ELSEVIER

Decision Support Systems 31 (2001) 205–222

Decision Support Systems

www.elsevier.com/locate/dsw

# Dynamic rule refinement in knowledge-based data mining systems

Sang C. Park [a], Selwyn Piramuthu [b], Michael J. Shaw [c,*]

[a] *Industrial Management Department, Korea Advanced Institute of Science and Technology, 373-1 Kusong-Dong, Yusong-Ku, Taejon 305-701, South Korea*

[b] *The Wharton School, University of Pennsylvania, 1300 Steinberg Hall–Dietrich Hall, Philadelphia, PA 19104-6366, USA*

[c] *Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champagne, Room 2051, 405 N. Mathews Avenue, Urbana, IL 61801, USA*

**Abstract**

The availability of relatively inexpensive computing power as well as the ability to obtain, store, and retrieve huge amounts of data has spurred interest in data mining. In a majority of data mining applications, most of the effort is spent in cleaning the data and extracting useful patterns in the data. However, a critical step in refining the extracted knowledge especially in dynamic environments is often overlooked. This paper focuses on knowledge refinement, a necessary process to obtain and maintain current knowledge in the domain of interest. The process of knowledge refinement is necessary not only to have accurate and effective knowledge bases but also to dynamically adapt to changes. KREFS, a knowledge refinement system, is presented and evaluated in this paper.

KREFS refines knowledge by intelligently self-guiding the generation of new training examples. Avoiding typical problems associated with dependency on domain knowledge, KREFS identifies and learns distinct concepts from scratch. In addition to improving upon features of existing knowledge refinement systems, KREFS provides a general framework for knowledge refinement. Compared to other knowledge refinement systems, KREFS is shown to have more expressive power that renders its applicability in more realistic applications involving the management of knowledge. © 2001 Elsevier Science B.V. All rights reserved.

*Keywords:* Knowledge refinement; Data mining

## 1. Introduction

The decreasing cost of computing, the ease of collecting and storing data, advances in DBMS technologies, as well as the extensive set of available analytical tools have been instrumental in generating interest in data mining applications. In addition to traditional top-down data analyses including well-founded application queries and report generation, bottom-up discovery-driven data analyses have been gaining popularity. Both individuals as well as organizations are beginning to explore possible ways to extract useful patterns that may be present in data to facilitate faster, more accurate, and better business decisions.

[*] Corresponding author. Tel.: +1-217-244-1266; fax: +1-217-244-8371.

*E-mail address:* m-shaw2@uiuc.edu (M.J. Shaw).

PII: S0167-9236(00)00132-9

Given the strategic advantage in data mining as a valuable decision support tool, the projected growth of data mining applications in a short period of time is not surprising. The Meta Group has estimated the market for data mining applications to reach US$8.1 billion by the year 2001 (Financial Post, October 1999). In a majority of data mining applications, a commonly known estimate is that between 70% and 80% of the resources are spent on pre-processing the data [8]. This includes integrating existing sources of data, supplementing the existing data with other necessary data, selecting the relevant data, preparing the data including data conversions, forming new attributes, as well as means to handle noisy, incomplete, duplicate, or missing data. Only the remaining small percentage is used for actually discovering patterns in the data. After all this, only a small fraction of the supposedly useful information discovered from the data are useful and actionable in reality. This is further exacerbated by the dynamic nature of most real-world environments that results in obsolescence of extracted knowledge. This necessitates careful examination and refinement of extracted knowledge over time. Thus, the process of knowledge refinement is necessary to maintain accurate, effective, and useful knowledge base that is dynamically updated as per changes in the environment. Knowledge refinement is especially critical in maintaining accurate and robust knowledge in a dynamic environment.

Notable knowledge refinement systems include SEEK [30], SEEK2 [15], FOIL [33], GOLEM [26], and KBANN [36]. Most of these systems are customized for specific applications, and application to other domains is difficult in general. The specific details of the strengths and limitations of these systems are discussed in Section 3.

We develop a general knowledge refining system, KREFS, to overcome limitations identified in existing systems. In the next section, we provide a brief overview of data mining. In Section 3, we discuss the strengths and weaknesses of existing knowledge refinement systems. An overview of KREFS, the proposed knowledge refinement system, is provided in Section 4. Comparative analysis of KREFS' relative performance over an existing knowledge refinement system is presented in Section 5. A real-world bankruptcy prediction data is used to illustrate the performance of KREFS in Section 6. Section 7 concludes this paper with a brief discussion.

## 2. Data mining

Data mining is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [13]. Specific methods used in data mining applications include statistical pattern recognition (e.g., Refs. [14,16,18]), association rules (e.g., Refs. [1,2,7,37]), recognizing sequential or temporal (time series) patterns (e.g., Refs. [4,5,22]), clustering or segmentation, (e.g., Ref. [12]), data visualization (e.g., Refs. [21,24,35]), and classification (e.g., Refs. [3,9,20]).

A typical example application for association rules is market basket analysis. The goal here is to find patterns across a large number of transactions to understand buying patterns. For example, in the financial domain, these methods can be used to analyze customers' account portfolios to identify financial services that are often utilized together. This information can then be used to create service packages targeting appropriate customers, to serve them more efficiently. While association rules involve transactions that occur at a single point in time, sequential or temporal analyses involve transactions that occur across time. Here, in addition to the characteristics of the transactions themselves, the order or sequence in which they occur is important. An application of this method could be to identify patterns and predict characteristics of future transactions based on current transactions, as in likely sequences of purchases for direct marketing. Clustering or segmentation methods identify groups of related records that are homogeneous or identical in some respect. An application could be to target segments of a population for a sales campaign based on their demographics and previous purchasing behavior. Data visualization is used to cluster related records based on certain dimensions of interest. However, the method gets unwieldy as the number of dimensions in the data increases.

Classification is perhaps one of the most popular method of choice in data mining applications. Methods such as decision trees, neural networks, and genetic algorithms have been widely used for classi-