Journal of Universal Computer Science, vol. 12, no. 6 (2006), 627-653
submitted: 31/5/06, accepted: 23/6/06, appeared: 28/6/06 © J.UCS

Data Mining Methods for Discovering Interesting Exceptions from an Unsupervised Table

Einoshin Suzuki
(Kyushu University, Japan
suzuki@i.kyushu-u.ac.jp)

Abstract: In this paper, we survey efforts devoted to discovering interesting exceptions from data in data mining. An exception differs from the rest of data and thus is interesting and can be a clue for further discoveries. We classify methods into exception instance discovery, exception rule discovery, and exception structured-rules discovery and give a condensed and comprehensive introduction.

Key Words: exception, instance, rule, interestingness, unexpectedness, pattern, data mining
Category: I.2.6 - Learning

1 Introduction

With the prevalence of computer systems such as WWW, huge data are increasing in number, size, and their degree of importance. Computers, which are progressing in terms of both hardware and software, provide an effective mean for analyzing such data. Consequently today we have reasons and means for efficiently analyzing huge data of considerable importance with a low cost.

Data mining or Knowledge Discovery in Databases (KDD) [Fayyad et al. 1996] represents a research field that views such data as a gigantic mine and try to find useful knowledge that corresponds to precious resource. One widely used definition is “Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” [Frawley et al. 1991]. In the past 15 years, the community of data mining has grown from a few academic workshops and project teams to numerous academic societies and companies.

Data mining is related with various research fields including machine learning, pattern recognition, database, statistics, artificial intelligence, knowledge acquisition, and data visualization [Fayyad et al. 1996]. Especially, it is generally agreed that machine learning, database, and statistics represent major fields. These three research fields have effective solutions in realizing inductive inference, processing huge data, and analyzing data, respectively, and thus are beneficial for other research fields. Data mining borrows techniques from these research fields but is devoted to the discovery of useful knowledge from huge data. For instance, learning from examples in machine learning typically concerns induction of global model of high accuracy from memory-resident data, while rule discovery in data mining rather concerns discovery of local tendencies of high value from disk-resident data [Mitchell 1999].

Finding useful knowledge from huge data is a nontrivial task and is recognized to follow a KDD process model, which represents an iterative application of data engineering, pattern extraction, and pattern interpretation methods. Data engineering, which consists of data measurement and data preprocessing, produces a processed data, from which patterns are extracted. A pattern is defined as an expression that includes a set of variables and can possibly be turned into useful knowledge after interpretation and deployment by humans. We believe that pattern extraction corresponds to the core of the KDD process because it represents a challenging task that requires techniques of the three major fields.

Pattern extraction consists of a model/pattern structure, a score function, an optimization/search method, and a data management strategy [Hand et al. 2001]. The four issues represent how to represent a pattern, how to evaluate the goodness of a pattern, how to look for patterns, and how to use data that typically reside in disk, respectively. As the notion of usefulness cannot be defined precisely, a score function is defined so that it is expected to capture "interestingness", which represents potential usefulness.

Exception is a class of knowledge that attracts much attention in the data mining community. An exception differs from the rest of data and thus is interesting and can be a clue for further discoveries. History of science shows that a new theory is typically built by finding an exception to a theory and extending it to explain the exception. In this paper, we conduct a survey on methods for finding interesting exceptions. Our survey can be viewed as an extension of [Suzuki 2004a] and borrows several ideas from [Suzuki 2004b].

2 Exception Discovery

Because there exist a variety of approaches for finding interesting exceptions, we give an intuitive but general definition of our problem. We define an interesting exception as something different from most of the rest. The objective of an exception discovery task is to obtain a set Π of exceptions given data D and additional information α. Here α typically represents domain-specific criteria such as expected profit or domain knowledge and an element of Π represents an exception π.

The case in which D represent a table, alternatively stated "flat" data, is most extensively studied in data mining. On the other hand, the case when data D represent structured data such as time-series data and text data typically necessitates a procedure for handling such a structure. In order to focus on the interestingness aspect, we limit our attention to the former case and recommend [Chakrabarti et al. 1998, Liu et al. 2001] to readers interested in finding interesting exceptions from structured data. In this case, D consists of n examples $e_1, e_2, \ldots, e_n$. An example e_i is described with m attributes $a_1, a_2, \ldots, a_m$ and an attribute a_j takes one of $|a_j|$ values $v_{j,1}, v_{j,2}, \ldots, v_{j,|a_j|}$. We represent $e_k = (w_{k,1}, w_{k,2}, \ldots, w_{k,m})$.

Various forms of patterns have been studied in data mining though much of the efforts have been spent on rules, classifiers, clusters, and examples. In data mining, a rule typically represents a probabilistic tendency between a premise and a conclusion thus possibly represents an exceptional tendency. It should be noted that a set of rules related to each other might represent an exceptional tendency more appropriately than a single rule. A classifier represents a global model for predicting the class of an example but can be safely ignored as we limit our attention to finding exceptions. A cluster consists of similar examples and thus possibly represents a set of exception instances. An example or an instance is, however, considered more appropriate for representing an exception than a cluster because it can be an outlier i.e. an example that is substantially different from other examples. In this paper, we classify exception discovery into exception instance discovery, exception rule discovery, and exception structured-rules discovery. In exception instance discovery, an exception π represents a generalized example $e'_j = (w'_{j,1}, w'_{j,2}, \ldots, w'_{j,m})$, where $w'_{j,k} = w_{j,k}$ or $w'_{j,k}$ corresponds to a wild card (i.e. $*$ in a UNIX command). In exception rule discovery, an exception π represents a rule $y \rightarrow x$, where each of x and y represents a logical expression with a set of value v specifications $a = v$'s to an attribute a. In exception structured-rules discovery, an exception π represents a set of rules related to each other.

From another important viewpoint, exception discovery can be classified into supervised, which employs class information in the discovery process, and unsupervised, which does not. We note that the latter can lead to more unexpected discoveries and thus focus on unsupervised discovery in this survey. Therefore, supervised methods such as [Freund and Schapire 1996, Sugaya et al. 2001, Yamanishi and Takeuchi 2001, Yamada et al. 2003, Jumi et al. 2004, Jumi et al. 2005] and research topics such as cost-sensitive classification [Elkan 2001], fraud detection from known frauds [Chan and Stolfo 1998, Fawcett and Provost 1997, Lee et al. 1998] are excluded from this survey.

In the remainder of this paper, we will overview exception instance discovery, exception rule discovery, and exception structured-rules discovery in Sections 3, 4, and 5, respectively, then we give some conclusions in Section 6.

3 Exception Instance Discovery

3.1 Overall View

In this section, we overview the case that a pattern represents an instance or a part of an instance. Such a pattern is typically called an outlier, which has attracted attention of researchers in statistics even before the proliferation of data mining research. While there are various definitions of an outlier, [Hawkins 1980] states that "an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism".

Before exception instance discovery became popular, outlier detection could be classified into a distribution-based approach, a computational geometry approach, and a machine learning approach. These approaches differ from discovery of interesting exception instances in that they consider outliers as noise. For instance, in machine learning, several clustering methods such as CLARANS [Ng and Han 1994], DBSCAN [Ester et al. 1995], BIRCH [Zhang et al. 1996], and CURE [Guha et al. 1998] consider outliers as something that prevents accurate clustering. In other words, outliers are considered as useless by-products of clustering and are not the primary objective of the algorithms.

One of the earliest data mining methods devoted to discovery of interesting exception instances is [Arning et al. 1995]. It is a heuristic method that takes D in which $E \equiv \{e_1, e_2, \ldots, e_n\}$; a sequence S of ν subsets $E_1, E_2, \ldots, E_\nu$ with $2 \leq \nu \leq n$, $E_j \subseteq E$, and $E_{j-1} \subset E_j$; a user-defined dissimilarity function $\mathcal{D}_{\mathcal{S}} : \mathcal{P}(E) \rightarrow \Re_0^+$ with respect to S; a cardinality function $\mathcal{C} : \mathcal{P}(E) \rightarrow \Re_0^+$ with $(\forall E_1, E_2 \subseteq E)$ $E_1 \subset E_2 \Rightarrow \mathcal{C}(E_1) < \mathcal{C}(E_2)$ as input. Then based on its $O(n)$ algorithm, it discovers a sequential exception E_{x} which is defined as $(\forall E_j$ occurring in $S)$ $SF(E_{\mathrm{x}}) \geq SF(E_{\mathrm{j}})$ in terms of a smoothing factor $SF(E_j) \equiv \mathcal{C}(E_j - E_{j-1})(\mathcal{D}_S(E_j) - \mathcal{D}_S(E_{j-1}))$. Though the method lacks a solid theoretical foundation and extensive validation, it represents a pioneering work in discovery of interesting instances.

Currently, discovery of interesting exception instances can be classified into either a distance-based approach, a density-based approach, a projection-based approach, and a distribution-based approach. We define that the distance-based approach employs a distance function in discovering an interesting exception instance. The density-based approach detects an interesting exception instance based on the local density of the instance's neighborhood. In the projection-based approach, an outlier is detected in a subspace of the example space that is spanned by given attributes. We define that the distribution-based approach employs a probabilistic distribution in discovering an interesting exception instances. In this approach, an outlier is typically an observation that deviates from a standard distribution. These approaches overlap; e.g. a method in the density-based approach often employs a distance function and thus we use the decision list shown in Figure 1. We will see these approaches and discuss miscellaneous issues in the remainder of this section.

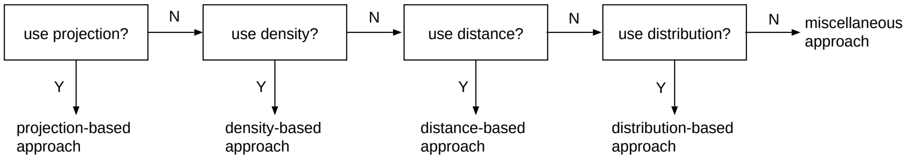

Figure 1: Decision list for classifying a method for discovery of interesting exception instances

3.2 Distance-based Approach

Knorr and Ng are honored to have opened a way for outliers to "first-class citizen" with their distance-based approach [Knorr and Ng 1998, Knorr et al. 2000]. Their work focuses on detection of interesting exception instances and differs from the previous methods in that they deal with disk-resident large-scale data. Their definition of an outlier can be stated as follows: "an example e in a dataset D is a $DB(p, r)$ outlier if at least a fraction p of the examples in D lies greater than distance r from e." They assume that each attribute a_i represents a numerical attribute and use the L_P-norm distance. Typically, p is near 1. Thus, intuitively an example is considered unusual if its neighboring examples are far away. They also showed that their notion of distance-based outlier generalizes many notions from the distribution-based approach.

As researchers from the database community, one of the main interests of Knorr and Ng seems to be algorithmic issues for processing disk-resident data with a small number of disk scans and thus a small amount of execution time. For the problem of finding all outliers given specific values for p and r, they proposed an algorithm that partitions the example space into cells, i.e. hypercubes with edge length $\frac{r}{2\sqrt[P]{m}}$, and exploits populated cells each of which contains no less than $n(1-p)$ examples to prune cells in their neighborhood. Another main interest seems to concern the discovery of interesting instances. Applications for finding exceptional NHL hockey players from record statistics and finding exception trajectories from video-surveillance data are appealing and promising.

According to Ramaswamy et al., choosing appropriate values for p and r is a nontrivial task and the definition of Knorr and Ng does not provide a degree of being an outlier to examples [Ramaswamy et al. 2000]. Therefore, Ramaswamy et al. proposed a novel definition for distance-based outliers based on the distance of a point (i.e. an example represented with numerical attributes) to its kth nearest neighbor. Their definition of an outlier can be stated as follows: "given an input data set with n points, parameters ν and k, a point e is a D_{ν}^{k} outlier if there are no more than $\nu - 1$ other points e' such that $D^k(e') > D^k(e)$", where they use $D^k(p)$ to denote the distance of point e from its kth nearest neighbor.

Although distance-based methods are effective in a number of applications, it typically requires a large amount of computation time. Straightforward algorithms based on pairwise comparison of examples typically require $O(n^2)$ distance computations. This quadratic scaling becomes a real problem in data sets with more than millions of records [Bay and Schwabacher 2003]. Partition-based algorithms such as that of Knorr and Ng are inefficient when m is large. Another drawback is the inappropriateness of a distance measure in high-dimensional space, i.e. the curse of dimensionality [Aggarwal and Yu 2001, Aggarwal 2001]. We will present countermeasures for this problem in Section 3.4. Distance measures based on solid theoretical foundations such as [Bennett et al. 1998] might be a solution.

3.3 Density-based Approach

The distance-based approach sometimes gives counterintuitive results. For instance, in Figure 2, both e_1 and e_2 can be regarded as an outlier according to the definition of Hawkins in Section 3.1. However, a method in the distance-based approach such as [Knorr and Ng 1998, Knorr et al. 2000] cannot recognize e_2 as an outlier because it uses a specific value of r for all examples and thus neglects local distributions of examples.

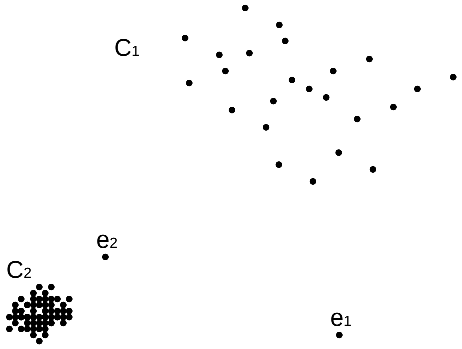

Figure 2: Motivating example for local outlier in [Breunig et al. 2000]. [Knorr and Ng 1998, Knorr et al. 2000] cannot recognize e_2 as an outlier.

To circumvent this problem, Breunig et al. proposed to assign to each example a degree of being an outlier based on the degree of how isolated the example is with respect to the surrounding neighborhood [Breunig et al. 2000]. This idea corresponds to considering the density of the neighborhood of the example and thus initiated the density-based approach. It should be noted that another degree of being an outlier was proposed by Ramaswamy et al. in the same conference as described in Section 3.2.

The degree is called LOF (local outlier factor) of an example and is defined as follows. First, they define the k-distance of an example e (denoted by k-$distance(e)$) as the distance $d(e, o)$ to an example o such that for at least k examples $o' \neq e$ it holds that $d(e, o') \leq d(e, o)$ and for at most $k - 1$ examples $o' \neq e$ it holds that $d(e, o') < d(e, o)$. Intuitively, k-distance represents the sparsity around the example e. Next, the k-distance neighborhood $N_k(e)$ of e is defined so that it contains every example whose distance from e is no greater than the k-distance:

$$N_k(e) = \{q \neq e | d(e, q) \leq k\text{-}distance(e)\}. \quad (1)$$

Third, they define the reachability distance $reach - dist_k(e, o)$ of example e with respect to example o as follows:

$$reach - dist_k(e, o) = \mathrm{MAX}\{k\text{-}distance(o), d(e, o)\}. \quad (2)$$

where $\mathrm{MAX}(x, y)$ represents the larger one of x and y. This definition alleviates effects of noise in the neighborhood of e. Fourth, the local reachability density $lrd_k(e)$ of an example e is defined as follows:

$$lrd_k(e) = \frac{|N_k(e)|}{\sum_{o \in N_k(e)} reach - dist_k(e, o)}. \quad (3)$$

Finally, the LOF $LOF_k(e)$ of e is defined as follows:

$$LOF_k(e) = \frac{\sum_{o \in N_k(e)} \frac{lrd_k(o)}{lrd_k(e)}}{|N_k(e)|}. \quad (4)$$

The computation of LOF values for every example requires a large number of k-nearest neighbors searches and can be computationally expensive. In [Breunig et al. 2000], they show three possibilities and use a variant of X-tree, which is an index method of complexity $O(n \log n)$. Jin et al. restrict the problem to finding only ν most outstanding local outliers, i.e. the top-ν examples that are most likely to be local outliers according to their LOFs [Jin et al. 2001]. To compress the data they introduced the concept of “micro-cluster”, which comes from BIRCH [Zhang et al. 1996] and was later generalized to data squashing [Dumouchel et al. 1999]. The idea is to compress the data into small clusters, represent each small cluster using statistical information, and use the upper and lower bounds for the

LOF of each example for pruning. The method showed significant improvement over X-tree-based methods in terms of computation time in experiments using synthetic data.

The density-based approach inherits many advantages of the distance-based approach and seems to fit our intuition as Figure 2 shows. One of the major drawbacks of the density-based approach is its large computation time as we have seen and another major drawback is its inappropriateness in high-dimensional space. One-class support vector machines (1-SVM), which is based on the notion of large margin classifier can be a solution for the latter [Schölkopf 2001]. Systematic comparison using real data would be interesting though a faster version of 1-SVM should be used.

3.4 Projection-based Approach

We have seen in previous sections that the distance-based approach and the density-based approach are inappropriate in high-dimensional space. Aggarwal and Yu attribute the reason to the facts that most of the points are likely to lie in a thin shell about any other point [Beyer 1999] and abnormal deviations may be embedded in some lower-dimensional space of the example space [Aggarwal and Yu 2001]. They proposed to find outliers in a subspace of the example space because a subspace is dense, is likely to be less affected by the noise, and only a subset of attributes are meaningful in a typical application.

Intuitively, an outlier is defined to reside in a subspace with abnormally low density in their method. They discretize each attribute into ϕ equi-depth ranges; thus, each range is expected to contain a fraction $f = 1/\phi$ of the examples. They assume that the central limit theorem can be applied to the problem of estimating the number of examples contained in a k-dimensional cube $\mathcal{D}$ and propose an evaluation index that they call the sparsity coefficient $S(\mathcal{D})$ as follows:

$$S(\mathcal{D}) = \frac{n(\mathcal{D}) - Nf^k}{\sqrt{Nf^k(1-f^k)}}, \tag{5}$$

where $n(\mathcal{D})$ represents the number of points in $\mathcal{D}$. As $n(\mathcal{D})$ is assumed to fit a normal distribution with mean Nf^k and standard deviation $\sqrt{Nf^k(1-f^k)}$, the normal distribution tables can be used to quantify the level of significance for a point to deviate significantly from average behavior.

The projection-based approach is supported by persuasive motivations and avoids some drawbacks of the other approaches. A serious shortcoming is that the number of cubes is $\Omega(2^m)$ which prevents any methods from finding all outliers. Though this problem can be tolerated by using heuristic search, it cannot avoid overlooking interesting exception instances. The projection-based approach for finding interesting exception instances is related with exception rule discovery and exception structured-rules discovery which we will see in the Sections 4 and 5, respectively.

3.5 Distribution-based Approach

An important application of discovery of interesting exception instances is fraud detection because fraudulent behaviors are rare and deviate from the norm. Though most of the methods for fraud detection are supervised, an unsupervised method has several advantages because it necessitates no labeled examples which is often costly and it can potentially detect novel types of frauds. Outlier detection has a long history in statistics [Hawkins 1980] but has largely focused on data that is univariate and data with a known or parametric distribution. In discovery of interesting exception instances, a distribution-based approach assumes multivarite data of a large number of examples and is often adaptive because fraudsters change their behaviors to avoid being detected.

One of the earliest unsupervised methods for fraud detection deals with cellular fraud calls [Burge and Shawe-Taylor 1997, Burge and Shawe-Taylor 2001]. It updates a set of prototypes of phone calls using a nonparametric model based on a maximal entropy principle. Important notions such as the use of decay parameters to weaken the influence of previous calls, incremental updates for online learning, and differential analysis between a current behavior profile and a behavior profile history using the Hellinger distance [Pitman 1979] were introduced. Several interesting examples in the application domain were shown in [Burge and Shawe-Taylor 1997], although no systematic experiments were reported. The absence might be due to the confidential nature of the project under which the work was done.

Inspired by [Burge and Shawe-Taylor 1997], Yamanishi et al. proposed Smart Shifter [Yamanishi et al. 2000], which can use parametric models for profiles and measures the difference of the input data and profiles. It assumes that an example $(e_{\mathrm{Cat},t}, e_{\mathrm{Con},t})$ has a time stamp t and is divided into vectors $e_{\mathrm{Cat},t}$ and $e_{\mathrm{Con},t}$ described with categorical and continuous attributes, respectively. Given the tth input $(e_{\mathrm{Cat},t}, e_{\mathrm{Con},t})$, Smart Shifter first identifies the cell in a multi-dimensional histogram that $e_{\mathrm{Cat},t}$ falls into and updates the histogram density using an extension of Laplace estimation that discounts past examples. Then, Smart Shifter updates the finite mixture model for that cell based on an extension of the EM algorithm and a prototype updating algorithm both of which discount past examples to obtain $p^{(t)}(e_{\mathrm{Con}}|e_{\mathrm{Cat}})$, where the superscript (t) represents the tth function. Finally, Smart Shifter calculates a score for the example on the basis of the models before and after the update using the Hellinger score. [Yamanishi et al. 2000] shows systematic experiments for network intrusion detection and interesting outliers from medical pathology data. They also proposed a unifying framework for detecting outliers and change points from non-stationary time series data [Yamanishi and Takeuchi 2002], but we omit methods for structured data as we stated in Section 2.

The distribution-based approach is based on a solid theoretical foundation, is typically time-efficient, and often provides on-line detection. One of its major drawbacks is that it typically assumes a small number of attributes, say between 2 and 7, because even a large number of examples would be insufficient to estimate a statistical model of higher dimensions accurately. We believe that using other approaches for feature selection for the distribution-based approach is a promising avenue for effective methods.

3.6 Miscellaneous Issues

An important issue of discovery of interesting exception instances is to provide explanation of the discovered instances because the user would be perplexed given a long list of instances. [Knorr and Ng 1999] attempts to provide intensional knowledge, by which the authors mean a description or an explanation of why an identified outlier is exceptional. They extend their problem of detecting $DB(p, r)$ outliers to discover also the subspace of the example space and classify outliers into strongest, weak, and trivial. Intuitively, a strongest outlier resides in a space of which subspaces contain no outliers. Thus, they introduced the notion of dominance among outliers. A weak outlier is not an outlier in any of its subspaces but is not a strongest outlier. A trivial outlier is also an outlier in at least one of its subspaces and thus is considered uninteresting. The explanation of an outlier corresponds to the subspace and the kind of the outlier. Note that this method is related with the projection-based approach in Section 3.4 though it goes beyond identification to explanation.

Rules learned from a set of outliers can be interesting and might deepen our understanding of multi-database mining by providing possible explanation of the outliers. Zhong et al. propose peculiar rules, each of which is learned from a relatively small number of peculiar values[1] [Zhong et al. 1999, Zhong et al. 2003]. Roughly speaking, a value is peculiar if it represents a peculiar case described by a relatively small number of examples and is very different from those of other examples in a data set. For instance, in Japan, the areas of arable land and forest are very large in Hokkaido (1209Kha and 5355 Kha, respectively) but very small in Tokyo (12 Kha and 18Kha, respectively) and in Osaka (80 Kha and 59 Kha, respectively). These values are very different from other values in the attributes and thus are regarded as the peculiar value. From these peculiar values, peculiar rules such as the following are learnt in the method:

```
ArableLand(large) & Forest(large) -> PopulationDensity(low)
ArableLand(small) & Forest(small) -> PopulationDensity(high)
```

[1] The original expression in [Zhong et al. 1999, Zhong et al. 2003] is “peculiar data”.

Finding the peculiar values is based on conceptual distance between two attribute values. They use a threshold that is defined with the mean and the variance for each attribute in the process. The RVER (Reverse Variant Entity-Relationship) model is used to represent the peculiar values and the conceptual relationships among the peculiar values discovered from multiple databases. In [Zhong et al. 1999], experiments with Japan-survey, web-log, weather, and supermarket data show interesting examples of peculiar rules.

Roughly speaking, the methods presented so far model "normal" examples by extracting information from given data. However, when looking for an outlier in a data set, it often happens that a qualitative description on normal behavior is available. Angiulli et al. propose a complementary method that exploits such descriptions given in the form of logical rules to define "normal" examples [Angiulli et al. 2004]. As an example, suppose a bank approves a loan request that amounts greater than 50K Euro if an endorsement is provided as a guarantee by a reliable third party. This constraint can be encoded using a logical rule like the following:

```
Approved(L)
      <- ReqLoan(L,C,A), A > 50K, Endorsement(L,P), Reliable(P).
```

Here Approved(Loan ID), ReqLoan(Loan ID, Customer, Amount), Endorsement(Loan ID, Party), and Reliable(Party) represent records an approved loan request, a loan request, an endorsement, and the guaranteeing party, respectively. If a loan request l_1 that is approved by an unreliable party p_1 is approved, their method views p_1 as an outlier because it violates the logical rule.

Angiulli et al.'s method represents a subjective method for outlier detection because it is dependent on domain knowledge provided in the form of a set of logical rules. They analyze the time complexity of their method and show that most of their operations are intractable. This fact shows that it is required to design efficient heuristics for practically solving detection problems in real cases. Their main contribution of [Angiulli et al. 2004] is an introduction of a novel approach. Validation in several fields such as bioinformatics, fraud detection, network robustness analysis, and intrusion detection remains future work.

4 Exception Rule Discovery

4.1 Overall View

In Artificial Intelligence (AI), which is an uppercategory of machine learning, rule is the most extensively studied pattern. A rule represents a local probabilistic tendency in D and can be represented as $y \rightarrow x$. Here, y and x are called a premise and a conclusion, respectively, and each of them specifies a subspace of the example space. For instance, $(a_1 = v_{1,1}) \vee (a_2 = v_{2,2}) \rightarrow (a_3 = v_{3,1}) \wedge (a_4 =$

$v_{4,4}$) is a rule. A rule can be classified into either logical or probabilistic, and this paper is concerned with the latter. A probabilistic rule can have a confidence $\Pr(x|y)$ smaller than 1, i.e. $\Pr(x|y) < 1$, while a logical rule necessitates $\Pr(x|y) = 1$. Note that here we use each of x and y to represent a set of examples that reside in the corresponding subspace. In the rest of this paper, we use this notation.

Since a rule is represented as a combination of attribute values, there are $\Omega(m3^{m-1})$ kinds of rules. Finding interesting rules has been an important research topic in data mining. Since interestingness is often related with unexpectedness, exception rules are considered as promising candidates of interesting rules. As in Section 2, we skip a precise definition of an exception rule. Even before the proliferation of data mining research, AI had various methods for handling exception rules such as circumscription [McCarthy 1980]. At that time, however, the main interest was not put on efficient discovery.

In the reminder of this section, we will overview mainly two approaches devoted to discovery of interesting exception rules. An objective approach employs only given data, while a subjective approach also employs information supplied by the user. We will also explain integrated methods that exploit other learning and/or discovery methods.

4.2 Objective Approach

EXPLORA is a versatile data mining system that supplies various functionalities including a rule searcher, a change detector, and a trend detector [Hoschka and Klösgen 1991]. Its "subgroup discovery" capability, in its most fundamental form, represents one of the earliest objective approaches for discovery of interesting exception rules. They consider the following template to characterize a discovered pattern: "**Target group** shows **outstanding behavior** within **population** for **subpoplulation** in **year**". For instance, an instance of this type of template is given by "**Persons having initiated an insurance policy at company A** are **highly overrepresented** within **the clients of A** for **high-income people in the South** in **1987**". Note that searching for a group that exhibits an outstanding behavior in its mother population is equivalent to discovering an exception rule. To represent an "outstanding behavior", the authors propose several possibilities such as "overrepresented", "highly overrepresented", and "strong increase in the percentage of target objects". The former two possibilities are defined in terms of $(p - p_0)/s > 3$ and $(p - p_0)/s > 5$, where p, p_0, and s represent the percentage of target objects in the subpopulation, the percentage of target objects in the population, and the standard deviation for p, respectively. Alternate evaluation functions were proposed in [Klösgen 1996].

Subgroup discovery can be stated as discovery of interesting subgroups that deviate from their mother population. Various extensions and applications have been proposed after EXPLORA. They include application to multi-relational data and sampling [Wrobel 1997], application to a medical domain and heuristic search [Gamberger and Lavrač 2002a, Gamberger and Lavrač 2002b]. Typically, such a method employs domain knowledge and/or assumes user guidance in search due to the expensive nature of the search process. In that sense, these methods can be regarded as belonging to the subjective approach or the integrated approach in reality.

The idea of discovering rules each of which deviates from its mother population can be found in many methods though in general they are not regarded as discovery of interesting exception rules. For instance, ITRULE algorithm employs an information-theoretic evaluation function to measure the degree of interestingness of a rule [Smyth and Goodman 1992]. The function measures the amount of information compressed by a rule $y \rightarrow x$ and thus essentially depends on the difference of the code lengths $-\log_2 P(x)$ and $-\log_2 P(x|y)$ of x without and with $y \rightarrow x$, respectively. We will explain this function in Section 5.2. Another example can be found in discovery of quantitative association rules, where the conclusion represents a mean or a standard deviation of a target quantitative attribute [Aumann and Lindell 1999]. The methods searches for rules each of which exhibits a deviating value in the conclusion to the mother population.

The objective approach is free from overlooking useful knowledge due to an inapropriate use of domain knowledge. Moreover it can be applied to problems in which few or no domain knowledge exists. These advantages are realized at the sacrifice of search efficiency, which can be prohibitive for large-scale problems. It should be noted that the outcome of the objective approach is often unexpected but happens to be uninteresting.

4.3 Subjective Approach

To discriminate the subjective approach from the objective approach, we first explain [Sarawagi et al. 1998], which serves as a support system for On-Line Analytical Processing (OLAP). OLAP software helps analysts and managers gain insight of the database by accepting queries from them and returning a snapshot of the database as an OLAP cube. The process is interactive and often referred as data exploration, in which analysts and managers exploit a set of hierarchies each of which is associated with an attribute. They perform operations such as drill-down (zooming into more detailed levels of hierarchies), roll-up (zooming out to less detailed levels), and selection (choosing a subset of dimension members). [Sarawagi et al. 1998] proposes a discovery-driven method of data exploration, which colors cells (i.e. regions defined by a set of attribute=value's) of interest using three kinds of surprise value:

- SelfExp: represents the surprise value of the cell relative to other cells at the same level of aggregations.

- InExp: represents the degree of surprise somewhere beneath this cell if we drill down from the cell.
- PathExp: represents the degree of surprise for each drill-down path from the cell.

The degrees of surprise are defined with means and standard deviations as in EXPLORA. We classify this method to the subjective approach because user's guidance and the attribute hierarchies are mandatory while objective approach methods in the previous section do not necessarily require such subjective information.

Silberschatz and Tuzhilin proposed evaluation functions for subjective interestingness of a discovered pattern [Silberschatz and Tuzhilin 1996]. In this work, they mainly consider "unexpectedness" and try to evaluate it assuming that the user describes pieces of knowledge as a set of beliefs. Beliefs are classified into "soft beliefs" each of which degree is subject to change according to discovered patterns and "hard beliefs" each of which always holds. A degree of a soft belief represents the extent that the user believes the belief. Such degrees have several candidates including conditional probabilities in the Bayesian approach and certainty factors in the Dempster-Shafer theory. Consider the case of being supplied a new fact E when another fact ξ supports a soft belief α, then the degree $P(\alpha|E,\xi)$ associated with α is updated based on Bayes' theorem in the Bayesian approach:

$$P(\alpha|E,\xi) = \frac{P(E|\alpha,\xi)P(\alpha|\xi)}{P(E|\alpha,\xi)P(\alpha|\xi) + P(E|\overline{\alpha},\xi)P(\overline{\alpha}|\xi)}.$$

The interestingness of a discovered rule r is represented by the degree of influence to the set of beliefs. If r contradicts to a set of hard beliefs, r is regarded as interesting. If r contradicts to a set B of soft beliefs, the interestingness $I(r,B)$ of r is given as follows:

$$I(r,B) \equiv \sum_{\alpha \in B} \omega_i |P(\alpha|r,\xi) - P(\alpha|\xi)|,$$

where ω_i represents a normalized weight for each belief α. Subjective measure of interestingness, though no rule discovery algorithm is considered, represents a general pioneering work for discovering exception rules based on domain knowledge.

Later Tuzhilin together with Padmanabhan proposed a method for discovering unexpected rules [Padmanabhan 1998]. Let XA represent $X \wedge A$, given a belief $X \rightarrow Y$, the method first discovers all rules $XA \rightarrow B$ each of which satisfies the conditions for association rules (support $\Pr(XAB)$ and confidence $\Pr(B|XA)$ are greater than their respective thresholds [Agrawal et al. 1996]) and

B contradicts to Y. Next the method obtains more general and more unexpected rules $X'A \rightarrow B$ by generalizing X to X'. For instance, given a belief "professional → weekend" (i.e. a professional tends to go shopping during weekends rather than weekdays), the method might discover "December $\wedge$ professional → weekday" (i.e. a professional tends to go shopping during weekdays rather than weekends in December), then "December → weekday" (i.e. one tends to go shopping during weekdays rather than weekends in December). This method has been extended to discover the minimal set of unexpected patterns relative to given beliefs [Padmanabhan 2000]. These methods can be regarded as logic-based as they depend on a binary relation called logical contradiction. A logic-based method can be complementary to other statistics-based methods, and empirical evidence on the synergy of the two approaches should be investigated.

In the previous methods, domain knowledge is used only in a relatively simple manner. Liu et al. have proposed a method that ranks rules according to their degrees of interestingness based on fuzzy matching as a post-processing of rule discovery [Liu et al. 1999a]. Domain knowledge is given as a set of rules each of which is associated with fuzzy patterns in this framework and the similarity and dissimilarity of a discovered rule to the rules is obtained by fuzzy matching. Intuitively, the similarity between a pair of rules is defined as a combination of similarities between the corresponding pair of the premises and the corresponding pair of the conclusions. Similarly, a pair of rules are judged dissimilar either when (1) the premises are similar but the conclusions are different; (2) the conclusions are identical, the attributes in the premises are the same, but their values are different; or (3) the conclusions are identical but the attributes in the premises are different. Since there are multiple similarities and dissimilarities, there are several rankings of rules. In this method, a rule that is dissimilar to domain knowledge can be regarded as an exception rule by its definition. As Liu et al. admit, describing domain knowledge as a set of rules with fuzzy patterns is a difficult task for the user.

In order to remedy this difficulty, Liu et al. have proposed a language in which a user can express domain knowledge as "impressions" and an algorithm that evaluates the interestingness of a discovered rule based on a set of impressions [Liu et al. 1997]. Similar to their previous method, this method serves as a post-processing of rule discovery. Contrary to their previous method that employs relatively concrete domain knowledge such as "monthly salary $\geq$ 5,000 \$ → loan = approved", this method employs abstract domain knowledge such as "high monthly salary often implies loan approval". A user can employ impressions including the following formats, where C_j, a, and C_{sub} represent a class, an attribute, and a set of classes, respectively:

1. $a <\!\!\rightarrow C_j$: if the value of a is small then C_j is likely to occur.
2. $a >\!\!\rightarrow C_j$: if the value of a is large then C_j is likely to occur.

3. $a <<\rightarrow C_j$: if the value of a is within a certain range then C_j is likely to occur.

4. $a| \rightarrow C_{sub}$: there is a relation between a and C_{sub}.

5. $a[S] \rightarrow C_j$: if the value of a is an element of a set S then C_j is likely to occur.

It is possible to specify an impression that involves multiple attributes. For instance, an impression "savings $>$ $\wedge$ age $<<$ $\rightarrow$ loan = approved" represents that the loan is likely to be approved if the value of the saving attribute is large and the value of the age attribute is within a certain range. Furthermore, a user can specify an impression in which only a part of its conditions hold. Discovered rules are ranked according to the result of their matching to a set of impressions. Similarly to the fuzzy matching method several kinds of rankings exist and a rule that violates a set of impressions can be regarded as an exception rule. The specification of domain knowledge in an abstract level seems promising as it tolerates problems called the bottleneck of knowledge acquisition.

The problem of interactions among user-supplied domain knowledge has been an important issue in AI. A logic-based approach is typically brittle to noisy data while a heuristic approach lacks of theoretical foundations. A Bayesian network, on the other hand, assumes noise in data and has a solid theoretical foundation. Jaroszewicz et al. assumed that the user supplies the domain knowledge as a Bayesian network and proposed a discovery method for finding unexpected patterns in data relative to the Bayesian network [Jaroszewicz and Simovici 2004, Jaroszewicz and Scheffer 2005]. In the methods, the degree of interestingness of an itemset is defined as the absolute difference between its supports estimated from data and from the Bayesian network. Note that Bayesian networks have an advantage of representing full joint probability distributions, allowing for practically feasible probabilistic inference from those distributions. They developed efficient algorithms for evaluating the degree of interestingness of a collection of frequent itemsets, for finding all attribute sets with a given minimum degree of interestingness, and for finding an approximately most interesting unexpected patterns. The method seems to be effective as it tackles the problem of interactions among user-supplied domain knowledge in a general and solid manner. Questions remain on the availability of a Bayesian network, possible remedies when the Bayesian network is wrong, and necessity of defining a full joint probability distribution of the data.

The pros and cons of the subjective approach are the reverse of those of the objective approach in the previous section. We believe that a specification of domain knowledge in an abstract level and a use of inference such as Bayesian network are promising. Another avenue will be tight-coupling the rule discovery with other steps of the KDD process model. Versatile systems such as EXPLORA

[Hoschka and Klösgen 1991,Klösgen 1996] and DM-II [Liu et al. 1999b SIGKDD] as well as query languages such as Rule-QL [Tuzhilin and Liu 2002] belong to this line of research.

5 Exception Structured-rules Discovery

5.1 Overall View

We overviewed methods devoted to discovery of interesting exception rules in the previous section. In such a method, a discovered pattern is represented by an exception rule, which means that the discovered pattern does not include the base of the exception rule. Here we use the term "base" as a pattern that represents a normal behavior and thus gives an explanation on the exceptionality of its corresponding exception rule. For instance, the fact that a jobless person is rarely issued a credit card can be viewed as the base on an exception rule `occupation=jobless and asset=high -> credit card=issued`. Obviously, finding exception rules with their bases is a more time-consuming task than finding exception rules.

In AI, exceptions have long attracted attention of researchers. Before proliferation of data mining research, however, efficient systematic discovery of both exceptions and their bases was prohibitive due to various reasons, mainly due to the limited capability of hardware. We believe that our work on rule pair discovery in Section 5.2 represents one of the earliest methods for such discovery. A rule pair consists of an exception rule and its strong rule, which represents the base of the exception rule, this thus can be regarded as structured-rules. In this section, we explain methods for discovery of interesting exception structured-rules by overviewing our rule pair discovery, several modifications related to break of monotonicity, and our systematic search guided by a meta pattern.

5.2 Rule Pair

We have invented the rule pair and proposed several methods for discovering a set of rule pairs [Suzuki and Shimura 1996,Suzuki 1996,Suzuki 1997,Suzuki and Kodratoff 1998,Suzuki and Żytkow 2000,Suzuki 2002]. Let an atom be an event representing a value assignment or a range assignment to an attribute. A discovered pattern is represented by a rule pair $r(x,\ x',\ Y,\ Z)$.

$$r(x,\ x',\ Y,\ Z) \equiv (Y \rightarrow x,\ Y \wedge Z \rightarrow x') \qquad (6)$$

where each of Y and Z represents a conjunction of atoms, and each of x and x' represents an atom. Here x and x' have the same attribute but different values. We call $Y \rightarrow x$ a strong rule, $Y \wedge Z \rightarrow x'$ an exception rule, and $Z \rightarrow x'$ a reference rule.

Smyth, in his rule discovery system ITRULE, has proposed the quantity $J(x;\ y)$ of information compressed by a rule $y \rightarrow x$ as a measure J of interestingness [Smyth and Goodman 1992].

$$J(x;\ y) = \Pr(y)\ j(x;\ y) \tag{7}$$

$$\text{where } j(x;\ y) = \Pr(x|y)\log_2 \frac{\Pr(x|y)}{\Pr(x)} + \Pr(\overline{x}|y)\log_2 \frac{\Pr(\overline{x}|y)}{\Pr(\overline{x})} \tag{8}$$

We defined our measure of interestingness of a rule pair as a product of the J-measure of a strong rule and the J-measure of an exception rule [Suzuki and Shimura 1996]. We derived an upper bound of our measure and proposed an efficient algorithm that performs a branch-and-bound search based on it. We introduced several probabilistic constraints, since when $\Pr(x'|Z)$ is large, our exception rule exhibits low unexpectedness [Suzuki 1996]. We, together with Kodratoff, have considered unexpectedness from a different perspective and proposed a novel probabilistic criterion, which mainly considers the number of counter-examples [Suzuki and Kodratoff 1998].

We then proposed a method in which we specify thresholds $\theta_1^{\mathrm{S}},\ \theta_1^{\mathrm{F}},\ \theta_2^{\mathrm{S}},\ \theta_2^{\mathrm{F}},\ \theta_2^{\mathrm{I}}$ for probabilistic criteria of a rule pair. Since a rule pair discovered from 10,000 examples exhibits different reliability from another rule pair discovered from 100 examples, it is inappropriate to use a ratio $\widehat{\Pr}(\cdot)$ in a data set as a probabilistic criterion. Therefore, we considered a true probability $\Pr(\cdot)$ for each probabilistic criterion, and obtained a set of rule pairs each of which satisfies discovery conditions with the significance level δ [Suzuki 1997, Suzuki 2002]. In the following, $\mathrm{MIN}(x, y)$ represents the smaller one of x and y:

$$\begin{aligned}\Pr[\quad & \Pr(Y) \geq \theta_1^{\mathrm{S}},\ \Pr(x|Y) \geq \mathrm{MAX}(\theta_1^{\mathrm{F}}, \widehat{\Pr}(x)),\ \Pr(YZ) \geq \theta_2^{\mathrm{S}}, \\ & \Pr(x'|YZ) \geq \mathrm{MAX}(\theta_2^{\mathrm{F}}, \widehat{\Pr}(x')), \Pr(x'|Z) \leq \mathrm{MIN}(\theta_2^{\mathrm{I}}, \widehat{\Pr}(x'))\ ] \geq 1 - \delta. \end{aligned} \tag{9}$$

Confirming (9) for each rule pair numerically is time-consuming because (9) contains five true probabilities. Our method overcomes these difficulties by obtaining analytical solutions based on simultaneous reliability estimation of true probabilities. We have also proposed an efficient discovery algorithm based on pruning.

We briefly explain our endeavor with our method in a data mining contest with the meningitis data set [Suzuki and Tsumoto 2000]. The data set consists of 140 patients each of whom is described by 38 attributes and has been made public as a benchmark problem to the data mining community. Our method has discovered 169 rule pairs from a pre-processed version of this data set. These rule pairs were inspected by Dr. Tsumoto, who is a domain expert, and each rule pair was assigned a five-rank score for each of the following evaluation criteria:

- Validness: the degree that the discovered pattern fits the domain knowledge.

- Novelty: the degree that the discovered pattern does not exist in the domain knowledge.
- Usefulness: the degree that the discovered pattern is useful in the domain.
- Unexpectedness: the degree that the discovered pattern partially contradicts the domain knowledge.

For the scores, five and one represent the best score and the worst score, respectively. We show the results classified by the attributes in the conclusions in Table 1. From the table, we see that the average scores of the discovered rule pairs are high for several attributes in the conclusions. As Dr. Tsumoto admits, this result is considered to come from the structure of a rule pair, which seems to be useful for discovery of interesting patterns.

Table 1: Average performance of the proposed method with respect to attributes in the conclusion. The column "#" represents the number of discovered rule pairs.

attribute	#	validness	novelty	unexpectedness	usefulness
(all)	169	2.9	2.0	2.0	2.7
CULTURE	2	1.0	1.0	1.0	1.0
C_COURSE	1	1.0	1.0	1.0	1.0
RISK	1	1.0	1.0	1.0	1.0
CT_FIND	36	3.3	3.0	3.0	3.2
EEG_FOCUS	11	3.0	2.9	2.9	3.3
Course (G)	8	1.8	2.0	2.0	1.8
FOCAL	18	3.1	2.2	2.7	3.0
LOC_DAT	11	2.5	1.8	1.8	2.5
Diag2	72	3.0	1.1	1.1	2.6
CULT_FIND	4	3.3	4.0	4.0	3.5
KERNIG	4	2.0	3.0	3.0	2.0
SEX	1	2.0	3.0	3.0	2.0

Our method has been also applied to a 1994 bacterial test data set (20,919 examples, 135 attributes, 2 classes) [Suzuki 2000b]. We have found that discovery of interesting patterns from the data set requires further pre-processing that considers distribution of attribute values and cause and effect relationships. However, this application shows that our method is adequate in terms of efficiency in exception rule mining from a relatively large-scale data set.

5.3 Break of Monotonicity

If an exception rule is represented by $yz \rightarrow x$, the value of $\Pr(x|yz)$ typically differs from those of $\Pr(x|y)$ and $\Pr(x|z)$ considerably. For instance when the value of $\Pr(x|yz)$ is nearly 1, the values of $\Pr(x|y)$ and $\Pr(x|z)$ are often nearly 0 or $\Pr(x)$. In such a case the exception rule $yz \rightarrow x$ is said to break the monotonicity of the rules $y \rightarrow x$ and $z \rightarrow x$.[2]

Okada in his rule induction method, which he calls a cascade model, extends this idea to see the distribution of all values $v_{j,1}, v_{j,2}, \ldots, v_{j,|a_j|}$ associated with the attribute a_j in a conclusion [Okada 1999,Okada 2000]. The discovered pattern is similar to our rule pair but he considers a squared sum of the differences of occurrence probabilities of the values as an evaluation function. Thus, in this framework the conclusion goes beyond a specification of a value to an attribute and corresponds to a specification of a distribution of values to the attribute. We think that there surely exist applications that are appropriate for this type of rules, although [Okada 1999, Okada 2000] seem to be based on theoretical motivations.

Liu et al. have proposed a method that discovers a set of interesting rules based on statistical tests [Liu et al. 1999b]. They modified the confidence condition $\widehat{\Pr}(x|y) \geq \theta_{\rm c}$ of association rule $y \rightarrow x$ discovery to an existence of correlation of x and y based on a χ^2 test. A rule that also satisfies the other support condition $\widehat{\Pr}(xy) \geq \theta_{\rm s}$ is classified into one of the following three categories:

1. Positive correlation (direction 1): if x and y are correlated, and $\widehat{\Pr}(xy) /(\widehat{\Pr}(x)\widehat{\Pr}(y)) > 1$.
2. Negative correlation (direction -1): if x and y are correlated, and $\widehat{\Pr}(xy) /(\widehat{\Pr}(x)\widehat{\Pr}(y)) < 1$.
3. No correlation (direction 0): if x and y are not correlated.

This method deals with the problem of discovering interesting rules $yz \rightarrow x$ of which yz and x are positively correlated. If the directions of $y \rightarrow x$, $z \rightarrow x$, and $yz \rightarrow x$ are either of (1) 0, 0, 1, (2) -1, -1, 1, (3) -1, 0, 1, and (4) -1, 1, 1, the $yz \rightarrow x$ is considered to be possibly interesting. Liu et al. define a direction setting rule as a rule that satisfies one of the four conditions, and propose an algorithm for discovering a set of direction setting rules. Experiments using 30 data sets show that the method can reduce the number of discovered rules considerably. We consider that this method is deeply related to our rule pair discovery.

Yugami et al. proposed to discover a set of exception rules each of which shows a high confidence value, although any rules each of which premise is a subset

[2] The term "break of the monotonicity" is named by A. Tuzhilin.

of the premise do not show high confidence values [Yugami et al. 2000]. For instance, their method DIG (Discover Interesting rules with Grouping attribute values) discovers the following rule:

$$\begin{aligned} & \text{cap\_color} \in \{\text{brown}, \text{red}\} \wedge \text{stalk\_root} = \text{bulbous} \rightarrow \text{edible} : \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{confidence value } 100\%, \\ \text{where } & \text{cap\_color} \in \{\text{brown}, \text{red}\} \rightarrow \text{edible} : \text{confidence value } 50\%, \\ & \text{stalk\_root} = \text{bulbous} \rightarrow \text{edible} : \text{confidence value } 51\%. \end{aligned}$$

As Yugami admits, his exception rule is related to our rule pair discovery but should be interpreted differently, since each atom in the premise has little influence in predicting the conclusion.

DIG employs a relative degree of the confidence of the rule compared to the case in which no interaction exists among conditions in the premise as a measure of interestingness. As we can see from the above example, the number of candidates of a rule is large since a premise of a rule is represented by a conjunction of "attribute $\in$ a set of values". By using its efficient algorithm, DIG obtains all rules each of which satisfies user-specified conditions on the number of conditions in the premise, support, confidence, and the degree of interestingness. DIG has been applied to the "mushroom" (8,124 examples, 22 attributes, 2 classes), "satimage" (6,435 examples, 36 attributes), and "letter recognition" (20,000 examples, 16 attributes) data sets in the UCI Machine Learning Repository [Blake 1999]. The results show that DIG can discover rules each of which is at least interesting from the statistical viewpoint in a practical time.

We note that detecting a break of monotonicity without giving the normal behavior corresponds to exception structured-rules discovery. It is obvious that each method has applications that are especially suited for the method. We raise the question of how many kinds of methods exist, and the following section shows an attempt to answer this question.

5.4 Meta Pattern for Promising Patterns

Żytkow and Suzuki classified exception structured-rules for discovery of interesting patterns based on a meta pattern and proposed an efficient algorithm that discovers all structured-rules [Suzuki and Żytkow 2000, Suzuki and Żytkow 2005]. This work represents a systematic generation of all exception structured rules. In the approach, a discovered pattern is defined based on a rule triple $t(y,\ x,\ \alpha,\ \beta,\ \gamma,\ \delta)$, which represents the meta pattern, using literals $x,\ y,\ z$. A strong rule, an exception rule, and a reference rule are defined as $y \rightarrow x$, $\alpha \not\rightarrow \beta$, and $\gamma \rightarrow \delta$, respectively:

$$t(y,\ x,\ \alpha,\ \beta,\ \gamma,\ \delta) = (y \rightarrow x,\ \alpha \not\rightarrow \beta,\ \gamma \rightarrow \delta), \tag{10}$$

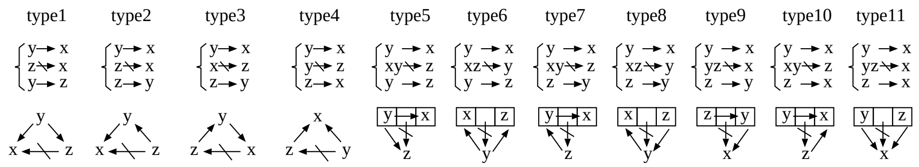

Figure 3: Classification of the rule triples. A rectangle on the top center for a rule triple represents a conjunction of literals in the top right and left. For instance, the three rectangles in type 11 represent, from the left to the right, "y", "$y \wedge z$", and "z".

where each of α, β, γ, and δ represents a meta variable that is instantiated by variables x, y, and z, resulting a definition of various exception structured-rules. Here $y \rightarrow x$ represents a rule and shows that $\widehat{\Pr}(y)$ and $\widehat{\Pr}(x|y)$ are greater than their respective thresholds. On the other hand, $\alpha \not\rightarrow \beta$ represents a negative rule and shows that $\widehat{\Pr}(\alpha)$ is greater than its threshold, and $\widehat{\Pr}(\beta|\alpha)$ is smaller than its threshold.

Under appropriate assumptions, the discovered patterns can be classified into the eleven structures which are shown in Figure 3. The algorithm efficiently searches rule triples with pruning. Experiments using 15 UCI data sets show that the pruning method is effective and the kinds of exception structured-rules which seem interesting (types 2, 5, 8, 9, and 11) are rarely discovered. Our intuition is that the rareness has some connection to the degree of interestingness. A formal analysis, however, seems to require strong assumptions.

6 Conclusions

The interests in the AI community has shifted from study of intelligence itself to desiderata on intelligent behaviors. Discovery of interesting patterns, as it is deeply related to the interests, represents an important research avenue in AI. From another viewpoint, discovery is at the same time a highly intellectual and rewarding activity. We believe that this research issue will keep on attracting attention especially around exception discovery .

In data mining, what is important is not each individual discovery but clarification and realization of principles of discovery (cf. http://www.cs.uvm.edu/~icdm/10Problems/10Problems-05.pdf). We anticipate further development of pattern representation, evaluation function, and their integration with other steps of the KDD process model toward a unifying theory of discovery of interesting exceptions.

Acknowledgments

This work was partially supported by the grant-in-aid for scientific research on fundamental research (B) 18300047 from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

[Aggarwal and Yu 2001] Aggarwal, C. C., Yu, P. S.: "Outlier Detection for High Dimensional Data", *Proc. 2001 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD)* (2000), 37–46.

[Aggarwal 2001] Aggarwal, C. C.: "Re-designing Distance Functions and Distance-Based Applications for High Dimensional Data", *SIGMOD Record, 30, 1* (2001), 13–18.

[Agrawal et al. 1996] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I.: "Fast Discovery of Association Rules," *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Menlo Park, Calif. (1996), 307–328.

[Angiulli et al. 2004] Angiulli, F., Greco, G., Palopoli, L.: "Detecting Outliers via Logical Theories and its Data Complexity", *Discovery Science, Lecture Notes in Artificial Intelligence 3245 (DS-2004)*, Springer-Verlag (2004), 101-113.

[Arning et al. 1995] Arning, A., Agrawal, R., Raghavan, P.: "A Linear Method for Deviation Detection in Large Databases", *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, AAAI Press, Menlo Park, Calif. (1996), 164–169.

[Aumann and Lindell 1999] Aumann, Y. and Lindell, Y.: "A Statistical Theory for Quantitative Association Rules", *Proc. Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (1999), 261–270.

[Bay and Schwabacher 2003] Bay, S. D., Schwabacher, M. : "Mining distance-based outliers in near linear time with randomization and a simple pruning rule", *Proc. Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2003), 29–38.

[Bennett et al. 1998] Bennett, C.H., Gács, P., Li, M., Vitányi, P., Zurek, W.: "Information Distance", *IEEE Trans. Inform. Theory, 44, 4* (1998), 1407–1423.

[Berka 2003] Berka, P. : ECML/PKDD 2003 Discovery Challenge, Download Data about Hepatitis, *http://lisp.vse.cz/challenge/ecmlpkdd2003/* (current April 26th, 2003).

[Beyer 1999] Beyer, K. S., Goldstein, J., Ramakrishnan, R., Shaft, U.: "When is "Nearest Neighbor" Meaningful?", *Proc. Seventh International Conference on Database Theory (ICDT)* (1999), 217–235.

[Blake 1999] Blake, C., Merz, C.J., Keogh, E.: "UCI Repository of Machine Learning Databases", *http://www.ics.uci.edu/~mlearn/MLRepository.html*, Univ. of Calif. Irvine, Dept. Information and CS (current May 5, 1999).

[Breunig et al. 2000] Breunig, M. M., Kriegel, H.-P., Ng, R. T., Sander, J. G: "LOF: Identifying Density-Based Local Outliers", *Proc. 2000 ACM SIGMOD International Conference on Management of Data* (2000), 93–104.

[Burge and Shawe-Taylor 1997] Burge, P. and Shawe-Taylor, J.: "Detecting Cellular Fraud Using Adaptive Prototypes", *AAAI-97 Workshop on AI Approaches to Fraud Detection and Risk Management* (1997), 1–8.

[Burge and Shawe-Taylor 2001] Burge, P. and Shawe-Taylor, J.: "Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection", *J. Parallel and Distributed Computing, 61* (2001), 915–925.

[Chakrabarti et al. 1998] Chakrabarti, S., Sarawagi, S., Dom., B.: "Mining Surprising Patterns using temporal description Length", *Proc. 24th VLDB Conf.* (1998), 606-617.

[Chan and Stolfo 1998] Chan, P. K., Stolfo, S. J.: “Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection” *Proc. Fourth Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, (1998), 164–168.

[Dumouchel et al. 1999] DuMouchel, W. et al.: ”Squashing Flat Files Flatter”, *Proc. Fifth ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD)* (1999), 6–15.

[Elkan 2001] Elkan, C. : “The Foundation of Cost-sensitive Learning”, *Proc. Seventeenth Intl. Joint Conf. on Artificial Intelligence (IJCAI)* (2001), 973–978.

[Ester et al. 1995] Ester, M., Kriegel, H.-P., Xu, X.: “A Database Interface for Clustering in Large Spatial Databases”, *Proc. First International Conference on Knowledge Discovery and Data Mining (KDD-95)* (1995), 94–99.

[Freund and Schapire 1996] Freund, Y., Schapire, R. E.: “Experiments with a New Boosting Algorithm”, *Proc. Thirteenth Int'l Conf. on Machine Learning (ICML)* (1996), 148–156.

[Fayyad et al. 1996] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P.: “From Data Mining to Knowledge Discovery: An Overview”, *Advances in Knowledge Discovery and Data Mining*, eds. U. M. Fayyad *et al.*, AAAI/MIT Press, Menlo Park, Calif. (1996), 1–34.

[Fawcett and Provost 1997] Fawcett, T., Provost, F.: “Adaptive Fraud Detection”, *Data Mining and Knowledge Discovery, 1, 3* (1997), 291–316.

[Frawley et al. 1991] Frawley, W. J., Piatetsky-Shapiro, G., Matheus, C. J.: “Knowledge Discovery in Databases: An Overview”, *Knowledge Discovery in Databases*, AAAI/MIT Press, Menlo Park, Calif. (1991), 1–27.

[Gamberger and Lavrač 2002a] Gamberger, D., Lavrač, N.: “Descriptive Induction through Subgroup Discovery: A Case Study in a Medical Domain”, *Proc. Nineteenth Int'l Conf. on Machine Learning (ICML)* (2002), 163–170.

[Gamberger and Lavrač 2002b] Gamberger, D., Lavrač, N.: “Generating Actionable Knowledge by Expert-guided Subgroup Discovery”, *Principles of Data Mining and Knowledge Discovery, LNAI 2431 (PKDD)*, Springer (2002), 163–174.

[Guha et al. 1998] Guha, S., Rastogi, R., Shim, K.: “CURE: An Efficient Clustering Algorithm for Large Databases” *Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD 1998)* (1998), 73–84.

[Hand et al. 2001] Hand, D., Mannila, H., Smyth, P.: *Principles of Data Mining*, MIT Press (2001).

[Hawkins 1980] Hawkins, D.M.: *The Identification of Outliers.* Chapman and Hall. London (1980).

[Hoschka and Klösgen 1991] Hoschka, P., Klösgen, W.: “A Support System for Interpreting Statistical Data”, *Knowledge Discovery in Databases*, AAAI/MIT Press, Menlo Park, Calif. (1991), 325–345.

[Jaroszewicz and Simovici 2004] Jaroszewicz, S., Simovici, D. A.: “Interestingness of Frequent Itemsets using Bayesian Networks as Background Knowledge”, *Proc. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2004), 178–186.

[Jaroszewicz and Scheffer 2005] Jaroszewicz, S., Scheffer, T.: “Fast Discovery of Unexpected Patterns in Data, Relative to a Bayesian Network”, *Proc. Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2005), 118–127.

[Jin et al. 2001] Jin, W., Tung, A. K. H., Han, J.: “Mining Top-n Local Outliers in Large Databases”, *Proc. Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2001), 293–298.

[Jumi et al. 2004] Jumi, M., Suzuki, E., Ohshima, M., Zhong, N., Yokoi, H., Takabayashi, K.: “Spiral Discovery of a Separate Prediction Model from Chronic Hepatitis Data”, *Proc. Third International Workshop on Active Mining (AM)* (2004), 1–10.

[Jumi et al. 2005] Jumi, M., Suzuki, E., Ohshima, M., Zhong, N., Yokoi, H., Takabayashi, K.: "Multi-strategy Instance Selection in Mining Chronic Hepatitis Data", *Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence 3488 (ISMIS-2005)*, Springer-Verlag (2005), 475-484.

[Klösgen 1996] Klösgen, W.: "Explora: A Multipattern and Multistrategy Discovery Approach", *Advances in Knowledge Discovery and Data Mining*, eds. U. M. Fayyad *et al.*, AAAI/MIT Press, Menlo Park, Calif. (1996), 249–271.

[Knorr and Ng 1998] Knorr, E. M., Ng, R. T.: "Algorithms for Mining Distance-Based Outliers in Large Datasets", *Proc. 24th International Conference on Very Large Data Bases (VLDB'98)* (1998), 392-403.

[Knorr and Ng 1999] Knorr, E. M., Ng, R. T.: "Finding Intensional Knowledge of Distance-Based Outliers", *Proc. 25th International Conference on Very Large Data Bases (VLDB'99)* (1999), 211–222.

[Knorr et al. 2000] Knorr, E. M., Ng, R. T., Tucakov, V.: "Distance-Based Outliers: Algorithms and Applications", *VLDB J.*, 8, 3-4 (2000), 237–253.

[Lee et al. 1998] Lee, W., Stolfo, S. J., Mok, K. W.: "Mining Audit Data to Build Intrusion Detection Models", *Proc. Fourth Int'l Conf. on Knowledge Discovery and Data Mining (KDD)* (1998), 66–72.

[Liu et al. 1997] Liu, B., Hsu, W., Chen, S.: "Using General Impressions to Analyze Discovered Classification Rules", *Proc. Third Int'l Conf. on Knowledge Discovery and Data Mining (KDD)* (1997), 31–36.

[Liu et al. 1999a] Liu, B. *et al.*: "Finding Interesting Patterns Using User Expectations", *IEEE Trans. Knowledge and Data Eng., 11, 6* (1999), 817–832.

[Liu et al. 1999b] Liu, B., Hsu, W., and Ma, Y.: "Pruning and Summarizing the Discovered Associations", *Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)* (1999), 125–134.

[Liu et al. 1999c] Liu, B., Hsu, W., Ma, Y., and Chen, S.: "Mining Interesting Knowledge using DM-II", *Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)* (1999), 430–434.

[Liu et al. 2001] Liu, B., Ma, Y., Yu, P. S.: "Discovering unexpected information from your competitors' web sites", *Proc. Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2001), 144–153.

[McCarthy 1980] J. McCarthy: "Circumscription - a form of nonmonotonic reasoning", *Artificial Intelligence, 13*, (1980), 27–39.

[Mitchell 1999] T.M. Mitchell, Machine Learning and Data Mining, CACM **42** (1999), 31–36.

[Ng and Han 1994] Ng, R. T., Han, J.: "Efficient and Effective Clustering Methods for Spatial Data Mining", *Proc. 20th International Conference on Very Large Data Bases (VLDB'94)* (1994), 144–155.

[Okada 1999] Okada, T.: "Rule Induction in Cascade Model Based on Sum of Squares Decomposition", *Principles of Data Mining and Knowledge Discovery (PKDD), LNAI 1704*, Springer-Verlag (1999), 468–474.

[Okada 2000] Okada, T.: "Efficient Detection of Local Interactions in the Cascade Model", *Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence 1805 (PAKDD-2000)*, Springer-Verlag (2000), 193-203.

[Padmanabhan 1998] Padmanabhan, B., Tuzhilin, A.: "A Belief-Driven Method for Discovering Unexpected Patterns", *Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, AAAI Press, Menlo Park, Calif. (1998), 94–100.

[Padmanabhan 2000] Padmanabhan, B., Tuzhilin, A.: "Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns", *Proc. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2000), 54–63.

[Pitman 1979] Pitman, E.: *Some Basic Theory for Statistical Inference*, London, Chapman and Hall (1979).

[Quinlan 1993] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, San Mateo, Calif., Morgan Kaufmann (1993).

[Ramaswamy et al. 2000] Ramaswamy, S., Rastogi, R., Shim, K.: “Efficient Algorithms for Mining Outliers from Large Data Sets”, *Proc. 2000 ACM SIGMOD Int’l Conf. on Management of Data (SIGMOD)* (2000), 427–438.

[Sarawagi et al. 1998] Sarawagi, S., Agrawal, R., Megiddo, N.: “Discovery-Driven Exploration of OLAP Data Cubes”, *Proc. Sixth International Conference on Extending Database Technology (EDBT’98), LNCS 1377*, Springer (1998), 168–182.

[Schölkopf 2001] Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A. J., Williamson, R. C.: “Estimating the support of a high-dimensional distribution”, *Neural Computation, 13* (2001), 1443–1472.

[Silberschatz and Tuzhilin 1996] Silberschatz, A., Tuzhilin, A.: “What Makes Patterns Interesting in Knowledge Discovery Systems”, *IEEE Trans. Knowledge and Data Eng., 8, 6* (1996), 970–974.

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

[Spenke 2001] Spenke, M.: “Visualization and Interactive Analysis of Blood Parameters with InfoZoom”, *Artificial Intelligence in Medicine, 22, 2* (2001), 159–172.

[Sugaya et al. 2001] Sugaya, S., Suzuki, E., Tsumoto, S.: “Instance Selection Based on Support Vector Machine for Knowledge Discovery in Medical Database”, *Instance Selection and Construction for Data Mining*, H. Liu and H. Motoda (eds.), Kluwer, Norwell, Mass. (2001), 395–412.

[Suzuki and Shimura 1996] Suzuki, E., Shimura, M.: “Exceptional Knowledge Discovery in Databases Based on Information Theory”, *Proc. Second Int’l Conf. Knowledge Discovery and Data Mining (KDD)*, AAAI Press, Menlo Park, Calif. (1996), 275–278.

[Suzuki 1996] Suzuki, E.: “Discovering Unexpected Exceptions: A Stochastic Approach”, *Proc. Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery (RSFD)* (1996), 225–232.

[Suzuki 1997] Suzuki, E.: “Autonomous Discovery of Reliable Exception Rules”, *Proc. Third Int’l Conf. Knowledge Discovery and Data Mining (KDD)*, AAAI Press, Menlo Park, Calif. (1997), 259–262.

[Suzuki and Kodratoff 1998] Suzuki, E., Kodratoff, Y.: “Discovery of Surprising Exception Rules Based on Intensity of Implication”, *Principles of Data Mining and Knowledge Discovery, LNAI 1510 (PKDD)*, Springer (1998), 10–18.

[Suzuki and Tsumoto 2000] Suzuki, E., Tsumoto, S.: “Evaluating Hypothesis-Driven Exception-Rule Discovery with Medical Data Sets”, *Knowledge Discovery and Data Mining, LNAI 1805 (PAKDD)*, Springer, Berlin (2000), 208–211.

[Suzuki 2000a] Suzuki, E. (ed.): *Proc. Int’l Workshop of KDD Challenge on Real-world Data (KDD Challenge 2000)* (2000).

[Suzuki 2000b] Suzuki, E.: “Mining Bacterial Test Data with Scheduled Discovery of Exception Rules”, *Proc. Int’l Workshop of KDD Challenge on Real-world Data (KDD Challenge)*, Kyoto, Japan (2000), 34–40.

[Suzuki and Żytkow 2000] Suzuki, E., Żytkow, J. M.: “Unified Algorithm for Undirected Discovery of Exception Rules”, *Principles of Data Mining and Knowledge Discovery, LNAI 1910 (PKDD)*, Springer (2000), 169–180.

[Suzuki 2002] Suzuki, E.: “Undirected Discovery of Interesting Exception Rules”, *International Journal of Pattern Recognition and Artificial Intelligence, 16, 8* (2002), 1065–1086.

[Suzuki 2004a] Suzuki, E.: “Discovering Interesting Exception Rules with Rule Pair”, *Proc. ECML/PKDD-2004 Workshop W8 on Advances in Inductive Rule Learning* (2004), 163–178.

[Suzuki 2004b] Suzuki, E.: “Evaluation Scheme for Exception Rule/Group Discovery”, *Intelligent Technologies for Information Analysis*, Springer, Berlin (2004), 89–108.

[Suzuki and Żytkow 2005] Suzuki, E., Żytkow, J. M.: “Unified Algorithm for Undirected Discovery of Exception Rules”, *International Journal of Intelligent Systems,* *20, 7* (2005), 673–691.

[Tuzhilin and Liu 2002] Tuzhilin, A. and Liu, B.: "Querying Multiple Sets of Discovered Rules", *Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2002), 52–60.

[Wrobel 1997] Wrobel, S.: "An Algorithm for Multi-relational Discovery of Subgroups", *Principles of Data Mining and Knowledge Discovery (PKDD), LNCS 1263*, Springer-Verlag (1997), 78–87.

[Yamada et al. 2003] Yamada, Y., Suzuki, E., Yokoi, H., Takabayashi, K.: "Decision-tree Induction from Time-series Data Based on a Standard-example Split Test", *Proc. Twentieth International Conference on Machine Learning (ICML)*, (erratum http://www.slab.dnj.ynu.ac.jp/ erratumicml2003.pdf) (2003), 840–847.

[Yamanishi et al. 2000] Yamanishi, K., Takeuchi J., Williams, G. and Milne, P.: "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms", *Proc. Sixth ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD)* (2000), 320–324.

[Yamanishi and Takeuchi 2001] Yamanishi, K., Takeuchi, J.: "Discovering Outlier Filtering Rules from Unlabeled Data: Combining a Supervised Learner with an Unsupervised Learner", *Proc. Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2001), 389–394.

[Yamanishi and Takeuchi 2002] Yamanishi, K., Takeuchi, J.: "A Unifying Framework for Detecting Outliers and Change Points from Non-stationary Time Series Data", *Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2002), 676–681.

[Yugami et al. 2000] Yugami, N., Ohta, Y., Okamoto, S.: "Fast Discovery of Interesting Rules", *Knowledge Discovery and Data Mining, LNAI 1805 (PAKDD)*, Springer, Berlin (2000), 17–28.

[Zhang et al. 1996] Zhang, T., Ramakrishnan, R., Livny, M.: "BIRCH: An Efficient Data Clustering Method for Very Large Databases", *Proc. 1996 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD)* (1996), 103–114.

[Zhong et al. 1999] Zhong, N., Yao, Y. Y., Ohsuga, S.: "Peculiarity Oriented Multi-Database Mining", *Principles of Data Mining and Knowledge Discovery (PKDD), LNAI 1704*, Springer-Verlag (1999), 136–146.

[Zhong et al. 2003] Zhong, N., Yao, Y. Y., Ohshima, M.: "Peculiarity Oriented Multi-Database Mining", *IEEE Transaction on Knowledge and Data Engineering, 15, 4* (2003), 952–960.