Validation of Concept Representation from Data
– Concept Representation and Knowledge Discovery in Medical Databases –

Shusaku Tsumoto
Department of Medical Informatics,
Shimane Medical University, School of Medicine
89-1 Enya-cho, Izumo 693-8501 Japan
E-mail: tsumoto@computer.org

Abstract

Conventional studies on knowledge discovery in databases (KDD) shows that combination of rule induction methods and attribute-oriented generalization is very useful to extract knowledge from data. However, attribute-oriented generalization in which concept hierarchy is used for transformation of attributes assumes that a given hierarchy is consistent. Thus, if this condition is violated, application of hierarchical knowledge generates inconsistent rules. In this paper, first, we show that this phenomenon is easily found in data mining contexts: when we apply attribute-oriented generalization to attributes in databases, generalized attributes will have fuzziness for classification. Then, we introduce two approaches to solve this problem, one process of which suggests that combination of rule induction and attribute-oriented generalization can be used to validate concept hiearchy. Finally, we briefly discuss the mathematical generalization of this solution in which context-free fuzzy sets is a key idea.

1 Introduction

Conventional studies on machine learning[11], rule discovery[2] and rough set methods[5, 13, 14] mainly focus on acquisition of rules, the targets of which have mutually exclusive supporting sets. Supporting sets of target concepts form a partition of the universe, and each method search for sets which covers this partition. Especially, Pawlak's rough set theory shows the family of sets can form an approximation of the partition of the universe. These ideas can easily extend into probabilistic contexts, such as shown in Ziarko's variable precision rough set model[19]. However, mutual exclusiveness of the target does not always hold in real-world databases, where conventional probabilistic approaches cannot be applied.

In this paper, first, we show that these phenomena are easily found in data mining contexts: when we apply attribute-oriented generalization to attributes in databases, generalized attributes will have fuzziness for classification, which causes rule induction methods to generate inconsistent rules. Then, we introduce two solutions. The first one is to introduce aggregation operators to recover mathematical consistency. The other one is to introduce Zadeh's linguistic variables, which describes one way to represent an interaction between lower-level components in an upper level components and which gives a simple solution to deal with the inconsistencies. Finally, we briefly discuss the mathematical generalization of this solution in which context-free fuzzy sets is a key idea. In this inconsistent problem, we have to take care about the conflicts between each attributes, which can be viewed as a problem with multiple membership functions.

2 Attribute-Oriented Generalization and Fuzziness

In this section, first, a probabilistic rule is defined by using two probabilistic measures. Then, attribute-oriented generalization is introduced as transforming rules.

2.1 Probabilistic Rules

2.1.1 Accuracy and Coverage

In the subsequent sections, we adopt the following notations, which is introduced in [10].

Let U denote a nonempty, finite set called the universe and A denote a nonempty, finite set of attributes, i.e., $a : U \to V_a$ for $a \in A$, where V_a is called the domain of a, respectively.Then, a decision table is defined as an information system, $A = (U, A \cup \{d\})$.

The atomic formulas over $B \subseteq A \cup \{d\}$ and V are expressions of the form $[a = v]$, called descriptors over B, where $a \in B$ and $v \in V_a$. The set $F(B, V)$ of formulas over B is the least set containing all atomic formulas over

B and closed with respect to disjunction, conjunction and negation.

For each $f \in F(B,V)$, f_A denote the meaning of f in A, i.e., the set of all objects in U with property f, defined inductively as follows.

1. If f is of the form $[a = v]$ then, $f_A = \{s \in U | a(s) = v\}$
2. $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \vee g_A$; $(\neg f)_A = U - f_a$

By the use of this framework, classification accuracy and coverage, or true positive rate is defined as follows.

Definition 1
Let R and D denote a formula in $F(B,V)$ and a set of objects which belong to a decision d. Classification accuracy and coverage(true positive rate) for $R \to d$ is defined as:

$$\begin{aligned} \alpha_R(D) &= \frac{|R_A \cap D|}{|R_A|}(= P(D|R)), \ \textit{and} \\ \kappa_R(D) &= \frac{|R_A \cap D|}{|D|}(= P(R|D)), \end{aligned}$$

where $|A|$ denotes the cardinality of a set A, $\alpha_R(D)$ denotes a classification accuracy of R as to classification of D, and $\kappa_R(D)$ denotes a coverage, or a true positive rate of R to D, respectively.

1

2.1.2 Definition of Rules

By the use of accuracy and coverage, a probabilistic rule is defined as:

$R \stackrel{\alpha,\kappa}{\to} d \quad s.t. \quad R = \wedge_j \vee_k [a_j = v_k], \ \alpha_R(D) \geq \delta_\alpha, \ \kappa_R(D) \geq \delta_\kappa$.

This rule is a kind of probabilistic proposition with two statistical measures, which is an extension of Ziarko's variable precision model(VPRS) [19].[2]

It is also notable that both a positive rule and a negative rule are defined as special cases of this rule, as shown in the next subsections.

[1] Pawlak recently reports a Bayesian relation between accuracy and coverage[8]:

$$\begin{aligned} \alpha_R(D)P(D) &= P(R|D)P(D) = P(R,D) \\ &= P(R)P(D|R) = \kappa_R(D)P(R) \end{aligned}$$

This relation also suggests that *a priori* and *a posteriori* probabilities should be easily and automatically calculated from database.

[2] This probabilistic rule is also a kind of *Rough Modus Ponens*[7].

2.2 Attribute-Oriented Generalization

Rule induction methods regard a database as a decision table[5] and induce rules, which can be viewed as reduced decision tables. However, those rules extracted from tables do not include information about attributes and they are too simple. In practical situation, domain knowledge of attributes is very important to gain the comprehensiblity of induced knowledge, which is one of the reasons why databases are implemented as relational-databases[1]. Thus, reinterpretation of induced rules by using information about attributes is needed to acquire comprehensive rules. For example, terolism, cornea, antimongoloid slanting of palpebral fissures, iris defects and long eyelashes are symptoms around eyes. Thus, those symptoms can be gathered into a category "eye symptoms" when the location of symptoms should be focused on. symptoms should be focused on. The relations among those attributes are hierarchical as shown in Figure 1. This process, grouping of attributes, is called attribute-oriented generalization[1].

Attribute-oriented generalization can be viewed as transformation of variables in the context of rule induction. For example, an attribute "iris defects" should be transformed into an attribute "eye symptoms=yes".It is notable that the transformation of attributes in rules correspond to that of a database because a set of rules is equivalent to a reduced decision table. In this case, the case when eyes are normal is defined as "eye symptoms=no". Thus, the transformation rule for iris defects is defined as:

$$[\textit{iris-defects} = \textit{yes}] \to [\textit{eye-symptoms} = \textit{yes}] \qquad (1)$$

In general, when $[A_k = V_l]$ is a upper-level concept of $[a_i = v_j]$, a transforming rule is defined as:

$$[a_i = v_j] \to [A_k = V_l],$$

and the supporting set of $[A_k = V_l]$ is:

$$[A_i = V_l]_A = \bigcup_{i,j} [a_i = v_j]_a,$$

where A and a is a set of attributes for upper-level and lower level concepts, respectively.

2.3 Examples

Let us illustrate how fuzzy contexts is observed when attribute-oriented generalization is applied by using a small table (Table 1). Then, it is easy to see that a rule of "Aarskog",

$$[\textit{iris-defects} = \textit{yes}] \to \textit{Aarskog} \quad \alpha = 1.0, \kappa = 1.0$$

- *Location*
  - Head · · ·
  - Face · · ·
  - Eye :
    - *terolism*
      - *hyper*
      - *normal*
      - *hypo*
    - *cornea*
      - *megalo*
      - *large*
      - *normal*
    - *antimongoloid slanting of palpebral fissures*
      - *yes*
      - *no*
    - *iris defects*
      - *yes*
      - *no*
    - *eyelashes*
      - *long*
      - *normal*
    - · · ·
  - *Noses* · · ·
  - · · · · · ·

Figure 1: An Example of Attribute Hierarchy

Table 1: A Small Database on Congenital Disorders

U	round	telorism	cornea	slanting	iris-defects	eyelashes	class
1	no	normal	megalo	yes	yes	long	Aarskog
2	yes	hyper	megalo	yes	yes	long	Aarskog
3	yes	hypo	normal	no	no	normal	Down
4	yes	hyper	normal	no	no	normal	Down
5	yes	hyper	large	yes	yes	long	Aarskog
6	no	hyper	megalo	yes	no	long	Cat-cry

DEFINITIONS: round: round face, slanting: antimongoloid slanting of palpebral fissures, Aarskog: Aarskog Syndrome, Down: Down Syndrome, Cat-cry: Cat Cry Syndrome.

is obtained from Table 1.

When we apply transforming rules shown in Figure 1 to the dataset of Table 1, the table is transformed into Table 2. Then, by using transformation rule 1, the above rule is transformed into:

$$[\textit{eye-symptoms} = \textit{yes}] \rightarrow Aarskog.$$

It is notable that mutual exclusiveness of attributes has been lost by transformation. Since five attributes (telorism, cornea, slanting, iris-defects and eyelashes) are generalized into *eye-symptoms*, the candidates for accuracy and coverage will be (2/4, 2/3), (2/4, 3/3), (3/4, 3/3), (3/4, 3/3), (3/3, 3/3) and (3/4, 3/3), respectively. Then, we have to select which value is suitable for the context of this analysis.

In [12], Tsumoto selected the minimum value in medical context: accuracy is equal to 2/4 and coverage is equal to 2/3.

Thus, the rewritten rule becomes the following probabilistic rule:

$$[\textit{eye-symptoms} = \textit{yes}] \rightarrow Aarskog,$$
$$\alpha = 3/4 = 0.75, \kappa = 2/3 = 0.67.$$

This examples show that the loss of mutual exclusiveness is directly connected to the emergence of fuziness in a dataset. It it notable that the rule used for transformation is a deterministic one. When this kind of transformation is applied, whether applied rule is deterministic or not, fuzziness will be observed. However, no researchers has pointed out this problem with combination of rule induction and transformation.

It is also notable that the conflicts between attributes with respect to accuracy and coverage corresponds to the vector representation of membership functions shown in Lin's context-free fuzzy sets[4].

Table 2: A Small Database on Congenital Disorders (Transformed)

U	eye	eye	eye	eye	eye	eye	class
1	no	no	yes	yes	yes	yes	Aarskog
2	yes	yes	yes	yes	yes	yes	Aarskog
3	yes	no	no	no	no	no	Down
4	yes	yes	no	no	no	no	Down
5	yes	yes	yes	yes	yes	yes	Aarskog
6	no	yes	yes	yes	no	yes	Cat-cry

DEFINITIONS: eye: eye-symptoms

2.4 What is a problem ?

The illustrative example in the last subsection shows that simple combination of rule induction and attribute-oriented generalization easily generates many inconsistent rules. One of the most important features of this problem is that simple application of transformation violates mathematical conditions.

Attribute-value pairs can be viewed as a mapping in a mathematical context, as shown in Section 2. For example, in the case of an attribute "*round*", a set of values in "*round*", $\{yes, no\}$ is equivalent to a domain of "round". Then, since the value of *round* for the first example in a dataset, denoted by "1" is equal to 1, $round(1)$ is equal to *no*. Thus, an attribute is a mapping from examples to values. In a reverse way, a set of examples is related to attribute-value pairs:

$$round^{-1}(no) = \{1, 6\}.$$

In the same way, the following relation is obtained:

$$eyeslashes^{-1}(normal) = \{3, 4\}.$$

However, simple transformation will violate this condition on mapping because transformation rules will change different attributes into the same name of generalized attributes. For example, if the following two transformation rules are applied:

$$\begin{aligned} round &\rightarrow \text{eye-symptoms}, \\ eyeslashes &\rightarrow \text{eye-symptoms}, \\ normal &\rightarrow no, \\ long &\rightarrow yes, \end{aligned}$$

then the following relations are obtained:

$$\begin{aligned} \text{eye-symptoms}^{-1}(no) &= \{1, 6\}, \\ \text{eye-symptoms}^{-1}(no) &= \{3, 4\}, \end{aligned}$$

which leads to contradiction. Thus, transformed attribute-value pairs are not mapping because of one to many correspondence.

In this way, violation is observed as generation of logically inconsistent rules, which is equivalent to mathematical inconsistencies.

3 Solutions

3.1 Join Operators

In Subsection 2.3, since five attributes (telorism, cornea, slanting, iris-defects and eyelashes) are generalized into *eye-symptoms*, the candidates for accuracy and coverage will be (2/4, 2/3), (2/4, 3/3), (3/4, 3/3), (3/4,3/3), (3/3, 3/3), and (3/4, 3/3), respectively. Then, we show one approach reported in [12]. the minimum value is selected: accuracy is equal to 2/4 and coverage is equal to 2/3. This selection of minimum value is a kind of *aggregation*, or *join* operator. In join operators, conflict values will be integrated into one values, which means that one to many correspondence is again transformed into one to one correspondence, which will recover consistencies.

Another example of join operators is "average". In the above example, the average of accuracy is 0.71, so if the average operator is selected for aggregation, then the accuracy of the rule is equal to 0.71. This solution can be generalized into context-free fuzzy sets introduced by Lin[4], which is shown in Section 4.

3.2 Zadeh's Linguistic Variables

3.2.1 Concept Hierarchy and Information

Another solution is to observe this problem from the viewpoint of information. After the application of transformation, it is clear that some information is lost. In other words, transformation rules from concept hierarchy are kinds of projection and usually projection loses substantial amounts of information. Intuitively, over-projection is observed as fuzziness.

For example, let me consider the following three transformation rules:

$$[Round = yes] \rightarrow [\text{Eye-symptoms} = yes],$$
$$[\text{Iris-Defects} = yes] \rightarrow [\text{Eye-symptoms} = yes],$$
$$[Telorism = hyper] \rightarrow [\text{Eye-symptoms} = yes]$$

One of the most important questions is whether eyes only contribute to these symptoms.

Thus, one way to solve this problem is to recover information on the hierarchical structure for each symptoms. For example, let us summarize the components of each symptom and corresponding accuracy into Table 3.

It is notable that even if components of symptoms are the same, the values of accuracy are not equal to each other. These phenomena suggest that the degrees of contribution of components are different in those symptoms. In the above examples, the degrees of contribution of Eye in $[Round = yes]$, $[Iris - Defects]$ and $[Telorism]$ are estimated as 1/2 (0.5), 3/3 (1.0) and 2/3(0.67), respectively.

3.2.2 Linguistic Variables and Knowledge Representation

Zadeh proposes linguistic variables to approximate human linguistic reasoning[16, 17, 18]. One of the main points in his discussion is that when human being reasons hierarchical structure, he/she implicitly estimates the degree of contribution of each components to the subject in an upper level.

In the case of a symptom $[Round = yes]$, this symptom should be described as the combination of Eye, Nose and Frontal part of face. From the value of accuracy in Aarskog syndromes, since the contribution of Eye in [Round=yes] is equal to 0.5, the linguistic variable of $[Round = yes]$ is represented as:

$$[Round = yes] =$$
$$0.5 * [Eye] + \theta * [Nose] + (0.5 - \theta) * [Frontal],$$

where 0.5 and θ are degrees of contribution of eyes and nose to this symptom, respectively. It is interesting to see that the real hierarchical structure is recovered by Zadehfs linguistic variable structure, which also suggests that linguistic variables captures one aspect of human reasoning about hierarchical structure. Especially, one important issue is that Zadeh's linguistic variables, although partially, represent the degree of interactions between sub-components in the same hierarchical level, which cannot be achieved by simple application of object-oriented approach.

Another important issue is that the degree of contribution, which can be viewed as a subset of a membership function, can be estimated from data. Estimation of membership function is one of the key issues in application of fuzzy reasoning, but it is a very difficult to extract such membership functions from data and usually they are given by domain experts[9].

In summary, these two important issues suggest that a dataset can be used to validate a concept hierarchy. If some inconsistencies are observed after transformation by a given hierarchy, then some information are thought to be lost in the process of transformation.[3] From the observation of lost information, we can go further into the next step to construct more consistent hierarchy or knowledge representation. Combination of rule induction methods and attribute- oriented generalization may play an important role in validation. Also, it may be possible to measure the quality of concept representation from data. Although this topic is not discussed in this paper, evaluation of concept representation is very important for us to construct complete and sound concept representation. Construction of terminology and concept representation should be adaptive because our medical knowledge is dynamic and new knowledge is coming everyday.

4 Functional Representation of Context-Free Fuzzy Sets

Lin has pointed out problems with multiple membership functions and introduced relations between context-free fuzzy sets and information tables[4]. The main contribution of context-free fuzzy sets to data mining is that information tables can be used to represent multiple fuzzy membership functions. Usually when we meet multiple membership functions, we have to resolve the conflicts between functions. Lin discusses that this resolution is bounded by the context: min, maximum and other fuzzy operators can be viewed as a *context*. The discussion in Section 2 illustrates Lin's assertion. Especially, when we analyze relational-database, transformation will be indispensable to data mining of multi-tables. However, transformation may violate mutual exclusiveness of the target information table. Then, multiple fuzzy membership functions will be observed.

Lin's context-free fuzzy sets shows such analyzing procedures as a simple function as shown in Figure 4. The important parts in this algorithm are the way to construct a list of membership functions and the way to determine whether this algorithm outputs a metalist of a list of membership functions or a list of numerical values obtained by application of fuzzy operators to a list of membership functions.

[3] In this approach, we assume that data does not include errors of experts' decisions.

Table 3: Components of Symptoms

Symptoms	Components	Accuracy
$[Round = yes]$	: [Eye, Nose, Frontal]	$\alpha = 1/2$
$[Iris - Defects = yes]$	: [Substructure of Eye]	$\alpha = 3/3$
$[Telorism = hyper]$	: [Eye, Nose, Frontal]	$\alpha = 2/3$

5 Conclusions

This paper shows that combination of attribute-oriented generalization and rule induction methods generate inconsistent rules and proposes one solution to this problem. It is surprising that tranformation of attributes will easily generate this situation in data mining from relation databases: when we apply attribute-oriented generalization to attributes in databases, generalized attributes will have fuzziness for classification. In this case, we have to take care about the conflicts between each attributes, which can be viewed as a problem with linguistic uncertainty or multiple membership functions. Finally, the author pointed out that these contexts should be analyzed by using two kinds of fuzzy techniques: one is introduction of aggregation operators, which can viewed as those on multiple membership functions. The other one is linguistic variables, which captures the degree of contribution.

This paper is still a preliminary research on combination of medical terminology and KDD methods. Further work should be done, but this combination may be useful to construct and evaluate medical terminology and concept representation. It will be our future work to introduce a measure for evaluation of terminologies and to formalize validation of terminology and concept representation from data.

References

[1] Y.D. Cai, N. Cercone and J. Han, Attribute-oriented induction in relational databases. in: Shapiro, G. P. and Frawley, W. J. (eds), *Knowledge Discovery in Databases*, AAAI press, Palo Alto, CA, pp.213-228 (1991).

[2] Fayyad, U.M., et al.(eds.): *Advances in Knowledge Discovery and Data Mining*, AAAI Press (1996).

[3] Lin, T.Y. Fuzzy Partitions: Rough Set Theory, in *Proceedings of IPMU'98*, Paris, pp. 1167-1174, 1998.

[4] Lin, T.Y. Context Free Fuzzy Sets and Information Tables *Proceedings of EUFIT'98*, Aachen, pp.76-80, 1998.

[5] Pawlak, Z., *Rough Sets*. Kluwer Academic Publishers, Dordrecht, 1991.

[6] Pawlak, Z. Conflict analysis. In: *Proceedings of the Fifth European Congress on Intelligent Techniques and Soft Computing (EUFIT'97)*, pp.1589–1591, Verlag Mainz, Aachen, 1997.

[7] Pawlak, Z. Rough Modus Ponens. *Proceedings of IPMU'98* , Paris, 1998.

[8] Pawlak, Z. Rough Sets and Decision Analysis, *Fifth IIASA workshop on Decision Analysis and Support*, Laxenburg, 1998.

[9] Pedrycz, W. and Gomide, F. *An Introduction to Fuzzy Sets – Analysis and Design*, MIT Press, MA, 1996.

[10] Skowron, A. and Grzymala-Busse, J. From rough set theory to evidence theory. In: Yager, R., Fedrizzi, M. and Kacprzyk, J.(eds.) *Advances in the Dempster-Shafer Theory of Evidence*, pp.193-236, John Wiley & Sons, New York, 1994.

[11] *Readings in Machine Learning*, (Shavlik, J. W. and Dietterich, T.G., eds.) Morgan Kaufmann, Palo Alto, 1990.

[12] Tsumoto, S. Knowledge Discovery in Medical Databases based on Rough Sets and Attribute-Oriented Generalization. *Proceedings of IEEE-FUZZ99*, IEEE Press, Anchorage, 1998.

[13] Tsumoto, S. Automated Induction of Medical Expert System Rules from Clinical Databases based on Rough Set Theory *Information Sciences* **112**, 67-84, 1998.

[14] Tsumoto, S., Automated Discovery of Plausible Rules based on Rough Sets and Rough Inclusion, *Proceedings of PAKDD'99*, (in press), LNAI, Springer-Verlag.

[15] Zadeh, L.A., Toward a theory of fuzzy information granulation and its certainty in human reasoning and fuzzy logic. *Fuzzy Sets and Systems* **90**, 111-127, 1997.

```
procedure Resolution of Multiple Memberships;
  var
    i : integer; L_a, L_i : List;
    A: a list of Attribute-value pairs (multisets:bag);
    F: a list of fuzzy operators;
  begin
    L_i := A;
    while (A ≠ {}) do
      begin
        [a_i = v_j](k) = first(A);
        Applend μ([a_i = v_j](k)) to L_[a_i=v_j]
          /* L_[a_i=v_j]: a list of membership function
             for attribute-value pairs */
        A := A − [a_i = v_j](k);
      end.
    if (F = {}) then
      /* Context- Free */
        return all of the lists L_[a_i=v_j];
    else
      /* Resolution with Contexts*/
      while (F ≠ {}) do
        begin
          f = first(F);
          Apply f to each L_[a_i=v_j];
            μ_f([a_i = v_j]) = f(L_[a_i=v_j])
          Output all of the membership functions
            μ_f([a_i = v_j])
          F := F − f;
        end.
  end {Resolution of Multiple Memberships};
```

Figure 2: Resolution of Multiple Fuzzy Memberships

[16] Zadeh, L.A., The concept of linguistic variable and its application to approximate reasoning. *Information Sciences* **8**, 199-249 (part I), 1975.

[17] Zadeh, L.A., The concept of linguistic variable and its application to approximate reasoning. *Information Sciences* **8**, 301-357 (part II), 1975.

[18] Zadeh, L.A., The concept of linguistic variable and its application to approximate reasoning. *Information Sciences* **9**, 43-80 (part IIII), 1976.

[19] Ziarko, W., Variable Precision Rough Set Model. *Journal of Computer and System Sciences*. 46, 39-59, 1993.