Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 33 (2007) 1110–1116

Expert Systems with Applications

www.elsevier.com/locate/eswa

Ranking discovered rules from data mining with multiple criteria by data envelopment analysis

Mu-Chen Chen *

Institute of Traffic and Transportation, National Chiao Tung University 4F, No. 118, Section 1, Chung Hsiao W. Road, Taipei 10012, Taiwan, ROC

Abstract

In data mining applications, it is important to develop evaluation methods for selecting quality and profitable rules. This paper utilizes a non-parametric approach, Data Envelopment Analysis (DEA), to estimate and rank the efficiency of association rules with multiple criteria. The interestingness of association rules is conventionally measured based on support and confidence. For specific applications, domain knowledge can be further designed as measures to evaluate the discovered rules. For example, in market basket analysis, the product value and cross-selling profit associated with the association rule can serve as essential measures to rule interestingness. In this paper, these domain measures are also included in the rule ranking procedure for selecting valuable rules for implementation. An example of market basket analysis is applied to illustrate the DEA based methodology for measuring the efficiency of association rules with multiple criteria.

Keywords: Data mining; Association rules; Interestingness; Data envelopment analysis; Multiple criteria

1. Introduction

Data mining techniques have become widespread in business. Moreover, various rules may be obtained using data mining techniques, and only a small number of these rules may be selected for implementation due, at least in part, to limitations of budget and resources. Association rule mining differs from traditional machine learning techniques by permitting decision makers to pick from the many potential models that can be supported by the data (Webb & Zhang, 2005). Generally, association rule mining discovers all rules that meet certain sets of criteria or constraints, such as minimum support and minimum confidence, rather than generating a single model that best matches the data.

Evaluating the interestingness or usefulness of association rules is important in data mining. In many business applications, it is necessary to rank rules from data mining due to the number of quality rules (Tan & Kumar, 2000) and business resource constraint (Choi, Ahn, & Kim, 2005). Selecting the more valuable rules for implementation increases the possibility of success in data mining. For example, in market basket analysis, understanding which products are usually bought together by customers and how the cross-selling promotions are beneficial to sellers both attract marketing analysts. The former makes sellers to provide appropriate products by considering the customers' preferences, and the later allows sellers to gain increased profits by considering the sellers' profits. Customers' preferences can be measured based on support and confidence in association rules. On the other hand, seller profits can be assessed using domain related measures such as sale profit and cross-selling profit associated with the association rules.

Since high value products are relatively uncommonly bought by customers, a rule that is profitable to sellers may not be discovered by setting constraints of minimum support and minimum confidence in the mining process.

[*] Tel.: +886 2 2349 4967; fax: +886 2 2349 4953.
E-mail address: ittchen@mail.nctu.edu.tw

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.08.007

Cohen et al. (2000) described a good example of this, namely the *Ketel vodka and Beluga caviar* problem. Although, most customers infrequently buy either of these two products, and they rarely appear in frequent itemsets, their profits may be potentially higher than many lower value products that are more frequently bought. Another example regarding the interesting infrequent itemsets is described in Tao, Murtagh, and Farid (2003). The association rule of [*wine* ⇒ *salmon*, 1%, 80%] may be more interesting to analysts than [*bread* ⇒ *milk*, 3%, 80%] despite the first rule having lower support. The items in the first rule typically are associated with more profit per unit sale.

From the examples of *Ketel vodka and Beluga caviar* and *wine and salmon*, infrequent itemsets may be interesting for certain applications provided that domain information is considered (Tao et al., 2003; Webb & Zhang, 2005). However, the traditional association rule mining algorithms (Agrawal, Imielinski, & Swami, 1993; Srikant & Agrawal, 1997) cannot classify such infrequent products to interesting itemsets since the subjective domain knowledge is ignored. A lower threshold can be set to identify the infrequent itemsets with a high value. However, numerous association rules are consequently generated, and it is extremely difficult for analysts to select the useful rules between them.

In previous studies dealing with the discovery of subjectively interesting association rules, most approaches require manual input or interaction by asking users to explicitly distinguish between interesting and uninteresting rules (Liu, Hsu, Chen, & Ma, 2000). Liu et al. briefly reviewed these existing approaches. The measures of interestingness are specified as constraints in the mining process, and only the rules that satisfied these constraints are retrieved. Klemetinen, Mannila, Ronkainen, Toivonen, and Verkamo (1994) proposed an item constraint, which describes the occurrence of certain items in the conditional (right hand side) and consequent (left hand side) parts. Srikant, Vu, and Agrawal (1997) also proposed a mining algorithm that considered the item and item hierarchy constraints specified by analysts. Moreover, Lakshmanan, Han, and Pang (1998) extended the approach developed by Srikant et al. to consider much more complicated constraints, including domain, class, and SQL-style aggregate constraints. The approach developed by Ng et al. can support constraint based, human-centered exploratory mining of association rules. Goethals and Van den Bussche (2000) also proposed an interactive approach based on querying conditions within the association rule mining process.

Liu et al. (2000) proposed an approach to assist analysts in finding interesting rules from a set of mined association rules by analyzing the rules using the domain information. The mined rules are then ranked according to two subjective interestingness measures, *unexpectedness* and *actionability*. The degree of unexpectedness of rules can be measured by the extent to which they surprise the analyst (Liu & Hsu, 1996; Silberschatz & Tuzhilin, 1996). Meanwhile, the degree of actionability can be measured by the extent to which analysts can use the discovered rules to their advantage. The system developed by Liu et al. (2000) is an interactive and iterative post-processing technique. This system first asks analysts to specify their existing domain knowledge, and then analyzes the discovered rules to identify the potentially interesting ones. However, Liu et al. focused on unexpected rules, which are measured by unexpectedness.

Choi et al. (2005) proposed a group decision making approach based on Analytic Hierarchy Process (AHP) to rank the association rules generated from data mining. This approach would construct a consensus provided that a group of managers work together to select discovered. The rule quality can be improved by considering both objective criteria and subjective preferences of managers. However, this approach encounters a problem of requiring considerable human interaction to find out the weights of criteria by aggregating the opinions of various managers.

Most existing association rule mining algorithms take the measure of large support to find frequent itemsets, and all items are considered to have equal weight (Tao et al., 2003). Therefore, these approaches are unsuitable for discovering the interesting infrequent itemsets as described in the above two examples. Tao et al. developed an approach that used an improved model of weighted support. In the approach of weighted association rule mining, itemsets are no longer simply counted as they appear in a transaction, and the subjective measures (e.g., profit) are also adopted for rule evaluation.

Most of the abovementioned approaches focus on computation efficiency by embedding the subjective constraints in the mining procedure to prune the search space. However, a huge amount of subjective domain knowledge may exist, which can be considered as potential subjective constraints and interestingness measures. It is sophisticated to determine the subjective constraints and interestingness measures before discovering some rules. Provided that the constraints are not adequately stated, the interesting rules may not be discovered after the mining procedure. Additionally, rule interestingness may be a relative measure, but not an absolute one. Generally, decision makers can suitably select interesting rules for implementation after making comparisons between some potential rules.

In data mining, it is substantial to bring together the statistic based rule extraction and profit based action to meet the enterprises' objectives (Wang, Zhou, & Han, 2002). This paper aims at using a non-parametric approach, Data Envelopment Analysis (DEA), to estimate and rank the efficiency (interestingness or usefulness) of association rules with multiple criteria. The interestingness of association rules is measured by multiple criteria involving support, confidence and domain related measures. This paper uses DEA as a post-processing approach. After the rules have been discovered from the association rule mining algorithms, DEA is used to rank those discovered rules based on the specified criteria. The remainder of this paper is organized as follows. Section 2 introduces the concept of association rules. Section 3 then presents the DEA method.