ORIGINAL ARTICLE

Application of market–basket analysis on healthcare

Abishek B. Rao[1] · Jammula Surya Kiran[1] · Poornalatha G[1]

Received: 6 May 2020 / Revised: 18 April 2021 / Accepted: 20 August 2021 / Published online: 27 August 2021

Abstract Data analysis plays a vital role in the present era as it helps us to understand the patterns by exploring it in meaningful ways. Market—basket is one of the main methods used to find frequently occurring items in a transactional database and many researchers use the Apriori algorithm for this purpose. This paper presents the application of Market Basket Analysis to the healthcare section. The present work tries to find frequent diseases that occur together in an area by using the Apriori algorithm. This could help the residents of an area to be more cautious about the frequently occurring diseases and take all possible precautionary measures to safeguard their health. In addition, it could also help the doctors so that, they are ready with required medications to treat the patients.

Keywords Frequent disease · Frequent itemset · Healthcare · Market-basket analysis · Common symptom

1 Introduction

In general, thousands of transactions are recorded in our day to day life, which includes, various fields such as marketing, finance, health care, etc. In the present-day scenario, the usage of manual records for storing any sort of information has reduced due to advancements in technology.

Every Organization, Educational Institution, Company are maintaining their database to store their records. Every database related to that domain has an ample amount of data related to it, for example, educational institution database has the information related to every student of their institution, which helps to maintain the records of all students as well as understanding the hidden patterns.

As these databases consists of several thousands of records, it is impossible to read the database manually, and extract the knowledge out of it. So, various data mining techniques (association, clustering, classification, and prediction) are used to extract hidden and useful knowledge from the database.

Market Basket Analysis is one of the main techniques of data mining that focuses on “purchase patterns” of the customers. The aim of this analysis is, to identify the interesting patterns by considering the products bought together by the customers. To identify the patterns, analysis of the products bought together by the customers is performed. This requires details on the recorded transactions, which includes the information of customers such as, the item bought together by a customer during a visit to the supermarket. This kind of study gives a clear idea about the items preferred by customers.

Data mining has immense potential for exploring the meaningful and hidden patterns in the medical database also. In the present work, an online repository of healthcare

✉ Poornalatha G
poornalatha.g@manipal.edu

Abishek B. Rao
abishekbrao1996@gmail.com

Jammula Surya Kiran
suryakiranjammula4919@gmail.com

[1] Information and Communication Technology Department, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

dataset which includes Electronic Health Records (EHR) of several patients, the information about patients' details from different countries, diseases he/she is suffering from, year in which the patient got affected by the diseases etc. is used. The current work identifies all the frequent diseases that are occurring in an area by analyzing this medical data set.

The remaining part of the paper is organized as follows: the literature review is presented in Sect. 2, details of the methodology are discussed in Sect. 3, results and observations are detailed in Sect. 4, the conclusion is given in Sect. 5 followed by the references at the end.

2 Literature review

Market Basket Analysis is the search for meaningful associations in a customer purchased data. The Market Basket Analysis is done using different data mining algorithms.

Market Basket Analysis can be implemented by using the Apriori algorithm (Borgelt and Kruse 2002; Nengsih 2015). This algorithm determines the frequent data or item set from the transaction database. It determines the frequent item sets from the database using candidate item set generation.

Market Basket Analysis can also be implemented using the FP-tree algorithm (Chavan et al. 2014). It visualizes the data in the form of a tree. It creates an FP tree using two iterations over data. This algorithm generates frequent item sets without the generation of candidate item sets.

Market Basket Analysis using the FP-growth algorithm (Min 2010); (Yongmei and Yong 2009) uses a technique in which it will find long frequent patterns with several short recursive modes, then it connects the least frequent item sets to supply good selectivity. It reduces the search overhead.

Market Basket Analysis is implemented using the FP Bonsai algorithm (Gayathri 2017). It is implemented to mine the frequent patterns that will result in efficient frequent item sets that highlights monotone constraints. The usage of the FP bonsai algorithm has proven to be having higher efficiency and results in lesser execution time.

The Apriori and FP growth are combined to form frequent pattern mining of item sets (Singh et al. 2014). It concentrates on the web patterns of server log files. This comparison can be applied for Market Basket Analysis as well.

Many authors of which, few are given below have used different techniques and algorithms: Reference (Bhargav et al. 2014) uses a neural network to implement Market Basket Analysis. Reference (Trnka 2010)-(Han et al. 2000) uses an approach of single-layer feed-forward partially connected neural network. It takes less time since repeated scanning of the database is not done and helps in improving the efficiency. Reference (Masrani and Poornalatha 2018) uses the Naïve Bayes algorithm.

Apriori is also used in medical research with the aim of reducing the manual effort of the health care people in analyzing the data related medical domain. For example, to detect the schistosomiasis disease (Ali et al. 2019), and to find frequent diseases (Ilayaraja and Meyyappan 2013). An efficient technique for Apriori algorithm in medical data mining is proposed by Khan et al. (2019). A novel approach was proposed for disease prediction using weighted association rules in Lakshmi and Vadivu 2019. Association rule mining is used (Harahap et al. 2018) to recommend the medical need. Table 1, summarizes the above mentioned existing research works.

As evident from the Table 1, there are many applications based on the market-basket analysis. However, analysis of medical data plays an important role, as it has the potential to ease the work of doctors. The existing work focuses on finding frequent diseases, recommending the need or medication, try to predict the disease based on the symptoms etc. However, limited focus is observed in finding the diseases that could occur frequently together. Hence, the present work considers the application of market-basket analysis to health care, in particular to find frequently occurring diseases. Accordingly, a disease data set with 43 several types of diseases is considered. As the number of diseases was 43 and of various kinds, the traditional Apriori algorithm is used in the present work. In addition, high support count is considered to reduce many database scans required by the basic Apriori algorithm.

3 Methodology

The procedure adopted in the present work is as shown in Fig. 1 (Saunders et al. 2009). The philosophies, approaches, strategies, choices, time horizons, techniques and procedures which are useful for the present work are highlighted in the diagram. The various components are as follows:

Longitudinal The approach used here is a longitudinal approach, in which data is studied for a longer period.

*Mono method*The method used in the proposed paper is the Apriori. Apriori is applied to the dataset, and reduction is applied to mine frequent patterns among the diseases that are occurring in a particular year.

Experiment The experiment consists of two phases. They are Candidate Itemset generation and Pruning. The Candidate Itemset generation, generates all candidate Item sets from the given set of items, and Pruning is carried out based on the user defined minimum support. Pruning

Table 1 Literature review summary

Algorithm	Purpose	Remarks
Apriori (Borgelt and Kruse 2002; Nengsih 2015)	Market-basket analysis	Finds frequent item sets. Requires more number of database scans
FP-tree (Chavan et al. 2014; Failed 2010; Yongmei and Yong 2009; Gayathri 2017)	Market-basket analysis	Finds frequent item sets. Visualizes data as a tree. Requires less number of database scans
Apriori and FP-tree (Failed 2014)	Application of market-basket analysis	Finds the frequent web patterns
Neural network based (Bhargav et al. 2014; Trnka 2010; Han et al. 2000)	Market-basket analysis	Avoids repeated database scans
Naïve Bayes (Masrani and Poornalatha 2018)	Application of market-basket analysis	Analyzes the twitter sentiments
Apriori based (Ali et al. 2019; Ilayaraja and Meyyappan 2013; Khan et al. 2019; Lakshmi and Vadivu 2009; Harahap et al. 2018),	Application of market-basket analysis	Analyzes medical domain related data

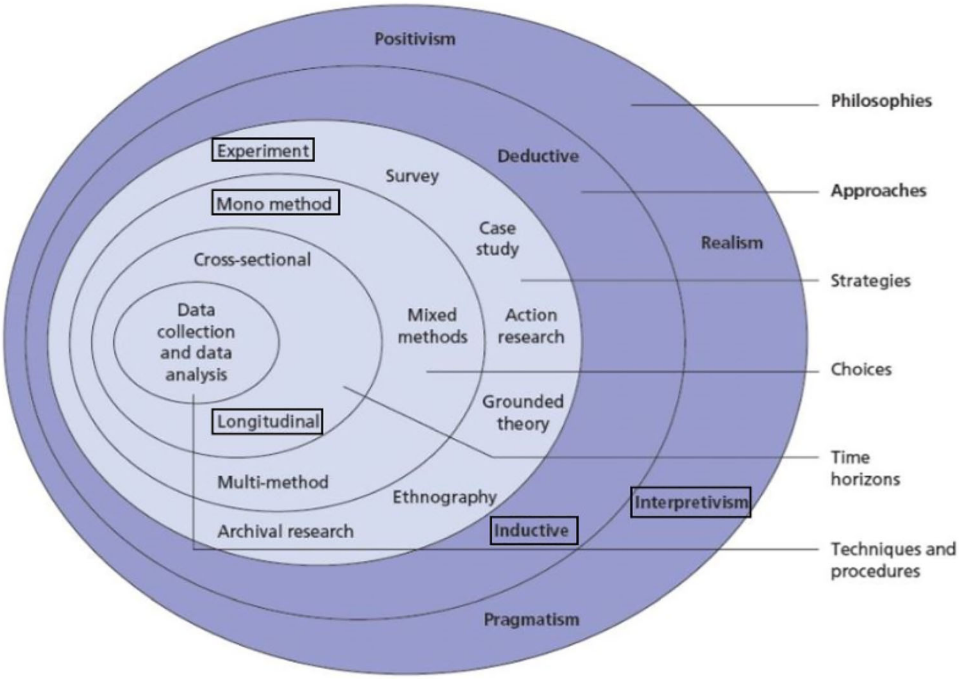

Fig. 1 Onion diagram of the present method (Saunders et al. 2009)

method follows a property in which an itemset is frequent if and only if all its subsets are frequent.

Inductive The dataset, which is going to be analyzed, is found to fit into the existing algorithm. So, the inductive approach is followed.

Interpretive The results obtained in the experiment gives frequent diseases occurring in a particular year.

Figure 2 depicts all the steps involved in the current method. The detailed steps are given below:

Data collection and data analysis

Data collection or Dataset related to the present work is taken from an online repository consisting of patients Electronic Health Records (EHR) that includes forty-three diverse types of diseases from which the patients are suffering.

Candidate itemset generation

In this step, all 1-item disease sets are considered as, an initial candidate itemset and the support count of each item is calculated. The items satisfying the user defined minimum support are placed in 1-item disease frequent set.

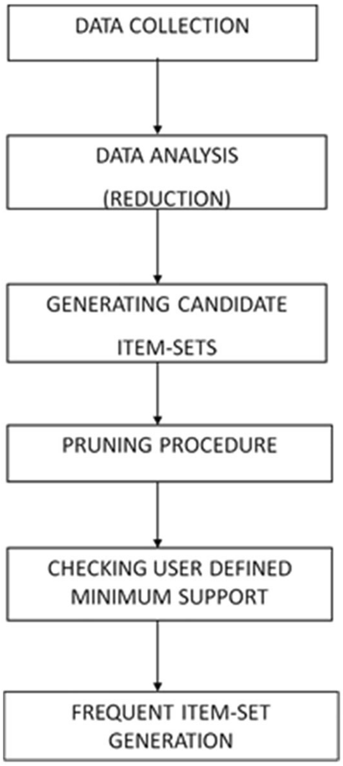
DATA COLLECTION
DATA ANALYSIS (REDUCTION)
GENERATING CANDIDATE ITEM-SETS
PRUNING PROCEDURE
CHECKING USER DEFINED MINIMUM SUPPORT
FREQUENT ITEM-SET GENERATION

Fig. 2 methodology steps

After 1-item frequent set calculation, other frequent item sets are determined by using the Apriori algorithm.

C. *Pruning procedure*

The Pruning step removes all infrequent items based on the subsets of an itemset. After 1-item frequent set calculation, the first candidate itemset generation is used to generate Candidate Item sets immediately. After generating candidate item sets, for each pass, Pruning procedure is used to eliminate all infrequent diseases.

D. *Checking user defined minimum support*

After Pruning, all infrequent item sets whose subsets are not found are removed and Support is calculated for the item in candidate set, the items, which are having user defined minimum support are placed in, frequent disease itemset.

E. *Frequent itemset generation*

Finally, frequent item sets of all the passes of the algorithm are combined to give Global frequent itemset, which gives all the frequent diseases occurring in a particular year.

4 Result and observations

After applying the Apriori algorithm to the dataset taken, frequent disease item sets are found. All frequent item sets: 1-item, 2-item, 3-item, etc. are generated from the candidate item sets. These frequent disease item sets, help in taking preventive measures to overcome diseases since frequent disease item sets are associated with common symptoms. Table 2 shows the sample input data set and Table 3 shows the sample output obtained. The minimum support count considered for this experiment is 95%.

Consider the 1-item frequent set from the result, which refers to the diseases, which are occurring often (refer first row of Table 2). Consider the 2-item frequent set {1,7} from the result where 1 refers to Amoebiasis and 7 refers to Botulism, these diseases have common symptoms like abdominal pain, nausea, and fever. Consider the 3-item frequent set {1,7,8} from the result, where 1 refers to the

Table 2 Sample input data set of diseases

ID	Disease name
1	Amoebiasis
2	Anaplasmosis and ehrlichiosis
3	Babesiosis
4	Cholera
5	Botulism, other
6	Botulism, foodborne
7	Botulism, wound
8	Brucellosis
9	Campylobacteriosis
10	Chlamydia
11	Ciguatera
12	Dengue
13	Coccidioidomycosis
14	Creutzfeldt-Jakob disease and other transmissible spongiform encephalopathies
15	Cryptosporidiosis

Table 3 Frequent item sets

Sl. no	Frequent itemset size	Frequent sets of diseases
1	1	{1}, {7}, {8}, {9}, {11}, {12},{13}, {17}, {19}, {20}, {22}, {23}, {27}, {34, {38}, {40}, {42}, …
2	2	{1,7}, {1,8}, {1,9}, {1,11}, {1,12}, {1,13}, {1, 17}, {1,19}, {1, 20}, …
3	3	{1,7,8}, {1,7,9}, {1,7,11} {1,7,12}, {1,7,13}, {1,7,17},…
4	4	{1,7,8,9}, {1,7,8,11}, {1,7,8,12}, {1,7,8,17}, …
5	5	{1,7,8,9,11}, {1,7,8,9,12}, {1,7,8,9,13}, {1,7,8,9,19}, …

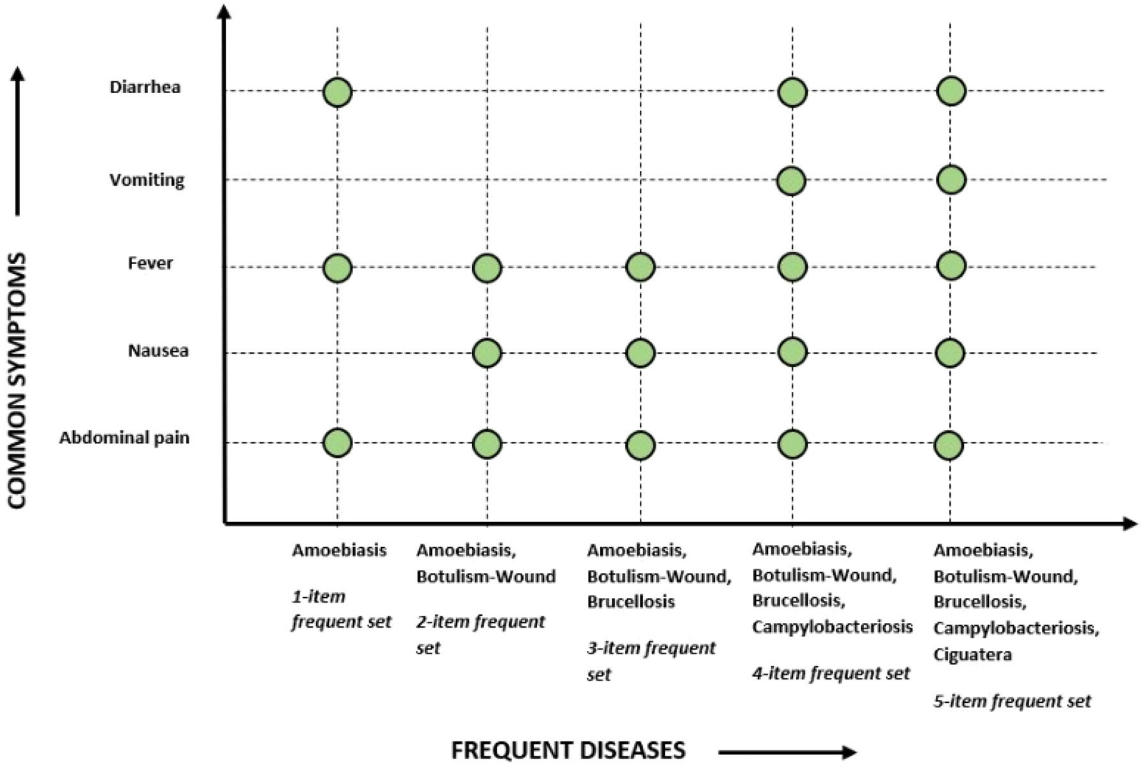

Fig. 3 A web plot showing frequent diseases with the associated symptoms

disease Amoebiasis, 7 refers to the disease Botulism from a wound, and 8 refers to the Brucellosis, these diseases have common symptoms like nausea, vomiting, and abdominal pains. Since there are frequent itemsets, one more 3-itemset {1,7,20} can be considered which refers to the diseases Amoebiasis, Botulism, and HIV respectively. All these diseases contain common symptoms like fever and abdominal pain. Consider the 4-item frequent set {1,7,8,11} from the result which refers to diseases like Amoebiasis, Botulism, Brucellosis, and Ciguatera Fish poisoning disease respectively, all these diseases have a common symptom of nausea and vomiting in general. If frequent item sets like 5-item, 6-item are considered then the diseases, having the common symptoms will not be more specific. For the diseases in the less frequent item sets, the common symptoms of the diseases will be more specific than the higher frequent item sets.

Similarly, all other diseases that occur frequently together are found by the current work as shown in Table 3. By looking into this information, doctors can plan the required necessary medications well in advance to ensure the safety of their patients. Health workers can create awareness among common citizens regarding the same so that, they can protect themselves from being affected with possible diseases. Thus, the Table 3 gives information regarding the frequently occurring diseases as well as diseases that occur together frequently. A web plot of the sample frequent diseases and their associated symptoms is depicted in Fig. 3.

5 Conclusion

The frequent itemsets from the result give the frequently occurring diseases, which are having common symptoms that help the doctors to give proper medications according to the symptoms of the disease. As the proposed work can give details of commonly occurring diseases based on year, the medical or health authorities can take preventive measures to take care of the possible frequently occurring diseases. In addition, an awareness could be created among citizens about these diseases, their common symptoms etc. so that, they can take required precautionary measures.

Funding Open access funding provided by Manipal Academy of Higher Education, Manipal.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Ali Y, Farooq A, Alam TM, Farooq MS, Awan MJ, Baig TI (2019) Detection of schistosomiasis factors using association rule mining. IEEE Access 7:186108–186114

Bhargav A, Mathur RP, Bhargav M (2014) "Market basket analysis using artificial neural network," International Conference for Convergence for Technology, pp. 1–6

Borgelt C, Kruse R (2002) Induction of association rules: apriori implementation. Compstat 1:395–400

Chavan G, Gaikwad N, Samal T, Sonule A, Palivela H, and Patil AC (2014) "Various mining techniques defined for mining product valuation instances in market basket data," International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE), pp. 1–6

Gayathri B (2017) "Efficient market basket analysis based on FP-Bonsai," 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 788–792

Han J, Pei J, and Yin Y (2000) "Mining frequent patterns without candidate generation," ACM SIGMOD International Conference on Management of Data, pp.1–12.

Harahap M, Husein AM, Aisyah S, Lubis FR, Wijaya BA (2018) Mining association rule based on the diseases population for recommendation of medical need. J of Phys Conf Series 1007:12–17

Ilayaraja M, and Meyyappan T (2013) "Mining medical data to identify frequent diseases using Apriori algorithm," 2013 International conference on Pattern Recognition, Informatics, and Mobile Engineering, pp.194–199

Khan MA, Pradhan SK, Fatima H (2019) An efficient technique for Apriori algorithm in medical data mining. Lecture Notes Netw. Syst. Springer Nature 32:187–195

L. Min, W. Chunyan, and Y. Yuguang, "The Research of FP-Growth Method Based on Apriori Algorithm in MDSS," *2010 International Conference on Digital Manufacturing & Automation*, Changsha, pp. 770–773, 2010.

Lakshmi KS, Vadivu G (2019) A novel approach for disease comorbidity prediction using weighted association rule mining. J Ambient Intell Humanized Comput. https://doi.org/10.1007/s12652-019-01217-1

Masrani M, and Poornalatha G (2018), "Twitter Sentiment Analysis using modified naïve bayes algorithm," International Conference on Information Systems Architecture and Technology, pp.171–181.

Nengsih W (2015) "A comparative study on market basket analysis and apriori association technique," 3rd International Conference on Information and Communication Technology (ICoICT), 461–464

Saunders M, Lewis P, and Thornhill A (2009) "Research methods for business students", Fifth Edition, Pearson Education Limited, England

Singh AK, Kumar A, and Maurya A K (2014) "An empirical analysis and comparison of apriori and FP- growth algorithm for frequent pattern mining," 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, pp. 1599–1602, 2014.

Trnka A, "Market Basket Analysis with Data Mining methods (2010)" International Conference on Networking and Information Technology, pp. 446–450

Yongmei L and Yong G (2009) "Application in market basket research based on FP-growth algorithm," 2009 WRI World Congress on Computer Science and Information Engineering, pp. 112–115

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.