International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

A Classified Medical Infertility Dataset using High Utility Item Set Mining

Suvarna U, Srinivas Y

Abstract*: *The most modern technological innovations led towards generating lots of data, which is either redundant or of imperative use. To mine the meaningful information from this huge repository, Data mining techniques will be of vital importance. This article aims at mining the useful patterns from this enormous repository and presents some possible solutions while treating the patients suffering with various problems of infertility. A Classified High utility item set mining with Naïve Bayes classification (CHUIM-NB) is proposed for classifying the data, which will be of productive usage to the Medical Practitioners during the treatment of the patients. The proposed model has three stages: the stage1 aims at generating the training data, the second stage aims at proposing a two phase algorithm for producing high utility item set and also the rules for association mining (CHUIM) and in the third stage, the Naives classification model (CHUIM-NB) is considered for the effective diagnoisis/ treatment.

Index Terms*: *High utility itemset mining (HUIM), Classification, Naïve bayes Classification, InVirto Fertilization (IVF).

I. INTRODUCTION

In today's social world, every married couple, aspire for having children to be a part of their family and the dream for children became a binding thing in their lives. Regrettably, not all the women could fulfill this dream, due to infertility. To treat infertility, various treatments are presently available, in order to neutralize the hormonal imbalance, stimulate the ovulation, providing progesterone's and immunoglobins to balance and make the female fertile capable for conception. One such treatment is Invirto fertilization (IVF) as in Fig 1. In IVF, there are major differences in treating the patients like IVF /IUI /ICSI etc, the developed system helps to suggest the most suitable treatments for a patient based on the information provided. If a system, that can read the behavior of the previous patients and their treatments and suggest the appropriate & accurate treatment to undergo, which helps both the doctor/clinic and the patient in the following ways:

a) saves the patient number of attempts to the treatment
b) saves patient investment on the treatment
c) Increases the success rate of the doctor/clinic treating the patient with their predictive analysis.
d) A positive feedback of the patients and confidence of the doctors treating them will become enormous.

The rest of the paper is organized as follows: section 2 discusses about the brief review of literature carried out in this direction. In section 3, general concepts of IVF, dataset, associative classification, high utility item set mining have been discussed. Section 4 explains the problem statement of CHUIM. The proposed algorithm of CHUIM is explained in section 5. The experimental results can be found in section 6 and section 7 concludes the work.

II. RELATED WORKS

Keeping in view of the seriousness of the problem, several authors and researchers have presented many articles with a due focus of pointing out the ideas to increase the fertility rate. Among most the works already presented, methods based on association rules, high utility mining and classified association are discussed in this section to present the concept in its right perspective. Traditional Association rule mining (ARM) approached is proposed by Agarwal et al,(1994), however, the main limitation of this method is that it could not able to produce the associations among the attributes accurately and thereby generated numerous number of rules and eventually failed to handle large datasets. The authors utilized Aprioiri algorithm to find the frequent items, which needs too many scans of database and generated numerous items and from the items, generating meaning rules from these huge item set lacked efficiency. To overcome the limitations, the quantitative association rules were developed for large relational datasets and presented by Agarwal et al,(1996), which were able to handle large datasets but not the candidate item sets and however, the main problem identified was frequent patterns .Tseng et al, (2013), Liu et al (2005) Yao et al,(2004) Ahmed et al,(2009) to address major issues with large very number of items generated from Apriori to FP growth where the patterns were filtered to UP growth. The authors presented the article in a two phase manner that identifies the candidate item sets and eliminates the low utility item sets. Liu et al (2012) HUI_Miner uses a single phase unlike the above two phase algorithms; the author uses a utility list which doesn't require any candidate generation. A Fast algorithm Viger et al, (2014) FHM is similar to HUI Miner but uses a depth first search and finds a search space for HUIs with optimization methods called EUCS to prune the search spaces. Next, the association rule which effectively mines with fewer candidates and less number of scans is considered for the integration with the classifier algorithm.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Suvarna U, Dept of IT, Gitam University, Visakhapatnam, AP, India.
Srinivas Y, Dept of IT, Gitam University, Visakhapatnam, AP, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Many efficient classifiers are used to draw conclusions and predict the right outcomes. Every classification algorithm serves different purpose of outcomes of which the popular are classifiers like decision trees algorithms like ID3, C4.5 Quinlan et al, (1979,1993), Hunt et al,(1966) that creates complex trees and do not generalize well .Another algorithm is naïve Bayes which assumes the independence of every pair of feature with the help of a conditional probability and logistic regression classification Brieman et al, (1984) which are used for a single outcome, With the k-nearest neighbor technique, this is done by evaluating the k number of closest neighbors Xindong et al, (2007) and there are gradient descent classification where it supports linear models but suffers with sensitivity to feature scaling and support vector machines Vapnik et al, (1995) and Cristianini et al, (2000) which are considered a good classifier because of its high generalized performance without the a priori knowledge and is insensitive to the outliers. For our model, Naive Bayes is more suitable because of the multiple outcomes and categorical attributes in our dataset.

Associative Classification is a process of integrating both classification and association rules for predicting even accurate result class. Liu et al (2002) proposed an integration algorithm integration of classifier with association rules were explained by Yoda et al, (1997) &Liu et al (1997),but ineffective for attributes with continuous values and was a major issue, where c4.5 classification and ARM techniques using Apriori algorithm ,Ma et al,(1998) CBA, classification based on multiple association rules Li et al (2001)CMAR , classification based on predictive association rules (CPAR), a multi-class based on association rule Thabtah et al,(2005) MCAR were all similar with little advancements in the association classification in increasing the effectiveness.

III. BACKGROUND

A. *In virto Fertilization*

Infertility: It's a perception that if the female is not able to conceive, it is the problem of female alone and always male factors are taken granted and never considered. But, childlessness is caused due to a range of evaluating factors with respect to both female and male partners. Infertility is defined as not being able to conceive for a reproductive age woman, which is normally considered to be 35 years. After evaluating female and male fertility factors for clinical need to perform routine tests, Arlene et al, (2011) they turn out with treatment options for infertile couples which is explained in the below table I. The table has a value "unexplained", where the suggestion or prediction of a particular treatment is still a challenge. In this paper, we explained only one procedure IVF for understanding the medical infertility problem. The remaining methods like ICSI, IUI are not explained. In the dataset, alternate IVF classes are included. Various clinics have different procedures to perform this IVF process, Raju et al, 2005 have suggested the stages of performing the IVF procedure is in fig 2. The first stages include the suppression of the natural release of hormones during the 18^{th} to 24^{th} day from the start of periods(for female), stimulates the ovaries to produce the follicles

I: Male & female infertility factors and suggested treatments

Patient condition	Suggested Treatments
Female Infertility:	
Ovaries failures	Ovulation Induction (OI)
Tubal factor	Controlled ovarian stimulation(COS) with IVF
Endometriosis	Intra Uterine Insemination(IUI) or COS with IVF
Male Infertility:	
Male subfertility	IUI with or without OI
Female Infertility:	
Unexplained	IUI or IVF or COS with IVF or Intracytoplasmic sperm injection(ICSI)

Invirto Fertilzation (IVF) which is commonly known as a Test tube baby method.

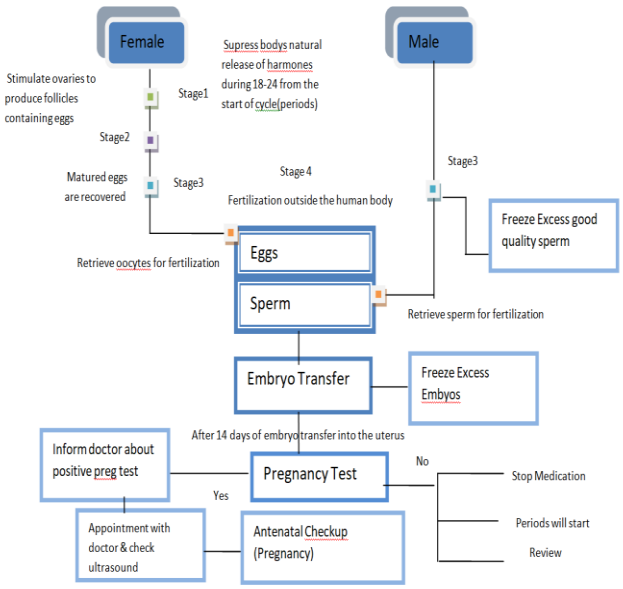


Fig 1: IVF procedure (not considered OI,IUI, ICSI or COS IVF in this process)

containing eggs and later the matured eggs are recovered, which are called oocytes.Parallely they collect the sperm from male and wash it and separate them from poor quality sperm and produce high quality sperms, the required number of sperms will be used for fertilization, remaining excess sperms will be kept frozen for future use in sperm banks. The next stage is retrieve oocytes and sperms together for fertilization,outside the human body (in a test tube).Once the embryo is formed, excess embryos are preserved frozen and pregnancy test is conducted after 14 days of the embryo transfer to uterus, and based on the test positive or negative, the medication is stopped or consult doctor for antenatal checkups respectively.

II: Attributes for the relation IVF_dataset (synthetic dataset)

Attributes	Description
Season	Seasons in which the analysis was performed 1) winter, 2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1)
Sex	sex (male or female)
Age	Age at the time of analysis. 18-38, above 38yrs, IVF treatment not much suggested for females.
Child_disease	Childish diseases (ie , chicken pox, measles, mumps, polio) 1) yes, 2) no. (0, 1)
Accidents	Any accident or serious trauma 1) yes, 2) no. (0, 1)
Surgeries	Any Surgical intervention 1) yes, 2) no. (0, 1)
Fever	High fevers in the last year 1) less than three months ago, 2) more than three months ago, 3) no.
Alcohol	Frequency of alcohol consumption 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never
Smoke	Smoking habit 1) never, 2) occasional 3) daily.
Sit_hours	Number of hours spent sitting per day 1-16
Condition	Condition based on female and male infertility factors {ovaries_fail, tubal factors, endometriosis, malesubfertility, unexplained}
Weight	{Underweight, normal BMI ,Overweight, obese}
Attempts	Attempts of treatment already done from 0-5, above 5. More than 5attempts i.e, 6 is not suggested. (0,1)
Diagnosis	{IUI, IVF , COS-IVF, ICSI, OI}

In this paper, we have considered a IVF synthetic dataset with 14 attributes and 10,000 patient records. Initially, the dataset is partitioned into 2 groups, training and test data. The training dataset is considered and high utility item set is applied to mine the most suitable trained data and then apply Naïve Bayes classification model to suggest the result class**.** The IVF dataset is collected with the various facts and information about the patient and the doctor needs to examine the patient with these details **There are** associative classifiers like CAR, CBA, CPAR, CMAR which uses a combination of advanced association to work on classifiers for more accuracy. Through our CHUIM, we can reduce the training dataset with valid and accurate records to generate the optimal outputs and also the rules can be minimized as it generates top k results based on CHUIM

B. IVF dataset

A dataset is very large volume of related data that can be used as an input file for a problem that has raw data and process it to find some productive results. For our model, we have used an IVF dataset collected from UCI repository and edited to 14 attributes based on the infertility factors available in table I and 10,000 patient records are made based on the predicted and unexplained factors options. The IVF dataset has14 attributes that is useful in predicting the correct diagnosis for the patients would look like the following table II. The Description of each attribute is briefed in the table II and the values used for all the attributes are either real or nominal values. Nominal values are more useful for association kind of problems, which is why this dataset is accurate for testing the model. Table III is the IVF dataset in tabular format with a sample of few listed records with all the attributes

C. Associative classification:

Association and classification are two important and different data mining techniques that are used with various applications in getting the processed results.

III: IVF_dataset (a part of dataset is shown in the table)

Season	Sex	Age	Child_disease	Accidents	Surgeries	Fever	Alcohol	Smoke	Sit_hours	condition	weight	Attempts	Diagnosis
spring	male	35	yes	yes	no	less_3_months	hardly_or_never	never	11	malesubfertility	obese	3	IUI_with_OI
spring	female	23	yes	no	no	less_3_months	several_week	daily	1	unexplained	under_weight	1	ICSI
spring	male	32	yes	no	yes	less_3_months	everyday	ocassional	1	malesubfertility	normal	1	IUI_with_OI
spring	female	30	no	no	no	more_3months	several_week	daily	1	endometriosis	normal	2	COS_with_IVF
summer	female	23	yes	no	no	more_3months	hardly_or_never	ocassional	6	endometriosis	normal	0	IUI
summer	female	25	yes	no	no	no_fever	everyday	ocassional	4	unexplained	normal	0	ICSI
summer	female	27	no	no	no	less_3_months	several_week	ocassional	1	ovaries_fail	normal	0	OI
fall	female	23	yes	no	no	more_3months	several_week	never	2	endometriosis	obese	0	IUI
winter	female	24	yes	no	no	less_3_months	several_week	never	3	ovaries_fail	normal	1	OI
winter	female	30	no	no	no	more_3months	hardly_or_never	daily	4	endometriosis	normal	5	COS_with_IVF
winter	female	28	no	no	no	less_3_months	hardly_or_never	daily	1	tubal_factor	normal	2	COS_with_IVF
fall	female	21	yes	yes	yes	more_3months	once_week	daily	5	tubal_factor	under_weight	0	IVF
fall	female	22	yes	yes	yes	more_3months	several_week	daily	5	endometriosis	normal	0	IUI
fall	female	19	yes	yes	yes	no_fever	several_week	daily	1	tubal_factor	normal	0	IVF
fall	female	27	no	yes	yes	no_fever	once_week	daily	4	endometriosis	normal	3	IVF
spring	female	23	no	no	no	no_fever	several_week	ocassional	12	tubal_factor	obese	0	IVF
spring	female	24	no	no	no	more_3months	everyday	ocassional	10	ovaries_fail	obese	4	OI
fall	female	21	yes	yes	no	less_3_months	several_week	ocassional	14	endometriosis	normal	2	IUI
fall	female	33	no	no	no	more_3months	several_week	ocassional	15	ovaries_fail	obese	4	OI
winter	female	34	yes	no	no	no_fever	several_week	ocassional	12	tubal_factor	over_weight	4	IVF
winter	female	35	no	no	no	more_3months	everyday	ocassional	13	ovaries_fail	obese	4	OI
winter	female	36	yes	no	no	no_fever	once_week	ocassional	15	unexplained	over_weight	0	IVF
winter	female	32	yes	no	no	no_fever	hardly_or_never	daily	16	tubal_factor	obese	0	COS_with_IVF

Associative classification combines these two techniques and aims at the best associative rules to be listed and classified with most accurate results. In this paper we aimed at the application of a medical infertility data,where the target class (class Diagnosis in table III) is to suggest the best trestment possible which is a class label (IUI,IVF,OI,etc.. in table III) for the patient who arrives at the clinic using the integrated model. Association Classification typically consists of two steps. The first step aims at the subset of association rules which are both accurate and frequent association rules and the second step performs the rules for classification using classification model or a classifier.

Classification

Classification is a concept which is concerned with prediction using a class model or a classifier and involves two steps. The first by taking the dataset and predetermine a training set and test set(for eg: 70% training set & 30% test set) and the second step by applying the model to predict the future classes(class labels) for the test data.We can perform classification using if-then statements ,decision trees, Bayesian classification, neural networks, support vector machines. This technique is purely a supervised learning method. The most suitable classification algorithm for the proposed model is Naïve bayes because it is the most simple and effective way to get accurate results for categorical attributes and large datasets with multiple class labels.

Associative rule mining

Associative rule mining (ARM) is a procedure that aims at frequent patterns, associations or correlations from the datasets, mostly used with transactional databases in market research and many other applications. The Association rules are the statements that discover relationships between unrelated data and extract the interesting measures for finding frequent patterns. The Association is best suited for categorical or nominal data rather numerical.ARM has been replaced with high utility itemset mining from past decades where it rules out taking the interesting measures like support and confidence with an additional measure called utility. There are various HUIMs developed past few years that perform well with respect to memory usage, run time and accuracy. The proposed algorithm uses minimal high utility itemset mining developed by Viger et al ,(2016)

IV. PROBLEM STATEMENT AND PROPOSED MODEL

A. *Problem Statement*

Definition 1. Association rule: Let D be a transactional Dataset ,

T= {Tid_1, Tid_2, Tid_3,……Tid_n} are the transactions of the dataset , where Tid_1,Tid_2,…….Tid_j are the number of transactions in the dataset and let I = {It_1,It_2,It_3,…..It_n} are the set of items in the transaction dataset. It_1,It_2,….It_j are the number of items in the transaction database.

An association rule of the form X => Y, where the position of item X is an antecedent and item Y is a consequent respectively.

A rule X => Y, where X is an item in I of a tranasaction T in a database D ,where X⊆T, and X and Y are items such that X ⊂ I , $Y \subset I$, *and* $X \cap Y = \varnothing$

Example 1: X => Y can be of the forms buys(mobile) => buys(charger) or {Age(>20),person(student)} =>{buys(mobile)} or {Time(5pm),buys(mobile) }=> {buys(charger)}

Definition 2. High utility itemset mining: An itemset X is a high utility itemset if utility of X ≥ min_util, ,where min_util is the utility threshold measure

Definition 3. Utility of an item 'It' in a transaction Tid , util(It,Tid):The product of internal utility (iutil) and external utility (eutil) of an item "It" of itemset x in a daaset D.

$\mathrm{Util(It, Tid)} = \mathrm{eutil(It, Tid)} \times \mathrm{iutil(It, Tid)}$. Where "It" is i^{th} item of Items and Tid is an i^{th} transaction.

Definition 4. Internal utility, iutil(It): The quantitative measure of each item in the transaction T in a database D ,where X⊆T

Example 2: The below table IV is a transactional dataset where no. of transactions are 4, Tid_1,Tid_2, Tid_3 and Tid_4 and the no. of items are It_1, It_2, It_3. A Tid1 is a transaction record that have 10 internal utility in item 1 (it_1) , For Transaction Tid1, iutil(it_1)=10 , iutil(it_3)=1.

IV: A Transaction dataset Dt

TID	Items in a transaction		
	It_1	It_2	It_3
Tid_1	10	0	1
Tid_2	0	8	0
Tid_3	0	0	2
Tid_4	0	2	6

V: A Profit table for external utility

Items	It_1	It_2	It_3
Profit	3	2	1

Definition 5. External utility, eutil(It): The profitable measure of each item in x in a transaction T, where X ⊆ It

Example 3: The table V is a utility table where items are It_1, It_2, It_3 and each item is given a profit. External utility eutil(it_1) is 3 and so on.

Therefore for Tid_1, util(It_1)= 10×3=30

Definition 6. Total Utility util(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.

$$\mathrm{util}(x) = \sum_{It \in X \wedge Tid \in D} \sum_{It \in X} util(It, Tid)$$

Definition 7: Transaction-Weighted Utilization (TWUtil) measure, which provides an upper-bound on the utility of itemsets and is anti-monotonic property. The transaction-weighted utilization (TWU) of an itemset X is defined as the sum of the transaction utilities of transactions containing X, i.e. TWUtil(X)

$$= \sum_{Tid} util(Tid)$$

Definition 8: 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 ⊂ X.

Property 1: (Pruning search space using the TWUtil). From Liu et al, (2012) , Let X be an itemset, if TWUtil(X) < min_util, then X and the supersets of X are always low at utility.

Property 2 (utility of an itemset using its util-list). The utility of an itemset is the sum of internal utility values in its utility-list. i.e, iutil values.

Property 3 (Pruning search space using util-lists). Let X be an itemset.

Let the extensions of X be ExtX ,the itemsets 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.
The Definition 1 to definition8 discusses the general support, confidence and utility measures that are used in high utility itemset mining and how to effectively extract min HUIs and property 1-3 are useful in making an utility structure without producing any candidate generations. The definitions from Definition 9 to Definition 12 disccusses on how to define the Classified Association rule to apply to a classifier.

VI : IVF Dataset is converted to [Attribute, value] pair

Tid	[season-winter]	…	[alcohol, never]	…	[smoke, occasional]	**Diagnosis**
1	1		1		1	**IVF**
2	0		0		1	**COS_with IVF**
…						**IUI**

Definition 9. An [Attribute , value] pair Transaction dataset , Dt is a dataset from Da.
Let A={ A_1, A_2,….., A_{n-1}} are the attributes in Dt and C= A_n is the Result attribute that has the Class labels .i.e, n^{th} attribute of the Dt. And Let V= { v_1, v_2,….., v_n} are values in each attribute A_i. Then the itemset {It_1, It_2,….} of the Dt is { [A_1,v_1], [A_2,v_2],….. } and the last item is called the class label C ={c_1, c_2,….., c_n}
For example: From table VI, [season, winter] is an item and diagnosis is the result class and class labels are IVF, IUI, etc..
Conditions:
a) Attribute which is pair (attribute, value), is used in place of Item. For example (Season, winter) is an item formed from Attribute Season and value winter in Table VI
b) An Attribute set is equivalent to Itemset for example : [Season,winter], [Smoke, occasional] .
c) Support count of Attribute [Ai, vi] is number of matching rows of Attribute. For example: [Season, winter] count is 30% in Dt.
d) High utility itemsets will be the Attribute sets whose Attribute (Ai, vi) passes the *minsup* threshold if support count {[Ai, vi]…[Aj, vj]}, ≥*minsup*.
For Example: [Season,Winter] has support count= 30% and [Smoke, occasional] has support count=60% and min_sup=50%, then we only consider the [smoke,occasional] to be the high utility itemset.
Definition 10. New Association rule: A rule X=> Y is an association rule, where the position of item X is an antecedent and item Y is a consequent respectively.
A rule X => Y, where X is an item having { A_1, A_2,….., A_{n-1}} of a tranasactional dataset Dt ,where X⊆A, and Y the consequent is the last item in Dt ,which is a class label C.
The new association rule will be of the form { A_1, A_2,….., A_{n-1}} => {c_1, c_2,….., c_n}
Definition11. Classified Association rule[CA Rule]: [Modified Definition 8] , 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
Conditions: a) {{[Ai, vi], …,[Aj, vj]}, C } is called rule attribute.
b)Support count of rule attribute {{[Ai, vi], …,[Aj, vj]}, C } is the number of matching rows of dataset.
c) Rule attribute {{[Ai, vi], …,[Aj, vj]}, C } ≥minconf .
Definition 12. A Classifier, Given two rules r_i and r_j and r_i > r_j ,when
a) Conf of rule r_i is > r_j
b) Conf of rule r_i is = r_j , but Sup of rule r_i is > r_j
c) Both Conf and sup of rule r_i is = r_j , but r_i generated earlier than r_j
Where Conf is the confidence & Sup is the support measure.

B. *Proposed Model*

Our proposed model Classified High utility item set mining (CHUIM) with naïve bayes classification, CHUIM-NB (Fig 2) modifies the association rules into Classified association rules. Initially the stage 1 is a Dataset D is divided into a training set 70% (Da) and test set 30%(Db) (as in Fig 2) and the next stage we convert the dataset to be convenient for association ,which is a transactional kind of database (Dt) as shown in table VI. Here the phase1 starts. We give a profit table (table V) which is addition of simple weights/profits to all the items of Dt and name it weigtedDt or WDt. We then apply the high utility itemset mining using a utility threshold to find all the frequent items, This is the stage where we remove all the infrequent items, thereby the combination of low profit rules are removed indirectly. For CHUIM, We use an utility list structure ,a Depth first search

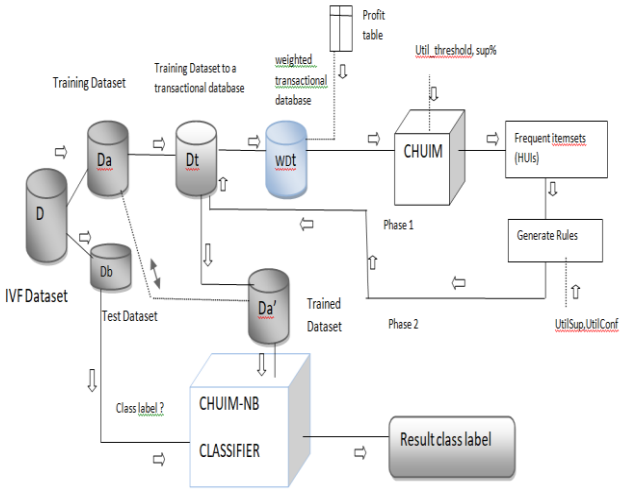


Fig 2: CHUIM-NB proposed model

and EUCS which is performed with a hash map of hash maps where (a,b,c) such that c≠0 are kept. This is very fast and a single database scanning method of HUIM. The results of phase 1 produced are HUIs that are minimal and highly profitable. In phase 2, We generate the rules with the available HUIs and supplement the utilsup, utilconf and prune the items in Dt that are not minimal HUIs. At stage 3, Dt (new set of items [At_i, Vt_i]), is converted to Da' which is not a transactional database anymore,

but the regular dataset with trained data records ([A_1, A_2,….., A_{n-1}]) and we generate Classified association rules. ({ A_1, A_2,….., A_{n-1}}) => {c_1, c_2,….., c_n}. With this Da' and the test dataset Db, using naïve bayes classification, we get the result class label from multiple labels.

V. PROPOSED ALGORITHM

Phase 1: This phase has three algorithms, the HUI algorithm, the search algorithm and the construction of utility list algorithm.

Phase 1 ,Algorithm 1: HUI Algorithm

Input: WDt dataset (Dt dataset & profit table) , min_util

Output: High utility itemsets (HUIs)

1. Scan the training transactional dataset Dt and calculate TWutil(x) of each item It_i such that TWutil(x) ≥ min_util
2. Let SI be the set of items that satisfy TWutil(x) ≥ min_util and the total order of set of items *It**
3. Sort the SI items into ascending order of TWutil(x) values.
4. Scan D to buid utility list & build EUCS for pruning and search space for all It_i , where $It_i \in$ It* such that sum({ It_i ,*utillist.iutil)* ≥ min_util
5. Search(*It*,min_util,EUCS)*
6. Remove the items from Dt where TWutil(x) < min_util
7. For each itemset $It_x \in$ SI do
8. For each $It_i \in$ It_x , SI has high utility itemsets
9. Return HUIs

The above algorithm 1 takes the dataset and profit table and min_utlity threshold to generate HUIs. We calculate the transaction weighted utility of item set and compare with min_utilty threshols and keep the set of all such items in SI , after sorting to increasing order .we build an utility list,prune and search using EUCS (Estimated utility cooccurence structure, where TWutil(A,B)=C is mentioned in a structure as (a,b,c)) from algorithm 2. Further we get the items that are sum of all high profitable items that satisfy sum(It_i ,*utillist.iutil)* >= min_util, we use effective searching & pruning strategy using EUCS and Depth first search to remove the items where TWutil(x) < min_util and finally return the HUIs which are minimal and each item is $It_i \in$ It_x and item set $It_x \in$ SI .

Phase 1 Algorithm 2: Search & construct algorithm

Input: Itemset A, ExA is the extensions of A, min_util, EUCS

Output: set of HUIs

1. For each itemset $A_i \in$ ExA
2. if Sum of A_i utillist internal and external utils sum{$A_i$utillist.iutil} + sum{$A_i$utillist.eutil} ≥ min_util then
3. ExA ←∅;
4. For each $A_j \in$ ExA such that j has set of items of It* as i
5. If there exists (i,j,c) ∈ EUCS such that c ≠0, c ≥ min_util
6. $A_{ij} = A_i \cup A_j$
7. Construct utillist of A_{ij} , set A_{ij}.Utillist ←∅
8. For each record $e_i \in A_i$.Utillist
9. If there exists $e_j \in A_j$.Utillist and e_i.tid = e_{ij}.tid and
10. If A_i.Utillist≠ 0 ,Search element e ∈ A.Utillist such that e.tid = e_j.tid
11. e_{ij} ← (e_i.tid,e_i.iutil + e_j.iutil,e_j.eutil)
12. otherwise e_{ij} ← (e_i.tid,e_i.iutil + e_j.iutil−e.iutil,e_j.eutil
13. A_{ij}.Utillist ← A_{ij}.Utillist ∪ e_{ij}
14. return UtillistA_{ij}
15. ExA = ExA ∪ A_{ij}
16. if sum{A_{ij}utillist.iutil} ≥ min_util then Return A_j

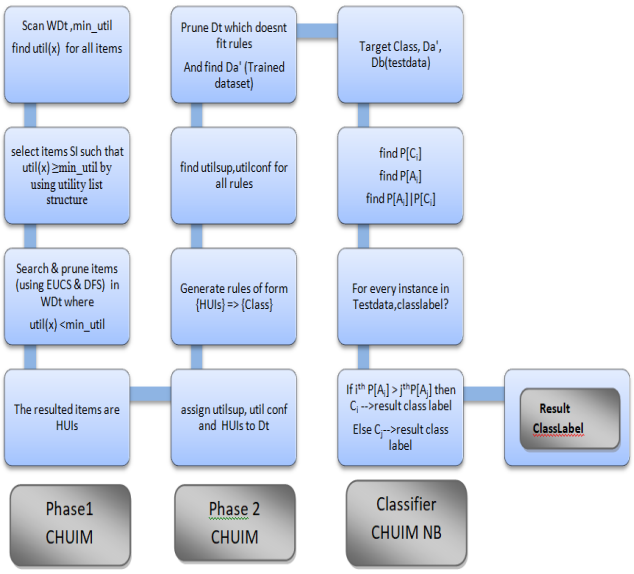


Fig 3: Algorithm flow of CHUIM-NB model

Phase 2:

Phase 2: CHUIM algorithm

input : high utility itemsets(HUIs) , util_sup, util_conf

Output: generate subset of rules for classification CARs , Dt (pruned)

i.e, {few SI rules} => Class {c}

1. SI←{SI rules} => Class {c}
2. Sort rules in a descending order
3. For every SI in Dt, count←0
4. If Dt not empty & SI ruleset not empty
5. Find all new SI's with all class labels of the form

{SI rules} => {class}

6. Compute utilsup and utilconf
7. Remove the rules where rule utilsup ≥ minsup, utilconf ≥ minconf

8. Remove the items with the the rules in Dt
9. Generated subset of association rules ARs
10. Return Dt with minimal attributes

Classifier CHUIM Naïve Bayes: CHUIM NB:

Input: Da', Class C , Db

Output: a class member from the result class

1. For every distinct attribute,value pair [Ai,vi] in Dt
2. For every class label in result class Ci
3. P(Ci) =no.of occurences/ total instances in Dt.
4. P([ai,vi] | ci)=P([ai,vi]) ∪ P(Ci) / P(Ci), where i=1,2,3,….n
5. Ts ←Take instance to test without class label from Db
6. If (P[ai,vi]|ci) > (P[aj,vj])|Cj) where j⊆i then
7. C_i←resultclass
8. Else
9. C_j←resultclass

The complete algorithm flow of phase 1, phase 2 and then the classification using naïve bayes flow can be viewed at Fig 3 for better understanding.

VI. EXAMPLE & EXPERIMENTAL RESULTS

A. *Example*

Step 1: Lets consider only a sample of above IVF dataset with 2 attributes and one class with only 2 labels showed below in table VIII, the 2 attributes season have 3 seasons and weight again 2 categories.

VIII: Sample IVF Dataset "D" for example.

s.no	Season	weight	Diagnosis
1	Winter	Obese	IVF
2	Summer	normal	IUI
3	Spring	normal	IVF
4	Winter	normal	IUI
5	Winter	Obese	IVF
6	Winter	Normal	IVF
7	Spring	Normal	IUI
8	Winter	Normal	?

Step 2: Now Dataset D has Training Data, Da:rows 1-7 in table VIII and Test data, Db are separated like table IX

IX: D is seperated as Da and Db for classification

Da	s.no	Season	weight	Diagnosis
	1	Winter	obese	IVF
	2	Summer	normal	IUI
	3	Spring	normal	IVF
	4	Winter	normal	IUI
	5	Winter	obese	IVF
	6	Winter	Normal	IVF
	7	Spring	Normal	IUI
Db	s.no	Season	weight	Diagnosis
	8	Winter	Normal	?

Step 3: The next step is to convert Da to Dt (A transactional database as in Table IV) and the new Dt table would look like table XI. And let the profit/weight table W for the attributes is as shown in table X. Both table X & table XI are WDt.

X: Profit /Weight table W)

Items	[season_ winter]	[season_su mmer]	[season_sp ring]	[weight_o bese]	[weight_nor mal]
Profit/ weight	60	20	40	70	80

Step 4: Phase 1 starts from WDt, Where Dt with W (weighted/profit table) .

XI: Dt of sample IVF Dataset for example

Tid	[season _winter]	[season _summer]	[season _spring]	[weigh t_obese]	[weight_ normal]	Diagnosis
1	1	0	0	1	0	IVF
2	0	1	0	0	1	IUI
3	0	0	1	0	1	IVF
4	1	0	0	0	1	IUI
5	1	0	0	1	0	IVF
6	1	0	0	0	1	IVF
7	0	0	1	0	1	IUI

Sup[season_winter]=4/7=57.14%,
Sup[season_summer]=1/7=14.2%
Sup[season_spring]=2/7=28.5%,
Sup[weight_obese]=2/7=28.5%
Sup[weight_normal]=5/7=71.4%, Iutil[season_winter]=4, Iutil[season_summer]=1,Iutil[season_spring]=2,Iutil[weight_obese]=2,Iutil[weight_normal]=5, and eutil is in table X.
And utility(x) =iutil * eutil
since the quantities of each transaction is 1 from table XI, the sum of all quantities with number of occurrences would be the number of occurrences of item in the dataset. Now Support% and utility factors can be shown in the table XII.

XII: CHUIM table with items and sup%, utility factors

s.no	Items	Sup%	Utility
1	Season_winter	57.14	4*60=240
2	Season_summer	14.2	1*20=20
3	season_spring	28.5	2*40=80
4	weight_obese	28.5	2*70=140
5	weight_normal	71.4	5*80=400

Step 5:Let us suppose the util_threshold =150 and min_sup%=25%,

Then we can prune the items Season_summer due to < min_sup% and also < util_threshold and the item season_spring < util_threshold but has min_sup%. This is the completion of phase 1 and the output is:

Step 6: The HUIs (frequent items) that are formed in CHUIM are

Season_winter
weight_obese
weight_normal

Step 7:In Phase 2 , we generate rules from the HUIs, and provide class , min_sup and confidence for items.The class has two classes IUI& IVF. We need the rules of the form [Atrribute] -> C, with the rules as follows: Season -> IVF , Season -> IUI for Winter, Weight -> IUI, weight -> IUI for obese & normal. And the pruned attributes are highlighted in table XIII for Dt.

XIII: Dt with highlighted pruning attributes

Tid	[season_winter]	[season_summer]	[season_spring]	[weight_obese]	[weight_normal]	Diagnosis
1	1	0	0	1	0	IVF
2	0	1	0	0	1	IUI
3	0	0	1	0	1	IVF
4	1	0	0	0	1	IUI
5	1	0	0	1	0	IVF
6	1	0	0	0	1	IVF
7	0	0	1	0	1	IUI

Step 7:

XIV: Dt pruning the generated rules that doesn't have util_sup,util_conf

Rules	Sup%	Conf%
Season_winter ->IVF	42.85	75
Season_winter ->IUI	14.28	25
Weight_obese->IVF	28.57	0
Weight_obese->IUI	0	0
Weight_normal ->IVF	28.57	40
Weight_normal ->IUI	42.85	60

P[season_winter] =4/7,P[Weight_obese]=2/7,P[Weight_normal]=5/7, P[season_winter,IVF]=3/7,P[season_winter,IUI]=1/7, P[Weight_obese,IVF]=2/7,P[Weight_obese,IUI]=0/7, P[Weight_normal,IVF]=2/7, P[Weight_normal,IVF]=3/7

Step 8: If util_sup=25% and util_conf=50%, then only season_winter ->IVF and weight_normal->IUI are the rules which prouduce high utility results.The highlighted portion of table XIV are pruned.

Step 9: Now convert Da to Da' with the above generated rules from table XIII and the final Da' is table XV which is the trained dataset.

XV: Da' Trained dataset.

s.no	Season	Weight	Diagnosis
1	Winter	Obese	IVF
2	Summer	Normal	IUI
3	Spring	Normal	IVF
4	Winter	Normal	IUI
5	Winter	Obese	IVF
6	Winter	Normal	IVF
7	Spring	Normal	IUI

Step 10: Now Da' is compared with Db and the treatment for Winter, normal ? , and the result classlabel is IUI, from table XV. [We apply Naïve which is a conditional probability of X => y, where X is an attribute and Yis a class and then conclude the result class label].

A. *Experiment results*

Accuracy is one of the performance measure which gives the accurate rate of success of the algorithm. The following algorithms are compared for classification accuracy using weka tool Sharma et al,(2013) for NB and J48(decision tree) and manually worked with java on CBA, MCAR and Hadi et al, (2018) HAR with attribute selection and using min support and confidence measures . We have crossed folded 10 times and everytime we use the 70% and 30% datsets 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. we consider the accuracy of the holdout method which is the remaining test segment after the folds and the percentage of dividing the trained and test datas.We considered the precision, recall, f-measure available with weka tool to compute the Accuracy % when compared to other Association Classification Algorithms with CHUIM-NB is as shown in Fig 4. We can observe the performance of CHUIM-NB is comparatively better than the discussed algorithms. Using the ROC formats, calculating the sensitivity (a true positive rate with correctly identified instances) and specificity(true negative rate , the incorrect instances identified) we can get the accuracy % and all the parameters are then worked with various confidence , min_conf% to see the variations of the algorithm behavior which is shown in Fig 5. We can see if the min_conf=100% the specificity is also 100% and there is a gradual increase in accuracy% and sensitivity% , as the confidence is reducing, there is a sudden change from min_conf 85% and more as confidence is reducing to 80%. The best results can be viewed for this algorithm when the min_conf is between 85%-100%.

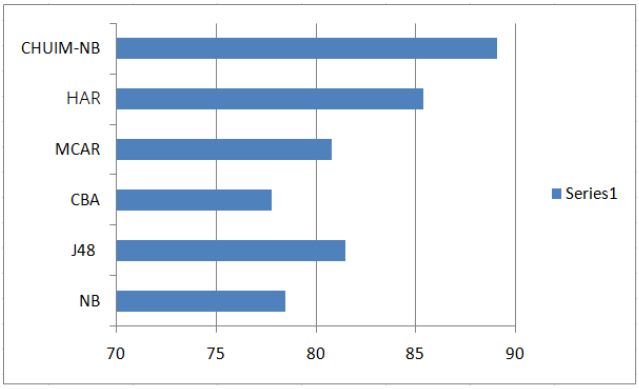
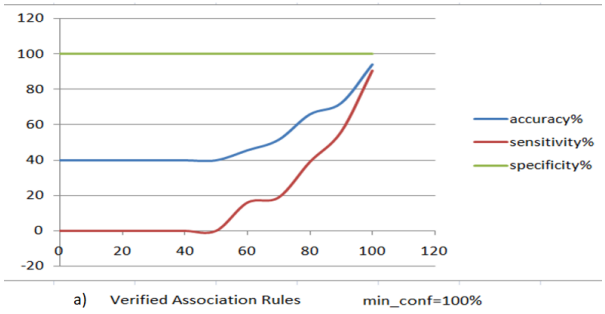
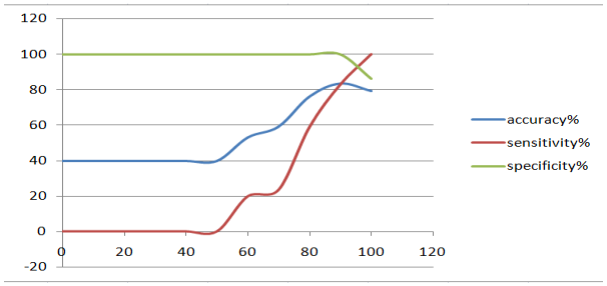


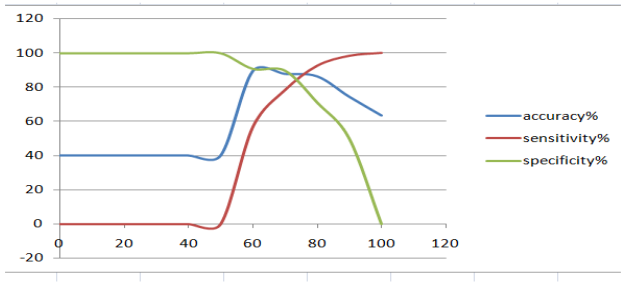
Fig 4: Classification Accuracy % and Comparison algorithms



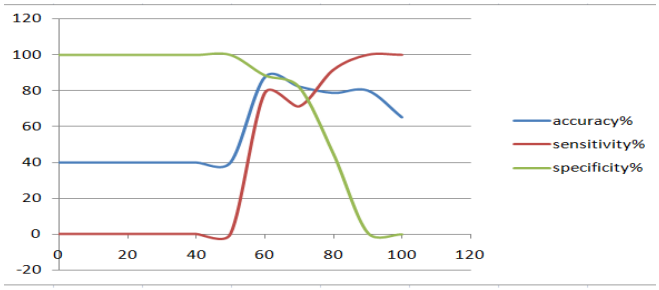
a) Verified Association Rules min_conf=100%



b) Verified Association rules with min_conf=95%



c) Verified Association Rules min_conf=85%



d) Verified association rules min_conf=80%

Fig 5: #verified association rules with various min_conf% a) with 100%, b)with 95%, c)with 85% and d)with 80%

Note: the graph is actually made for #verified rules from 60 to 100 and made the values static for accuracy % as 40%, sensitivity% as 0% and specificity as 100% from #verified association rules from 0 to 50, to visualize the graph better.

VII. CONCLUSION

The integrated association & classification methods that have evolved in the recent times (like MCAR, HAR) to effectively prune the unwanted rules and search for the effective class label for the test data. Our CHUIM-NB has been performing better than the above said methods due to its list structure search spacing & pruning mtechniques for finding the best rules and is very effective even with the memory usage and execution time, especially for dense and categorical datasets.

ACKNOWLEDGMENT

This work has been supported by “KRISHNA IVF CLINIC RESEARCH LAB, VISHAKAPATNAM” for providing the real datasets.

REFERENCES

1. Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).
2. Srikant, R., & Agrawal, R. (1996, June). Mining quantitative association rules in large relational tables. In *Acm Sigmod Record* (Vol. 25, No. 2, pp. 1-12). ACM.
3. Yoda, K., Fukuda, T., Morimoto, Y., Morishita, S., & Tokuyama, T. (1997, August). Computing Optimized Rectilinear Regions for Association Rules. In *KDD* (Vol. 97, pp. 96-103).
4. Liu, B., Hsu, W., & Chen, S. (1997, August). Using General Impressions to Analyze Discovered Classification Rules. In *KDD* (pp. 31-36). Tseng, V. S., Shie, B. E., Wu, C. W., & Philip, S. Y. (2013).
5. Efficient algorithms for mining high utility itemsets from transactional databases. *IEEE transactions on knowledge and data engineering*, *25*(8), 1772-1786.
6. Tseng, V. S., Wu, C. W., Shie, B. E., & Yu, P. S. (2010, July). UP-Growth: an efficient algorithm for high utility itemset mining. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 253-262). ACM.
7. Liu, Y., Liao, W. K., & Choudhary, A. (2005, May). A two-phase algorithm for fast discovery of high utility itemsets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*(pp. 689-695). Springer, Berlin, Heidelberg.
8. Yao, H., Hamilton, H. J., & Butz, C. J. (2004, April). A foundational approach to mining itemset utilities from databases. In *Proceedings of the 2004 SIAM International Conference on Data Mining* (pp. 482-486). Society for Industrial and Applied Mathematics.
9. Ahmed, C. F., Tanbeer, S. K., Jeong, B. S., & Lee, Y. K. (2009). Efficient tree structures for high utility pattern mining in incremental databases. *IEEE Transactions on Knowledge and Data Engineering*, *21*(12), 1708-1721.
10. Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.
11. Hunt, E. B., Marin, J., & Stone, P. J. (1966). Experiments in induction.
12. Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. *Expert systems in the micro electronics age*.
13. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.
14. Breiman, L., Friedman, J. H., & Olshen, R. A. (2009). Stone, cj (1984) classification and regression trees. *Wadsworth, Belmont, California*.
15. Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
16. Vapnik, V. (1995). The nature of statistical learning theory Springer New York Google Scholar.
17. Ma, B. L. W. H. Y., & Liu, B. (1998, August). Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*.
18. Soni, S., & Vyas, O. P. (2010). Using associative classifiers for predictive analysis in health care data mining. *International Journal of Computer Applications*, *4*(5), 33-37.
19. Li, W., Han, J., & Pei, J. (2001, November). CMAR: Accurate and efficient classification based on multiple class-association rules. In *icdm* (p. 369). IEEE.
20. Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, *15*(1), 55-86.
21. 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.
22. Hadi, W. E., Al-Radaideh, Q. A., & Alhawari, S. (2018). Integrating associative rule-based classification with Naïve Bayes for text classification. *Applied Soft Computing*, *69*, 344-356.
23. Liu, M., & Qu, J. (2012, October). Mining high utility itemsets without candidate generation. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 55-64). ACM.
24. Fournier-Viger, P., Wu, C. W., Zida, S., & Tseng, V. S. (2014, June). FHM: Faster high-utility itemset mining using estimated utility co-occurrence pruning. In *International symposium on methodologies for intelligent systems* (pp. 83-92). Springer, Cham.
25. Fournier-Viger, P., Lin, J. C. W., Wu, C. W., Tseng, V. S., & Faghihi, U. (2016, September). Mining minimal high-utility itemsets. In *International Conference on Database and Expert Systems Applications* (pp. 88-101). Springer, Cham.
26. GKovacs, G. T. (1999). What factors are important for successful embryo transfer after in-vitro fertilization?. *Human Reproduction*, *14*(3), 590-592.
27. Raju, G. A. R., Haranath, G. B., Krishna, K. M., Prakash, G. J., & Madan, K. (2005). Vitrification of human 8-cell embryos, a modified protocol for better pregnancy rates. *Reproductive biomedicine online*, *11*(4), 434-437.
28. Templeton, A., Morris, J. K., & Parslow, W. (1996). Factors that affect outcome of in-vitro fertilisation treatment. *The Lancet*, *348*(9039), 1402-1406.
29. Sahoo, J., Das, A. K., & Goswami, A. (2015). An efficient approach for mining association rules from high utility itemsets. *Expert systems with Applications*, *42*(13), 5754-5778.
30. Sharma, T. C., & Jain, M. (2013). WEKA approach for comparative study of classification algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(4), 1925-1931.
31. Arlene J. Morales, (2011). Understanding Infertility, evaluation and treatment options,Fertility Specialists medical Group, Inc, A presentation.

AUTHORS PROFILE

Suvarna U Research scholar, Gitam University,Vishakapatnam, AP, India
Email: viksonio@gmail.com

Srinivas Y Professor, Gitam University, Vishakapatnam, AP, India
Email: sriteja.y@gmail.com