Association Rule Mining based on a Modified Apriori Algorithm in Heart Disease Prediction

Anirudh Batra
Mohanasundaram R
Rishin Haldar

SCOPE – School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

Abstract – Heart diseases is the principal source of death in numerous nations. To limit this amount of deaths can be a tedious task since it will involve a significant change in our lifestyles, and in some cases, it may occur due to circumstances beyond our control. Nevertheless, this number can be reduced by using an efficient detection technique. This is where data mining comes in. Although several tests have to be conducted in order to detect heart disease with accuracy, this number of tests can be qualified using data mining. This study aims at introducing a more efficient version of Apriori algorithm and extracting several hidden patterns from a dataset gathered from hospitals and clinics which are significant in the prediction of heart diseases.

Key Words – heart disease, data mining, apriori, patterns, prediction

1. INTRODUCTION

Every year, a huge amount of medical data is accumulated by the healthcare industry. Using Apriori algorithm and by setting association rules, we can reduce the amount of deaths due to heart diseases. By analysing the efficiency of the legacy Apriori algorithm, a modified algorithm has been proposed to improve the efficacy of the Apriori algorithm by limiting the scale of the candidate item set.

For the purposes of this study, a Heart Disease Data Warehouse has been created containing heart patients' data which has been obtained from several conducted tests.

With the help of Data Mining, informative data can be extracted from bulk raw data which can be interpreted by humans. Association Rule Mining is regarded as one of the most resourceful applications of data mining. This is because it makes it possible to discover useful patterns and item relationships. One major step in Association Rule Mining is finding a frequent item set using the threshold support value and forming the association rules by using the specified confidence and the frequent itemset. The first step is Pre-processing in which missing values are dealt with. Then, binning is used to divide the data into several bins based on medical expert recommendations.

2. LITERATURE SURVEY

A Modified Apriori Algorithm for Fast and Accurate Generation of Frequent Item Sets by K.A.Baffour, C. Osei-Bonsu, A.F. Adekoya

This paper focuses on one of the two steps of the Apriori algorithm, i.e., generation of candidate item sets. The existing Apriori Algorithm has several shortcomings. Some of them are- the generation of a plethora of item sets, the need to perform many DB scans, along with the production of several combinations that never occur in the DB. A novel and modified version of the Apriori Algorithm is proposed which significantly reduces the number of DB passes using a row-wise combination generation technique.

Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation by Sheila A. Abaya

The focus of this paper lies in improving the efficacy of the existing Apriori algorithm, mainly by optimizing the database access. This was made possible by introducing minor modifications in the code, different set sizes and set frequencies. This resulted in a faster generation of possible frequent item sets. This was made possible by reducing the number of database passes needed, which was down by a significant amount as compared to the traditional Apriori algorithm.

A Modified Apriori Algorithm for mining Frequent Pattern and Deriving Association Rules using Greedy and Vectorization Method by Arpita Lodha, Vishal Shrivastava

In this paper, a new approach is introduced for finding frequent item sets by using a greedy and vectorization technique which reduces the time consumed by 79%. Further, the number of rules generated are also limited, thus removing the redundant ones.

Efficient Implementations of Apriori and Eclat by Borgelt, C

This paper discusses the need to reduce the humungous amount of item sets, which render naïve approaches inviable because of their unacceptable execution time. It also elucidates the similarities and differences between two algorithms: Apriori and Eclat. Also, depending upon the minimum support value, when to use either of the two algorithms has been mentioned to obtain the maximum efficiency.

3. APRIORI ALGORITHM

This algorithm has three steps:

1. For item I from 1 to n do
2. For each set Jn such that for each h (h belongs to Jn) that occurs in at least k baskets do
3. Examine the data to find whether the set Jn occurs in at least k baskets

In case of this algorithm, plenty of time is spent in accessing the database for matches, hence, its efficiency can be subjected to further improvement.

4. MODIFIED APRIORI ALGORITHM

The Classical Apriori Algorithm (CAA) has been changed to predict heart diseases using medical data mining. This algorithm is needed to find the frequent item sets using which the association rules are generated. Frequent item sets are the item sets that have the minimum specified support in the given dataset.

```
Algorithm: Modified Apriori
Inputs     : Medical dataset (Ck) and support threshold (min_support)
Output     : Locally frequent diseases that satisfy support
Preprocess: Associate numeric values with discrete values (Pk)
Lk: frequent itemset of size k
k=1
while Lk != emptyset
  Pk+1 = Medical data generated from Lk
  For each transaction t in T
    Increment the count of medical data in Pk+1 contained in t
    Lk+1 = medical data in Pk+1 with min_support
  end
end
return UkLk
```

Figure 1: Modified Apriori Algorithm (MAA)

This algorithm makes use of an existing dataset (taken from hospitals and clinics) and minimum support as inputs. Before the algorithm is used, the data is pre-processed to convert it into an easier format for processing (numeral to discrete values).

5. PROPOSED SYSTEM

The data needed for this study was a sample subset of about 1000 entries collected from 25 medical establishments (hospitals and clinics) in India, under the supervision of the National Health Ministry. 8 attributes are used, 7 of them being considered as inputs predicting the future state of "Diagnosis".

Attributes	Descriptions	Encoding\Values	Feature
Age	Age in years	28-66	Numeric
Chest Pain Type	It signals heart attack and has four different conditions: Asymptotic, Atypical Angina, Typical Angina, and without Angina.	Asymptotic = 1 Atypical Angina = 2 Typical Angina = 3 Non-Angina = 4	Nominal
Rest Blood Pressure	Patient's resting blood pressure in mm Hg at the time of admission to the hospital	94-200	Numeric
Blood Sugar	Below 120 mm Hg- Normal Above 120 mm Hg- High	High = 1 Normal = 0	Nominal Binary
Rest Electrocardiographic	Normal, Left Ventricular Hypertrophy (LVH) ST_T wave abnormality	Normal=1 Left Vent Hyper = 2 ST_T wave abnormality = 3	Nominal
Maximum Heart Rate	maximum heart rate attained in sport test	82-188	Numeric
Exercise Angina	It includes two conditions of positive and negative	Positive = 1 Negative = 0	Nominal Binary
Diagnosis	It includes two conditions of positive and negative	Positive = 1 Negative = 0	Nominal Binary

Figure 2: Attributes in dataset

After the dataset had been prepared, data pre-processing was done in order to transform the raw data into an understandable format. All the databases were stored on a server using MySQL Client software. A Minimum Support Threshold (MST) is used in discovering frequent item sets. The MST is generally taken as user input. But in this study, measures of central tendency were used to calculate the MST.

MST = (max + min)/2

For example,

ITEM	OCCURRENCE
I1	7
I2	7
I3	6
I4	2
I5	4
I6	4

From this table, max=7 and min=2. Therefore, MST=(7+2)/2 =4.5 ~ 5

Figure 2 depicts the architecture of the proposed MAA.

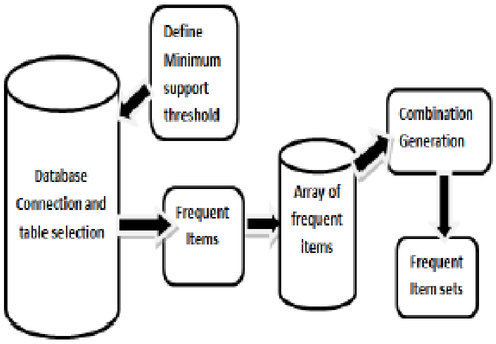

Figure 3: Architecture of proposed MAA

Now, in addition to introducing a modified Apriori algorithm, two additional steps have been added to further improve the efficiency, namely – Orthogonal Matching Pursuit (OMP) algorithm and Vectorization.

The OMP is an iterative greedy algorithm that constructs an approximation through an iterative procedure. At each step (iteration), a locally optimum solution is chosen as is done in case of any greedy approach. During each iteration, a column vector in A is found which resembles a residual vector r the most. OMP relies on the hope that all the locally optimum solutions would result into a globally optimal solution.

Vectorization is nothing but a linear transformation tool which converts a matrix into a column vector.

This proposed framework, as shown in figure 3, uses a greedy data transformation approach to reduce the size of the transaction and on top of that, applies vectorization to speed up the algorithm. After this is done, the proposed MAA is used to generate the association rules.

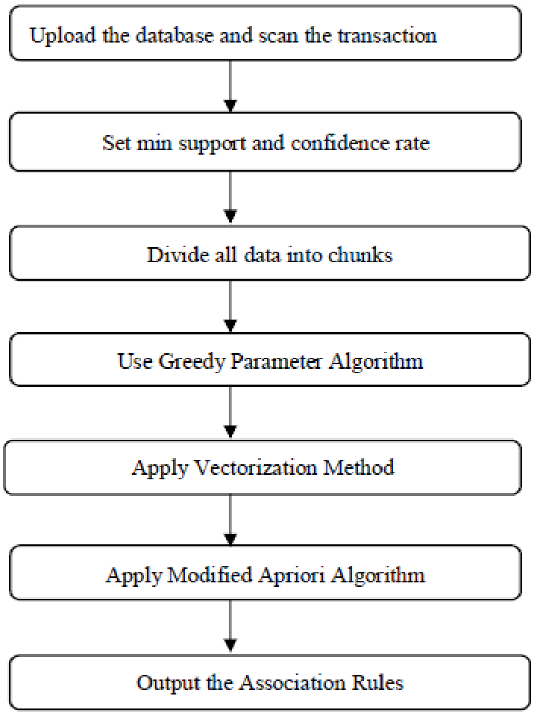

Figure 4: Proposed framework

6. DESIGN SPECIFICATIONS

A. Hardware Components
   A PC with:
   i) 4 GB RAM
   ii) Core 2 dual processor
   iii) Running windows 7 OS
B. Software Components
   i) Net Beans IDE
   ii) MySQL Database Server

7. RESULTS

The results of implementing the stipulated framework has been juxtaposed with the Classical Apriori Algorithm in this section.

Table 1 compares the execution time of CAA and MAA, and also shows the percentage improvement with respect to the number of transactions.

Table 1: Execution Time

Sr. No.	No of Transactions	Execution Time in Apriori (in seconds)	Execution Time in Modified Apriori (in seconds)	Percentage Improvement
1	100	1.067334	0.2152	80%
2	200	2.006108	0.4234	79%
3	500	5.172584	1.1544	78%
4	1000	10.157086	2.1014	79%

In figure 5, a graph depicting the number of transactions with respect to the execution time for both the algorithms are shown.

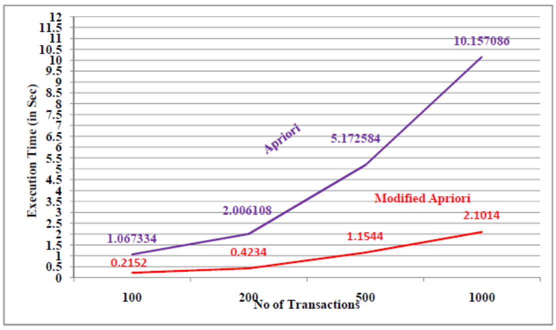

Figure 5: No of transactions v/s Execution Time

Figure 6 lists all the association rules given as output by the MAA. Although the result seen would be identical in the case of CAA, the only difference is a much lower execution time (which makes a significant difference in case of a large number of transactions).

8. CONCLUSION

A novel and modified Apriori is introduced in this paper which reduces the number of database passes, and thus the execution time. Apart from changing the Apriori algorithm, two steps: OMP and Vectorization were added in order to further optimize the whole process. The results were evident in the graph shown depicting the execution time and how it fares with the number of transactions. Now, using this new proposed algorithm, association rules were generated to predict heart diseases using the dataset.

9. FUTURE WORKS

In this paper, the proposed MAA has only been used on a limited data for heart disease prediction. But observing its efficiency even when a large number of transactions are involved, it can easily be used on a much large dataset. Also, the domain it is applicable to shouldn't be restricted to heart diseases and it should be used in other domains too.

Figure 6: Association Rules

Rules
Healthy rules:
If{Sex=female∩exercise_induced_angina=fal∩number_of_vessels_colored=0∩ thal = nom} => class healthy (conf., 0.98).
If {Sex = female ∩ fasting_blood_sugar = fal ∩ exercise_induced_angina = fal ∩ number_of_vessels_colored = 0} => class healthy (conf., 0.98)
If{Sex=female∩exercise_induced_angina=fal∩number_of_vessels_ colored = 0} => class healthy (conf., 0.98).
If {Sex = female ∩ fasting_blood_sugar = fal ∩ exercise_induced_angina = fal ∩ thal = norm} => class healthy (conf., 0.95).
If {Resting_blood_pres less or = '(115.2, 136.4]'∩exercise_induced_angina = fal∩ number_of_vessels_colored =0 ∩thal = norm}=>class healthy(conf., 0.94)
. **Sick Rules:**
If {Chest_pain_type = asympt ∩ slope = flat ∩ thal = rev} => class sick (conf., 0.96).
If {Chest_pain_type=asympt ∩ exercise_induced_angina=TRUE ∩ thal=rev} => class sick (conf., 0.94).

10. REFERENCES

[1] K.A.Baffour, C. Osei-Bonsu, A.F. Adekoya. A Modified Apriori Algorithm for Fast and Accurate Generation of Frequent Item Sets. International journal of scientific & technology research volume 6, issue 08, august 2017

[2] Arpita Lodha, Vishal Shrivastava. A Modified Apriori Algorithm for Mining Frequent Pattern and Deriving Association Rules using Greedy and Vectorization Method. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 4, Issue 6, June 2016.

[3] Sheila A. Abaya. Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation. International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012.

[4] D. Kerana Hanirex, and M.A. Dorai Rangaswamy. 2011. Effi-cient Algorithm for Mining Frequent Itemsets using Clustering Techniques. International Journal on Computer Science and Engi-neering (IJCSE) Vol. 3 No. 3 March 2011.

[5] P.Purdon, D. Gucht and D. Groth,‖ Average Case Performance of the Apriori Algorithm,‖ Society for Industrial and Applied Mathematics (2004) Vol. 33 No.5 pp. 1223-1260

[6] Jiawei Han, MichelineKamber, “Data Mining, Concepts and Techniques”, ISBN 978-81-312-0535-8, Elsevier India Private Limited, 2006.

[7] T. Junfang, “An Improved Algorithm of Apriori Based on Transaction Compression,” vol. 00, pp. 356–358, 2011.

[8] Goswami D.N. et. al. “An Algorithm for Frequent Pattern Mining Based On Apriori” (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 942-947.

[9] Ibrahim Umar Said, Jamila M. Muhammad, Manoj Kumar Gupta. Intelligent Heart Disease Prediction System by Applying Apriori Algorithm. International Journal of Advanced Research in Computer Science and Software Engineering.

[10] Mohammed Abdul Khaleel, Sateesh Kumar Pradhan, G.N.Dash. Finding Locally Frequent Diseases Using Modified Apriori Algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013.

[11] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules”, Seminar Presentation at University of Tokyo, 2004

[12] E. Barati et al., “A Survey on Utilization of Data Mining Approaches for Dermatological (Skin) Diseases Prediction”, Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI): March Edition, 2011

[13] AbdelghaniBellaachia and Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques”

[14] G.Subbalakshmi et al., "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering (IJCSE)

[15] N.DEEPIKA et al., "Association rule for classification of Heart-attack patients", International Journal of Advanced Engineering Sciences and Technologies, Vol No. 11, Issue No. 2, 253 – 257