IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.542

Predicting Relations for the State of Vaccine for Covid Using High Utility Itemset Mining

U Suvarna[1], Harika Gorantla[2], Pragathi Sai Kalluri[3], Sri Lakshmi Sneha Gaddam[4], Tejaswi Battu[5]

Associate Professor, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India[1]

U.G. Students, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India[2,3,4,5]

ABSTRACT: The present pandemic situation due to COVID-19 is giving challenges to the medical fraternity in treating the patients and finding the right vaccine. We have tried to attempt an algorithm that predicts which combinations of drugs can be used to covid affected patients, and for non covid people to get a suggested vaccine. This can be a relief to survive from this pandemic outbreak. This prediction is done with the help of a data mining technique called High Utility Itemset Mining (HUIM) where utility factors of all attributes are retrieved and their relations are predicted and extract high utility items that influence & using the existing data, we can predict the right vaccine for patients.HUIM is an advanced version of associate rule mining (ARM), Unlike Frequent itemset mining which has support and confidence measures, HUIM focuses on utility measures and not just the frequency patterns. The HUIM finds interesting patterns and relations that are useful to identify the cause and impact of that vaccine on the patient. Our covidCl_HUIM algorithm helps to first extract the important items that are affecting the patient, and next predicts the kind of vaccine suggested.

KEYWORDS: ARM, Support, Confidence, HUIM, Utility-measure, Internal utility, External Utility, Frequent-itemsets.

I.INTRODUCTION

Huge amounts of data were collected by retailers on a daily basis and are stored in the databases for years together and of no business interest. Now, Retailers are interested to analyze their data to learn about the behaviour pattern of their customers purchase. This gave life to new business intelligence –Data Mining. Such data which is stored can be used to encourage variety of applications related to business like market promotions, store management, work flow management and fraud detections. Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, and reduce risks and more. The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases," the term "data mining" wasn't coined until the 1990s. But its foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions). What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power. Over the last decade, advances in processing power and speed have enabled us to move beyond manual, tedious and time-consuming practices to quick, easy and automated data analysis. The more complex the data sets collected, the more potential there is to uncover relevant insights. Retailers, banks, manufacturers, telecommunications providers and insurers, among others, are using data mining to discover relationships among everything from price optimization, promotions and demographics to how the economy, risk, competition and social media are affecting their business models, revenues, operations and customer relationships. Data mining allows you to sift through all the chaotic and repetitive noise in your data. Understand what is relevant and then make good use of that information to assess likely outcomes. Accelerate the pace of making informed decisions. Association rule mining (ARM) is one concept of data mining which can extract this intelligence. Frequent itemset mining (FIM) is used to find frequent items of business interest from transactional database. FIM couldn't find the utility itemsets (profitable to biz) because FIM is limited by the intendment of the discovered itemsets and quantity is not considered (all items are viewed as having the same importance).High utility itemset mining (HUIM) introduced utility, a measure which is a weight /profit associated to each itemset. Each items quantity and profit per item must also be considered unlike FIM which only considers support.

Definitions:

1.1 Frequent itemsets :

To find frequent itemsets we need to know the itemset and its support measures.

A **Frequent item (FreqI)** , is the itemset has the specified
minimum support, and the percentage of transactions containing the itemset.

1.2 Itemset

An Itemset is a collection of one or more items that appear in a transaction database.
The main objective of FIM is to find frequent itemsets that appear in a transactional
database.

1.3 K-itemset

An itemset that contains k items in its itemset.
For e.g: {P} -- 1 itemset, {P, Q, R} -- 3
itemset, and so on. The Total no.of possible items is ,
For e.g: 3-itemset{P,Q,R} = { P , Q, R, PQ, PR, QR, PQR }

1.4 Frequent Itemset Mining

An itemset whose Support(S) is more than or equal to minimum support threshold.

1.5 Support_count()

The frequency of an occurrence of an itemset.

1.6 Support (Sup) %

The percentage of fraction of the transactions that contain an itemset.

2. HUIM (High Utility Item/Pattern Mining)

An itemset is said to be High utility itemset if it is frequent item and possess utlity factors.In Simple terms, it is an extension to Frequent item mining with Utility.

Anything that’s profitable, satisfactory to consumer experiences from a product or any service is called utility. It is a qualitative concept, and to quantify it to be simple, it’s nothing but the profit or weight associated with every itemset in the database.

Utility has two parts, internal utility & an external utility for every item, it can be defined as a product of External Utility (eu) and Internal Utility(iu),where eu is the cost of every item in transactions and iu is the no. of the items(quantity) in transactions.

2.1 Utility =iu * eu

For eg: If a transaction T1 is having PQRS items, then item P has its cost , 20Rs and also no.of P's purchased in that transaction are 5.. Likewise , for Q, R, S, etc..
Therefore Utility (P) = iu(P) * eu (P) = 20 * 5 = 100

2.2 Total Utility Tutil(x): The utility of an itemset in a database D is the sum of the utilities of X in all the transactions having X in D.
Tutil(x)= $\sum_{It \in X \wedge Tid \in D} \sum_{It \in X} util(It, Tid)$

2.3 Transaction-Weighted Utilization (TW_Util) measure, which provides an upper-bound on the utility of itemsets and is anti-monotonic property. The transaction-weighted utilization (TW_Util) of an itemset X is defined as the sum of the transaction.
utilities of transactions containing X, i.e. TW_Util(X) =￼

2.4 $\sum_{Tid} util(Tid)$ Minimal High Utility items (min_HUIs) : An item x is said to be min_HUI if util(x) min_util and there should not be an item Y which is subset of X and holds util(Y) min_util, where Y Ì X.

2.5 Pruning search space using util-lists: Let X be an itemset. Let the extensions of X be ExtX ,theitemsets formed by appending an item y to X such that y ∈ X. If the sum of iutil and eutil values in util(X) <min_util, where X and its extensions of X are low utility.

2.6 Classified Association rule[CA Rule]:   CA Rules are of form {[Ai, vi], …,[Aj, vj]}□C where C Class-Label {c_1, c_2,….., c_n}. Where antecedent is itemset and consequent is the result class.

2.7 Naïve bayes theorem

II.LITERATURE SURVEY

In data mining, we have different techniques and these are classification, clustering, association rule mining. Association rule mining is one of the most important techniques of data mining. It is used to discover the frequently occurring patterns in the database. The main aim of association rule mining is to find out the interesting relationships and correlations among the different items of the database[1] .The AIS algorithm was the first algorithm proposed by Agrawal, Imielinski, and Swami for mining association rule. It focuses on improving the quality of databases together with necessary functionality to process decision support queries [2]. The volume of data is increasing dramatically as the data generated by daily activities. Therefore, mining association rules from massive amount of data in the database is interested for many industries which can help in many business decision making processes. The techniques for discovering association rules from the data have traditionally focused on identifying relationships between items, which show some aspect of human behavior[3]. Association rules are used in many fields to find out the patterns in the data. With the help of patterns, we can find out how many combinations of events occur at the same time [1]. Association rules are if/then statements that help to uncover relationships between unrelated data in a database. Applications of association rules are basket data analysis, classification, cross-marketing, clustering, catalog design, and loss-leader analysis etc[2].Generally association rules are expressed in the form of X=>Y. Here X and Y are the itemsets in the database. X is called as an antecedent and Y is called as consequent. Association rule mining consists of two basic measures and these are: support (s) and confidence (c)[1].Support(s) is defined as the proportion of records that contain X U Y to the overall records in the database. The amount for each item is augmented by one, whenever the item is crossed over in different transaction in database during the course of the scanning.
support(XY) = support sum of XY / overall records in database

Confidence(c) is defined as the proportion of the number of transactions that contain X U Y to the overall records that contain X, where, if the ratio outperforms the threshold of confidence, an association rule X => Y can be generated.
confidence(X/Y) = support(XY) / support(X) [4]

The HUIM which used utility list to save the dynamic data and extract the high utility items , prunes the unimportant items and later uses these utility items for classification using Naïve bayes theorem [10]. The similar method is applied with different measures in our algorithm.

Dataset

A medical Covid dataset with the following attributes are taken for analaysis to perform the prediction as shown in Table 1.

S.No	**Test**	**Impression**
1	RTPCR	If RTPCR is positive, Covid +
2	SPO2	If SPO2 < 93% , not normal
3	Temp	If Temp > 100.5, not normal
4	DDimer	0.1-0.5 mg/L
5	CT severity Score	< 9, Mild 9 to 15 Moderate, >15 Severe
6	CORADS	1 – no [no infection] 2 – low [infection] 3 – intermediate [unclear covid +] 4 –High [suspicion to covid+] 5- Very high [Covid +] 6 -PCR +
7	ESR	Male 0-6 mm Femal 0-7mm
8	Hb1AC	Non-diabetic 4.8-6.0 Good control 6.1-7.0 Target therapy 7.1-8.0 Change therapy >8.0
9	BP	120 systolic /80 diastolic normal
10	CRP	6.0 mg/dl normal

Table 1.Dataset key attributes and its result

A CSV file with patient details looks like the Table 2., which is then given internal utility based on the impression of dataset and the profit table gives the extrenal utility values as shown in Table 3. If you observe the data, the vaccine is not suggested if RTPCR is positive and a time limit along with suggested vaccine is given “Vac_after_30days” which means vaccine to be given after 30 days and there is “Rpt_Vac_30”, if the patient has RTPCR is positive and other vitals are fine, basic medications (numbered in the list), for eg: 1-VitD3, 2- Paracetemol, 3- Limcee, 4-multivitamin, 11- remidisivier etc…

Here in this table 2, the 1st 7 rows are trained data and the 8th row is the test data.

Pid	**RTPCR**	**BP**	**Temp**	**SPo2**	**DDimer**	**ESR**	**CORADS**	**Severity_Score**	**Hb1ac**	**CRP**	**Dose**	**Time**	**Medications**	**Vaccine**
1	positive	140/20	99	95	0.3	25	5	10/25	5	9	0	Vac_after_30dys	1,4,7,9	Covishield
2	negative	120/80	100	94	0.6	36	3	7/25	9	11	2	Nil	2	No_vaccine
3	positive	135/73	100	98	0.4	80	2	6/25	11	70	1	Rpt_Vac_30days	1,2,4,8,9	covaxin
4	positive	132/91	100.5	91	0.6	50	5	13/25	8	11	0	Vac_after_30dys	2,3,5,7,8,10	Sputnik
5	negative	120/69	101	94	0.3	60	5	6/25	9	10	0	immediate	1,3,6,7	Covishiled
6	negative	140/95	100	95	0.3	23	Nil	Nil	11	6	0	immediate	1,3,7	covishield
7	negative	112/87	98	98	0.3	11	Nil	Nil	5	5	2	Nil	nil	No_vaccine
8	negative	120/80	98	96	0.6	70	3	7	10	56	3	Nil	Nil	**?**

Table2: sample dataset

The above is converted to transactions with the profit table to calculate the utility values of the items, so that the most utility values can help us predict the best vaccine. Below is the table of profit/external values Table 3 and transactions table Table 4

RTPCR	BP	Temp	SPo2	DDimer	ESR	CORADS	Severity_Score	Hb1ac	CRP
89	10	23	60	56	18	80	93	30	19

Table 3: External Utility / Profit table

Pid	Items	Total_utility
1	RTPCR (9), BP (2) , Temp(1), CORADS (3), Severity(7)	1735
2	Temp(2),SPO2(1),CORADS(1),DDimer(2),Severity(2)	484
3	RTPCR(8), BP(4),ESR(3),hbalc(2), CRP(6)	980
4	RTPCR(9),BP(6),Temp(5),Ddimer(3),CORADS(6),severity(10)	2554
5	SPO2(1),ESR(1),CORADS(2)	238

Table 4: Transaction table of patients

Now calculate TWU of all the items in the transaction table:

RTPCR = 5269
BP= 5269
Temp= 4773
SPO2= 722
DDimer= 3038
ESR= 1218
Corads= 5011
Severity_level= 4773
Hb1ac= 980
CRP= 980

If the utility >1000 , consider the items to be high utility items.
Now only the records of patients having these attributes mentioned will be taken for naïve bayes theorem and calculate the probability to find the Vaccine "label"

So,
A new patient Id = "8", as shown in above table 2, will be suggested "Covaxin" from the algorithm using Covid Classified High Utility Itemset Mining algorithm (Covid_clHUIM) algorithm.

III.PROPOSED SYSTEM

HUIM is an advanced version of associate rule mining (ARM),Unlike Frequent itemset mining which has support and confidence measures, HUIM focuses on utility measures and not the frequency patterns. This application predicts which combination of pharmaceutical products can be used to survive from this pandemic.The application takes the patient symptoms and medicinal reports as an input and gives the suggested vaccine accordingly.

High utility pattern mining is an emerging data science task, which consists of discovering patterns having a high importance in databases.A popular application of high utility itemset mining is to discover all sets of items purchased together by customers that yield a high profit.

An itemset is said to be High Utility itemsetif it is a frequent item in the database and possess utility factors.
The main challenge in mining HUIs is to maintain Downward closure Property. Downward closure property tells that all the subsets of a frequent itemset must be frequent and similarly all the supersets of infrequent itemset cannot be frequent. To maintain the downward closure the TWU concept (above mentioned definition 2.3) is used and find the high utility items. Using the list structure that is mentioned in above definition 2.5, we can form the items in their increasing order and unlike CHUIM [10], we have introduced an additional measure "K", which retrieves the Top K values in the increasing order of the TWU, which yields better results compared to CHUIM.

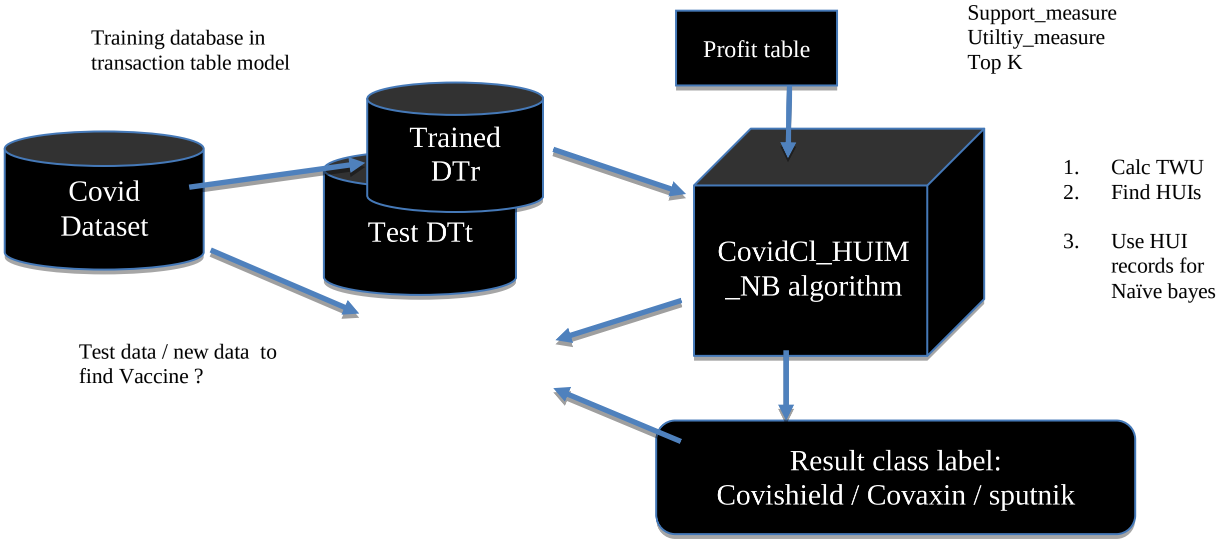

Fig 1 Proposed model

A covid dataset is taken where patients had covidpositive, negative or other health issues and various symptoms and this existing data is taken for analysis to predict which vaccine can be suggested to people who are positive/negative at what time period, which kind of vaccine is suitable with the help of proposed model.

The new data or test data (80- 20 rule) is separated and the model is designed with Test DTr and remaining testing data is in Test DTt, The Test DTr data is converted to transaction table using the CA rule section 2.6 as mentioned above, and the input is a profit table, support measure, Top K items and using these, we find the TWU values of each item(attribute) and find the high utility items.

After finding the high utility items using utility measure, we use the top K value to further filter the high utility items. Using the attributes that are obtained, we take the trained dataset records based on these attribute values existence and use the naïve bayestheorem to identify the result class, if the vaccine suggested is covishield or covaxin or sputnik.

Algorithm

The proposed algorithm has three phases with sub algorithms namely HUI algorithm, Search & construct algorithm, CovidCl_HUIM algorithm and Naïve bayes classification algorithm.

Phase 1 :HUI_Search

I/p: Transactional dataset DTr, profit table,utility_measure, topK

O/p: High utility itemset (HUI) items.

1. Let S be the set of items that satisfy TWU(item) ≥ utility_measureand the total order of set of items *Ti**

2. For each item setTi$_{\text{item}}$ S do

3. For eachTi_jTi$_{\text{item}}$ do

4. Scan the transactional dataset DTr only once

5. Calculate total weighted utility(TWU), TWUt(item) for each items Ti in DTr

such that the TWU(item) ≥ utility_measure

```
6. Sort the S items in the increasing order of TWU(item)
7. Use pruning and search space for all $Ti_j$
 8. Call Search($Ti*$, utility_measure, search_space)
 9. S= HighUitiltyItems(HUIs)
```

Phase 2: CovidCl_HUIM Algorithm

```
I/p:High utility itemsets(HUIs) , utility_measure , top K
O/p: generate subset of rules for classification ,DTr (pruned)
1. S←{S Rules} => Class {DTr}
2. Sort Rules in a descending order
3. For every S in DTr, Scount←0
4. If  DTr not empty and Sruleset not empty
 5. Find all new S's with all class labels of the form
{SRules} => {class_label}
6. Compute util_sup and util_conf
7. Remove the rules where rule util_sup $\geq$ util_measure,
                                        util-conf $\geq$ util_measure
8. Remove the items with the the rules in DTr
9.  Remove the items with Top K value
10. Generated subset of association rules
11. Return  DTr with minimal attributes
```

Phase 3: CovidCl_HUIMNB Algorithm

```
I/P: DTt, S
O/P: a class member from the result class
1.For every distinct attribute,value pair [attk,valk] in DTr
2.For every class label in result class Clk
3.P(Clk) =no.ofoccurences/ total instances in DTr.
4.P([attk,valk] | clk)=P([attk,valk]) ∪ P(Clk) / P(Clk), where k=1,2,3,….n
5.DT ←Take instance to test without class label from DTr
6.If (P[attk,valk]|clk) > (P[attk,valk])|Cli) where i⊆k then
7.$Cl_k$←result_class
8.Else$Cl_i$←result_class
```

The algorithm is divided into 3 phases, where the 1st phase calculates the high utility items by pruning the items when the utility measure is not reached and after sorting them in TWU order, the topK is also further removed and we get the high utility itemsets as a result and in the second phase, these HUIs are given as input to form as a class label form in order to find the suggested vaccine.

Once it is replaced, we apply naïvebayes in the third phase to find the resulted label.

IV.EXPERIMENTAL RESULTS

Accuracy is one of the performance measures which gives the accurate rate of success of the algorithm. The following algorithms are compared for classification accuracy using on CHUIM, CBA, MCAR[11] with attribute selection and using min support and confidence measures. We have crossed folded 10 times of 10,000 patient records and every time we use the 80% and 20% datasets as trained and tested data and with the stratified cross validation, we get the correctly classified instances and incorrect ones and a detailed work on the accuracy as shown in fig 2. The accuracy, specificity and sensitivity are more when the conf_measure is 95-100% as shown in the fig 3.

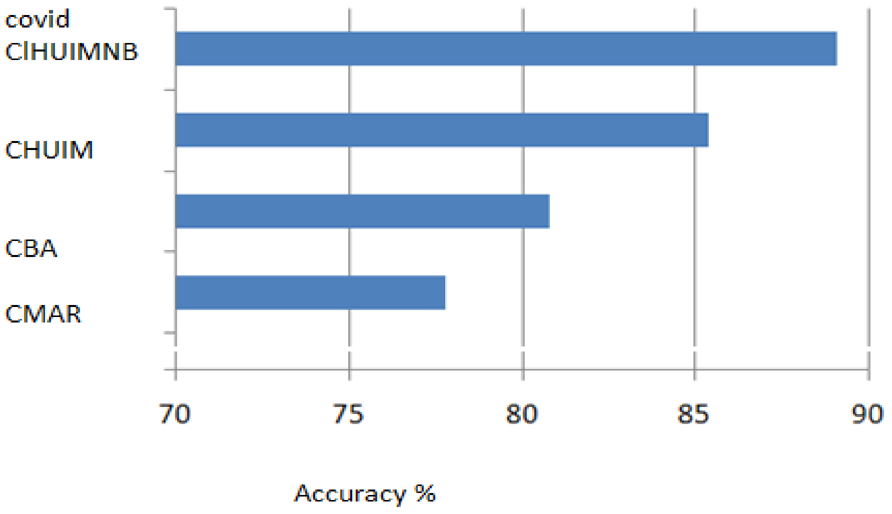

fig 2. Accuracy comparison of various algorithms

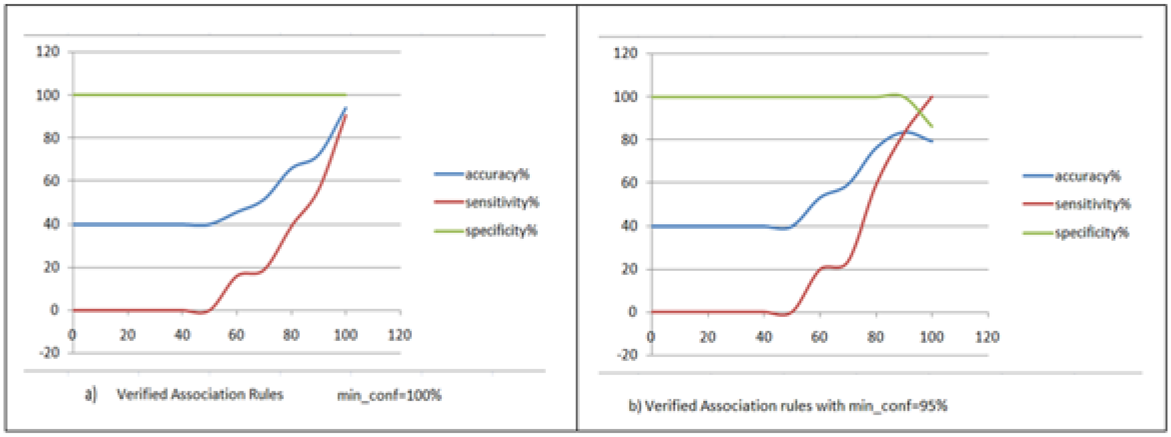

Fig 3: ACCURACY, SENSITIVITY, SPECIFICITY of CovidCl_HUIM_NB algorithm

V.CONCLUSION

High utility itemsets (HUIs) mining along with classification algorithms , widely called as associative classifiers show a significant improvement compared to normal classification methods, when used in emsemble ways, the result shown are impressive.

| DOI: 10.15680/IJIRCCE.2021.0907051|

There is a new measure added in the algorithm of CHUIM and used in CovidCLHUIM for better results, The comparison study showed the proposed algorithm is effective in both time and space complexity, but there can be more improvement in the space complexity with a different datastructure combination which we are focusing in the next experiments.

REFERENCES

[1] JagmeetKaur ,NeenaMadan. Association Rule Mining: A Survey . International Journal of Hybrid Information Technology Vol.8, No.7 (2015), pp.239-242 .
[2] Trupti A. Kumbhare, Prof. Santosh V. Chobe .An Overview of Association Rule Mining Algorithms. International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 927-930
[3] Surbhi K. Solanki, Jalpa T. Patel .A Survey on Association Rule Mining. 2015 Fifth International Conference on Advanced Computing & Communication Technologies.
[4] T. Karthikeyan and N. Ravikumar. A Survey on Association Rule Mining .International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1.
[5] U. Suvarna, Y. Srinivas .:Efficient High-Utility Itemset Mining Over Variety of Databases: A Survey, 803-807(2018).
[6]Jiawei Han, MichelineKamber, Jian Pei.Data Mining Concepts and Techniques, Third edition.
[7]Fournier-Viger, P., Chun-Wei Lin, J., Truong-Chi, T., &Nkambou, R. (2019). A Survey of High Utility Itemset Mining. Lecture Notes in Computer Science, 1–45.
[8] Tseng, V.S., Wu, C.-W., Viger, Yu, P.S.: Efficient algorithms for mining top-k high utility itemsets. IEEE Trans. Knowl. Data Eng. **28**(1), 54–67 (2016)
[9]Sudip Bhattacharya1, Deepty Dubey2,High Utility ItemsetMining,International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 8,August 2012).
[10] U Suvarna, Y. Srinivas: A Classified Medical Infertility Dataset using high utility itemset mining, International Journal of Recent Technology and Engineering (IJRTE),(ISSN:2277-3878,Volume 8, Issue-2, July 2019).
[11] Thabtah, F., Cowling, P., &Peng, Y. (2005, January). MCAR: multi-class classification based on association rule. In *Computer Systems and Applications, 2005. The 3rd ACS/IEEE International Conference on* (p. 33). IEEE

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details