International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-2 Issue-4 June-2012

Finding the Chances and Prediction of Cancer through Apriori Algorithm with Transaction Reduction

Shashank Singh[1], Manoj Yadav[2], Hitesh Gupta[3]
Department of Computer Science, PCST, Bhopal, India[1,2,3]

Abstract

Frequent pattern mining is an important task of data mining. It is essential for mining association, relevant and interesting links. In addition, it is widely used in data classification, clustering and other data mining tasks. Many effective, scalable algorithms have been developed in terms of frequent pattern mining. The Apriori algorithm is a classical frequent item sets generation algorithm and a milestone in the development of data mining. In this paper we apply the apriori algorithm with transaction reduction on cancer symptoms. We consider five different types of cancer and according to the classification we generate the candidate sets and minimum support to find the spreading of cancer. By this we can find the symptoms by which the cancer is spreading more and also about the highest spreading cancer type.

Keywords

Apriori, cancer, cancer symptoms, pattern mining

1. Introduction

Extraction of frequent patterns in transaction-oriented database is crucial to several data mining tasks such as association rule generation, time series analysis, classification, etc. Most of these mining tasks require multiple passes over the database and if the database size is large, which is usually the case, scalable high performance solutions involving multiple processors are required.

Being firstly introduced and researched by Agrawal et al. [1] in 1993, association rule mining is one of the important problems in data mining. The numbers of mined frequent item sets and association rules are usually enormous, especially in dense databases as well as with the small minimum supports and confidences. As usual, users only take care of the smaller sets of frequent item sets and association rules that satisfy given properties or constraints. The post-processing to select them wastes much time and can be repeated many times until users get the results that they desire. Hence, the problem of mining frequent item sets and association rules with constraints is reality. It has been receiving attentions of many researchers.

In [2][3][4][5][6] some authors researched on mining frequent item sets and association rules from the viewpoint of the user's interaction with the system. Srikant et al. [7] considered the problem of integrating constraints in the form of Boolean expression that appoint the presence or absence of items in rules.

The importance of data mining is increasing exponentially since last decade and in recent time where there is very tough competition in the market where the quality of information and information on time play a very crucial role in decision making of policy has attracted a great deal of attention in the information industry and in society as a whole. There is very large amount of data availability in real world and it is very difficult to excess the useful information from this huge database and provide the information to which it is needed within time limit and in required pattern. So data mining give the way to remove the noise from data and extract information from huge database and present it in the form in which it is needed for each specific task. The use of data mining is very vast. It is very helpful in application like to know the trend of market, fraud detection, and shopping pattern of customers, production control and science exploration etc. in one sentence data mining is mining of knowledge from huge amount of data. Using data mining we can predict the nature or behavior of any pattern. There is several research works which has been adopting data mining techniques in several ways. In [8][9] author suggested data mining techniques in mobile devices for better computing. In [10] predictive apriori and distributed grid based apriori algorithms was suggested for knowledge extraction. In [11] new optimization algorithm called APRIORI-IMPROVE based on the insufficient of apriori was proposed to save time and space. In [12] Rough set theory apriori was presented solves the problems of apriori algorithm to improve the efficiency of the methodology/algorithm. In [13] apriori algorithm was used for biological sequence analysis.

The remaining of this paper is organized as follows. We discuss Apriori Algorithm in Section 2. In Section 3 we discuss about Frequent Item set Mining. In section 4 we discuss about Recent Scenario. In section 5 we discuss about the proposed approach. Conclusions are given in Section 6. Finally references are given.

2. Apriori Algorithm

In computer science and data mining, Apriori[1] is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions, or having no timestamps (DNA sequencing).

As is common in association rule mining, given a set of item sets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. 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 candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data.
The Apriori Algorithm is an influential algorithm for mining frequent item sets for boolean association rules. The key concepts are following:

- Frequent Item sets: The sets of item which has minimum support.
- Apriori Property: Any subset of frequent item set must be frequent.
- Join Operation: To find Lk , a set of candidate k-item sets is generated by joining Lk-1 with itself.
- Find the frequent item sets: the sets of items that have minimum support. A subset of a frequent item
- set must also be a frequent item set • i.e., if {AB} is a frequent item set, both {A} and {B} should be a frequent item set. Iteratively find frequent item sets with cardinality from 1 to k (k-item set)
- Use the frequent item sets to generate association rules.

3. Frequent Item Set Mining

Frequent Item Set Mining is a method for market basket analysis. It aims at finding regularities in the shopping behavior of customers of supermarkets, mail-order companies, on-line shops etc. More specifically: Find sets of products that are frequently bought together.

Possible applications of found frequent item sets:

- Improve arrangement of products in shelves, on a catalog's pages etc.
- Support cross-selling (suggestion of other products), product bundling. Fraud detection, technical dependence analysis etc.
- Often found patterns are expressed as association rules, for example: If a customer buys bread and milk, then she/he will probably also buy cheese.

Let B={I1,….Im} be a set of items. This set is called the item base. Items may be products, special equipment items, service options etc. Any subset I subset B is called an item set. An item set may be any set of products that can be bought (together). Let T = (t1,….tn) with $\forall k$; $1 \leq k \leq n$:t_k subset of B be a vector of transactions over B. This vector is called the transaction database. A transaction database can list, for example, the sets of products bought by the customers of a supermarket in a given period of time. Every transaction is an item set, but some item sets may not appear in T. Transactions need not be pair wise different: it may be $t_j=t_k$ for $j\neq k$. T may also be defined as a bag or multiset of transactions. The set B may not be explicitly given, but only implicitly assigned as a bag or multiset of transactions. The set B may not be explicitly given, but only implicitly as $B = U_{k=1}^{n} t_k$.

4. Recent Scenario

In 2010, Jitao Zhao et al. [14] analysis the characteristics of medical data, and propose a general framework for medical data mining, which consists of medical data selection, medical data preprocessing, medical data mining and knowledge evaluation. The framework could provide the methodical steps for scholars who are interested in

the related researches of data mining and medical domain.

In 2010, Qin Li et al. [15] proposed to discover closed frequent item sets with a simple linear list structure called the Frequent Pattern List (FPL) in transaction database. The approach selects representation patterns from candidate item sets to reduce combinational space of frequent patterns. By performing two operations, signature vertex conjunction and vertex counting, it simplify the process of closed item sets generation.

In 2011, Hnin Wint Khaing et al. [16] presented an efficient approach for the prediction of heart attack risk levels from the heart disease database. Firstly, the heart disease database is clustered using the K-means clustering algorithm, which will extract the data relevant to heart attack from the database. This approach allows mastering the number of fragments through its k parameter. Subsequently the frequent patterns are mined from the extracted data, relevant to heart disease, using the MAFIA (Maximal Frequent Itemset Algorithm) algorithm. The machine learning algorithm is trained with the selected significant patterns for the effective prediction of heart attack. They have employed the ID3 algorithm as the training algorithm to show level of heart attack with the decision tree. The results showed that the designed prediction system is capable of predicting the heart attack effectively.

5. Proposed Approach

We are using apriori algorithm for finding the frequency of cancer in a human being. For this we first present the apriori algorithm. Apriori is a seminal algorithm proposed by R. Agrawal nad R. Srikant in 1994 for mining frequent item sets for Boolean association rules. In this paper we present apriori with transaction reduction approach for finding cancer from the cancer symptoms.

Apriori Algorithm: (by Agrawal et al at IBM Almaden Research Centre) can be used to generate all frequent Item set:

Pass 1

1. Generate the candidate itemsets in C_1
2. Save the frequent itemsets in L_1

Pass k

1. Generate the candidate itemsets in C_k from the frequent itemsets in L_{k-1}
   1. Join L_{k-1} p with L_{k-1}q, as follows: insert into C_k select $p.\text{item}_1, p.\text{item}_2, \ldots, p.\text{item}_{k-1}, q.\text{item}_{k-1}$ from L_{k-1} p, L_{k-1}q where $p.\text{item}_1 = q.\text{item}_1, \ldots p.\text{item}_{k-2} = q.\text{item}_{k-2}, p.\text{item}_{k-1} < q.\text{item}_{k-1}$
   2. Generate all (k-1)-subsets from the candidate itemsets in C_k
   3. Prune all candidate itemsets from C_k where some (k-1)-subset of the candidate itemset is not in the frequent itemset L_{k-1}
2. Scan the transaction database to determine the support for each candidate itemset in C_k
3. Save the frequent itemsets in L_k

For better understanding our approach we take the example which is shown below in the following Tables:

Table 1: Leukemia

Symptoms of Leukemia
Fever
Weakness and Fatigue
Swollen or Bleeding Gums
Headaches
Swollen Tonsils
Bone Pain
Paleness
Red Spots
Weight Loss

Table 2: Lung Cancer

Symptoms of Lung Cancer
Shortness of Breadth
Chest Pain
Hemoptysis
Shoulder Pain
Paralysis
Stroke

Table 3: Oral Cancer

Symptoms of Oral Cancer
White Patches
Mixed Red and White Patches
Red Patches
Bleeding
Loose Teeth
Lump in Neck

Table 4: Skin Cancer

Symptoms of Skin Cancer
A small lump (spot or mole) that is shiny, waxy, pale in color, and smooth in texture.
A red lump (spot or mole) that is firm
A sore or spot that bleeds or become crusty
Rough and scaly patches on the skin
Flat scaly areas of the skin that are red or brown
Any new growth that is suspicious

Table 5: Thyroid Cancer

Symptoms of Thyroid Cancer
A lump, or nodule, in the front of the neck
Hoarseness or difficulty speaking in a normal voice;
Swollen lymph nodes, especially in the neck;
Difficulty swallowing or breathing
Pain in the throat or neck

For better understanding of our approach we taken five types of cancer and their symptoms and then apply apriori algorithm on them which is shown in the below tables. A, B, C, D and E subsequently represent the type of cancer which is given in the table 1 to table 5. T1, T2, T3, T4 and T5 are the patient checkup transaction, if the symptoms from 1 to 5 shown in table 1 to table 5 is present we put the value 1 otherwise we put 0. We apply the apriori algorithm on table 6 which is shown from table 7.

Table 6: Patients Transaction

TID	A	B	C	D	E
T1	1	1	1	0	0
T2	1	1	1	1	1
T3	1	0	1	1	0
T4	1	0	1	1	1
T5	1	1	1	1	0

Table 7: Pass1:C1

Item set(X)
A
B
C
D
E

Table 8: L1

Itemset(X)	Supp(X)
A	100%
B	60%
C	100%
D	80%
E	40%

Table 9: Pass 2 C2

Itemset(X)
A,B
A,C
A,D
A,E
B,C
B,D
B,E
C,D
C,E
D,E

Table 10: L2

Itemset(X)	Supp(X)
A,B	60%
A,C	100%
A,D	80%
A,E	40%
B,C	60%
B,D	40%
B,E	20%
C,D	80%
C,E	40%
D,E	40%

Nothing pruned since all subsets of these item sets are infrequent.

Table 11: L2 after saving only the frequent item sets

Itemset(X)	Supp(X)
A,B	60%
A,C	100%
A,D	80%
A,E	40%
B,C	60%
B,D	40%
C,D	80%
C,E	40%
D,E	40%

To create C3 only look at items that have the same first item (in pass k, the first k - 2 items must match)

Table 12: C3

Itemset(X)
A,B,C

A,B,D
A,C,D
A,C,E
A,D,E
B,C,D
C,D,E

Pruning eliminates ABE since BE is not frequent Scan transactions in the database.

Table 13: L3

Itemset(X)	Supp(X)
A,B,C	60%
A,B,D	40%
A,C,D	80%
A,C,E	40%
A,D,E	40%
B,C,D	40%
C,D,E	40%

First k - 2 = 2 items must match in pass k = 4

Table 14: Pass 4 C4

Itemset(X)
A,B,C,D
A,C,D,E

After that we perform pruning on the minimum support to find the probability of the symptom.

6. Conclusion

Frequent item set mining is one of the most important areas of data mining specifically in association. Existing implementation of Apriori based algorithms focus on cancer dataset. We consider the item values in generating the frequency based on attribute. Each attribute in the dataset may consist of multiple values. The computation is based on Apriori concept. This algorithm uses the concept of Apriori with transaction reduction algorithm, which finds the symptoms of cancer efficiently.

References

[1] R. Agrawal, T. Imielinski, and N. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the ACM SIGMOID, pp. 207-216, 1993.

[2] T.N. Anh, D.V. Hai, T.C. Tin, and L.H. Bac, "Efficient Algorithms for Mining Frequent Itemsets with Constraint," in Proceedings of the third international conference on knowledge and systems engineering, 2011.

[3] R.J. Bayardo, R. Agrawal, and D. Gunopulos, "Constraint-Based Rule Mining in Large, Dense Databases," Data Mining and Knowledge Discovery, Kluwer Academic Pub., vol. 4, No. 2/3, pp. 217–240, 2000.

[4] G. Cong and B. Liu, "Speed-up Iterative Frequent Itemset Mining with Constraint Changes," ICDM, pp. 107-114, 2002.

[5] A.J. Lee, W.C. Lin, and C.S.Wang, "Mining Association rule with multi-dimensional constraints," Journal of Systems and Software, no. 79, pp. 79-92, 2006.

[6] R.T. Nguyen, V.S. Lakshmanan, J. Han, and A. Pang, "Exploratory Mining and Pruning Optimizations of Constrained Association Rules," in Proceedings of the 1998 ACM-SIG-MOD Int'l Conf. on the Management of Data, pp. 13-24, 1998.

[7] R. Srikant, Q. Vu, and R. Agrawal, "Mining association rules with item constraints," in Proceeding KDD'97, pp. 67-73, 1997.

[8] Dubey, A.K.; Shandilya, S.K., "A comprehensive survey of grid computing mechanism in J2ME for effective mobile computing techniques," Industrial and Information Systems (ICIIS), 2010 International Conference on , vol., no., pp.207,212, July 29 2010-Aug. 1 2010.

[9] Ashutosh Dubey and Shihir Shandilya," Exploiting Need Of Data Mining Services in Mobile Computing Environments", Computational Intelligence and Communication Networks (CICN), 2010.

[10] Sumithra, R.; Paul, S., "Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery," Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on , pp.1,5, 29-31 July 2010.

[11] Rui Chang; Zhiyi Liu, "An improved apriori algorithm," Electronics and Optoelectronics (ICEOE), 2011 International Conference on, vol.1, no., pp.V1-476, V1-478, 29-31 July 2011.

[12] Chen Chu-xiang; Shen Jian-jing; Chen Bing; Shang Chang-xing; Wang Yun-cheng, "An Improvement Apriori Arithmetic Based on Rough Set Theory," Circuits, Communications and System (PACCS), 2011 Third Pacific-Asia Conference on , vol., no., pp.1,3, 17-18 July 2011.

[13] Fan Min; Youxi Wu; Xindong Wu, "The Apriori property of sequence pattern mining with wildcard gaps," Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference, pp.138-143, December 2010.

[14] Jitao Zhao and Ting Wang, "A General Framework for Medical Data Mining", 2010 International Conference on Future Information Technology and Management Engineering.

[15] Qin Li and Sheng Chang, "Generating Closed Frequent Itemsets with the Frequent Pattern List", IEEE 2010.
[16] Hnin Wint Khaing," Data Mining based Fragmentation and Prediction of Medical Data", IEEE 2011.

I had done BE(IT) from RKDF IST Bhopal in 2007, currently I am pursuing ME(CSE) from Patel College of Science and Technology Bhopal.