American Journal of Computer Science and Information Engineering
2015; 2(1): 1-6
Published online March 20, 2015 (http://www.aascit.org/journal/ajcsie)

AASCIT American Association for Science and Technology

Keywords
Breast Cancer Database,
Distributed Database,
Association Rule Mining,
Cryptography Algorithm,
Privacy

Received: February 20, 2015
Revised: March 5, 2015
Accepted: March 6, 2015

Breast Cancer Prediction Through DHBCA Algorithm in Horizontal Partitioned Database

Raghvendra Kumar[1], Prasant Kumar Pattnaik[2], Yogesh Sharma[1]

[1]Faculty of Engineering Technology, Jodhpur National University, Jodhpur, Rajsthan, India
[2]School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India

Email address

raghvendraagrawal7@gmail.com (R. Kumar), patnaikprasant@gmail.com (P. K. Pattnaik), yogeshsharma@gmail.com (Y. Sharma)

Citation

Raghvendra Kumar, Prasant Kumar Pattnaik, Yogesh Sharma. Breast Cancer Prediction Through DHBCA Algorithm in Horizontal Partitioned Database. *American Journal of Computer Science and Information Engineering*. Vol. 2, No. 1, 2015, pp. 1-6.

Abstract

According to the world health organization (WHO) cancer will became leading cause of woman death of the world wide. Breast cancer has become the most hazardous type of cancer among woman in the world. Early detection of breast cancer is essential in reducing life losses. In this paper mining rules to mine the attributes relationship. The support and confidence value are used to estimated from all item set. Minimum support and confidence value are used to predict rule from the list datasets. Our proposed work is applicable where the numbers of sites are greater than two and each site want to calculate the global rule and global confidence from the database without disclosing their private information to other sites presents in the global environments. Our proposed algorithm Double Hash Based Cryptography Algorithm (DHBCA) to provide the highest privacy in the breast cancer database and also zero percent of data leakage.

1. Introduction

Data mining [8], also known as knowledge discovery in databases is defined as "the extraction of implicit, previously unknown, and potentially useful information from data" [9, 10, 11]. It encompasses a set of processes performed automatically, whose task is to discover and extract hidden features from large datasets. Rule mining [7, 8] is one of the most studied problems in machine learning and data mining. Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. The goal of the global rule mining algorithms is to construct a model from a set of training data whose target class labels are known and then this model is used to mine unseen instances. The Rule mining of Breast Cancer [1, 2, 3, 4, 5] data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors. Breast cancer is one of the most common cancers among women. Breast cancer is one of the major causes of death in women when compared to all other cancers. Cancer is a type of diseases that causes the cells of the body to change its characteristics and cause abnormal growth of cells. Most types of cancer cells eventually become a mass called tumor. The occurrence of breast cancer is increasing globally. It is a major health problem and represents a significant worry for many women. Early detection of breast cancer is essential in reducing life losses. However earlier treatment requires the ability to detect breast cancer in early stages. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones. The automatic diagnosis of breast cancer is an important, real

world medical problem. Thus, finding an accurate and effective diagnosis method is very important. In recent years machine learning methods have been widely used in prediction, especially in medical diagnosis. Medical diagnosis is one of major problem in medical application. The Rule mining of Breast Cancer [6, 7] data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors. A major mining of problems in medical science involves the diagnosis of disease, based upon various tests performed upon the patient. For this reason the use of Rule mining in medical diagnosis is gradually increasing when consider the database that are distributed among different sites in the distributed environment.

2. Related Work

Numerous works in literature related with cancer disease diagnosis uses data mining techniques have. This paper presents a novel and more efficient DHBCA algorithm [7, 8, 9]. It analyzes and considers Apriori algorithm [8]. It's the algorithm to mine association rules. Breadth-first search [8] strategy is used to counting the support of item sets, confidence of the item set and uses a candidate generation function which exploits the downward closure property of support and confidence.

Apriori algorithm:

1. Generating item sets that pass a minimum support threshold.
2. Generating rules that pass a minimum confidence threshold.
3. Bottom up approach is used, where frequent subsets are extended one item at a time a step known as Candidate generation. When no further successful extensions are found, it terminates.
4. Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently.

Apriori algorithm gets large frequent item sets through the combination and pruning of small frequent item sets. The principle of the algorithm is: firstly calculates the support of all item sets in candidate item set Ck obtained by Lk-1, the support of the item set is greater than or equal to the minimum support, frequent k-item set considered as the candidate k-item set, that is Lk, then all frequent k-item sets combined into a new candidate item set Ck+1, level by level, until finds large frequent item sets.

Three constraints were introduced to decrease the number of patterns.

1. Necessary attributes to appear on only one side of the rule.
2. It segregates attributes into uninteresting groups.
3. Number of attributes in a rule is restricted.

Two groups of rules envisaged the presence or absence of cancer disease in four specific cancer arteries.

Data mining methods may aid the clinicians in the predication of the survival of patients and in the adaptation of the practices consequently. Early breast cancer can in some cases present as breast pain or a painful lump. Breast cancer is most frequently discovered as an asymptomatic nodule on a mammogram A lump under the arm or above the collarbone that does not go away may be present. When breast cancer has invaded the dermal lymphatic's small lymph vessels of the skin, it can resemble skin inflammation is known as inflammatory breast cancer where it is blocking lymphatic vessels and this can cause some symptoms around the breast, as well as an orange peel texture to the skin referred to as paid' orange. It may also have been no previous signs of breast cancer and the cancer might be missed in screening mammograms. Changes in the appearance or shape of the breast can raise suspicions of breast cancer. One of the symptoms of breast cancer is Paget's disease of the breast. It presents as eczematous skin changes at the nipple and it is a late manifestation of an underlying breast cancer. Some breast symptoms do not turn out to represent underlying breast cancer. Some breast diseases such as fibrocystic mastopathy, functional mastodynia, and mastitis and fibro adenoma of the breast are more common causes of breast symptoms. A new breast symptom should be taken seriously by both patients and their doctors by the possibility of an underlying breast cancer at almost any age. Occasionally, breast cancer presents as metastatic disease i.e. cancer that has spread beyond the organ. Metastatic breast cancer will cause symptoms that depend on the location of metastasis. Some common sites of metastasis include bone, liver, lung, and brain. Weight loss can occasionally herald an occult breast cancer as symptoms of fevers or chills. Joint pains or bone can sometimes be manifestations of metastatic breast cancer as jaundice or neurological symptoms. The uncommon symptom with metastatic breast cancer is Pleural effusions. Since these symptoms can also be manifestations of many other illnesses. The extraction of significant patterns from the cancer disease data warehouse is presented in this section. The cancer disease data warehouse contains the screening clinical data of cancer patients. Initially, the data warehouse is preprocessed to make the mining process more efficient. In our proposed study, it uses preprocessing in order to handle missing values. Then applying equal interval binning with approximate values based on medical expert advice to Pima Indian breast cancer data [9, 10, 11]. DHBCA algorithm is applied to generate the rules. Also consider important measure confidence. Calculating the significant items for all frequent patterns with the aid of the approach proposed. The frequent patterns are selected with confidence greater than predefined threshold. These frequent patterns can be used in the design and development of breast cancer prediction system.

Pruning-Classification Association Rule DHBCA. DHBCA Combines minimum frequency items with minimum frequency item sets. It first deletes infrequent items from item sets and then it classifies those item sets based on the frequency of item sets to discover frequent item sets. The candidate item sets counts are greatly

reduced and item sets need not to be combined or decomposed then the operation time and memory requirement could be decreased accordingly. It has significant advantage in mining association rule at large volumes of items and small frequency of item sets. Association rules are nothing different from classification rules except that does not predict only class labels but also predict any other attribute. It is used to produce a combination of attributes. Those different association rules convey different regularities that trigger in the dataset and generally predict the different things and so many association rule generated from even the data set is small. By keeping such rules which are applicable reasonably large number of instances based on coverage and accuracy criteria. An association rule is the number of instances for which it predicts correctly this is often called its support. Confidence is the number of instances that it predicts correctly and expressed as proportion of all instances to which it applies called accuracy. The user have to specify the minimum coverage and accuracy values and look for only those rules whose values are at least of the specified minimum value.

3. Proposed Work

The cancer disease severity prediction system is tested using the breast cancer diagnosis datasets. The dataset is downloaded from the UCI (University of California) [3, 4, 5] machine learning repository [4]. It provides information about the breast cancer patient diagnosis information [11, 12]. The class information and associated symptom details are provided in the dataset. The dataset is also constructed with noise records. Rule mining process is performed on the data sets. The dataset attribute details are given in table. Rule mining operations can be tested using the dataset when the database is distributed among different sites and each site want to collaborate there global result. Assumption for the proposed work are taken as the breast cancer database is horizontally partitioned and distributed among sites and the total number of sites is greater than or equal to four ($n \geq 4$). The sites are considered as trusted site and all the site contain their own private data and no other site will be able to know other site data .In this method, basically, hash based secure sum technique has been used. In secure sum each site will determine their own data value and send to predecessor site that near to original site and this goes on till the original site collects all the value of data after that the parent site will determine the global confidence and it also not necessary that the result found is globally frequent or infrequent depending on value which will create after collecting all the value. We have considered four sites S_1, S_2, S_3, S_4 where the sites are interchanging its position with another by following the algorithm. The secure sum techniqueis based on changing neighbors in each round of segment computation. The number of the site S_1 is selected as the initiator site which starts the computation by distributing the data segment. The number of parties for this technique must be four or more. When all the rounds of segments summation are completed the sum is announced by the forthcoming site. The steps are as follows. This algorithm applicable when the number of sites greater than or equal to 4 ($n \geq 4$) and all the sites are arranged in the ring topology format. Table1 Contains the list of attributes and their description and Table2, Table3, Table4 and Table5 are shows the breast cancer databases that are distributed in horizontal partitioned manner among the different sites, in horizontal partition only transaction is divided.

Table 1. *Attributes and their description*

S/No	Attribute Name	Description
1	Pid	Patient Identifier Number
2	Ct	Clump Thickness
3	UC Size	Uniformity of Cell Size
4	UC Shape	Uniformity of Cell Shape
5	Ma	Marginal Adhesion
6	Sece	Single Epitaxial Cell Size
7	Bn	Bare Nuclei
8	Bc	Bland Chromatic
9	Nn	Normal Nuclei
10	M	Mitoses
11	Class	Class

Table 2. *Breast cancer database for site1*

T/No	Item Set
1	Pid, Ct, UC Size, Nn, Class
2	Pid, UC Shape, M, Sece, Bn, Bc
3	M, Ma, Ct, Pid
4	Class, Sece, Ma
5	UC Shape, UC Size
6	Nn, M, Sece
7	Pid, Ct, Bn
8	Bc, Pid, M
9	Sece, Pid, UC Shape, UC Size

Table 3. *Breast cancer database for site2*

T/No	Item Set
1	Ct, UC Size, Nn, Class
2	Pid, UC Shape, M, Sece
3	M, Ma, Ct, Pid, Bn, Bc
4	Class, Sece, Ma
5	UC Shape, UC Size, , Bn, Bc
6	Nn, M, Sece, UC Size
7	Pid, Ct, Bn, Bc
8	Bc, Pid, M, , UC Size
9	Sece, Pid, UC Shape, UC Size

Table 4. *Breast cancer database for site3*

T/No	Item Set
1	Pid, Ct, UC Size, Nn, Class
2	Pid, UC Shape, M, Sece, Bn, Bc, Nn, Class
3	M, Ma, Ct, Pid, Nn, Class
4	Class, Sece, Ma
5	UC Shape, UC Size, Nn, Class
6	Nn, M, Sece, UC Shape, UC Size
7	Pid, Ct, Bn, UC Shape, UC Size
8	Bc, Pid, M
9	Sece, Pid, UC Shape, UC Size

Table 5. *Breast cancer database for site4*

T/No	Item Set
1	Pid, Ct, UC Size, Nn, Class
2	Pid, UC Shape, M, Sece, Bn, Bc, UC Size, Nn
3	M, Ma, Ct, Pid, , UC Size, Nn
4	Class, Sece, Ma, UC Size, Nn,
5	UC Shape, UC Size
6	Nn, M, Sece, UC Size
7	Pid, Ct, Bn
8	Pid, Bc, Pid, M
9	Sece, Pid, UC Shape, UC Size

4. Double Hash Based Cryptography Algorithm (DHBCA)

For this we are using one of method to find the global result of confidence [8].

There is mainly number of steps to find the global confidence.

Step1:- Each site will calculate their frequent item sets and infrequent item sets and store the data value on memory.

Step2:- Each site will generate their own random number because we are using hash based secure sum protocol so that each site have two random number one of its own and other is received by previous site.

Step3:- Now the site 1 will calculate the partial confidence value by using the following formula.

PC_j= Support(X and Y) – Minimum Confidence*Support (X) *|DB|+RN1-RNn+Mask Value /* where RN is random number*/.

After that site1 calculate the mask value

$PC_j = PC_j$ + Mask Value

Step4:- Site 2 compute the PC_j for each item received the list using the formula

$PC_{j=}$ PC_j + Support(X and Y) – Minimum Confidence*Support (X) *|DB|+Rn1-Rn (i-1)

Step5:- After that the value of PCj calculated by site 2 send to next coming site and after that all the value is send to the original site and that original site will calculate all global Confidence.

Step6:- Site 1 will find whether that global Confidence is grater then zero or not if the value is grater then zero then it will be global frequent rule otherwise is infrequent.

Step7:- Like that the entire site will calculate the will calculate the global confidence site 3 site 4 site 5…….site n.

Step8:- Finally the site 1 calculate the global confidence by using the formula

Global Confidence (GC) = Partial Confidence-Mask Key

Step9:- At last the site 1 will send the calculated value of global confidence and global frequent item set to all other site in the horizontal partition.

Step10:- Starting site broadcast the global confidence value to all the sites presents in the network.

Support count PId=6, Support count Ct=3, Support count UC Size=3, Support count UC Shape=3, Support count Ma=1, Support count Sece=4, Support count Bn=1, Support count Bc=2, Support count Nn=2, Support count M=4, Support count Class=2. Support PId=6/9=0.66, Support Ct=3/9=0.33, Support UC Size=3/9=0.33, Support UC Shape=3/9=0.33, Support Ma=1/9=0.11, Support Sece=4/9=0.44, Support Bn=1/9=0.11, Support Bc=2/9=0.22, Support Nn=2/9=0.22, Support M=4/9=0.44, Support Class=2/9=0.22. Consider the minimum support 40% then select those item set whose support greater than or equal to that minimum support. Then selected number of item sets are {Pid, Sece, M}, then consider two item sets, Then support (Pid and Sece)=2, support (Pid and M)=3, support (Sece and M)=2, then support (Pid and Sece)=2/9=0.22, support (Pid and M)=3/9=0.33, support (Sece and M)=2/9=0.22. All the support count value is less than the considered value so neglect all the two pair item sets. Now calculate confidence, confidence= support (Pid and Sece)/ support (Pid) =2/6=0.33*100=33%, confidence=support (Pid and M)/ support (Pid)=3/6=0.50*100=50%, confidence=support (Sece and M)/ support (Sece)=2/4=0.50*100=50%.

Support count PId=5, Support count Ct=3, Support count UC Size=4, Support count UC Shape=3, Support count Ma=1, Support count Sece=4, Support count Bn=3, Support count Bc=4, Support count Nn=2, Support count M=4, Support count Class=2. Support PId=5/9=0.55, Support Ct=3/9=0.33, Support UC Size=4/9=0.44, Support UC Shape=3/9=0.33, Support Ma=1/9=0.11, Support Sece=4/9=0.44, Support Bn=3/9=0.33, Support Bc=4/9=0.44, Support Nn=2/9=0.22, Support M=4/9=0.44, Support Class=2/9=0.22. Consider the minimum support 40% then select those item set whose support greater than or equal to that minimum support. Then selected number of item sets are {Pid, UC Size, Sece, Bc M}, then consider two item sets, Then support count (Pid and UC Size)=2, support count (Pid and Sece)=2, support count (Pid and Bc)=3, support count (Pid and M)=3, support count (UC Size and Sece)=2, support count (UC Size and Bc)=2 support count (UC Size and M)=2, support count (Sece and Bc)=0, support count (Sece and M)=2, support count (Bc and M)=2 then support (Pid and UC Size)=2/9=0.22, support (Pid and Sece)=2/9=0.22, support (Pid and Bc)=3/9=0.33, support (Pid and M)=3/9=0.33, support (UC Size and Sece)=2/9=0.22, support (UC Size and Bc)=2/9=0.22, support (UC Size and M)=2/9=0.22, support (Sece and Bc)=0/9=0.00, support (Sece and M)=2/9=0.22, support (Bc and M)=2/9=0.22. All the support count value is less than the considered value so neglect all the two pair item sets. Now calculate the confidence of each rule, Confidence=support (Pid

and UC Size)/ support (Pid) =2/5=0.40*100=40%, confidence=support (Pid and Sece)/ support (Pid) =2/5=0.40*100=40%, confidence=support (Pid and Bc)/ support (Pid) =3/5=0.60*100=60%, confidence=support (Pid and M)/ support (Pid) =3/5=0.60*100=60%, confidence=support (UC Size and Sece)/ support (UC Size) =2/4=0.50*100=50%, confidence=support (UC Size and Bc)/ support (UC Size) =2/4=0.50*100=50%, confidence=support (UC Size and M)/ support (UC Size) =2/4=0.50*100=50%, confidence=support (Sece and Bc)/ support (Sece) =0/4=0%, confidence=support (Sece and M)/ support (Sece)=2/4=0.50*100=50%, confidence=support (Bc and M)/ support (Bc) =2/4=0.50*100=50%.

Support count PId=6, Support count Ct=3, Support count UC Size=5, Support count UC Shape=5, Support count Ma=2, Support count Sece=4, Support count Bn=2, Support count Bc=2, Support count Nn=5, Support count M=4, Support count Class=5. Support PId=6/9=0.66, Support Ct=3/9=0.33, Support UC Size=5/9=0.55, Support UC Shape=5/9=0.55, Support Ma=2/9=0.22, Support Sece=4/9=0.44, Support Bn=2/9=0.22, Support Bc=2/9=0.22, Support Nn=5/9=0.55, Support M=4/9=0.44, Support Class=5/9=0.55. Consider the minimum support 40% then select those item set whose support greater than or equal to that minimum support. Then selected number of item sets are {Pid, UC Size, UC Shape, Sece,Nn, M, Class}, then consider two item sets, Then support count (Pid and UC Size)=3, support count (Pid and UC Shape)=3, support count (Pid and Sece)=2, support count (Pid and Nn)=4, support count (Pid and M)=3, support count (Pid and Class)=3, support count (UC Size and UC Shape)=4, support count (UC Size and Sece)=2, support count (UC Size and Nn)=2, support count (UC Size and M)=1, support count (UC Size and Class)=2, support count (UC Shape and Sece)=3, support count (UC Shape and Nn)=3, support count (UC Shape and M)=2 support count (UC Shape and Class)=2, support count (Sece and Nn)=2, support count (Sece and M)=2, support count (Sece and Class)=2, support count (Nn and M)=3, support count (Nn and Class)=4, support count (M and Class)=1, then support (Pid and UC Size)=3/9=0.33, support (Pid and UC Shape)=3/9=0.33, support (Pid and Sece)=2/9=0.22, support (Pid and Nn)=4/9=0.44, support (Pid and M)=3/9=0.33, support (Pid and Class)=3/9=0.33, support (UC Size and UC Shape)=4/9=0.44, support (UC Size and Sece)=2/9=0.22, support (UC Size and Nn)=2/9=0.22, support (UC Size and M)=1/9=0.11, support (UC Size and Class)=2/9=0.22, support (UC Shape and Sece)=3/9=0.33, support (UC Shape and Nn)=3/9=0.33, support (UC Shape and M)=2/9=0.22,support (UC Shape and Class)=2/9=0.22, support (Sece and Nn)=2/9=0.22, support (Sece and M)=2/9=0.22, support (Sece and Class)=2/9=0.22, support (Nn and M)=3/9=0.33, support (Nn and Class)=4/9=0.44, support (M and Class)=1/9=0.11. All the support count value is greater than the considered value, so these are the list of attributes to select for two pair {support (UC Size and UC Shape), support (Pid and Nn), support (Nn and Class)}. Confidence=support (Pid and UC Size)/ support (Pid) =3/6=0.50*100=50%, confidence=support (Pid and UC Shape)/ support (Pid) =3/6=0.50*100=50%, confidence=support (Pid and Sece)/ support (Pid) =2/6=0.33*100=33%, confidence=support (Pid and Nn)/ support (Pid) =4/6=0.66*100=66%, confidence=support (Pid and M)/ support (Pid) =3/6=0.50*100=50%, confidence=support (Pid and Class)/ support (Pid) =3/6=0.50*100=50%, confidence=support (UC Size and UC Shape)/ support (UC Size)=4/5=0.8*100=80%,confidence= support (UC Size and Sece)/ support (UC Size) =2/5=0.40*100=40%, confidence=support (UC Size and Nn)/ support (UC Size) =2/5=0.40*100=40%, confidence=support (UC Size and M)/ support (UC Size) =1/5=0.20*100=20%, confidence=support (UC Size and Class)/ support (UC Size) =2/5=0.40*100=40%, confidence=support (UC Shape and Sece)/ support (UC Shape)=3/5=0.60*100=60%, confidence=support (UC Shape and Nn)/ support (UC Shape) =3/5=0.60*100=60%, confidence=support (UC Shape and M)/ support (UC Shape) =2/5=0.40*100=40%, confidence=support (UC Shape and Class)/ support (UC Shape)=2/5=0.40*100=40%, confidence=support (Sece and Nn)/ support (Sece) =2/4=0.50*100=50%, confidence=support (Sece and M)/ support (Sece) =2/4=0.50*100=50%, confidence=support (Sece and Class)/ support (Sece)=2/4=0.50*100=50%, confidence=support (Nn and M)/ support (Nn) =3/5=0.60*100=60%, confidence=support (Nn and Class)/ support (Nn)=4/5=0.80*100=80%, confidence=support (M and Class)/ support (M)=1/4=0.25*100=25%.

Support count PId=6, Support count Ct=3, Support count UC Size=7, Support count UC Shape=3, Support count Ma=2, Support count Sece=3, Support count Bn=2, Support count Bc=2, Support count Nn=5, Support count M=4, Support count Class=2. Support PId=6/9=0.66, Support Ct=3/9=0.33, Support UC Size=7/9=0.77, Support UC Shape=3/9=0.33, Support Ma=2/9=0.22, Support Sece=3/9=0.33, Support Bn=2/9=0.22, Support Bc=2/9=0.22, Support Nn=5/9=0.55, Support M=4/9=0.44, Support Class=2/9=0.22. Consider the minimum support 40% then select those item set whose support greater than or equal to that minimum support. Then selected number of item sets are {Pid, UC Size, Nn, M}, then consider two item sets, Then support (Pid and UC Size)=3, support (Pid and Nn)=3, support (Pid and M)=3, support (UC Size and Nn)=5, support (UC Size and M)=2, support (Nn and M)=3, then support(Pid and UC Size)=3/9=0.33, support (Pid and Nn)=3/9=0.33, support(Pid and M)=3/0.99, support(UC Size and Nn)=5/9=0.55, support (UC Size and M)=2/9=0.22, support (Nn and M)=3/9=0.33. All the support count value is greater than the considered value, so these are the list of attributes to select for two pair {support(UC Size and Nn)}. Now calculate the confidence, confidence=support (Pid and UC Size)/ support (Pid)=3/6=0.50*100=50%, confidence=support (Pid and Nn)/ support (Pid)=3/6=0.50*100=50%, confidence=support (Pid and M)/ support (Pid)=3/6=0.50*100=50%, confidence=support (UC Size and Nn)/Support(UC Size)=5/7=0.71*100=71%,

Confidence=support (UC Size and M)/ Support(UC Size)=2/7=0.28*100=28%, confidence=support (Nn and M)/ support (Nn)=3/5=0.6*100=60%, then support(Pid and UC Size)=3/9=0.33, support (Pid and Nn)=3/9=0.33, support(Pid and M)=3/0.99, support(UC Size and Nn)=5/9=0.55, support (UC Size and M)=2/9=0.22, support (Nn and M)=3/9=0.33, now select those attributes whose confidence is greater than or equal to 50% then the number of selected items confidence=support (Pid and Nn)/ support (Pid)=3/6=0.50*100=50%, confidence=support (Pid and M)/ support (Pid)=3/6=0.50*100=50%, confidence=support (UC Size and Nn)/Support(UC Size)=5/7=0.71*100=71%, confidence=support (Nn and M)/ support (Nn)=3/5=0.6*100=60%.

Cryptography technique use to minimize the loss of information and privacy. So here we use double hash based technique to provide the security to the distributed database. After finding the item sets frequent or infrequent, but we are applying this technique for both frequents and infrequent item sets. Then we use the double hash function is used to make the association rule more secure so following method is used to provide the security to the database using following formula to calculate the mask value, Mask value is calculated by using two different hash functions, Key1=Hash (key) =key mod N, where Key1 security key, And after that Mask key has calculated which store value after hashing, Mask key = Hash2 (Key1) =Key+M^{key1}, In this we are using three horizontal partition distributed databases for finding privacy preserving association rule mining when no party is considered as a trusted party. There are three existing site1, site 2, site3 and site4, and posses' different database DB1, DB2, DB3 and DB4 respectively. And the minimum confidence is 50 % for all the sites. Let us consider Key =110, M=2, Hash key=key mod M, Mask key=hash key- M^{key}, Hash key=110 mod 2=0, Mask key=110-2^0=109, consider the following rule to calculate the global confidence of the rule (Pid →UC Size), consider the random number of site1, site2, site3 and site4 are 1,2 3, and 4, and all the sites are arrange in ring topology network. Local confidence for site1, PC1= Support (Pid →UC Size) – Minimum Confidence*Support (Pid) *|DB|+RN1-RNn+Mask Value, PC1= 0– 0.50*6 *9+1-4+109=0-27+1-4+109=79,PC2=2-0.50*5*9+2-1+79=59.5,PC3=3-0.50*6*9+3-2+59.5=36.5,

PC4=3-0.50*6*9+4-3+36.5=13.5. Global Confidence (GC) = Partial Confidence-Mask Key=13.5-10.9=-95.5, in this case the global confidence is less than zero it means that this rule is globally infrequent rule but may be that is locally frequent rule and we are not able to say that this is the rule for the breast cancer. Preserving privacy in association rule mining when the database is distributed horizontally among n (n>2) number of sites when no trusted party is considered. A replica which adopts a hash based secure sum cryptography technique to find the global association rules is propose by preserving the privacy constraints. Double hashing function is adopted to enhance the privacy further. The proposed replica capably finds global frequent rule sets even when no site can be treated as trusted.

5. Conclusion

Rule mining has become a significant tool for knowledge discovery. The association rule mining and cryptography techniques are integrated under the global rule mining process. The weighted association rule mining is carried out on the data with class labels. By using DHBC Algorithm, the system is designed to predict cancer severity levels inside the woman breast when the data distributed among the different sites. The system performs weighted rule mining on labeled data values. Symptom based weight assignment is performed. Global Confidence based filtering model. Disease and its severity level are globally predicted with highest privacy and also zero percent of data leakage.

References

[1] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

[2] A.Bellachia and E.Guvan,"Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006.

[3] A. Endo, T. Shibata and H. Tanaka (2008), Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computing and Human Sciences, vol.13, pp.11-16.

[4] Breast Cancer Wisconsin Data [online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancerwisconsin/ breast-cancer-wisconsin.data.

[5] Brenner, H., Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. Lancet. 360:1131–1135, 2002.

[6] D. Delen, G. Walker and A. Kadam (2005), Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, vol.34, pp.113-127.

[7] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco: Morgan Kaufmann; 2005.

[8] J. Han and M. Kamber, Data Mining-Concepts and Technique (The Morgan Kaufmann Series in Data Management Systems), 2nd ed. San Mateo, CA: Morgan Kaufmann, 2006.

[9] Mitchell, T. M., Machine Learning, McGraw-Hill Science/Engineering/Math, 1997

[10] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[11] Razavi, A. R., Gill, H., Ahlfeldt, H., and Shahsavar, N., Predicting metastasis in breast cancer: comparing a decision tree with domain experts. J. Med. Syst. 31:263–273, 2007.

[12] S.B.Kotsiantis and P.E.Pintelas,"Combining Bagging and Boosting", International Journal of Information and Mathematical Sciences, 1:4 2005.