Expert Systems with Applications 36 (2009) 8503–8508

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A new method for ranking discovered rules from data mining by DEA

Mehdi Toloo [a,*], Babak Sohrabi [b], Soroosh Nalchigar [b]

[a] Department of Mathematics, Islamic Azad University of Central Tehran Branch, Tehran, Iran
[b] Department of Information Technology Management, Faculty of Management, University of Tehran, Tehran, Iran

ARTICLE INFO

Keywords:
Data mining
Data envelopment analysis
Association rule
Interestingness

ABSTRACT

Data mining techniques, extracting patterns from large databases have become widespread in business. Using these techniques, various rules may be obtained and only a small number of these rules may be selected for implementation due, at least in part, to limitations of budget and resources. Evaluating and ranking the interestingness or usefulness of association rules is important in data mining. This paper proposes a new integrated data envelopment analysis (DEA) model which is able to find most efficient association rule by solving only one mixed integer linear programming (MILP). Then, utilizing this model, a new method for prioritizing association rules by considering multiple criteria is proposed. As an advantage, the proposed method is computationally more efficient than previous works. Using an example of market basket analysis, applicability of our DEA based method for measuring the efficiency of association rules with multiple criteria is illustrated.

1. Introduction

With the rapid growth of databases in many modern enterprises data mining has become an increasingly important approach for data analysis. In recent years, the field of data mining has seen an explosion of interest from both academia and industry (Olafson, Li, & Wu, 2008). Increasing volume of data, increasing awareness of inadequacy of human brain to process data and increasing affordability of machine learning are reasons of growing popularity of data mining (Marakas, 2004).

One of the main objectives of data mining is to produce interesting rules with respect to some user's point of view. This user is not assumed to be a data mining expert, but rather an expert in the field being mined (Lenca, Meyer, Vaillant, & Lallich, 2008). The problem of discovering association rules has received considerable research attention and several fast algorithms for mining association rules have been developed (Srikant, Vu, & Agrawal, 1997). Using these techniques, various rules may be obtained and only a small number of these rules may be selected for implementation due, at least in part, to limitations of budget and resources (Chen, 2007). According to Liu, Hsu, Chen, and Ma (2000) the interestingness issue has long been identified as an important problem in data mining. It refers to finding rules that are interesting/useful to the user, not just any possible rule. Indeed, there exist some situations that make necessary the prioritization of rules for selecting and concentrating on more valuable rules due to the number of qualified rules (Tan & Kumar, 2000) and limited business resources (Choi, Ahn, & Kim, 2005). According to Chen (2007), 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 (Chen, 2007).

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 (Chen, 2007). Srikant et al. (1997) presented three integrated algorithms for mining association rules with item constraint. Moreover, Lakshmanan et al. (1998) extended the approach presented by Srikant et al. to consider much more complicated constraints, including domain, class, and SQL-style aggregate constraints. Liu et al. (2000) presents an Interestingness Analysis System (IAS) to help the user identify interesting association rules. In their proposed method, they consider two main subjective interestingness

* Corresponding author. Tel.: +98 9122390003.
E-mail addresses: m_toloo@yahoo.com (M. Toloo), Bsohrabi@ut.ac.ir (B. Sohrabi), Nalchigar@ut.ac.ir (S. Nalchigar).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.10.038

measures, unexpectedness and actionability. Choi et al. (2005), using analytic hierarchy process (AHP) presented a method for association rules prioritization which considers the business values which are comprised of objective metric or managers' subjective judgments. They believed that proposed method makes synergy with decision analysis techniques for solving problems in the domain of data mining. Nevertheless this method requires large number of human interaction to obtain weights of criteria by aggregating the opinions of various managers. Chen (2007) developed their work and proposed a data envelopment analysis (DEA) based methodology for ranking association rules while considering multiple criteria. During his ranking procedure, he uses a DEA model, proposed by Cook and Kress (1990), to identify efficient association rules. Then, he applies another DEA model, developed by Obata and Ishii (2003), to discriminate efficient association rules. It should be noted that his proposed method requires the first model to be solved for all DMUs and the second model to be solved for efficient DMUs. As a drawback, this approach requires considerable number of linear programming (LP) models to be solved. Moreover, this approach includes some redundant computations and considerations. Therefore there is a need for a method which is able to rank association rules more efficiently. This paper tries to fill the gap by developing a new integrated DEA model which is able to identify most efficient association rule by solving only one mixed integer linear programming (MILP) and proposing a new method for ranking association rules with multiple criteria. The proposed method is computationally efficient and helps user to get fast results.

DEA is a non-parametric linear programming based technique for measuring the relative efficiency of a set of similar units, usually referred to as decision making units (DMUs). Because of its successful application and case studies, DEA has gained too much attention and widespread use by business and academy researchers. Evaluation of data warehouse operations (Mannino, Hong, & Choi, 2008), selection of flexible manufacturing system (Liu, 2008), assessment of bank branch performance (Camanho & Dyson, 2005), examining bank efficiency (Chen, Skully, & Brown, 2005), analyzing firm's financial statements (Edirisinghe & Zhang, 2007), measuring the efficiency of higher education institutions (Johnes, 2006), solving facility layout design (FLD) problem (Ertay, Ruan, & Tuzkaya, 2006) and measuring the efficiency of organizational investments in information technology (Shafer & Byrd, 2000) are examples of using DEA in various areas. Similar to Chen (2007), 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 main contribution of this paper is to develop a new integrated DEA model for finding most efficient association rule (by solving only one LP) and to propose a new method for ranking discovered association rules of data mining.

The rest of this paper is organized as follows. In section 2, briefly, association rule is described. Section 3, presents DEA models and section 4 discuss a previous method for ranking association rules. Our proposed method is introduced in section 5. Then, applicability of our method is illustrated in section 6. The paper closes with some concluding remarks in section 7.

2. Association rule

Association rule mining, introduced by Agrawal, Imielinski, and Swami (1993), has been widely used from traditional business applications such as cross-marketing, attached mailing, catalog design, loss-leader analysis, store layout, and customer segmentation to e-business applications such as the renewal of web pages and web personalization (Choi et al., 2005).

Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form $X \Rightarrow Y$, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contains X to contain Y. An example of an association rule is: "40% of transactions that contain bread also contain milk; 3% of all transactions contain both these items". Here 40% is called the confidence of the rule, and 3% the support of the rule. It should be noted that associations may include any number of items on either side of the rule. An efficient algorithm is required that restricts the search space and checks only a subset of all association rules, yet does not miss important rules (Chen, 2007). Many algorithms can be used to discover association rules from data to extract useful patterns. Apriori algorithm is one of the most widely used and famous techniques for finding association rules (Agrawal & Srikant, 1994; Agrawal et al., 1993). Apriori operates in two phases. In the first phase, all itemsets with minimum support (*frequent* itemsets) are generated. This phase utilizes the downward closure property of support. In other words, if an itemset of size k is a frequent itemset, then all the itemsets below $(k-1)$ size must also be frequent itemsets. Using this property, candidate itemsets of size k are generated from the set of frequent itemsets of size $(k-1)$ by imposing the constraint that all subsets of size $(k-1)$ of any candidate itemset must be present in the set of frequent itemsets of size $(k-1)$. The second phase of the algorithm generates rules from the set of all frequent itemsets.

Association rule mining is a popular technique for market basket analysis, which typically aims at finding buying patterns for supermarket, mail-order and other customers. By mining association rules, marketing analysts try to find sets of products that are frequently bought together, so that certain other items can be inferred from a shopping cart containing particular items. Association rules can often be used to design marketing promotions, for example, by appropriately arranging products on a supermarket shelf and by directly suggesting to customers items that may be of interest (Chen, 2007).

3. DEA models

DEA is a data-oriented approach for relatively evaluating the performance of a group of entities referred to DMUs. It was introduced by Charnes, Cooper, and Rhodes (1978) based on Farell's pioneering work. They generalized the single-output to single-input ratio definition of efficiency to multiple inputs and outputs. In their original DEA model, Charnes, Cooper and Rhodes (CCR model) proposed that the efficiency of a DMU can be obtained as the maximum of a ratio of weighted outputs to weighted inputs, subject to the condition that the same ratio for all DMUs must be less than or equal to one. The DEA model must be run n times, once for each unit, to get the relative efficiency of all DMUs. The envelopment in CCR is constant returns to scale meaning that a proportional increase in inputs results in a proportionate increase in outputs. Banker, Charnes, and Cooper (1984) developed the BCC model to estimate the pure technical efficiency of decision making units with reference to the efficient frontier. It also identifies whether a DMU is operating in increasing, decreasing or constant returns to scale. So CCR models are a specific type of BCC models.

Assume that there are n DMUs, $(\mathrm{DMU_j} : j = 1, 2, \ldots, n)$ which consume m inputs $(\mathbf{x}_i : i = 1, 2, \ldots, m)$ to produce s outputs $(\mathbf{y}_r : r = 1, 2, \ldots, s)$. The BCC input oriented (BCC-I) model evaluates the efficiency of $\mathrm{DMU_o}$, DMU under consideration, by solving the following linear program: