEE Transcript on Knowledge and Data Engineering, 1998)

[3]. N.N. Binh, H.T. Bao, T. Morita: *Study of a Mixed Similarity Measure for Classification and Clustering.* 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD'99, Beijing, April 1999. Lecture Notes in Artificial Intelligence 1574, pp.375-379, Springer-Verlag, 1999.

[4]. N.N. Binh, H.T. Bao: *A Mixed Similarity Measure in Linear Time and Space Computation for Distance-Based Methods.* 4th European Conference on Principles of Data Mining and Knowledge Discovery PKDD 2000, Lyon, September 13-16, 2000. Lecture Notes in Artificial Intelligence 1910, pp. 211 - 220, Springer, 2000.

[5]. N.N Bình, H.T. Bảo, N.Đ. Dũng: *Computing a Mixed Similarity Measure for Distance-based Methods.* Proc. Of Scientific papers, the 19th Scientific Conference of HUT, Section of Maths and Informatics, pp.12-16, Hanoi, 2001 (in Vietnamese).

[6]. J.R. Quinlan: *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.

[7]. J.R. Quinlan: *Data Mining Tools See5 and C5.0.* (http://www.rulequest.com/see5-info.html, RULEQUEST RESEARCH 1997-2004)

[8]. The UCI Machine Learning Site and Repository (http://www.ics.uci.edu/~mlearn/, http://www.ics.uci.edu/~mlearn/MLSummary.html