Simultaneous Reliability Evaluation of Generality and Accuracy for Rule Discovery in Databases

Einoshin Suzuki

Electrical and Computer Engineering,
Yokohama National University,
79-5, Tokiwadai, Hodogaya, Yokohama 240-8501, Japan.
suzuki@dnj.ynu.ac.jp

Abstract

This paper presents an algorithm for discovering conjunction rules with high reliability from data sets. The discovery of conjunction rules, each of which is a restricted form of a production rule, is well motivated by various useful applications such as semantic query optimization and automatic development of a knowledge base. In a discovery algorithm, a production rule is evaluated according to its generality and accuracy since these are widely accepted as criteria in learning from examples. Here, reliability evaluation for these criteria is mandatory in distinguishing reliable rules from unreliable patterns without annoying the users. However, previous discovery approaches have either ignored reliability evaluation or have only evaluated the reliability of generality, and consequently, tend to discover a huge number of rules. In order to circumvent these difficulties we propose an approach based on a simultaneous estimation. Our approach discovers the rules that exceed pre-specified thresholds for generality and accuracy with high reliability. A novel pruning method is employed for improving time efficiency without changing the discovery outcome. The proposed approach has been validated experimentally using 21 benchmark data sets from the UCI repository.

Introduction

The discovery of production rules and their relatives is one of the most important topics in KDD due to its generality. Among the relatives, a conjunction rule (Smyth & Goodman 1992), in which A and B represent a conjunction of atoms and a single atom respectively, and an association rule (Agrawal *et al.* 1996), in which A and B represent a set of examples and every attribute is binary, are considered as the most important. The discovery of conjunction rules is well motivated by various useful applications such as semantic query optimization (Hsu & Knoblock 1996) and automatic development of a knowledge-base (Smyth & Goodman 1992).

Since the usefulness of a production rule cannot be formalized explicitly, one of the most important points in discovering such rules from a database is how to evaluate their "interestingness" i.e. their potential usefulness. Such evaluation methods for the interestingness of a production rule are classified into two categories: one assumes a single criterion which is a combination of generality and accuracy (Klösgen 1996; Smyth & Goodman 1992), and the other specifies minimum thresholds for generality and accuracy (Agrawal *et al.* 1996). Here, these methods employ point-estimated probabilities in the evaluation, and are therefore indifferent to the size of a database and to the reliability of a discovered rule. Typically, careful examination for the interestingness is desirable in order to keep the number of discovered rules as small as possible, since a huge number of rules can be discovered from a database. A precise evaluation of reliability would considerably reduce the effort of the users for checking the discovery output, since only a small number of rules with high reliability are discovered in this case. However, previous methods of reliability evaluation employed in KDD (Agrawal *et al.* 1996; Chan & Wong 1991; Siebes 1994) estimate the confidence interval of a single probability, and are thereby inadequate for the evaluation of accuracy since it is defined as a conditional probability.

In order to circumvent these difficulties we propose a novel approach in which conjunction rules are discovered according to their reliability levels based on a simultaneous evaluation of generality and accuracy. This approach employs the normal approximations of multinomial distributions. Experimental results confirm that the proposed approach is effective both in filtering out less reliable rules and in reducing computational time.

Description of the Problem

Let an example e_i be a description about an object stored in a data set in the form of a record, then a data set contains n examples $e_1, e_2, \cdots, e_n$. An example e_i is represented by a tuple $< a_{i1}, a_{i2}, \cdots, a_{im} >$ where $a_{i1}, a_{i2}, \cdots, a_{im}$ are values for m discrete attributes. Here, continuous attributes are supposed to be discretized using an existing method such as (Dougherty

et al. 1995; Fayyad & Irani 1993). An event representing, in propositional form, a single value assignment to an attribute will be called an atom.

A conjunction rule (Smyth & Goodman 1992) $r(\mu)$ is a production rule which represents a probabilistic correlation or causality between its premise and conclusion, and is represented by

$$r(\mu) \equiv A_\mu \rightarrow c, \quad (1)$$
$$A_\mu \equiv a_1 \wedge a_2 \wedge \cdots \wedge a_\mu, \quad (2)$$

where a_i and c are single atoms. In this paper, we consider the problem of finding a set of conjunction rules from a data set.

In learning from examples, generality and accuracy are considered as the most general criteria for evaluating the goodness of a hypothesis. In case of a conjunction rule $A_\mu \rightarrow c$, these two criteria correspond to the point-estimated probabilities $\hat{p}(A_\mu)$ and $\hat{p}(c|A_\mu)$ respectively (Smyth & Goodman 1992).

Existing methods for evaluating the generality and the accuracy of a production rule can be classified into two approaches: the single expression approach such as (Klösgen 1996; Smyth & Goodman 1992) which assumes a single criterion defined by a combination of two criteria, and the simultaneous approach such as (Agrawal *et al.* 1996) which specifies a minimum threshold for each criterion.

Besides, several methods have been proposed to evaluate the reliability of generality (Agrawal *et al.* 1996; Chan & Wong 1991; Siebes 1994), which estimate the true probability $p(A_\mu)$ of the premise. These methods can be employed for a modified version of the simultaneous approach: specify a minimum threshold for $p(A_\mu)$ with some confidence level and another minimum threshold for $\hat{p}(c|A_\mu)$. Note that a smaller number of rules are discovered in this modified version since the lower-bound of the true probability $p(A_\mu)$ is smaller than $\hat{p}(A_\mu)$.

We take the simultaneous approach due to its generality, and for a more detailed evaluation of a rule, consider both the reliability of its generality and the reliability of its accuracy. We discover the rules $r(\mu)$ of which true probabilities $p(A_\mu)$ and $p(c|A_\mu)$ are greater than or equal to their respective minimum thresholds θ_S, θ_F with a confidence level $1-\delta$.

$$\Pr\{p(A_\mu) \geq \theta_S,\ p(c|A_\mu) \geq \theta_F\} \geq 1-\delta. \quad (3)$$

Evaluation of Reliability

Previous Methods

In KDD, the Chernoff bound is frequently used in assessing the reliability of a discovered production rule (Agrawal *et al.* 1996; Siebes 1994). Consider the problem of estimating the true probability p of an atom from a data set. Let the probability obtained by point estimation be $\hat{p}$, and $\ln(x)$ be the natural logarithm of x. Then according to the theorem, the $1-\delta$ confidence interval for the probability p is given as follows (Siebes 1994).

$$\hat{p} - \sqrt{\frac{1}{n}\ln\left(\frac{2}{\delta}\right)} \leq p \leq \hat{p} + \sqrt{\frac{1}{n}\ln\left(\frac{2}{\delta}\right)}. \quad (4)$$

This problem, however, can be also resolved by the normal approximations of the binomial distributions. Let α_δ satisfy

$$\frac{1}{2\pi}\int_{-\alpha_\delta}^{\alpha_\delta} \exp\left(\frac{-x^2}{2}\right) dx = 1-\delta. \quad (5)$$

Then the confidence interval is given by

$$\hat{p} - \alpha_\delta\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + \alpha_\delta\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \quad (6)$$

Note that (4) and (6) are for estimating a single probability. Therefore, neither (4) nor (6) are adequate for our problem since (3) contains a conditional probability.

Normal Approximations of the Multinomial Distributions

The above discussions suggest that we should estimate the confidence region of the probabilities related to (3). This paper employs the normal approximations of the multinomial distributions. First, atoms D_1, D_2, D_3 are defined as follows.

$$D_1 \equiv c \wedge A_\mu, \quad (7)$$
$$D_2 \equiv \overline{c} \wedge A_\mu, \quad (8)$$
$$D_3 \equiv \overline{A_\mu}. \quad (9)$$

Since the case $\exists\ i\ \hat{p}(D_i) = 0$ is easily handled by the normal approximations of the binomial distributions, we first show the results for the case $\forall\ i\ \hat{p}(D_i) \neq 0$. From (7) $\sim$ (9), the events D_1, D_2, D_3 are mutually exclusive and exhaustive, hence we can assume that their respective numbers of occurrence (x_1, x_2, x_3) are multinomially distributed. Let (u_1, u_2, u_3) be the respective numbers of the examples (x_1, x_2, x_3) in the data set, and

$$\vec{u} \equiv (u_1, u_2), \quad (10)$$
$$\vec{x} \equiv (x_1, x_2). \quad (11)$$

Assuming that n is enough large, the above multinomial distribution is approximated by a 2-dimensional normal distribution, of which probability density function is given by

$$f(\vec{x}) = \frac{1}{2\pi|\boldsymbol{H}|^{1/2}} \exp\left\{-\frac{{}^t(\vec{x}-\vec{u})\boldsymbol{H}^{-1}(\vec{x}-\vec{u})}{2}\right\}. \quad (12)$$

Here, $\boldsymbol{H}$ and $\boldsymbol{H}^{-1}$ represent the covariance matrix and its inverse matrix given by (13) and (14) respectively, and ${}^t\boldsymbol{G}$ is the transposed matrix of a matrix $\boldsymbol{G}$.

$$\boldsymbol{H} = \frac{1}{n}\begin{pmatrix} u_1(n-u_1) & -u_1u_2 \\ -u_1u_2 & u_2(n-u_2) \end{pmatrix}, \quad (13)$$

$$\boldsymbol{H}^{-1} = \frac{1}{n-u_1-u_2}\begin{pmatrix} \frac{n-u_2}{u_1} & 1 \\ 1 & \frac{n-u_1}{u_2} \end{pmatrix}. \quad (14)$$

Consider a region covered by an ellipse

$$V_\delta : {}^t(\vec{x}-\vec{u})\boldsymbol{H}^{-1}(\vec{x}-\vec{u}) \leq \beta_\delta^2, \quad (15)$$

which satisfy

$$\Pr(\vec{x} \in V_\delta) = 1-\delta. \quad (16)$$

This ellipse V_δ corresponds to the $1-\delta$ confidence region of $\vec{x}$, and β_δ can be calculated by numerical integration. From (14) and (15), V_δ is given as follows.

$$\frac{1}{n-u_1-u_2}\left\{\frac{n-u_2}{u_1}(x_1-u_1)^2 + \frac{n-u_1}{u_2} (x_2-u_2)^2 + 2(x_1-u_1)(x_2-u_2)\right\} \leq \beta_\delta^2. \quad (17)$$

From (7) $\sim$ (9), (3) represents a problem of judging whether (18) and (19) always hold in (17).

$$\frac{x_1+x_2}{n} \geq \theta_{\mathrm{S}}, \quad (18)$$

$$\frac{x_1}{x_1+x_2} \geq \theta_{\mathrm{F}}. \quad (19)$$

Since an expression on the left hand side in (18) or (19) being a constant represents a line, the maximum and the minimum of each expression occur at the extremes of the ellipse. Let λ be an undetermined multiplier, then according to the Lagrange's multiplier method, an extremum of an expression f under $g=0$ satisfy

$$\left(\frac{\partial f}{\partial x_1}\frac{\partial f}{\partial x_2}\right) = \lambda\left(\frac{\partial g}{\partial x_1}\frac{\partial g}{\partial x_2}\right). \quad (20)$$

Let g be the ellipse in (17), and f_1 and f_2 be the expression on the left hand side in (18) and (19) respectively. Appling (20) for $f=f_1$ and $f=f_2$, we obtain

$$\left(1-\beta_\delta\sqrt{\frac{1-\hat{p}(A_\mu)}{n\hat{p}(A_\mu)}}\right)\hat{p}(A_\mu) \geq \theta_{\mathrm{S}}, \quad (21)$$

$$\left(1-\beta_\delta\sqrt{\frac{\hat{p}(\bar{c},A_\mu)}{\hat{p}(c,A_\mu)\{(n+\beta_\delta^2)\hat{p}(A_\mu)-\beta_\delta^2\}}}\right) \hat{p}(c|A_\mu) \geq \theta_{\mathrm{F}}. \quad (22)$$

Similarly, we can show that (21) and (22) also hold for $\exists\, i\ \hat{p}(D_i)=0$. Therefore, we obtain a set of conjunction rules each of which satisfies (21) and (22).

Discovery Algorithm

In our algorithm, a discovery task is viewed as a search problem, in which a node of a search tree represents a conjunction rule $r(\mu)$. Let $\mu=0$ represents the state in which the premise of a rule $r(\mu)$ contains no atoms, then we define that $\mu=0$ holds in a node of depth 1, and as the depth increases by 1, an atom is added to the premise.

A depth-first search method is employed to traverse this tree, and the maximum value M of μ is given by the user. To alleviate the inevitable inefficiency of depth-first search, we employ the following theorem. Based on the theorem, the nodes which satisfy at least one of the stopping criteria (23) and (24) are not expanded without altering the algorithm's output.

Theorem 1 *Let the rule of the current node be $r(\mu')$. If the rule $r(\mu')$ satisfies either (23) or (24), no rules $r(\mu)$ of the descendant nodes satisfy both (21) and (22).*

$$\left(1-\beta_\delta\sqrt{\frac{1-\hat{p}(A_{\mu'})}{n\hat{p}(A_{\mu'})}}\right)\hat{p}(A_{\mu'}) < \theta_{\mathrm{S}}, \quad (23)$$

$$\left(1-\beta_\delta\sqrt{\frac{1-\hat{p}(A_{\mu'})}{n\hat{p}(A_{\mu'})}}\right)\hat{p}(c,A_{\mu'}) < \theta_{\mathrm{S}}\theta_{\mathrm{F}}. \quad (24)$$

Proof Assume a rule $r(\mu)$ of a descendant node satisfies both (21) and (22). First, function $(1-\beta_\delta\sqrt{(1-x)/n/x})$ increases monotonically from n, $\beta_\delta > 0$. Then, contradictions can be derived from (21), (22), $\hat{p}(A_{\mu'}) \geq \hat{p}(A_\mu)$ and $\hat{p}(c,A_{\mu'}) \geq \hat{p}(c,A_\mu)$. □

In rule discovery, we often discover similar rules $A_\mu \rightarrow c$ and $A'_\mu \rightarrow c$, where A'_μ is a specialization of A_μ. For an easy interpretation of the results, such a rule $A'_\mu \rightarrow c$ is not outputted if $\hat{p}(c|A'_\mu) \leq \hat{p}(c|A_\mu)$.

Application to Data Sets

The proposed method was tested with data sets from several domains, including 21 benchmark data sets (Merz *et al.* 1996) in the machine learning community.

We have compared our method with the previous methods, where the parameters were set to $M=3$, $\delta = 0.05$, $\theta_{\mathrm{S}}=0.1$ and $\theta_{\mathrm{F}}=0.9$. In the experiments, a continuous attribute is discretized in advance by (Fayyad & Irani 1993). Figure 1 shows the ratio of the number of discovered rules between the three approaches and the approach without reliability evaluation. A bullet ($\bullet$), a times ($\times$) and a triangle ($\triangle$) correspond to our approach, the approach based on the Chernoff bound and the approach based on the normal approximations of the binomial distributions respectively. In the figure, a bar chart represents the number of rules discovered by the approach without reliability evaluation in a logarithmic scale, and data sets, which are shown along the horizontal axe, are sorted with respect to these numbers.

From figure 1, we note that the proposed approach reduces a considerable number of rules in many data sets compared with the previous approaches. In the "german" data set, for example, it reduces 88% and 81% of the rules compared with the approach based on the normal approximations of the binomial distributions and with the approach based on the Chernoff bound respectively. Since 2,352 rules are discovered by

the no evaluation approach from this data set, the proposed method is effective in reducing the user's effort of verifying the discovered rules. Same conclusion can be obtained from the experiments with the "satellite" data set, in which the proposed method reduces more than 51,128 rules compared with the other methods. The proposed method, since it evaluates more information than the other methods, reduces a larger number of rules or at least the same number of rules. It reduces only a small fraction of rules in the "nursery" and "shuttle" data set, the reason of which is difficult to be explained since it concerns the distribution of the attribute values in the data sets. A simple explanation is that these data sets have a small number of rules with low reliability to be reduced since the numbers of attributes are relatively small (at most 10) and the numbers of examples are large (at least 12,960). However, such a data set, which has a large number of examples and a small number of attributes and attribute values, is considered to be rare in KDD.

In the three algorithms presented in the previous experiments, discovery is regarded as a search problem, and its execution time is known to be approximately propositional to the number of searched nodes. Figure 2 shows the ratio of the number of searched nodes between the three approaches and the approach without reliability evaluation. To be fair, we have derived stopping criteria which correspond to theorem 1 for every method, and have employed them in the experiments.

From figure 2, we note that the proposed approach reduces a considerable number of nodes in many data sets compared with the previous approaches. In the "satellite" data set, for example, it reduces 80% and 77% of the nodes compared with the approach based on the normal approximations of the binomial distributions and with the approach based on the Chernoff bound respectively. The proposed method, since it evaluates more information than the other methods, reduces a larger number of nodes or at least the same number of nodes. The effectiveness of the approach, however, is difficult to be analyzed since it relates complicatedly with the probability distribution of the attribute values in the data set.

From the above experiments, the proposed approach is superior to or at least as effective as the existing approaches in terms of filtering out less reliable rules and pruning the searched nodes. We can safely conclude that our approach reduces both the user's effort and the computational time, and is therefore mandatory for the efficient discovery of reliable rules.

Conclusion

This paper has described an approach based on a stochastic estimation for discovering rules with high reliability. The previous approaches, since they have neglected reliability evaluation or have only evaluated the reliability of generality, had problems of discovering a huge number of unnecessary rules, causing a considerable overload to their users. Our approach remedies this problem by a simultaneous estimation of the confidence region based on the normal approximations of the multinomial distribution. Consequently, our approach filters out rules with low reliability by evaluating both the reliability of generality and the reliability of accuracy. Moreover, we have derived stopping criteria to improve search efficiency without altering the discovery results.

The proposed approach has been applied to 21 data sets from the UCI repository (Merz *et al.* 1996). Experimental results show that our approach reduces both the number of discovered rules and the number of searched nodes compared with the existing approaches. The proposed approach is especially effective in reliable rule discovery in databases with a large number of attributes and attribute values. Moreover, it would be also effective when the computational time is limited due to its use of newly-derived stopping criteria.

References

Agrawal, R., Mannila, H., Srikant, R., *et al.* 1996. Fast Discovery of Association Rules, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, pp. 307–328.

Chan, K. C. C. and Wong, A. K. C. 1991. A Statistical Technique for Extracting Classificatory Knowledge from Databases, *Knowledge Discovery in Databases*, AAAI Press/The MIT Press, pp. 107–123.

Dougherty, J., Kohavi, R. and Sahami, M. 1995. Supervised and Unsupervised Discretization of Continuous Features, *Proc. of ICML-95*, pp. 194–202.

Fayyad, U. M. and Irani, K. B. 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, *Proc. of IJCAI-93*, pp. 1022–1027.

Hsu, C. and Knoblock, C. A. 1996. Using Inductive Learning to Generate Rules for Semantic Query Optimization, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/ The MIT Press, pp. 425–445.

Klösgen, W. 1996. Explora: A Multipattern and Multistrategy Discovery Approach, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/ The MIT Press, pp. 249–271.

Merz, C. J. and Murphy, P. M. 1994. UCI Repository of machine learning databases, *http:// www.ics.uci.edu/ ~mlearn/MLRepository. html*, Univ. of California, Dept. of Information and Computer Sci.

Siebes, A. 1994. Homogeneous Discoveries Contain No Surprises: Inferring Risk-profiles from Large Databases, *AAAI-94 Workshop on Knowledge Discovery in Databases*, pp. 97–107.

Smyth, P. and Goodman, R. M. 1992. An Information Theoretic Approach to Rule Induction from Databases, *IEEE Trans. on Knowledge and Data Eng.*, 4 (4), pp. 301–316.

Figure 1: Performance of the 3 methods with respect to the number of discovered rules. The left scale is for line graphs, and the right scale is for bar charts.

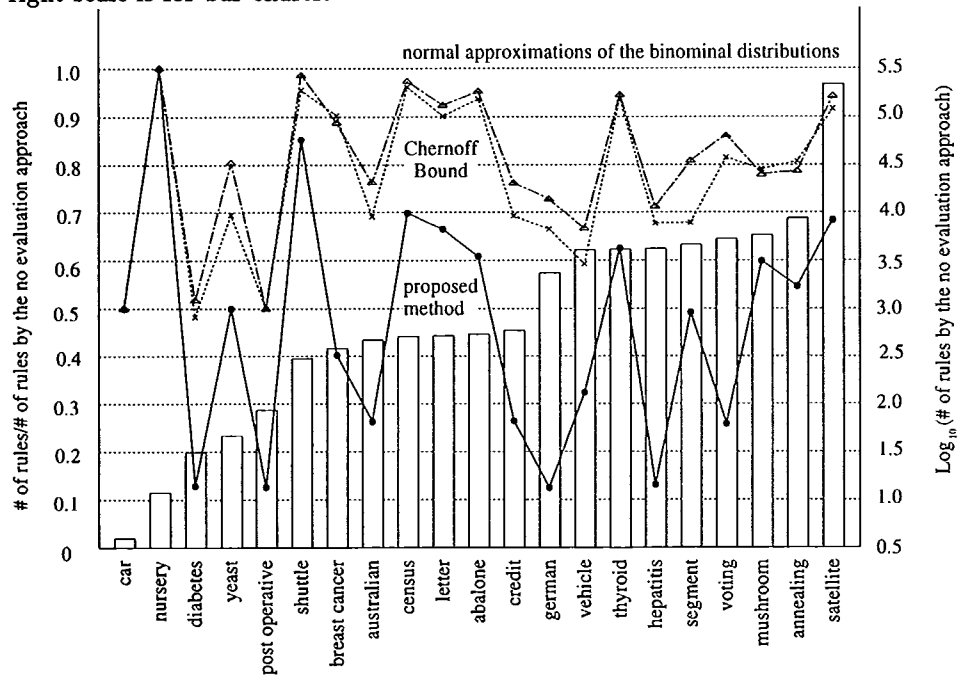

Figure 2: Performance of the 3 methods with respect to the number of searched nodes. The left scale is for line graphs, and the right scale is for bar charts.