# Exceptional Knowledge Discovery in Databases based on Information Theory

**Einoshin Suzuki**
Division of Electrical and
Computer Engineering, Faculty of Engineering,
Yokohama National University,
156, Tokiwadai, Hodogaya,
Yokohama, 240, Japan.
suzuki@dnj.ynu.ac.jp

**Masamichi Shimura**
Dept. of Computer Science, Graduate School
of Information Science and Engineering,
Tokyo Institute of Technology,
2-12 Ohokayama, Meguro,
Tokyo, 152, Japan.
shimura@cs.titech.ac.jp

## Abstract

This paper presents an algorithm for discovering exceptional knowledge from databases. Exceptional knowledge, which is defined as an exception to a general fact, exhibits unexpectedness and is sometimes extremely useful in spite of its obscurity Previous discovery approaches for this type of knowledge employ either background knowledge or domain-specific criteria for evaluating the possible usefulness, i.e. the interestingness of the knowledge extracted from a database It has been pointed out, however, that these approaches are prone to overlook useful knowledge.

In order to circumvent these difficulties, we propose an information-theoretic approach in which we obtain exceptional knowledge associated with general knowledge in the form of a rule pair using a depth-first search method. The product of the ACEs (Average Compressed Entropies) of the rule pair is introduced as the criterion for evaluating the interestingness of exceptional knowledge. The inefficiency of depth-first search is alleviated by a branch-and-bound method, which exploits the upper-bound for the product of the ACEs. MEPRO, which is a knowledge discovery system based on our approach, has been validated using the benchmark databases in the machine learning community.

## Introduction

Recently, databases have grown remarkably both in size and in number. Consequently, increasing attention has been paid to the automatic extraction of knowledge from them, i.e. Knowledge Discovery in Databases (KDD) (Frawley et al. 1991). In KDD, the discovered knowledge can be classified into two categories: general knowledge, which holds for numerous examples, and exceptional knowledge, which represents an exception to general knowledge. For instance, "a jumbo jet is a safe means of transportation" is a piece of general knowledge, while "a jumbo jet which does not satisfy condition X is a dangerous means of transportation" is a piece of exceptional knowledge.

Although exceptional knowledge is often overlooked, it represents a different fact from general knowledge and can be extremely useful. Among the approaches for discovering such useful exceptional knowledge, well-known systems include EXPLORA (Hoschka & Klösgen 1991), which employs background knowledge for evaluating the knowledge extracted from a database, and KEFIR (Piatetsky-Shapiro & Matheus 1994) which employs domain-specific criteria.

Since a huge amount of knowledge can be embedded in a database, the discrimination of possibly useful or *interesting* knowledge is one of the most important topics in the KDD community. Especially in the case of discovering exceptional knowledge hidden in databases, the most crucial problem is to define appropriate criteria for evaluating the interestingness of the extracted knowledge (Piatetsky-Shapiro & Matheus 1994). As described above, previous criteria either require background knowledge or are inherently domain-specific. However, the use of such background knowledge can in fact hinder the discovery of interesting knowledge (Frawley et al. 1991). Furthermore, it is difficult to find such criteria in some domains.

In order to circumvent these difficulties, we propose a novel approach which employs neither background knowledge nor domain-specific criteria.

## Rule Pair

Let an example $e_i$ be a description about an object stored in a database in the form of a record, then a database 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.

Consider the problem of finding $K$ pieces of knowledge $\{r_1, r_2, \cdots, r_K\}$. We can view a piece of knowledge $r_i$ to be discovered from a database as represented by a **rule pair** $r(\mu, \nu)$:

$$r(\mu,\nu) \equiv \begin{cases} Y_\mu & \rightarrow \quad x \\ Y_\mu \wedge Z_\nu & \rightarrow \quad x' \end{cases} \qquad (1)$$

where $Y_\mu = y_1 \wedge y_2 \wedge \cdots \wedge y_\mu$, $Z_\nu = z_1 \wedge z_2 \wedge \cdots \wedge z_\nu$. Here, $x$, $x'$, $y_i$, and $z_i$ are **atoms**, each of which is an event representing, in propositional form, a single value assignment to an attribute. Atoms $x$ and $x'$ have the same attribute but different values.

Since an if-then rule in equation (1) represents correlation or causality between its premise and conclusion, every rule pair is assumed to satisfy the following inequalities

$$\mathrm{p}(x|Y_\mu) > \mathrm{p}(x),\ \mathrm{p}(x'|Y_\mu \wedge Z_\nu) > \mathrm{p}(x'), \qquad (2)$$

where $Y_\mu = y_1 \wedge y_2 \wedge \cdots \wedge y_\mu$, $Z_\nu = z_1 \wedge z_2 \wedge \cdots \wedge z_\nu$.

A rule pair of equation (1) can be interpreted as "if $Y_\mu$ then $x$, but if $Y_\mu$ and $Z_\nu$ then $x'$". Since the event $Y_\mu$ occurs more frequently than $Y_\mu \wedge Z_\nu$, the rule $Y_\mu \rightarrow x$ represents a piece of general knowledge, and is thus called a **general rule**. On the other hand, the rule $Y_\mu \wedge Z_\nu \rightarrow x'$ represents the associated piece of exceptional knowledge, and is thus called an **exceptional rule**.

## ACEP: average compressed entropy product

From the point of view of information theory, the rule $Y_\mu \rightarrow x$ indicates that each of the $n\mathrm{p}(x, Y_\mu)$ examples has a code length of $-\log_2 \mathrm{p}(x|Y_\mu)$, which is smaller than the original length, $-\log_2 \mathrm{p}(x)$, and each of the $n\mathrm{p}(\overline{x}, Y_\mu)$ examples, a code length of $-\log_2 \mathrm{p}(\overline{x}|Y_\mu)$ instead of $-\log_2 \mathrm{p}(\overline{x})$. The use of the reduced code length, or the compressed entropy, allows us to measure the information content of an if-then rule quantitatively. The entropy per example compressed by the rule, $\mathrm{ACE}(x, Y_\mu)$, which is called the **Average Compressed Entropy (ACE)**, is given as follows.

$$\begin{aligned}\mathrm{ACE}(x, Y_\mu) \equiv\ & \mathrm{p}(x, Y_\mu) \log_2 \frac{\mathrm{p}(x|Y_\mu)}{\mathrm{p}(x)} \\ & + \mathrm{p}(\overline{x}, Y_\mu) \log_2 \frac{\mathrm{p}(\overline{x}|Y_\mu)}{\mathrm{p}(\overline{x})} \end{aligned} \qquad (3)$$

A rule of large information content is useful in the sense that it gives a compact representation for data stored in a database. Since ACE is a measure for the information content of a rule, it can be considered as a function for the usefulness of the rule. Therefore, the interestingness of a rule extracted from a database is evaluated by its ACE. Since ACE increases monotonously as $\mathrm{p}(x)$ decreases, as $\mathrm{p}(x|Y_\mu)$ increases, or as $\mathrm{p}(x, Y_\mu)$ increases, it can be also viewed as a unified criterion for evaluating the unexpectedness, stability, and generality of a rule. Actually, Smyth (Smyth & Goodman 1991) showed various desirable properties of ACE as a criterion for evaluating the interestingness of an if-then rule extracted from a database.

However, an exceptional rule $Y_\mu \wedge Z_\nu \rightarrow x'$, whose ACE is high, may not be "interesting" if the ACE of the associated general rule $Y_\mu \rightarrow x$ is extremely low. That is, the interestingness of an exceptional rule depends not only on its ACE but also on the ACE of the associated general rule. It is reasonable therefore to represent the interestingness of an exceptional rule in terms of both the above ACEs. Note that interestingness should increase as the ACEs increase, and decrease when they decrease. Among the functions which satisfy these requirements, the add-sum $\mathrm{ACE}(x, Y_\mu) + \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ and product $\mathrm{ACE}(x, Y_\mu) \cdot \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ are considered as the simplest formulations.

Let us analyze the appropriateness of these functions as evaluation criteria for the interestingness of exceptional knowledge. Consider the case in which the maximums of both ACEs for constant $x$ and $x'$ occur, since we are interested in the rule pairs whose ACEs are close to their respective maximum values. From equation (1) and (2), the following equations (4)~(6) are obtained.

$$\begin{aligned}&\mathrm{ACE}(x, Y_\mu) \\ &= (a+b)\log_2\left(\frac{a+b}{a+b+c+d+e+f}\frac{1}{\mathrm{p}(x)}\right) \\ &\quad +(c+d+e+f) \\ &\quad \cdot \log_2\left(\frac{c+d+e+f}{a+b+c+d+e+f}\frac{1}{\mathrm{p}(\overline{x})}\right) \end{aligned} \qquad (4)$$

$$\begin{aligned}&\mathrm{ACE}(x', Y_\mu \wedge Z_\nu) \\ &= c\ \log_2\left(\frac{c}{a+c+e}\frac{1}{\mathrm{p}(x')}\right) \\ &\quad +(a+e)\log_2\left(\frac{a+e}{a+c+e}\frac{1}{\mathrm{p}(\overline{x'})}\right) \end{aligned} \qquad (5)$$

$$\frac{a+b}{a+b+c+d+e+f} > \mathrm{p}(x),\ \frac{c}{a+c+e} > \mathrm{p}(x'), \qquad (6)$$

where $a = \mathrm{p}(x, Y_\mu, Z_\nu)$, $b = \mathrm{p}(x, Y_\mu, \overline{Z_\nu})$, $c = \mathrm{p}(x', Y_\mu, Z_\nu)$, $d = \mathrm{p}(x', Y_\mu, \overline{Z_\nu})$, $e = \mathrm{p}(\overline{x \vee x'}, Y_\mu, Z_\nu)$, and $f = \mathrm{p}(\overline{x \vee x'}, Y_\mu, \overline{Z_\nu})$. Note that the following inequalities hold for these variables.

$$\begin{aligned}&a, b, c, d, e, f \geq 0,\ a+b \leq \mathrm{p}(x),\ c+d \leq \mathrm{p}(x'), \\ &\qquad e+f \leq \mathrm{p}(\overline{x \vee x'}) \end{aligned} \qquad (7)$$

A simple calculation shows that both $\mathrm{ACE}(x, Y_\mu)$ and $\mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$ are maximized when $b = \mathrm{p}(x)$ and $a = d = e = f = 0$. Let $U$ and $V$ be the maximum value of $\mathrm{ACE}(x, Y_\mu)$ and $\mathrm{ACE}(x', Y_\mu \wedge Z_\nu)$, respectively. From equation (4) and (5), we obtain

$$\begin{aligned} U &= \mathrm{p}(x)\log_2\frac{1}{\mathrm{p}(x)+c} + c\ \log_2\left(\frac{c}{\mathrm{p}(x)+c}\frac{1}{\mathrm{p}(\overline{x})}\right), \\ V &= c\ \log_2\frac{1}{\mathrm{p}(x')}, \end{aligned} \qquad (8)$$

where from equation (6) and (7),

$$0 \leq c \leq \mathrm{p}(x'),\ c < \mathrm{p}(\overline{x}). \qquad (9)$$

A simple calculation shows that the maximum of the add-sum $U+V$ for constant $x$ and $x'$ occurs when either $\mathrm{ACE}(x, Y_\mu) = 0$ or $\mathrm{ACE}(x', Y_\mu \wedge Z_\nu) \approx 0$. The add-sum function, therefore, is inappropriate as a criterion for interestingness since its maximum value is dominated by one of the ACEs. Actually, using this function to

determine interestingness in some databases will yield results which contain useless knowledge.

On the other hand, the product $U \cdot V$ can be proved to possess no such shortcomings, and thus the product of the ACEs, **Average Compressed Entropy Product (ACEP)**, can be considered as one of the simplest functions appropriate for evaluating the interestingness of an exceptional rule. Therefore, the interestingness function is defined by the ACEP, $\mathrm{ACEP}(x, Y_\mu, x', Z_\nu)$.

$$\begin{aligned}&\mathrm{ACEP}(x, Y_\mu, x', Z_\nu)\\ &\quad \equiv \mathrm{ACE}(x, Y_\mu) \cdot \mathrm{ACE}(x', Y_\mu \wedge Z_\nu) \qquad (10)\end{aligned}$$

## Discovery Algorithm

Consider a discovery algorithm which generates $K$ rule pairs, where $K$ is a user-specified parameter. The generated rule pairs are the $K$ most interesting ones in the database as defined by ACEP. In the algorithm, a discovery task is viewed as a search problem, in which a node of a search tree represents a rule pair $r(\mu, \nu)$ of equation (1). A depth-first search method with maximum depth $D$ is employed to traverse this tree.

Let $\mu = 0$ and $\nu = 0$ represent the state in which the premises of a rule pair $r(\mu, \nu)$ contain no $y_i$ or no $z_i$ 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 general or exceptional rule. A node of depth 2 is assumed to satisfy $\mu = 1$ and $\nu = 0$; a node of depth 3, $\mu = \nu = 1$; and a node of depth $l$ ($\geq 4$), $\mu + \nu = l - 1$ ($\mu, \nu \geq 1$). Therefore, a descendant node represents a rule pair $r(\mu', \nu')$ where $\mu' \geq \mu$ and $\nu' \geq \nu$. According to the following theorem, an upper-bound exists for the ACEP of this rule pair.

**Theorem 1** *Let* $\mathrm{H}(\alpha) \equiv [\alpha/\{(1+\alpha)\mathrm{p}(\overline{x})\}]^{2\alpha}/\{(1+\alpha)\mathrm{p}(x)\}$, $\alpha_1$ *and* $\alpha_2$ *satisfy* $\mathrm{H}(\alpha_1) > 1 > \mathrm{H}(\alpha_2)$, *and* $ACEP = \mathrm{ACEP}(x, Y_{\mu'}, x', Z_{\nu'})$. *If* $\mathrm{H}(\mathrm{p}(x', Y_\mu, Z_\nu)/\mathrm{p}(x, Y_\mu)) < 1$ *then,*

$$\begin{aligned}ACEP \;<\; & \alpha_2 \mathrm{p}(x, Y_\mu)^2 \left\{ \log_2 \left( \frac{1}{1+\alpha_1} \frac{1}{\mathrm{p}(x)} \right) + \alpha_1 \right.\\ & \left. \cdot \log_2 \left( \frac{\alpha_1}{1+\alpha_1} \frac{1}{\mathrm{p}(\overline{x})} \right) \right\} \log_2 \frac{1}{\mathrm{p}(x')} \;, \qquad (11)\end{aligned}$$

*else*

$$\begin{aligned}ACEP \;\leq\; & \left\{ \mathrm{p}(x, Y_\mu) \log_2 \left( \frac{\mathrm{p}(x, Y_\mu)}{\mathrm{p}(x, Y_\mu) + \mathrm{p}(x', Y_\mu, Z_\nu)} \right.\right.\\ & \left. \cdot \frac{1}{\mathrm{p}(x)} \right) + \mathrm{p}(x', Y_\mu, Z_\nu)\\ & \cdot \log_2 \left( \frac{\mathrm{p}(x', Y_\mu, Z_\nu)}{\mathrm{p}(x, Y_\mu) + \mathrm{p}(x', Y_\mu, Z_\nu)} \right.\\ & \left.\left. \cdot \frac{1}{\mathrm{p}(\overline{x})} \right) \right\} \mathrm{p}(x', Y_\mu, Z_\nu) \log_2 \frac{1}{\mathrm{p}(x')} \;. \qquad (12)\end{aligned}$$

In other words, if the upper-bound for the current node is lower than $ACEP_K$ (the $K$th highest ACEP of the discovered rule pairs), no rule pair exists whose ACEP is higher than $ACEP_K$ in its descendant nodes. This law tells us that there is no need to expand such descendant nodes and that these nodes can be safely cut off. To alleviate the inevitable inefficiency of depth-first search, a Branch-and-Bound Method (BBM) based on $ACEP_K$ is employed in our approach.

## Application to Databases

The proposed method was implemented as MEPRO (database Miner based on average compressed Entropy PROduct criterion), and tested with data sets from several domains, including the voting records database (Murphy & Aha 1994).

The voting records database consists of voting records in a 1984 session of Congress, each piece of data corresponding to a particular politician. The class variable is party affiliation (republican or democrat), and the other 16 attributes are yes/no votes on particular motions such as Contra-aid and budget cuts. Table 1 shows the results of asking MEPRO for the 10 best rule pairs, where the maximum search depth $D$ is restricted to 8. A comma and $C$ in the table represent conjunction and the premise of the general rule respectively, while the columns $xY$ and $Y$ are the respective actual number of occurrences of the event $x \wedge Y$ (conclusion and premise) and $Y$ (premise).

From table 1, we note that interesting exceptional knowledge emerges, confirming that the system is adequate for the task. According to the second rule pair, 91 % of the 253 congressmen who voted "yes" to "adoption" were democrats. However, 17 of these (who voted "yes" to "physician" and "satellite" in addition to "adoption") were republicans. It is found, from this rule pair, that even republicans vote "yes" to "adoption". The premise of this exceptional rule, which can be viewed as giving a partial definition of these republicans, is highly interesting.

The maximum depth should be large enough so that MEPRO investigates rule pairs whose premises have sufficient numbers of atoms. However, in depth-first search, the number of rule pairs grows exponentially as the depth increases. In this section, we show experimental evidence which suggests that BBM is quite effective in alleviating such inefficiency.

Figure 1 shows a plot of the ratio of the number of nodes pruned by BBM to the total number of nodes visited by depth-first search with depth $D$. The database chosen for this evaluation was the "voting" database. The system was run with six different values of $D$ (3, 4, $\cdots$, 8) and three values of $K$ (10, 50, 100). Note that the ratio decreases as $K$ increases; actually it is 0 if $K$ is equal to or greater than the number of nodes within depth $D$. The figure shows that BBM is more effective with a larger depth, e.g. it reduces by more than 80 % of the number of nodes searched when $D = 8$. This is especially important since we must go deeper in the tree to obtain useful exceptional knowledge.

| Rank | Rule pair | $p(x\|Y)$ | $p(x)$ | $xY$ | $Y$ | $ACE$ | $ACEP$ |
|---|---|---|---|---|---|---|---|
| 1 | adoption=yes → physician=no | 0.87 | 0.57 | 219 | 253 | 0.175 | 0.0115 |
| | $\mathcal{C}$, party=rep → physician=yes | 1.00 | 0.41 | 22 | 22 | 0.066 | |
| 2 | adoption=yes → party=demo | 0.91 | 0.61 | 231 | 253 | 0.195 | 0.0105 |
| | $\mathcal{C}$, physician=yes, satellite=yes → party=rep | 1.00 | 0.39 | 17 | 17 | 0.054 | |
| 3 | satellite=yes → physician=no | 0.82 | 0.57 | 197 | 239 | 0.118 | 0.0104 |
| | $\mathcal{C}$, party=rep → physician=yes | 0.95 | 0.41 | 37 | 39 | 0.088 | |
| 4 | party=demo → salvador=no | 0.75 | 0.48 | 200 | 267 | 0.135 | 0.0101 |
| | $\mathcal{C}$,nicaraguan=no,crime=yes → salvador=yes | 0.97 | 0.49 | 37 | 38 | 0.075 | |
| 5 | crime=yes → party=rep | 0.64 | 0.39 | 158 | 248 | 0.105 | 0.0101 |
| | $\mathcal{C}$, physician=no → party=demo | 0.97 | 0.61 | 74 | 76 | 0.095 | |
| 6 | adoption=yes → party=demo | 0.91 | 0.61 | 231 | 253 | 0.195 | 0.0099 |
| | $\mathcal{C}$, physician=yes, synfuels=no, south-africa=yes → party=rep | 1.00 | 0.39 | 16 | 16 | 0.050 | |
| 7 | salvador=yes → party=rep | 0.74 | 0.39 | 157 | 212 | 0.182 | 0.0098 |
| | $\mathcal{C}$, physician=no → party=demo | 0.98 | 0.61 | 41 | 42 | 0.054 | |
| 8 | crime=yes → salvador=yes | 0.78 | 0.49 | 194 | 248 | 0.151 | 0.0097 |
| | $\mathcal{C}$,physician=no, satellite=yes, nicaraguan=yes → salvador=no | 0.95 | 0.48 | 35 | 37 | 0.064 | |
| 9 | nicaraguan=yes → party=demo | 0.90 | 0.61 | 218 | 242 | 0.169 | 0.0096 |
| | $\mathcal{C}$, physician=yes, synfuels=no → party=rep | 1.00 | 0.39 | 18 | 18 | 0.057 | |
| 10 | satellite=yes → party=demo | 0.84 | 0.61 | 200 | 239 | 0.094 | 0.0095 |
| | $\mathcal{C}$, physician=yes, salvador=yes → party=rep | 1.00 | 0.39 | 32 | 32 | 0.100 | |

Table 1: The 10 best rule pairs from the voting records database.

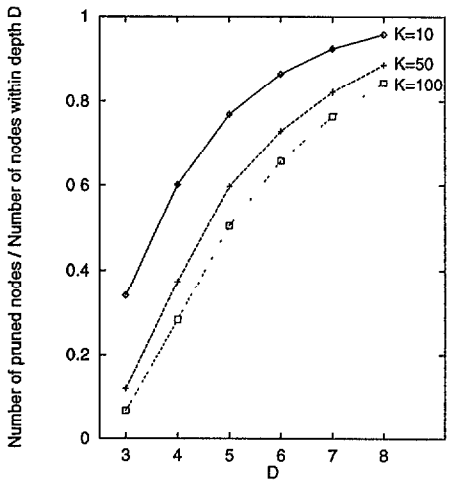


Figure 1: Performance of BBM with varying depth $D$ and number of target rule pairs $K$.

## Conclusion

This paper has described an approach for finding exceptional knowledge using the criterion ACEP (Average Compressed Entropy Product), which requires neither pre-supplied background knowledge nor domain-specific criteria. Consequently, our KDD system MEPRO is immune from the problem of overlooking useful knowledge inherent in the previous approaches which employ either background knowledge or domain-specific criteria. Moreover, we have derived the upper-bound for ACEP and used this in a BBM (Branch-and-Bound Method) to improve search efficiency without altering the discovery results.

Our MEPRO system has been applied to several benchmark databases in the machine learning community. Experimental results show that our system is promising for the efficient discovery of interesting exceptional knowledge. MEPRO is effective in exceptional knowledge discovery in databases where it is difficult to obtain background knowledge a priori. Moreover, it would discover unknown and useful exceptional knowledge in databases where such knowledge is left undiscovered due to the unpredictable misuse of user-supplied background knowledge.

## References


Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. 1991. Knowledge Discovery in Databases: An Overview. In Knowledge Discovery in Databases, 1-27. Piatetsky-Shapiro, G., and Frawley, W. J. (eds). AAAI Press/ The MIT Press.

Hoschka, P., and Klösgen, W. 1991. A Support System For Interpreting Statistical Data. In Knowledge Discovery in Databases, 325-345. Piatetsky-Shapiro, G., and Frawley, W. J. (eds). AAAI Press/ The MIT Press.

Murphy, P. M., and Aha, D. W. 1994. UCI Repository of machine learning databases. http://www.ics.uci.edu/~mlearn/ MLRepository.html. Dept. of Information and Computer Science, University of California.

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

Smyth, P., and Goodman, R. M. 1991. Rule Induction Using Information Theory. In Knowledge Discovery in Databases, 159-176. Piatetsky-Shapiro, G. and Frawley, W. J. (eds). AAAI Press/ The MIT Press.