IGI PUBLISHING

ITB15219

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, **Data Mining and Knowledge Discovery Technologies** edited by D. Taniar © 2008, IGI Global

Chapter II

Current Interestingness Measures for Association Rules: What Do They Really Measure?

Yun Sing Koh, Auckland University of Technology, New Zealand
Richard O'Keefe, University of Otago, New Zealand
Nathan Rountree, University of Otago, New Zealand

Abstract

Association rules are patterns that offer useful information on dependencies that exist between the sets of items. Current association rule mining techniques such as apriori often extract a very large number of rules. To make sense of these rules we need to order or group the rules in some fashion such that the useful patterns are highlighted. The study of this process involves the investigation of an "interestingness" in the rules. To date, various measures have been proposed but unfortunately, these measures present inconsistent information about the interestingness of a rule. In this chapter, we show that different metrics try to capture different dependencies among variables. Each measure has its own selection bias that justifies the rationale for preferring it compared to other measures. We present an experimental study of the behaviour of the interestingness measures such as lift, rule interest, Laplace, and information gain. Our experimental results verify that many of these measures

are very similar in nature. From the findings, we introduce a classification of the current interestingness measures.

Introduction

Interestingness measures are divided into two types: objective and subjective measures. Objective measures are based on probability, statistics, or information theory. They use a data driven approach to assess the interestingness of a rule. They are domain independent and require minimal user participation. Objective measures emphasise conciseness, generality, reliability, peculiarity, or diversity of the rules found. Some objective measures are symmetric with respect to the permutation of items, while others are not. From an association rule mining perspective, symmetric measures are often used for itemsets whereas asymmetric measures are applied to rules. Using these measures each association rule is treated as an isolated rule and they are not compared against each other. Subjective measures take into account both the data and the user of these data. Hence, subjective measures require access to domain knowledge on the data. These measures determine whether a rule is novel, actionable, and surprising. A rule is interesting if it is both surprising and actionable. However, this is a highly subjective view as actionable is determined by both the problem domain and the user's goals (Silberschatz & Tuzhilin, 1995).

In this chapter, we only concentrate on objective measures, as they do not need expert domain knowledge. A large number of rules may be extracted as we lower the minimum support threshold or increase the number of items in the database. For this reason, the number of possible association rules grows exponentially with the number of items and the complexity of the rules being considered. For this reason, objective measures are used to rank, order, and prune the rules for presentation.

Currently there are more than 50 objective measures proposed and at present, there are a number of reviews conducted to make sense of the interestingness measures for association rules (Geng & Hamilton, 2006; McGarry, 2005; Tan & Kumar, 2000). Here we make two major contributions. We present a new visualisation technique to visualise and evaluate the current objective measures and also discuss the suitability of these measures in detecting meaningful rules.

Most objective measures are probability based. They are normally functions of a 2×2 contingency table. Table 1 shows the contingency table for $A \rightarrow B$ in dataset D. Here $n(AB)$ denotes the number of transactions containing both A and B in dataset D or count(AB,D). N denotes the total number of transactions or $|D|$. For this pur-