International Journal of Applied Research

ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 8.4
IJAR 2021; 7(5): 166-169
www.allresearchjournal.com
Received: 04-03-2021
Accepted: 17-04-2021

Reena Hooda
Assistant Professor, CSE,
Indira Gandhi University
Meerpur, Rewari, Haryana - India

Corresponding Author:
Reena Hooda
Assistant Professor, CSE,
Indira Gandhi University
Meerpur, Rewari, Haryana - India

Applicability of association rules in finding correlations among covid-19 patient data and importance of fuzzy set theory in predictions

Reena Hooda

Abstract

The current paper highlight the applicability of association rules in predicting the relation between the different symptoms of Corona and this disease and also the closeness of these symptoms with the reports showing death or cure by the confidence measure and certainty measures. The paper explains the association rule measures like support, confidence, and lift as well as certainty factor to predict the associations among the factors or the attributes of the disease by taking the 3 random rules. The measures successfully predict the ratio between the different symptoms that whether they are dependent to each other, independent or the substitute. The database is a dummy database of patients containing some common symptoms taken from the official site of Centers for Disease Control and Prevention. The paper further emphasizes the importance of set theory and fuzzyfication of the data to get more realistic inferences.

Keywords: Fuzzy set, support, certainty factor, confidence, rule

1. Introduction

Data mining as the name defined is a useful instrument to take out the hidden and useful information that benefits the analyzer as well as the user of this information. The main goal of the data mining is to discover the new and useful patterns in the real world datasets [2] the discovery of the information and the inferences must be meaningful so that can be disseminated in research or enhancing the sale in business. Mining makes the gathered data in a systematic way sometimes as a summary data, visualization data or in different views or sights as per the user requirements. Association rule mining is the applicability of association rules to mine the massive transactions in order to find the frequent itemsets, gaining the knowledge about patterns choices etc. [1] Association mining works on non-numeric data to infer the meaningful rules [1]. The following are some applications of Association rules:

1. Market Basket Analysis to check the frequent items and buying patterns in order to maximize the sale [10]. The promotion design, layouts in store, cross-selling is benefited with such type of pattern analysis by putting the matching items near to each other or by putting the items with most popular items [10, 13].
2. It can be also used to find out the behavior of a person, likings and categorize those, for instance, if a person likely to watch the movie, he also likely to purchase the corns [11]. Thus interesting patterns and buying habits can be discovered easily [8].
3. It can be further beneficial to correlate the web pages to a person that he generally visited the sites and pages so what advertisements can be displayed on those pages [4]. Online shopping for instance, a catalog can be designed based on such information [8].
4. Mining the data to find the hidden information, association rule mining is best suited for non-numeric data analysis through the generation of rules. The rule is also enhanced by many researchers for non-numeric data too. The items in databases are put as attributes or the dimensions and a transaction shows the record for those elements in a row [12].

2. Association rule mining on patient database

In the basic Association rule mining database is created for COVID 19 patients based on the symptoms given [14] and putting Boolean values 0 and 1 in the rows.

1 value symptom is there otherwise 0 indicating the absence of the symptoms. Symptoms are the attributes or the element of the database and rows indicating the patient record given in table 1:

Table 1: Shows dummy database COVID-19 patients with symptoms [14] and reports.

Patient ID	Confusion	Cough	Fatigue	Difficulty breathing	Fever	Body aches	Smell/ taste loss	Throat	Vomit	Chest pain	Skin Blue/ Pale Spots	Not able to stay awake	Corona - 1(yes) 0(No)	Reports Cured-1, death -0
P1	1	1	1	1	1	1	0	0	0	0	0	0	0	1
P2	0	1	0	1	1	0	1	1	0	0	0	0	1	1
P3	0	0	0	0	0	1	0	1	1	1	0	1	0	1
P4	1	1	0	0	1	0	1	1	0	0	1	1	1	0
P5	1	1	0	0	0	0	1	0	0	1	1	1	1	0
P6	1	0	0	1	0	0	1	0	1	0	1	0	1	1

Table 2: Shows the Age vector given separately to reduce the size

Patient ID	Age
P1	28
P2	40
P3	66
P4	55
P5	35
P6	67

The support & confidence, lift are counted on the rigid formula based on the set theory like union, set subset, proper subset, tree and ratios for X =>Y.

2.1 Applications of set theory

The rules and data items are described in the form of sets in association rule mining. Moreover the support confidence and lift etc. methods are also based on the set theory. Following is the representation of the set notations used in association rule mining:

1. Capital letter is the name of the set.
2. Small letter say "i" is the constant used for iteration
3. A=set of items contained as attributes/properties/elements/columns in patient database.
4. P is a set of patients P = { P1, P2, P3, P4, P5, P6}
5. Each row of patient database represents a record having some attribute. That means attributes are the properties [1].
6. Set A contains all the column names of the database: A= {Confusion, Cough, Fatigue, Difficulty breathing, Fever, Body aches, Smell/ taste loss, Throat, Vomit, Chest pain, Skin Blue/ Pale Spots, Not able to stay awake, Corona -1(yes) 0(No), Reports Cured-1, death -0} [14].
7. Each P contains some attributes represented by Boolean values in data base for instance:
   P1= {Confusion, Cough, Fatigue, Difficulty breathing, Fever, Body aches, Smell/ taste loss, Reports}
   i.e. P1= {I1, I2, I3, I4, I5, I6, I7, I8} where I is the itemset (patientset), and count can be represented by "i".
   Where I ⊆ A that means "I" is the subset of A, "I" belongs to A ot "I" contained in A or A is the domain of "I". if "I"<A and all elements of "I" are the subset of A that means "I" is the proper subset of A such that I ⊂ A.
8. XUY contains the Union symbol that means all the common elements of X & Y and all the elements of X only and all the element of Y only or in short X or Y. X∩Y means only the common items of X and Y.
9. If any attribute (Symptoms) is present in the patient, that is represented as 1 otherwise represented as 0.
10. X and Y are the itemsets (patientset) for each patient and used to define the rules for mining. Here the rule is for example X=> Y where the X and Y are the set of attributes (Symptoms shortly); X patientset is called antecedent (Left Hand side) and Y patientset is called the consequent of X (on Right Hand Side)

3. Computing association rule measures and discussions

The important measure support, confidence, lifts are counted on the COVID -19 patient databases given in table 1. Where the value 1 of an attribute indicates the presence of attribute and 0 is the absent, reports column also shows the patient died or cured where the cured represented by 1 and died represented by 0 should keep in mind while counting the different measures for the rule containing the attribute "reports". There are 3 rules are selected randomly and the measures support confidence and lift are computed to check whether the rule is realistic, whether the symptoms are correlated or independent & factor for other diseases than corona.

Support (X) = $\frac{Frequency\ of\ X\ i.e.Count}{Total\ No.of\ Transacctions\ (T)}$ [12].

Confidence (X=>Y) = $\frac{Support\ of\ X\ and\ Y\ i.e.Support(XUY)}{Support\ (X)}$ [9].

Lift= $\frac{Support(XUY)}{Support\ (X) \times Support\ (Y)}$ [9].

Conviction (Y=>X) = $\frac{1-Support(Y)}{1-Confidence\ (X=>Y)}$ [9].

Let's consider 3 rules

1. X= {Cough, Difficulty breathing, Fever} where the X Count is 2/6; Y= {Corona} where Y count is 4/6; from database of COVID-19 patients in table1.
2. X= {Confusion, Smell/ taste loss, Skin Blue/ Pale Spots} where the X Count is 3/6; Y= {Reports} where Y count is 4/6; from database of COVID-19 patients in table1.
3. X= {Confusion, Smell/ taste loss, Skin Blue/ Pale Spots} where the X Count is 3/6; Y= {Not able to stay awake, Corona} where Y count is 2/6; from database of COVID-19 patients in table1.

Rule 1

Support(X)= $\frac{Frequency\ of\ X\ i.e.Count}{Total\ No.of\ Transacctions\ (T)}$ [12]= $\frac{2}{6}$ =.3333 or 33.33%

Support (Y) = $\frac{Y\ Count}{Total\ No.of\ Transacctions\ (T)}$ [12]= $\frac{4}{6}$ =.6666 or 66.66%

Confidence (X=>Y) = $\frac{Support\ of\ X\ and\ Y\ i.e.Support(XUY)}{Support\ (X)}$ [9]

$Support(XUY)$= $\frac{Frequency\ of\ X\ and\ Y\ i.e.(\text{Cough,Difficulty breathing,Fever,Corona})}{Total\ No.of\ Transacctions\ (T)}$
= $\frac{1}{6}$ =.1666 or 16.66%

So, Confidence (X=>Y) = $\frac{.1666}{.3333}$ =.4998 =49.98% or 50% approx.

That indicates that 50% of the cases if these symptoms arise will be resulted in Corona.

Lift= $\frac{Support(XUY)}{Support\ (X) \times Support\ (Y)}$ [9] =$\frac{.1666}{.3333\times.6666}$
=$\frac{.1666}{.3333\times.6666}$ =$\frac{.1666}{.2222}$ =.7497=74.97% means attributes appears less than expected as it is <1.

Rule 2

Support (X) = $\frac{Frequency\ of\ X\ i.e.Count}{Total\ No.of\ Transacctions\ (T)}$ [12]= $\frac{3}{6}$ =.5 or 50%

Support (Y) = $\frac{Y\ Count}{Total\ No.of\ Transacctions\ (T)}$ [12]= $\frac{4}{6}$ =.6666 or 66.66%

Confidence (X=>Y) = $\frac{Support\ of\ X\ and\ Y\ i.e.Support(XUY)}{Support\ (X)}$ [9]

$Support(XUY)$= $\frac{Frequency\ of\ X\ and\ Y\ i.e.(\text{Confusion,Smell/ taste loss,Skin Blue/ Pale Spots,Reports})}{Total\ No.of\ Transacctions\ (T)}$
= $\frac{1}{6}$ =.1666 or 16.66%

So, Confidence (X=>Y) = $\frac{.1666}{.5}$ =.3332 =33.32% approx.
That indicates that 33.32% of the cases if these symptoms arise will be resulted in cured and no death in 33.3% only (very few cases are cured) that means these symptoms are critical and may cause death (1 is cure and death is represented by 0 in table 1 and we have to count 1 for the XUY i.e.Confusion, Smell/ taste loss, Skin Blue/ Pale Spots, Reports).
Lift= $\frac{Support(XUY)}{Support\ (X) \times Support\ (Y)}$ [9] = $\frac{.1666}{.5\times.6666}$
= $\frac{.1666}{.3333}$ =.4998= 50% approx.; means attributes appears less than expected as it is <1. The X and Y are negatively correlated. Means more symptoms less cured.

Rule 3

Support (X) = $\frac{Frequency\ of\ X\ i.e.Count}{Total\ No.of\ Transacctions\ (T)}$ [12]= $\frac{3}{6}$ =.5 or 50% Approx.

Support (Y) = $\frac{Y\ Count}{Total\ No.of\ Transacctions\ (T)}$ [12]= $\frac{2}{6}$ =.3333 or 33.33%
Confidence (X=>Y) = $\frac{Support\ of\ X\ and\ Y\ i.e.Support(XUY)}{Support\ (X)}$ [9]

Support(XUY)= $\frac{\text{Frequency of X and Y i.e.(Confusion,Smell/ taste loss,Skin Blue/ Pale Spots,Not able to stay awake,Corona)}}{\text{Total No.of Transacctions (T)}}$ = $\frac{2}{6}$ =.3333 or 33.33%

So, Confidence (X=>Y) = $\frac{.3333}{.5}$ =.6666 =66.66%=67% approx.
That indicates that 67% approx. of the cases if these symptoms arise will be resulted in corona and patient will not be able to stand.

Lift = $\frac{Support(XUY)}{Support\ (X) \times Support\ (Y)}$ [9] =$\frac{.3333}{.5\times.3333}$= $\frac{.1666}{.1666}$ = 1; means attributes X and Y are independent.

3.1 Certainty factor for the rules

Another important measure of association mining is the certainty factor [2] that increases the belief about the accuracy of the rules selected. 2 ways to count the certainty factor:

1. If the Confidence(X=>Y)>Support(Y). For example in rule 3

Then Certainty factor = $\frac{\text{Confidence (X=>Y)}-Support\ (Y)}{1-Support\ (Y)}$ [2]= $\frac{.6666-.3333}{1-.3333}$ =$\frac{.3333}{.6667}$

=.4999= 49.99%=50% Approx. (positive)

2. And if Confidence(X=>Y) ≤ Support(Y); For example rule 2:

Then Certainty factor = $\frac{\text{Confidence (X=>Y)}-Support\ (Y)}{Support\ (Y)}$ [2]= $\frac{.3332-.6666}{.6666}$

=$\frac{-0.3334}{.6666}$ =-0.50 = -50% Approx. (negative)

If the certainty factor is positive that means the elements are dependent to each other as in rule 3 in which confidence is also > 50%, if the certainty factor is negative (as in rule 2 where the confidence is also too small) then the elements are negatively correlated or act as a substitute, if the certainty factor is 0 then the elements are independent [2], this factor gives more accurate results.

3.2 Issues in Boolean-valued Data

1. In efficient in case of huge amount of data, for example in Table 2, age attribute cannot be defined in Boolean as it is not atomic rather it is a vector that contains a range of values like 1-20,21-35, 36-45, 46-60, 61-72, above 72, it needs a range or the category so taken separately in Table 2.
2. The minimum support set by the user may be inappropriate, so affect the generation of the rules and can give wrong prediction about some rules and frequency patterns of the items or the attributes [1].
3. Support percentage and minimum support to find the frequent itemsets (patientset) is based on yes or not in

range. This rigidness sometimes produced unnatural results and barred the interestingness of the rules

3.3 Applicability of fuzzy set theory

As compare to rigid crisp set, fuzzy set provides a range to become the member of a set.in place of Yes(1) or No (0),it gives a range of values [0,….,1] or shortly [0,1] and gives an option to be a member based on degree of membership given by membership function [1]. The fuzzy set concept introduced in association rule mining to make the association rules works on numeric data too [1]. The crisp boundaries first converted into the fuzzy boundaries with the use of membership function [1] For the computations apriori (breadth first search algorithm) becomes slow If the number of items or elements and transactions are very large in a database, so to make the fast processing dataset is partitioned into different clusters. Again the partitioning can be a fuzzy [1] means some of the elements of clusters may overlap, violating the basic definition of clustering i.e. discretization.

The basic steps in this case includes clustering (partitioning), fuzzy dataset and partition P= {P1, P2, ….,Pn} where "µ" is membership function and "i" is the constant. Attributes can be categorical attributes like male or female and quantitative attributes, categorical attributes are converted into the consecutive integers and the quantity attributes are converted into the discrete intervals using equi-depth partitioning [1].

The fuzzy set theory removes the gap between human reasoning and rigidness of the sets to represent the knowledge in more realistic way. It is better to incorporate fuzzy logic in data mining to infer optimal and realistic information [2]. In the table 1, the patient have either the symptoms present or absent, however it may be the symptoms are present a low degree or high degree, this important information is not present in the database shown in table1. So to get more realistic and accurate information, it is better to convert these values into fuzzy using the membership degree so that more realistic and accurate information can be used to get the realistic values of association rule measures like support, confidence, lift etc.

4. Conclusions

Association rules mining is the best approach to find the relations among the symptoms of the corona disease and find their dependencies, patterns and novelties through the key measures of association rules i.e. support, confidence, lift and certainty factor. Fuzzification is must as even in the database given in Table 1, some of the symptoms may overlapped and some of them are independent, some are highly associated with the disease and some of them not necessarily the true symptoms of corona. So better to use the degree of membership of each attribute in the set rather than just 0 or 1 and even in when the measures are counted specially against the 3 categories <1, >1 or =1 as for instance .998 is < 1 however it is near to 1 also, so to predict the results better, fuzzyfication is good. Such study can even help in early detection of the crucial symptoms and controlling the spread of the disease. Thus the idea can help the researchers and developers to analyze the data more accurately, predict the right inference rules and control the spread of disease. The future scope the work is to convert the Boolean values of patient data in to fuzzy data.

5. References

1. Anand V Saurkar, Gode SA. Association Rule Mining with Fuzzy Logic: an Overview International Journal of Science and Research (IJSR) 2015, 4(6). https://www.ijsr.net/archive/v4i6/SUB155265.pdf
2. Miguel Delgado, Nicolás Marín, Daniel Sánchez, María-Amparo Vila. Fuzzy Association Rules: General Model and Applications. Ieee Transactions On Fuzzy Systems 2003, 11(2). https://core.ac.uk/download/pdf/208378877.pdf
3. Lekha A, Srikrishna CV, Viji Vinod. Fuzzy Association Rule Mining. Journal of Computer Science 2015;11(1):71-74. https://thescipub.com/pdf/jcssp.2015.71.74.pdf
4. Régis Pierrard, Jean-Philippe Poli, Céline Hudelot. A Fuzzy Close Algorithm for Mining Fuzzy Association Rules. 2018. ffhal-01698352v2f. https://hal.archives-ouvertes.fr/hal-01698352/file/fuzzy-close-algorithm.pdf
5. Amal Moustafa, Badr Abuelnasr, Mohamed Said Abougabal. Efficient mining fuzzy association rules from ubiquitous data streams, Alexandria Engineering Journal 2015;54(2):163-174. ISSN 1110-0168, https://doi.org/10.1016/j.aej.2015.03.015. (https://www.sciencedirect.com/science/article/pii/S1110016815000290)
6. Michal Burda. Interest Measures for Fuzzy Association Rules Based on Expectations of Independence, Advances in Fuzzy Systems, vol. 2014, Article ID 197876, 7 pages, 2014. https://doi.org/10.1155/2014/197876, https://www.hindawi.com/journals/afs/2014/197876/
7. Aly El-Semary, Janica Edmonds, Jesús González-Pino, Mauricio Papa. Applying Data Mining of Fuzzy Applying Data Mining of Fuzzy Association Rules to Network Intrusion Association Rules to Network Intrusion Detection Detection. Center for Information Security Department of Computer Science University of Tulsa, Tulsa, OK 74104. http://book.itep.ru/depository/fuzzy/IAW2006-06-1.pdf
8. https://paginas.fe.up.pt/~ec/files_0506/slides/04_AssociationRules.pdf
9. Wikipedia contributors. (2021, April 14). Association rule learning. In Wikipedia, The Free Encyclopedia. Retrieved 03:30, April 23, 2021, from https://en.wikipedia.org/w/index.php?title=Association_rule_learning&oldid=1017733408
10. Geeksforgeeks 2018. Association Rule. Retrieved April 25, 2021, from https://www.geeksforgeeks.org/association-rule/
11. Surya Remanan (Nov., 2018). Association Rule Mining the data. Retrieved April 25, 2021, from https://towardsdatascience.com/association-rule-mining-be4122fc1793
12. https://www.javatpoint.com/association-rule-learning
13. Annalyn Ng 2016.Association Rules and the Apriori Algorithm: A Tutorial. https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html
14. Centers for Disease Control and Prevention (Feb 2021). Symptoms of COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html