Multi-Class Based Frequent and Infrequent Pattern Mining Model for Medical Data Sets

[1]Sujatha Kamepalli
Associate Professor, IT Department,
VFSTR,
Vadlamudi, Guntur district, Andhra Pradesh,
India- 522213.

Abstract: Data is growing exponentially in all the disciplines including in the medical domain. Daily patients are coming and medical industry stores the data corresponding to those patient's and uses the data to make clinical decision, prevent medication error, and improve patients' outcomes. The medical environment is information rich but weaker in terms of knowledge. In order to extract the hidden relationships and trends, need powerful tools. This paper proposes three models in order to find frequent and infrequent patterns from complex medical databases. In each model new measures are developed and implemented on Medical Pulmonary Embolism Database and Medical Rheumatoid Arthritis Database. The proposed models have less computational time compared to the traditional pattern mining models on medical database.

IndexTerms- **Data Mining, Knowledge, Medical Domain, Pattern Mining, Medical Pulmonary Embolism Database, Medical Rheumatoid Arthritis Database**

I. INTRODUCTION

In all the fields data is growing tremendously, for that medical field is not exceptional. Health industry stores a large amount of patient's data and uses the data to make clinical decision, prevent medication error, and improve patients' outcomes [3]. The medical environment is information rich but weaker in terms of knowledge and powerful tools to identify hidden relationship & trends in medical data. There is an instant need of new generation of computational theories and tools to assists human in extracting useful knowledge from rapidly growing volumes of digital data. For this, implementation of data mining tools makes the extraction process efficient. Owing to the great advantages various organizations in various fields using data mining technology.

Data mining plays a vital role in the medical field by predicting various associative patterns. It is the process of selecting, exploring large amounts of data to find hidden, previously unknown patterns [1]. Data mining is used to analyze different types of data by using available data mining tools. There are different data mining techniques, such as classification, regression, clustering and association rules that are applied on data sets for predicting useful information from large amounts of data [2].

Data mining techniques have a wide scope of applicability in the field of disease diagnosis and prognosis and hidden biomedical and health care patterns. With the help of this further intelligent systems and decision support systems can build for better earlier diagnosis prediction of different diseases [4]. A number of benefits are provided by the data mining. Some of them are as follows: it plays a very important role in the detection of fraud and abuse, provides better medical treatments at reasonable price, detection of diseases at early stages, intelligent healthcare decision support systems etc. Some of the services provided by the data mining techniques in healthcare are: number of days of stay in a hospital, ranking of hospitals, better effective treatments, fraud insurance claims by patients as well as by providers, readmission of patients, identifies better treatments methods for a particular group of patients, construction of effective drug recommendation systems, etc [5]. Due to all these reasons researchers are greatly influenced by the capabilities of data mining. In the healthcare field researchers widely used the data mining techniques [6]. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system [11].

II. LITERATURE SURVEY

Parvez Ahmad, Saqib Qama, Syed Qasim Afser Rizvi, in Techniques of Data Mining In Healthcare: A Review" reviews various Data Mining techniques such as classification, clustering, association, regression in health domain. It also highlights applications, challenges and future work of Data Mining in healthcare [6].

Johannes K. Chiang Sheng-Yin Huang, in Multidimensional Data Mining for Healthcare Service Portfolio Management With regard to healthcare service management, this research aims at providing a novel data schema and an algorithm to solve the aforementioned problems. A forest of concept taxonomies is used as the data structure for representing healthcare associations patterns that consist of concepts picked up from various taxonomies. Then, the mining process is formulated as a combination of finding the large item sets, generating, updating and output the association patterns. Crucial mechanisms in each step will be clarified. At last, this paper presents experimental results regarding efficiency, scalability, information loss, etc. of the proposed approach to prove the advents of the approach [7].

Priyanka N, Dr.Pushpa RaviKumar in Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree intends to adopt Naïve Bayes & Decision tree – two data mining techniques for the effective prediction of Heart disease. It compares the efficiency & accuracy of the two techniques to decide among them the best [8].

K.M. Mehedi Hasan Sonet, Md. Mustafizur Rahman, Pritom Mazumder, Abid Reza, Rashedur M Rahman in Analyzing Patterns of Numerously Occurring Heart Diseases Using Association Rule Mining an association based rule mining technique has been used to identify such hidden patterns of the most commonly occurring heart diseases namely Unstable Angina (UA), Myocardial Infarction(MI), Coronary Heart Disease(CHD) etc. among Bangladeshi people and unraveling the hidden information by analyzing the results. Basically, other researchers in this field used the classification and clustering methods of data mining by which they could predict the chance of occurring heart diseases and clustered them to identify the dependency of one attribute to another. The trends or patterns for heart diseases may vary depending on sex, age, socioeconomic condition, demographic regions and so on. The objective of our work is to find out those hidden trends or patterns. Therefore, we have chosen association rule mining technique to find those patterns or trends among patients depending on their age, sex, regions and socioeconomic condition [9].

Stephen E. Brossette, Alan P. Sprague, J. Michael Hardin, Ken B. Waites, Md, Warren T. Jones, Stephen A. Moser,Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. The authors first illustrate the need for automated pattern discovery and data mining in hospital infection control and public health surveillance. Next, they define association rules, explain how those rules can be used in surveillance, and present a novel process and system— the Data Mining Surveillance System (DMSS)—that utilize association rules to identify new and interesting patterns in surveillance data. Experimental results were obtained using DMSS to analyze Pseudomonas aeruginosa infection control data collected over one year (1996) at University of Alabama at Birmingham Hospital. Experiments using one-, three-, and six-month time partitions yielded 34, 57, and 28 statistically significant events, respectively. Although not all statistically significant events are clinically significant, a subset of events generated in each analysis indicated potentially significant shifts in the occurrence of infection or antimicrobial resistance patterns of *P*. aeruginosa. The new process and system are efficient and effective in identifying new, unexpected, and interesting patterns in surveillance data. The clinical relevance and utility of this process await the results of prospective studies currently in progress[10].

Siri Krishan Wasan ,Vasudha Bhatnagar and Harleen Kaurthe Impact Of Data Mining Techniques On Medical Diagnostics We identify a few areas of healthcare where these techniques can be applied to healthcare databases for knowledge discovery. In this paper we briefly examine the impact of data mining techniques, including artificial neural networks, on medical diagnostics [11].

B.Venkatalakshmi, M.V Shivsankar Heart Disease Diagnosis Using Predictive Data mining this project intends to design and develop diagnosis and prediction system for heart diseases based on predictive mining. Number of experiments has been conducted to compare the performance of various predictive data mining techniques including Decision tree and Naïve Bayes algorithms. In this proposed work, a 13 attribute structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree and Naive Bayes have been applied and their performance on diagnosis has been compared. Naïve Bayes outperforms when compared to Decision tree [12].

Ionuț Țăranu Data mining in healthcare: decision making and precision The trend of application of data mining in healthcare today is increased because the health sector is rich with information and data mining has become a necessity. Healthcare organizations generate and collect large volumes of information to a daily basis. Use of information technology enables automation of data mining and knowledge that help bring some interesting patterns which means eliminating manual tasks and easy data extraction directly from electronic records, electronic transfer system that will secure medical records, save lives and reduce the cost of medical services as well as enabling early detection of infectious diseases on the basis of advanced data collection. Data mining can enable healthcare organizations to anticipate trends in the patient's medical condition and behavior proved by analysis of prospects different and by making connections between seemingly unrelated information. The raw data from healthcare organizations are voluminous and heterogeneous. It needs to be collected and stored in organized form and their integration allows the formation unite medical information system. Data mining in health offers unlimited possibilities for analyzing different data models less visible or hidden to common analysis techniques. These patterns can be used by healthcare practitioners to make forecasts, put diagnoses, and set treatments for patients in healthcare organization [13].

III. PROPOSED METHOD

In recent years, the widespread availability of complex data has led to an awareness of the importance of mining complex data. In order to provide new understanding of certain domains such as medical, e-commerce etc., association mining has become important in the research area of data mining. One important data mining task is to discover interesting but implicit hidden patterns that might exist in high dimensional data. Discovering relevant patterns from complex data is much more difficult than discovering patterns in traditional data.

3.1.Frequent and Infrequent Item Set Mining

Let I= {i1, i2, . . ., im} be a set of data items. A transactional database T= {t1,t2, . . ., tn} is a set of transactions, where each transaction tq ($q \in [1, n]$) is a set of items in I and is characterized by a transaction ID (tid). An item set I is a set of data items. More specifically, a k-item set is a set of k items in I. The support (or occurrence frequency) of an item set is the number of transactions containing I in T.

i. An item set I is frequent, if its support is greater than or equal to a predefined maximum support threshold ξ.

ii. An item set I is infrequent, if its support is less than or equal to a predefined maximum support threshold ξ.

3.2.Algorithm

Complex data has multiple attributes with different classes and values. Complex data is partitioned based on the instance properties. Proposed multi-class based infrequent mining algorithm is used to find interesting patterns and its relationship within instances.

The multi class based algorithm which finds infrequent high utility associations in complex data. The algorithm first finds the class based sub data sets from the given complex data set. Then for each class based sub data set it generates transaction based Boolean matrices and class based Boolean matrices. After that it generates the 1 to n-item sets candidate item sets. For each item set in the list it calculates the lift value. If the calculated lift value is greater than or equal to minimum threshold then it gets the associated items list from D and finds the correlations between the items and the associated Boolean matrices. If the correlation between item and the associated item satisfies min threshold then if the correlation between item and the associated item satisfies min confidence then adds the item and associated Boolean occurrences to PAR. Else if the correlation between item and the associated item satisfies negative items support and confidence then add to IAR and inserted into CP Tree. If the correlation between item and the associated item satisfies min threshold in case of negative correlation then add to IAR and inserted into CP Tree. Finally it gets patterns from the constructed CP Tree.

INPUT: Complex Data D, Minimum threshold ($\rho_{\min}$)
OUTPUT: Infrequent Patterns
METHOD:

1. Extract class based sub data sets CSD i.e. SD1, SD2,…………SDn.
2. For each dataset in CSD
3. Generate Transaction based Boolean Matrix TBM.
4. Generate Class based Boolean Matrix CBM1; CBM2…CBMn Here N is Number of classes.
5. Generate 1-item sets to n-item sets CS.
6. Scan the data base CS and extract 1-item sets (f1)
7. For each item in 1-item list, calculate lift and check the condition
8. if (lfv ≥ ρmin) then
9. Get the associated item list in D and find the correlation between the item and the associated Boolean matrices.
10. if (the correlation between item and the associated item satisfies min threshold) then
11. if(the correlation between item and the associated item satisfies min confidence) then
12. PAR ← PAR ∪{*item*,ϕ1….ϕk}
13. else if (the correlation between item and the associated item satisfies negative items support and confidence) then
14. IAR ← IAR ∪ { ¬ ϕ1…¬ ϕk}
15. Insert CPtree(IAR,ρmin)
16. if (the correlation between item and the associated item satisfies min threshold in case of negative correlation) then (σcorr≤ -ρmin)
17. if {conf(ϕ1, ¬ϕk) ≥ *conf min)* then
18. IAR ← IAR ∪ { ϕ1, ¬ϕk }
19. Insert CPtree(IAR, ρmin)
20. if (conf(¬ϕ1,ϕ*k*) ≥ *conf*min) then
21. IAR ← IAR ∪ {(¬ϕ1,ϕ*k* }
22. Insert CPtree(IAR, ρmin)
23. Infrequent Rules ← getPatterns(CPtree)

Lift calculates the ratio between the rules support and confidence of the item set in the rule consequent based on the each selected class. Lift= Probability of occurrence of an item in samples of ith class/ Probability of occurrence of an item in samples of ith class.
Prob(Ci/Di): Probability of occurrence of an item in samples of ith class .
Prob(Ci,D): Probability of occurrence of an item in a dataset of ith class.

Correlation finds the rank correlations among different item sets and their Boolean occurrences.

$$\sigma_{corr} = Correlation(\mathrm{i}, \varphi_1, \varphi_k) = |\mathrm{D}_i|\,\mathrm{lift}(\mathrm{i}, \phi_1) - |\mathrm{D}_i|\,\mathrm{lift}(\mathrm{i}, \phi_2) / |\mathrm{D}| \sqrt{\mathrm{lift}(\mathrm{i}, \phi_1)^2 - \mathrm{lift}(\mathrm{i}, \phi_2)^2}$$

3.3.Medical Database

In this model, the heart failure biomedical database is taken from https://bioportal.bioontology.org/ontologies/PEandhttps://bioportal.bioontology.org/ontologies/RAO. This database contains disease association treatments for association Mining. The major issue in this model is sparsity and null patterns as shown below.

Table. 3.1. Sample Medical Pulmonary Embolism Database (PE)

Class ID	Preferred	Synonyms	Definition	Obsolete	CUI	Semantic	Parents	http://bm	http://bm	http://data.bioontology
http://bm	Eliquis			FALSE			http://bmi.utah.edu		C3530466	j.0:C3530466
http://bm	heparin therapy			FALSE			http://bmi.utah.edu		C0522794	j.0:C0522794
http://bm	peripheral vascular alpha-recep			FALSE			http://bmi.utah.edu		T0000271	j.0:T0000271
http://bm	catheter-directed rotational er			FALSE			http://bmi.utah.edu		T0000277	j.0:T0000277
http://bm	tPA			FALSE			http://bm	alteplase |	C0032143	j.0:C0032143
http://bm	Xarelto			FALSE	Null Values		http://bmi.utah.edu		C3159309	j.0:C3159309
http://bm	continuous aspiration of thror			FALSE	or		http://bmi.utah.edu		T0000280	j.0:T0000280
http://bm	right heart catheterization-thro			FALSE	Sparse Values		http://bmi.utah.edu		T0000272	j.0:T0000272
http://bm	Intravenous fluids			FALSE			http://bm	intraveno	C1289919	j.0:C1289919
http://bm	tinzaparin			FALSE			http://bmi.utah.edu		C0216278	j.0:C0216278
http://bm	Embolectomy			FALSE			http://bm	rotational	C0162575	j.0:C0162575
http://bm	lanoteplase			FALSE			http://bmi.utah.edu		C0753753	j.0:C0753753
http://bm	History of inferior vena cava fil			FALSE			http://bm	vena cava	C1998116	j.0:C1998116
http://bm	Anti factor Xa measurement			FALSE			http://bm	monitorin	C0427612	j.0:C0427612
http://bm	VENA CAVA INFERIOR FILTER PI			FALSE			http://bm	inferior ve	C0750159	j.0:C0750159
http://bm	supplemental oxygen therapy			FALSE			http://bm	suppleme	C0184633	j.0:C0184633
http://bm	fluindione			FALSE			http://bm	flunidione	C0117899	j.0:C0117899
http://bm	interventional treatment			FALSE			http://bmi.utah.edu		T0000270	j.0:T0000270
http://bm	novel oral anticoagulant			FALSE			http://bm	Novel ora	T0000276	j.0:T0000276
http://bm	International Normalized Ratic			FALSE			http://bm	Internatio	C0525032	j.0:C0525032
http://bm	dabigatran			FALSE			http://bmi.utah.edu		C2348066	j.0:C2348066
http://bm	antiplatelet therapy			FALSE			http://bmi.utah.edu		C1096021	j.0:C1096021

Table. 3.2. Sample Medical Rheumatoid Arthritis Database (RA)

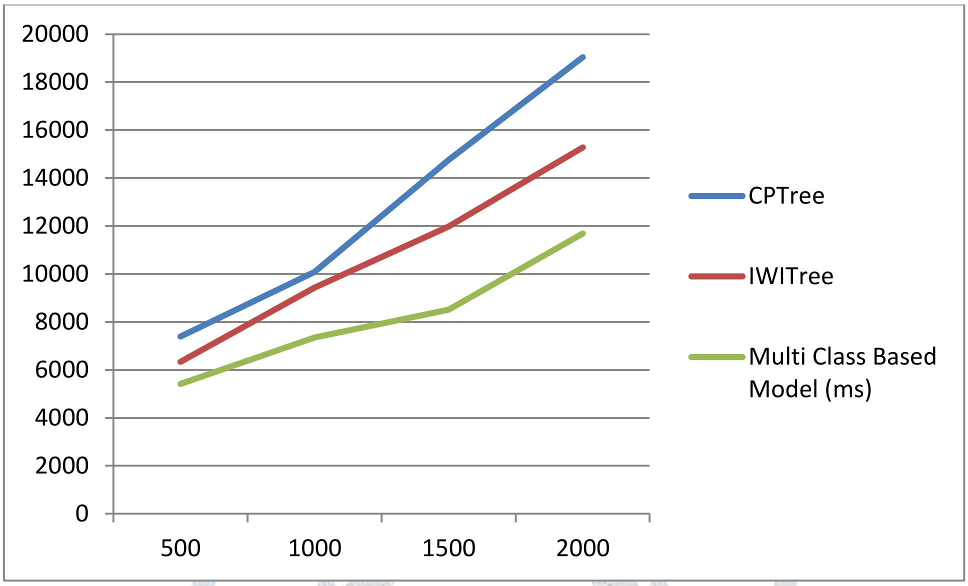
Class ID	Preferred	Synonyms	Definition	Obsolete	CUI	Semantic	Parents	http://bm	http://bm	http://data.bioontology.
http://bm	salsalate			FALSE			http://bmi.utah.edu		C0073983	j.0:C0073983
http://bm	ROSE HIPS			FALSE			http://bm	rose-hip	C0772037	j.0:C0772037
http://bm	Relaxation Therapy			FALSE			http://bm	mindfulne	C0282333	j.0:C0282333
http://bm	Coach in sports activity acciden			FALSE			http://bm	physical-a	C0277717	j.0:C0277717
http://bm	certolizumab pegol			FALSE	NULL VALUES		http://bm	certolizun	C1872109	j.0:C1872109
http://bm	Weight-Bearing Strengthening			FALSE	OR		http://bm	high-inter	C2350268	j.0:C2350268
http://bm	morphine			FALSE	SPARSITY VALUES		http://bmi.utah.edu		C0026549	j.0:C0026549
http://bm	LEF 537			FALSE			http://bm	LEF	C0672527	j.0:C0672527
http://bm	yttrium-90			FALSE			http://bmi.utah.edu		C0303596	j.0:C0303596
http://bm	Exercise			FALSE			http://bm	high-inter	C0015259	j.0:C0015259
http://bm	tilidine			FALSE			http://bmi.utah.edu		C0040219	j.0:C0040219
http://bm	preventive therapy			FALSE			http://bm	preventiv	C0033107	j.0:C0033107
http://bm	cardiovascular risk reduction			FALSE			http://bmi.utah.edu		T0000286	j.0:T0000286
http://bm	tofacitinib			FALSE			http://bmi.utah.edu		C2930696	j.0:C2930696
http://bm	Arthroscopic synovectomy kne			FALSE			http://bm	arthrosco	C0408192	j.0:C0408192
http://bm	flurbiprofen			FALSE			http://bmi.utah.edu		C0016377	j.0:C0016377
http://bm	stress management			FALSE			http://bmi.utah.edu		C0150788	j.0:C0150788
http://bm	methylprednisolone			FALSE			http://bmi.utah.edu		C0025815	j.0:C0025815
http://bm	emotion regulation therapy			FALSE			http://bmi.utah.edu		T0000295	j.0:T0000295
http://bm	Cyclophosphamide			FALSE			http://bmi.utah.edu		C0010583	j.0:C0010583
http://bm	metacarpophalangeal joint			FALSE			http://bm	metacarp	C0025525	j.0:C0025525
http://bm	hydrocodone			FALSE			http://bmi.utah.edu		C0020264	j.0:C0020264

Table. 3.3. Medical Database Performance Evaluation

Patterns	CPTree	IWITree	Multi Class Based Model (ms)
500	7385.59	6335.11	5409.103
1000	10086.3	9433.54	7345.744
1500	14757.4	11977.7	8508.678

2000	19040.3	15275.6	11692.71

Table 3.3. Illustrates the performance analysis of existing models with the proposed model in terms of candidate sets generation time and pattern mining time. From the table, it is observed that the proposed model has less computational time compared to the traditional pattern mining models on medical database.



IV. CONCLUSIONS

Discovering relevant patterns from complex data is much more difficult than discovering patterns in traditional data. Complex data has multiple attributes with different classes and values. Complex data is partitioned based on the instance properties. Proposed multi-class based infrequent mining algorithm is used to find interesting patterns and its relationship within instances. Lift calculates the ratio between the rules support and confidence of the item set in the rule consequent based on the each selected class. Correlation finds the rank correlations among different item sets and their Boolean occurrences. In this model, the heart failure biomedical database is taken; it is observed that the proposed model has less computational time compared to the traditional pattern mining models on medical database.

REFERENCES

[1] Dave Smith, "Data Mining in the Clinical Research Environment", available at. http://www.sas.com/ .

[2]. Damtew A., "Designing a predictive model for heart disease detection using data mining Techniques", A Thesis Submitted to the School of Graduate Studies of Addis Ababa University, 2011.

[3]. Daniyal, Wei-Jen Wang, Mu-Chun Su , Si-Huei Lee , Ching-Sui Hung ,Chun-Chuan Chen, “A guideline to determine the training sample size when applying big data mining methods in clinical decision making”, Proceedings of IEEE International Conference on Applied System Innovation 2018, ISBN 978-1-5386-4342-6, pp:678-681.

[4]. Narander Kumar, Sabita Khatri, “Implementing WEKA for medical data classification and early disease prediction”, 3rd IEEE International Conference on "Computational Intelligence and Communication Technology" (IEEE-CICT 2017), 978-1-5090-6218-8, pp:1-6.

[5]. H. C. Koh and G. Tan, “Data Mining Application in Healthcare”, Journal of Healthcare Information Management, vol. 19, no. 2, (2005).

[6]. Parvez Ahmad, Saqib Qama, Syed Qasim Afser Rizvi, “Techniques of Data Mining In Healthcare: A Review”, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15, June 2015, pp: 38-50.

[7]. Sheng-Yin Huang, Multidimensional Data Mining for Healthcare Service Portfolio Management", 978-1-4673-5214-7.
[8]. Priyanka N, Dr.Pushpa RaviKumar, "Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree", International Conference on circuits Power and Computing Technologies 978-1- 5090-4967- 7/17/$31.00 © 2017 IEEE.
[9] K.M. Mehedi Hasan Sonet, Md. Mustafizur Rahman, Pritom Mazumder, Abid Reza, Rashedur M Rahman," Analyzing Patterns of Numerously Occurring Heart Diseases Using Association Rule Mining", The Twelfth International Conference on Digital Information Management (ICDIM 2017) September 12-14, 2017, Kyushu University, Fukuoka, Japan.
[10]. Stephen E. Brossette, Alan P. Sprague, J. Michael Hardin, Ken B. Waites, Md, Warren T. Jones, Stephen A. Moser, "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance", Journal of the American Medical Informatics Association Volume 5 Number 4 Jul / Aug 1998 pp: 373-381.
[11]. Siri Krishan Wasan,Vasudha Bhatnagar and Harleen Kaur," The Impact Of Data Mining Techniques On Medical Diagnostics Data", Science Journal, Volume 5, 19 October 2006 Pp: 119-126.
[12]. B.Venkatalakshmi, M.V Shivsankar," Heart Disease Diagnosis Using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3, March 2014ISSN (Online) : 2319 – 8753 ISSN (Print) : 2347 – 6710, pp:1873-1877
[13]. Ionuț ȚĂRANU," Data mining in healthcare: decision making and precision", Database Systems Journal vol. VI, no. 4/2015pp:33-40.

CORRESPONDING AUTHOR'S PROFILE:

Dr.K. Sujatha completed her Ph.D. in April 2018, Krishna University, Machilipatnam, A.P. Her interested areas of research are Data Mining, Cloud Computing, and Data Analytics. She is appreciated by adjudicators for her publications in research. She has 20 International Journal Publications with high Impact Factors and Indexing (Scopus). She has two national journal publications. She attended for number of AICTE sponsored workshops and attended for a number of FDPs. She is member in Indian Association of Engineers (IAE). She is Life Time Member in International Association of Engineering & Technology for Skill Development (IAETSD). She has a total of 13 years' experience in teaching. She is working as Associate Professor in Vignan's Foundation for Science, Technology and Research, Vadlamudi, Guntur District. A.P.

Site this article as:

Dr. Sujatha Kamepalli, "Multi-Class Based Frequent and Infrequent Pattern Mining Model for Medical Data Sets", Journal of Emerging Technologies and Innovative Research (JETIR),……………………