DM Assignment 05

9.2.2016

Adithya Abraham Philip

1PI13CS008

Overview

The objective of the assignment is to experiment with the FPGrowth implementation present in the Weka API. To gain a better understanding of the statistical measures used in association rule mining, and the FPGrowth algorithm implementation in Weka, three questions have been asked, whose answers have been provided in this document.

Questions

1. What do the measures **Leverage** and **Conviction** mean?

Leverage is a measure of the independence (or rather, lack thereof) of two items in a set of transactions. For a rule **L=>R** it is given by the formula **P(L,R) - P(L).P(R)**, where **P** is probability, **L** and **R** are items (or sets of items), and **P(L,R)** is the probability of a transaction containing both **L** and **R.** A negative leverage would imply that they are negatively correlated, a positive leverage that they are positively correlated, and the closer the value is to 0, the lesser the degree of either correlation (0 would mean completely independent). It is a symmetric measure. Range is [-1,1].

Conviction for a given rule (**L=>R)** is calculated as the ratio of the probability of **L** occurring without **R** if they were independent, to the actual probability with which **L** occurs without **R** i.e **P(L)*P(!R)/P(L,!R)**. If conviction is 1, that would mean **L** and **R** are independent. A conviction close to 0, would imply that there is a negative correlation between **L** and **R,** and a conviction greater than 1 would imply there is a positive correlation between L and R. It is not a symmetric measure.

2. How are Leverage and Conviction calculated in Weka?

Leverage is calculated by first computing the probability of the rule being correct **(totalSupport/totalTransactions)**, then calculating the expected probability of the rule being correct if they were independent i.e

(premiseSupport/totalTransactions)*(consequenceSupport/totalTransactions), and then subtracting the latter from the former.

Conviction is calculated by first computing **Sup(premise)*Sup(!consequence)/total_transactions** as the numerator, **(Sup(premise) - totalSupport + 1)** as the denominator, then dividing the former by the latter. The **+ 1** in the denominator is not part of the original formula, but is used to prevent a division by zero error in the program, in case the confidence is exactly 1.

3. Notice that Weka can print out a string representation of a rule. Suppose you wanted to change default way in which a rule is printed, which method in which class needs to be modified?

Sample rule: **[vegetables=t, biscuits=t]: 1764 ==> [fruit=t]: 1404 conf:(0.8) <lift:(1.24)> lev:(0.06) conv:(1.76)**

Modifying the **toString()** method in the **FPGrowth.AssociationRule** class would change the way the rule is printed.