Fuzzy Decision Tree Induction Approach for Mining Fuzzy Association Rules

Rolly Intan and Oviliani Yenty Yuliana

Informatics Engineering Department
Petra Christian University, Surabaya, Indonesia
rintan@peter.petra.ac.id, ovi@peter.petra.ac.id

Abstract. Decision Tree Induction (DTI), one of the Data Mining classification methods, is used in this research for predictive problem solving in analyzing patient medical track records. In this paper, we extend the concept of DTI dealing with meaningful fuzzy labels in order to express human knowledge for mining fuzzy association rules. Meaningful fuzzy labels (using fuzzy sets) can be defined for each domain data. For example, fuzzy labels *poor disease*, *moderate disease*, and *severe disease* are defined to describe a condition/type of disease. We extend and propose a concept of fuzzy information gain to employ the highest information gain for splitting a node. In the process of generating fuzzy association rules, we propose some fuzzy measures to calculate their support, confidence and correlation. The designed application gives a significant contribution to assist decision maker for analyzing and anticipating disease epidemic in a certain area.

Keywords: Data Mining, Classification, Decision Tree Induction, Fuzzy Set, Fuzzy Association Rules.

1 Introduction

Decision Tree Induction (DTI) has been used in machine learning and in data mining as a model for prediction a target value based on a given relational database. There are some commercial decision tree applications, such as the application for analyzing a return payment of a loan for owning or renting a house [15] and the application of software quality classification based on the program modules risk [16]. Both applications inspire this research to develop an application for analyzing patient medical track record. The Application is able to present relation among (single/group) values of patient attribute in decision tree diagram. In the developed application, some domains of data need to be utilized by meaningful fuzzy labels. For example, fuzzy labels *poor disease*, *moderate disease*, and *severe disease* describe a condition/type of disease; *young*, *middle aged* and *old* are used as the fuzzy labels of ages. Here, a fuzzy set is defined to express a meaningful fuzzy label. In order to utilize the meaningful fuzzy labels, we need to extend the concept of (*crisp*) DTI using fuzzy approach. Simply, the extended concept is called *Fuzzy Decision Tree* (FDT). To generate FDT from a normalized database that consists of several tables, there are several

C.S. Leung, M. Lee, and J.H. Chan (Eds.): ICONIP 2009, Part II, LNCS 5864, pp. 720–728, 2009.

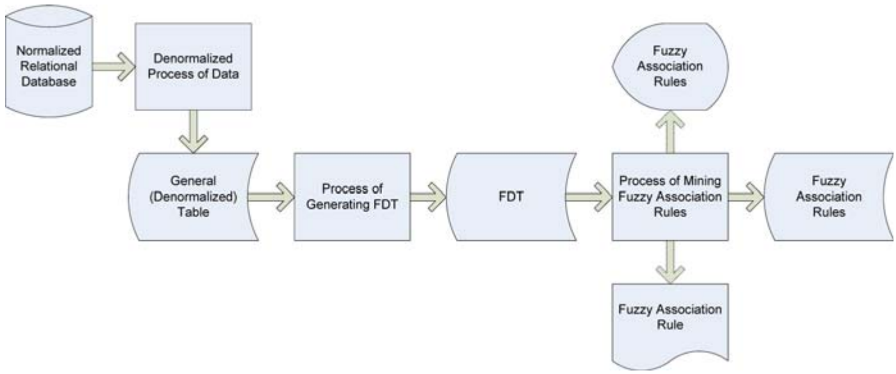

Fig. 1. Process of Mining Association Rules

sequential processes as shown in Fig. 1. First is the process of joining tables known as *Denormalization of Database* as discussed in [4]. The process of denormalization can be provided based on the relation of tables as presented in Entity Relationship Diagram (ERD) of a relational database. Result of this process is a general (denormalized) table. Second is the process of constructing FDT generated from the denormalized table.

In the process of constructing FDT, we propose a method how to calculate fuzzy information gain by extending the existed concept of (*crisp*) information gain to employ the highest information gain for splitting a node. The last is the process of mining fuzzy association rules. In this process, fuzzy association rules are mined from FDT. In the process of mining fuzzy association rules, we propose some fuzzy measures to calculate their support, confidence and correlation. Minimum support, confidence and correlation can be given to reduce the number of mining fuzzy association rules. The designed application gives a significant contribution to assist decision maker for analyzing and anticipating disease epidemic in a certain area.

The structure of the paper is the following. Section 2 as main contribution of this paper is devoted to propose the concept and algorithm for generating FDT. Section 3 proposes some equations of fuzzy measures that play important role in the process of mining fuzzy association rules. Section 4 demonstrates the algorithm and in a simple illustrative results. Finally a conclusion is given in Section 5.

2 Fuzzy Decision Tree Induction (FDT)

Based on type of data, we may classify DTI into two types, namely crisp and fuzzy DTI. Both DTI are compared based on Generalization-Capability [14]. The result shows that Fuzzy Decision Tree (FDT) is better than Crisp Decision Tree (CDT) in providing numeric attribute classification. Fuzzy Decision Tree formed by the FID3, combined with Fuzzy Clustering (to form a function member) and validated cluster (to decide granularity) is also better than Pruned Decision Tree. Here, Pruned

Decision Tree is considered as a Crisp enhancement [13]. Therefore in our research work, disease track record analyzer application development, we propose a kind of FDT using fuzzy approach.

An information gain measure [1] is used in this research to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or impurity in these partitions. In order to process crisp data, the concept of information gain measure is defined in [1] by the following definitions.

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for i=1,…, m). Let s_i be the number of samples of S in class C_i. The expected information needed to classify a given sample is given by

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

where p_i is the probability that an arbitrary sample belongs to class C_i and is estimated by s_i/s.

Let attribute A have v distinct values, $\{a_1, a_2, ..., a_v\}$. Attribute A can be used to partition S into v subsets, $\{S_1, S_2, ..., S_v\}$, where S_j contains those samples in S that have value a_j of A. If A was selected as the test attribute then these subsets would correspond to the braches grown from the node containing the set S. Let s_{ij} be the number of samples of class C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + ... + s_{mj}}{s} I(s_{1j}, ..., s_{mj}) \tag{2}$$

The term $\frac{s_{ij} + ... + s_{mj}}{s}$ acts as the weight of the jth subset and is the number of samples in the subset divided by the total number of samples in S. The smaller the entropy value, the greater the purity of the subset partitions.The encoding information that would be gained by branching on A is

$$Gain(A)=I(s_1, s_2,..., s_m) - E(A) \tag{3}$$

In other words, *Gain(A)* is the expected reduction in entropy caused by knowing the values of attribute A.

When using the fuzzy value, the concept of information gain as defined in (1) to (3) will be extended to the following concept. Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values, v_i (for i=1,…, m), defining m distinct classes, C_i (for i=1,…, m). And also suppose there are n meaningful fuzzy labels, F_j (for j=1,…, n) defined on m distinct values, v_i. $F_j(v_i)$ denotes membership degree of v_i in the fuzzy set F_j . Here, F_j (for j=1,…, n) is defined by satisfying the following property:

$$\sum_{j}^{n} F_j(v_i) = 1, \forall \mathrm{i} \in \{1,...m\}$$

Let β_j be a weighted sample corresponding to F_j as given by

$\beta_j = \sum_{i}^{m} \det(C_i) \times F_j(v_i)$, where $\det(C_i)$ is the number of elements in C_i.

The expected information needed to classify a given weighted sample is given by

$$I(\beta_1, \beta_2, ..., \beta_n) = -\sum_{j=1}^{n} p_j \log_2(p_j) \tag{4}$$

where p_j is estimated by β_j/s.

Let attribute A have u distinct values, $\{a_1, a_2, …, a_u\}$, defining u distinct classes, B_h (for h=1,…, u). Suppose there are r meaningful fuzzy labels, T_k (for k=1,…, r), defined on A. Similarly, T_k is also satisfy the following property.

$$\sum_{k}^{r} T_k(a_h) = 1, \forall h \in \{1,...,u\}$$

If A was selected as the test attribute then these fuzzy subsets would correspond to the braches grown from the node containing the set S. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{k=1}^{r} \frac{\alpha_{1k} + ... + \alpha_{nk}}{s} I(\alpha_{1k}, ..., \alpha_{nk}) \tag{5}$$

Where α_{jk} be intersection between F_j and T_k defined on data sample S as follows.

$$\alpha_{jk} = \sum_{h}^{u} \sum_{i}^{m} \min(F_j(v_i), T_k(a_h)) \times \det(C_i \cap B_h) \tag{6}$$

Similar to (4), $I(\alpha_{ik}, …, \alpha_{nk})$ is defined as follows.

$$I(\alpha_{1k}, ..., \alpha_{nk}) = -\sum_{j=1}^{n} p_{jk} \log_2(p_{jk}) \tag{7}$$

where p_{jk} is estimated by α_{jk}/s.

Finally, the encoding information that would be gained by branching on A is

$$Gain(A) = I(\beta_1, \beta_2, …, \beta_n) - E(A) \tag{8}$$

Since fuzzy sets are considered as a generalization of crisp set, it can be proved that the equations (4) to (8) are also generalization of equations (1) to (3).

3 Mining Fuzzy Association Rules from FDT

Association rules are kind of patterns representing correlation of attribute-value (items) in a given set of data provided by a process of data mining system. Generally, association rule is a conditional statement (such kind of *if-then rule*). Performance or interestingness of an association rule is generally determined by three factors, namely *confidence*, *support* and *correlation* factors. Confidence is a measure of certainty to assess the validity of the rule. The support of an association rule refers to the percentage of relevant data tuples (or transactions) for which the pattern of the rule is true. Correlation factor is another kind of measures to evaluate correlation between two entities.

Related to the proposed concept of FDT as discussed in Section 2, the fuzzy association rule, $T_k \Rightarrow F_j$ can be generated from the FDT. The confidence, support and correlation of $T_k \Rightarrow F_j$ are given by

$$\text{confidence}(T_k \Rightarrow F_j) = \frac{\sum_h^u \sum_i^m \min(F_j(v_i), T_k(a_h)) \times \det(C_i \cap B_h)}{\sum_h^u T_k(a_h) \times \det(B_h)} \tag{9}$$

$$\text{support}(T_k \Rightarrow F_j) = \frac{\sum_h^u \sum_i^m \min(F_j(v_i), T_k(a_h)) \times \det(C_i \cap B_h)}{s} \tag{10}$$

$$\text{correlation}(T_k \Rightarrow F_j) = \frac{\sum_h^u \sum_i^m \min(F_j(v_i), T_k(a_h)) \times \det(C_i \cap B_h)}{\sum_h^u \sum_i^m F_j(v_i) \times T_k(a_h) \times \det(C_i \cap B_h)} \tag{11}$$

To provide a more generalized multidimensional fuzzy association rules as proposed in [6], it is started from a single table (relation) as a source of data representing relation among item data. Formally, a relational data table [12] R consists of a set of tuples, where t_i represents the i-th tuple and if there are n domain attributes D, then $t_i = (d_{i1}, d_{i2}, \cdots, d_{in})$. Here, d_{ij} is an atomic value of tuple t_i with the restriction to the domain D_j, where $d_{ij} \in D_j$. A relational data table R is defined as a subset of the set of cross product $D_1 \times D_2 \times \cdots \times D_n$, where $D = \{D_1, D_2, \cdots, D_n\}$. Tuple t (with respect to R) is an element of R. In general, R can be shown in Table 1.

Now, we consider χ and ψ as subsets of fuzzy labels. Simply, χ and ψ are called fuzzy datasets. A fuzzy dataset is a set of fuzzy data consisting of several distinct fuzzy labels, where each fuzzy label is represented by a fuzzy set on a certain domain attribute. Formally, χ and ψ are given by $\chi = \{F_j \mid F_j \in \Omega(D_j),\ \exists\ j \in \mathrm{N}_n\}$ and

Table 1. A Schema of Relational Data Table

Tuples	D_1	D_2	$\cdots$	D_n
t_1	d_{11}	d_{12}	$\cdots$	d_{1n}
t_2	d_{21}	d_{22}	$\cdots$	d_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_s	d_{s1}	d_{s2}	$\cdots$	d_{sn}

$\psi = \{F_j \mid F_j \in \Omega(Dj), \exists\ j \in \mathrm{N}_n\}$, where there are n domain data, and $\Omega(D_j)$ is a fuzzy power set of D_j. In other words, F_j is a fuzzy set on D_j. The confidence, support and correlation of $\chi \Rightarrow \psi$ are given by

$$\text{support}(\chi \Rightarrow \psi) = \frac{\sum_{i=1}^{s} \inf_{F_j \in \chi \cup \psi} \{F_j(d_{ij})\}}{s} \tag{12}$$

$$\text{confidence}(\chi \Rightarrow \psi) = \frac{\sum_{i=1}^{s} \inf_{F_j \in \chi \cup \psi} \{F_j(d_{ij})\}}{\sum_{i=1}^{s} \inf_{F_j \in \chi} \{F_j(d_{ij})\}} \tag{13}$$

$$\text{correlation}(\chi \Rightarrow \psi) = \frac{\sum_{i=1}^{s} \inf_{F_j \in \chi \cup \psi} \{F_j(d_{ij})\}}{\sum_{i=1}^{s} \inf_{A_j \in \chi} \{A_j(d_{ij})\} \times \inf_{B_k \in \psi} \{B_k(d_{ik})\}} \tag{14}$$

4 FDT Algorithms and Results

The research is conducted based on the Software Development Life cycle method. The application design conceptual framework is shown in Fig 1. An input for developed application is a single table that is produced by denormalization process from a relational database. The main algorithm for mining association rule process, i.e. Decision Tree Induction, is shown in Fig 2. Furthermore, the procedure for calculating information gain, to implementing equation (4), (5), (6), (7) and (8), is shown in Fig 3. Based on the highest information gain the application can develop decision tree in which the user can display or print it. The rules can be generated from the generated decision tree. Equation (9), (10) and (11) are used to calculate the interestingness or performance of every rule. The number of rules can be reduced based on their degree of support, confidence and correlation compared to the minimum value of support, confidence and correlation determined by user.

```
For i=0 to the total level
    Check whether the level had already split
    If the level has not yet split Then
      Check whether the level can still be split
      If the level can still be split Then
      Call the procedure to calculate information gain
        Select a field with the highest information gain
        Get a distinct value of the selected field
        Check the total distinct value
        If the distinct value is equal to one Then
        Create a node with a label from the value name
        Else
        Check the total fields that are potential to become a current test attribute
        If no field can be a current test attribute Then
           Create a node with label from the majority value name
        Else
           Create a node with label from the selected value name
        End If
      End If
  End If
End If
 End for
 Save the input create tree activity into database
```

Fig. 2. The Generating Decision Tree Algorithm

```
Calculate gain for a field as a root
Count the number of distinct value field
For i=0 to the number of distinct value field
    Count the number of distinct value root field
    For j=0 to the number of distinct value root field
        Calculate the gain field using equation (4) and (8)
    End For
    Calculate entropy field using equation (5)
End For
Calculate information gain field
```

Fig. 3. The Procedure to Calculate Information Gain

In this research, we implement two data types as a fuzzy set, namely alphanumeric and numeric. An example of alphanumeric data type is *disease*. We can define some meaningful fuzzy labels of *disease*, such as *poor disease*, *moderate disease*, and *severe disease*. Every fuzzy label is represented by a given fuzzy set. The *age* of patients is an example of numeric data type. *Age* may have some meaningful fuzzy labels such as *young* and *old*. Fig 4 shows an example result of FDT applied into three domains (attributes) data, namely *Death*, *Age* and *Disease*.

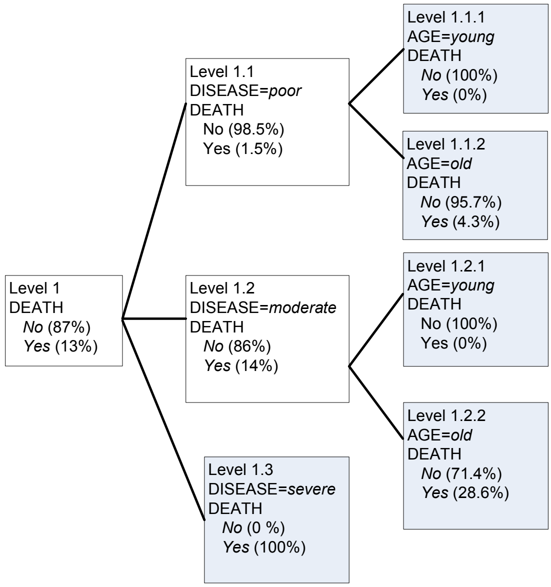

Fig. 4. The Generated Decision Tree

5 Conclusion

The paper discussed and proposed a method to extend the concept of Decision Tree Induction using fuzzy value. Some generalized formulas to calculate information gain ware introduced. In the process of mining fuzzy association rules, some equations ware proposed to calculate support, confidence and correlation of a given association rules. Finally, an algorithm was briefly given to show the process how to generate FDT.

Acknowledgment

This research was supported by research grant PHB (244/SP2H/PP/DP2M/III/2008) and HIKOM (25/SP2H/PP/DP2M/V/2009) from Indonesian Higher Education Directorate.

References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. The Morgan Kaufmann Series (2001)
2. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, New Jersey (1995)
3. Intan, R.: An Algorithm for Generating Single Dimensional Association Rules. Jurnal Informatika (Terakreditasi SK DIKTI No. 56/DIKTI/Kep/2005) 7(1) (May 2006)

4. Intan, R.: A Proposal of Fuzzy Multidimensional Association Rules. Jurnal Informatika (Terakreditasi SK DIKTI No. 56/DIKTI/Kep/2005) 7(2) (November 2006)
5. Intan, R.: A Proposal of an Algorithm for Generating Fuzzy Association Rule Mining in Market Basket Analysis. In: Proceeding of CIRAS (IEEE). Singapore (2005)
6. Intan, R.: Generating Multi Dimensional Association Rules Implying Fuzzy Valuse. In: The International Multi-Conference of Engineers and Computer Scientist, Hong Kong (2006)
7. Gunawan, O.P.: Perancangan dan Pembuatan Aplikasi Data Mining dengan Konsep Fuzzy c-Covering untuk Membantu Analisis Market Basket pada Swalayan X (in Indonesian) Final Project (2004)
8. Zadeh, L.A.: Fuzzy Sets and systems. International Journal of General Systems 17, 129–138 (1990)
9. Agrawal, R., Imielimski, T., Swami, A.N.: Mining Association Rules between Sets of Items in Large Database. In: Proccedings of ACM SIGMOD International Conference Management of Data, pp. 207–216. ACM Press, New York (1993)
10. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: Proccedings of 20th International Conference Very Large Database, pp. 487–499. Morgan Kaufmann, San Francisco (1994)
11. Codd, E.F.: A Relational Model of Data for Large Shared Data Bank. Communication of the ACM 13(6), 377–387 (1970)
12. Benbrahim, H., Amine, B.: A Comparative Study of Pruned Decision Trees and Fuzzy Decision Trees. In: Proceedings of 19th International Conference of the North American, Atlanta, pp. 227–231 (2000)
13. So, Y.D., Sun, J., Wang, X.Z.: An Initial comparison of Generalization-Capability between Crisp and fuzzy Decision Trees. In: Proceedings of the First International Conference on Machine Learning and Cybernetics, pp. 1846–1851 (2002)
14. ALICE d'ISoft v.6.0 demonstration, `http://www.alice-soft.com/demo/al6demo.htm` (Accessed: 31 October 2007)
15. Khoshgoftaar Taghi, M., Liu, Y., Seliya, N.: Genetic Programming-Based Decision Trees for Software Quality Classification. In: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, California, pp. 374–383 (2003)