Using Background Knowledge as a Bias to Control the Rule Discovery Process

Ning Zhong[1], Juzhen Dong[1], and Setsuo Ohsuga[2]

[1] Dept. of Information Eng., Maebashi Institute of Technology, Japan
[2] Dept. of Infor. and Computer Science, Waseda University, Japan

Abstract. This paper investigates a way of using background knowledge in the rule discovery process. This technique is based on Generalization Distribution Table (GDT for short), in which the probabilistic relationships between concepts and instances over discrete domains are represented. We describe how to use background knowledge as a bias to adjust the prior distribution so that the better knowledge can be discovered.

1 Introduction

Over the last two decades, many researchers have investigated inductive methods [3] for learning *if-then* rules and concepts from instances. According to the value of information, these methods can be divided into two types. The first type is based on the *formal* value of information; that is, the real meaning of data is not considered in the learning process. ID3 and Prism are the typical methods of this type [5, 1]. Although *if-then* rules can be discovered by using the methods, it is difficult to use background knowledge in the learning process. The other type of inductive methods is based on the *semantic* value of information; that is, the real meaning of data must be considered by using some background knowledge in the learning process. Dblearn is a typical method belonging to this type [3]. It can discover rules by means of background knowledge represented by concept hierarchies, but if there is no background knowledge, it can do nothing. The question is *"how can both the formal value and the semantic value be considered in a discovery system?"*. Unfortunately, so far there is not any inductive learning method that can consider both of the formal value and the semantic value of information at the same time. It is clear that an ideal rule discovery system should have such feature, that is, on one hand, background knowledge can be used flexibly in the discovery process; on the other hand, if no background knowledge is available, it can also work on the formal value of data.

In [7, 8], we proposed a new methodology called GDT (Generalization Distribution Table), for learning *classification* rules in data with uncertainty and incompleteness. The main features of the GDT are

- It can predict unseen instances and represent explicitly the uncertainty of a rule including the prediction of possible instances in the strength of the rule.

[3] The discussion in this paper is limited to *attribute value learning*, which is a major type of inductive learning.

D.A. Zighed, J. Komorowski, and J. Zytkow (Eds.): PKDD 2000, LNAI 1910, pp. 691–698, 2000.

– It can flexibly select biases for search control, and background knowledge can be used as a bias to control the creation of a GDT and the rule induction process.

In [9], we discussed the first feature of the GDT. This paper investigates the second feature of the GDT, that is, a way of using background knowledge in the rule discovery process. This paper is organized as follows: First Section 2 is a brief review of the basic concepts of the GDT. Section 3 discusses how to adjust the prior distribution by background knowledge. Section 4 gives a real world example to show the effects of usage of background knowledge. Finally in Section 6 we summarize our work and point out the future research direction.

2 Generalization Distribution Table (GDT)

The central idea of our methodology is to use a variant of transition matrix, which is called *Generalization Distribution Table (GDT)*, as a hypothesis search space for generalization, in which the probabilistic relationships between concepts and instances over discrete domains are represented [7, 8]. A GDT consists of three components:

The first is *possible instances*, which are denoted in the top row of a GDT, are all possible combinations of attribute values in a database. The number of the possible instances is $\prod_{i=1}^{m} n_i$, where m is the number of attributes, n_i is the number of different attribute values in each attribute i.

The second is *possible generalizations* for instances, which are denoted in the left column of a GDT, are all possible cases of generalization for all possible instances. "$*$", which specifies a wild card, denotes the generalization for instances[4]. For example, the generalization $*b_0c_0$ means the attribute a is unimportant for describing a concept. In other words, if $a = \{a_0, a_1\}$ and both $a_0b_0c_0$ and $a_1b_0c_0$ can describe a concept, attribute a does not seem to be important, that is, from $\{b_0c_0\}$, the concept can be described, and so we use the generalization $*b_0c_0$ to describe the concept, and let $*b_0c_0$ represent the set $\{a_0b_0c_0,\ a_1b_0c_0\}$. The number of the possible generalizations is $\prod_{i=1}^{m}(n_i+1) - \prod_{i=1}^{m} n_i - 1$.

The third is *probabilistic relationships* between the possible instances and the possible generalizations, which are represented in the elements G_{ij} of a GDT, are the probabilistic distribution for describing the strength of the relationship between every possible instance and every possible generalization. The prior distribution is equiprobable, if any prior background knowledge is not used. Thus, it is defined by the Eq. (1), and $\sum_j G_{ij} = 1$:

$$\begin{aligned} G_{ij} &= p(PI_j|PG_i) \\ &= \begin{cases} \dfrac{1}{N_{PG_i}} & \text{if } PI_j \in PG_i \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

[4] For simplicity, the wild card will be omitted in some places in this dissertation.

where PI_j is the jth possible instance, PG_i is the ith possible generalization, and N_{PG_i} is the number of the possible instances satisfying the ith possible generalization, that is,

$$N_{PG_i} = \prod_{k\in\{l|\ PG[l]=*\}} n_k \tag{2}$$

where $PG_i[l]$ is the value of the kth attribute in the possible generalization PG_i, $PG[l] = *$ means that PG_i doesn't contain attribute l.

Furthermore, for convenience, letting $E = \prod_{k=1}^{m} n_k$, Eq. (1) can be changed into the following form:

$$G_{ij} = p(PI_j|PG_i) = \begin{cases} \dfrac{\displaystyle\prod_{k\in\{l|\ PG[l]\neq *\}} n_k}{E} & \text{if } PI_j \in PG_i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

because of

$$\begin{aligned} \frac{1}{N_{PG_i}} = \frac{1}{\displaystyle\prod_{k\in\{l|\ PG[l]=*\}} n_k}. &= \frac{\displaystyle\prod_{k\in\{l|\ PG[l]\neq *\}} n_k}{\displaystyle\prod_{k=1}^{m} n_k} \\ = \frac{\displaystyle\prod_{k\in\{l|\ PG[l]\neq *\}} n_k}{E}. \end{aligned}$$

Since E is a constant for a given database, the prior distribution $p(PI_j|PG_i)$ is directly proportional to the product of the numbers of values of all attributes contained in PG_i.

Thus, in our approach, the basic process of hypothesis generation is to generalize the instances observed in a database by searching and revising the GDT. Here, two kinds of attributes need to be distinguished: *condition* attributes and *decision* attributes (sometimes called class attributes) in a database. Condition attributes as possible instances are used to create the GDT, but the decision attributes are not. The decision attributes are normally used to decide which concept (class) should be described in a rule. Usually a single decision attribute is all that are required.

3 Adjusting the Prior Distribution by Background Knowledge

One of the main features of the GDT methodology is that biases can be selected flexibly for search control, and background knowledge can be used as a bias to control the creation of a GDT and the rule discovery process. This section explains how to use background knowledge as a bias to adjust the prior distribution for learning much better knowledge.

As stated in Section 2, when no prior background knowledge as a bias is available, as default, the occurrence of all possible instances is equiprobable, and the prior distribution of a GDT is shown in Eq. (1). However, the prior distribution can be adjusted by background knowledge, and will be un-equiprobable after the adjustment. Generally speaking, the background knowledge can be given in

$$a_{i_1j_1} \Rightarrow a_{i_2j_2}, \qquad Q,$$

where $a_{i_1j_1}$ is the j_1th value of attribute i_1, and $a_{i_2j_2}$ is the j_2th value of attribute i_2. $a_{i_1j_1}$ is called the *premise* of the background knowledge, $a_{i_2j_2}$ is called the *conclusion* of the background knowledge, and Q is called the *strength* of the background knowledge. It means that Q is the probability of occurrence of $a_{i_2j_2}$ when $a_{i_1j_1}$ occurs. $Q = 0$ means that *"$a_{i_1j_1}$ and $a_{i_2j_2}$ never occur together"*; $Q = 1$ means that *"$a_{i_1j_1}$ and $a_{i_2j_2}$ always occur in the same time"*; while $Q = 1/n_{i_2}$ means that the occurrence of $a_{i_2j_2}$ is the same as the case of without background knowledge, where n_{i_2} is the number of values of attribute i_2. For each instance PI (or each generalization PG), let $PI[i]$ (or $PG[i]$) denote the entry of PI (or PG) corresponding to attribute i. For each generalization PG such that $PG[i_1] = a_{i_1j_1}$ and $PG[i_2] = *$, the prior distribution between the PG and related instances will be adjusted. The probability of occurrence of attribute value $a_{i_2j_2}$ is changed from $1/n_{i_2}$ to Q by background knowledge, so that, for each of the other values in attribute i_2, the probability of its occurrence is changed from $1/n_{i_2}$ to $(1-Q)/(n_{i_2}-1)$. Let the adjusted prior distribution be denoted by p_{bk}. The prior distribution adjusted by the background knowledge "$a_{i_1j_1} \Rightarrow a_{i_2j_2}, \quad Q$" is

$$p_{bk}(PI|PG) = \begin{cases} p(PI|PG) \times Q \times n_{i_2} & \text{if } PG[i_1] = a_{i_1j_1}, PG[i_2] = *, \\ & \quad PI[i_2] = a_{i_2j_2} \\ p(PI|PG) \times \dfrac{1-Q}{n_{i_2}-1} \times n_{i_2} & \text{if } PG[i_1] = a_{i_1j_1}, PG[i_2] = *, \\ & \quad \exists j (1 \le j \le n_{i_2}, j \ne j_2)\ PI[i_2] = a_{i_2j} \\ p(PI|PG) & otherwise \end{cases} \tag{4}$$

where coefficients of $p(PI|PG)$, $Q \times n_{i_2}$, $\frac{1-Q}{n_{i_2}-1} \times n_{i_2}$, and 1 are called *adjusting factor* (AF for short) with respect to the background knowledge "$a_{i_1j_1} \Rightarrow a_{i_2j_2}, \quad Q$". They explicitly represent the influence of a piece of background knowledge to the prior distribution. Hence, the adjusted prior distribution can be denoted by

$$p_{bk}(PI|PG) = p(PI|PG) \times AF(PI|PG), \tag{5}$$

and the AF is

$$AF(PI|PG) = \begin{cases} Q \times n_{i_2} & \text{if } PG[i_1] = a_{i_1 j_1}, PG[i_2] = *, \\ & \quad PI[i_2] = a_{i_2 j_2} \\ \dfrac{1-Q}{n_{i_2}-1} \times n_{i_2} & \text{if } PG[i_1] = a_{i_1 j_1}, PG[i_2] = *, \\ & \quad \exists j (1 \le j \le n_{i_2}, j \ne j_2)\ PI[i_2] = a_{i_2 j} \\ 1 & \mathit{otherwise.} \end{cases} \tag{6}$$

So far, we have explained how the prior distribution is influenced by only one piece of background knowledge. We then consider the case that there are several pieces of background knowledge such that for each i $(1 \le i \le m)$ and each j $(1 \le j \le n_i)$, there is at most only one piece of background knowledge with a_{ij} as its conclusion.

Let S be the set of all pieces of background knowledge to be considered. For each generalization PG, let

$B[S, PG] = \{i \in \{1, \ldots, m\}|$
$\exists i_1 (1 \le i_1 \le m)\ \exists j_1 (1 \le j_1 \le n_{i_1})\ \exists j (1 \le j \le n_i)$
[(there is a piece of background knowledge in S with $a_{i_1 j_1}$ as its premise and with a_{ij} as its conclusion) & $PG[i_1] = a_{i_1 j_1}$ & $PG[i] = *$] },

and for each $i \in B[S, PG]$, let

$J[S, PG, i] = \{j \in \{1, \ldots n_i\}|\ \exists i_1 (1 \le i_1 \le m)\ \exists j_1 (1 \le j_1 \le n_{i_1})$
[(there is a piece of background knowledge in S with $a_{i_1 j_1}$ as its premise and with a_{ij} as its conclusion) & $PG[i_1] = a_{i_1 j_1}$ & $PG[i] = *$] }.

Then, we must use the following *adjusting factors* AF_S with respect to all pieces of background knowledge,

$$AF_S(PI|PG) = \prod_{i=1}^{m} AF_i(PI|PG) \tag{7}$$

where

$$AF_i(PI|PG) \tag{8}$$
$$= \begin{cases} Q_{ij} \times n_i & \text{if } i \in B[S, PG],\ j \in J[S, PG, i], \text{ and } PI[i] = a_{ij} \\ \dfrac{1 - \sum\limits_{j \in J[S,PG,i]} Q_{ij}}{n_i - |J[S, PG, i]|} \times n_i & \text{if } i \in B[S, PG], \\ & \quad \forall j (j \in J[S, PG, i])[PI[i] \ne a_{ij}] \\ 1 & \text{otherwise} \end{cases}$$

where for each i $(1 \le i \le m)$ and each j $(1 \le j \le n_i)$, Q_{ij} denotes the strength of the background knowledge (in S) with a_{ij} as its conclusion.

Although Q can be any value from 0 to 1 in principle, giving an exact value of Q is difficult, and the more the background knowledge, the more difficult to calculate the prior distribution. Hence, in practice, if "$a_{i_1j_1} \Rightarrow a_{i_2j_2}$" with higher possibility, we treat that Q is 1, that is, $a_{i_2j_2}$ occurs but other values of attribute i_2 do not, when $a_{i_1j_1}$ occurs. In contrast, if "$a_{i_1j_1} \Rightarrow a_{i_2j_2}$" with lower possibility, we treat that Q is 0, that is, $a_{i_2j_2}$ does not occur but other values of attribute i_2 occur in equiprobable, when $a_{i_1j_1}$ occurs. Furthermore, if several pieces of background knowledge with higher possibility, and the conclusions of them belong to the same attribute i_2, all of the attribute values (conclusions) are treated as occurrence in equiprobable, but other values in attribute i_2 are treated as no occurrence.

4 An Application

This section describes a result of an experiment in which background knowledge is used in the learning process to discover rules from a meningitis database [2]. The database was collected at the Medical Research Institute, Tokyo Medical and Dental University. It has 140 instances, each of which is described by 38 attributes that can be categorized into present history, physical examination, laboratory examination, diagnosis, therapy, clinical course, final status, risk factor etc. The task is to find important factors for diagnosis (bacteria and virus, or their more detail classifications) and predicting prognosis. A more detailed explanation on this database could be found at http://www.kdel.info.eng.osaka-cu.ac.jp/SIGKBS.

For each of the decision attributes: DIAG2, DIAG, CULTURE, C_COURSE, and COURSE(Grouped), we run our rule discovery system named GDT-RS, which is a way of implementation of the GDT methodology by combining the GDT with rough sets [10], on it twice: using background knowledge and without using background knowledge, to acquire the rules respectively. For the discretization of continuous attributes, an automatic discretization [4] is used.

4.1 Background Knowledge Given By a Medical Doctor

The experience of a medical doctor can be used as background knowledge:

> If the brain wave (EEG-WAVE) is normal, the focus of brain wave (EEG_FOCUS) is never abnormal;
> If the number of white blood cells (WBCs) is high, the inflammation protein (CRP) is also high.

In the following list, a part of the background knowledge given by a medical doctor is described:

- Never occurring together.

$$\begin{array}{ll} EEG\_WAVE(normal) & \Leftrightarrow EEG\_FOCUS(+) \\ CSF\_CELL(low) & \Leftrightarrow Cell\_Poly(high) \\ CSF\_CELL(low) & \Leftrightarrow Cell\_Mono(high) \end{array}$$

– Occurring with lower possibility.

$WBC(low) \Rightarrow CRP(high)$
$WBC(low) \Rightarrow ESR(high)$
$WBC(low) \Rightarrow CSF\_CELL(high)$
$WBC(low) \Rightarrow Cell\_Poly(high)$

– Occurring with higher possibility.

$WBC(high) \Rightarrow CRP(high)$
$WBC(high) \Rightarrow ESR(high)$
$WBC(high) \Rightarrow CSF\_CELL(high)$
$EEG\_FOCUS(+) \Rightarrow FOCAL(+)$
$EEG\_WAVE(+) \Rightarrow EEG\_FOCUS(+)$

Here “high” in brackets denoted in the background knowledge means that the value is greater than the maximal value in the normal values range; and “low” means that the value is less than the minimal value in the normal values range.

4.2 Comparing the Results

The effects of usage of the background knowledge in GDT-RS are as follows:

First, some candidates of rules, which are deleted due to lower strengths during rule discovery without background knowledge, are selected. For example, $rule_1$ is deleted when no background knowledge is used, but after using the background knowledge stated above, it is reserved because its strength increased 4 times.

$rule_1$:

$$ONSET(acute) \wedge ESR(\leq 5) \wedge CSF\_CELL(> 10) \wedge CULTURE(-) \rightarrow VIRUS(E).$$

Without using background knowledge, the strength S of $rule_1$ is 30*(384/E). In the background knowledge given above, there are two clauses related to this rule:

- Never occurring together
$CSF\_CELL(low) \Leftrightarrow Cell\_Poly(high)$
$CSF\_CELL(low) \Leftrightarrow Cell\_Mono(high).$

By using automatic discretization to continuous attributes $Cell\_Poly$ and $Cell\_Mono$, the attribute values in each of attributes are divided into two groups: high and low. Since the high groups of $Cell\_Poly$ and $Cell\_Mono$ do never occur when $CSF\_CELL(low)$ occurs, the product of the numbers of attribute values is decreased to $E/4$, and the strength S is increased to $S = 30 * (384/E) * 4$.

Second, using background knowledge also causes some rules to be replaced by others. For example, the rule

$rule_2$:

$DIAG(VIRUS(E)) \wedge LOC[4, 7) \rightarrow EEG\_abnormal, \quad S = 30/E$

can be discovered without background knowledge, but after using the background knowledge stated above, it is replaced by

$rule_{2'}$:

$EEG\_FOCUS(+) \wedge LOC[4,\ 7) \rightarrow EEG\_abnormal, \quad S = (10/E) * 4.$

The reason is that both of them contain the same instances, but the strength of $rule_{2'}$ becomes larger than that of $rule_2$.

The result has been evaluated by a medical doctor. According to his opinions, both $rule_1$ and $rule_{2'}$ are reasonable, and $rule_{2'}$ is much better than $rule_2$.

Although the similar results can be obtained from the meningitis database by using GDT-RS and C4.5 if such background knowledge is not used, it is difficult that such background knowledge is used in C4.5 [6].

5 Conclusion

We presented a new way of using background knowledge in the rule discovery process. The theoretical and experimental results show that our methodology is very flexible. That is, on one hand, background knowledge can be used flexibly in the discovery process so that the better knowledge can be discovered; on the other hand, if no background knowledge is available, it can also work on the formal value of data.

Our future work includes cooperatively using such background knowledge and different types of background knowledge as biases in more different aspects of the rule discovery process to learn much better knowledge.

References

1. J. Cendrowska, "PRISM: An Algorithm for Inducing Modular Rules", *International Journal of Man-Machine Studies*, Vol.27, (1987) 349-370.
2. Dong, J.Z., Zhong, N., and Ohsuga, S. "Rule Discovery from the Meningitis Database by GDT-RS" (Special Panel Discussion Session on Knowledge Discovery from a Meningitis Database) *Proc. the 12th Annual Conference of JSAI* (1998) 83-84.
3. J. Han, Y. Cai, and N. Cercone, "Data-Driven Discovery of Quantitative Rules in Relational Databases", *IEEE Trans. Knowl. Data Eng.*, Vol.5, No.1 (1993) 29-40.
4. S.H. Nguyen and H.S. Nguyen, "Discretization Methods in Data Mining", L. Polkowski and A. Skowron (eds.) *Rough Sets in Knowledge Discovery*, Vol.1, Physica-Verlag (1998) 451-482.
5. J.R. Quinlan, "Induction of Decision Trees", *Machine Learning*, Vol.1 (1986) 81-106.
6. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).
7. N. Zhong and S. Ohsuga, "Using Generalization Distribution Tables as a Hypotheses Search Space for Generalization". *Proc. 4th Inter. Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD-96)* (1996) 396-403.
8. N. Zhong, J.Z. Dong, and S. Ohsuga, "Discovering Rules in the Environment with Noise and Incompleteness", *Proc. 10th Inter. Florida AI Research Symposium (FLAIRS-97)* edited in the *Special Track on Uncertainty in AI* (1997) 186-191.
9. N. Zhong, J.Z. Dong, and S. Ohsuga, "Data Mining: A Probabilistic Rough Set Approach", L. Polkowski and A. Skowron (eds.) *Rough Sets in Knowledge Discovery*, Vol.2, In Studies in Fuzziness and Soft Computing series, Vol.19, Physica-Verlag (1998) 127-146.
10. N. Zhong, J.Z. Dong, and S. Ohsuga, "GDT-RS: A Probabilistic Rough Induction System", Bulletin of International Rough Set Society, Vol.3, No.4 (1999) 133-146.