Evaluating the performance of apriori and predictive apriori algorithm to find new association rules based on the statistical measures of datasets.

[1]Mukesh Sharma,[2]Jyoti Choudhary, [3]Gunjan Sharma
[1]Associate.Professor, [2]Assistant.Professor, [3]Mtech Scholar ,
Department of Computer Science and Engineering
The Technological Institute of Textile and Science,Bhiwani-127021, Haryana – India

Abstract:

Recently ,various advancements has emerged in the field of data mining. One of the hottest topic in this area is mining for association rules from the existing massive collection of datasets. The pattern obtained from these databases are used in various fields like super market sales-prediction, fraud detection and weather forecasting etc. So it is necessary that only strong rules are mined by using appropriate algorithm. In this paper, out of the various existing algorithms of association rule mining, two most important algorithm i.e. apriori and predictive apriori algorithm are chosen for experiment. Their performance is compared based on the interesting measures using weka3.7.5 which is a java based machine learning tool. After that ,various statistical measures are calculated of different datasets and then based on the comparison of algorithms and statistical measures of data, new rules are generated using see5 tool.

Keywords and Phrases: Data mining, Association rules, predictive apriori, machine learning, apriori etc.

1 Introduction

Data mining ,now a days, is the most important field of computer science and it deals with the process of extracting information from a data set and transform it into an understandable structure for further use. The mining process is an iterative sequence of steps. As the data is collected from various sources so the data is not clean. Presence of noise can disturb the predicting procedure. Therefore, Cleaning of data has to be performed first. As the data belongs to different sources integration is to be done. Not all the data is required to the user, therefore data selection should be done and then the data should be transformed to the required form for mining process. Finally, the Data Mining Engine with the help of knowledge base uses various tools for mining the data repository which contains the transformed data for pattern evaluation. Association rule mining is one of the most important technique of data mining and it finds the hidden patterns from the massive database. This technique finds the association between the items of the data file in the form of rules. The knowledge obtained from this technique is used for different applications like super market sales-prediction, medical diagnosis, fraud detection and financial forecast etc. So it become important to mine strong and interesting rules which are useful for the user.

2 Various association rule mining algorithms

(a) Apriori algorithm

Apriori is an algorithm proposed by R. Agrawal and R Srikant in 1993 [1] for mining frequent item sets for boolean association rule. The name of algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties. Apriori employs an iterative approach known as level-wise search, where k item set are used to explore (k+1) item sets. There are two steps in each iteration. The first step generates a set of candidate item sets. Then, in the second step the occurrence of each candidate set in database is counted and then pruning of all disqualified candidates (i.e. all infrequent item sets) is done. Apriori uses two pruning technique, first on the bases of support count (should be greater than user specified support threshold) and second for an item set to be frequent , all its subset should be in last frequent item set The iterations begin with size 2 item sets and the size is incremented after each iteration. The algorithm is based on the closure property of frequent item sets: if a set of items is frequent, then all its proper subsets are also frequent[2]. This algorithm is easy to implement and parallelized but it has the major disadvantage that it requires various scans of databases and is memory resident.

(b) Predictive apriori algorithm

This algorithm searches with an increasing support threshold for the best 'n' rules concerning a support-based corrected confidence value[3]. A rule is added if the expected predictive accuracy of the rule is among the 'n' best and it is not subsumed by a rule with at least the same expected predictive accuracy. This is also a confidence based association rule but in this rules ranked are sorted according to “predictive accuracy”. It tries to maximize predictive accuracy of an association rule rather than confidence in apriori.

(c) Tertius algorithm

Tertius is basically a first order logic discovery algorithm. Tertius employs a complete top-down A* search over the space of possible rules[4]. If there are A attributes with on the average V values and search for rules with up to n literals, the number of possible rules is of the order (AV)^n.

3 Various interestingness measures

(a) Support

Support for ARM is introduced by R.Agrawal in 1993[1] and it is defined as the proportion of transactions in the data set which contain the itemset.. It measures the frequency of association, i.e. how many times the specific item has been occurred in a dataset. An itemset with greater support is called frequent or large itemset. In terms of probability theory ,it can be expressed as:

Support = P (A ∩ B) = number of transactions containing both A and B /Total number of transactions

(b)Confidence

Confidence measures the strength of the association rules . It is defined as the ratio of the number of transactions that include all items in a particular frequent item set to the number of transactions that include all items in the subset. It determines how frequently item B occurs in the transaction that contains A. Confidence expresses the conditional probability of an item. The definition of confidence is

$$\text{Confidence} = P(A \mid B) = \frac{P(A \cap B)}{P(A)}$$

(c)Predictive Accuracy

Predictive accuracy is generally used for the Predictive Apriori rule measurement. According to Scheffer , definition of predictive accuracy is as follows: Let **D** be a data file with **n** number of records. If **[x → y]** is an Association Rule which is generated by a static process **P** then the predictive accuracy of [**x** →**y**] is **c([x → y])**=**P[n]** satisfies **y**|**n** satisfies **x**]where distribution of **r** is govern by the static process **P** and the Predictive Accuracy is the conditional probability of **x→n** and **y→n**.

4 Experiments

In this research ,various steps are followed first of all, two algorithms of association rule mining are compared using different measures of accuracy on 15 different datasets. The datasets are taken from the uci repository. Then various statistical measures are calculated using matlab and then based on the compared algorithms result and statistical measure result, new rules are generated with the help of See5 tool.

(a) Data preprocessing

Firstly data is preprocessed, which means raw data is prepared into a format which can be used for further processing. So,15 uci datasets are chosen which do not contain any missing values and also noiseless. Then preprocessing technique “unsupervised discretization” is applied on the datasets using weka 3.7.5.This technique is applied for converting a range of numeric attributes into nominal attributes.

(b) Association rules

Then on the preprocessed data ,the apriori and predictive apriori algorithms are applied on the datasets for generating the rules. Top 10 rules are taken for the experiment, and based on the rules, average confidence and average predictive accuracy of apriori and predictive apriori algorithms are calculated. The details are given in the table 2. Out of these two algorithms ,predictive apriori performs better.

(c) Dataset statistical measures

In this step, different central tendency measures like mean, median and mode and various statistical measures of datasets are calculated using matlab. The average of statistical measures of all the attributes are taken as global measure of the dataset characteristics. Here table 1 shows statistical measures.

Table 1: Statistical measures

Measure	Notation
Arithmetic mean	Mean
Median	Median
Mode	Mode
Variance	variance
Standard deviation	std_dev
Interquartile range	iqr
Range	range
Average deviation	ave_dev

Table 2: Comparison of algorithms

Datasets	Priori	Predictive Apriori	Better Algorithm
cmc	0.964	0.994	predictive apriori
ecoli	0.998	0.985	apriori
Haberman	0.96	0.974	predictive apriori
iris	0.992	0.991	apriori
tae	0.992	0.991	predictive apriori
vehicle	0.981	0.959	apriori
spect_test	0.921	0.994	predictive apriori
solar_flare	0.964	0.994	predictive apriori
ppd	0.94	0.993	predictive apriori
breast_w	0.98	0.994	predictive apriori
diabetes	0.975	0.986	predictive apriori
page_blocks	1	0.994	apriori
contact_lenses	1	0.744	apriori
hayes_roth	1	0.98	apriori
glass	0.981	0.987	predictive apriori

(1) Mean

The mean (or average) of a set of data values is the sum of all of the data values divided by the number of data values.

$$\text{Mean} = \frac{\text{sum of all data values}}{\text{Number of data values}}$$

Symbolically,

$\bar{x} = \sum x/\text{n}$

Where $\bar{x}$ is the mean of the set of x values, $\sum x$ is the sum of all the x values, and n is the number of x values.

(2) Median

The median of a set of data values is the middle value of the data set when it has been arranged in ascending order. That is, from the smallest value to the highest value.

(3) Mode

The mode of a set of data values is the value(s) that occurs most often. For eg the mode of these numbers 48,44,48,45,42,49,48 is 48.

(4)Variance

The variance of a data set is the arithmetic mean of the squared differences between the values and the mean.

(5) Standard deviation

Standard deviation is defined as the square root of the variance. The standard deviation measures the spread of the distribution about the mean.

(6) Interquartile range

Interquartile range is defined as the difference between the 75th percentile and the 25th percentile.

(7) Range

Range is measured by taking the difference between the highest value and the lowest value of a dataset. For eg the range of the dataset 41,37,30,20,8,22,46, 43,33,5 is 41.

(8) Average deviation

Average deviation is defined as the arithmetic mean of the absolute deviations and absolute deviation is further defined as the absolute difference between each data value and the arithmetic mean.

Table 3: By using various statistical measures and table 1 following table is constructed

Datasets	**Mean**	**Median**	**Variance**	**Std_Dev**	**Ave_Dev**	**Iqr**	**Range**	**Mode**	**Class**
Cmc	1.769	1.583	3.203	0.801	0.669	1.135	2.958	1.208	two
ecoli	0.299	0.222	0.0602	0.206	0.148	0.159	0.906	0.148	one
Haberman	3.849	3.533	11.3	1.447	1.057	1.466	7.866	3.466	two
Iris	2.122	2.064	0.748	0.744	0.653	1.4	2.471	1.385	one
Tae	6.532	5.75	32.891	3.614	3.051	5.625	14.625	5.875	two
vehicle	96.292	91.272	1657.334	18.709	15.58	28.455	98.773	90	one
spect_test	0.375	0.086	0.211	0.456	0.421	0.826	1	0.086	two
solar_flare	0.253	0.1515	0.118	0.321	0.235	0.212	1	0.151	two
Ppd	0.334	0.354	0.156	0.369	0.308	0.489	1	0.333	two
breast_w	2.856	1.5	7.084	2.523	2.02	2.9	8.2	1.1	two
diabetes	40.026	34.096	1682.745	22.927	17.209	27.965	157.05	25.028	two
page_blocks	168.951	52.491	1914499	560.214	186.705	129.79	15546	14.666	one
contact_lenses	0.388	0.277	0.226	0.474	0.434	0.777	1	0.111	one
hayes_roth	1.026	0.93	0.372	0.488	0.408	1	1.5	0.75	one
glass	6.399	6.35	0.44	0.513	0.361	0.451	2.912	6.123	two

5 Rule generation

This is based on the various statistical measures of different datasets as given in table3 , rules are generated using see5 data mining tool. See5 is a sophisticated data mining tool for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions. Out of the various statistical measures of the datasets ,see5 generates the rules based on the mean, median and range as shown in figure1.

Figure1:rules generated by see5

```
See5 [Release 2.09]          Thu Aug 02 22:20:57
2012
-------------------


   Options:
         Rule-based classifiers
         Do not use global pruning
         Pruning confidence level 99%

Read 15 cases (9 attributes) from try.data

Rules:

Rule 1: (6, lift 1.5)
         mean <= 40.026
         range > 2.471
         ->  class two  [0.875]

Rule 2: (2, lift 1.3)
         median <= 0.1515
         ->  class two  [0.750]

Rule 3: (2, lift 1.9)
         mean > 40.026
         ->  class one  [0.750]

Rule 4: (5/1, lift 1.8)
         median > 0.1515
         range <= 2.471
         ->  class one  [0.714]

Default class: two


Evaluation on training data (15 cases):

               Rules
           ----------------
            No      Errors

             4    1( 6.7%)   <<


            (a)   (b)    <-classified as
           ----  ----
             8     1    (a): class two
                   6    (b): class one

     Attribute usage:
          73%  range
          53%  mean
          47%  median

    Time: 0.0 secs
```

6 Result and Conclusion

For Apriori algorithm:-
If mean>40.026 ,median > 0.1515 and range <= 2.471 of a dataset then choose apriori algorithm.

For predictive apriori algorithm:-
If mean <= 40.026,range > 2.471 and median <= 0.1515 of a dataset then choose predictive apriori algorithm.

Association rule mining is really the emergeable topic now a days. Researchers aim to find the best and strong association rules. This paper firstly compares the performance of apriori and predictive apriori and concluded that predictive apriori performs better based on the predictive accuracy and then various statistical measures are calculated. However, the main focus of this research is to generate new rules. Therefore, this research recommends an algorithm by analyzing the mean, median and range of a dataset for finding the new association rules which are applied on the various datasets. This research can be further enhanced by considering more association rules algorithms and other statistical measures.

7 References:

[1] R.Agrawal and R.Srikant,"Fast algorithms for mining association rules",Proc. Of the 20th Int Conference on very large databases,Santingo,Chile,September 1994.

[2] Goswami D.N and Chaturvedi Anshu "An Algorithm for Frequent Pattern Mining Based On Apriori" (IJCSE) International Journal on Computer Science and Engineering
Vol. 02, No. 04, 2010, 942-947

[3] Tobias Scheffer: Finding Association Rules That Trade Support Optimally against Confidence. In: 5th European Conference on Principles of Data Mining and Knowledge Discovery, 424-435, 2001.

[4]P.A. Flach and N.Lachiche,"confirmation-guided discovery of first-order rules with tertius,"Kluwer Academic Publishers .-The Netherlands,Vol.42-pp. 61-95,2001.

[5] S. Mutter, M. Hall and E. Frank"Using Classification to Evaluate the Output of Confidence based Association Rule mining"Lecture notes in Artificial Intelligence,Advances in Artificial Intelligence - AI 2004. Berlin, Springer, vol. 3339. - pp. 538-549. 2004.

[6] Weka , http://www.cs.waikato.ac.nz/ml/weka/.

[7] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference
on Management of Data, 207-216.

[8] Li Yang,Mustafa Sanver; Mining Short Association Rules with One Database Scan; Int'l Conf. on Information and
Knowledge Engineering; June 2004.

[9] Li Yang, Mustafa Sanver; Mining Short Association Rules with One Database Scan; Int'l Conf. on Information and Knowledge Engineering; June 2004.

[10] see5, http://www.rulequest.com/download.html

[11] Freitas, A., 'On rule interestingness measures', Knowledge Based Systems 12(5-6), 309–315 (1999).

[12] Assaf Schuster, Ran Wolff, and Dan Trock; Distributed Algorithm for Mining Association Rules; IEEE Int'l Conf. on Data Mining; November 2003

[13] R.J. Bayardo, Jr. Efficiently mining long patterns from databases. In L.M. Haas and A. Tiwary, editors, Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, volume 27(2) of SIGMOD Record, pages 85–93. ACM Press, 1998.

[14] Jiawei Han and Micheline Kamber, (2006), "Data Mining: Concepts and Techniques, 2nd edition ".Elsevier publications

[15] Ying Liu, Wei – Keng Liao, Alok Choudhary,(2005), " A fast high utility itemsets mining algorithm ", UBDM 2005, Proceedings of the 1st international workshop on utility-based data mining, ACM, NY.