Are We Really Discovering "Interesting" Knowledge From Data?

Alex A. Freitas
Computing Laboratory, University of Kent
Canterbury, CT2 7NF, UK
A.A.Freitas@kent.ac.uk
http://www.cs.kent.ac.uk/~aaf

Abstract

This paper is a critical review of the literature on discovering comprehensible, interesting knowledge (or patterns) from data. The motivation for this review is that the majority of the literature focuses only on the problem of maximizing the accuracy of the discovered patterns, ignoring other important pattern-quality criteria that are user-oriented, such as comprehensibility and interestingness. The word "interesting" has been used with several different meanings in the data mining literature. In this paper interesting essentially means novel or surprising. Although comprehensibility and interestingness are considerably harder to measure in a formal way than accuracy, they seem very relevant criteria to be considered if we are serious about discovering knowledge that is not only accurate, but also useful for human decision making. The paper discusses both data-driven methods (based mainly on statistical properties of the patterns) and user-driven methods (which take into account the user's background knowledge or believes) for discovering interesting knowledge. Data-driven methods are discussed in more detail because they are more common in the literature and are more controversial. The paper also suggests future research directions in the discovery of interesting knowledge.

1 Introduction

A well-known definition of knowledge discovery is as follows [Fayyad et al. 1996]:

"Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

Although this definition is often quoted in the literature, in general it has not been taken very seriously by the research community. This claim is supported by the fact that the vast majority of data mining works focus on discovering knowledge that is accurate – e.g., trying to maximize predictive accuracy in the classification task. This can be considered as aiming to discover valid patterns, and perhaps aiming at discovering "potentially useful" patterns – to the extent that we believe that there is a high positive correlation between the accuracy of a pattern and its usefulness to the user. However, in practice the correlation between predictive accuracy and usefulness of discovered patterns is not so clear, and the focus on maximizing predictive accuracy does not seem to improve the chances of discovering novel, ultimately understandable patterns in the data. Actually, it is often the case that focusing on maximizing predictive accuracy only – ignoring other criteria to evaluate the quality of patterns – significantly harms the discovery of understandable, novel and useful knowledge. A few examples can illustrate this point, as follows.

[Brin et al. 1997] found, in a Census dataset, several rules which were very accurate but were also useless, because they represented obvious patterns in the data, such as "five years olds don't work", "unemployed residents don't earn income from work" and "men don't give birth". [Tsumoto 2000] found 29,050 rules, out of which only 220 (less than 1% of them) were considered interesting or unexpected by the user. These two works are examples of the fact that high accuracy is not a sufficient condition for the usefulness or interestingness (novelty or surprisingness) of a pattern. In addition, high accuracy is not always a necessary condition for the usefulness or interesting of a pattern. For instance, [Wong & Leung 2000] found rules with 40-60% confidence that were considered, by senior medical doctors, novel and more accurate than the knowledge of some junior doctors.

The goal of this paper is to contribute to a better understanding of the limitations of the concepts and techniques used to discover comprehensible and interesting patterns. Hence, this paper can be considered a critical review of the literature on the discovery of comprehensible, interesting patterns. By "interesting" we mean novel or surprising. Note that we consider interestingness and comprehensi-

bility to be different quality criteria, since patterns such as "men don't give birth" are comprehensible but not interesting at all. Hence, this paper focuses on two out of the four pattern-quality criteria mentioned in Fayyad et al.'s definition. Concerning the other two criteria, we interpret "valid" essentially as "accurate", a pattern-quality criterion that is not discussed here because it is already extensively discussed in the literature; and we follow [Silberchatz & Tuzhilin 1996] in using surprisingness or novelty as a proxy for usefulness, because usefulness is a concept whose formalization seems elusive.

The remainder of this paper is organised as follows. Section 2 discusses the discovery of comprehensible (understandable) patterns. Section 3 discusses the discovery of interesting (novel or surprising) patterns. Section 4 presents the conclusions and future research directions.

2 On the Discovery of Comprehensible Patterns

In many application domains, in order for the user to trust the discovered patterns and make important decisions based on them, it is usually necessary that the user understand those patterns. For instance, in principle a medical doctor should not blindly trust the diagnosis output by a black box classification algorithm and recommend a surgery for the patient based just on that automatic diagnosis. The doctor should interpret the discovered patterns in the context of her/his previous knowledge about the application domain. Similarly, a user would probably hesitate in investing a large amount of money in a financial application based on some pattern automatically discovered by a black box prediction algorithm. In addition, in some applications the reason for a decision must be explained for legal reasons, which requires that the patterns on which the decision was based be understandable.

This does not mean that comprehensibility is always important. In principle the need for understandable patterns arises when the patterns will be used to support a decision to be made by a human user. In some applications the discovered patterns will be automatically used by a machine rather than support a human decision, and so they do not need to be understandable. A typical example is the pattern recognition task of automatically recognizing the post code in a letter and sending the letter to a pigeon hole containing letters for the appropriate destination.

In any case, in applications where a human user would like to make important, strategic decisions based on the discovered patterns, intuitively the comprehensibility of the discovered patterns improves the potential usefulness of those patterns – although of course just comprehensibility by itself is not guarantee that the patterns will be really useful to the user. Despite the importance of comprehensibility, there has been little progress towards techniques that improve the comprehensibility of discovered patterns. In general we can say that some knowledge representations lend themselves more naturally to comprehensible patterns than others. For instance, most researchers would agree that representations such as decision trees, IF-THEN rules or Bayesian networks tend to be more comprehensible than, say, neural networks or support vectors. However, as pointed out by [Pazzani 2000], there is no consensus on which of these representations is the most comprehensible in general, and there seems to be no cognitive psychology study comparing the comprehensibility of different representations from the point of view of human users. Pazzani also suggests some cognitive psychology-related criteria for evaluating pattern comprehensibility, such as the criterion that the pattern should be consistent with the user's prior knowledge, but there has been relatively little work in this area. In any case, note that although the criterion of consistency with prior knowledge tends to improve comprehensibility, intuitively it tends to hinder the discovery of novel or surprising patterns – see Section 3.

As for the usual measure of "comprehensibility" or "simplicity" often used in the literature, which consists of measuring the size (number of conditions or nodes) of a rule set, decision tree or Bayesian net, it should be noted that this is just a measure of syntactical simplicity, which is very different from semantic simplicity (which would need to involve the meaning of the attributes in the conditions or nodes of the rules, decision tree or Bayesian net). In any case, if a large number of patterns are discovered, one possibility to reduce the user's cognitive workload in interpreting the discovered patterns consists of selecting a subset of the most "interesting" (novel or surprising) patterns – using, for instance, some of the methods discussed in Section 3 – and show just those selected patterns to the user.

3 On the Discovery of Interesting (Novel or Surprising) Patterns

There are two basic approaches to discover novel or surprising (unexpected) patterns, namely the user-driven (or "subjective") approach and the data-driven (or "objective") approach. In essence, the user-driven approach is based on using the domain knowledge, believes or preferences of the user; whilst the data-driven approach is based on statistical properties of the patterns. Hence, the data-driven approach is more generic, independent of the application domain. This makes it easier to use this approach, avoiding difficult issues associated with the manual acquisition of the user's background knowledge and its transformation into a computational form suitable for a data mining algorithm. On the other hand, the user-driven approach tends to be more effective at discovering truly novel or surprising knowledge to the user, since it explicitly takes into account the user's background knowledge. This raises the question of to what extent the data-driven approach is effective in discovering interesting patterns to the user – an issue that will be discussed in subsection 3.2.

3.1 User-Driven Methods for Discovering Interesting Patterns

A classic example of user-driven method for discovering interesting patterns is the use of user-specified templates in the context of association rules [Klementtinen et al. 1994]. In this case the user essentially specifies inclusive templates – indicating which items the user is interested in (among a large number of items available in the database) – and restrictive templates – indicating which items the user is not interested in. Then an association rule is considered interesting if it matches at least one inclusive template and it matches no restrictive template.

Another example of user-driven method is the use of user-defined general impressions [Liu et al. 1997], [Romao et al. 2004]. In this case the user specifies general impressions in the form of IF-THEN rules, such as "IF (salary = high) AND (education_level = high) THEN (credit = good)". Note that this is a *general impression* because its conditions are not precisely defined. By contrast, the data mining algorithm is supposed to produce rules with well-defined conditions, such as "salary > £50K". Once such rules are produced by the data mining algorithm, the system can match the rules with the general impressions, in order to find surprising rules. In particular, if a rule and a general impression have similar antecedents ("IF part") but different consequents ("THEN part"), the rule can be considered surprising, in the sense of contradicting a user's belief (general impression). For instance, the rule "IF (salary > £50k) AND (education_level ≥ BSc) AND (Mortgage = yes) THEN (credit = bad)" would be considered surprising with respect to the aforementioned general impression.

3.2 Data-Driven Methods for Discovering Interesting Patterns

There are more than 50 measures of rule quality that have been called rule "interestingness" measures in the literature. A review of these measures can be found in [Hilderman & Hamilton 2001], [Tan et al. 2002]. One classical example of these data-driven rule interestingness measures is the one proposed by [Piatetsky-Shapiro 1991], defined as Interest = $|A \cap C| - (|A| \times |C|) / N$, where $|A \cap C|$ is the number of examples satisfying both the rule antecedent A and the rule consequent C, $|A|$ ($|C|$) is the number of examples satisfying the rule antecedent A (rule consequent C), and N is the total number of examples. Hence, Interest is a measure of the deviation from statistical independence between A and C. Note that it measures the symmetric correlation between A and C, and not an asymmetric implication, i.e., Interest has the same value for the two "opposite" rules: IF A THEN C, IF C THEN A.

Until a few years ago, in general works proposing data-driven rule interestingness measures implicitly assumed that such measures were correlated with the user's real, subjective interest in the rules, and typically papers using those measures did not report any subjective evaluation of the rules by the user. More recently, some works have reported the results of experiments to assess to what extent the values of data-driven rule interestingness measures are correlated with the real, subjective interest of the user. The methodology used for this assessment can be summarized in three steps, namely: (a) rank the discovered rules according to each of a number of data-driven rule interestingness measures; (b) show (a subset of) the discovered rules to the user, who assigns an "interestingness score" to each rule based on her/his subjective interest in the rule; and (c) measure the linear correlation (or another measure of association) between the ranking of each data-driven rule interestingness measure and the real, subjective human interest on the rules. A couple of experiments following the basic idea of this methodology are as follows.

[Ohsaki et al. 2004] have done experiments with 39 data-driven rule interestingness measures, involving rules discovered from a hepatitis dataset. They report the results of two experiments. In the first one the highest correlation between a rule interestingness measure (out of the 39 measures) and the user's real interest was just 0.48, and only one measure had a correlation greater than or equal to 0.4 (on a scale from –1 to +1). In the second experiment the highest correlation was again 0.48, and only four rule interestingness measures had a correlation greater than or equal to 0.4. (It should be noted, though, that the paper also reports other indicators of performance of the rule interestingness measures, according to which those measures seem to obtain better results.) [Carvalho et al. 2005] have done experiments with 11 data-driven rule interestingness measures, involving 8 datasets and one user for each dataset. Out of the 88 reported correlation values (involving 11 rule interestingness measures × 8 users), 31 correlation values were greater than or equal to 0.6. The correlation values associated with each measure varied considerably across the 8 datasets / users, so that no single rule interestingness measure performed consistently well across all datasets / users. In addition, more recent results reported in [Carvalho 2005], in experiments involving 45 users (9 datasets and 5 users per dataset), suggest that, overall, the correlation between data-driven measures and real human interest is considerably lower than the correlation results obtained with 8 users in [Carvalho et al. 2005].

The aforementioned results support the intuitive argument that it is difficult to use a purely data-driven approach for discovering patterns that are truly novel or surprising to the user. There are some works that try to reduce this strong limitation of the data-driven approach, using not only statistical properties of the rules but also concepts or ideas that intuitively seem more likely to lead to the discovery of interesting patterns – although the extent to which these ideas capture real human interest seems somewhat controversial. Let us now briefly review some of these works.

One approach consists of automatically learning which combination of a number of data-driven rule interestingness measures is a good predictor of real human interest, as proposed by [Abe et al. 2005]. This work involves a kind of "meta-learning", constructing a meta-dataset where each meta-example corresponds to a classification rule discovered from the a give dataset, the 39 predictor meta-attributes are values of 39 data-driven rule interestingness measures for each of the meta-examples (rules) and the class meta-attribute is the user's real, subjective interest in each of the rules. (So, this is a hybrid data/user-driven approach.) The values of the class meta-attribute are manually specified by the user in the meta-training set and automatically predicted by the algorithm in the meta-test set. The authors applied five different classification algorithms to the meta-dataset, and report that the best predictive accuracy – measured by leave-one-out – was 81.6%. This seems a good result, but it should be noted that different classification algorithms selected different meta-attributes for the classification model.

Another approach consists of using a data-driven rule interestingness measure that is "more surprisingness-oriented" than the mere use of statistical properties, in particular discovering exception rules, as follows. Let R_1 be a general rule of the form "IF $Cond_1$ THEN $Class_1$", and let R_2 be an exception rule of the form "IF $Cond_1$ AND $Cond_2$ THEN $Class_2$", where $Cond_1$, $Cond_2$ are conjunctions of conditions. Note that rule R_2 is a specialization of, and predicts a different class from, rule R_1. Hence, R_2 is an exception of R_1. In this kind of data-driven interestingness method, the exception rule R_2 can be considered an interesting rule if both R_2 and its generalized rule R_1 have a high predictive accuracy. Rule interestingness measures based on these ideas are discussed e.g. in [Suzuki & Kodratoff 1998], [Suzuki 2004]. The rationale for this exception-based approach is that users tend to know the general data relationships in their application domain, but are less likely to know exceptions to those general relationships. Hence, exception rules tend to be more surprising or novel to users than general rules. A real-world example involves car accident data [Suzuki 2004], where, in addition to the known general rule "IF (used_seat_belt = yes) THEN (injury = no)", the system also discovered the surprising exception rule "IF (used_seat_belt = yes) AND (passenger = child) THEN (injury = yes)".

Another surprisingness-oriented data-driven method consists of discovering instances of Simpson's paradox in data, as follows. Let the event *C* be the apparent "cause" of an event *E*, the "effect". Simpson's paradox occurs if the event *C* increases the probability of the event *E* in a given population *Pop* and, at the same time, decreases the probability of event *E* in *every* subpopulation of *Pop* [Pearl 2000]. Let Z and ¬Z denote two complementary values of a confounding variable, representing complemen-

tary properties describing two subpopulations of *Pop*. Then, mathematically Simpson's paradox occurs if the following 3 inequalities hold for a given data set:

$$P(E \mid C) > P(E \mid \neg C),\ P(E \mid C, Z) < P(E \mid \neg C, Z),\ P(E \mid C, \neg Z) < P(E \mid \neg C, \neg Z),$$

where $P(X \mid Y)$ denotes the conditional probability of X given Y.

A classic example of Simpson's paradox occurred in a comparison of tuberculosis deaths in New York City and Richmond, Virginia, in 1910. Overall, the tuberculosis mortality rate of Richmond was higher than New York's one. However, the opposite was observed when the data was partitioned according to two racial categories: white and non-white. In both the white and non-white categories, Richmond had a lower mortality rate. In this example, the events C and ¬C are Richmond and New York, the event E is tuberculosis death, and the events Z and ¬Z are the categories white and non-white. A number of other occurrences of the paradox in real-world data are reported in [Fabris & Freitas 1999], [Freitas & Fabris 2006], [Kohavi 2005]. The two works by Fabris & Freitas also describe algorithms that systematically search for occurrences of Simpson's paradox in data.

Although Simpson's paradox is well-known by statisticians, it is usually very surprising to data mining *users*, who typically have no formal statistical training. This makes the automatic detection of Simpson's paradox one of the few data-driven methods for discovering patterns that are likely to be considered surprising according to a user's subjective evaluation [McGarry 2005].

4 Conclusions and Future Research Directions

This paper presented a critical review of the current concepts and methods used for discovering comprehensible and interesting (novel or surprising) patterns in data. This is an important topic, because most works focus only on maximizing pattern accuracy (since accuracy is easier to measure), ignoring other aspects of pattern quality that, although harder to measure, are clearly related to the usefulness of the discovered patterns to the user.

We have discussed several methods for discovering interesting patterns, based on either a data-driven or a user-driven approach. The data-driven approach is normally easier to implement, but, since it does not take into account the user's domain knowledge, it has difficulty in discovering truly interesting knowledge to the user. In particular, recent results suggest that the effectiveness of a number of data-driven rule interestingness measures has been overrated in the literature. Three kinds of method that try to overcome some limitations of a data-driven approach based only on statistical properties of the data have been discussed, in particular: (a) a "meta-learning" method using a classification algorithm to learn which combination of data-driven rule interestingness measures best predicts the user's rule interest; and methods oriented towards the discovery of surprising patterns, namely: (b) the discovery of exception rules (which are less likely to be known by users than general rules); and (c) the discovery of instances of Simpson's paradox (which tend to be surprising to the user due to the nature of the "paradox"). However, even in the case of these methods there is not enough empirical evidence in the literature to show that they are effective in discovering patterns that are really interesting to the user, since most of the papers on these methods do not report the subjective evaluation of the discovered patterns by the user.

One research direction would be to try to significantly reduce the bottleneck of the user-driven approach, the manual acquisition of the user's background knowledge, by using text mining to automatically generate background knowledge about the application domain from the published literature. For instance, instead of asking the user to specify a comprehensive set of general impressions representing her/his background knowledge, in principle (at least in some application domains) a text mining algorithm could automatically extract general impressions from the literature. Presumably the user should still be in the loop to validate the general impressions discovered by the text mining algorithm, but intuitively it would be easier for the user to validate automatically-discovered general impressions than to specify a large number of general impressions herself/himself.

Another research direction would be to develop methods for discovering interesting patterns from the start of the KDD process – i.e. in the data preparation phase, rather than methods to be applied in the data mining phase or in the knowledge post-processing phase. For instance, current attribute selection methods in general are designed for maximizing the predictive accuracy of the data mining algo-

rithm, and those methods normally show no concern for the interestingness (novelty or surprisingness) of the patterns to be discovered by the data mining algorithm.

References

[Abe et al. 2005] H. Abe, S. Tsumoto, M. Ohsaki, T. Yamaguchi. Evaluating a rule evaluation support method with learning models based on objective rule evaluation indices. *Proc. 5th Int. Conf. Hybrid Intelligent Systems (HIS-2005)*, 169-174. PUC-Rio, Rio de Janeiro, Brazil, Nov. 2005.

[Brin et al. 1997] S. Brin, R. Motwani, J.D. Ullman, S. Tsur. Dynamic itemset counting and implication rules for market basket data. *Proc. KDD-97.* AAAI Press.

[Carvalho 2005] D.R. Carvalho. A decision tree / genetic algorithm to cope with the problem of small disjuncts. (In Portuguese). *Ph.D. Thesis*. Federal University of Rio de Janeiro, Brazil, Dec. 2005.

[Carvalho et al. 2005] D.R. Carvalho, A.A. Freitas, N.F. Ebecken. Evaluating the correlation between objective rule interestingness measures and real human interest. *Proc. PKDD-2005, LNAI 3721,* 453-461. Springer.

[Fabris & Freitas 1999] C.C. Fabris and A.A. Freitas. Discovering surprising patterns by detecting instances of Simpson's paradox. In: *Research and Development in Intelligent Systems XVI,* 148-160. Springer.

[Fabris & Freitas 2006] C.C. Fabris and A.A. Freitas. Discovering surprising instances of Simpson's paradox in hierarchical multi-dimensional data. *Int. J. on Data Warehousing & Mining*, *2(1)*, pp. 26-48.

[Fayyad et al. 1996] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery: an overview. *In: Advances in Knowledge Discovery and Data Mining*, 1-34. AAAI Press.

[Hilderman & Hamilton 2001] R.J. Hilderman and H.J. Hamilton. *Knowledge Discovery and Measures of Interest*. Kluwer.

[Kohavi 2005] R. Kohavi. Focusing the mining beacon: lessons and challenges from the world of e-commerce. *Invited talk at PKDD-2005*. www.kohavi.com. Visited on Jan. 2006.

[Klementtinen et al. 1994] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. *Proc. 3rd Int. Conf. on Information and Knowledge Management*, 401-407.

[Liu et al. 1997] B. Liu, W. Hsu, S. Chen. Using general impressions to analyze discovered classification rules. *Proc. KDD-97*, 31-36. AAAI Press.

[McGarry 2005] K. McGarry. A survey of interestingness measures for knowledge discovery. *Knowledge Engineering Review J.*, *20(1)*, 39-61.

[Ohsaki et al. 2004] M. Ohsaki, S. Kitaguchi, K. Okamoto, H. Yokoi, T. Yamaguchi. Evaluation of rule interestingness measures with a clinical dataset on hepatitis. *Proc. PKDD-2004*, 362-373. Springer.

[Pazzani 2000] M.J. Pazzani. Knowledge discovery from data? *IEEE Intellig. Sys.,* Mar/Apr. 2000, pp. 10-13.

[Pearl 2000] J. Pearl. *Causality: models, reasoning and inference*. Cambridge Univ. Press.

[Piatetsky-Shapiro 1991] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. *In: Knowledge Discovery in Databases*, 229-248. AAAI/MIT Press.

[Romao et al. 2004] W. Romao, A.A. Freitas, I.M.S. Gimenes. Discovering Interesting Knowledge from a Science & Technology Database with a Genetic Algorithm. *Applied Soft Computing 4*, 121-137.

[Silberchatz & Tuzhilin 1996] S. Silberchatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Trans. Knowledge and Data Engineering, 8(6).*

[Suzuki 2004] E. Suzuki. Discovering interesting exception rules with rule pair. *Proc. Workshop on Advances in Inductive Rule Learning at PKDD-2004*, 163-178.

[Suzuki & Kodratoff 1998] E. Suzuki and Y. Kodratoff. Discovery of surprising exception rules based on intensity of implication. *Proc. PKDD-98, LNAI 1510*, 10-18. Springer.

[Tan et al. 2002] P-N. Tan, V. Kumar, J. Srivastava. Selecting the right interestingness measure for association patterns, 32-41. *Proc. ACM SIGKDD KDD-2002*. ACM Press.

[Tsumoto 2000] S. Tsumoto. Clinical knowledge discovery in hospital information systems: two case studies. *Proc. PKDD-2000*, *LNAI 1910*, 652-656. Springer.

[Wong & Leung 2000] M.L. Wong and K.S. Leung. *Data mining using grammar based genetic programming and applications*. Kluwer.