International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT2063109

Study to Determine Adverse Diseases Pattern using Rare Association Rule Mining

Keerti Shrivastava[1], Varsha Jotwani[2]
[1] Assistant Professor, ITM University, Gwalior, Madhya Pradesh, India
[2] Associate Professor, Ravindranath Tagore University, Bhopal, Madhya Pradesh, India

ABSTRACT

Data mining is a method for finding patterns from repositories that remain hidden, unknown but fascinating. It has resulted in a number of strategies and emphasizes the detection of patterns to identify patterns that occur frequently, seldom and rarely. With their implementations, the work has improved the efficiency of the techniques. Yet typical methods for data mining are limited to databases with static behavior. The first move was to investigate similarities between the common objects through association rules mining. The original motivation for the search for these guidelines was the consumers ' shopping patterns in transaction data for supermarkets. This attempts to classify combinations of items or items that influence the presence likelihood of other items or items in a transaction. The request for rare association rule mining has improved in current years. The identification of unusual data patterns is critical, including medical, financial, or security applications. This survey seeks to give an analysis of rare pattern mining strategies, which in general, comprehensive and constructed. We discuss the issues in the quest for unusual rules using conventional association principles. Because mining rules for rare associations are not well known, special foundations still need to be set up.
Keywords : Data Mining, Association Rule Mining, Rare Patterns, Adverse Disease.

I. INTRODUCTION

The main objective of association rule mining (ARM) is to establish relations between a number of objects in a transaction database. Agrawal et al. (1993) developed the association Rule mining with the aim of extracting interesting similarities, repeated trends, associations or unexpected structures between transaction database sets of items or other data repositories. Such a relationship is depended on the co-occurrence of the objects in the database rather than on the inherent properties of data themselves. Such connections between objects are also referred to as association rules. In a database, there are two main types of rules: regular and unusual. The organization rules often and seldom include dissimilar information around the database. Frequent rules concentrate on often occurring patterns, while rare comments focus on seldom occurring patterns. In numerous areas, events often occurring may be less significant than occurrences that seldom occur, given the fact that regular patterns reflect the known and the predicted, whereas uncommon patterns may be unpredictable or unfamiliar to domain experts. Types of mining rare items provide detection of relatively rare diseases, the prediction of collapse of telecommunication equipment and comparisons amongst supermarket goods which are rarely purchased. The predicted regular responses to drugs in the field of medicine are less important than unusual, uncommon reactions suggesting adverse reactions or interactions with drugs. Assume that a patient symptom database includes a rare set of increased heart rate, fever, skin contusion, low blood compression in which all

conditions other than low blood pressure are normal [1]. Today many research is using data mining techniques for knowledge discovery from massive data Data mining has many functions like classification, association, clustering and prediction. In this paper, we implemented an association rule-based algorithm to extract knowledge from medical repositories for predicting the diseases. Appropriate and automated decision-making system can help in achieving less death rate. Hence the medical data set is used in this work.

An association rules is referred to as true uncommon rule if it supports a minimum supportive supply denoted in mines, i.e.supp(r), and its trust exceeds the minimum trust denoted in minconf, i.e.conf(r)= minconf. Uncommon AR is typically necessary to meet at the same time minimum user support and minimum confidence.

For many applications, like medicine & biology, unusual rules are very important[2]. But the main problem with both generations of rare rules is, even producing uninteresting rules a single mineral value will not classify all rarity. Many minuscule values are being used to overcome this problem[2] but they still have the same problem and some unusual laws have been discarded. This led us to develop a technique for the collection of all rare rules without creating interesting rules for rare associations.

Association Rule

AR is a rule-based method of machine learning to discover important relationships in databases of variables. Association rules are principles that help reveal associations in the database, connection database or another knowledge repository between unrelated data. ARs are applied to identify connections amongst objects often applied in conjunction. Besides the above examples, association rules are commonly used in many fields of practice, including.

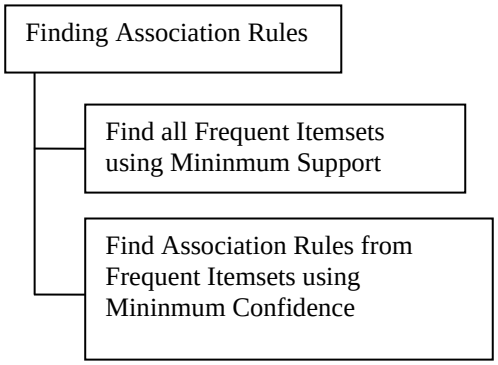


Figure 1: Generating Association Rules [3]

Cloud mining, identification of intrusions, continuous manufacturing, and bioinformatics. In comparison to sequence mining, association rule learning does not normally consider the order of things within or through transactions. The organization rules, support or trust have two basic requirements. The ties & rules made by analyzing data for commonly applied patterns are defined.

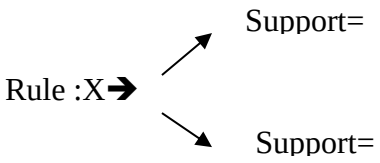
i. Support

Support is an indicator of how often items in databank appear. Support is characterized as a percentage of transactions containing items in the database. The objects are supported in example database because itemset occurs in 20 percent of all transactions (1 of 5 transactions). The statement is a collection of pre-conditions so that it evolves (rather than more inclusive) more restrictively.[4]

ii. Confidence

The confidence value of a rule is the share of transactions comprising confidence which contains confidence. For instance: a rule has confidence in the database, which ensures that for 100 percent of transactions involving butter and bread, the rule is codetermined. Trust is the measure of confidence of the rule in relation to a collection of transactions').

Remember that this implies the union support for X and Y products. We usually think of the probability of events and not other things. This is a bit confusing. The probability of an occurrence involving an element set is to be rewritten as well.[4]Thus trust can be seen as an estimation of the conditional probability of having the RHS of rules in transactions given such transactions also include the LHS. [5]



Rare association rule mining

In the last few years, rare ARM (RARM) has been very successful. The research field in the field of association regulating mining is extremely demanding. This recognizes relationships that have little funding but are very trustworthy. Applications that use these seldom-specific connections are fraudulent use of credit cards, network failure identification, education data or medical diagnostics. It's like discovering precious gems at open fields to discover unique connections. The different mining algorithms of the Rare Association

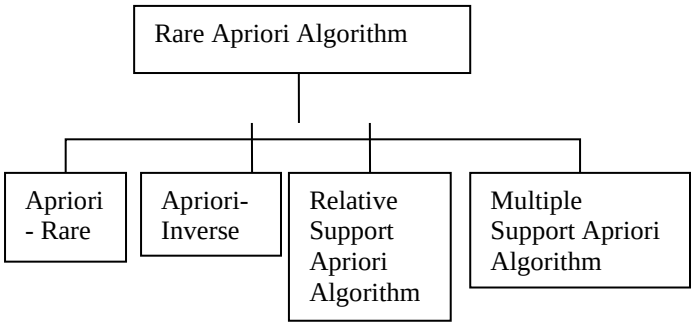


Fig 1 : Rare Association Rule Mining Algorithms

A. Apriori-Inverse

a) This uses the idea of full support to produce candidate products instead of minimal support.

b) Candidate places of interest are below a median support but below an absolute minimum support value.

c) Rule X if the above-mentioned laws are considered regular rules, sup(X) < maxsup and sup(x) > minabssup d).

d) Apriori-Inverse creates uncommon laws that take no objects beyond maxsup into consideration.

e) The Apriori-Inverse has all entirely inconsistent regulations. Rules that fall below maximum user support level but above minimum trust level that is determined by users are referred to as sporadical rules.

B. Apriori-Rare

a) The key purpose of this algo is to generate both common and uncommon elements.

b) The Apriori algorithm is modified to produce a special minimum item collection. b).

c) A subset called support account is used to find the support count of a certain element set.

d) Ri = Rare items (Supportcount Minsup)

C. RSAA algorithms (Relative Support Apriori Algorithm)

a) A) This algorithm uses relative help. Relative support (RSup) is distinct as any data set and supported by the element I represented as sup(i):

$$RSup(i1,i2,i3,\ldots i \quad \frac{Sup(i1,i2,i3,\ldots}{Min(Sup(i1),Sup}$$

b) This algorithm raises the support threshold for too low-frequency articles but reduces the support threshold for high-frequency articles.

D. Multiple supports Apriori (MsApriori)

a) The MIS can be sponsored for each object in the database).

b) The itemset is considered a regular array if actual support of item set exceeds minimum MIS values of items in collection.. [6]

Adverse Diseases

In recent times, medical practitioners have received considerable attention from different life-threatening illnesses, cardiovascular diseases, cancers or tumors or digestive diseases. The global mortality rate of 17.3 million people per year due to these hazardous conditions is expected to grow to further than 23.6 million by 2030.

Medics have received substantial attention in recent times due to various life-threatening illnesses, cardiovascular diseases, cancers or tumors or digestive diseases. The worldwide mortality rate of 17.3 million people a year is expected to increase to over 23.6 million by 2030, due to these hazardous conditions.

In multiple applications of the medical domain computer intelligence technology, particularly disease diagnosis, is widely used. A sub-section addresses the information technology for risk analysis or detection of diseases available in the literature.

A. Cardiovascular Diseases

Cardiovascular Disease (CVD) is a disease that affects people around the world's hearts and/or blood vessels. This is the main cause of death or inefficiency particularly in us in numerous European countries. Initial signs of CVD were identified by a soldier who suffered an injury in a Vietnamese war. The official estimated that approximately 90% of the wounded soldiers in Vietnam were found in the arteries. It is interesting that all these soldiers were below the age of 20.[7]

B. Breast Cancer

“The general statement of malignant tumor occurs in cells of breast & travels into surrounding taxis. Breast cancer affects one woman in 8 lifetime. "Breast cancer The disorder typically is apparent in women, but also in men. Figures from the United States indicate that there were only 40 thousand deaths in one year in the United States caused by breast cancer. [8]

C. Hepatitis

The virus is responsible for acute and chronic hepatitis B infection and one of the World Health Organisation's most serious human health issues and kills thousands per year. Different viruses of the Hepadnaviridae community infect non-human primates, other animals, and certain birds. Although most non-human primate virus isolates are phylogenetically alike to human hepatitis B virus, origins of those are uncertain, like human genotypes. The human hepatitis B virus may have originated from primates. Nevertheless, it is possible. To reduce human risk, it is crucial to know if such viruses are normal to humans or primates.[9].

II. Literature Survey

El-Houssainy A. Rady and Ayman S. Anwar [2019] In this sense, the use of effective data mining techniques exposes and removes secret data from clinical & patient laboratory data, can be useful to help doctors improve accuracy of diagnosis stage of disease. Results have been compared for the application of probabilistic neural networks (PNN), multilayer perceptron (MLP), SVM, & radial base function (RTF) algorithms. PNN algos provide improved classification & prediction efficiency to assess the extent of chronic renal disease & results have been compared. [10].

Ilkim Ecem Emre et al. [2018] It is designed by approaching the problem from a different point of view to help develop new solutions. The purpose of this analysis is to examine the impact of ARF experiencing in infancy using data mining methods

on the basis of cardiac diseases. DM classification techniques were used and five algorithms were tested. Pathways with a naïve Bayes classifier, decision trees (CART, C4.5, C5.0, C5.0 enhanced,) & random algos were evaluated on archives of patients diagnosed by ARF. The results of the derived algorithms were compared. The findings from various algorithms were compared to the use-value evaluation parameters of the model. best results were shown to use the hold-out approach (80% preparation, 20% testing) rendering to the CART model [11].

Anindita Borah and Bhabesh Nath [2018] proposed an effective ID framework founded on SVM integration by aspect enhancement. In particular, logarithm marginal density ratio transformation is applied to new characteristics by the aim of finding new & high-quality transformed training information; SVM sync was utilized to construct an ID model. Experimental outcomes suggest that our suggested method may perform well & has a great competitive advantage related to additional existing techniques in terms of detection rate, accuracy, training speed, & false alarm rate [12].

Jeong-woo Kim et al. [2017] Propose a tool for the detection of genes associated with diseases with MeSH terminology and association rules. Through evaluating Me SH words we have identified genes or collected data based on association rules on gene-gene interactions. We generated gene-gene networks & recognized genes for disease by combining the extracted interactions. We employed five cancers, including the prostate, lungs, breast, stomach & colorectal cance, as suggested, This has shown that the suggested approach is more effective than previously published methods for the detection of genes linked to disease & candidates for disease. We identified twenty genes per disease in this analysis. We presented among them 34 major candidate genes that help the relationship between candidate genes and diseases [13].

Aicha Boutorh and Ahmed Guessoum [2016] new hybrid intelligent methodology for addressing dimensional, based on the ARM & Neural Networks (NN), is proposed. On the one hand, Grammar Evolution (GE) optimized ARM is applied to pick most educational characteristics and to minimize dimensioning by parallel elimination of associations across SNPs in 2 separate case & control sample datasets. NN is used for effective classification to complement the previous mission. For setting parameters of 2 combined technologies, a Genetic algorithm (GA) is applied. Suggested GA-NN-GEARM method applied in the NCB I Gene Expression Omnibus (GEO) website to four separate SNP sets of datasets. In some cases, the generated model has achieved a high classification accuracy, exceeding 100%, and conducted multiple selection techniques in conjunction with different classifiers. [14].

Vladimir Ivancevi ˇ c´ et al. [2015] Caries of the initial childhood (ECC) is a potentially serious childhood disease worldwide. The obtainable results are typically centered on a logistic regression model, nonetheless, DM can be applied to extract extra data after the same data set, especially for ARM. Methods: ECC data were obtained in the 10 percent South Backa-to-area (Vojvodina, Serbia) cross-sectional study of pre-school children. Association rules by association rules have been derived after data. Highly ranked association laws have identified risk factors. Results: Male gender, often breastfeeding, a high order of birth, language or low body weight at birth are all dominant risk factors that have been found. Only male children were significantly affected by the ECC's low health awareness by parents. Conclusions: many kinds of literature that support the validity of methods confirm risk factors discovered [15].

U. Y. Bhatt and P. A. Patel [2015] Proposed a method based on Maximum restriction to produce the Rare Tree Structure Association Law. Temporary results

show which MCRP-Tree takes less time than the current algorithm for law generation, and identifies more interesting rarity objects [16].

A. Soltani and M. Akbarzadeh [2014] An important cogency-inspired measure is proposed to use a recent confabulation-inspired ARM (CARM) algorithm. Cogency is only determined on the basis of the likelihood of parallel items and so the proposed association rules for algorithm mines are just one file transfer. Thanks to its cogent approach, the proposed algorithm is also more effective in managing rare items. In order to assess the proposed algo, the problem of associative classification has been used. CARMs from the Irvine machine learning repository is tested through both simulated and real benchmark data sets. Experiment shows which, due to unique file access, proposed algos is consistently faster than the conditional regular pattern growth algorithm but requires less memory space. Statistical analysis also demonstrates the superiority of the method in unbalanced data sets for classifying minority groups [17].

S. Song et al. [2014] Proposes a hybrid clustering-ARM method for recognizing the population risk trend for an adverse event associated with chronic disease. Classification Association Rules (CARs) for cardiovascular (CVD) production are advanced from training data and are grouped on the basis of common case status that follows the precedent rule. Test cases are then allocated to rule clusters for risk groups with common CVD risk factors. Method is shown with a sample of data collected from the Biological Specific and Data Repository Information Coordination Center (BioLINCC) of American National Heart, Lung or Blood Institute. [18].

Jesmin Nahar et al. [2013] Examines the healthy and ill factors contributing to male and female heart diseases. Such variables will be defined using Association Rule Mining, a calculation intelligence technique, and the biological database UCI Cleveland, together with the three algorithms for the ruling class, Apriori, Predictive Apriori, and Tertio, is considered. By evaluating the available information on safe and ill people, women are seen to be less likely to have heart disease than men and take trust as a predictor. The healthy and sick conditions characteristics were also recognized. Factors including asymptomatic chest pain and the incidence of angina caused by exercise indicate that both men and women have the ability to have cardiovascular disease. Nevertheless, resting ECG as standard or hyper & slope as flat are only possible high-risk factors for women. single rule that expresses ECG hyper was exposed to be an important factor for men. For women, ECG rest is an important distinctive factor for predicting cardiovascular disease. The no. of vessels with colors o's & an olden increase of below or equivalent to 0.56 suggest healthy status for both genders as compared with the health status of men & women [19].

H. D. Vargas Cardona et al. [2012] Presents a NEUROZONE software system containing the two principal applications: firstly, the study of microelectrode recording (MER) signals facilitates the on-line identification of brain structure; and secondly, proceedings and analyzes of off-line databases permitting for incorporation of newly qualified automated recognition classifiers. The software supports the analysis done during a procedure by a physician and aims to decrease the potential adverse side effects of insufficient target identification. The software can also help specialists to mark surgical records to form an off-line database or to boost record numbers in the current off-line database. NEUROZONE was verified on the Subthalamic Nucleus (STNs) positive identification by up to 85 percent using a Bayes naive classification at Eje Cafetero's Institute for Epilepsy and Parkinson (Colombia). [20].

III. Conclusion

The analysis of large no. of data is carried out using association techniques. It is most commonly used DM techniques in healthcare. ARM provides an automated way to find new disease information. This work aims to reduce the search space so that laws are exponentially exploded while information loss is reduced. They explain briefly the laws of association and the rare association algorithm, different types of adverse diseases. There are various algorithms for association rules. Certain of ARMs Rules algorithms mentioned in this document have been discussed in this paper. The law of rarity applies to the rule of association between regular and rare elements or between rare elements. Existing work on the use of uncommon association rules is based on the notion that at the start of the mining process the entire data is available. In order to properly diagnose and heal, disease detection is required at an early stage. Heart disease, cancer & hepatitis diagnostic systems are costly, time-consuming & bug-prone. More than qualified data contained in databases, the clinical decisions regarding the diagnosis of illness primarily are rooted in the expertise and experience of medical experts. This could lead to unnecessary mistakes resulting in undue medical expenses that have an effect on patients ' quality of service. Health care systems produce large volumes of hidden information that can not be identified using traditional methods. Hence, to tackle these problems we will try to work on those issues.

IV. REFERENCES

[1]. Y. Koh and S.D. Ravana, "Unsupervised Rare Pattern Mining: A Survey", ACM Trans. Knowl. Discov. Data. 2016, pp. 1-31.

[2]. R. U. Kiran and P. K. Reddy. Mining rare association rules in the datasets with widely varying items' frequencies. The 15th International Conference on Database Systems for Advanced Applications Tsukuba, Japan, April 1-4, 2010.

[3]. IshNathJhaSamarjeet Borah, An Analysis on Association Rule Mining Techniques, International Conference on Computing, Communication and Sensor Network (CCSN) 2012

[4]. ManishaGirotra, KanikaNagpalSaloniinochaNeha Sharma Comparative Survey on Association Rule Mining Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 84 – No 10, December 2013

[5]. Sotiris Kotsiantis, DimitrisKanellopoulos, AssociationRules Mining: A Recent Overview, GESTS International Transactions on ComputerScience and Engineering, Vol.32 (1), 2006, pp. 71-82

[6]. Anjana.k1, Dr. Jithendranath Mungara, "Review on Various Rare Association Rule Mining Algorithms in Big Data", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 6 Issue IV, April 2018- Available at www.ijraset.com

[7]. Mohammad-Hossein Biglu, 1 Mostafa Ghavami, 2,* and Sahar Biglu, "Cardiovascular diseases in the mirror of science", J Cardiovasc Thorac Res. 2016; 8(4): 158–163.

[8]. Mussarat Yasmin, Muhammad Sharif, Sajjad Mohsin, "Survey Paper on Diagnosis of Breast Cancer Using Image Processing Techniques", Research Journal of Recent Sciences, Vol. 2(10), 88-98, October (2013).

[9]. Souza bf1, Drexler jf2, lima rs3, rosário mde o1, Netto em4, "theories about evolutionary origins of human hepatitis b virus in primates and humans", Braz J Infect Dis. 2014 Sep-Oct;18(5):535-43.

[10]. Rady, E.-H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data

mining algorithms. Informatics in Medicine Unlocked, 100178.

[11]. Emre, İ. E., Erol, N., Ayhan, Y. İ., Özkan, Y., & Erol, Ç. (2018). The Analysis of the Effects of Acute Rheumatic Fever in Childhood on Cardiac Disease With Data Mining. International Journal of Medical Informatics, Volume 123, March 2019, Pages 68-75.

[12]. Borah, A., & Nath, B. (2018). Identifying risk factors for adverse diseases using dynamic rare association rule mining. Expert Systems with Applications, 113, 233–263.

[13]. Kim, J., Bang, C., Hwang, H., Kim, D., Park, C., & Park, S. (2017). IMA: Identifying disease-related genes using MeSH terms and association rules. Journal of Biomedical Informatics, 76, 110–123.

[14]. Boutorh, A., & Guessoum, A. (2016). Complex diseases SNP selection and classification by hybrid Association Rule Mining and Artificial Neural Network-based Evolutionary Algorithms. Engineering Applications of Artificial Intelligence, 51, 58–70.

[15]. Ivančević, V., Tušek, I., Tušek, J., Knežević, M., Elheshk, S., & Luković, I. (2015). Using association rule mining to identify risk factors for early childhood caries. Computer Methods and Programs in Biomedicine, 122(2), 175–181.

[16]. U. Y. Bhatt and P. A. Patel, "An effective approach to mine rare items using Maximum Constraint," 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2015, pp. 1-6.

[17]. A. Soltani and M. -. Akbarzadeh-T., "Confabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets," in IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 11, pp. 2053-2064, Nov. 2014.

[18]. S. Song, J. Warren, and P. Riddle, "Profiling Cardiovascular Disease Event Risk through Clustering of Classification Association Rules," 2014 IEEE 27th International Symposium on Computer-Based Medical Systems, New York, NY, 2014, pp. 294-299.

[19]. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y.-P. P. (2013). Association rule mining to detect factors that contribute to heart disease in males and females. Expert Systems with Applications, 40(4), 1086–1093.

[20]. H. D. Vargas Cardona et al., "NEUROZONE: On-line recognition of brain structures in stereotactic surgery - application to Parkinson's disease," 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, 2012, pp. 2219-2222

Cite this article as :

Keerti Shrivastava, Varsha Jotwani, "Study to Determine Adverse Diseases Pattern using Rare Association Rule Mining ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.519-526, May-June-2020. Available at doi : https://doi.org/10.32628/CSEIT2063109
Journal URL : http://ijsrcseit.com/CSEIT2063111