Mining of Medical Data to Identify Risk Factors of Heart Disease Using Frequent Itemset

H K Shifali, Dr. B. Srinivasu, Rajashekar Shastry, B N Ranga Swamy

H K Shifali M.Tech
Dept. of Computer Science and Engineering, Stanley College of Engineering and Technology for Women
Telangana - Hyderabad, India.
Dr. B. Srinivasu Associate Professor, Rajashekar Shastry Assistant Professor
Dept. of Computer Science and Engineering, Stanley College of Engineering and Technology for Women
Telangana - Hyderabad, India.
B N Ranga Swamy Scientist 'F'
Defence Electronics Research Laboratory (D.L.R.L)
Telangana - Hyderabad, India.

---***---

Abstract – - *Data mining techniques are used in the field of medicine for various purposes. The function of data mining in biomedical information is remarkable one. Mining association rule is one of the interesting topics. It was first proposed for market basket analysis. In this project a method has been implemented to identify risk factors having heart disease through frequent Itemset. Frequent itemsets are generated based on the attributes and minimum support value. The extracted frequent itemsets help the medical practitioner to make diagnostic decisions and determine the risk level of patients at an early stage. The proposed method can be applied to any medical dataset to predict the risk factors with risk level of the patients based on chosen factors.*

Key Words*: *Data Mining, Heart Disease, Frequent Itemset, prediction, missing values,

1. INTRODUCTION

Heart disease or cardiovascular disease is the class of diseases that involve the heart or blood vessels (arteries and veins). Heart disease was considered to be a man's problem, but now it is recognized as number one killer of women, just as it is of men.

As described by National Heart Lung and Blood Institute echo can detect possible blood clots inside the heart, fluid buildup in the pericardium (the sac around the heart), and problems with the aorta But, the interpretation of echo recordings remains a challenge as it is time consuming. In order to solve this and many other problems in the health sector related to disease diagnosis. Data mining can be a solution by generating rules from those enormous datasets which can be used in echo readings. mining include diverse techniques to extract the information from a large database [1].

1.2 Problem Description

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. Data mining have shown a promising result in prediction of heart disease. It is widely applied for prediction or classification of different types of heart disease. The key objective of the work is,

- To identify key patterns or features from the dataset.
- To Identify and select attributes that are more relevant in relation to heart disease diagnosis.
- To analyze the results of the selected model with the help of domain expert.

1.3 Basics and Background Knowledge

Data Mining is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction [2].

Healthcare data analysis: healthcare data is diversified due to the high dimensionality of the data comprising of medical history records of patients having different symptoms. **Textual data analysis:** Textual data is commonly available in electronic documents or on social networks.

1.3.1 Association Rule Discovery

The association task for data mining is the job of finding which attributes "go together." Association rules are of the form "If antecedent, then consequent," together with a measure of the support and confidence associated with the rule.

$$\text{Support}(X \rightarrow Y) = P(XUY) \tag{Eq 1.1}$$

$$\text{Confidence}(X \rightarrow Y) = P(X/Y) \tag{Eq 1.2}$$

Association differs from classification in two ways: they can "predict" any attribute, not just the class, and they can predict more than one attribute's value at a time [3].

1.3.2 Frequent Itemset

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. Frequent pattern mining was first proposed by Agrawal et al.[4] for market basket analysis in the form of association rule mining.

1.3.3 Heart Disease

Heart disease or cardiovascular disease is the greatest scourge afflicting the world. As with previous scourge- bubonic plague, yellow fever, and smallpox; cardiovascular disease not only strikes down a significant fraction of the population without warning but causes prolonged suffering and disability in an even larger number. Simultaneously, remarkable progress has been made in preventing and treating cardiovascular diseases by medical and surgical means Braunwald et al., [5].

1.3.4 Major Types of Heart Disease

There are many types of heart disease, but in this study the researcher chooses to discuss five types that are common to happen. As described by Cindy [6], five common types of heart disease are discussed below.

Congenital heart disease-heart doesn't pump adequate blood to the other organs in the body. **Congestive heart failure -**heart disease which is passed down through the family.**Coronary heart disease (ischemic heart disease)- -**damage to the heart that happens because its blood supply is decreased.**Pulmonary heart disease-**disease that comes from lung, pulmonary, disorder. **Rheumatic heart disease** derives from strep throat infections.

2. LITERATURE SURVEY

Numerous works related to heart disease diagnosis using data mining techniques have motivated this study.

K S Kavitha et al [7] performed a work on "Modeling and design of evolutionary neural network for heart disease detection" an alternative solution for complex medical diagnosis in detection of heart disease where human knowledge is apprehended in a general fashion.

AH Chen et al [8] developed a new system to predict heart disease that can assist medical professionals in predicting heart disease status based on the clinical data of patients. An artificial neural network algorithm was developed for classifying heart disease based on 13 important clinical features.

Resul Das et al [9] introduced a methodology using SAS Base Software9.1.3 for diagnosing of the valvular heart disease. A neural networks ensemble method creates new models by combining the posterior probabilities or the predicted values from multiple predecessor models.

Carlos Ordonez [10] introduced an algorithm to searches for association rules on a training set reduces the number of rules using search constraints, and finally validates them an independent test set. The discovered rules are evaluated with support value, confidence value and lift. **Dubey et al.** [11] performed a work on suggest the early detection may help in providing better treatment and can cure. They also suggest that the framework based on data mining and optimization may predict diseases properly.

U.Chandrasekhar et al. [12] surveys various latest frequent pattern mining algorithms on data streams to understand various advantages and disadvantages, so they provides a way of using new insights in the direction of frequent pattern. The combination of optimization and data mining is also suggested has gotten much consideration and has been joined in numerous enhancement issues, to be specific the system directing, voyaging salesperson, quadratic task, and asset assignment issues.

3. ARCHITECTURE

This section deals with the proposed method, system architecture and short description of design process.

3.1 Proposed Method

We are proposing a method to predict the risk of heart disease based on patience history. In this approach we do not require training module, directly we can extract the frequent patterns by which we can generate the attributes that cause heart disease and rise early warnings to patience. In this method attributes which are in columns and patient records which are in rows are eliminated if they do not satisfy support value and which leads to dimension reduction from further analysis, if they do not satisfy the chosen rules.

3.2 System Architecture

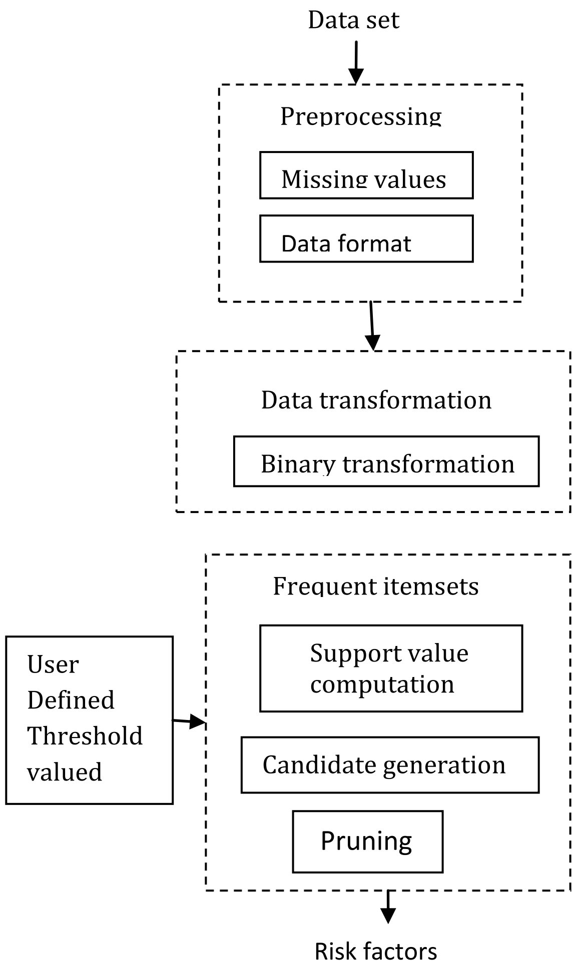


Figure 3.1 System Architecture

Data is preprocessed that involves transforming raw data into an understandable format, where missing value and data format are performed. This format is further converted to the binary values based on the conditions applied to the attribute. Then pruning is used to reduce the size of dataset by removing infrequent itemsets which further helps in generating the next itemset.

3.3 SHORT DESCRIPTION OF DESIGN PROCESS

Data Set

A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. A data set is organized into some type of data structure. In a database, for example, a data set might contain a collection of business data (names, salaries, contact information, sales figures, and so forth). The database itself can be considered a data set, as can bodies of data within it related to a particular type of information, such as sales data for a particular corporate department.

Attributes

An attribute is a property or characteristic of an object. Attribute values are numbers or symbols assigned to an attribute. Attributes may be classified into two main types

- **Numeric Attributes**

A numeric attribute is one that has a real-valued or integer-valued domain. Numeric attributes that take on a finite or countable infinite set of values are called discrete, whereas those that can take on any real value are called continuous. As a special case of discrete, if an attribute has as its domain the set {0;1}, it is called a binary attribute.

- **Categorical Attribute**

A categorical attribute is one that has a set-valued domain composed of a set of symbols.

Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Missing Value

An instance where in no data is present for the variable in question. Although this is common, it creates problems in using the remaining data to predict with any certainty the futures affiliated with the data also called missing data.

Data Format

Converting the dataset file and values into the format for destination data system

Data Transformation

Data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system. Where the data is converted into binary values based on the conditions applied to the attribute to reduce the dimensionality of the dataset encoding mechanism

Pruning

Pruning is used to reduce the size dataset by removing infrequent itemsets which further helps in generating the next itemset.

Dimension Reduction

Dimension Reduction refers to the process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely. These techniques are typically used while solving machine learning problems to obtain better features for a classification or regression task. It helps in data compressing and reducing the storage space required, It fastens the timeless dimensions leads to less computing.

Support

It is the probability of item or item sets in the given transactional data base: Support(X) = n(X) / n Where n is the total number of transactions in the database and n(X) is the number of transactions that contains the item set X. Therefore, Support (X=>Y) = Support (XUY).

Risk Factors

A condition or habit that makes a person more likely to develop a disease can also increase the chances that an existing disease will get worse.

4. IMPLEMENTATION

Based on the proposed system architecture in figure 3.1, the methodology is constructed into four modules (1) data set collection (2) preprocessing (3) data transformation (4) frequent Item set generation.

4.1 Dataset Description

The dataset used in this project is obtained from UCI Repository of Cleveland database [36] containing 303 patience's record. Below are the list of attributes and there values.

Table 4.1 Attributes of Heart Disease Dataset

S.no	Attribute name	Type	Values
1	Age	Continuous	In years
2	Sex	Discrete	1=male 2=female

3	C_P	Discrete	Value1=typical angina
			Value2=atypical angina
			Value3=non-anginal pain
			Value4=asymptomatic

4	Trestbps	Continuous	In mm Hg on Admission to the Hospital
5	Chol	Continous	In mg/dl

6	Fbs	Discrete	Fasting blood sugar>
			120mg/dl
			True=1
			False=0

7	RestECG	Discrete	Value0=normal
			Value1=having ST-T
			Wave abnormality
			Value2=showing probable

8	Thalach	Continous	Maximum heart rate achieved

9	Exang	Discrete	Value1=yes
			Value0=no

10	Old peak	Continuous	In number values

11	Slope	Discrete	Value1=up sloping
			Value2=flat
			Value3=down sloping

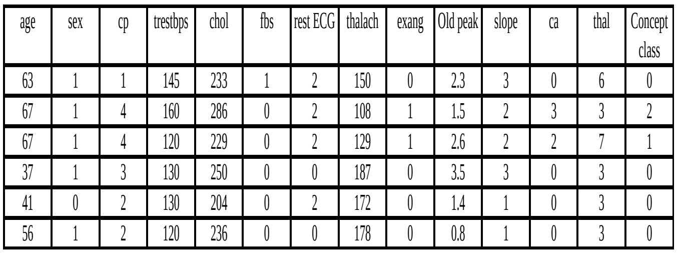
12	CA	Discrete	0-3 value

13	Thal	Discrete	Value3=normal
			Value6=fixed
			Value7=reversible
			Defect

14	Concept Class	Discrete	5 types of values 0-4

The dataset is used by authors for different purpose; they have used it to find the prediction of heart disease with different algorithms and technique. Our aim is to find the risk factor that causes heart disease. The Cleveland database contains 75 attributes in which 14 attributes are considered.

Table 4.2 Sample Dataset Structure



age	sex	cp	trestbps	chol	fbs	rest ECG	thalach	exang	Old peak	slope	ca	thal	Concept class
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0

4.2 Preprocessing

Analyzing data that has not been carefully screened can produce misleading results. If there is much irrelevant and redundant information or noisy and unreliable data reduces the performance. Removing such noisy data is preprocessing. This module consists of two sub modules as mentioned below.

1. Missing Value
2. Data Format

4.2.1 Missing value

Missing values are a common problem in real datasets. There are many possible explanations for why a data value may be unavailable the measurements were simply not made, human or machine error in processing a sample, and error in transmitting or storing data values into their respective records and thus different methods for handling this problem have been developed. The dataset file containing 14 attributes and 303 patient's records are saved in data file format. We have identified missing values with '?' and have ignored them by replacing it with '0'.

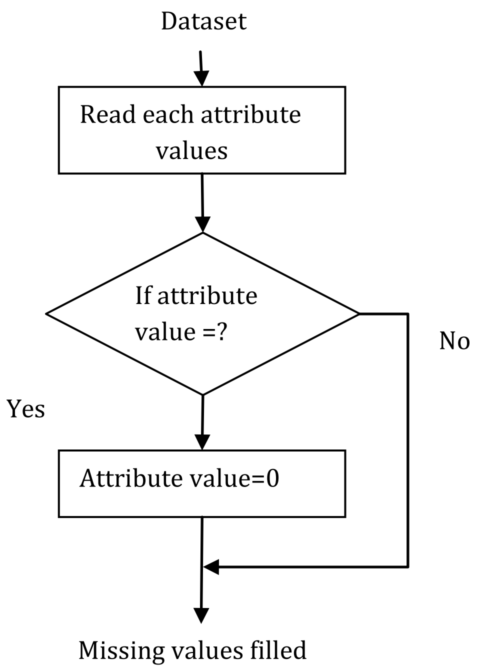


Figure 4.1 Work Flow of Missing Values

4.2.2 Data Format

The dataset which is collected from UCI repository are raw data. The data values are in string format it has to be converted into float format and further converted into integer this helps in implementation of the further modules.

4.3 Binary Transformation

The pre-processed data is taken as input and the data is transformed into suitable format so that mining algorithms can be directly applied. In this process, the data are transformed into 1 and 0 based on the conditions mentioned below which represents presence and absence of attributes that cause heart disease. Following are the possible conditions for Heart Disease.

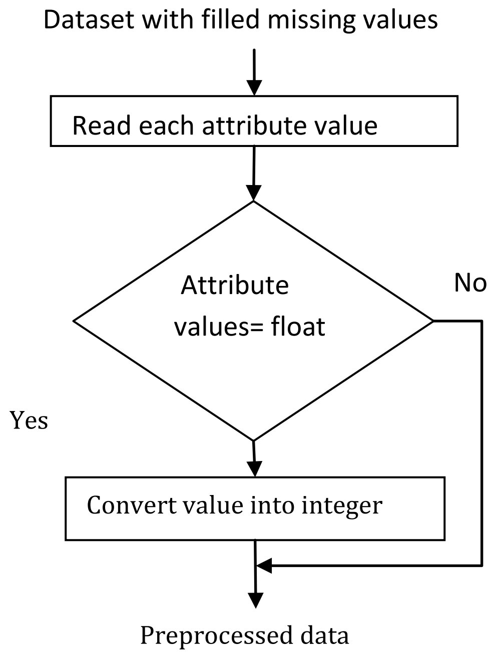

Dataset with filled missing values
Read each attribute value
Attribute values= float
No
Yes
Convert value into integer
Preprocessed data

Figure 4.2 Work Flow of Data Format

4.4 Frequent Itemset

In this module, proposed method is implemented to binary transformed file to generate frequent itemsets (1, 2, 3 and 4) where the 14 attributes and key attribute patients id are considered and values are calculated by summation the attribute values and patience data is deleted if entire row or column has zero values. Logic AND operation is applied after generation of itemset 1 to get itemsets 2, 3 and 4.

Procedure

1. Open the binary transformed file.
2. Read each attribute value including patient id.
3. Add each attribute value.
4. Find the sum of all the attribute values.
5. Compute support value using

$$Support = \frac{Sum\ value\ of\ attributes}{Total\ number\ of\ records} \qquad \text{(Eq 4.1)}$$

6. Eliminate the attributes which do not satisfy support value against threshold.
7. Find the maximum of sum values of all the attributes.

8. Apply logic AND operation between maximum sum value of attribute with all the other attributes.
9. Delete the row containing zero i.e. by deleting the rows, patients who are not affected by the attributes are eliminated for further analysis.
10. Repeat step 4 to 9.
11. Finally, frequent itemsets (1, 2, 3, and 4) are generated.
12. All four frequent itemset show the pairs of risk factors causing heart disease.

5. EXPERIMENT RESULTS

5.1 Result Analysis

The dataset used in this project is Cleveland Clinic Foundation. Heart disease dataset is available at UCI repository. The data set has 76 raw attributes. However, all of the published experiments only refer to 13 of them and one more attribute which specify the concept class. The data set contains 303 patients' records.

During implementation we found 6 missing values which represents as '?' in attributes slope and ca. In slope and ca the missing values are 4 and 2 as mentioned in the below table 5.1.

Table 5.1 Missing Values

Attributes name	Missing values
Slope	4
CA	2
	Total=6

The overall performance of the proposed approach for the generation of frequent itemsets based on the support calculation against support threshold.

We set the threshold to see the variations of frequent itemsets from the above table we estimate that as the threshold increases frequent itemsets decreases. The table 5.2 it depicts for 40% support threshold frequent itemset-1 generates 11 risk factors whereas, for 50% support threshold frequent itemset-1 generates 7 risk factors we obtained four frequent itemsets. Similarly in frequent itemsets – 2, 3 and 4 for threshold 40% 10, 13 and 6 respectively and for threshold 50% 7, 6, 3 respectively.

Table 5.2: Generation of Frequent Itemsets

Support threshold	Iteration Number	Selected Attribute
40%	1	[1,2,3,4,5,7,8,10,11,13]
	2	[1,8,2,8,3,8,4,8,5,8,7,8,10,8,11,8, 13,8]
	3	[1,2,8,1,3,8,1,4,8,1,5,8,1,7,8,1,11 8,2,4,8,2,13,8]
	4	[1,2,4,8]
50%	1	[1,2,4,5,8,11]
	2	[1,8,2,8,4,8,5,8,11,8]
	3	[1,2,8,1,4,8]
	4	[1,2,4,8]

Table 5.3 Number of Lexical chains generated.

Iteration numbers	No of frequent itemsets at threshold 40%	No of frequent itemsets at threshold 50%
1	11	7
2	10	6
3	13	3
4	6	0

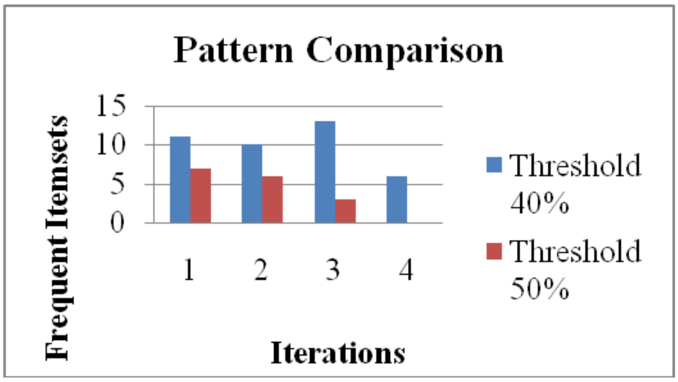


Figure 5.1 Pattern Comparison

6. CONCLUSIONS

A new method has been developed to generate the frequent itemsets based on the heart disease dataset which is collected from UCI repository Cleveland Clinical

Foundation. It helps to identify the risk factors from the extracted itemsets that cause heart disease. This helps medical practitioners to alert the patients with the risk factors that are obtained. This proposed method is implemented with Java programming language. We have generated four frequent itemsets. In first frequent itemset attribute 8 i.e. Thalach was found common in maximum patients, hence it was important attribute and was compared with all the other attributes using logic AND operation to generate second frequent itemsets. In second frequent itemset attribute (1, 8) i.e. Age and Thalach found common in maximum number of patients. Third frequent itemsets includes the attribute which occurs among maximum number of patients in itemset-2.This is repeated until dataset is exhausted. Finally frequent itemset will contain the attributes which causes heart attack.

As a future work, the researcher can plan to perform additional experiments with more dataset or real time dataset using different data mining techniques; algorithms to build a model that can specific heart disease types and a predictive model.

ACKNOWLEDGEMENT

To Mom
Thanks for the support and strength
Without you I am nothing.
To Dad
Thanks for the encouragement and moral support.
I thank my back bone Support my friends.

REFERENCES

1. M. Ilayaraja Medical Data Mining Method to Predict Risk Factors of Heart Attack and Raise Early Warning to Patients, 2015.
2. Oded, M. and Lior, R., "Data Mining and Knowledge Discovery Handbook", Springer Science and Business Media Inc., New York, 2005.
3. Witten H. Ian and Frank, E., "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition. Morgan Kaufmann Publishers, San Francisco, 2005.
4. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM-SIGMOD international conference on management of data (SIGMOD'93), Washington, DC, pp 207–216
5. Braunwald, E., Douglas, P. Zipes, Peter, L., Robert, B., "Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine", Third Edition, Harcourt Brace Jovanovich Inc., New York, 1988.
6. Cindy H., "5 Common Types of Heart Disease", EzineArticles.com. Available at http://ezinearticles.com/?5--Common--Types--of--Heart--Disease& id= 1073496, 2008.
7. K.S. Kavitha, K.V. Ramakrishnan and Manoj Kumar Singh. "Modelling and design of evolutionary neural network for heart disease detection" International Journal of Computer Science Issues, ISSN (Online): 1694-0814,Vol. 7, Issue.5, 2010.
8. AH Chen, SY Huang, PS Hong, CH Cheng and EJ Lin. "HDPS: Heart Disease Prediction System" Computing in Cardiology, Vol.38, pp.557-560, 2011.
9. Resul Dasa, Ibrahim Turkoglub and Abdulkadir Sengur. Diagnosis of valvular heart disease through neural networks ensembles" Computer Methods and Programs in Biomedicine, pp.185–191, 2009.
10. Carlos Ordonez."Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction" Transactions on Information Technology in Biomedicine, 2006.
11. Dubey A, Patel R, Choure K, "An efficient data mining and ant colony optimization technique (DMACO) for heart disease prediction", International Journal of Advanced Technology and Engineering Exploration.; 1(1):1-6, 2014.
12. U.Chandrasekhar, Sandeep Kumar. K, Yakkala Uma Mahesh," A Survey of latest Algorithms for Frequent Item set Mining in Data Stream", International Journal of Advanced Computer Research (IJACR),Volume-3 Number-1 Issue-9 March-2013.

BIOGRAPHIES

H K Shifali Received Her M.Tech Degree in Software Engineering from Osmania University. Her Research Interest is in Data Mining.