

Prediction of Heart Diseases Risk through Frequent Itemsets in Data Mining

Ramya Sri M M (14CSR151), Renuka P (14CSR158) and Rajeshkumar R (14CSL275)
and Kumaravel T, Assistant Professor
Department of Computer Science and Engineering
Kongu Engineering College(Autonomous)
Perundurai, Erode-638 060.

Abstract-According to recent survey by WHO organisation 17.5 million people dead each year. It will increase up to 75 million in the year 2030. Data mining is process which is used to analyses number of data sets and then extracts the meaning of data. Huge amount of data is generated for the prediction of heart diseases. Data mining provides methods and techniques for transformation of the data into useful information for decision making. These techniques can make process fast and take less time to predict the heart disease with more accuracy. Based on the chosen symptoms and minimum support value frequent itemsets are generated. The extracted frequent itemsets help the medical practitioner to make diagnostic decisions and determine the risk level of patients at an early stage.

Keywords: Frequent Itemsets; Heart Disease Prediction; Data Mining; Binary Dataset; Medical Data Mining.

1. INTRODUCTION

Data mining has already been established as a novel field for exploring hidden patterns in the huge dataset [1]. Day by day, large numbers of patients are visiting the hospitals for the purpose of various treatments. Number of patients' records are increasing in every department in the hospital. In medical field, data mining algorithms are used to mine the hidden knowledge in the dataset of the medical domain [2]. To find out the frequent itemsets, minimum support value has been used. Support value of the itemset greater than or equal to minimum support value is called frequent itemset. If an itemset is frequent, then all of its subsets also must be frequent [3].

A few kinds of heart disease are cardiovascular diseases, Bradycardia, Irregular or extra heartbeats, heart attack, coronary heart disease and Stroke. Stroke is a type of heart disease; it is caused by narrowing, blocking, or hardening of the blood vessels that go to the brain or by high blood pressure [4]. An obstruction is a heart defect that partly or completely blocks the flow of blood.

In this paper, a new mining method is proposed to predict the risk level of heart disease based on chosen symptoms by analyzing the heart disease dataset. The predictions of this method will help the medical practitioners in making diagnostic decisions to save lives of patients at risk.

2. LITERATURE REVIEW

Usha Rani et al. has introduced pincer search algorithm to discover the maximum frequent itemset [5]. It also reduces number of times the database is scanned. Frequent itemset mining without the generation of conditional frequent pattern trees was expressed by Meera Narvekar et al. [6]. The desired association rules are also discovered from the frequent itemset. Alagugowri et al. developed a predicting system to predict the heart disease [7]. KMeans clustering technique is used to distinguish the risky and non-risky factors to categorize. Tzung-Pei Hong et al. developed MFFP-Tree Fuzzy Mining Algorithm to find out the linguistic frequent Itemsets [8]. Marghny et al. has developed a new method to mine frequent itemset by avoiding the costly candidate generation-and-test processing. It also compresses essential information about all itemset, minimal and maximal length of frequent itemsets and database scans repeatedly [9]. Jahangir Kabir et al. proposed a novel method to determine maximal frequent itemsets with genetic algorithm [10]. The weighted support measure is introduced by Subrata Bose et al. that adopted a balanced approach to mine frequent patterns [11]. To mine frequent closed sequential pattern in temporal transaction data, Antonio Gomariz et al. proposed a ClaSP algorithm [12]. To mine frequent itemset based on nodesets, an efficient FIN algorithm was developed by Zhi-Hong Deng et al. [13]. Hai Duong et al. developed a new algorithm with double constraints to find out all frequent itemsets [14]. Mengchi Liu et al. proposed a HUIMiner (High Utility Itemset Miner) algorithm to mine high utility itemset [15]. Umair Shafique et al. implemented three

various algorithms (Neural Network, Decision Tree and Naïve Bayes) to discover interesting patterns from heart patients’ data. The results reveal that the Naïve Bayes algorithm has the highest accuracy among them [16]. Darshan M. Tank has proposed an algorithm to reduce pruning operations. Deepa S. Deshpande proposed a novel method for mining association rule using patter generation. To find out frequent feature set, the Boolean operations for pattern generation is adopted [17]. Zhou Zhiping et al. introduced matrixbased sorting index association rules algorithm to find the frequency k-itemsets directly. It discovers k-itemsets directly when frequent item sets are higher [18]. Amr Jadi et al. proposed an algorithm to predict and mitigate the risks by using runtime monitoring with neural networks [19]. Sallam Osman Fageeri et al. introduced a binary-based technique to find out the frequent itemsets that outperforms classic Apriori algorithm in terms of running time [20]. To estimate the size of candidate itemsets in Apriori based algorithms, linear algebra method was used by Savo Tomoviü et al. [21]. Sen Su et al. has designed differentially private FIM algorithm to offer high time efficiency rather than achieve high data utility and degree of privacy [22]. This survey indicates that many algorithms were developed by researchers to generate frequent itemsets. A new method proposed in this paper generates frequent itemsets efficiently based on chosen symptoms and support value.

3.DATA SOURCE

For this research work we are going to consider dataset with 1000 patients record. This dataset contains 19 symptoms as shown in Table 1.

They are the symptoms of various heart diseases, namely Atherosclerotic Diseases(AD), Heart Arrhythmias (HA), Dilated Cardiomyopathy (DC), Valvular Hear Diseases (VHD) and Heart Infection(HD)[24].

Table 1. Symptoms of Various Heart Diseases

Symptoms ID	Symptoms (Attribute)Name
1.	Chest pain (angina)
2.	Shortness of breath
3.	Pain, numbness, weakness or coldness in your legs or arms if the blood vessels in those parts of your body are narrowed
4.	Pain in the neck, jaw, throat, upper abdomen or back

5.	Fluttering in your chest
6.	Racing heartbeat (tachycardia)
7.	Slow heartbeat (bradycardia)
8.	Lightheadedness
9.	Fainting (syncope) or near fainting
10.	Breathlessness with exertion or at rest
11.	Swelling of the ankles and feet
12.	Swelling in your legs
13.	Fatigue
14.	Irregular heartbeats that feel rapid, pounding or fluttering
15.	Fever
16.	Swelling in your abdomen
17.	Changes in your heart rhythm
18.	Dry or persistent cough
19.	Skin rashes or unusual spots

4. PROPOSED METHODOLOGY

Heart disease is one of the leading dangerous diseases. Every day, the count of patients affected by heart diseases is increasing. In most hospitals, the medical records of patients with various diseases are maintained in electronic medium. It is very difficult to extract the useful information from the vast volume of records manually. Nowadays, several data mining algorithms are developed to extract the useful knowledge from massive data. In this paper, a method to predict the patients under risk based on the chosen symptoms by analyzing the heart disease dataset. Also, it discovers the risk level of those patients. The proposed algorithm avoids the generation of unnecessary itemsets. It removes the factors (itemsets) that do not satisfy the support value. Rows (record) having zero value in the entire column (itemset) is removed from further analysis. It simplifies the collection of frequent itemsets generation and improves the efficiency of itemsets generation. It saves execution time by avoiding unnecessary comparisons. Figure 1 shows the flow of the proposed method.

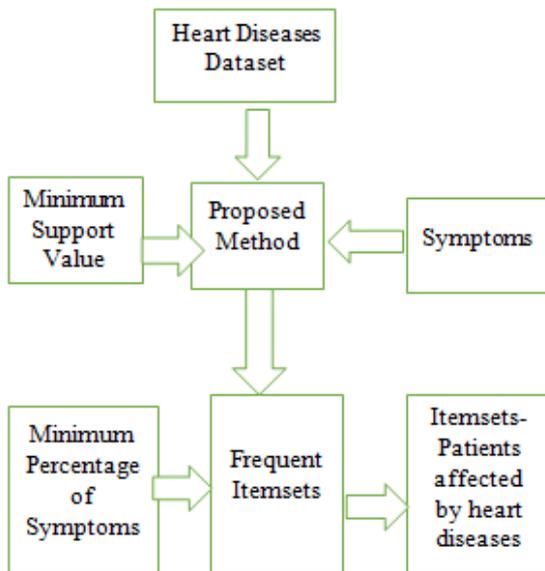


Fig. 1. Block Diagram of Proposed Method

4.1 Algorithm

Input D - Heart Disease Dataset

ms - Minimum support threshold value (0.1 - 1)

mps - Minimum percentage of symptoms (0.1 - 1)

s – Symptoms

Output

Fk - Frequent Itemsets Method

Step 1:

ts – Total number of symptoms (s)

k = ts

Combine all the chosen symptoms ‘s’ using priority and logical AND operation. Then finds a zero value in the combined columns and deletes that particular row. Find the sum value of this column. Calculate the support value for column using the formula:

$$\text{Support}(S) = \frac{\text{Sum Value of Column}}{\text{Total Number of Records}}$$

if $S < ms$ then terminate the process

else

Combined column consider as frequent itemset Fk. Then perform logical AND with Fk and all other columns.

Repeat steps 2 to 5 until dataset is null.

Step 2:

Find the sum values in each column of the table.

Step 3:

Calculate the support value for each column using the formula:

$$\text{Support}(S) = \frac{\text{Sum Value of Column}}{\text{Total Number of Records}}$$

Step 4: If $S < ms$ then delete the column from the table.

$k = k+1$

Add the column in table to Frequent Itemset Fk

Step 5:

Combines the column Max(Fk) with all other attributes without repetitive using logical AND operation. Delete the row having zero value in the entire column.

Step 6:

Calculate the percentage of symptoms for each frequent itemsets using the formula:

$$\text{Percentage of Symptoms} = \frac{\text{Number of Symptoms appear in Fk}}{\text{Total Number of Number Symptoms in Diseases}}$$

(i1...m, k1...n)

If $PS \geq mps$ then all these frequent itemsets are indicate risk factors of patients having heart disease.

Convert the dataset into binary format denoting the presence or absence of symptom that causes heart disease as 1 or 0 respectively. Dataset with minimum support value, minimum percentage of symptoms and the symptoms are given as input to the proposed method. In the first step, combine all the chosen symptoms (columns) using priority and logical AND operation. Then find a zero value in the combined columns and delete that particular row. Find the sum value of the combined column and calculate the support value of this column using the formula mentioned in the algorithm. The process has been terminated if the support value has been less than minimum support value. Otherwise the combined column is considered as frequent itemset Fk. Then perform the logical AND operation with frequent itemset Fk with all other columns. In the next two steps, find the sum values in each column of the table and calculate the support value for each column using the formula mentioned above. In step 4, check the support value of the each column with user given minimum support value. Delete the column which has the support value less than the minimum support value. All the column are considered as frequent itemset Fk+1 after deleting the unsatisfying column. In the step 5, find the Column which

has the maximum sum value, combines the column with all other columns using logical AND operation. Delete the row having zero value in the entire column. Perform this process from steps 2 to 5 repeatedly until dataset becomes null. Finally, it generates the maximum possible length of the itemsets F_k . In step 6, calculate the percentage of symptoms for each frequent itemsets using the formula mentioned in the step 6 of the algorithm. Finally, it extracts all the itemsets which has percentage of symptoms greater than or equal to user given minimum percentage of symptoms. The extracted itemsets are used to predict the patients who will be affected by the heart disease with risk level.

5.RESULT AND DISCUSSION

In this paper, a method is developed to generate the frequent itemsets based on the symptoms given by the user. It helps to identify the patients who is at risk of heart diseases from the extracted itemsets. The developed method is successfully implemented with Java programming language. The training dataset includes data of 1000 patients affected by heart related diseases with 19 clinical attributes. List of symptoms (attributes) of the dataset are shown in Table 1.

Performance of the proposed method is compared with existing methods to establish the efficiency of the proposed method. Table 2 shows the comparison with Apriori Algorithm [23] and Association Rule Mining Algorithm Based on Pattern Generation [17]. Existing methods generate all the possible itemsets in each iteration, thus increasing the computation time in comparisons and memory requirement. The proposed method does not generate the unnecessary itemsets in each frequent itemsets. Numbers of itemsets are very much reduced in the proposed method. This is the major research contribution in the proposed method. The limitation of the existing techniques is overcome by the developed method.

Plot in Fig. 2 is drawn to have statistical information on number of patients affected by 19 different symptoms. Majority of the patients are affected by Breathlessness with exertion or at rest and the lowest number of patients are affected with Swelling in your legs symptoms for the chosen dataset.

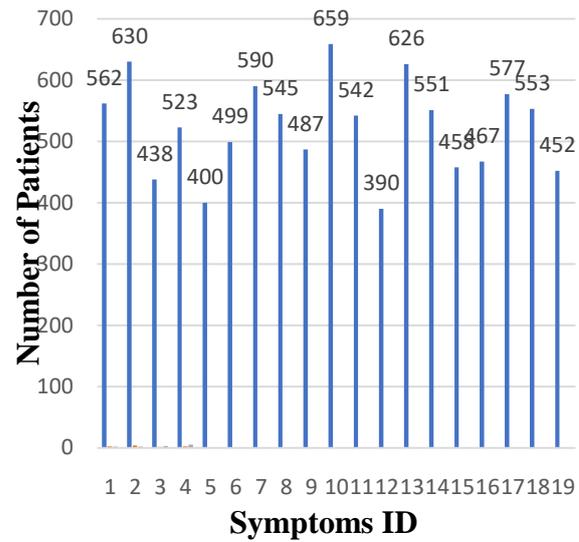


Fig. 2. Number of patients affected by different symptoms causing heart disease

Plot in Fig. 3 shows the results for minimum support value=0.1. It predicts the number of patients affected by a combination of chosen symptoms and also affected by atleast 60% of specific symptoms of a disease. Any symptom in the chosen dataset may be included in the combination. The zero value indicates the number of patients who are affected by chosen symptoms but below 60% of symptoms of a disease. From the plot, it is evident that more number of patients are affected by atleast 60% of the symptoms of Valvular Heart Disease (VHD) and combination of symptoms like chest pain, shortness of breath, fatigue or fainting.

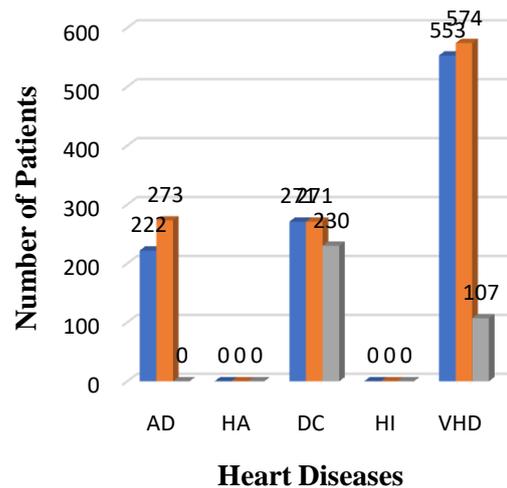


Fig. 3. Analysis of Heart Disease Dataset against combinations of symptoms with $ms=0.1$ and $mps=0.6$

■ Based on the Symptoms: Chest Pain, Shortness of Breath, Fatigue

■ Based on the Symptoms: Chest Pain, Shortness of Breath, Fainting

■ Based on the Symptoms: Chest Pain, Lightheadedness, Fatigue

Table 2 shows the result with the given support value 0.1. Existing methods generate all possible itemsets in each frequent itemset. The drawback of these methods is two-fold. Firstly, it generates very large itemsets in each frequent itemset which requires more storage space to retain all the sets. Secondly, it includes symptoms which are not relevant to particular combination of symptoms used for analysis. Number of itemsets increases inordinately when the disease bears more number of symptoms. This is the major drawback of the existing methods, which is overcome in the proposed method. In this analysis, a particular combination contains three symptoms which is assumed as frequent itemset-3. Therefore, the first two itemsets are not generated in the proposed method. The proposed method ensures that the symptoms of a chosen combination are included in all itemset of the each frequent itemset. Other itemsets which do not include them are ignored. It also ensures that the symptoms included in previous frequent itemset are included in the next frequent itemset provided those symptoms affect majority of the patients. In this way, the proposed method finds frequent itemset quickly compared to existing methods.

Table 2. Comparison of Proposed Method with Existing Methods.

Frequent Itemsets	Existing Methods			
	Apriori Algorithm [23]	IMSIA Algorithm [15]	Semi-Apriori Technique [18]	Association Rule Mining Algorithm Based on Pattern Generation [14]
Itemset -1	19	19	19	19
Itemset -2	171	171	171	171
Itemset -3	741	741	741	741
Itemset -4	1218	1218	1218	1218
Itemset -5	922	922	922	922
Itemset -6	386	386	386	386

Itemset -7	77	77	77	77
Itemset -8	6	6	6	6

Proposed Method

Frequent Itemsets	Based on the symptoms: Chest pain, Shortness of Breath, Fatigue	Based on the symptoms: Chest pain, Shortness of Breath, Fainting	Based on the symptoms: Chest pain, Lightheadedness, Fatigue
Itemsets-1	-	-	-
Itemsets-2	-1	-1	-1
Itemsets-3	16	15	14
Itemsets-4	12	12	12
Itemsets-5	7	7	9
Itemsets-6	3	3	2
Itemsets-7	-	-	1
Itemsets-8			

6.CONCLUSION

Medical data mining plays a vital role in the diagnosis of diseases and in life saving decisions. It is essential to find frequent itemset from patient data to predict the symptoms causing dangerous diseases. In this paper, an efficient method is provided to find frequent itemsets and risk level to predict the patients who will be affected by heart disease. The developed method analyses and predicts the number of patients at risk level. It provides a rapid aid to the medical practitioner in making emergency decisions to save the lives of patients at risk level. In the proposed method, symptoms representing columns and patient records representing rows are removed from further analysis, if they do not satisfy the chosen rules. The proposed method is applied over a heart disease dataset of 1000 records of patients suffering from various heart related diseases. The prediction results are encouraging and the efficiency of the method in frequent itemset generation is better than existing methods.

7. REFERENCES

- [1] William Carroll; G. Edward Miller, Disease among Elderly Americans: Estimates for the US civilian non institutionalized population, 2010,” Med. Expend. Panel Surv., no. June, pp. 1–8, 2013
- [2] Saurav Mallik, Anirban Mukhopadhyay, Ujjwal Maulik. RANWAR: Rank-Based Weighted Association Rule Mining from Gene Expression and Methylation Data. IEEE Transactions on NanoBioscience 2013; 14:59-66.
- [3] Darshan M. Tank. Improved Apriori Algorithm for Mining Association Rules. International Journal of Information Technology and Computer Science (IJITCS) 2014; 6:15-23.
- [4] Tom Dent, “Predicting the risk of coronary heart disease”, PHG foundation publisher, 2010.
- [5] Usha Rani G, Vijaya Prakash R, Govardhan A. Mining Multilevel Association Rule Using Pincer Search Algorithm. International Journal of Scientific Research 2013;
- [6] Meera Narvekar, Shafaque Fatma Syed. An Optimized Algorithm for Association Rule Mining using FP Tree. International Conference on Advanced Computing Technologies and Applications 2015; 45:101-110.
- [7] Alagugowri S, Christopher T. Enhanced Heart Disease Analysis and Prediction System [EHDAPS] Using Data Mining. International Journal of Emerging Trends in Science and Technology 2014; 1:1555-1560.
- [8] Tzung-Pei Hong, Chun-Wei Lin, Tsung-Ching Lin. The MFFP-Tree Fuzzy Mining Algorithm to Discover Complete Linguistic Frequent Itemsets. International Journal of Computational Intelligence 2014; 30:145–166.
- [9] Marghny H, Mohamed, Mohammed M, Darwieesh. Efficient Mining Frequent Itemsets Algorithms. International Journal of Machine Learning and Cybernetics 2013; 5:823-833.
- [10] Mir Md. Jahangir Kabir, Shuxiang Xu, Byeong Ho Kang, Zongyuan Zhao. A Novel Approach to Mining Maximal Frequent Itemsets Based on Genetic Algorithm. 9th International Conference on Information Technology and Applications (ICITA), at Sydney, Australia, 2014.
- [11] Subrata Bose and Subrata Datta. Frequent Pattern Generation in Association Rule Mining using Weighted Support. Third International Conference on Computer, Communication, Control and Information Technology (C3IT) 2015; 1-5.
- [12] Antonio Gomariz, Manuel Campos, Roque Marin, Bart Goethals. ClaSP: An Efficient Algorithm for Mining Frequent Closed Sequences. Advances in Knowledge Discovery and Data Mining 2013; 7818:50-61.
- [13] Zhi-Hong Deng, Sheng-Long Lv. Fast Mining Frequent Itemsets using Nodesets. Expert Systems with Applications 2014; 41:4505-4512.
- [14] Hai Duong, Tin Truong, Bay Vo. An Efficient Method for Mining Frequent Itemsets with Double Constraints. Engineering Applications of Artificial Intelligence 2014; 27:148-154.
- [15] Mengchi Liu, Junfeng Qu. Mining High Utility Itemsets without Candidate Generation. Proceedings of the 21st ACM International Conference on Information and Knowledge Management 2012; 55-64
- [16] Umair Shafique, Fiaz Majeed, Haseeb Qaiser, Irfan Ul Mustafa. Data Mining in Healthcare for Heart Diseases. International Journal of Innovation and Applied Studies 2015; 10:1312-1322.
- [17] Deepa S. Deshpande. A Novel Approach for Association Rule Mining using Pattern Generation. International Journal of Information Technology and Computer Science(IJITCS) 2014; 6:59-65.
- [18] Zhou Zhiping, Wang Jiefeng. An Improved Matrix Sorting Index Association Rule Data Mining Algorithm. 33rd Chinese Control Conference Proceedings, China 2014; 28-30.
- [19] Amr Jadi, Hussein Zedan, Turki Alghamdi. Risk Management based Early Warning System for Healthcare Industry. International Conference on Computer Medical Applications (ICCA) 2013; 1-6.
- [20] Sallam Osman Fageeri, Rohiza Ahmad, Baharum B. Baharudin. A Semi-Apriori Algorithm for Discovering the Frequent Itemsets. International Conference on Computer and Information Sciences (ICCOINS) 2014; 1-5.
- [21] Savo Tomoviü, Predrag Stanišiu. Upper Bounds on the Number of Candidate Itemsets in Apriori Like Algorithms. 3rd Mediterranean Conference on Embedded Computing (MECO) 2014; 260-263.
- [22] Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, Fangchun Yang. Differentially Private Frequent Itemset Mining via Transaction Splitting. IEEE Transactions on Knowledge and Data Engineering 2015; 27:1875-1891.
- [23] Ilayaraja M, Meyyappan T. Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm. Proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering 2013; 194–199.
- [24] <http://www.mayoclinic.org/diseases-conditions/heart-disease/basics/symptoms/con-20034056>.