Chapter 9

Automatic Item Weight Generation for Pattern Mining and its Application

Yun Sing Koh
The University of Auckland, New Zealand

Russel Pears
Auckland University of Technology, New Zealand

Gillian Dobbie
The University of Auckland, New Zealand

ABSTRACT

Association rule mining discovers relationships among items in a transactional database. Most approaches assume that all items within a dataset have a uniform distribution with respect to support. However, this is not always the case, and weighted association rule mining (WARM) was introduced to provide importance to individual items. Previous approaches to the weighted association rule mining problem require users to assign weights to items. In certain cases, it is difficult to provide weights to all items within a dataset. In this paper, the authors propose a method that is based on a novel Valency model that automatically infers item weights based on interactions between items. The authors experiment shows that the weighting scheme results in rules that better capture the natural variation that occurs in a dataset when compared with a miner that does not employ a weighting scheme. The authors applied the model in a real world application to mine text from a given collection of documents. The use of item weighting enabled the authors to attach more importance to terms that are distinctive. The results demonstrate that keyword discrimination via item weighting leads to informative rules.

DOI: 10.4018/978-1-4666-2148-0.ch009
INTRODUCTION

Association rule mining (Agrawal et al., 1993) aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in transactional databases. The relationships are not based on the inherent properties of the data themselves but rather based on the co-occurrence of the items within the database. There has been much work carried out in this area (Ashrafi et al., 2004; Daly & Taniar, 2004; Cokrowijoyo & Taniar, 2005; Koh et al., 2006; Ashrafi et al., 2007; Tzanis & Berberidis, 2007; Taniar et al., 2008; Giannikopoulos et al., 2010).

The original motivation for seeking association rules came from the need to analyze supermarket transactional data also known as market basket analysis. An example of a common association rule is bread → butter. This indicates that a customer buying bread would also buy butter. Association rules have been widely used in a wide variety of domains, however, traditional rule mining techniques are vulnerable to the “rule explosion problem”. Even modest sized datasets can produce thousands of rules, and as datasets get larger, the number of rules produced becomes unmanageable. This highlights a key problem in association rule mining; keeping the number of generated itemsets and rules in check, whilst identifying interesting rules amongst the plethora generated.

In the classical model of association rule mining, all items are treated with equal importance. In reality, most datasets are skewed with imbalanced data. By applying the classical model to these datasets, important but critical rules which occur infrequently may be missed. For example consider the rule: stiff neck, fever, aversion to light → meningitis. Meningitis occurs relatively infrequently in a medical dataset, however if it is not detected early the consequences can quickly become fatal. Recent research (Cai et al., 1998; Sun & Bai, 2008; Wang et al., 2000; Yan & Li, 2006) has used item weighting to emphasize such rules that rarely manifest but are nonetheless very important. For example, items in a market basket dataset may be weighted based on the profit they generate. However, most datasets do not come with preassigned weights and so the weights must be manually assigned, which is time consuming and maybe error-prone. Research in the area of weighted association rule mining has concentrated on formulating efficient algorithms for exploiting pre-assigned weights rather than deducing item weights from a given transactional database. We believe that it is possible to deduce the relative importance of items based on their interactions with each other. In application domains where expert’s input on item weights is either unavailable or impractical, an automated approach to assigning weights to items can contribute significantly to distinguishing high value rules from those with low value.

In this paper we discuss two major issues that are relevant to the field of weighted association rule mining. Firstly, we present a scheme that automates the process of assigning weights to items. The weights assignment process is underpinned by a “Valency model” that we propose. The model considers two factors: purity and connectivity. The purity of an item is determined by the number of items that it is associated with over the entire transactional database, whereas connectivity represents the strength of the interactions between items. We will elaborate on the Valency model later in the paper in Section 3. Secondly, association rules produced by the Valency model are evaluated through a scheme based on Principal Components Analysis. The formulation of this interest measure was motivated by the fact that none of the popularly used interest measures such as Confidence and Lift was able to capture differences between rules with highly weighted items from those with lowly weighted ones. We also apply the model to a real-world scenario, i.e., text mining. Text mining approaches typically extract keywords using measures of importance such as frequency of occurrence or measures such as TF-IDF (term frequency–inverse document frequency) to rank words in terms of their significance. While such measures are useful we are more interested in