Mining Data with Group Theoretical Means

Gabriele Kern-Isberner
University of Dortmund, Germany

INTRODUCTION

Knowledge discovery refers to the process of extracting new, interesting, and useful knowledge from data and presenting it in an intelligible way to the user. Roughly, knowledge discovery can be considered a three-step process: preprocessing data; data mining, in which the actual exploratory work is done; and interpreting the results to the user. Here, I focus on the data-mining step, assuming that a suitable set of data has been chosen properly.

The patterns that we search for in the data are plausible relationships, which agents may use to establish cognitive links for reasoning. Such plausible relationships can be expressed via association rules. Usually, the criteria to judge the relevance of such rules are either frequency based (Bayardo & Agrawal, 1999) or causality based (for Bayesian networks, see Spirtes, Glymour, & Scheines, 1993). Here, I will pursue a different approach that aims at extracting what can be regarded as structures of knowledge — relationships that may support the inductive reasoning of agents and whose relevance is founded on information theory. The method that I will sketch in this article takes numerical relationships found in data and interprets these relationships as structural ones, using mostly algebraic techniques to elaborate structural information.

BACKGROUND

Common sense and expert knowledge is most generally expressed by rules, connecting a precondition and a conclusion by an if-then construction. For example, you avoid puddles on sidewalks because you are aware of the fact that if you step into a puddle, then your feet might get wet; similarly, a physician would likely expect a patient showing the symptoms of fever, headache, and a sore throat to suffer from a flu, basing his diagnosis on the rule that if a patient has a fever, headache, and sore throat, then the ailment is a flu, equipped with a sufficiently high probability.

If-then rules are more formally denoted as conditionals. The crucial point with conditionals is that they carry generic knowledge that is applicable to different situations. This fact makes them most interesting objects in artificial intelligence, in a theoretical as well as in a practical respect. For instance, a sales assistant who has a general knowledge about the preferences of his or her customers can use this knowledge when consulting any new customer.

Typically, two central problems have to be solved in practical applications: First, where do the rules come from? How can they be extracted from statistical data? And second, how should rules be represented? How should conditional knowledge be propagated and combined for further inferences? Both of these problems can be dealt with separately, but it is most rewarding to combine them, that is, to discover rules that are most relevant with respect to some inductive inference formalism and to build up the best model from the discovered rules that can be used for queries.

MAIN THRUST

This article presents an approach to discover association rules that are most relevant with respect to the maximum entropy methods. Because entropy is related to information, this approach can be considered as aiming to find the most informative rules in data. The basic idea is to exploit numerical relationships that are observed by comparing (relative) frequencies, or ratios of frequencies, and so forth, as manifestations of interactions of underlying conditional knowledge.

My approach differs from usual knowledge discovery and data-mining methods in various respects:

- It explicitly takes the instrument of inductive inference into consideration.

- It is based on statistical information but not on probabilities close to 1; actually, it mostly uses only structural information obtained from the data.
- It is not based on observing conditional independencies (as for learning causal structures), but aims at learning relevant conditional dependencies in a nonheuristic way.
- As a further novelty, it does not compute single, isolated rules, but yields a set of rules by taking into account highly complex interactions of rules.
- Zero probabilities computed from data are interpreted as missing information, not as certain knowledge.

The resulting set of rules may serve as a basis for maximum entropy inference. Therefore, the method described in this article addresses minimality aspects, as in Padmanabhan and Tuzhilin (2000), and makes use of inference mechanisms, as in Cristofor and Simovici (2002). Different from most approaches, however, it exploits the inferential power of the maximum entropy methods in full consequence and in a structural, nonheuristic way.

Modelling Conditional Knowledge by Maximum Entropy (ME)

Suppose a set $R^* = \{(B1|A1)[x1], \ldots, (Bn|An)[xn]\}$ of probabilistic conditionals is given. For instance, R^* may describe the knowledge available to a physician when he has to make a diagnosis. Or R^* may express common sense knowledge, such as "Students are young with a probability of (about) 80%" and "Singles (i.e., unmarried people) are young with a probability of (about) 70%", the latter knowledge being formally expressed by $R^* = \{ (young|student)[0.8], (young|single)[0.7] \}$.

Usually, these rule bases represent incomplete knowledge, in that a lot of probability distributions are apt to represent them. So learning or inductively representing the rules, respectively, means to take them as a set of conditional constraints and to select a unique probability distribution as the best model that can be used for queries and further inferences. Paris (1994) investigates several inductive representation techniques in a probabilistic framework and proves that the principle of maximum entropy (ME-principle) yields the only method to represent incomplete knowledge in an unbiased way, satisfying a set of postulates describing sound common sense reasoning. The entropy H(P) of a probability distribution P is defined as

$$H(P) = - \Sigma_w P(w) \log P(w),$$

where the sum is taken over all possible worlds, w, and measures the amount of indeterminateness inherent to P. Applying the principle of maximum entropy, then, means to select the unique distribution $P^* = ME(R^*)$ that maximizes H(P) among all distributions P that satisfy the rules in R^*. In this way, the ME-method ensures that no further information is added, so the knowledge R^* is represented most faithfully.

Indeed, the ME-principle provides a most convenient and founded method to represent incomplete probabilistic knowledge (efficient implementations of ME-systems are described in Roedder & Kern-Isberner, 2003). In an ME-environment, the expert has to list only whatever relevant conditional probabilities he or she is aware of. Furthermore, ME-modelling preserves the generic nature of conditionals by minimizing the amount of information being added, as shown in Kern-Isberner (2001).

Nevertheless, modelling ME-rule bases has to be done carefully so as to ensure that *all* relevant dependencies are taken into account. This task can be difficult and troublesome. Usually, the modelling rules are based somehow on statistical data. So, a method to compute rule sets appropriate for ME-modelling from statistical data is urgently needed.

Structures of Knowledge

The most typical approach to discover interesting rules from data is to look for rules with a significantly high (conditional) probability and a concise antecedent (Bayardo & Agrawal, 1999; Agarwal, Aggarwal, & Prasad, 2000; Fayyad & Uthurusamy, 2002; Coenen, Goulbourne, & Leng, 2001). Basing relevance on frequencies, however, is sometimes unsatisfactory and inadequate, particularly in complex domains such as medicine. Further criteria to measure the interestingness of the rules or to exclude redundant rules have also been brought forth (Jaroszewicz & Simovici, 2001; Bastide, Pasquier, Taouil, Stumme, & Lakhal, 2000; Zaki, 2000). Some of these algorithms also make use of optimization criteria, which are based on entropy (Jaroszewicz & Simovici, 2002).