ESTIMATION OF ASSOCIATION MINING TO DETERMINE THE PERFORMANCE OF APRIORI ALGORITHM

P.Kalaiselvi PG Student, Dr. C.Nalini Associate Professor , Department of Computer Science and Engineering.
Bharath University, Selaiyur, Chennai-600073, India
kalai.adi@gmail.com , drnalinichadhabaram@gmail.com

Abstract

In this paper, the performance of the different association methods like Apriori and Tertius are compared. For comparative analysis, we use the Super Market Dataset. We consider the dataset with multiple itemsets and multiple customers. The main objective of this paper is to increasing the sales and reduces the cost of itemset. In this work Association rule mining approach is developed to determine the minimum support and minimum count. Association Rule Problem is to identify all
association rules X ⇒ Y with a minimum support and confidence.

Keywords: Data Mining, Association, Itemset, Support ,Confidence .

1. INTRODUCTION

Data mining and knowledge discovery [2] (data mining or KDD for short) has emerged to be one of the most vivacious areas in information technology in the last decade. It has boosted a major academic and industrial campaign crossing many traditional areas such as machine learning, database, and statistics, as well as emergent disciplines, for example, bioinformatics. As a result, KDD has published thousands of algorithms and methods.

Compared with the booming fact in academia, data mining applications in the real world has not been as active, vivacious and charming as that of academic research. This can be easily found from the extremely imbalanced numbers of published algorithms versus those really workable in the business environment. That is to say, there is a big gap between academic objectives and business goals, and between academic outputs and business expectations. However, this runs in the opposite direction of KDD's original intention and its nature. It is also against the value of KDD as a discipline, which generates the power of enabling smart businesses and developing business intelligence for smart decisions in production and living environment. For instance, academic researchers do not really know the needs of business people, and are not familiar with the business environment.

Knowledge discovery is further expected to migrate into actionable knowledge discovery (AKD). AKD targets knowledge that can be delivered in the form of business-friendly and decision-making actions, and can be taken over by business people seamlessly. However, AKD is still a big challenge to the current KDD research and development.

Data mining is a powerful paradigm of extracting information from data. It can help enterprises focus on important information in their data warehouse. Data mining is also known as Knowledge Discovery in Databases (KDD). It involves the extraction of hidden pattern to predict future trends and behaviors which allow businesses to make proactive, knowledge-driven decisions.

The current vast development in ubiquitous computing, cloud computing and networking across every sector and business has made data mining emerging as one of the most active areas in Information and Communication Technologies (ICT) as data and its deep analysis becomes an important issue for enhancing the soft power of an organization, its production systems, decision making and performance.

The primary goal of machine learning is to derive general patterns from a limited amount of data. The majority of machine learning scenarios generally fall into one of two learning tasks: supervised learning or unsupervised learning [4].

The supervised learning task is to predict some additional aspect of an input object. Examples of such a task are the simple problem of trying to predict a person's weight given their height and the more complex task of trying to predict the topic of an image given the raw pixel values. One core area of supervised learning is the classification task.

The subject is introduced briefly as following, In section 2, formulates the problem. In section 3, the experimental results and analysis. We present the conclusion in section 4.

2. ASSOCIATION RULE

Association rules are always defined on binary attributes ,such as those we used in the sample database to represent subscriptions to magazines, so in fact we would have to flatten the table mentioned above before we could execute an association algorithm. Association rules are only used in data mining . The definition for association mining is Set of items: I={I1,I2,…,Im},Transactions: D={t1,t2, …, tn}, tj I,Itemset: {Ii1,Ii2, …, Iik} I, Support of an itemset: Percentage of transactions which contain that itemset, Large (Frequent) itemset: Itemset whose number of occurrences is above a threshold.

- Association Rule (AR): implication X ⇒ Y where X,Y ⊆ I and X ∩ Y = ;
- Support of AR (s) X ⇒ Y: Percentage of transactions that contain X ∪Y
- Confidence of AR (α) X ⇒ Y: Ratio of number of transactions that contain X ∪Y to the number that contain X

The main objective of association mining is increase sales and reduce the cost.

2.1 Association Mining

The aim objective of association Mining is extract interesting correlations, frequency itemset, association among set of items in transaction database and relational database. A rules are used in telecommunication network Classification, Cluster, etc.

2.1.1 Association Rule Mining Task

1. Given a set of transactions T, the goal of association rule mining is to find all rules having

(i) support ≥ minsup threshold

(ii) confidence ≥ minconf threshold

2. Brute- force approach:

(i)List all possible association rules

(ii) Compute the support and confidence for each rule Prune rules that fail the minsup and minconf thresholds.

2.1.2 Mining Association Rules

Observations:

- All the above rules are binary partitions of the same itemset: {Milk, Bread, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements.

Association Rule Techniques

1. Find Large Itemsets.
2. Generate rules from frequent itemsets.

3. PROPOSED APRIORI ALGORITHM

Apriori Algorithm is a classic algorithm for learning association rule. This algorithm is mainly designed to operate on database containing transactions. For example: Collections of items bought by customer. Apriori is a bottom up approach. Apriori algorithm uses breadth first search and hash tree structure for efficient support and confidence.

3.1 Apriori Algorithm

Input :D database of transaction minimum support, minimum count

Output: L Frequent Itemset

• Tables:

Lk = Set of k- itemsets which are frequent

Ck = Set of k- itemsets which could be frequent

• Method:

Init. Let k=1

Generate L1 (frequent itemsets of length 1)

Repeat until no new frequent itemsets are identified

a) Generate C(k+1) candidate itemsets from Lk frequent itemsets

b) Count the support of each candidate by scanning the DB

c) Eliminate candidates that are infrequent, leaving only those that are frequent

3.2 Improve Apriori Efficiency

1. Hash-based itemset counting

2.Transaction reduction

3. Partitioning

4. EXPERIMENTAL ANALYSIS AND RESULTS

In this section, first we collect the dataset. Apply the dataset in Weka tool to find Classification results. The data were collected for Super Market details.

Key idea of algorithm is begin by generating or finding frequent itemsets with just one item(1-itemsets) and to recursively generate frequently itemset with 2 item, then with 3 itemsets and so on till no new itemset.

Support-> A=>B =P(AUB)

Confidence-> -A=>B =P(B/A)

4.1 Example of Rules:

SNO	ITEMS
C1	Milk, Egg, Bread
C2	Milk, Egg, Bread, Chip
C3	Milk, Egg, Chip
C4	Egg, Bread, Chip
C5	Bread, Milk, Chip,bear

Transaction database 1

4.1.1 First Candidate item set

ITEM	*SUPPORT COUNT*
Milk	*3*
Egg	4
Bread	4
Chip	4
Bear	*1*

$\frac{If,then}{total}$ =S $\frac{if,then}{if}$ =C

Assume Support %= 50

Minimum Support %=50

We select items whose support is $\geq minimum\,Support$

Support ratio $=\frac{Support}{Total\,number\,of\,Transaction}$

$= \frac{1}{5}100$ =20 %

After eliminating the bear

4.1.2 Frequent 1 – itemset and its support

ITEM	*SUPPORT COUNT*
Milk	*3*
Egg	4
Bread	4
Chip	4

Frequent 1 itemset 1

4.1.3 Second Candidate -item set

ITEM	*SUPPORT COUNT*
Milk, Egg	*3*
Milk, Bread	2
Milk, Chip	2
Egg, Bread	3
Egg, Chip	3
Bread, Chip	*3*

Cantidate2 itemset 2

After eliminating minimum support count {milk, bread}, {milk, Chip}

4.1.4 Frequent 2-itemset

ITEM	*SUPPORT COUNT*
Milk, Egg	*3*
Egg, Bread	3
Egg, Chip	3
Bread, Chip	*3*

Frequent 2 itemset

4.1.5 Third Candidate -itemset

ITEM	*SUPPORT COUNT*
Milk, Egg, Bread	*2*
Milk, Egg, Chip	2
Milk, Bread ,Chip	1
Egg, Bread, Chip	*2*

4.1.6 Frequent 4-itemset

ITEM	*SUPPORT COUNT*
Milk, Egg, Bread, Chip	*3*

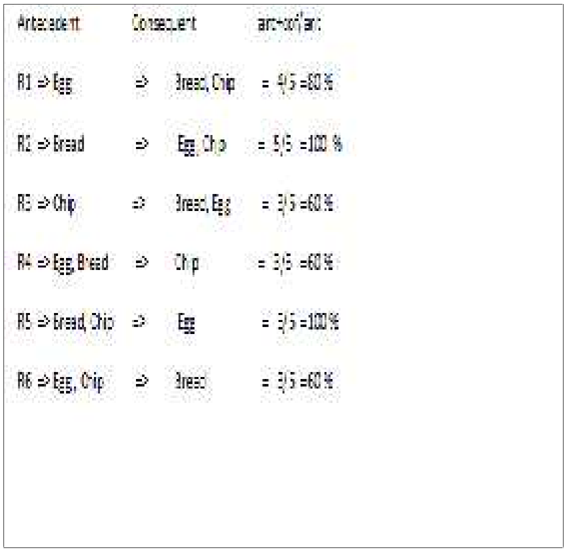
Frequent 2 itemset

L=L1 U L2 U L3 UL4

Milk, Egg, Bread, Chip

Consider the frequent 3 itemset Egg, Bread, Chip Subset is {Egg}, {Bread}, {Chip}, {Egg, Bread}, {Egg, Chip}, {Bread, Chip}

4.2 Comparison of support and minimum count



Antecedent	Consequent		ant+conf/ant
R1 => Egg	=>	Bread, Chip	= 4/5 =80 %
R2 => Bread	=>	Egg, Chip	= 5/5 =100 %
R3 => Chip	=>	Bread, Egg	= 3/5 =60 %
R4 => Egg, Bread	=>	Chip	= 3/5 =60 %
R5 => Bread, Chip	=>	Egg	= 3/5 =100 %
R6 => Egg, Chip	=>	Bread	= 3/5 =60 %

Compare to all the itemset {Bread, Chip},{Egg, Chip},{Egg}….

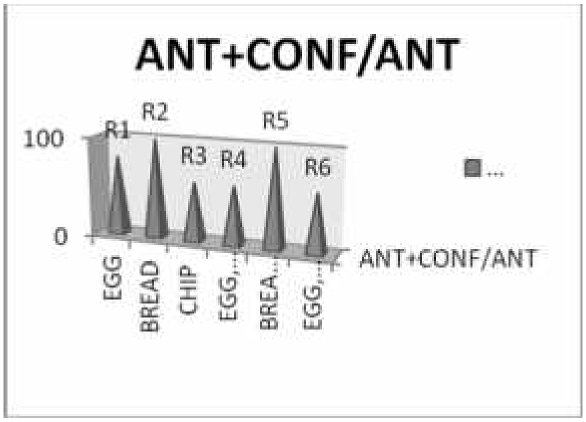


FIG:1 ITEMSETS COMPARISION

The performance of the classifiers depends on the characteristics of the data to be classified. The Random sub-sampling, k-fold cross validation and bootstrap method. In our study, we have selected k-fold cross validation for evaluating the classifiers. In k-fold cross validation, the initial data are randomly partitioned into k mutually exclusive subset or folds d1,d2,…,dk, each approximately equal in size. The training and testing is performed k times. In the first iteration, subsets d2, …, dk collectively serve as the training set in order to obtain a first model, which is tested on d1; the second iteration is trained in subsets d1, d3… dk and tested on d2; and so no.

5 PERFORMANCE EVALUATION FOR ASSOCIATION RULE

5.1 Apriori weka analysis

Minimum support: 0.3 (1 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 5

Best rules found:

1. ITEMS=Milk, Egg, Bread 1 ==> SNO=C1 1 conf:(1)

2. SNO=C1 1 ==> ITEMS=Milk, Egg, Bread 1 conf:(1)

3. ITEMS=Milk, Egg, Bread, Chip 1 ==> SNO=C2 1 conf:(1)

4. SNO=C2 1 ==> ITEMS=Milk, Egg, Bread, Chip 1 conf:(1)

5. ITEMS=Milk, Egg, Chip 1 ==> SNO=C3 1 conf:(1)

6. SNO=C3 1 ==> ITEMS=Milk, Egg, Chip 1 conf:(1)

7. ITEMS=Egg, Bread, Chip 1 ==> SNO=C4 1 conf:(1)

8. SNO=C4 1 ==> ITEMS=Egg, Bread, Chip 1 conf:(1)

9. ITEMS=Bread, Milk, Chip,bear 1 ==> SNO=C5 1 conf:(1)

10. SNO=C5 1 ==> ITEMS=Bread, Milk, Chip,bear 1 conf:(1)

5.2 Tertius weka analysis

1. /* 0.666667 0.000000 */ ITEMS = Milk, Egg, Bread ==> SNO = C1

2. /* 0.666667 0.000000 */ ITEMS = Milk, Egg, Bread, Chip ==> SNO = C2

3. /* 0.666667 0.000000 */ ITEMS = Milk, Egg, Chip ==> SNO = C3

4. /* 0.666667 0.000000 */ ITEMS = Egg, Bread, Chip ==> SNO = C4

5. /* 0.000000 0.800000 */ TRUE ==> ITEMS = Milk, Egg, Bread

6. /* 0.000000 0.800000 */ TRUE ==> ITEMS = Milk, Egg, Bread, Chip

7. /* 0.000000 0.800000 */ TRUE ==> ITEMS = Milk, Egg, Chip

8. /* 0.000000 0.800000 */ TRUE ==> ITEMS = Egg, Bread, Chip

9. /* 0.000000 0.800000 */ TRUE ==> ITEMS = Bread, Milk, Chip, bear

10. /* 0.000000 0.800000 */ TRUE ==> SNO = C1

11. /* 0.000000 0.800000 */ TRUE ==> SNO = C2

12. /* 0.000000 0.800000 */ TRUE ==> SNO = C3

13. /* 0.000000 0.800000 */ TRUE ==> SNO = C4

14. /* 0.000000 0.800000 */ TRUE ==> SNO = C5

Number of hypotheses considered: 70

Number of hypotheses explored: 70

6. Weka Tool:

Weka is a software environment that integrates several machine learning tools within a common framework and a uniform GUI. Classification and summarization are the main data mining tasks supported by the weka system.

7. CONCLUSION

Association rule approach for determining the performance of different methods has been developed to achieve high efficiency and classification sales accuracy and reduce the cost of itemset. The proposed system consists of two stages. 1. Frequent Itemset Generation 2. Rule Generation. After Finding Large itemsets and Generate rules from frequent itemsets. Generate all itemsets whose support ≥ minsup. Experimental results indicate that the proposed model with Association rule in Super Market dataset can be classified. In future may implement the same concept with advance association mining techniques. Advance technique are Generalized Association Rules, Multiple minimum supports,Multiple-Level Association Rules, Quantitative Association Rules, Using multiple minimum supports, Correlation Rules, Sequential patterns mining , Graph mining and so on.

ACKNOWLEDGEMENTS

The author would like to thank the management of Bharath University, India for the support and encouragement for this research work.

REFERENCES

[1].Jun Du, Charles X. Ling, Asking Generalized Queries to Domain Experts to Improve Learning, IEEE Transaction on Knowledge and Data Engineering, Vol.22, No,16, June 2010.

[2].J. Du and C. X. Ling. Active learning with generalized queries. In Proceedings of the 2009 IEEE International Conference on Data Mining, 2009.

[3]..L. Cao and C. Zhang, Domain-driven data mining: A practical methodology. International Journal of Data Warehousing and Mining, 2(4):49–65, 2006.

[4].Simon Tong, Active Learning: Theory and Applications, 2005.

[5]. S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In ICML '06: Proceedings of the 23rd international conference on Machine learning, pages 417–424, New York, NY, USA, 2006. ACM.

[6].G. Druck, G. S. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In S. H. Myaeng, D. W. Oard, F. Sebastian, T. S. Chua, M. K. Leong, S. H. Myaeng,D. W. Oard, F. Sebastian, T. S. Chua, and M. K. Leong, editors,SIGIR, pages 595–602. ACM, 2008.

[7]Abd. Wahab, M. H, Siraj, F and Yusoff, N. (2004). Log Mining Using Generalize Association Rules. In Proceedings of Master Final Project 2004 Presentation, UUM, Malaysia.

[8] Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the International ACM SIGMOD Conference, Washington DC, USA, pages 207–216.

[9] Agrawal, R. and Srikant, R. (1994). Fast Algorithm for Mining Association Rules. Proc. of the 20th VLDB Conference. Pp 487-499.

[10] Agrawal, R., and Srikant, R. (1995). Mining Sequential Patterns. In Proc. of the Eleventh International Conference on Data Engineering (ICDE), Taiwan. Pp 3-14.

[11] Batista, P and Silva, M (2001). Prospector dos Dados de Acesso a um Servidor de Noticias na Web, 2nd Coferencia sobre Redes de Compotators, Evora, Portogal.

[12] Boon Lay, C, Khalid, M and Yusof, R. (!999). Intelligent Database by Neural Network and Data Mining. In Proc. of Artificial Intelligent Applications in Industry, Kuala Lumpur. Pp 201-219.

[13] Borgelt, C. (2004). Apriori: Finding Association Rules/Hyperedges with the Apriori Algorithm. School of Computer Science, University of Magdeburg.

[14]Chen, M.-S., Jan, J., Yu, P.S. (1996). Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, (8:6). Pp 866.883.