A hybrid methodology for Knowledge mining with the application PRM

Dr. Ritu Bhargava
Lecturer, Computer Science
Sohpia Girls' College Ajmer, India
drritubhargava92@gmail.com

Abhishek Kumar
Assistant Professor Aryabhatta College of Engg. & Research Center, Ajmer India
abhishekkmr812@gmail.com

Prerna Pareek
Research scholar
MJRP University Jaipur
prerna.pareek2907@gmail.com

Abstract*—Data mining is a very effective technique for retrieval of knowledge from huge datasets. This technique is becoming an increasingly very important tool to find the key data from heterogeneous data or data sets such as Data warehouses, Data Marts, Data repositories etc. The extracted data must be operated to understand and analyze it. These operations may be classification, association or any kind of Data mining related tools. The Classification based on Predictive Association Rules (CPAR), we have used in our case. It is rule-based learning. We have performed various Association based test with Classification techniques to analyze the data and find the results. PRM uses the key concept of decreasing weights rather than removing records. The CPAR is more efficient in rule-based learning. The paper makes bases from mathematical rigors used for data mining techniques and we have performed the operation with respect to the basic equations and techniques. The efficient data mining using PRM has been applied in this particular work.

***Keywords*— Data mining, CPAR, PRM, Association Rule.**

I. INTRODUCTION

Data mining is very innovative and classic technology that is primarily used for mining or dredging and meaning information from the huge database[1,2,3]. It basically deals with dealing scattered and irrelevant data and makes it pertinent enough to get in converted into knowledge as all information is not knowledge. Data mining is the result of elongated research work advancement. This method got in to shape when large business data were stored in meaningless order so in order to get those data into pertinent order and making it for the knowledgeable pattern. Frequent example mining is an all-around examined issue that intends to find the connections among items in light of their events in exchanges. Because of the development of uses that include dubious information, conventional ways to deal with mine frequent examples may not be relevant for a few genuine applications. As of late, quality research has been directed to relate the events of items in unverifiable databases for applications like informal communities, sensor systems, protein-protein connection examination and wrong studies.[4,5] Vulnerability could likewise emerge from concealing of information protection concerns. Not at all like certain information where the events of items in exchanges are unequivocal, in an indeterminate database.[6,7] In this paper, we center around three essential issues in information mining including unverifiable information. In particular, we propose two calculations for weighted frequent examples mining, a disjunctive association rules mining calculation, and two calculations for finding both conjunctive and disjunctive causal tenets mining, all from unverifiable information. The importance weight-based example mining methodology can be felt in numerous areas, for example, biomedical information examination where the causes of diseases are one quality as well as a blend of qualities; web traversal design mining where the effect of each site page is extraordinary; et cetera. In this way, numerous calculations have been proposed for measuring items in terms of their hugeness.

Predictive rule mining is an algorithm which transforms basically the FOIL algorithm to better accuracy and effectiveness.PRM is basically driven by a rule despite removing, the weight is reduced with the factor multiplication process. What happened in case of Foil, it contracted more rules and each positive instances is usually sheltered more than once. The primary work of PRM is to minimize the workload complexity of FOIL [8, 9] while selecting every literal during rule establishing process, The PRM elects only optimum literals and eliminates all other .there are very few literals which are frequent in nature and having similar gains.There are usually numerous rules with the same accuracy completely remaining dataset oriented.The optimum rule among them may not be considered as the best rule based on whole datasets. The way in which PRM works it selects generally one of them which can further results in missing some prominent rules.[10]

II. LITERAUTURE SURVEY

Association rules mining is a typical data mining issue that investigates the connections among items in light of their events in exchanges. Customary ways to deal with min frequent examples may not be material for a few genuine applications. Rather than deterministic or certain data where the events of items in exchanges are unmistakable, in an unverifiable database. For this situation, the recurrence of an item (or itemsets) is ascertained as the normal number of events of the item (itemsets) in the exchanges. Authors, [11] in this, mining data from a database are the principal point of data mining. The most pertinent data because of data mining is getting relations among different things. All the mining independent frequent itemsets is the resulting factor in the mining of various data itemsets.[9] Numerous algorithms talked about in the previous works require different sweep of the database to get the data on different sub ventures of the algorithm which ends up troublesome. Here Author is proposing an algorithm combined predictive rule mining approach with FOIL and CPAR,[10] It should mine huge amount of data from a database just in one iteration. It utilizes Lexicographic requesting of data thing esteems and frequent itemsets in different domain which are connected to their consistent pattern analysis. Authors [11] in this, the developing a methodology and improvement of data mining innovations result in genuine risk to the security of individual or bulk data. The most recent research point in data mining called predictive rule mining was considered in recent years. The primary work of PRM is to minimize the workload complexity of FOIL while selecting every literal during rule establishing process, The PRM elects only optimum literals and eliminates all other .there are very few literals which are frequent in nature and having similar gains.There are usually numerous rules with same accuracy completely remaining dataset oriented .The optimum rule among them may not be considered as the best rule based on whole datasets. The essential thought of PRM is to adjust the data such that to perform data mining algorithms successfully with security of individual data contained in the dataset. Presently a day investigations of PRM for the most part discussed on how to improve the security hazard brought by data mining tasks, be that as it may, actually, undesirable exposure of individual data may likewise occur during the time dependent data, data distributing, furthermore, gathering information.[12] This think about focus on the rules eliminating issues associated to data mining from a more extensive point of view and concentrate different approaches that can ensure individual data. In specific, discover four unique kinds of clients associated with data mining applications, to be specific, data mining Authority, data supplier, data collecting authority, and decision maker authority. Every client, talk about his protection concerns and the strategies that can be received to ensure touchy data. At that point quickly display the pros and cons of related research points, audit best in class approaches, also, exhibit a few contemplations on future research bearings. Other than investigating the deletion of eliminated rules approaches for each type of client, audit the hypothetical methodologies, which are proposed for breaking down the interchanges among distinctive clients in a data mining situation, every one of whom has his own valuation on the individual data. By separating the task of various clients concerning rules eliminating process of individual data, it will give some valuable bits of knowledge into the investigation of PRM. Authors [13] in this, Data mining is utilized for mining valuable data from enormous datasets and discovering important groupings from the data. More establishments are presently utilizing data mining methods in a day to life. PRM mining method has turned into a critical in the field of explore. Different advances have been executed to enhance the execution of predictive rule mining algorithms. These investigations gives the preliminaries of essential ideas about regular succession eliminated rules and present an overview of the advancements. An exploratory outcome indicates preferred execution over FOIL & CPAR [15]. So here focus on PRM algorithm that support the deletion of rules after elimination so it will reduce the complexity but at the same time will make removal of some important data as well.]. The above talked about investigation and insight about the illness and attributes has been utilized by numerous researchers in different work for better groupings and expectations. The information mining technique is totally reliant on the datasets just the better properties and measure of data can better break down the execution of the different algorithm. there is constantly one restriction related with information mining activity that is the means by which proficient the datasets and their traits are for instance now and again qualities are less and in a few information availabilities is less in the two cases the precision of forecast and classification will endure. [16]

, Various works have been performed identified with wellbeing information prediction. Authors [12] Have proposed an approach in which example and connection were portrayed identified with coronary heart sickness. Despite the fact that the work can be utilized substantially more methods to enhance the precision, the creators clarify the manifestations and treatment in an extremely successful way.

Silbershatz H, et al [15] has proposed more refined work with the use of molecule swarm calculation and feed forward back spread, In the work particularly information mining methods have been connected for the forecast of the appropriate outcome in view of the age, sexual orientation and different highlights of the patients.

Simons et.al & Shahwan-Akl [16,17] has mulled over a joined procedure over an enormous dataset utilizing diverse qualities in the datasets like circulatory strain, cholesterol, age, sexual orientation, and heartbeat rate and so forth. This work portrays the forecast with more precision similarly and having blend and investigation of numerous calculations like guileless Bayes classifier and choice tree and KNN. It was recommended that when qualities are diminished then no one but exactness can be accomplished.

Bhargava et.al [18] utilized the predetermined number of properties for the given datasets, it depended on fundamentally the side effects and the passing rate because of infections .the work basically in light of the dataset records which was least as far as the odd data, this work has not anticipated according to high computing capacities of the calculation.

Albeit numerous works in this area have been performed for ideal outcomes and better expectation characteristics .high finish rate has been accomplished however on less trait and littler datasets. This missing dimension has a tendency to propose a model of the procedure with the blend of numerous classic algorithms keeping in mind the end goal to get them all the more wide measurement of results and errors of classifiers.

III. METHODOLOGY

There are eight primary attributes in the dataset. Each deals with the exception of Disease is the fundamental characteristics of the Heart Disease. Each reason is ordered by some predefined measures and parameters. These measures are sorted by the making result more knowledge oriented. [15]

```
Proposed Algorithm
Procedure Predictive Rule Mining
Set the weight threshold  of every example to 1
The rule set R ⟵ ϕ so that min threshold can be set
Total Weight ⟵ Total Predictive weight
A ⟵  Compute PN Array from D, dataset
While Total Weight (P) > δ. Total Weight
      N ‘ --- - N,  P’ ---P, A’----A
Ruler -----empty rule
While true
      Optimum  literal p will be found   according to A ‘
If gain (p) < min_gain then break
      Append p to r
For each example t in P’ N’ not satisfying r’s body
      Remove t from P’ or N’
      Change A ‘ according to the removal of t
End
End
R---- R {r}
For each example t in P satisfying r’s body
      t.weight   ⟵   α t.weight
change A according to the weight decreased
end
end
return R
```

IV. RESULT

Predictable attribute

1. Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1:>50% diameter narrowing (has heart disease) [18]

Key Attribute

Patient Id- Patient’s identification number

Input Attributes

1. Age- in years
2. Sex (Value 1: Male; value 0:Female)
3. Chest Pain Type (Value 1: typical type 1 angina, value 2: typical type angina, vale 3: non-angina pain; value 4: asymptomatic)
4. Fasting Blood sugar (value 1:>120mg/dl; Value 0:<120 mg/dl)
5. Resting- Resting electrographic results (value 0: normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
6. Exchange- exercise-induced angina (Value 1: yes, Value 0: no)
7. Slope-The slope of the peak exercise ST segment (Value 1: unsloping; value 2: flat; value 3: downsloping)
8. CA- No. of major vessels colored by fluoroscopy (Value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholesterol (mg/dl)
12. maximum heart rate achieved
13. Old peak- ST depression induced by exercise
14. Smoking (value 1: past, value 2: current; value 3: never)
15. Obesity (value 1: yes; value 0: no)

The patient is evaluated using these attributes. Diagnosis value 0 means no heart disease; value 1 means has heart disease.

Statistics of Dataset

- Number of Attributes: 15 normal attributes 1 key attribute to being used by algorithms.
- Number of Tuples (For Training): 1728
- Number of Tuples (For Testing): 1525

Class-wise distribution of data (For Training Set)

Diagnosis	**No. of Patients**
Value: 0 (no heart disease)	1210
Value: 1 (has heart disease)	518

Table-1. Class-wise Distribution in Training Set

Class-wise distribution of data (For Test Set)

Diagnosis	No. of Patients
Value: 0 (no heart disease)	1108
Value: 1 (has heart disease)	417

Table-2. Class-wise Distribution in Test Set

The parameters for the individual algorithms are as follows:

FOIL: In this particular case Maximum of three attributes in the antecedent of a rule are considered for testing the dataset compatibility.

PRM: The Minimum gain threshold is equal to 0.7, total weight threshold is equal to 0.05, and the decay factor considered =2/3.

CPAR: The Minimum gain threshold is equal to 0.7, and the total weight threshold is equal to 0.05, whereas the decay factor =2/3.

Note also that the optimum five rules are used, in each case, when classifying a test record and that for CPAR a similarity ratio of 1:0.99 was used.

A. PRM Algorithm

Testing Mode	Accuracy	No. of Rules Generated	Generation
50:50	99%	4	0.02
75:25	99.29%	5	0.02

Table3.Comparison of PRM algorithm with different measures

B. CPAR algorithm

Testing Mode	Accuracy	No. of Rules Generated	Generation
50:50	98.14%	4	0.02
75:25	99.29%	5	0.02

Table- 4: Comparison of CPAR algorithm with different measures

Above Results Show that CPAR is better Algorithm in terms of rule generation and Classification Accuracy. But when the parameters are dynamic and PRM is compared with FOIL then PRM has better results as shown in above tables.

V.CONCLUSION AND FUTURE WORK

The paper focuses on the mathematical rigors and efficient and meaningful analysis of larger datasets to get at some specific results. We have analyzed the data of diseases in such a way that the confidence support and confusion matrix created with knowledge receiving.The PRM algorithm gets better classification and results as far as bigger datasets are concerned. Basically, authentic data sets have been taken with nominal attributes. The future work can be considered the PRM algorithm with the combination of the different approach in order to improve the accuracy of the class results. The work basically explained that after application of PRM with other efficient algorithms like CPAR and FOIL will always provide better results as compared to the application of PRM over the same datasets. The future work can be predicting the knowledge out of huge dataset using the combination of different related methodology.

References

[1] Feng Gui, Yunlong Ma, Feng Zhang, Min Liu, Fei Li, Weiming Shen, Hua Bai, "A Distributed Frequent Itemset Mining Algorithm Based on Spark", Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)

[2] Hongjian Qiu, Yihua Huang, Rong Gu, Chunfeng Yuan, "YAFIM: A Parallel Frequent Itemset Mining Algorithm with Spark", 2014 IEEE 28th International Parallel & Distributed Processing Symposium Workshops

[3] Paladugu S (2010) Temporal mining framework for risk reduction and early detection of chronic diseases. The University of Missouri-Columbia.

[4] Obenshain MK (2004) Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology 25: 690-695.

[5] Shillabeer A, Roddick JF (2006) Towards role based hypothesis evaluation for health data mining. Electronic. Journal of Health Informatics 1: 1-9.

[6] Porter T, Green B (2009) Identifying Diabetic Patients: A Data Mining Approach.

[7] Panzarasa S, Quaglini S, Sacchi L, Cavallini A, Micieli G, et al. (2010) Data mining techniques for analyzing stroke care processes. In the Proc. of the 13th World Congress on Medical Informatics.

[8] Li L, Tang H, Wu Z, Gong J, Gruidl M, et al. (2004) Data mining techniques for cancer detection using serum proteomic profiling. Artificial intelligence in medicine 32: 71-83.

[9] Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications 36: 7675-7680.

[10] Lakshmi K, Krishna MV, Kumar SP (2013) Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability. International Journal of Scientific and Research Publications 3: 1-10.

[11] Centers for Disease Control and Prevention (2013) Chronic Disease Prevention and Health Promotion. Accessed 27 September 2013, from http://www.cdc.gov/nccdphp/

[12] U.S department of health and human services (2005) High Blood Cholesterol What you need to know.

[13] Department of Health & Aging AG (2012) Seniors and Aged Care Australia websites have been replaced.

[14] Heller RF, Chinn S, TunstallPedoe HD, Rose G (1984) How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. British Medical Journal 288: 1409-1411.

[15] Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, et al. (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97: 1837-1847.

[16] Simons LA, Simons J, Friedlander Y, McCallum J, Palaniappan L (2003) Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia 178: 113-116.

[17] Shahwan-Akl L (2010) Cardiovascular disease risk factors among adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing 1: 1-7.

[18] Dr.Neeraj Bhargava, Aakanksha Jain, Abhishek Kumar, Dac-Nhuong Le, (2017) pages 35 – Detection of Malicious Executables Using Rule-Based Classification Algorithms DOI: http://dx.doi.org/10.15439/978-83-949419-2-5