Implementation of Apriori Algorithm in Health Care Sector: A Survey

Divya Jain[1], Sumanlata Gautam[2]
[1, 2] Department of Computer Science, ITM University, Gurgaon, (Haryana), India
[1]divyajain1890@gmail.com, [2]sumanlataguatam@itmindia.edu

***ABSTRACT* - This research presents a review of the implementation of Apriori Algorithm on different healthcare datasets using machine learning tool Weka. Association rule mining greatly helps to identify trends and patterns from huge databases. This research paper analyses the various results generated by implementing the Association Apriori Algorithm in Weka and compares its performance with other algorithms. In this paper we have evaluated the results generated by this algorithm executed in Weka. With the results obtained, we find that although there are significant drawbacks associated with Apriori Algorithm, it is found to be the simplest algorithm for mining frequent itemsets and that it achieves better results than Predictive Apriori Algorithm and Tertius Algorithm. This paper reviewed the research papers which mainly concentrated on implementation of Apriori Algorithm on different healthcare datasets.**

***Keywords*- HealthCare, Association Rule Mining, AprioriAlgorithm, Weka.**

I. INTRODUCTION

A vast amount of information is present in the world. The databases are full of extensive data but everybody is interested in the information that can solve the objective of the problem. Data mining is a technique that mines data effectively to get useful information.Data mining plays a crucial role in mining of healthcare data. Healthcare data can be collected from various hospitals. The assimilated data can be used to analyze the patient reports which help in identifying the patterns present in the databases which further helps to get information about various diseases present, their symptoms, causes, remedies and precautions that can help to prevent the occurrence of various diseases. Medical researchers have been interpreting data from time to time and data gets modified almost every day. The main concern is to be updated about the changes in patterns of data. Data mining techniques help in achieving this objective. Apriori association technique has been proven to be effective in finding various trends in healthcare databases. This paper presents a review of implementation of Apriori Algorithm in Weka on datasets collected from different hospitals from specific places at different course of time.

II. DATA MINING IN HEALTH CARE SECTOR

Healthcare sector is a massive area which deals with data about hospitals, patients, doctors,medical devices and equipment's. The management of large health care data poses a great challenge to the researchers. The usage of data mining and machine learning techniques has revolutionized the healthcare organizations. The field of data mining helps to discover hidden patterns by bringing a set of machine learning tools and techniques. It is useful in evaluating the effectiveness of medical treatments [1]. Data mining techniques like classification, association, clustering are applied to healthcare datasets to analyse data to improve health policy-making, early detection of disease outbreaks and preventing the occurrence of various diseases. Data mining provides healthcare authorities an additional source of knowledge for effective decision-making [2]. The information provided by data mining methods can help healthcare insurers detect fraud and abuse and healthcare organizations can make better customer relationship management decisions. Further physicians can identify effective treatments with best practices and patients will receive better and more affordable healthcare services.

III. ASSOCIATION RULE MINING

Association rule mining is one of the most important data mining techniques to produce strong association rules in the database.There are many association rule algorithms like Apriori Algorithm, Eclat Algorithm, and FP-growth Algorithm [3]. An important characteristic of association rule mining is that it divides the problem of mining into sub-problems to do efficient computing. One problem finds frequent itemsets from database and the other problem generates association rules from the database [4]. There are two rule evaluation metrics for association algorithms - support and confidence. These parameters greatly affect the rules produced from the dataset. The frequent itemsets produced must satisfy the constraints of minimum support and minimum confidence. Support for an association rule is the measure of how frequently a particular item or set of items occurs in a dataset [5]. Confidence represents the strength of an association rule [6]. The objective of association algorithms is to generate different rules with minimum support and minimum confidence. The main problem with association rule mining is that it works well with categorical attributes but is not efficient when used with numeric attributes.

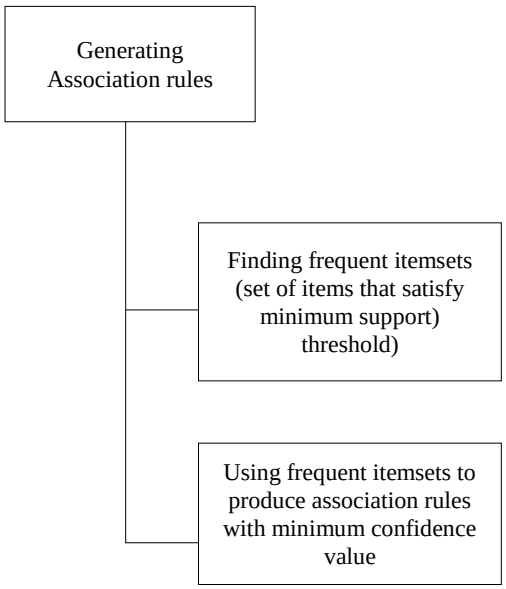

Figure 1: Association Rule Generation

IV. CONCEPT OF APRIORI ALGORITHM

It is the fundamental and most important algorithm for mining frequent itemsets. It was first given by Agrawaland Srikant in 1994 [7].It is a level wise algorithm which works in an iterative fashion to discover all frequent itemsets in a database. It uses prior knowledge of frequent itemsets properties [8]. Frequent itemsets are the sets of items that satisfy minimum support threshold.This algorithm takes only categorical input and associates attributes present in the dataset. There is a property associated with this algorithm called "Apriori Property" which states that any subset of frequent itemsets is also a frequent itemset. For example, if {x,y,z} is a frequent set then the sets { {x},{y},{z} }, { { x,y },{ x,z },{ y,z }} must also be frequent.The execution of this algorithm is organized in two phases. In the first stage, the candidates are generated and in the next phase frequent itemsets are generated [9]. The generated large itemsets are used to produce association rules from database.

V. LITERATURE SURVEY

PrasannaDesikan, et al. [2**],**gives an introduction to healthcare management and an overview of how data mining helps in management of healthcare data. Authors tell about current trends and prominent models for detection of various diseases. Various types of data used in hospitals like HL7, EHR, EMR, ENR etc. are discussed efficiently. They draw attention towards the new challenges faced by data mining to aid in healthcare management. Data mining helps in the detection of fraud and abuse, healthcare resource management and diagnosis and treatment of various diseases. It also helps healthcare organisations in making better customer relationship management decisions.Ms.Shweta and Dr. Kanwal Garg [10], emphasize mainly on using Apriori Algorithm for frequent item set mining. Authors have discussed the problem of frequent itemset mining and have addressed the mining using Apriori Algorithm with an example. Authors have implemented Apriori Algorithm on a bank dataset in Weka and then a comparison is made between the results of Apriori Algorithm, Predictive Apriori Algorithm and Tertius Algorithm. In future, these algorithms can also be used in other domains. These algorithms can be combined to get improved algorithm which can be used in any real- time application. In [11], association rule mining is used to generate strong association rules by executing Apriori Algorithm on real time datasets. In this paper, authors have exemplified Apriori Algorithm on a dataset by finding frequent itemsets and then generating association rules from frequent sets. In each iteration, frequent patterns are identified by using the candidate generation and pruning steps. At every step, it is taken care that "Apriori Property" of the algorithm is not violated. With the implementation of the algorithm they predict the occurrence of different diseases in a particular area and what type of people are affected by a particular disease is found out. Certain correlations between parameters - age, sex, environmental conditions and humidity have been realized in the dataset. In addition, an architecture of associative classifier is given in which different association rules produced by different categories are provided as input to classifier which predicts the particular class which serve as output.The outcome shows three different strong association rules with the result that poor hygiene and harmful environmental conditions in males between the age of 30 and 60 are likely to be ill with the contagious disease. Also it has been found that contagious disease has no relation with family history of a patient.In future, we can apply Apriori Algorithm without candidate generation on the dataset to get around the problem of examining database again and again which may give better results. We can also use frequent pattern tree structure to avoid costly scan of the dataset. P. Kasemthaweesab, et al. [12**],**proposed a research work of discovering association rules that relates diabetic mellitus with complications to obtain valuable information for the treatment of adult people. The authors have examined the patients undergoing from diabetes mellitus with multiple complications in Thailand. The complications studied are ophthalmic, renal, neurological and peripheral circulatory complications. The main goal is to analyse patients' profile such as gender, age, occupation through the implementation of Apriori Algorithm in Weka. Before applying association technique, the data of 65,535 patients is normalized to 29,823 records. The outcome of applying Apriori Algorithm in Weka give association rules based on gender, occupation complications and age of a person. The research generates association rules which prove helpful in diagnosing disease complications.
Following results are produced:

1. There is a likelihood of both male and female to be affected by Insulin-independent Type II diabetes mellitus with complications.
2. There is a chance to be diagnosed with type 2 diabetes mellitus with complications if the age of a person is between 50 and 79 or if the person does housework, trade and has no occupation.
3. If the individual is a female suffering from type 2 diabetes mellitus, the person would probably be having

neurological complications if she is 50-79 years old, ophthalmic complications if she is 60-79 years old, renal complications if she is 70-79 years old or if the person is female hireling.

4. There may be an incidence of ophthalmic, neurological and peripheral circulatory complications in type 2 diabetes mellitus patients if they have no occupation.

In future, we can study more complications associated with diabetes mellitus and can apply other data mining techniques to get more interesting relationships between diabetes and other diseases. This research has studied only four complicated states. We can work on more complications associated with diabetes mellitus. The results produced by the research can be used for the prevention of various diseases and for a reliable treatment for patients who are already suffering from diabetic complications.M. Ilayaraja, et al. research work [13] proposed a method using Apriori Algorithm to find how various diseases occur frequently in a particular geographical area during the year 2012. The training data consisting of 1216 records is collected from Hospital Management System (HIS) which includes patient's details – name, age, disease, district, location, time, date, etc. The implementation has been done with machine learning tool Weka on 29 different diseases. After examining results in Weka with minimum support value as 0.35 and minimum confidence value as 0.9, it is found that 4 different diseases, namely, Heart Disease, Pain, Smoking and Whooping Cough occur many times independently in the year 2012. Moreover, Heart Disease and Smoking largely hit the patients at the same time. The outcome also reveals that Heart disease, Liver Disease, Overweight and smoking occur simultaneously. Further, the results show the occurrence of a specific disease in a particular month and reveals that some patients are suffering from the same disease in a particular month. The experimented results can be used by doctors to arrive at good decisions concerning frequently occurring diseases. In future, we can combine other data mining techniques with the proposed method to get better results and to have better interpretation of generated results.The research proposed by B. M. Patil, R. C. Joshi and DurgaToshniwal[14] presents a method which makes use of application of association and classification techniques on numeric data to find whether a patient is likely to be affected by diabetes or not. The research work underlined the significance of discretization and pre-processing of data. Authors have presented a work on enhancing the discretising concept and modified equal width binning interval approach to deal with continuous variables. The Pima Indian diabetes dataset initially consisted of 768 female patients whose age is at least 21. With the help of pre-processing, instances are reduced to 625 and unwanted attributes are removed. Based on medical expert advice, data variables are binned into various categories using approximate equal interval binning. As Apriori Algorithm can only deal with categorical data, all 625 instances are converted to categorical data. Then Apriori Algorithm based on appropriate support and confidence values is implemented in WEKA to get associated rules from the database. The algorithm generates ten strongest rules which indicate the presence of diabetes. Similarly, other ten best rules are generated indicating absence of diabetes disease. This paper does not take attributes which influence diabetes. Future work needs to be done on finding factors which can give help in getting better results. S.M. Nuwangi, et al. [15**],** proposed a research work on diabetes disease with the help of association rule mining technique. In this paper, authors illustrate the global prevalence of diabetes. This research produces some new results that were earlier not given due weightage in medical field but are significantly important in medical field. The implementation has been done in Weka on a diabetes dataset consisting of 10,000 records taken from General Hospital of Sri Lanka with around 30 attributes considering both diabetic type 1 and type 2 patients. After performing data cleaning, pre-processing and filtering process, association rules are generated in Weka with proper discretization and error handling mechanisms. The objective is to find factors that are usually not known to doctors but which greatly affects diabetes. The research identifies unknown side effects of diabetes and factors which significantly affect diabetes.

The implementation in Weka gives the following findings:

1. The first association rule considers edema. According to the Rule 1 generated in Weka, it is found that female diabetic patients aging 57-75 years old having low cholesterol suffering from wheezes have higher chance to be diagnosed with edema. The effect would be stronger with diabetic type 1 patients than type 2 patients as the confidence of the rule is greater for diabetic type 1 patients than diabetic type 2 patients.Also, a direct relationship between diabetes and edema has been discovered. Likewise, the relationship is stronger with diabetic type 1 patients than type 2 patients.

2. The rule also depicts that there is a high probability that diabetic type 2 patients would be suffering from wheezes, which is not yet proven in medical field.

3. The first rule for diabetic type 1 patients also states that older people are more affected with type 1 diabetes than young people.

4. The second rule finds risk factors for wheezes. According to the literature, there is no link between wheezes and diabetes (both type 1 and 2 diabetic patients). But according to this research, authors have identified a strong relationship between the two diseases.According to the generated association rule, female diabetic patients who are housewives with low cholesterol and sodium level and diastolic blood pressure in the normal range suffering from edema have a higher probability to be diagnosed with wheezes.5. The third association rule considers risk factors for diabetes type 1 and type 2 patients. For diabetic type 2 patients, it states that females aging 39–56 years old with low cholesterol level and normal range of sodium level are likely to be ill with type 2 diabetes.In addition, this rule discovered a relationship between gender female and type 2 diabetes. Usually, doctors don`t consider this correlation but with the generated rule, authors have cited the presence of diabetes in females.For diabetic type 1 patients, the risk factors are same as for diabetic type 2 patients. One more risk factor is considered and that is potassium in between 2.5 and 3.6. The above risk factors have stronger relationship for diabetes type 1 than diabetes type 2 patients. The literature gives strong correlation between high potassium level, high sodium level and diabetes. But the rule has mentioned the relationship between low

ranges of potassium and sodium level and Diabetes type 1.The overall conclusion reveals a strong relationship between diabetes, edema and wheezes. Wheezes and edema are found to be unidentified side effects of diabetes. The study finds risk factors for diabetes and aftereffects of diabetes. Earlier studies have not taken gender as a strong factor that affects diabetes. But this research has identified gender female as a strong factor of diabetes disease. The research identifies a relationship between gender, age and diabetes. Till now, physicians are unaware of the link between diabetes and wheezes, gender and diabetes. This research brings a light on these important relationships.In future, we need to work on identifying more risk factors for type 1 diabetes. The study discovers more risk factors for diabetic type 2 patients than type 1 patients. We can expand the above study to get more interesting relationships between different diseases.In [16], authors have expanded the above mentioned research with the help of classification techniques to evaluate the results produced earlier from the association rule mining in Weka. The implementation is done on the same dataset of diabetic patients using decision trees to validate association rules produced earlier. Out of approximately 1000 rules produced during association rule generation, three best rules are chosen that strongly affects diabetic patients. To further discover patterns in the dataset, classification techniques are applied. Various decision trees were generated depicting strong relationship between wheezes, edema and diabetes. The results produced by decision tree analysis proved the results produced by association rule mining. The outcome shows a strong relationship between diabetes (both types), wheezes and edema. Also, it is found that the possibility of occurrence of edema for both diabetic type patients increases if they are suffering from wheezes. Besides that, wheezes is diagnosed more in cases when a patient is suffering from diabetes and edema than the people protected against edema. The gender female is found to be the major factor for the high FBS (Fasting Blood Sugar) level.In future, we can work on modifying Weka to get over the problem of visualizing complete decision tree. In [18], authors have implemented Apriori Algorithm, Predictive Apriori Algorithm and Tertius Algorithm on different healthcare datasets using Weka tool. The datasets studied are breast cancer, mushroom, larynx cancer, zoo, sunburn, imaginary disease, contact lenses, monk, soya bean and titanic datasets. The main goal is to compare the results of these algorithms and find out the parameters that affect the results and to study the problems associated with their implementation. The purpose is to compare the performance of the three association algorithms using Weka [17]. Furthermore appropriate conditions have been found to know when these algorithms give proper results. Besides, architecture for generating the association rules on datasets using Weka is also presented. The different case studies are executed in Weka to find the utility of association rule mining. The results of the case studies give following findings:

1. All three algorithms work well with nominal attributes but have problems with numeric attributes. So discretization needs to be done to convert numeric to nominal attributes.
2. Apriori Algorithm is based on rule parameters – support, confidence and number of cycles used, but these rule measures are not considered for Predictive Apriori Algorithm and Tertius Algorithm. The latter algorithm is based on the considered hypothesis and explored hypothesis.
3. The default number of best rules in Apriori Algorithm and in Predictive Apriori Algorithm is respectively 10 and 100 but this is not realized in all case studies.
4. In case of Apriori Algorithm, the number of cycles required to generate best rules has inverse relation with the value of minimum support being used and is independent of number of attributes and instances. The minimum support threshold has great effect on the best rules produced and also on the average size of frequent item sets.
5. In Apriori Algorithm, the number of best rules generated are independent of the number of instances and attributes but are dependent on the value minimum support taken but in case of Tertius Algorithm, the number of best rules produced is dependent on the number of confirmation value that has a default value of 10.
6. In Predictive Apriori Algorithm, the best rules depends on the dataset being used and the number of selected attributes .The greater the number of best rules, the greater the expected accuracy. A rule is added if the expected predictive accuracy of the particular rule is among 'n' number of best rules and it is not a part of another rule with at least the same expected predictive accuracy. However, in Tertius Algorithm the number of rules produced is based on the considered hypothesis .The run information for this algorithm is different from the other two algorithms.
7. During the implementation of Apriori Algorithm, it has been found that the larger the dataset, the larger the execution time of Apriori Algorithm, the lesser the number and size of frequent itemsets, the more memory needed to execute algorithm. Further, in the Tertius Algorithm, the number of datasets and the number of confirmation values are directly proportional to the number of hypothesis considered and explored.

The outcome reveals that Apriori Algorithm is the simplest algorithm out of the three algorithms but has issues associated with the longer time taken and efficiency relative to the size of database. The problem with Apriori Algorithm is that it requires repeated scan of the database. Predictive Apriori Algorithm has a favourable computational performance since it uses dynamic pruning technique. The Tertius Algorithm fails to execute all datasets in time. This algorithm has problems associated with memory and therefore needs high end configuration. In future, we can work on improving the efficiency of Apriori Algorithm and further enhancing the performance of the algorithms. A lot of improvement is needed to improve Tertius Algorithm by finding ways to make it run faster.

VI. CONCLUSION AND FUTURE SCOPE

This paper studies the implementation of Apriori Algorithm on different healthcare datasets in WEKA. The outcome of the study is that this algorithm can be efficiently used to discover hidden patterns and generate associated rules from datasets and is better than Predictive Apriori Algorithm and Tertius Algorithm. The study helps us to know that Apriori Algorithm is an easy implementation in Weka. The review of the implementation of Apriori Algorithm gives the results that can be used by physicians and patients for effective decision making. The execution in Weka gives best rules that determine the relation between various attributes in the datasets. The results can be used by various doctors and healthcare administrators to make more consistent understanding of diseases which occurs frequently.

Table 1. Summarized Conclusion of Literature Survey

Disease/Datasets Used	Work Done	Utility	Future Work
Real – time dataset of a certain area.	Apriori Algorithm is implemented in [11] to predict the occurrence of disease in a particular area. Three different strong association rules are generated with the data set by applying Apriori Algorithm.	An important relation between the attributes in the dataset is identified. The study reveals certain results which help to predict the chance of disease hit in a particular area.	Can apply Apriori Algorithm without candidate generation on the dataset to fix the difficulty of examining database repeatedly. Frequent pattern tree structure can also be used to avoid expensive scanning of database.
Breast cancer, mushroom, larynx cancer, zoo, sunburn, imaginary disease, contact lenses, monk, soya bean and titanic datasets.	Three different association algorithms-Apriori, Predictive Apriori and Tertius Algorithms are executed in Weka on different datasets to compare their results [18]. The performance analysis and suitability of the three algorithms is worked out.	Utility of the association rule mining is identified. Also it helps to find advantages and limitations of the three association algorithms. The results give knowledge about suitable conditions when these algorithms should be used.	Can work on improving the limitations of these algorithms. Other data mining tool like Tanagra can be used to get different interesting results.
29 different diseases	Developed a method using Apriori Algorithm and implemented the large medical dataset using the proposed technique [13] in Weka.	Identified frequent diseases affecting patients over a particular geographical area in a given time period. Results can be used by medical practitioners in making good decisions concerning frequently occurring diseases.	Can combine the given approach with other data mining techniques to have better assessment of results.
Diabetes	Presented a new method [14] to generate association rules on numeric data using Apriori Algorithm and classification technique on a diabetes dataset.	Ten strongest rules are generated for class “diabetes=yes”. Similarly, ten best rules are generated for class “diabetes=No”. These results give risk factors for diabetes.	Needs to find more factors that influence diabetes.

Diabetes Mellitus(DM) with ophthalmic, renal, neurological and peripheral circulatory complications	Association rules are discovered in [12] through the execution of Apriori Algorithm in Weka that relates diabetes mellitus with multiple complications in Thailand considering gender, age and occupation factors.	The results produced can be used for a reliable treatment for patients suffering from diabetic complications. Further, the outcome can be used for the prevention of various diseases and can help in better decision making in identifying disease complications.	Can work on more complication states associated with diabetes mellitus and can apply other techniques to get more results.
Diabetes, Edema and Wheezes	Association rules are generated in Weka to find the risk factors that contribute towards diabetes [15]. The generated rules are also used to find the unknown side effects of diabetes	The study gives results which earlier studies have not strongly considered. Different risk factors of diabetes are found and unidentified effects of diabetes are discovered. The relationship between gender and diabetes, wheezes and diabetes are important achievements of the study.	In future, we can work on finding more risk factors for type1 diabetes. We can work on finding association rules on other different diseases using the same approach.
Diabetes, Edema and Wheezes	Applied classification techniques using decision trees on the same dataset for the research [16] to evaluate the results produced through association.	The results of classification technique evaluates the earlier results produced through association. The results found an interesting relationship between edema and diabetic patients. Further, wheezes has a great effect on increasing the probability of occurring edema for both diabetic type patients.	Can work on modifying Weka to get over the problem of visualizing complete decision tree.

REFERENCES

[1] PrasannaDesikan, Kuo-Wei Hsu and JaideepSrivastava, Data Mining for healthcare Management, SIAM International Conference on Data Mining, Hilton Phoenix East\Mesa, Arizona, USA.

[2] Sotiris Kotsiantis, DimitrisKanellopoulos, "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

[3] Mohammad Mudassar Khan and Prof. AnandRajavat, "An Efficient Algorithm for Extracting Frequent Item Sets from a Data Set", International Journal of Advanced Research in Computer Science and Software Engineering 3(6), June - 2013, pp. 1373-1375.

[4] Arun K. Pujari, Data Mining Techniques

[5] Han Jiawei, MichelineKamber, Data Mining: Concepts and Technique. Morgan Kaufmann lPublishers, 2000

[6] MsShweta and Dr. KanwalGarg, "Mining Efficient Association Rules Through Apriori Algorithm Using

Attributes and Comparative Analysis of Various Association Rule Algorithm", In: Proceeding of IJARCSSE, ISSN 2277-128X, Vol. 3, Issue 6, June 2013.

[7] Smitha.T and V.Sundaram,Association Models for Prediction with Apriori Concept, International Journal of Advances in Engineering & Technology, Nov. 2012, Vol. 5, Issue 1, pp. 354-360.

[8] P.KasemthaweesabandW.Kurutach, "Association Analysis of Diabetes Mellitus (DM) With Complication States Based on Association Rules", 7th IEEE Conference on Industrial Electronics and Applications (ICIEA) 2012.

[9] M. Ilayaraja and T. Meyyappan, "Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm", In: Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 21-22 February.

[10] B. M. Patil, R. C. Joshi, DurgaToshniwal, Association rule for classification of type -2 diabetic patients, Second International Conference on Machine Learning and Computing, 9-11 Feb. 2010.

[11] S.M. Nuwangi, C. R. Oruthotaarachchi, J.M.P.P. Tilakaratna, H. A. Caldera, "Utilization of Data Mining Techniques in Knowledge Extraction for Diminution of Diabetes", Second Vaagdevi International Conference on Information Technology for Real World Problems (VCON), 9-11 Dec, 2010

[12] S.M.Nuwangi, C. R. Oruthotaarachchi, J.M.P.P. Tilakaratna& H. A. Caldera, "Usage of Association rules and Classification Techniques in Knowledge Extraction of Diabetes", 6th International Conference on Advanced Information Management and Service (IMS), Nov. 30 2010-Dec. 2 2010..

[13] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, San Francisco, 2 edition, 2005.

[14] A. Lekha, Dr. C V Srikrishna and Dr. VijiVinod, "Utility of Association Rule Mining: a Case Study using Weka Tool", 2013 International Conference on Emerging Trends in VLSI, Embedded System, Nano Electronics and Telecommunication System (ICEVENT),7-9 Jan, 2013