

Optimal Knowledge Refreshing Policies for the KDD Process

Xiao Fang

Department of Information, Operations and Technology Management, University of Toledo, Toledo, Ohio
43606, USA, xiao.fang@utoledo.edu

Olivia R. Liu Sheng

School of Accounting and Information Systems, University of Utah, Salt Lake City, Utah 84112, USA,
actos@business.utah.edu

Abstract

Knowledge Discovery in Databases (KDD) provides organizations necessary tools to sift through vast data stores to extract knowledge, which supports and improves organizational decision making. Most of the prior KDD research assumed that data are static, and focused on either the efficiency of the KDD process (e.g., designing more efficient data mining algorithms) or identifying new applications of KDD. However, data are dynamic in reality (i.e., new data are continuously accumulated). Knowledge discovered using KDD becomes obsolete over time as the discovered knowledge only relates the nature of data at the time KDD process was run. Newly added data could bring in new knowledge and invalidate some earlier discovered knowledge. To support effective and efficient decision making, knowledge discovered using KDD needs to be updated along with its dynamic data source. In this research, we focus on knowledge refreshing, which we define as the process of keeping knowledge discovered using KDD up-to-date with its dynamic data source.

Prior related research such as incremental data mining and data stream mining only addressed one step in the KDD process – the data mining step, and focused on

maintaining patterns over a dynamic data source. However, it is knowledge, not patterns that can support organizational decision making effectively. To support effective decision making, the KDD process needs to be completed. The KDD process cannot be fully automated, except for the data mining step. It requires people (e.g., domain experts) in cleaning data before mining it and extracting knowledge from patterns discovered using data mining. Hence, the KDD is a costly process, as personnel costs dominate equipment and computational costs. As a result, it may be impractical to run KDD whenever there is an update in a data source. Further, it may be unnecessary to run KDD whenever there is an update in a data source. Such a practice may often result in no new knowledge because successive snapshots of real world data sources are very likely to overlap considerably. On the other hand, running KDD too seldom could result in losing critical knowledge. This will have adverse effect on decision making. Therefore, it is critical to determine when to run KDD so as to optimize the trade-off between the cost of knowledge loss and the cost of running KDD. Our research studies the knowledge refreshing problem from this perspective.

In this research, we formally define a metric for measuring knowledge loss. Through extensive experiments with real world data sets, we find that the relationship between the amount of new data and the amount of knowledge loss can be characterized by a Weibull function. We next propose a Markov decision model for the knowledge refreshing problem. The optimal policy for determining when to run KDD over a finite time period can then be derived from the model. Key properties of the optimal policy are analyzed and demonstrated using real world data sets.