International Journal of Scientific & Engineering Research, Volume 8, Issue 4, April-2017
ISSN 2229-5518

An 0verview of Disease Analysis using Association Rule Mining

D. Sheila Freeda, and M. Lilly Florence

Abstract: Data mining is used to analysis of massive data from various sources and retrieving useful information. This information can be converted to knowledge discovery process. Today large amount of data are used to stored healthcare system. These system include doctor details, disease information, patient information.Data mining techniques are applied to the medical datasets to discover suitable analysis and comparison relating to diabetes dataset. In this paper the author is apply data mining techniques and to compare the algorithms to find out co-occurrence of disease with the help of Association rule mining and Apriori Algorithm.

Keywords: Association Rule Mining, Apriori Algorithm Data Mining.

I. INTRODUCTION

Today many research is using data mining techniques for knowledge discovery from massive data.Data mining has many functions like classification, association, clustering and prediction. In this paper we implemented association rule based Apriori algorithm to extract knowledge from medical repositories for predicting the diseases. Appropriate and automated decision making system can help in achieving less death rate. Hence the medical data set is used in this work. In this work the authors are trying to compare the data mining techniques to find a correlation among diseases, so that if a patient is suffering with one disease, he/she may get the correlated diseases also. So we can alert and prevent the patient from diseases. The data set is collected from various hospitals. The other sections of this paper is organized as; section 2 describes the previous work carried out in this research area, section 3 and 4 introduces the association rule mining and apriori algorithm. Section 5 describes the medical data what we have used in this work. Section 6 explains the results and finding outs. Conclusion and future work are briefs in section 7.

II. RELATED WORKS

In [7] the authors were an attempt to find any relationship in a clinical database. They also describes the data mining processes. In [8] the authors implemented association rules for medical data to find association among attributes.

The authors in [1] proposed apriori data mining based on association rule for finding frequency of disease by patients. Clinical stat correlation prediction was developed to extract information from healthcare database which predict the relationship among primary disease and secondary disease [2]. In [3] the authors presented a methodology to identify the locally frequent diseases by implementing apriori mining technique. In this work the author collected data from medical centre and generate a frequent pattern to identify the frequent disease. In [5] the authors studied current technique of knowledge discovery in medical database using data mining. They also carried out some comparative study among techniques and concluded that the association rule mining perform well for medical data set. In [6] , the author surveyed various application of association rule mining in medical data in areas of nosocomial infections, adverse drug reactions, etc. By understanding the previous work carried out by various researchers in the area of data mining of disease analysis, the current work of the author is unique and it is need and important for the society.

III. ASSOCIATION RULE

Association rule is a rule-based machine learning method for discovering interesting relations between variables in large databases. Association rules are if/then statements that help touncover relationships between unrelated data in a database, relational database or other information repository. Association rules are used to find the relationships between the objects which are frequently used together. In addition to the above example association rules are employed today in many application areas including
Web usage mining, intrusion detection, Continuous production, and bioinformatics. In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions. There are two basic criteria that association rules uses, support and .confidence. It identifies the relationships and rules generated by analyzing data for frequently used if/then patterns.

i). Support

Support is an indication of how frequently the itemset appears in the database.The support of with respect to is defined as the proportion of transactions in the database which contains itemset In the example database, the

D. Sheila Freeda
Assist.Prof, Spurthy College of science and management studies, Bangalore, India
M. Lilly Florence
Professor, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

itemset has a support of since it occurs in 20% of all transactions (1 out of 5 transactions). The argument of is a set of preconditions, and thus becomes more restrictive as it grows (instead of more inclusive).[3]

ii). Confidence

Confidence is an indication of how often the rule has been found to be true.The *confidence* value of a rule, with respect to a set of transactions , is the proportion of the transactions that contains which also contains Confidence is defined as:For example, the rule has a confidence of in the database, which means that for 100% of the transactions containing butter and bread the rule is correct (100% of the times a customer buys butter and bread, milk is bought as well).

Note that means the support of the union of the items in X and Y. This is somewhat confusing since we normally think in terms of probabilities of events and not sets of items. We can rewrite as the probability where and are the events that a transaction contains itemset and , respectively.[4]Thus confidence can be interpreted as an estimate of the conditional probability , the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.[3][5]

$$Rule: X \Rightarrow Y \quad \begin{cases} Support = \dfrac{frq(X,Y)}{N} \\ Confidence = \dfrac{frq(X,Y)}{frq(X)} \end{cases}$$

IV. AIS ALGORITHM

The AIS algorithm [1] was the first algorithm proposed by Agrawal, Imielinski, and Swami for mining association rule. It focuses on improving the quality of databases together with necessary functionality to process decision support queries. In this algorithm only one item consequent association rules are generated, which means that the consequent of those rules only contain one item, for example, rules like X ∩ Y ⇒Z can be generated but not the rules like X ⇒Y ∩ Z. The databases were scanned many times to get the frequent item sets in AIS. To make this algorithm more efficient, an estimation method was introduced to prune those item sets candidates that have no hope to be large, consequently the unnecessary effort of counting those item sets can be avoided. Since all the candidate item sets and frequent item sets are assumed to be stored in the main memory, memory management is also proposed for AIS when memory is not enough. In AIS algorithm, the frequent item sets were generated by scanning the databases several times. The support count of each individual item was accumulated during the first pass over the database. Based on the minimal support count those items whose support count less than its minimum value gets eliminated from the list of item. Candidate 2-itemsets are generated by extending frequent 1-itemsets with other items in the transaction. During the second pass over the database, the support count of those candidate 2-itemsets are accumulated and checked against the support threshold. Similarly those candidate (k+1)-item sets are generated by extending frequent k-item sets with items in the same transaction. The candidate item sets generation and frequent item sets generation processiterate until any one of them becomes empty.

1. Candidate itemsets are generated and counted on-the-fly as the database is scanned.
2. For each transaction, it is determined which of the large itemsets of the previous pass are contained in this transaction.
3. New candidate itemsets are generated by extending these large itemsets with other items in this transaction.

candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently.

It generates candidate item sets of length from item sets of length

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

Figure 2. Example of AIS algorithm

V.SETM ALGORITHM

In the SETM algorithm, candidate itemsets [2] aregenerated on-the-fly as the database is scanned, but counted at the end of the pass. Then new candidate itemsets are generated the same way as in AIS algorithm, but the transaction identifier TID of the generating transaction is saved with the candidate itemset in a sequential structure. It separates candidate generation process from counting. At the end of the pass, the support count of candidate itemsets is determined by aggregating the sequential structure. The SETM algorithm [4] has the same disadvantage of the AIS algorithm. Another disadvantage is that for each candidate itemset, there are as many entries as its support value.

1. Candidate itemsets are generated on-the-fly as the database is scanned, but counted at the end of the pass.
2. New candidate itemsets are generated the same way as in AIS algorithm, but the TID of the generating transaction is saved with the candidate itemset in a sequential structure.
3. At the end of the pass, the support count of candidate itemsets is determined by aggregating this sequential structure.

VI. APRIORI ALGORITHM

4. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as *candidate generation*), and groups of

Database

TID	Items
100	1,3,4
200	2,3,5
300	1,3,2,5
400	2,5

C1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

Itemset	Support
{1,3}	2
{1,4}	1
{3,4}	1
{2,3}	2
{2,5}	3
{3,5}	2
{1,2}	1
{1,5}	1

C2

Itemset	TID
{1,3,4}	100
{2,3,5}	200
{1,3,5}	300
{2,3,5}	300

Figure 3. Example of SETM algorithm [2]

Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent -length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.At each step, the algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. accesses a field of the data structure that represents candidate set , which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies.

1. We have to find out the frequent itemset using Apriori algorithm.
2. Then Association rules will be generated using min.suppot&min.confidence.

Items	Count Number
I1	7
I2	8
I3	6
I4	2
I5	3
I6	1

a) C1

Large 1 Items
I1
I2
I3
I5

b) L1

Items	Count Number
I1,I2	5
I1,I3	4
I1,I5	3
I2,I3	4
I2,I5	3
I3,I5	1

c) C2

Large 2 Items
I1,I2
I1,I5
I2,I5
I2,I3
I1,I3

d) L2

Items	Count Number
I1,I2,I5	3
I1,I2,I3	2

e) C3

Figure 4. Example of Apriori algorithm [3]

TABLE I
COMPARISON OF ASSOCIATION RULE MINING ALGORITHMS

Characteristics	AIS	SETM	Apriori
Data support	Less	Less	Limited
Speed in initial phase	Slow	Slow	High
Speed in later phase	Slow	Slow	Slow
Accuracy	Very Less	Less	Less

VII. CONCLUSION

There are various association rule mining algorithms. In this paper we have discussed six association rule mining algorithms with their example: AIS, SETM, Apriori, Comparison is donebased on the above performance criteria. Each algorithm has some advantages and disadvantages. In this work we have implemented association rule mining on apriori algorithm which will be more useful for medical data. We briefly explains the association rule and apriori algorithm. We have implemented the association rule on selected medical data with the help of simulation tool and we generated set of association rules. We examined all the generated

association rule and selected only 7 relevant rules as most important rule to our work.From the above comparison we can conclude that, SETM algorithm performsbetter than all other algorithms discussed here.

REFERENCES

[1] Qiankun Zhao, Sourav S. Bhowmick, *Association Rule Mining: A Survey*, Technical Report, CAIS, Nanyang Technological University,Singapore, 2003

[2] KomalKhurana, Mrs. Simple Sharma, *A Comparative Analysis of Association Rules Mining Algorithms*, International Journal ofScientific and Research Publications, Volume 3, Issue 5 , May 2013 ISSN 2250-3153

[3] IshNathJhaSamarjeet Borah, *An Analysis on Association Rule Mining Techniques*, International Conference on Computing,Communication and Sensor Network (CCSN) 2012

[4] ManishaGirotra, KanikaNagpalSaloniinochaNeha Sharma *Comparative Survey on Association Rule Mining Algorithms,* International Journal of Computer Applications (0975 – 8887) Volume 84 – No 10, December 2013

[5] Sotiris Kotsiantis, DimitrisKanellopoulos, *AssociationRules Mining: A Recent Overview*, GESTS International Transactions on ComputerScience and Engineering, Vol.32 (1), 2006, pp. 71-82

[6] GagandeepKaur, ShrutiAggarwal , *Performance Analysis of Association Rule Mining Algorithms*, International Journal ofAdvanced Research in Computer Science and Software Engineering,Volume 3, Issue 8, August 2013, ISSN: 2277 128X

[7] PratikshaShendge, Tina Gupta, *ComparitiveStudy of Apriori& FP Growth Algorithms*, Indian journal of research, Volume 2, Issue 3,March 2013.

[8] Ming-Syan Chen, Jiawei Han, P.S.Yu, *Data mining: an overview from a database perspective*, IEEE Transactions on Knowledge andData Engineering, Volume:8, Issue: 6 ISSN: 1041-4347, 866 - 883

[9] Parita Parikh, Dinesh Waghela, *Comparative Study of Association Rule Mining Algorithms, Parita Parikh* et al, UNIASCIT, Vol 2 (1),
2012, 170-172, ISSN 2250-0987.

[10]. Agrawal, R., Imielinski, T & Swami A. Mining association rules between sets of items in large databases, Proceedings of ACM SIFMOD international conference on management of data, New York, pp. 207-216, 1993.

[11] Hanauer, D. A., Rhodes, D.R. &Chinnaiyan, A.M. Exploring Clinical associations using –omics based enrichment analyses PLoS One, 4(4) E5203 2009.

[12] Ordonez, C., Ezquerra, N.F & Santana, C. A. Constraining and summarizing association rules in medical data., Knowledge and information Systems, 3, pp. 1-2, 2006.

[13] Paetz J &Brause R W, A frequent patterens tree approach for rule generation with categorical septic shock patient data. Proceddings of the second international symposium on medical data analysis, London: Springer – Verlag, pp 207-12, 2001.

[14] D. KeranaHanirex, and M.A. DoraiRangaswamy. 2011. Effi-cient Algorithm for Mining Frequent Itemsets using Clustering Techniques.