Discovering the Difference in the Usage of Symbol toward improving Data Quality in Databases

Tetsuya Yoshida, Kensuke Ohnishi, Shogo Nishida
Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
yoshida@sys.es.osaka-u.ac.jp

Abstract: *This paper proposes to discover the difference in the usage of symbol for data description toward improving the quality of data in databases. Large scale databases are usually constructed through the participation of multiple people with different backgrounds for the collection and description of data. Since the meaning represented by symbol or the usage of symbol in data description can be different depending on individuals due to their respective viewpoints, there can be inconsistency in the description of data which are stored in databases. To discover the difference in the usage of symbol for data description, decision trees for the classification of respective cases are first constructed. After that, mutual classification of diagnosis cases is utilized to find out the correlation between the nodes in decision trees so as not to destroy the original characteristics or structure of each decision tree.*

1 Introduction

In accordance with the development of technologies in databases and machine learning, various researches have been carried out in the field of Knowledge Discovery from Databases (KDD) [2, 7]. Although the importance of dealing with the whole process of KDD has been noticed, most researches have mainly focused on improving the quality of data mining algorithms. Several researches have investigated in the visualization of the result of data mining in the context of human computer interaction so that it can be more comprehensive to the user of KDD system as the post-processing of data mining. However, the pre-processing of data mining, that is, improving the quality of data in databases, has not been focused intensively so far. For instance, it is often noticed that data mining algorithms cannot discover effective knowledge or rules not because the quality of the algorithms themselves is low but because the quality of data stored in databases is low with noise and missing values.

When constructing large scale databases, usually multiple people with different backgrounds participate in the collection and specification of data. Since the meaning denoted by symbol or the usage of symbol in the description of data can be different depending on individuals due to their different points of view, there can be inconsistency in the description of data which are stored in databases. We have been investigating a framework and method for discovering the difference in the usage o f symbol among people [3, 8] in the context of CSCW (Computer Supported Cooperative Work) [6, 1] to faciliate mutual understanding in collaboration. In our previous approach, the system worked as an interactive system and suggested the cadidates or possibility for the difference. The system was effective to discover the difference in the context of human-computer interaction for small-scale databases since the users could take the initiative and responsibility for the identification of difference. However, just pointing out the possibility or candidates has the problem of hindering the scale up of the system to large scale databases, since the user can be confused with the false alarm problem.

This paper proposes to discover the difference in data description based on the correlation between decision trees toward improving the quality of data in databases. The structure of knowledge held by individuals is constructed based on the diagnosis cases which are specified by him/her and represented as the decision tree. To utilize the correlation between decision trees, the diagnosis cases are then classified by the decision trees which are constructed for the cases specified by the other and vice versa. After that, the discovery algorithms are applied to find out the difference in the usage of symbol based on the correlation between decision trees. Thus, in this approach the difference and correspondance between the usage of symbols in the description of data are discovered and clarified based on the result of the mutual classification of cases. Clarifying the difference and correspondance between the usage of symbol will also contribute to facilitating the mutual understanding between the people who participated in the description of data, since the difference in the usage of symbols or words often hinder communication and collaboration.

This paper is organized as follows. Sect. 2 explains our framework of dealing with the difference in the usage of symbol for data description. Sect. 3

briefly describes the method for discovering difference in the usage of symbol by preserving the individual characteristics of decision trees. Brief concluding remarks are described in Sect. 4.

2 Framework of Discovering Difference in Data Description

2.1 Difference in Data Description

This paper focuses on dealing with the difference in data description at the symbol level. Usually different symbols are used to represent different meanings, however, the same symbol can be used to represent different meaning depending on the viewpoint and the context in which the symbol is utilized to represent data, and different symbols can be used to represent the same meaning, vice versa. These kinds of difference can occur especially when people with different backgrounds and knowledge participate in the data description to register cases into lage-scale databases. The types of difference in data description dealt with in this paper are defined as follows:

- Type 1: different symbols are used to represent the same meaning.

- Type 2: the same symbol is used to represent different meaning.

Suppose there are an expert in electric engineering called Adam and an expert in mechanical engineering called Bob. When they register the data for the diagnosis of motor failure in databases, Adam might utilize the symbol "anomaly in voltage frequency" and Bob might utilize the symbol "anomaly in revolutions" in the description of cases for the same symptom. In general the above two symbols represent different meaning , however, they can be considered as representing the same meaning when they are used to represent cases in the context of the diagnosis of motor failure.

As another example, it is hardly the case that the ranges of temperature in which Adam feels "hot" and Bob feels "hot" are the same and usually there are some differences between the ranges. Thus, there can be some range in which Adam feels "hot" and Bob feels "warm". This means that there are differences in the meaning which is represented with the symbol "hot" between Adam and Bob and that there is some overlap between the meaning which is conveyed with the symbol "hot" by Adam and the symbol "warm" by Bob. Thus, this kind of difference in the usage of symbol can be embedded into the description of data in databases.

2.2 System Architecture

The kind of problem which can be dealt with in our approach is classification such as diagnosis and the *class* of cases is determined based on the *attributes* and *values*, which characterizes the cases [4]. The system tries to construct the decision tree for cases which is most effective for their classification based on the information theory. A node in decision trees holds the attribute to characterize cases. Each link below a node holds the value for the attribute at the node and cases are divided gradually by following links. The class of cases is determined at the leaf which is reached as the result of link following. ID3 algorithm [5] is utilized to construct decision trees since it is fast and widely utilized in the field of machine learning.

The system architecture which incorporates the descovery method to improve the quality of data in databases is shown in Fig 1. Currently the system requires that two users represent their knowledge as cases with their respective symbols and register them into respective databases. By accepting the cases as input, the system constructs decision trees for them and tries to discover the difference in the usage of symbols for attributes, values and classes, based on the correspondance between decision trees. The system also displays the decision trees which are constructed for the cases. so that the users can grasp the difference intuitively and modify their usage of symbols in knowledge representation toward improving the quality of data in databases with less noise and inconsistency.

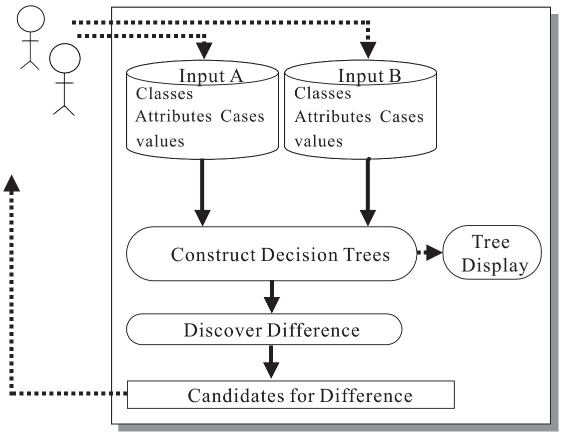

Figure 1: System Architecture

Difference in the usage of symbols for value means that the ranges covered by values differ and thus it is necessary to consider *to what extent* the ranges overlap. This is different from the difference in the usage of symbols for class and attribute in the sense that it is not sufficient to deal with it by considering whether the difference simply exists or not

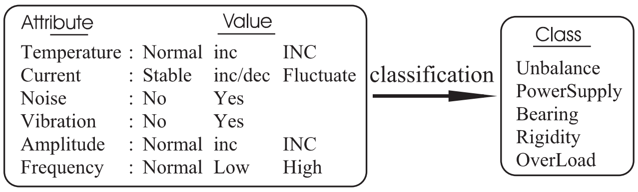

Figure 2: Input Cases for the System.

exist. Fig. 3 illustrates the relationship between the ranges of value "Normal/inc/INC" for the attribute "Temperature" for this kind of difference.

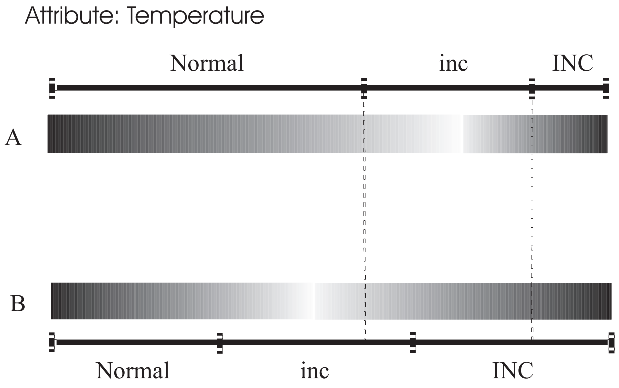

Figure 3: An example of difference for value.

Since there are 2 types of difference for 3 entities (but 2 types of difference for value symbols occur simultaneously), the system tries to discover 5 kinds of difference in the usage of symbols and shows the candidates for them in the descending order of the possibility. Hereafter, the kind of difference in the usage of symbols for class, attribute, and value is called **C1**, **C2**, **A1**, **A2**, **V**, respectively[1].

3 Methodology for Discovering the Difference in Data Description

It seems important not to destroy the original structure of decision tree to discover the difference in data description while sustaining the characteristics or viewpoint of individuals who specified the data with their own symbols. In general there can be several decision trees with the equivalent classification for a set of cases, however, ID3 [5] algorithm constructs the most simple tree in the sense that classification is carried out as fast as possible at the shallower or upper nodes in a decision tree by reducing the redundant information held in the cases. Thus, the attributes at upper nodes in decision tree can be considered as playing the more important role in the classification of cases for each user. However, constructing the decision tree by merging the set of cases from different users in [3, 8] can bring about the re-structuring of decision trees, which can loose the information about the importance of attributes in the original decision trees.

In this paper the relation between the leaves in two decision trees is called a "pair". If completely different symbols are utilized in the description of data depending on users, it is not necessary to think about the difference in the usage of symbols. Thus, this paper focuses on the situation in which there is some overlap between the set of diagnosis cases specified by the users. Here, overlap means that the cases represent the same knowledge for both users and there exist no difference in the usage of symbol at least for such overlapped cases. In our approach each case is thrown into the two decision trees and classified at some leaf in decision trees respectively, and the correspondance between the leaves is treated as a pair to discover the difference in the usage of symbol. A pair represents the agreement and disagreement about the classification for the same case, which is later utilized to discover the correspondance between the usage of symbol in the description of cases.

The algorithm for discovering the difference in data description based on pairs is as follows:

Step.1 determine the class of cases based on the decision tree by the other.

Step.2 construct a set of pairs.

Step.3 compare the set of pairs with the typical set of pairs when the difference the difference in the usage of symbol occurs.

All the algorithms for the kinds of difference described in Sect. 2.2 are designed based on the above steps. Due to the space shortage, only the algorithm for **C1**, namely, discovering the situation in which different symbols are used to represent the same class, is described in Sect. 3.1 in detail.

3.1 C1: Different Symbols are used to represent the Same Class

This kind of difference is represented at the leaf of decision tree (see Fig.4). When **C1** with respect to the class symbol Cc and Cd occurs, the difference can be discovered as the difference of symbols for leaves in decision trees. In Fig.4, classifying the set of diagnosis cases by Adam and Bob each other results in that the cases with class symbol Cc by Adam are classified as Cd in the decision tree for Bob and also the cases with class symbol Cd by Bob are classified as Cc in the other decision tree for Adam. The algorithm for discovering **C1** based on pairs is as follows:

[1] Since two kinds of difference for value symbol occur simultaneously.

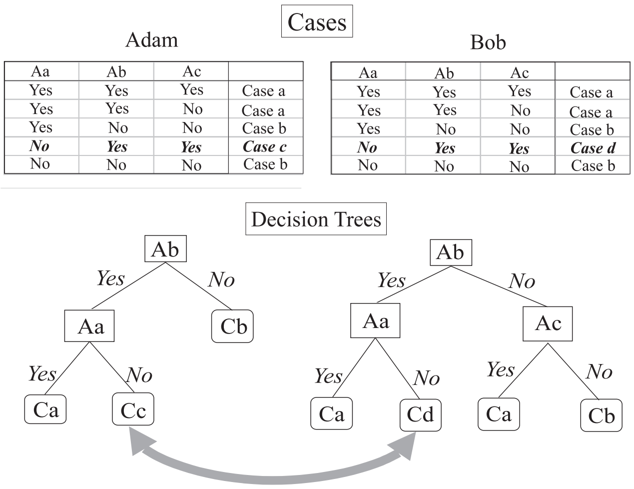

Figure 4: A pair with difference C1.

Step.1 Search the pairs with different class symbols. Suppose the class symbol Y is utilized for the counterpart of the class symbol X.

Step.2 Check the class symbol X', which corresponds to the counterpart of the pairs with class symbol Y.

Step.3 If X = X' for all the pairs found in **Step.1**, treat the pair as the difference **C1**. However, since it is rarely the case that X = X' for all the pairs, calculate the degree as:

$$\frac{\text{the number of pairs which satisfy the algorithm}}{\text{the number of pairs for the class}}$$

and treat it as a pair if the degree exceeds the specified threshold. The larger the degree, the more likely that the kind of difference occurs for the usage of symbols.

Step4. Repeat **Step.1** and **Step.3** and show the discovered pairs in the descending order of degree in **Step.3**.

4 Concluding Remarks

This paper has proposed to discover the difference in the usage of symbol for data description based on the correlation between decision trees toward improving the quality of data in databases. Besides the quality of data mining algorithms themselves, the low quality of data due to inconsistency in the description of data can also hinder the discovery of effective knowledge or rules from databases. By extending our previous approach [3, 8] which has been investigated in the context of CSCW (Computer Supported Cooperative Work) the discovery algorithms based on the correlation between decision trees for diagnosis cases are proposed to discover the difference in the usage of symbol for the description of data in databases. Experiments on the UC Irvine data depository are to be carried out to investigate the effectiveness of our approach as a first step toward improving the quality of data in databases by removing the difference and inconsistency in the usage of symbol for data description. We plan to extend the proposed approach in order to deal with the correlation between non-leaf nodes in decision trees in bottom-up fashion and to utilize it toward improving the quality of data in databases.

Acknowledgments

This research is partially supported by Grant in Aid for Scientific Research from the Japanese Ministry of Education, Science and Culture (10875080, 09450159), and Nissan Science Foundation.

References

[1] R. Baecker, editor. *Readings in Groupware and Computer-Supported Cooperative Work – Assisting Human-Human Collaboration*. Morgan Kaufmann, 1993.

[2] U.M. Fayyad, G. Piattetsky-Shapiro, P. Smyth, and R. Uthurusamy. From Data Mining to Knowledge Discovery: An Overview. In U.M. Fayyad, G. Piattetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. MIT Press, Cambridge, Massachusetts, 1996.

[3] T. Kondo, T. Yoshida, and S. Nishida. Design of the Interfaces to Detect Conceptual Difference among Different People. *Information Processing Society of Japan*, 39(5):1195–1202, May 1998. in Japanese.

[4] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Chichester, England, 1994.

[5] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[6] M. Stefik, G. Foster, D.G. Bobrow, K. Kahn, S. Lanning, and L. Suchman. Beyond the Chalkboard: Computer-Support for Collaboration and Problem Solving in Meeting. *Communications of the ACM*, 30(1):32–47, 1987.

[7] S.M. Weiss and N. Indurkhya. *Predictive Data Mining*. Morgan Kaufmann, 1998.

[8] T. Yoshida, T. Kondo, and S. Nishida. Discovering Conceptual Differences among People from Cases. In *The First International Conference on Discovery Science*, pages 162 – 173, 1998.