Autonomous Discovery of Reliable Exception Rules

Einoshin Suzuki

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

Abstract

This paper presents an autonomous algorithm for discovering exception rules from data sets. An exception rule, which is defined as a deviational pattern to a well-known fact, exhibits unexpectedness and is sometimes extremely useful in spite of its obscurity. Previous discovery approaches for this type of knowledge have neglected the problem of evaluating the reliability of the rules extracted from a data set. It is clear, however, that this question is mandatory in distinguishing knowledge from unreliable patterns without annoying the users. In order to circumvent these difficulties we propose a probabilistic estimation approach in which we obtain an exception rule associated with a common sense rule in the form of a rule pair. Our approach discovers, based on the normal approximations of the multinomial distributions, rule pairs which satisfy, with high confidence, all the specified conditions. The time efficiency of the discovery process is improved by the newly-derived stopping criteria. PEDRE, which is a data mining system based on our approach, has been validated using the benchmark data sets in the machine learning community.

Introduction

In data mining, an association rule (Agrawal *et al.* 1996), which is a statement of a regularity in the form of a production rule, represents one of the most important classes of the discovered knowledge due to its generality. An association rule can be classified into two categories: a common sense rule, which is a description of a regularity for numerous objects, and an exception rule, which represents, for a relatively small number of objects, a different regularity from a common sense rule (Suzuki & Shimura 1996) (Suzuki 1996). An exception rule exhibits unexpectedness and is often useful. For instance, the rule "using a seat belt is risky for a child", which represents exceptions to the well known fact "using a seat belt is safe", exhibited unexpectedness when it was discovered from car accident data several years ago, and is still useful. Moreover, an exception rule is often beneficial since it differs from a common sense rule which is often a basis for people's daily activity. For instance, suppose a species of poisonous mushrooms some of which are exceptionally edible. The exact description of the exceptions is highly beneficial since it enables the exclusive possession of the edible mushrooms.

Since an exception rule holds for a relatively small number of examples, the distinction of a reliable rule from a coincidental pattern is one of the most important issues in discovering this type of knowledge. However, such distinction was left to the users in the previous discovery systems (Piatetsky-Shapiro & Matheus 1994) (Klösgen 1996) (Suzuki & Shimura 1996) (Suzuki 1996). The evaluation of confidence by the users, depending on their subjective judgement, is unreliable and uncertain in case the discovered rules are numerous. In order to circumvent these difficulties we propose a novel approach in which exception rules are discovered according to their confidence level based on the normal approximations of the multinomial distributions. This approach can be called as autonomous, since an exception rule is discovered using neither users' confidence evaluation nor domain knowledge.

Description of the Problem

Let a data set contains n examples each of which expressed by m discrete attributes. An event representing, in propositional form, a single value assignment to an attribute will be called an atom. We define an association rule as the production rule of which the premise is represented by a conjunction of atoms and the conclusion is a single atom. In this paper, we consider the problem of finding a set of rule pairs each of which consists of an exception rule associated with a common sense rule. A rule pair $r(\mu, \nu)$ is defined as a pair of association rules as follows:

$$r(\mu,\nu) \equiv \begin{cases} A_\mu & \rightarrow c \\ A_\mu \wedge B_\nu & \rightarrow c', \end{cases} \quad (1)$$

$$A_\mu \equiv a_1 \wedge a_2 \wedge \cdots \wedge a_\mu, \quad (2)$$

$$B_\nu \equiv b_1 \wedge b_2 \wedge \cdots \wedge b_\nu. \quad (3)$$

In learning from examples, simplicity and goodness-of-fit are considered as the most general criteria for evaluating the goodness of a hypothesis (Smyth & Goodman 1992). In case of an association rule $A_\mu \rightarrow c$, these two criteria correspond to $p(A_\mu)$ and $p(c|A_\mu)$ respectively (Smyth & Goodman 1992). Existing methods for evaluating the simplicity and the goodness-of-fit of an association 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 the two criteria, and the simultaneous approach such as (Mannila *et al.* 1994), which specifies two minimum thresholds for both criteria. We take the latter approach due to its generality, and specify thresholds for the simplicity and the goodness-of-fit of both rules. Here, in order to consider the confidence level of the rules, we do not employ the probabilities $\hat{p}(A_\mu)$, $\hat{p}(c|A_\mu)$, $\hat{p}(A_\mu, B_\nu)$ and $\hat{p}(c'|A_\mu, B_\nu)$ obtained by the point estimation from the data set. We obtain rule pairs of which their true probabilities $p(A_\mu)$, $p(c|A_\mu)$, $p(A_\mu, B_\nu)$ and $p(c'|A_\mu, B_\nu)$ are greater than or equal to their respective thresholds with a probability of $1 - \delta$.

$$\Pr\{p(A_\mu) \geq \theta_1^{\mathrm{S}}\} \geq 1 - \delta, \quad (4)$$

$$\Pr\{p(c|A_\mu) \geq \theta_1^{\mathrm{F}}\} \geq 1 - \delta, \quad (5)$$

$$\Pr\{p(A_\mu, B_\nu) \geq \theta_2^{\mathrm{S}}\} \geq 1 - \delta, \quad (6)$$

$$\Pr\{p(c'|A_\mu, B_\nu) \geq \theta_2^{\mathrm{F}}\} \geq 1 - \delta. \quad (7)$$

Consider the case in which the conditional probability $p(c'|B_\nu)$ is large. In such a case, the exception rule $A_\mu \wedge B_\nu \rightarrow c'$ can be easily guessed from $B_\nu \rightarrow c'$, which we call the reference rule, and is not considered as unexpected. Therefore, we add the following condition to obtain truly unexpected exception rules.

$$\Pr\{p(c'|B_\nu) \leq \theta_2^{\mathrm{I}}\} \geq 1 - \delta. \quad (8)$$

From the above discussions, the problem dealt in this paper can be described as discovering, from a data set, the rule pairs $r(\mu, \nu)$ which satisfy (4) ~ (8).

Evaluation of Reliability

In data mining, the Chernoff bound and the normal approximations of the binomial distributions are frequently used in assessing the reliability of a discovered association rule (Agrawal *et al.* 1996) (Chan & Wong 1991) (Mannila *et al.* 1994) (Siebes 1994). However, these methods are for estimating a single probability, and cannot be applied to our problem since (5), (7) and (8) contain conditional probabilities.

This shows that we should estimate the confidence region of the probabilities related to (4) ~ (8). First, atoms $D_1, D_2, \cdots$, and D_8 are defined as follows:

$$D_1 \equiv c \wedge A_\mu \wedge B_\nu, \quad (9)$$

$$D_2 \equiv c' \wedge A_\mu \wedge B_\nu, \quad (10)$$

$$D_3 \equiv (\overline{c \vee c'}) \wedge A_\mu \wedge B_\nu, \quad (11)$$

$$D_4 \equiv c \wedge A_\mu \wedge \overline{B_\nu}, \quad (12)$$

$$D_5 \equiv \overline{c} \wedge A_\mu \wedge \overline{B_\nu}, \quad (13)$$

$$D_6 \equiv \overline{c'} \wedge \overline{A_\mu} \wedge B_\nu, \quad (14)$$

$$D_7 \equiv c' \wedge \overline{A_\mu} \wedge B_\nu, \quad (15)$$

$$D_8 \equiv \overline{A_\mu} \wedge \overline{B_\nu}. \quad (16)$$

Let an atom D_i which satisfy $\hat{p}(D_i) \neq 0$ be $E_1, E_2, \cdots$, and E_{k+1} respectively. Since the atoms $E_1, E_2, \cdots$, and E_{k+1} are mutually exclusive, the numbers of their simultaneous occurrences $(x_1, x_2, \cdots, x_{k+1})$ are multinomially distributed. Let $u_1, u_2, \cdots$, and u_{k+1} be the respective numbers of the examples of the atoms $E_1, E_2, \cdots$, and E_{k+1} in the data set, ${}^t\boldsymbol{G}$ be the transposed matrix of the matrix $\boldsymbol{G}$, and

$$\vec{u} \equiv (u_1, u_2, \cdots, u_k), \quad (17)$$

$$\vec{x} \equiv (x_1, x_2, \cdots, x_k). \quad (18)$$

Assuming that n is enough large, the above multinomial distribution is approximated by the k-dimensional normal distribution of which the frequency function is

$$f(\vec{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{H}|^{1/2}} \cdot \exp\left\{-\frac{{}^t(\vec{x} - \vec{u})\boldsymbol{H}^{-1}(\vec{x} - \vec{u})}{2}\right\} \quad (19)$$

Here, $\boldsymbol{H}$ represents the covariance matrix given by

$$\boldsymbol{H} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1k} \\ h_{21} & h_{22} & \cdots & h_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ h_{k1} & h_{k2} & \cdots & h_{kk} \end{pmatrix} \quad (20)$$

$$h_{ij} = \begin{cases} u_i(n - u_i)/n & (i = j) \\ -u_i u_j/n & (i \neq j). \end{cases} \quad (21)$$

Consider a k-dimensional region covered by an ellipsoid

$$V_\delta : {}^t(\vec{x} - \vec{u})\boldsymbol{H}^{-1}(\vec{x} - \vec{u}) \leq \beta(\delta, k)^2 \quad (22)$$

which satisfy

$$\Pr(\vec{x} \in V_\delta) = 1 - \delta. \quad (23)$$

This ellipsoid V_δ corresponds to the $1 - \delta$ confidence region of $\vec{x}$. The positive number $\beta(\delta, k)$ can be calculated by the numerical integration using the fact that the volume of V_δ is $\pi^{k/2}\beta^k/\Gamma(k/2+1)/|\boldsymbol{H}^{-1}|^{1/2}$ (Cramér 1966). From (9) ~ (15), (4) ~ (8) represent the problem of judging whether the following constraints always hold in (22), where $x(D)$ represents the number of occurrence of the atom D in the data set.

$$\frac{\sum_{i=1}^{5} x(D_i)}{n} \geq \theta_1^{\mathrm{S}}, \quad (24)$$

$$\frac{x(D_1) + x(D_4)}{\sum_{i=1}^{5} x(D_i)} \geq \theta_1^{\mathrm{F}}, \quad (25)$$

$$\frac{\sum_{i=1}^{3} x(D_i)}{n} \geq \theta_2^{\mathrm{S}}, \quad (26)$$

$$\frac{x(D_2)}{\sum_{i=1}^{3} x(D_i)} \geq \theta_2^{\mathrm{F}}, \qquad (27)$$

$$\frac{x(D_2)+x(D_7)}{\sum_{i=1}^{3} x(D_i)+\sum_{i=6}^{7} x(D_i)} \leq \theta_2^{\mathrm{I}}. \qquad (28)$$

Here, since each expression on the left hand sides in (24) $\sim$ (28) being a constant represents a plane, the maximum and the minimum of each expression occurs at the extremes on the hypersurface of the ellipsoid. Applying Lagrange's multiplier method to the ellipsoid of (22) and to each expression, we obtain the following results, where $\beta = \beta(\delta, k)$.

$$\left(1-\beta\sqrt{\frac{1-\hat{p}(A_\mu)}{n\hat{p}(A_\mu)}}\right)\hat{p}(A_\mu) \geq \theta_1^{\mathrm{S}}, \qquad (29)$$

$$\left(1-\beta\sqrt{\frac{\hat{p}(\overline{c},A_\mu)}{\hat{p}(c,A_\mu)\{(n+\beta^2)\hat{p}(A_\mu)-\beta^2\}}}\right) \cdot\hat{p}(c|A_\mu) \geq \theta_1^{\mathrm{F}}, \qquad (30)$$

$$\left(1-\beta\sqrt{\frac{1-\hat{p}(A_\mu,B_\nu)}{n\hat{p}(A_\mu,B_\nu)}}\right)\hat{p}(A_\mu,B_\nu) \geq \theta_2^{\mathrm{S}}, \qquad (31)$$

$$\left(1-\beta\sqrt{\frac{\hat{p}(\overline{c'},A_\mu,B_\nu)}{\hat{p}(c',A_\mu,B_\nu)\{(n+\beta^2)\hat{p}(A_\mu,B_\nu)-\beta^2\}}}\right) \cdot\hat{p}(c'|A_\mu,B_\nu) \geq \theta_2^{\mathrm{F}}, \qquad (32)$$

$$\left(1+\beta\sqrt{\frac{\hat{p}(\overline{c'},B_\nu)}{\hat{p}(c',B_\nu)\{(n+\beta^2)\hat{p}(B_\nu)-\beta^2\}}}\right) \cdot\hat{p}(c'|B_\nu) \leq \theta_2^{\mathrm{I}}. \qquad (33)$$

Therefore, our algorithm obtains rule pairs which satisfy all the constraints in (29) $\sim$ (33).

Discovery Algorithm

In our algorithm, a discovery task is viewed as a search problem, in which a node of a search tree represents a rule pair $r(\mu, \nu)$. Let $\mu = 0$ and $\nu = 0$ represent the state in which the premises of a rule pair $r(\mu, \nu)$ contain no a_i or no b_i respectively, then we define that $\mu = \nu = 0$ holds in a node of depth 1, and as the depth increases by 1, an atom is added to the premise of the exception rule or the common sense rule. A node of depth 2 is assumed to satisfy $\mu = 1$, $\nu = 0$, and a node of depth l (≥ 3), $\mu + \nu = l - 1$ ($\mu \geq 1$, $\nu \geq 0$).

A depth-first search method is employed to traverse this tree, and the maximum value M of μ and ν is given by the user. To improve the search efficiency, the nodes which satisfy at least one of the stopping criteria (34) $\sim$ (38) in theorem 1 are not expanded without altering the algorithm's output.

Theorem 1 *Let the rule pair of the current node be $r(\mu', \nu')$. If the rule pair satisfy one of (34) $\sim$ (38), no rule pairs $r(\mu, \nu)$ of the descendant nodes satisfy (29) $\sim$ (33).*

$$\left(1-\beta\sqrt{\frac{1-\hat{p}(A_{\mu'})}{n\hat{p}(A_{\mu'})}}\right)\hat{p}(A_{\mu'}) < \theta_1^{\mathrm{S}}, \qquad (34)$$

$$\left(1-\beta\sqrt{\frac{1-\hat{p}(A_{\mu'})}{n\hat{p}(A_{\mu'})}}\right)\hat{p}(c,A_{\mu'}) < \theta_1^{\mathrm{S}}\theta_1^{\mathrm{F}}, \qquad (35)$$

$$\left(1-\beta\sqrt{\frac{1-\hat{p}(A_{\mu'},B_{\nu'})}{n\hat{p}(A_{\mu'},B_{\nu'})}}\right)\hat{p}(A_{\mu'},B_{\nu'}) < \theta_2^{\mathrm{S}}, \qquad (36)$$

$$\left(1-\beta\sqrt{\frac{1-\hat{p}(A_{\mu'},B_{\nu'})}{n\hat{p}(A_{\mu'},B_{\nu'})}}\right)\hat{p}(c',A_{\mu'},B_{\nu'}) < \theta_2^{\mathrm{S}}\theta_2^{\mathrm{F}}, \qquad (37)$$

$$\left(1-\beta\sqrt{\frac{1-\hat{p}(A_{\mu'},B_{\nu'})}{n\hat{p}(A_{\mu'},B_{\nu'})}}\right)\hat{p}(B_{\nu'}) < \frac{\theta_2^{\mathrm{S}}\theta_2^{\mathrm{F}}}{\theta_2^{\mathrm{I}}}. \qquad (38)$$

Proof Assume a rule pair $r(\mu, \nu)$ satisfy (29) $\sim$ (33). First, the function $(1-\beta\sqrt{(1-x)/n/x})$ increases monotonously for $x > 0$ since $n, \beta > 0$. Here, (29) and $\hat{p}(A_{\mu'}) \geq \hat{p}(A_\mu)$ gives (the left hand side of (34)) $\geq (1-\beta\sqrt{(1-\hat{p}(A_\mu))/n/\hat{p}(A_\mu)})\ \hat{p}(A_\mu) \geq \theta_1^{\mathrm{S}}$, which contradict to (34). Contradiction to (35) $\sim$ (38) can be derived in a similar way. □

Application to Data Sets

The proposed method was implemented as PEDRE (Probabilistic Estimation-based Data mining system for Reliable Exceptions) and was tested with data sets from several domains. The results were quite successful. Here, we show the results using the mushroom data set from the UCI Repository.

The mushroom data set includes 22 descriptions and the edibility class of 8,124 mushrooms, each attribute having 2 to 12 values. Table 1 shows the rule pairs discovered by PEDRE, where the edibility class is the only attribute allowed in the conclusions and the parameters were set to $M = 3$, $\delta = 0.1$, $\theta_1^{\mathrm{S}} = 0.2$, $\theta_2^{\mathrm{S}} = 0.05$, $\theta_1^{\mathrm{F}} = 0.7$, $\theta_2^{\mathrm{F}} = 1.0$ and $\theta_2^{\mathrm{I}} = 0.5$.

The discovered rule pairs in the table show very interesting exceptions. According to the first rule pair, 72.9 % of the 1,828 mushrooms whose "bruises" is "f", "g-size" is "b" and "stalk-shape" is "e" are poisonous but 100 % of them are actually edible if the "stalk-root is "?". Each rule holds, with the 90 % confidence level, for at least 1,734 or 415 mushrooms with the conditional probability more than 70.3 % or 100 % respectively. Note that, from the reference rule, only 29.0 % of the mushrooms whose "stalk-root" is "?" are edible, which holds for at most 31.8 % with the 90 % confidence level. This shows that the discovered exception rule is truly unexpected.

The maximum value M of μ and ν should be large enough so that PEDRE investigates rule pairs whose premises have sufficient numbers of atoms. Figure 1 shows, in a logarithmic scale, the number of nodes searched by PEDRE with/without the stopping criteria (white/black bars). The parameters, except for M, were settled the same as in the previous experiments. The experiment lasted for about nine days with a Pentium Pro 200MHz equipped personal computer when

Table 1: The rule pairs with their associated reference rules discovered by PEDRE from the mushroom data set, where the edibility class is the only attribute allowed in the conclusions.

No.	common sense exception reference rule	(min) A (min)	$c \wedge A$	(min) $\hat{p}(c \mid A)$ (min) (max)
1	bruises=f, g-size=b, stalk-shape=e → class=p	1,828 (1,734)	1,332	0.729 (0.703)
	C, stalk-root=? → class=e	480 (415)	480	1.000 (1.000)
	stalk-root=? → class=e	2,480	720	0.290 (0.318)
2	g-attachment=f, stalk-root=? → class=p	2,288 (2,187)	1,760	0.769 (0.747)
	C, g-size=b, stalk-shape=e, veil-color=w → class=e	480 (415)	480	1.000 (1.000)
	g-size=b, stalk-shape=e, veil-color=w → class=e	2,636	1.232	0.467 (0.497)
3	stalk-root=?, sp-color=w → class=p	2,240 (2,139)	1,760	0.786 (0.764)
	C, g-attachment=f, g-size=b, stalk-shape=e → class=e	480 (421)	480	1.000 (1.000)
	g-attachment=f, g-size=b, stalk-shape=e → class=e	2,618	1,232	0.471 (0.498)
4	stalk-root=?, sp-color=w → class=p	2,240 (2,139)	1,760	0.786 (0.764)
	C, g-size=b, stalk-shape=e, veil-color=w → class=e	480 (421)	480	1.000 (1.000)
	g-size=b, stalk-shape=e, veil-color=w → class=e	2,636	1,232	0.467 (0.495)

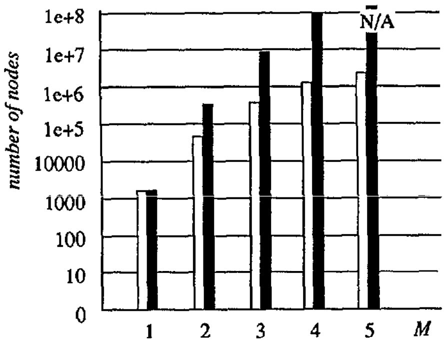

Figure 1: Performance of PEDRE with/without the stopping criteria (white/black bars).

the number of nodes was about 10^8. The figure shows that the stopping criteria are effective in improving the time efficiency. For instance, in case of $M = 4$, the criteria achieved 70 times of speed-up.

Conclusion

This paper has described an autonomous approach for finding reliable exception rules using the newly proposed probabilistic estimation method. The approach does not depend on the subjective evaluation of the reliability by humans, and discovers only the exception rules which satisfy the user-specified conditions confidentially. Consequently, our data mining system PEDRE is immune from the problem of misjudging an unreliable pattern as knowledge inherent in the previous approaches. Moreover, we have derived the stopping criteria to improve search efficiency without altering the discovery results. Experimental results show that our system is promising for the efficient discovery of reliable exception rules.

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.

Cramér, H. 1966. *Mathematical Methods of Statistics*, Princeton Univ. Press.

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.

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.

Mannila, H., Toivonen, H. and Verkamo, A., I. 1994. Efficient Algorithms for Discovering Association Rules, *AAAI-94 Workshop on Knowledge Discovery in Databases*, pp. 181–192.

Piatetsky-Shapiro, G. and Matheus, C., J. 1994. The Interestingness of Deviations, *AAAI-94 Workshop on Knowledge Discovery in Databases*, pp. 25–36.

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.

Suzuki, E. and Shimura, M. 1996. Exceptional Knowledge Discovery in Databases based on Information Theory, *Proc. of KDD-96*, pp. 275–278.

Suzuki, E. 1996. Discovering Unexpected Exceptions: A Stochastic Approach, *Proc. of RSFD-96*, pp. 225–232.