Discovering Unexpected Exceptions: A Stochastic Approach

Einoshin Suzuki suzuki@dnj.ynu.ac.jp

Division of Electrical and Computer Engineering, Faculty of Engineering,
Yokohama National University,
156, Tokiwadai, Hodogaya, Yokohama, 240, Japan.

Abstract: *This paper presents an algorithm for discovering exceptions from databases. An exception to a piece of common sense knowledge exhibits unexpectedness and is sometimes extremely useful in spite of its obscurity. Most of the previous discovery approaches for such exceptions 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 have proposed MEPRO based on an information-theoretic approach, and succeeded in discovering interesting exceptions. However, some of the exceptions discovered by MEPRO showed little unexpectedness due to the lack of the attention to this characteristic in the system.*
In this paper, we improve MEPRO for discovering unexpected exceptions, and propose it as MEPROUX. A general view of discovery as search is also given in order to compare the two systems in terms of a single framework. In this view, MEPRO and MEPROUX deal with the same discovery problem of finding a user-specified number of rule pairs each of which consists of an exception associated with a piece of common sense knowledge. The search strategies of the two systems are also identical in that they are depth-first search with a branch-and-bound method. The main difference of the systems lies in the evaluation criteria employed. Although the two criteria are defined by the information content of both knowledge in a rule pair, several stochastic constraints are introduced in MEPROUX for removing exceptions which can be easily predicted from the relationships inherent in the domain. The effectiveness and the efficiency of MEPROUX have been validated using several benchmark data sets in the machine learning community.

Keywords: Knowledge Discovery in Databases, Data Mining Method, Exception, Unexpectedness.

1 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) [3, 5]. One of the important goals of KDD is the discovery of common sense knowledge, which is well motivated by various useful applications such as the automatic development of a knowledge-base from a database. Various approaches [1, 8, 11, 13] have been proposed for discovering common sense knowledge. Another important goals of KDD is the discovery of unknown and useful exceptions [3]. Although an exception is often overlooked, it represents a different fact from common sense knowledge and can be extremely useful. Among the approaches for discovering such knowledge, well-known systems include EXPLORA [4], KEFIR [9] and MEPRO [12].

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 exceptions hidden in databases, the most crucial problem is to define appropriate criteria for evaluating the interestingness of the extracted knowledge [9]. In EXPLORA, a priori given background knowledge is employed to define the criteria, while KEFIR's criteria is essentially domain-specific. Appropriate use of background knowledge in the criteria or appropriate use of domain-specific criteria can enhance both the effectiveness and the efficiency of a KDD system. However, the use of such background knowledge can also hinder the discovery of interesting knowledge [3]. Furthermore, it is difficult to find appropriate criteria in some domains.

MEPRO, which employs neither background knowledge nor domain-specific criteria, has been proposed to circumvent these difficulties. While the previous systems can be regarded as exploiting user-supplied common sense knowledge in discovering exceptions, MEPRO extracts exceptions associated with their pieces of common sense knowledge. The extraction is based on a novel criterion for interestingness defined by the information content of both knowledge. Experimental results are promising, confirming that

MEPRO is an effective method for discovering interesting exceptions. However, some of the knowledge discovered by MEPRO showed little unexpectedness due to the lack of the attention to this characteristic in the system. In other words, the exceptions discovered by MEPRO are interesting, but some of them are predictable from the relationships inherent in the domain. In this paper, we modify MEPRO so that it discovers unexpected exceptions, and apply it to benchmark data sets in the machine learning community. The essentials of the modifications are the introduction of the stochastic constraints on the events predicted by the discovered knowledge. A general view of discovery as search is also given for comparing MEPRO with the modified system MEPROUX in terms of a single framework. The experiments showed that MEPROUX is a promising method for the effective and efficient discovery of unexpected and useful exceptions.

2 Discovery as Search

In order to compare MEPROUX with MEPRO in terms of a single framework, discovery is cast as a search problem in this paper. The astute readers may refer to the work done by Mitchell [6], in which generalization is viewed as search problem for characterizing various approaches in terms of the search strategies. Our view to discovery problem is almost identical, although the formalization of discovery method is slightly different due to the ambiguity of usefulness in discovery compared with the preciseness of accuracy in classification.

A discovery program accepts, from a data set, input instances represented in some language, which we call the instance language after Mitchell. Discovered knowledge corresponds to the possibly useful information extracted from the input instances. The knowledge is represented in a second language, which we shall call the knowledge language. Given the instance language and the knowledge language, the discovery problem is to output K pieces of knowledge by observing a set of instances in a data set, where K is a user-specified number.

This problem is essentially a search problem since the discovery task corresponds to examining the search space of possible solutions, which is defined by the knowledge language, to determine useful pieces of knowledge. Needless to say, the instances represented in the instance language plays a crucial role in the examination.

3 Rule Pair

The instance language employed in MERPO and MEPROUX is in an ordinary propositional representation. 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. The requirement for *discrete* valued attributes is dictated by the very nature of the rule-based representation. Hence, continuous attributes are supposed to be converted to nominal attributes using an existing method such as [2].

In the knowledge language, we view a piece of knowledge r_i to be discovered from a data set as represented by a **rule pair** which consists of an exception and a piece of common sense knowledge associated with it. Here, each piece of knowledge is represented by a stochastic if-then rule which states that the conclusion holds true with some probability if the premise holds true. This representation has been chosen since it is widely used in the A. I. community in spite of its simplicity. Hence, a piece of common sense knowledge is represented by $Y_\mu \rightarrow x$, where $Y_\mu = y_1 \wedge y_2 \wedge \cdots \wedge y_\mu$. Here, x and y_i are **atoms**, each of which is an event representing, in propositional form, a single value assignment to an attribute. On the other hand, an exception is represented by $Y_\mu \wedge Z_\nu \rightarrow x'$, where $Z_\nu = z_1 \wedge z_2 \wedge \cdots \wedge z_\nu$, and x' and z_i are atoms. Atoms x and x' have the same attribute but different values. To sum up, a node in the search space represents a rule pair $r(\mu, \nu)$, which is defined in the following way.

$$r(\mu,\nu) \equiv \left\{ \begin{array}{lcl} Y_\mu & \rightarrow & x \\ Y_\mu \wedge Z_\nu & \rightarrow & x' \end{array} \right. \quad (1)$$

Since a stochastic if-then rule represents correlation or causality between its premise and conclusion, every rule pair is assumed to satisfy the following inequalities.

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

4 GACE criterion

If discovery is viewed as a search problem, then discovery methods can be characterized in terms of the search methods that they employ. In our view, a search method consists of a criterion for evaluating the goodness of a node in the search space, and a search strategy for determining the traversing order of the nodes. This section gives an explanation of the evaluation criterion GACE employed in MEPRO and MEPROUX. We also describe the additional constraints introduced in MEPROUX for discovering unexpected exceptions.

Let $\overline{x}$ be the complement of x, and $\mathrm{p}(x, Y_\mu)$ be the joint probability of x and Y_μ, then 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.

$$\begin{aligned}\mathrm{ACE}(x, Y_\mu) \equiv\ & [\{-n\mathrm{p}(x, Y_\mu)\log_2 \mathrm{p}(x) \\ & -n\mathrm{p}(\overline{x}, Y_\mu)\log_2 \mathrm{p}(\overline{x})\} \\ & -\{-n\mathrm{p}(x, Y_\mu)\log_2 \mathrm{p}(x|Y_\mu) \\ & -n\mathrm{p}(\overline{x}, Y_\mu)\log_2 \mathrm{p}(\overline{x}|Y_\mu)\}]/n \\ =\ & \mathrm{p}(x, Y_\mu)\log_2 \frac{\mathrm{p}(x|Y_\mu)}{\mathrm{p}(x)} \\ & +\mathrm{p}(\overline{x}, Y_\mu)\log_2 \frac{\mathrm{p}(\overline{x}|Y_\mu)}{\mathrm{p}(\overline{x})} \end{aligned} \quad (3)$$

A rule of large information content is useful in the sense that it gives a compact representation for data stored in a data set. 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 data set 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 [10] showed various desirable properties of ACE as a criterion for evaluating the interestingness of an if-then rule extracted from a data set.

However, an exception $Y_\mu \wedge Z_\nu \rightarrow x'$, whose ACE is high, may not be "interesting" if the ACE of the associated piece of common sense knowledge $Y_\mu \rightarrow x$ is extremely low. That is, the interestingness of an exception depends not only on its ACE but also on the ACE of the associated piece of common sense knowledge. It is reasonable therefore to represent the interestingness of an exception 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 arithmetic mean $\{\mathrm{ACE}(x, Y_\mu) + \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)\}/2$ and the geometric mean $\sqrt{\mathrm{ACE}(x, Y_\mu) \cdot \mathrm{ACE}(x', Y_\mu \wedge Z_\nu)}$ are considered as the simplest formulations.

Let us analyze the appropriateness of these functions as evaluation criteria for the interestingness of an exception. 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)$\sim$(6) are obtained.

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

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

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

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.

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

From Lemma 1 in the appendix, 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

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

where from equation (6) and (7),

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

A simple calculation shows that the maximum of the arithmetic mean $(U+V)/2$ 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 arithmetic mean 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 data sets will yield results which contain useless and uninteresting knowledge.

On the other hand, the geometric mean $\sqrt{U \cdot V}$ can be proved to possess no such shortcomings, and thus the geometric mean of the ACEs, **Geometric mean of the Average Compressed Entropies (GACE)**, can be considered as one of the simplest functions appropriate for evaluating the interestingness of an exception. Therefore, the

interestingness function is defined by the GACE, $\mathrm{GACE}(x, Y_\mu, x', Z_\nu)$.

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

As we will describe in section 6, although MEPRO, which employed GACE criterion, succeeded in discovering interesting exceptions, some of them showed little unexpectedness. Careful analysis of the results led to the introduction of the following three constraints in the evaluation criterion of MEPROUX. Only rule pairs which satisfy all of these constraints are evaluated in MEPROUX.

First, we found that the rule $Z_\nu \rightarrow x'$, which we call the reference rule, plays an important role in determining the unexpectedness of the exception $Y_\mu, Z_\nu \rightarrow x'$. If the conditional probability $\mathrm{p}(x'|Z_\nu)$ is too high, it is easily predicted that the addition of the condition Z_ν to the premise of a piece of common sense knowledge $Y_\nu \rightarrow x$ changes the conclusion from x to x'. Various constraints on $\mathrm{p}(x'|Z_\nu)$ have been investigated, and the following constraint is currently employed in MEPROUX.

$$\mathrm{p}(x'|Z_\nu) \leq \mathrm{p}(x') + \frac{1 - \mathrm{p}(x')}{2} \quad (11)$$

Second, some of the exceptions had smaller conditional probability $\mathrm{p}(x'|Y_\mu, Z_\nu)$ compared to the conditional probability $\mathrm{p}(x'|Z_\nu)$ of the reference rule. These exceptions can be considered to exhibit meaningless combination of Y_μ and Z_ν in predicting x'. The following constraint has been introduced in order to circumvent this problem.

$$\mathrm{p}(x'|Y_\mu, Z_\nu) > \mathrm{p}(x'|Z_\nu) \quad (12)$$

Third, some of the rule $Y_\nu \rightarrow x$ could not be admitted as a piece of common sense knowledge since a more natural rule $Y_\nu \rightarrow x''$ existed where $\mathrm{p}(x|Y_\mu) < \mathrm{p}(x''|Y_\mu)$, and atoms x and x'' have the same attribute but different values. The objective of the following constraint is to remove such knowledge.

$$\forall x'' \mathrm{p}(x|Y_\mu) \geq \mathrm{p}(x''|Y_\mu) \quad (13)$$

5 Search Strategy

The search strategies employed in MEPRO and MEPROUX are identical: both of them represents a depth-first search with maximum depth D to traverse a search tree in which a node represents a rule pair $r(\mu, \nu)$ of equation (1). 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 in the search tree, and as the depth increases by 1, an atom is added to the premise of a rule in a rule pair. 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 (≥ 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 Theorem 1 in the appendix, an upper-bound exists for the GACE of this rule pair. In other words, if the upper-bound for the current node is lower than $GACE_K$ (the Kth highest GACE of the discovered rule pairs), no rule pair exists whose GACE is higher than $GACE_K$ 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 $GACE_K$ is employed in our systems.

6 Application to Data Sets

MEPRO and MEPROUX have been tested with data sets from several domains, including the voting records data set and the mushroom data set [7].

The voting records data set 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 and 2 shows the results of asking MEPROUX and MEPRO for the 10 best rule pairs respectively, where the maximum search depth D is restricted to 8. A comma and $\mathcal{C}$ in the table represent conjunction and the premise of the associated piece of common sense knowledge respectively, while x and Y are the conclusion and the premise, respectively.

From table 1, we note that unexpected interesting exceptions emerge, confirming that the system is adequate for the task. For instance, according to the fourth rule pair, 70 % of the 212 congressmen who voted "yes" to "salvador" voted "yes" to "education". However, 11 of these who voted "no" to "physician", "yes" to "religious" and "yes" to "nicaraguan" all voted "no" to "education". This exception can not be easily guessed from the reference rule, since the conditional probability of the rule is only 76 % while the probability of voting "no" to "education" is 54 %. Note that such a weak regularity is quite common in a data set. The premise of this exception, which can be viewed as giving a partial definition of these exceptional congressmen, is highly interesting.

On the other hand, exceptions in table 2 are interesting but some of them can be easily predicted from the reference rule. For example, the first rule pair shows that all the 22 republicans who voted "yes" to "adoption" voted "yes" to "physician", which is an exception to the piece of common sense knowledge: a congressman votes "no" to "physician" if he votes

Table 1: The 10 best rule pairs with their associated reference rules discovered by MEPROUX from the voting data set.

Rank	common sense knowledge exception reference rule	p($x|Y$)	p(x)	P(Y)	ACE	$GACE$
1	handicapped=no, crime=yes, → salvador=yes	0.84	0.49	0.40	0.154	
	$\mathcal{C}$, party=demo, physician=no, religious=yes, superfund=no → salvador=no	1.00	0.48	0.03	0.029	0.0672
	party=demo, physician=no, religious=yes, superfund=no → salvador=no	0.73	0.48	0.13		
2	salvador=yes, exports=no → education=yes	0.77	0.39	0.39	0.163	
	$\mathcal{C}$, party=demo, handicapped=yes, adoption=yes, physician=no, religious=yes → education=no	1.00	0.54	0.02	0.021	0.0582
	party=demo, handicapped=yes, adoption=yes, physician=no, religious=yes → education=no	0.71	0.54	0.09		
3	nicaraguan=yes → physician=no	0.87	0.57	0.56	0.175	
	$\mathcal{C}$, salvador=yes, religious=yes, satellite=yes → physician=yes	0.81	0.41	0.04	0.018	0.0567
	salvador=yes, religious=yes, satellite=yes → physician=yes	0.69	0.41	0.08		
4	salvador=yes → education=yes	0.70	0.39	0.49	0.138	
	$\mathcal{C}$, physician=no, religious=yes, nicaraguan=yes → education=no	1.00	0.54	0.03	0.023	0.0561
	physician=no, religious=yes, nicaraguan=yes → education=no	0.76	0.54	0.16		
5	handicapped=no, missile=no → education=yes	0.78	0.39	0.35	0.152	
	$\mathcal{C}$, adoption=yes, religious=yes, nicaraguan=yes → education=no	1.00	0.54	0.02	0.021	0.0560
	adoption=yes, religious=yes, nicaraguan=yes → education=no	0.71	0.54	0.18		
6	satellite=yes → party=demo	0.84	0.61	0.55	0.094	
	$\mathcal{C}$, salvador=yes, nicaraguan=yes, crime=yes → party=rep	0.89	0.39	0.04	0.033	0.0557
	salvador=yes, nicaraguan=yes, crime=yes → party=rep	0.68	0.39	0.06		
7	handicapped=no, crime=yes → salvador=yes	0.84	0.49	0.40	0.154	
	$\mathcal{C}$, party=demo, religious=yes, superfund=no → salvador=no	0.92	0.48	0.03	0.020	0.0552
	party=demo, religious=yes, superfund=no → salvador=no	0.71	0.48	0.13		
8	handicapped=no, crime=yes → salvador=yes	0.84	0.49	0.40	0.154	
	$\mathcal{C}$, physician=no, religious=yes, superfund=no → salvador=no	0.92	0.48	0.03	0.020	0.0552
	physician=no, religious=yes, superfund=no → salvador=no	0.71	0.48	0.13		
9	satellite=yes → physician=no	0.82	0.57	0.55	0.118	
	$\mathcal{C}$, salvador=yes, nicaraguan=yes, south-africa=yes → physician=yes	0.88	0.41	0.04	0.025	0.0546
	salvador=yes, nicaraguan=yes, south-africa=yes → physician=yes	0.70	0.41	0.05		
10	satellite=no, missile=no → party=rep	0.76	0.39	0.36	0.152	
	$\mathcal{C}$, adoption=yes, salvador=yes, immigration=no → party=demo	1.00	0.61	0.03	0.019	0.0543
	adoption=yes, salvador=yes, immigration=no → party=demo	0.80	0.61	0.06		

Table 2: The 10 best rule pairs with their associated reference rules discovered by MEPRO from the voting data set.

Rank	common sense knowledge exception reference rule	p($x|Y$)	p(x)	P(Y)	ACE	$GACE$
1	adoption=yes → physician=no	0.87	0.57	0.58	0.175	
	$\mathcal{C}$, party=rep → physician=yes	1.00	0.41	0.05	0.066	0.1070
	party=rep → physician=yes	0.97	0.41	0.39		
2	adoption=yes → party=demo	0.91	0.61	0.58	0.195	
	$\mathcal{C}$, physician=yes, satellite=yes → party=rep	1.00	0.39	0.04	0.054	0.1024
	physician=yes, satellite=yes → party=rep	0.95	0.39	0.09		
3	satellite=yes → physician=no	0.82	0.57	0.55	0.118	
	$\mathcal{C}$, party=rep → physician=yes	0.95	0.41	0.09	0.088	0.1018
	party=rep → physician=yes	0.97	0.41	0.39		
4	party=demo → salvador=no	0.75	0.48	0.61	0.135	
	$\mathcal{C}$,nicaraguan=no,crime=yes → salvador=yes	0.97	0.49	0.09	0.075	0.1007
	nicaraguan=no,crime=yes → salvador=yes	0.99	0.49	0.38		
5	crime=yes → party=rep	0.64	0.39	0.57	0.105	
	$\mathcal{C}$, physician=no → party=demo	0.97	0.61	0.17	0.095	0.1003
	physician=no → party=demo	0.99	0.61	0.57		
6	adoption=yes → party=demo	0.91	0.61	0.58	0.195	
	$\mathcal{C}$, physician=yes, synfuels=no, south-africa=yes → party=rep	1.00	0.39	0.04	0.050	0.0993
	physician=yes, synfuels=no, south-africa=yes → party=rep	0.99	0.39	0.18		
7	salvador=yes → party=rep	0.74	0.39	0.49	0.182	
	$\mathcal{C}$, physician=no → party=demo	0.98	0.61	0.10	0.054	0.0990
	physician=no → party=demo	0.99	0.61	0.57		
8	crime=yes → salvador=yes	0.78	0.49	0.57	0.151	
	$\mathcal{C}$,physician=no, satellite=yes, nicaraguan=yes → salvador=no	0.95	0.48	0.09	0.064	0.0985
	physician=no, satellite=yes, nicaraguan=yes → salvador=no	0.65	0.48	0.43		
9	nicaraguan=yes → party=demo	0.90	0.61	0.56	0.169	
	$\mathcal{C}$, physician=yes, synfuels=no → party=rep	1.00	0.39	0.04	0.057	0.0980
	physician=yes, synfuels=no → party=rep	0.98	0.39	0.32		
10	satellite=yes → party=demo	0.84	0.61	0.55	0.094	
	$\mathcal{C}$, physician=yes, salvador=yes → party=rep	1.00	0.39	0.07	0.100	0.0975
	physician=yes, salvador=yes → party=rep	0.93	0.39	0.39		

"yes" to "adoption". This exception exhibits the radical change of the conclusion caused by the additional condition "republicans", and is considered as highly interesting. However, it shows little unexpectedness since 97 % of republicans voted "yes" to "physician". From table 1 and 2, we can easily validate the effectiveness of the constraints introduced in section 4.

The mushroom data set includes 22 descriptions and the edibility class of 8124 mushrooms, each attribute having 2 to 12 values. While there were no restrictions on the attributes of the atoms in the previous experiment, users may also impose some constraints on them. In this experiment, for example, the edibility class is the only attribute allowed in the conclusions. Table 3 and 4 show the 8 most interesting rule pairs discovered by MEPROUX and MEPRO respectively, where the maximum search depth D is again set to 8.

The discovered rule pairs in table 3 also show interesting unexpected exceptions. According to the third rule pair, 74 % of the mushrooms whose "bruises" is "f" and "ring-number" is "o" are edible but 100 % of them are actually poisonous if the "ss-aring" is "f". This exception can not be easily predicted from the reference rule since its conditional probability is only 74 %. Again, exceptions in table 4 are interesting but some of them can be easily predicted from the reference rule. For example, the exception in the third rule pair shows little unexpectedness since the conditional probability of the reference rule is 100 %.

The maximum depth should be large enough so that MEPRO and MEPROUX investigate 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.

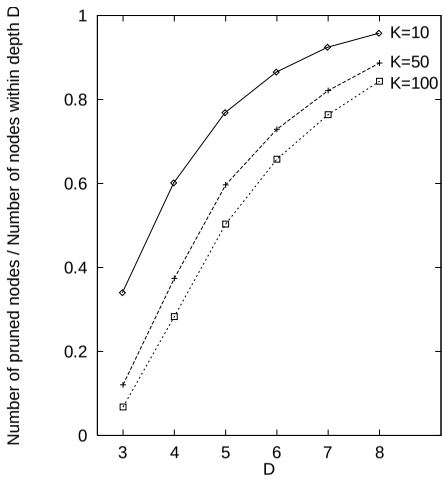

Figure 1: Performance of BBM with varying depth and number of target rule pairs K.

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 data set chosen for this evaluation was the "voting" data set described earlier. The systems were 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 exceptions.

7 Conclusion

In this paper, we have demonstrated the effectiveness of our proposed stochastic constraints for unexpected exceptions from the knowledge discovery viewpoint. An interpretation of discovery as search was given to compare the previous system MEPRO with the newly proposed system MEPROUX, in which the constraints are employed. Both systems were explained in detail, and two examples of discovering exceptions from a data set were also given for evaluating the effectiveness and the efficiency of the proposed system. The results were promising: MEPROUX produced truly unexpected exceptions efficiently.

Since MEPROUX requires no domain information except for a data set, it is effective in the exception discovery from the data sets where it is difficult to obtain background knowledge a priori. Moreover, it would discover unknown and useful exceptions from the data sets where such knowledge is left undiscovered due to the unpredictable misuse of user-supplied background knowledge. Ongoing work focuses on extensions and refinements of the basic MEPROUX system for applying it to larger problems. The ultimate goal of the study is the automatic discovery of some extremely useful exceptions which are unknown to mankind.

References

[1] Bergadano, F., Giordana, A. and Saitta, L. (1991). Integrated Learning in a Real Domain, in: Piatetsky-Shapiro, G. and Frawley, W. J. (eds.), *Knowledge Discovery in Databases*, AAAI Press/The MIT Press, pp.277-288.

[2] Dougherty, J., Kohavi, R. and Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features, in: *Proceedings of the 12th International Conference on Machine Learning*, Morgan Kaufmann, pp.194-202.

[3] Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J. (1991). Knowledge Discovery in

Table 3: The 8 best rule pairs with their associated reference rules discovered by MERPOUX from the mushroom data set, where the edibility class is the only attribute allowed in the conclusions.

Rank	common sense knowledge exception reference rule	p(x|Y)	p(x)	P(Y)	ACE	$GACE$
	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	0.52	0.132	
1	$\mathcal{C}$, ss-aring=f → class=e	1.00	0.52	0.05	0.048	0.0795
	ss-aring=f → class=e	0.74	0.52	0.07		
	bruises=f, veil-color=w, ring-number=o → class=p	0.77	0.48	0.52	0.132	
2	$\mathcal{C}$, ss-aring=f → class=e	1.00	0.52	0.05	0.048	0.0792
	ss-aring=f → class=e	0.74	0.52	0.07		
	bruises=f, ring-number=o → class=e	0.74	0.48	0.54	0.107	
3	$\mathcal{C}$, ss-aring=f → class=e	1.00	0.52	0.05	0.048	0.0713
	ss-aring=f → class=e	0.74	0.52	0.07		
	bruises=f, veil-color=w → class=p	0.72	0.48	0.56	0.096	
4	$\mathcal{C}$, ss-aring=f → class=e	1.00	0.52	0.05	0.048	0.0676
	ss-aring=f → class=e	0.74	0.52	0.07		
	bruises=f, g-attachment=f → class=p	0.72	0.48	0.56	0.095	
5	$\mathcal{C}$, ss-aring=f → class=e	1.00	0.52	0.05	0.048	0.0674
	ss-aring=f → class=e	0.74	0.52	0.07		
	stalk-root=?, sp-color=w → class=p	0.79	0.48	0.28	0.077	
6	$\mathcal{C}$, g-size=b → class=e	1.00	0.52	0.06	0.056	0.0659
	g-size=b → class=e	0.70	0.52	0.69		
	bruises=f, veil-color=w, sp-color=w → class=p	0.84	0.48	0.26	0.104	
7	$\mathcal{C}$, g-spacing=w, stalk-shape=e → class=e	1.00	0.52	0.04	0.039	0.0638
	g-spacing=w, stalk-shape=e → class=e	0.75	0.52	0.06		
	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	0.52	0.132	
8	$\mathcal{C}$, cap-color=w, stalk-root=e → class=e	1.00	0.52	0.03	0.030	0.0629
	cap-color=w, stalk-root=e → class=e	0.67	0.52	0.05		

Table 4: The 8 best rule pairs with their associated reference rules discovered by MERPO from the mushroom data set, where the edibility class is the only attribute allowed in the conclusions.

Rank	common sense knowledge exception reference rule	p(x|Y)	p(x)	P(Y)	ACE	$GACE$
	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	0.52	0.132	
1	$\mathcal{C}$,odor=n, sc-bring=w → class=e	1.00	0.52	0.11	0.107	0.1188
	odor=n, sc-bring=w → class=e	0.96	0.52	0.24		
	bruises=f, veil-color=w, ring-number=o → class=p	0.77	0.48	0.52	0.132	
2	$\mathcal{C}$,odor=n, sc-bring=w → class=e	1.00	0.52	0.11	0.107	0.1185
	odor=n, sc-bring=w → class=e	0.96	0.52	0.24		
	g-size=b → class=e	0.70	0.52	0.69	0.067	
3	$\mathcal{C}$,odor=f → class=p	1.00	0.48	0.19	0.205	0.1174
	odor=f → class=p	1.00	0.48	0.27		
	g-size=b → class=e	0.70	0.52	0.69	0.067	
3	$\mathcal{C}$,sp-color=h → class=p	1.00	0.48	0.19	0.205	0.1174
	sp-color=h → class=p	0.97	0.48	0.20		
	bruises=f, veil-color=w → class=p	0.72	0.48	0.56	0.096	
5	$\mathcal{C}$,odor=n, sc-bring=w → class=e	1.00	0.52	0.15	0.140	0.1160
	odor=n, sc-bring=w → class=e	0.96	0.52	0.24		
	bruises=f, g-attachment=f, ring-number=o → class=p	0.77	0.48	0.52	0.132	
6	$\mathcal{C}$,stalk-root=e → class=e	1.00	0.52	0.11	0.101	0.1156
	stalk-root=e → class=e	0.77	0.52	0.14		
	bruises=f, g-attachment=f → class=p	0.72	0.48	0.56	0.095	
7	$\mathcal{C}$,odor=n, sc-bring=w → class=e	1.00	0.52	0.15	0.140	0.1155
	odor=n, sc-bring=w → class=e	0.96	0.52	0.24		
	bruises=f, veil-color=w, ring-number=o → class=p	0.77	0.48	0.52	0.132	
8	$\mathcal{C}$,stalk-root=e → class=e	1.00	0.52	0.11	0.101	0.1153
	stalk-root=e → class=e	0.77	0.52	0.14		

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

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

[5] Matheus, C. J., Chan, P. K. C. and Piatetsky-Shapiro, G. (1993). Systems for Knowledge Discovery in Databases, *IEEE Transactions on Knowledge and Data Engineering*, 5 (6), pp.903-913.

[6] Mitchell, T. M. (1982). Generalization as Search, *Artificial Intelligence*, 18, pp.203-226.

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

[8] Piatetsky-Shapiro, G. (1991). Discovery, Analysis, and Presentation of Strong Rules, in: Piatetsky-Shapiro, G. and Frawley, W. J. (eds.), *Knowledge Discovery in Databases*, AAAI Press/The MIT Press, pp.229-248.

[9] Piatetsky-Shapiro, G. and Matheus, C. J. (1994). The Interestingness of Deviations, in: Fayyad, U. M. and Uthurusamy, R. (eds.), *AAAI-94 Workshop on Knowledge Discovery in Databases*, pp.25-36.

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

[11] Smyth, P. and Goodman, R. M. (1992). An Information Theoretic Approach to Rule Induction from Databases, *IEEE Transactions on Knowledge and Data Engineering*, 4 (4), pp.301-316.

[12] Suzuki, E. and Shimura, M. (1996). Exceptional Knowledge Discovery in Databases based on Information Theory, in: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp.275-278.

[13] Ziarko, W. (1991). The Discovery, Analysis, and Representation of Data Dependencies in Databases, in: Piatetsky-Shapiro, G. and Frawley, W. J. (eds.), *Knowledge Discovery in Databases*, AAAI Press/The MIT Press, pp.195-209.

Appendix

Lemma 1 *Let* $\sum_{i=1}^{n_1} a_i / \{\sum_{i=1}^{n_1} a_i + \sum_{i=1}^{n_2} b_i\} > \mathrm{p}(x)$. *Then*

$$G \equiv \sum_{i=1}^{n_1} a_i \log_2 \left(\frac{\sum_{i=1}^{n_1} a_i}{\sum_{i=1}^{n_1} a_i + \sum_{i=1}^{n_2} b_i} \frac{1}{\mathrm{p}(x)} \right) + \sum_{i=1}^{n_2} b_i \log_2 \left(\frac{\sum_{i=1}^{n_2} b_i}{\sum_{i=1}^{n_1} a_i + \sum_{i=1}^{n_2} b_i} \frac{1}{\mathrm{p}(\overline{x})} \right) \tag{14}$$

increases monotonously for each a_j, *and decreases monotonously for each* b_j.

Proof This lemma is easily proved by calculating the partial derivatives.

Theorem 1 *Let* $\mathrm{H}(\alpha) \equiv [\alpha/\{(1+\alpha)\mathrm{p}(\overline{x})\}]^{2\alpha}/\{(1+\alpha)\mathrm{p}(x)\}$, α_1 *and* α_2 *satisfy* $\mathrm{H}(\alpha_1) > 1 > \mathrm{H}(\alpha_2)$, *and* $GACE = \mathrm{GACE}(x, Y_{\mu'}, x', Z_{\nu'})$. *If* $\mathrm{H}(\mathrm{p}(x', Y_\mu, Z_\nu)/\mathrm{p}(x, Y_\mu)) < 1$ *then,*

$$GACE < \mathrm{p}(x, Y_\mu) \left[\alpha_2 \left\{ \log_2 \left(\frac{1}{1+\alpha_1} \frac{1}{\mathrm{p}(x)} \right) + \alpha_1 \log_2 \left(\frac{\alpha_1}{1+\alpha_1} \frac{1}{\mathrm{p}(\overline{x})} \right) \right\} \log_2 \frac{1}{\mathrm{p}(x')} \right]^{\frac{1}{2}}, \tag{15}$$

else

$$GACE \leq \left[\left\{ \mathrm{p}(x, Y_\mu) \log_2 \left(\frac{\mathrm{p}(x, Y_\mu)}{\mathrm{p}(x, Y_\mu) + \mathrm{p}(x', Y_\mu, Z_\nu)} \cdot \frac{1}{\mathrm{p}(x)} \right) + \mathrm{p}(x', Y_\mu, Z_\nu) \cdot \log_2 \left(\frac{\mathrm{p}(x', Y_\mu, Z_\nu)}{\mathrm{p}(x, Y_\mu) + \mathrm{p}(x', Y_\mu, Z_\nu)} \frac{1}{\mathrm{p}(\overline{x})} \right) \right\} \cdot \mathrm{p}(x', Y_\mu, Z_\nu) \log_2 \frac{1}{\mathrm{p}(x')} \right]^{\frac{1}{2}}. \tag{16}$$

Proof Since ACE is positive [10], if both the ACEs, $\mathrm{ACE}(x, Y_{\mu'})$ and $\mathrm{ACE}(x', Y_{\mu'} \wedge Z_{\nu'})$, are maximized, then $GACE$ is likewise maximized. Let $q = \mathrm{p}(x', Y_{\mu'}, Z_\nu)$, then from Lemma 1, this is the case when

$$\begin{aligned} \mathrm{p}(x, Y_{\mu'}, Z_\nu) &= \mathrm{p}(x, Y_\mu, Z_\nu), \\ \mathrm{p}(x, Y_{\mu'}, \overline{Z_\nu}) &= \mathrm{p}(x, Y_\mu, \overline{Z_\nu}), \\ \mathrm{p}(x', Y_{\mu'}, Z_{\nu'}) = q,\ \mathrm{p}(x', Y_{\mu'}, \overline{Z_\nu}) &= 0 \\ \mathrm{p}(\overline{x \vee x'}, Y_{\mu'}, Z_\nu) = \mathrm{p}(\overline{x \vee x'}, Y_{\mu'}, \overline{Z_\nu}) &= 0, \\ \mathrm{p}(x, Y_{\mu'}, Z_{\nu'}) = \mathrm{p}(\overline{x \vee x'}, Y_{\mu'}, Z_{\nu'}) &= 0, \end{aligned} \tag{17}$$

since these constraints do not restrict the range of q. The proof can be easily obtained by maximizing $GACE$ with respect to q where $0 \leq q \leq \mathrm{p}(x', Y_\mu, Z_\nu)$ and $q < \mathrm{p}(x, Y_\mu)\mathrm{p}(\overline{x})/\mathrm{p}(x)$.