Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 93 (2016) 12 – 18

Procedia Computer Science

6th International Conference On Advances In Computing & Communications, ICACC 2016, 6-8 September 2016, Cochin, India

Rule Power Factor: A New Interest Measure in Associative Classification

Ochin[a*], Suresh Kumar[b], Nisheeth Joshi[c]

[a,b]*Manav Rachna International University, Sec 43, Faridabad-121001, India*
[a,c]*Banasthali Vidyapith, Nawai, Rajasthan – 304022, India*

Abstract

In data mining it is generally anticipated that revealed knowledge should have characteristics of accuracy, reliability and interestingness. Most of the data mining algorithms find patterns that are accurate and reliable but might not be interesting. Interest measures are used to find the valued interesting rules which are useful to the user in effective decision making even in exceptional set of circumstances. A range of interest measures for rule mining have been suggested by researchers in the field of data mining to have different visualisations and analytics. In this paper, we have investigated a few of interest measures and proposed a new Interest measure with the name 'Rule Power Factor'. Experiments prove that this new interest measure is more informative and can act as a superset of 'Confidence' measure.

Peer-review under responsibility of the Organizing Committee of ICACC 2016

Keywords: Association rule mining, associative classification, lift, chi square, interest measure, data mining, rule power factor

1. Introduction

The field Associative Classification(AC) has tremendous opportunity and has many valued applications e.g. Market Basket Analysis, genetic epidemiological analysis, movies recommendation, heart disease prediction, cheminformatics, movies recommendation, ontology, bigdata, ubiqutous computing, Internet of Things. Associative Classification comprises with two techniques: Association rule mining (undefined) + Classification (Predefined).

* Corresponding author. Tel.: +91-0129-4198251.
E-mail address: ochinsharma2@gmail.com

Peer-review under responsibility of the Organizing Committee of ICACC 2016
doi:10.1016/j.procs.2016.07.175

Association rule mining is a very well-known technique by which various rules or association among variables can be defined. (Readers can refer paper [1] to read association rule mining in detail). Classification is a two phase process. In first phase, a sample data is collected from the data base and this is termed as training data. Using various constraints and techniques, a classification model is created to tag specific data with specific class label. When this model is justified on various parameters e.g completeness, accuracy, reliability on training data then test data is used to classify various data items with various predefined classes. Many researchers have proposed various interest measures for rule mining to fulfil the needs for effective decision making. Broadly, there are two types of classifications: objective and subjective. An objective measure has no awareness about the user. Mostly, objective measures are constructed on theories in probability, correlation and statistics. A subjective measure takes into account both the data and the user. This can be obtained by direct or indirect interaction with the user through the data mining process.

2. Related Work

There are various interest measures proposed by researchers. Each measure is basically to extract a specific pattern from the data. It depends on the need of pattern/data to be mined. For instance, user might be only interested in correlated data or or to find the number of data items dominating in the database. At many occasions, finding negative rules are useful to avoid misleading decisions. For complete list of interest measures readers may refer paper [12]. Consider table 1, containing transactional data set used in sub-sections from 2.1 to 2.5 to understand basic terminology and a few well known interest measures.

Table 1. Transactional data set

TID	T1	T2	T3	T4	T5
A	0	1	1	1	1
B	1	1	0	0	1

2.1. Support

Support is defined as the probability that transactions in the database contains items both the antecedent and the consequent of the rule.

support (A->B) = (Transactions containing both A and B items) / Total number of transactions.

support (A->B) => 2/5 = 40%

If minimum threshold for support is chosen low, large numbers of rules are generated and evaluation of such rules is complex and time consuming. And choosing minimum threshold value high can make the pattern skipped and can compromise the effective decision making. Support, confidence, rule generation are explained in detail in [1].

2.2 Confidence

Confidence is a measure that finds out the association among antecedence and precedence part of a rule. It ignores the total number of transactions while calculating confidence.

confidence(A->B)=Total number of transactions containing items A and B divided by total number of transactions containing item A. confidence (A->B) = support (A,B) / support(A)

confidence (A->B) = 2/4= 50%

2.3. Coverage

It measures how much database is covered by a rule A⇒ B. Hence, coverage is the number of transactions that satisfy the antecedent of a rule.

Coverage (A) = support (A) = 4 / 5 = 0.800