INTERNATIONAL JOURNAL FOR TRENDS IN ENGINEERING & TECHNOLOGY
VOLUME 3 ISSUE 1 –JANUARY 2015 - ISSN: 2349 - 9303

Summarization Techniques in Association Rule Data Mining For Risk Assessment of Diabetes Mellitus

Padmapriya.S[1]
[1]Arunai Engineering College, CSE,
vijipadmapriya@gmail.com

Jaya kumar.P[2]
[2]Arunai Engineering College, Department
jai13it@gmail.com

Abstract— At Early exposure of patients with dignified risk of developing diabetes mellitus is so hyper critical to the bettered prevention and global clinical management of these patients. In an existing system, apriori algorithm is used to find the itemsets for association rules but it is not efficient in finding itemsets and it uses only four association rules for finding the risk of diabetes mellitus so it have low precision. In this paper we are focusing to implement association rule mining to electronic medical records to detect set of danger factors and their equivalent or identical subpopulations that indicates patients at especially steep risk of progressing diabetes. Association rule mining accomplishes a very bulky set of rules for summarizing the EMR with huge dimensionability. We proposed a system in enlargement to combine risk of diabetes for the purpose of finding an suitable summary for this we use ten association rule and using the reorder algorithm for finding the itemsets and rules. For identifying the risk we considered four association rule set summarization techniques and organised a related calculation to support counselling with respect to their applicability merits and demerits and provide solutions to reduce the risk of diabetes. The above four methods having its fair strength but the bus algorithm developed the best acceptable summary.

Index Terms— Data Mining; Fuzzy Clustering means Algorithm; Association rule mining; Association rule summarization techniques.

1 INTRODUCTION

Diabetes may be ane sort of diseases characterized by High blood glucose level (blood sugar). If a person has diabetes mellitus, the body either doesn't manufacture enough hypoglycemic agent or the body is unable to use its own insulin. Aldohexose builds up within the blood and causes a condition that, if not controlled, will result in serious health complications such as stroke and even death. The chance of death for someone with diabetes mellitus is double the chance of someone of comparable age who doesn't have diabetes mellitus.

Diabetes may be a major reason for heart attack and stroke. Death values for heart attack and therefore the risk of stroke area unit concerning 2–4 times higher variant cluster adults with diabetes mellitus than variant cluster those without diabetes mellitus. UN agency reported that 67% U.S adults have diabetes mellitus additionally report having High blood force per unit area.1 for folks with diabetes mellitus, High blood glucose per unit area levels, High cholesterol level, and smoking increases the chance of heart attack condition and stroke. That risk may be reduced by dominant force per unit area and cholesterol levels and stopping smoking. In response to the pressing got to notice ascertained patients in datasets at High blood glucose risk of diabetes mellitus early, numerous diabetes mellitus risk indices (risk values) are developed. A number of specific of those indices (e.g. the Framingham score [15]) gained acceptance in clinical apply and area unit used as steering in remedy for illness.

Diabetes mellitus have three types. Type 1diabetes mellitus - the body doesn't manufacture hypoglycemic agent. Some 100% of all diabetes mellitus cases area unit sort one. Type 2diabetes mellitus in this the body doesn't manufacture enough hypoglycemic agent for correct operate. Some ninetieth of all cases of diabetes mellitus worldwide area unit of that sort. Gestational diabetes mellitus - that sort affects females throughout maternity. The most common diabetes mellitus symptoms embrace frequent voiding, intense thirst and hunger, weight gain, unusual weight loss, fatigue, cuts and bruises that don't heal male sexual pathology, symptom and tingling in hands and feet.

Association rules are implications that relate a set of potentially interacting conditions (e.g. max Body mass index and the occurrence of hypertension diagnosis) with eminent risk. The use of association rules is mainly favourable, because in addition to signifying the diabetes mellitus risk, they also readily provide the physician with a "justification", namely the related set of conditions. Namely co-morbid sickness, laboratory results, tablets and demographic information those are commonly available in electronic medical record (EMR) systems. With such an extensive set of factors of risk, the set of invented rule for data grows combinatorially large,

to a size that severely hinders interpretation. To overcome that challenge, we applied rule for data set summarization techniques to compress the original rule for data set into a most compact set that can be interpreted with ease.

Association rule mining is a method used to discover associations among the items. Applied to a cure for disease and condition, association rule can be viewed as finding phenotypes or etiologic pathways within population. They are interpretable, and they suggest interconnections between the factors of risk. Furthermost, they are rule for data, which makes them directly promote to execute in a clinical decision support system. While association rules can discover observed patients in dataset subpopulations (phenotypes) at mainly max risk of a Provide disease, they do not directly give us information about the efficacy of remedy for diseases. In that work, we have to extend the association rule data extracting process methodology to find subpopulations where the results at final of a remedy for disease at final differs variant group subpopulations, or differs from the in normal population.

2 RELATED WORK

In general, there is extensive literature on measuring the risk of diabetes mellitus. This section reviews about the some related work in order to explore the strengths and weakness of existing methods.

Rakesh Agrawal, Ramakrishnan Srikant [1] this paper propose two new algorithms for solving that problem that are fundamentally variant from the known other algorithms. It shows the comparative performance of the specified Apriori and AprioriTid algorithms against the AIS and SETM algorithms.

Foto Afrati, Aristides Gionis, Heikki Mannila [2] In that paper we have to address the issue of over large output size by introducing and studying the problem What are the k sets that best approximate a collection of frequent variable set of items . Our measure of approximating a collection of sets by k sets is defined to be the size of the collection covered by the k sets.

Eghbal G. Mansoori [3] Fuzzy grouping of data is superior to crisp grouping of data when the boundaries variant group the groups are vague and ambiguous. Propose a novel fuzzy rule for data-depends grouping of data algorithm (FRBC).

Aysel Ozgur, Pang-Ning Tan, and Viper Kumar [4] a framework for making regression models by using the rule for data for association of data. Propose a pruning scheme for redundant and insignificant rule for datas in the rule for data extraction step, and also a number of heuristics for making regression models.

Jian-Ping Mei and Lihui Chen [5] the original FCM uses Euclidean normally used to measure the object-to centroid distance. To propose a new fuzzy grouping of data approach called LinkFCM where an additional term is added into a fuzzy c-mean grouping of data type approach.

Existing systems intend to apply association rule mining to electronic medical records to determine sets of risk factors and their consequent subpopulations that symbolize patients at mainly elevated risk of increasing diabetes. Given the elevated dimensionality of EMRs, association rule mining generates a extremely huge set of rules which we have to summarize for simple medical use. We reviewed four association rule set summarization techniques and conducted a relative assessment to provide assistance concerning their applicability, strengths and weaknesses.

3 PROPOSED ALGORI THM

We present the proposed new fuzzy clustering of data algorithm that works on the variable set of items in the dataset of the Electronic medical records. The clinical application of association rule data mining is to find the set of items of health conditions that shows the consequent quantity of increased amount of risk of making diabetes mellitus. Association rule data extracting process in data mining used to describe extensive set of variable set of items resulted in an exponentially huge set of association rules formed. The main contribute is a comparative calculation of these enlarged summarization techniques that gives guidance to practitioners and Observed patients in datasets for choosing a relevant algorithm for a same type of problem in the domain.

Methodology

- Twelve association rule.
- Apriori hybrid algorithm.
- Fuzzy Clustering C Means.

A. Distributional association rule

A **distributional association rule** is defined by an itemset I and is an implication that for a continuous outcome y, its distribution between the affected and the unaffected subpopulations is statistically significantly different. For example, the rule {htn, fibra} indicates that the patients both presenting hypertension (high blood pressure) and taking statins (cholesterol drugs) have a significantly higher chance of progression to diabetes than the patients who are either not hypertensive or do not have statins prescribed. The distributional association rules are characterized by the following statistics. For rule R, let OR denote the observed number of diabetes incidents in the subpopulation DR covered by R. Let ER denote the expected number of diabetes incidents in the subpopulation covered by R.

$$ER = OR - i \in DRyi,.$$

Where yi is the martingale residual for patient i.

The **relative risk** of a set of risk factors that define R is

$$RR = OR/ER.$$

Table1: Values of the risk factors that appeared in any of the summarized rules.

Parameter	Weightage	Values
Male &Female	Age<30	0.1
	>30to<50	0.3
	Age>50&Age<70	0.7
	Age>70	0.8
Smoking	Never	0.1
	Past	0.3
	Current	0.6
Overweight	Yes	0.8
	No	0.1
Alcohol Intake	Never	0.1
	Past	0.3
	Current	0.6
Heart Rate	Low(<60bpm)	0.9
	Normal(60to100bpm)	0.1
	High(>100bpm)	0.9
Blood Sugar	High(>120&<400)	0.5
	Normal(>90&<120)	0.1
	Low(<90)	0.4
Bad Cholesterol	Very High>200	0.9
	High(160to200)	0.8
	Normal<160	0.1

Table2: Description of the risk factors that appeared in any of the summarized rules.

Factor	Description
bmi	body mass index
sbp	systolic blood pressure
dbp	diastolic blood pressure
hdl	high-density lipoprotein
tchol	total cholesterol
trigl	triglyceride
Medications:	
acearb	ACE inhibitor and angiotensin receptor
blocker	bb Beta-blocker
ccb	Calcium-channel blocker
diuret	Diuretics
fibra	Fibrates
statin	Statin
aspirin	Aspirin
Co-morbidities:	
htn	Hypertension
tobacco	Current smoker
ihd	Ischemic Heart Disease

B. Rule set and database summarization

The main aim of rule set summarization is to represent a set of rules I with a smaller set of rules A such that itemset I can be recovered from the itemset A with minimum loss of information and data. Data set summarization main aim is to indicate a data set D with a smaller set data set A of itemsets D can be recovered from A with minimal loss

C. Fuzzy Clustering C Means

In fuzzy clustering means the data elements in the item set can belong to more than one cluster. In this fuzzy clustering each and every point has a degree of belonging to as in fuzzy logic. It does not belonging entirely to just one cluster or same cluster.

D. Apriori Hybrid A lgorithm

Apriori and Apriori TID algorithms can be combined into a hybrid algorithm, called AprioriHybrid.

AprioriHybrid scales linearly with the number of transactions. In addition, the execution time decreases a little as the number of items in the database increases. As the average transaction size increases (while keeping the database size constant), the execution time increases only gradually. These experiments demonstrate the feasibility of using AprioriHybrid in real applications involving very large databases.

E. Extension to Account for Outcome and Patient Coverage

In this section, we discuss how we extended these techniques to incorporate the risk y of diabetes as manifested by the martingale residual. Since we are particularly interested in rules that predict high risk of diabetes, we can add $\bar{y}(I)$ the subpopulation mean risk of diabetes to the criterion with a weight λ that controls how much importance is assigned to the risk and how much to the other components of the criterion. Let $L*(I)$ be the resulting criterion and $L(I)$ the original criterion $L*(I) = -\lambda\bar{y}(I) + (1-\lambda)L(I)$.

Patient coverage is nothing but it denotes the total number of patients (or alternatively, cases) who are covered by any of the rules discovered by using the apriori hybrid in the summary set A. The sum squared prediction error denotes the how accurately the rules are discovered and how it can predict the risk of patients relative to the rule set. Restoration error denotes the how effectively we can restore our data set D from a summary rule set A.

F. Sum squared prediction error

It is one of the objective measures to evaluate the four summarization techniques used for summarization. The main goal is to assess the set of rules to identify how accurately this set of rules can be used to predict the risk of diabetes mellitus for the patients. At the end, first we have to calculate a "gold standard" estimate of each patients risk $\tilde{y}i$ depends upon the whole original rule set I after that compare the estimate $\hat{y}i$ obtained using the summary rule set to $\tilde{y}$. We have to calculate the "gold standard" estimate by using a boosted linear regression model using cross-validation. The predictors of the model are rules in the original rule set I and the outcome is the martingale residual y. Given a summary rule set A, which is an ordered set of rules, we make a prediction for patient i

through the first rule Ai that covers patient i. The predicted value is the subpopulation mean outcome on the training set.

ˆyi = ¯y (Ai) = meanj∈Dai yj.

The sum squared prediction error (SSPE) is the summed square difference between the risk predicted by the summary rule set ˆyi and the gold standard estimate ˜yi SSPE =I (ˆyi − ˜yi) 2.

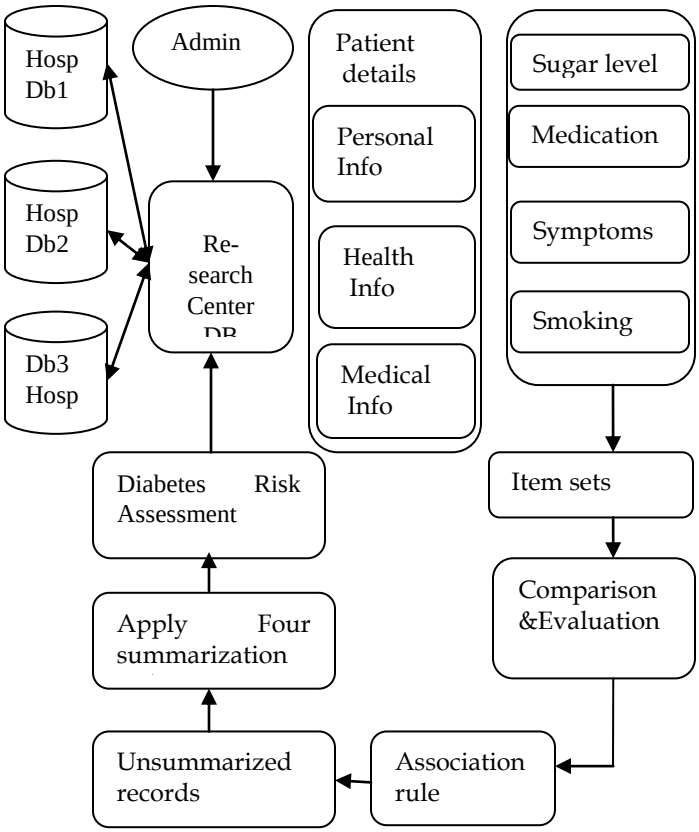
G. Summarization Techniques and Summarized Rule Set

Summarization is a key data mining concept which involves techniques for FInding a compact description of a dataset. Simple summarization methods such as tabulating the mean and standard deviations are often applied for data analysis, data visualization and automated report generation.Four summarization techniques are used.

we present the rule sets generated by the extended summarization algorithms. For each one algorithm, it provided the best suitable outcome because we used the parameter settings. For APRXCOLLECTION, we used α = .1, λ = 1; for RPGlobal, we used δ = .5, σ = .2, λ = .98; for Top-K, we used λ = .2; and for BUS, we used λ= 1. Note that λ notably varies from 1 single for Top-K, which previously takes the risk of diabetes into relation in the usual loss condition.

4 SYSTEM DESIGN

Figure 1 demonstrates the framework of our proposed approach.



A. Upload database:

Initially in our application there is no Database Records. We are going to implement summarization techniques in a Distributed Database not only in a single database. So have to ask permission to access the database of every Health Center Administrator or hospital administrator. In my application research center only find outs the risk of diabetes mellitus so the research center must get the database. For that it sends the request to several hospitals for accessing the hospital database.

At first research center sends the request to hospitals it reaches the hospitals for that request hospital admin replies whether it can access or not. If the hospital gives an access we can able to access the hospital database .Then the Observed patients in dataset database are fetched into the research center database with privacy preservation. Fetched database consists only needed information in clear manner now we can able to see the Observed patients in dataset medical information and id. The Specific Observed patients in dataset can be identified by means of their ID itself. After fetching the database we have to extract the relevant information for our project such as sugar level, BP, Medications etc. After extracting those values we have to form the variable set of items or grouping depends on the medications by using the fuzzy grouping of data techniques.

B. Discover Item sets and association rule

In fuzzy grouping data elements can belong to more than one cluster. In fuzzy grouping of data every point has a degree belonging to as in fuzzy logic rather than belonging completely to just one same cluster.

Next step is finding of association rules by using the apriori hybrid algorithm. We used the apriori hybrid algorithm, a variant of the well-known Apriori algorithm that discovers candidate set of items that contain specific items the item corresponding to the (binary) diabetes mellitus results at final in our case.

C. Un summarized rule for data

Next step is to find the unsummarized rule for data. It consists of the comparative risk and complete risk of Observed patients in datasets. These values are calculated depends on the sugar level,BP,BMI,Tablets etc.Every value consists of some of particular defined value depends on the gender and age by summing that values we can calculate the comparative risk and absolute risk.

D. Apply summarization techniques

We apply rule for data set summarization techniques namely APRX-COLLECTION, RPGlobal, TopK, BUS to evaluate the Risk of Diabetes mellitus. Predicition of Diabetes mellitus depends on Body condition, Tablets and Co., Morbites of the Observed patients in dataset subpopulation. While all four methods created reasonable summaries, every method had its clear benefits.

APRX-COLLECTION and RPGlobal primarily operate value on the aspect countenance of the rule for data with a primary objective of maximizing compression. TopK and BUS operate primarily on the Observed patients in datasets and their objective–

especially in case of TopK can be thought of as minimizing redundancy. Between the TopK and BUS, this allowed it to have better Observed patients in dataset coverage and better skill to reconstruct the original data base.

The APRX-COLLECTION algorithm finds supersets of the conditions (factors of risk) in the rule for data such that most subsets of the summary rule for data will be valid rule for data in the original (unsummarized) set. The RPGlobal summarization is similar to APRXCOLLECTION in that it is chiefly concerned with the aspect countenance of the rule for data, and then it performs a very aggressive compression. The Redundancy-Aware Top K (TopK) algorithm further reduces the redundancy in the rule for data set which was possible through operating on Observed patients in datasets rather than the aspect countenances of the rule for datas.BUS (as opposed to TopK) operates on the Observed patients in datasets and not on the rule for data, redundancy in terms of rule for data aspect countenance can occur.

APRX-COLLECTION

APRX-COLLECTION algorithm finds supersets of the conditions (risk factors) in the rule such that the majority subsets of the summary rule will be legal rules in the original (unsummarized) set and these subset rules imply related danger of diabetes

.

Table 3: Risk values of diabetes using APRX collection.

RR	ER	OR	RULE
1.96	**36.24**	**71**	Fibra accearb
1.34	**271.71**	**363**	Bmi trigal statin aspirin htn
1.19	**426.78**	**506**	Hdl trigl accerab aspirin htn
1.31	**348.92**	**457**	Bmi trigal accearb statin aspirin ihd
1.23	**534.58**	**660**	Bmi sbp ccb htn

RPGLOBAL

The most important limitations of APRX-COLLECTION were the redundancy in the set of rules and the intensity of the risk. The RPGlobal summarization is parallel to APRXCOLLECTION in that it is primarily concerned with the rule expression, and consequently it performs a exceedingly destructive compression. However, it addresses the two limitations by taking patient exposure into relation and by constructing the summary from rules in the original rule set (as different to an extended set).

Table 4: Risk values of diabetes using RP-Global.

RR	ER	OR	RULE
2.40	**21.70**	**52**	Fibra statin htn
1.5	**37.97**	**60**	Bmi hdl ihd
1.47	**45.52**	**67**	sbp statin htn tobacco
1.46	**317.03**	**464**	Bmi accearb htn
1.62	**32.16**	**52**	sbp tchol hdl trigl htn

TOP-K

The Redundancy-Aware Top K (Top-K) algorithm additionally reduces the redundancy in the rule set which was achievable during operating on patients slightly than the expressions of the rules. Top-K still achieves elevated compression rate.

Table 5: Risk values of diabetes using TOP-K

RR	ER	OR	RULE
2.40	**21.70**	**52**	Fibra
1.5	**37.97**	**60**	Bmi hdl ihd
1.47	**45.52**	**67**	sbp statin htn tobacco
1.46	**317.03**	**464**	Bmi accearb htn
1.62	**32.16**	**52**	sbp tchol statin hdl trigl htn

BUS

Bottom Up Summarization it is different from Top-K. BUS operates on the patients and it does not operate on the rules. Therefore, redundancy in terms of rule expression can happen. Conversely, BUS unambiguously reins the redundancy in the patient space during the parameter mandating the lowest number of new (before exposed) cases (patients with diabetes event) that require to be enclosed by every rule. Thus the compact variability in the rule expression does not transform into amplified redundancy.

Table 6 : Risk values of diabetes using BUS.

RR	ER	OR	RULE
2.34	**24**	**57**	bmi trigal acearb statin htn
2.10	**25**	**54**	Hdl trigal diuret aspirin htn
1.91	**56**	**107**	Bmi trigal accearb statin htn
1.54	**78**	**121**	bmi trigal tobacco htn
1.37	**39**	**54**	Dbp diuret htn

E. Send back results to hospital

Finally the summarization results calculated in APROX, RP GLOBAL, TOP-K and BUS are send back to the hospital for clinical guidance.

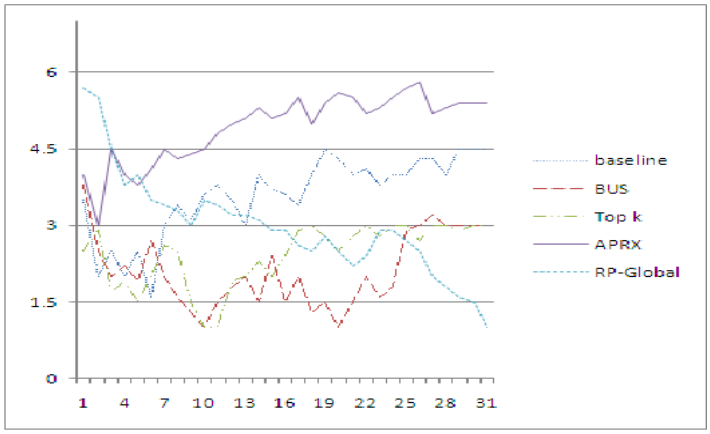
5 DISCUSSIONS

Risks are identified and the risk factors are given to the hospitals for suitable guidance to the patients.The risk factors are stored in the database.

6 RESULTS

Our proposed technique aim to assess the risk of diabetes mellitus for the patients. In this we are using the apriori hybrid algorithm to find the item set it discovers the item set in efficient manner. Fuzzy clustering used to discover the item set and

association rules in this the main advantage is rules are discovered using more than one cluster. The risk to assess the risk of diabetes mainly focuses on the medications of the patients. The below graph
is drawn by using the netbeans software.



Number of Rules

Fig. 2. Sum squared prediction error of the summarization methods as a function of the number of rules on cases.

7 CONCLUSION

In that project, we proposed the summarization for risk prediction of diabetes mellitus. That approach gives significant assistance to the Observed patients in datasets and doctors. It can expose hidden (unseen) clinical relationships and can suggest new patterns of conditions to forward prevention.

Association rule mining to discover sets of risk factors of risk and the consequent Observed patients in dataset subpopulations who are at consequently amplified risk of making progress to diabetes mellitus. An extreme number of association rule we give the clinical explanation of the results. For this scheme to be helpful, the number of rules is used for clinical explanation is build realistic and consistent.

ACKNOWLEDGMENT

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. “Fast algorithms for mining association rules”. In VLDB Conference, 1994.

[2]Eghbal G.Mansoori, “FRBC: A Fuzzy Rule-Based Clustering Algorithm,”proc IEEE transaction. Fuzzy sytems. Vol 19no.5, October 2011.

[3]Aysel Ozgur, Pang-Ning Tan, and Vipin Kumar. “RBA: An integrated framework for regression based on association rules”. In SIAM International.

[4]Jian-Ping Mei, Lihui Chen. “A Fuzzy approach for multitype relational data clustering,”IEEE transactions on Fuzzy systems, vol 20, No.2, April 2012.

[5]Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. “Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting”. BMC Medicine, 2011.

[6] Diabetes Prevention Program Research Group. “Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. The New England Journal of Medicine”, 346(6), 2002.

[7]Xiaoxin Yin and Jiawei Han. “CPAR: Classification based on predictive association rules”. In SIAM International Conference on Data Mining (SDM), 2003.

[8]Mohammad Al Hasan. “Summarization in pattern mining. In Encyclopedia of Data Warehousing and Mining”, (2nd Ed). Information Science Reference, 2008.

[9] R.Srikant, Q. Vu, and R. Agrawal. “Mining association rules with item constraints. In American Association for Artificial Intelligence (AAAI)”, 1997.

[10]Terry M. Therneau and Patricia M. Grambsch. “Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health”, Springer, 2010.

[11] Ruoming Jin, Muad Abu-Ata, Yang Xiang, and Ning Ruan. “Effective and efficient itemset pattern summarization: Regressionbased approach”. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2008.

[12] Bing Liu, Wynne Hsu, and Yiming Ma.”Integrating classification and association rule mining”. In ACM International Conference on Knowledge

[13] Xiaoxin Yin and Jiawei Han. “CPAR: Classification based on predictive association rules”. In SIAM International Conference on Data Mining (SDM), 2003.