A Probabilistic Logic-based Framework for Characterizing Knowledge Discovery in Databases

Ying Xie and Vijay V. Raghavan

The Center for Advanced Computer Studies, University of Louisiana at Lafayette
{yxx2098,raghavan}@cacs.louisiana.edu

In order to further improve the KDD process in terms of both the degree of automation achieved and types of knowledge discovered, we argue that a formal logical foundation is needed and suggest that Bacchus' probability logic is a good choice. By completely staying within the expressiveness of Bacchus' probability logic language, we give formal definitions of "pattern" as well as its determiners, which are "previously unknown" and "potentially useful". These definitions provide a sound foundation to overcome several deficiencies of current KDD systems with respect to novelty and usefulness judgment. Furthermore, based on this logic, we propose a logic induction operator that defines a standard process through which all the potentially useful patterns embedded in the given data can be discovered. Hence, general knowledge discovery (independent of any application) is defined to be any process functionally equivalent to the process specified by this logic induction operator with respect to the given data. By customizing the parameters and providing more constraints, users can guide the knowledge discovery process to obtain a specific subset of all previously unknown and potentially useful patterns, in order to satisfy their current needs.

1 Introduction

The knowledge Discovery in Databases (KDD) process is defined to be the non-trivial extraction of implicit, previously unknown and potentially useful patterns from data [3]. This concise definition implies the basic capabilities that a KDD system should have. In reality, however, the user of current KDD systems is only provided limited automation: on the one hand, he is required to specify in advance exactly which type of knowledge is considered potentially useful and thus need to be discovered; on the other hand, often, he has to make the novelty judgment from a glut of patterns generated by KDD systems.

Y. Xie and V.V. Raghavan: *A Probabilistic Logic-based Framework for Characterizing Knowledge Discovery in Databases*, Studies in Computational Intelligence (SCI) **6**, 87–100 (2005)
www.springerlink.com

In other words, the KDD system can neither support in the specification of knowledge types that are potentially useful nor that a specific piece of discovered knowledge is previously unknown. Thus, it cannot guarantee what it discovers is previously unknown and potentially useful. This deficiency offers big room for the improvement of KDD process in terms of both the degree of automation and ability to discover a variety of knowledge types. In the following subsections, we will show in detail that in order to achieve such improvement, a formal theoretical framework of KDD is needed, and suggest that Bacchus' probabilistic logic provides a good foundation for us to start from.

1.1 On "Previously Unknown Pattern"

One can imagine that without a specification of what is "already known", there is no way for KDD systems to judge what is "previously unknown". Nevertheless, the view of a "real world" of current KDD systems is just a set of facts or data, so that any discovered pattern with parameters higher than thresholds will be deemed novel by the KDD systems, though they may not be so to human users. This is one of the major reasons why identifying and eliminating uninteresting discovered patterns is always a hot topic in the practice of data mining. In other words, the lack of the ability to model "already known patterns" burdens the user with the task of novelty judgment. In order to improve the degree of automation of the knowledge discovery process, we need a comprehensive model of the "real world", which requires a unified formalism to represent both facts and patterns.

However, the formalization of the representation alone is not enough to solve the novelty judgment problem. The inference ability among patterns is also required. For example, assume that we only have the following two "already known patterns": 1) More than 80% of the students didn't get A in the test; 2) Every student living on campus got A. Now suppose we obtain another pattern through some KDD process: more than 80% of the students didn't live on campus. Does it belong to "previously unknown pattern" or not? Most of us may agree that it does not, because this pattern can be easily deduced from the other two. Therefore, in order to effectively identify "previously unknown pattern", the deductive ability of the formalism that helps in recognizing relationships among patterns (statistical assertions) is also a necessity.

1.2 On "Potentially Useful Pattern"

In almost all previous literature, "potential usefulness" was viewed as a user-dependent measure. This popular opinion implies that the user has already known or currently has the ability to know which type of pattern will be potentially useful. If we cannot say that there exists, to some degree, a paradox in the way the terms of "previously unknown" and "potentially