Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7751244

A COGENCY BASED APPROACH FOR DETERMINING APPROPRIATE COMBINATION OF MEDICAL ATTRIBUTES IN BREAST CANCER SURVIVABILITY

Ms. Sheethal Aji Mani
Research Scholar, Department of Computer Science, Presidency University
Bangalore, India, Email: sheethalabraham@gmail.com

Prof. Dr. T. K Thivakaran
Professor, Department of Computer Science, Presidency University
Bangalore, India, Email: tktpresidency4@gmail.com

One of the most common types of cancer that affects women is breast cancer. The early detection is very important for this disease. Eventhough genetic mutations and family history are associated with the high risk of developing breast cancer there can be other factors also which decide the chances for the breast cancer in women. For eg. aggreviated levels of CEA(Carcino embrionic Antigen) can be a pre-indication to the cancer and is associated with metastatic disease. However with a single biomarker or a medical attribute, the diagnosis may not be done appropriately and a combination of biomarkers or attributes related to the disease can certainly contribute in the prediction of breast cancer in an early stage. The selection of the appropriate combination of these medical attributes is a challenge as the knowledge about the best combination of these attributes will help the medical field experts and doctors to focus more on these findings and determine treatments based on these findings.

Various association rule mining techniques like classic Apriori Algorithm are endorsed to explore the best combinations of these attributes. But one of the major disadvantages of these traditional algorithms are,dataset need to be passed more than two times or atleast two times to identify the repeating and significant attributes. Therefore a cogency based association rule mining algorithm called CbARM algorithm is proposed and applied to the SEER dataset 2020(Nov). The concept of cogency refers to the likelihood that the presumed facts will be true if the conclusion is correct. One of the advantages of the proposed method is that it can be used to generate rules with just a single pass of the file. This is made possible by the construction of a knowledge link matrix. It is found that compared to Apriori algorithm more suitable combinations of medical attributes in terms of rules have been generated , once the CbARM algorithm is applied.

Keywords: *Breast cancer, Survivability, Association Rule Mining*

1. Introduction

Breast cancer is the most common type of cancer that affects people. It accounts for over 500,000 deaths annually. Due to the increasing number of women diagnosed with this disease, it has become the most common type of cancer in women [1]. Despite the various factors that can affect the development of the disease, breast cancer is still very treatable. Unfortunately, many of the deaths caused by this disease are due to the late detection. The economic and social factors that can influence the survival rate of patients can also be identified [2]. This can help improve the survival rate of cancer patients. If the medical experts find a combination of these factors that can increase the survival rate of patients, then the treatment options for these patients can be extended.

The concept of data refers to the gathered statistics and information that are used for further studies and references. This process can help create new theories and statements and it is very beneficial for various applications. For instance, if a market basket analysis is conducted, it can be very beneficial to analyze the patterns of buying behavior. This method can help the business develop new strategies and improve the efficiency of its operations.

Data mining is a process that involves analyzing the vast amount of information that is collected by an organization. This process is very beneficial for the development of new strategies and the efficiency of its operations. Since it is a complex process, there are various sub domains that are involved in the process. One of these is the frequent and infrequent mining. This process is commonly used to find hidden items in a dataset. The frequency of certain item sets is considered as a criterion that determines whether they should be considered along with regular or rare item sets. If the frequency of these items exceeds a certain threshold, then they should be added along with the regular or frequently appearing items.

Recognizing the rare or frequent item set is a significant step in the process, thereafter the next step is to perform association rule mining. This process involves finding out more useful rules that can help us reach a conclusion. A rule is a type of statistical procedure that can be used to analyze a group's buying behavior.

The antecedent and consequent of a rule are respectively shown in Table I. For instance, the six transactions shown in the table are composed of six different customers. While the fact is the antecedent, while the consequent is the outcome.

In the market basket, the transactions include various types of products such as Sausage, ghee, Mangoes, and Cookies. With the help of these sets, we can then extract rules that are helpful to analyze the purchasing behavior of a person.

For instance, the rule Mangoes → Sausage describes the occurrences where both Sausage and Mangoes are bought together. Similarly, the rule (Mangoes ,Sausage)→ ghee indicates the occurrences where both Sausage and Mangoes were taken with ghee. To find out if a rule is relevant,a matrix called confidence is considered. If the confidence of the rule is above the cutoff value , then the rule will be a strong rule

The two main measures used in the development of effective rules from the transactional data are support and confidence. Support refers to the number of times itemset appear in the dataset. On the other hand, confidence refers to the percentage of transactions that fulfill both X and Y. These two measures are used to analyze the relationship between the various elements in a dataset.

The paper is divided into two sections: first, it describes the various techniques that are used in the area of mining with association rules for the survival analysis among breast cancer patients. The second section focuses on the applications of these techniques in other areas. The paper also talks about the various aspects of the SEER dataset and the CbARM Algorithm's application in the analysis of it. In Section 3, the paper provides an overview of the data description and the procedure used for the extraction of the rules. The results of the research are summarized in Section 4 and 5. The conclusion of this study is also provided in Section 5.

2. Related Works

The paper also presents an overview of the various studies that are conducted on the use of rule mining in the perusal of breast cancer survival. These studies are complemented with other machine learning aswell as data mining techniques. Furthermore, the works that have been done in the past using this technique are covered in this section.

In the paper[7], Yu Duan and Fangfang Li present a case study that analyzed the SEER dataset and found out that there were over 300 association rules that are related to survived and non-survived patients. They also discovered that the presence of an EOD-Lymph Node plays a key role in the survival of patients.

The authors of this paper also discuss the importance of finding other attributes that can help determine the likelihood of survival for patients with a tumor type that is malignant.

In the paper[8], the authors of this study analyzed the effects of different classification algorithms on the accuracy of the results of the study. The four main methods that were used in the study were the RandomForest, NaiveBayes, SMO Poli-kernel, and Simple K-Means. The results of the study revealed that the SMO Poly-Kernel had the highest recall rate and precision.

Umesh and Ramachandra in [9] analyzed the Apriori Algorithm's application in the analysis of the SEER data for finding out the factors that can help determine the likelihood of a patient's survival after their cancer has returned. They discovered that the presence of various characteristics such as the age of menopause, lymph nodes, and tumor size can influence the likelihood of a patient's survival.

In his survey[10], Hamid Zand discussed the various methods used in data mining for predicting and diagnosing breast cancer. The paper analyzed the results of the study and found that the C4.5 decision tree algorithm performed well in predicting a patient's survival. He found that the NaiveBayes algorithm had a better accuracy than the other two classification algorithms

In the paper[11], the authors of the study, Wang Cheng, Liu-Ya Qin, and Zhang Lu, proposed a predictive model for breast cancer survival according to the decision tree algorithm. They discovered that the model performed well while the distribution of the data was equal.

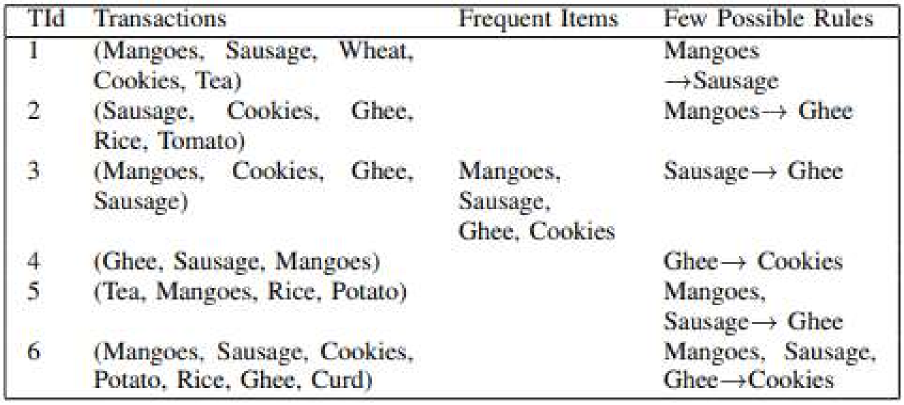
One of the most challenging factors when it comes to the development of effective rules is the time it takes to analyze the data. In this paper[12], the authors present two fast algorithms that can easily mine the rules from large databases. The Apriori and AprioriTid algorithms are respectively shown in the study. However, in rare cases, the support confidence framework for association rules is not ideal which is clarified in [13].

In 2004, a multi-objective approach was introduced [14] that involves extracting rules from a dataset using three measures: confidence, comprehensibility, and interestingness. The algorithm is powered by a genetic concept called the Genetic Algorithm. To perform sampling, the researchers randomly selected a few sets of rules from the data.

Another method was[15] introduced in 2010 for extracting association rules from incremental datasets. The GA approach breaks the data into multiple sections and discovers the association rules in each section. The results of the analysis are then analyzed and the support confidence framework is validated.

A different algorithm[16] was then introduced in 2016 that improves the memory and speed of the extraction process. The major concept of this method is to create a tree composed of the various items in the data. After the tree is created, the rules and frequent items are then extracted. This method requires two scans of the data.

A class conversion process is then used in [17,18] vertical mining to extract frequent itemsets. In 2017, a method called H-mine and H-struct [19] was introduced. These two data structures are used to mine frequent itemsets. In [20], the authors also introduced a set of measures called excl and lsup. These measures are based on class distribution and perform well when compared with support and confidence.



TId	Transactions	Frequent Items	Few Possible Rules
1	(Mangoes, Sausage, Wheat, Cookies, Tea)		Mangoes →Sausage
2	(Sausage, Cookies, Ghee, Rice, Tomato)		Mangoes→ Ghee
3	(Mangoes, Cookies, Ghee, Sausage)	Mangoes, Sausage, Ghee, Cookies	Sausage→ Ghee
4	(Ghee, Sausage, Mangoes)		Ghee→ Cookies
5	(Tea, Mangoes, Rice, Potato)		Mangoes, Sausage→ Ghee
6	(Mangoes, Sausage, Cookies, Potato, Rice, Ghee, Curd)		Mangoes, Sausage, Ghee→Cookies

Table 1: Market Basket Analysis: A Transaction Example

In [21], Akbarzadeh and Azadeh Soltani introduced a new algorithm that is more efficient when it comes to extracting both infrequent and frequent itemsets. This method, which only uses a one time scan of the data, allows users to easily generate rules.

3. Association Rule Mining

3.1 Related Concepts

One of the most common techniques used in data mining is the association rule mining. This method finds the associations between the various data items in a transaction. For instance, $I=\{it(1),it(2),it(3),....,it(m)\}$ is the itemset which is composed of m items, and the transaction set generated is $T =\{tr(1),tr(2),tr(3),....,tr(n)\}$ where $tr(1),tr(2)$ etc represents each transaction Consider an association rule generated as $P\rightarrow Q$, $P\varepsilon T$, $Q\varepsilon T$. Then $Support(P\rightarrow Q)= P(P \cup Q)$, $Confidence(P\rightarrow Q)=P(Q/P)=P(P \cup Q)/ P(P)$. For support and confidence a minimum value will be decided and the rules $P\rightarrow Q$ which satisfy $P(P \cup Q) \geq minsupport$ and $P(Q/P) \geq minconfidence$ are evaluated as interesting rules.

3.2 CbARM Algorithm

The concept of cogency is similar to that of the confabulation method. This method uses a single scan of the data to improve the memory efficiency. Because of its one time scan benefit the method will be best suitable for mining rules from the incremental data sets. In the following section the whole process will be explained in detail.

The first step in the process is to recognize the different items in the data. Then, the matrix is constructed using the information collected from the data. As the next level, extraction of the rules and frequent items need to be done.

When it comes to analyzing an incremental dataset, it can be challenging to keep track of the changes in the data due to the constant updating nature of the data. In traditional methods, the whole dataset is scanned every time a change is made. With an incremental dataset, the new items are usually smaller compared to the previous ones. This method, which only uses a single scan of the data, is more efficient when it comes to keeping track of the changes in the data. After the matrix is constructed, it only needs to be updated once new items are included to the dataset.

3.3 Item Identification and Support Calculation

Let I = {it(1),it(2),it(3),....,it(m)} be a collection of items and T = {tr(1),tr(2),tr(3),....,tr(n)} be a collection of transactions. The first step in the process is to identify the various items in the data. Then, we can calculate the support of each item and the frequency of its occurrence. The frequently and repeatedly occuring items or elements that cross a cutoff value of support are named as Frequent Items FR. FT1 is used to represent the first level subsets or fragments of these elements which can be named as Fragment Level 1.

$$FR = x \in I \mid support(x) \geq Smin$$

Where Smin is the support cutoff which is minimum

3.4 Building the matrix

Constructing a data matrix will be the next level. The order of the data matrix will be m X m , if there are m items present. For instance, if we consider m items, then (it)1, (it)2, (it)3,...(it)m, will represent the column heads as well as the row. To create the matrix, the values of each transaction will be assigned initially as zero. After that, the respective value will be incremented based on the presence of the two level subsets. Consider the following transactions

T1 ={P, Q, R}
T2 ={A, B, C, D, E}
T3 ={P, C, D, E}

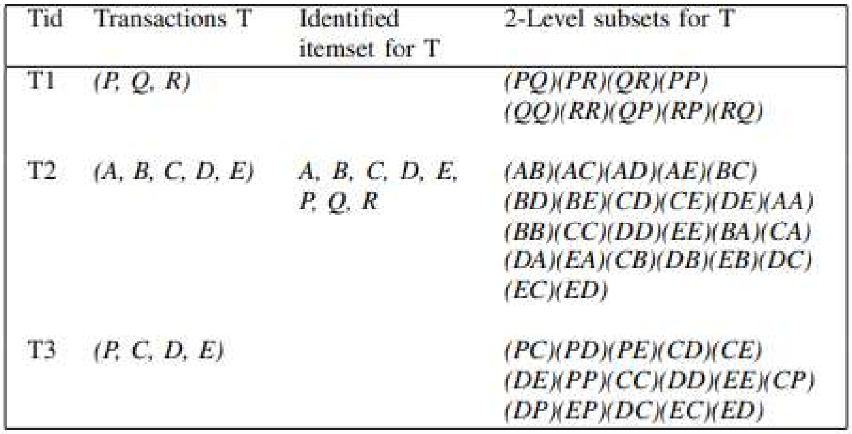
The Table 2 below shows the three transactions and the itemset that were associated with them. After the itemset is identified, a knowledge matrix can be generated by taking into account the rows and column heads. The link will be strengthened after the presence of the two level subsets. Algorithm 1 gives the steps to generate the matrix. An example matrix is also be shown in the Table 3. Steps to generate next level subsets and interesting rules given in Algorithm 2.

3.5 Cogency Calculation (P → Q)

The data matrix values are used at this level.

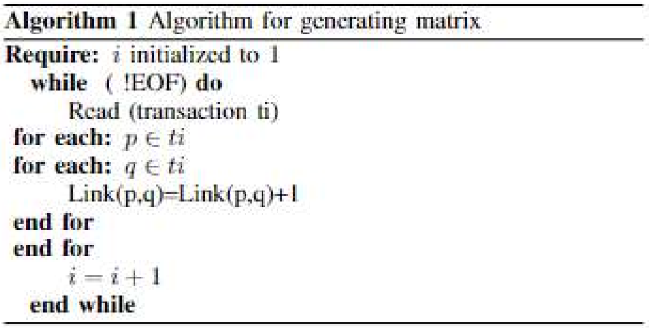
Suppose if P = {Cookies, Mangoes} and Q = {Sausage} then Cogency((Cookies, Mangoes) →Sausage) = Cogency on Cookies × Cogency on Mangoes

Cogency ((Cookies, Mangoes) →Sausage) = L (Cookies, Sausage) | L (Sausage , Sausage)× L (Mangoes, Sausage) | L (Sausage , Sausage)

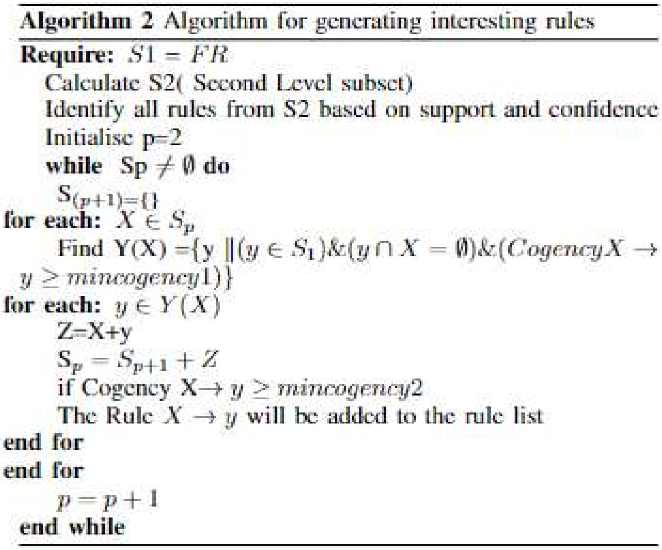


Tid	Transactions T	Identified itemset for T	2-Level subsets for T
T1	*(P, Q, R)*		*(PQ)(PR)(QR)(PP)(QQ)(RR)(QP)(RP)(RQ)*
T2	*(A, B, C, D, E)*	*A, B, C, D, E, P, Q, R*	*(AB)(AC)(AD)(AE)(BC)(BD)(BE)(CD)(CE)(DE)(AA)(BB)(CC)(DD)(EE)(BA)(CA)(DA)(EA)(CB)(DB)(EB)(DC)(EC)(ED)*
T3	*(P, C, D, E)*		*(PC)(PD)(PE)(CD)(CE)(DE)(PP)(CC)(DD)(EE)(CP)(DP)(EP)(DC)(EC)(ED)*

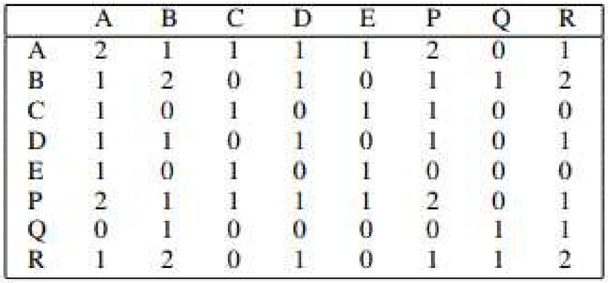
Table 2: Transactions and two level subsets



```
Algorithm 1 Algorithm for generating matrix
Require: $i$ initialized to 1
  while ( !EOF) do
      Read (transaction ti)
 for each: $p \in ti$
 for each: $q \in ti$
      Link(p,q)=Link(p,q)+1
 end for
 end for
      $i = i + 1$
  end while
```



```
Algorithm 2 Algorithm for generating interesting rules
Require: $S1 = FR$
  Calculate S2( Second Level subset)
  Identify all rules from S2 based on support and confidence
  Initialise p=2
  while Sp $\neq \emptyset$ do
    S$_{(p+1)}$={}
for each: $X \in S_p$
     Find Y(X) ={y $\|(y \in S_1)\&(y \cap X = \emptyset)\&(Cogency X \rightarrow y \geq mincogency1)\}$
for each: $y \in Y(X)$
     Z=X+y
     S$_p = S_{p+1} + Z$
     if Cogency X$\rightarrow y \geq mincogency2$
     The Rule $X \rightarrow y$ will be added to the rule list
end for
end for
     $p = p + 1$
 end while
```



	A	B	C	D	E	P	Q	R
A	2	1	1	1	1	2	0	1
B	1	2	0	1	0	1	1	2
C	1	0	1	0	1	1	0	0
D	1	1	0	1	0	1	0	1
E	1	0	1	0	1	0	0	0
P	2	1	1	1	1	2	0	1
Q	0	1	0	0	0	0	1	1
R	1	2	0	1	0	1	1	2

Table 3: Building matrix from a sample transaction

3.6 Incremental Datasets

Due to the increasing number of updations in the data, the classic methods turned out to be very inefficient when it comes to keeping track of the changes in the data. In most cases, the whole process takes several times more time than it should. With the cogency method, the matrix data is only updated.

4. Experimental Results

4.1 Data Preprocessing

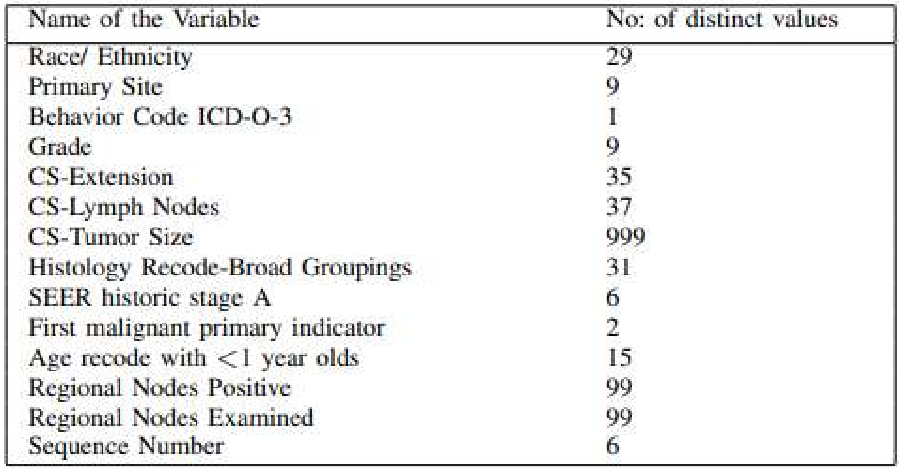
This paper presents the data set used for the study of the SEER Data set, which is a collection of cancer incidence statistics from various community-based cancer registry programs in the United States. These programs collect information on the characteristics of the patients, the initial tumor site, the stage of the disease, and the treatment options.

The data set contains various attributes such as the status of the patients, the initial tumor site, and the treatment options. After implementing the pre-classification method, the total number of records collected was 754652. Out of these, 34813 records were classified as Survived, while 9421 records were classified as Non-survived.

The classification process is carried out according to the following guidelines. If the status of the patients is positive for Survival Months and Vital Status Record which means, if Survival Months > 60 and Vital Status Record is 'alive', then the record is considered to be "survived." On the other hand, if the status of the patients is negative for Cause Of Death and Survival Months, then the record is considered to be "not survived." After implementing the pre-classification method, 14 attributes related to breast cancer were selected which is represented in Table 4.

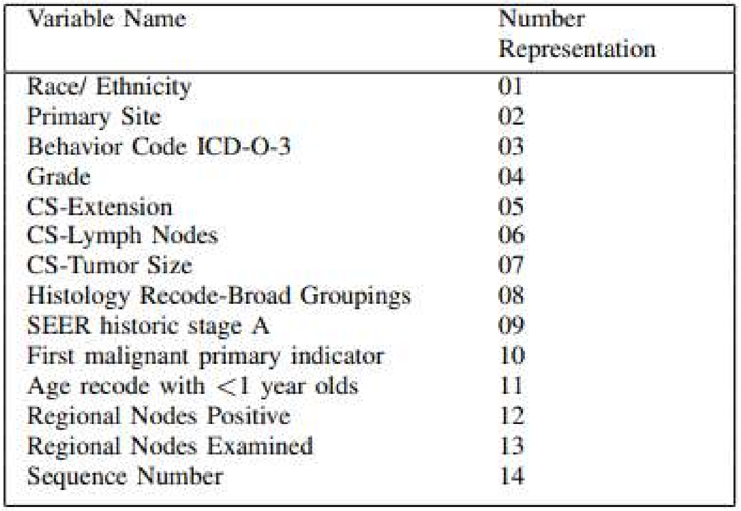
4.2 Transactional Dataset Conversion

Since the data set collected from the SEER Registries is not in a transactional format, we have to convert it into a new set of records. This process can be carried out by implementing the CbARM method, which only allows the use of transactional datasets. To perform this process, we have to select the various attributes from the data set using a number ranging from 1 to 14. In Table 5 new representations are given.



Name of the Variable	No: of distinct values
Race/ Ethnicity	29
Primary Site	9
Behavior Code ICD-O-3	1
Grade	9
CS-Extension	35
CS-Lymph Nodes	37
CS-Tumor Size	999
Histology Recode-Broad Groupings	31
SEER historic stage A	6
First malignant primary indicator	2
Age recode with <1 year olds	15
Regional Nodes Positive	99
Regional Nodes Examined	99
Sequence Number	6

Table 4: Attribute Names



Variable Name	Number Representation
Race/ Ethnicity	01
Primary Site	02
Behavior Code ICD-O-3	03
Grade	04
CS-Extension	05
CS-Lymph Nodes	06
CS-Tumor Size	07
Histology Recode-Broad Groupings	08
SEER historic stage A	09
First malignant primary indicator	10
Age recode with <1 year olds	11
Regional Nodes Positive	12
Regional Nodes Examined	13
Sequence Number	14

Table 5: Numerical representations for the attributes

4.3 Implementation Details

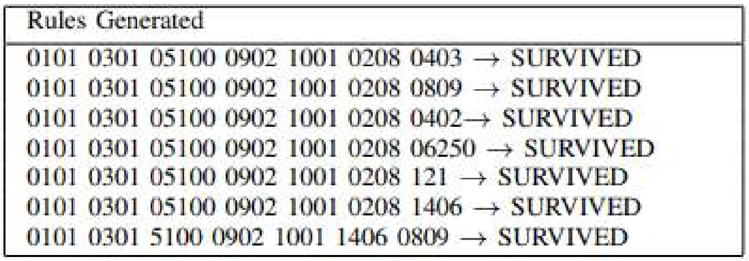
The proposed paper proposes the use of the CbARM Algorithm in order to perform an analysis on the data set collected from the Registries of the SEER. The process is carried out through the .NET Framework, which accepts the transaction data as input.

4.4 Association Rule Generation

The minimum support value taken in this work is 20%. There are total 9421 records in the NON-SURVIVED dataset and four samples of 9421 records in SURVIVED datset.

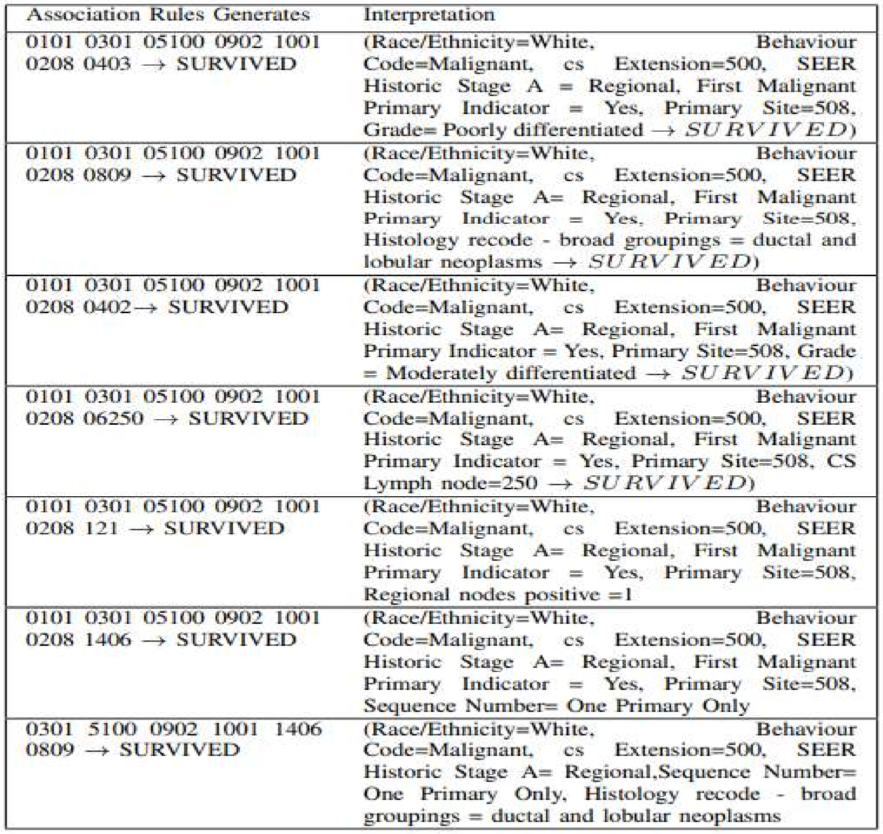
Since we have four samples of survived dataset (to balance the data) after applying CbARM Algorithm total 1326 common rules were generated . Among these 1326 rules 7 rules generated with 7 combination of attributes, 95 rules with 6 attribute combinations, 323 rules generated with 5 combination of attributes and 474 rules with 4 combinations are generated .Considering remaining rules 330 with three attribute and 97 rules with two attribute combinations are also generated. The 7 most important rules generated for the SURVIVED Dataset is given in the Table 6.

From the Table 6 the first rule (0101 0301 05100 0902 1001 0208 0403 → SURVIVED) indicates (Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500,SEER Historic Stage A= Regional, First Malignant Primary Indicator = Yes, Primary Site=508, Grade= Poorly differentiated → SURVIVED). In the same way the interpretation of these 7 rules are given below in the Table 7.



Rules Generated
0101 0301 05100 0902 1001 0208 0403 → SURVIVED
0101 0301 05100 0902 1001 0208 0809 → SURVIVED
0101 0301 05100 0902 1001 0208 0402→ SURVIVED
0101 0301 05100 0902 1001 0208 06250 → SURVIVED
0101 0301 05100 0902 1001 0208 121 → SURVIVED
0101 0301 05100 0902 1001 0208 1406 → SURVIVED
0101 0301 5100 0902 1001 1406 0809 → SURVIVED

Table 6: Rules generated with 7 attribute combinations

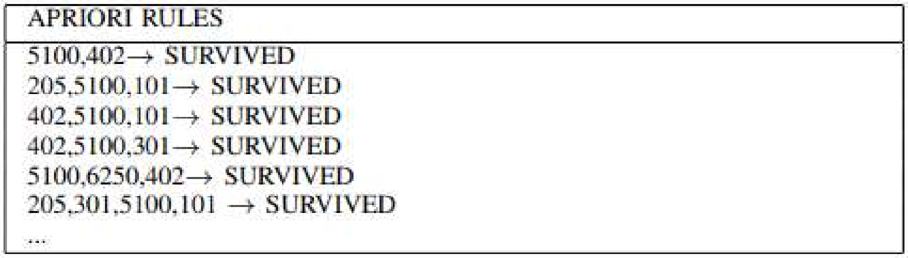


Association Rules Generates	Interpretation
0101 0301 05100 0902 1001 0208 0403 → SURVIVED	(Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500, SEER Historic Stage A = Regional, First Malignant Primary Indicator = Yes, Primary Site=508, Grade= Poorly differentiated $\rightarrow SURVIVED)$
0101 0301 05100 0902 1001 0208 0809 → SURVIVED	(Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500, SEER Historic Stage A= Regional, First Malignant Primary Indicator = Yes, Primary Site=508, Histology recode - broad groupings = ductal and lobular neoplasms $\rightarrow SURVIVED)$
0101 0301 05100 0902 1001 0208 0402→ SURVIVED	(Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500, SEER Historic Stage A= Regional, First Malignant Primary Indicator = Yes, Primary Site=508, Grade = Moderately differentiated $\rightarrow SURVIVED)$
0101 0301 05100 0902 1001 0208 06250 → SURVIVED	(Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500, SEER Historic Stage A= Regional, First Malignant Primary Indicator = Yes, Primary Site=508, CS Lymph node=250 $\rightarrow SURVIVED)$
0101 0301 05100 0902 1001 0208 121 → SURVIVED	(Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500, SEER Historic Stage A= Regional, First Malignant Primary Indicator = Yes, Primary Site=508, Regional nodes positive =1
0101 0301 05100 0902 1001 0208 1406 → SURVIVED	(Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500, SEER Historic Stage A= Regional, First Malignant Primary Indicator = Yes, Primary Site=508, Sequence Number= One Primary Only
0301 5100 0902 1001 1406 0809 → SURVIVED	(Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500, SEER Historic Stage A= Regional,Sequence Number= One Primary Only, Histology recode - broad groupings = ductal and lobular neoplasms

Table 7: 7 Combination Rules generated through CbARM and their interpretation

4.5 Comparison on rules generated through Apriori approach and CbARM Algorithm

The objective of this study is to analyze the potential of the CbARM algorithm to produce new rules and improve the efficiency of the classification process by comparing it with the traditional Apriori algorithm. Through the use of the SPMF implementation of the Apriori algorithm, we were able to test the Apriori algorithm against the SEER data set. Here the minimum support threshold was also taken as 20% and confidence value as 70%. The rules generated in the Apriori method is listed in the Table 8. In this method, the low support value resulted in the generation of many unnecessary rules. However, with the cogency value, the number of rules that are produced in the second method will be reduced, and more efficient rules will be added.



APRIORI RULES
5100,402→ SURVIVED 205,5100,101→ SURVIVED 402,5100,101→ SURVIVED 402,5100,301→ SURVIVED 5100,6250,402→ SURVIVED 205,301,5100,101 → SURVIVED ...

Table 8: Rules generated through APRIORI Algorithm

In Apriori Algorithm 52 interesting rules have been generated related to survivability in which 12 two attribute combination rules, 26 three attribute combination rules, 13 four attribute combination rules and one 5 attribute combination rule has been generated. A Comparison graph on the rules generated by Apriori and CbARM is depicted in Fig.1

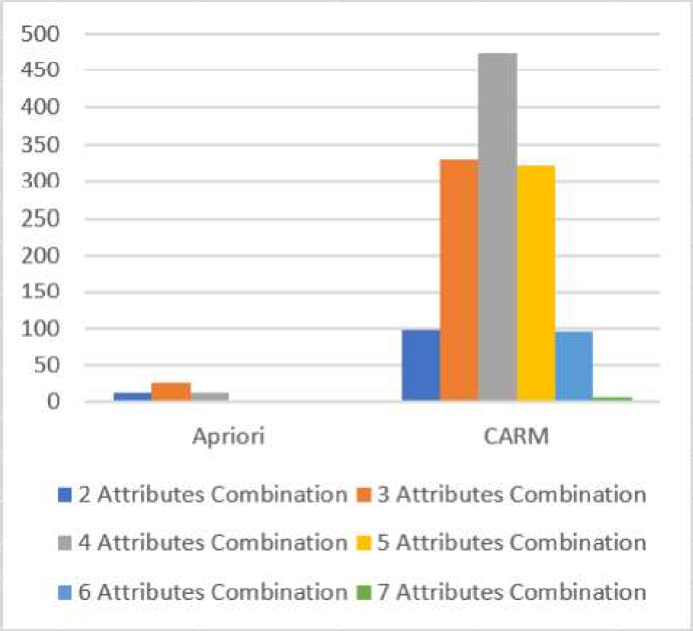


Fig 1: Comparison on rules generated by CbARM and APRIORI

5. Conclusion

In this work, we have applied the Confabulation based rule mining algorithm called CbARM to the recent 2020 November SEER datset and we obtained 1326 interesting association rules related to survivability. We found out Race/Ethnicity=White, Behaviour Code=Malignant, cs Extension=500,SEER Historic Stage A= Regional, First Malignant Primary Indicator = Yes, Primary Site=508, Grade= Poorly differentiated etc are very important attributes while considering the survivability. Knowing the data related to the survival of patients can help improve the quality of care for those suffering from this condition. It can also help medical experts identify the most effective treatment methods.

A comparison of CbARM algorithm with classic Apriori algorithm is also done in the paper. In Apriori, only till 5 attribute combinations are generated whereas in CbARM more efficient rules with 7 combinations are also generated and which inturn give lights to more significant attributes regarding the survivability in breast cancer patients. Moreover, compared to Apriori no need of scanning the dataset multiple times as CbARM demands only one scan of dataset as updations are made on to the constructed matrix.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021 May;71(3):209-49.

[2] Ali R, Mathew A, Rajan B. Effects of socio-economic and demographic factors in delayed reporting and late-stage presentation among patients with breast cancer in a major cancer hospital in South India. Asian Pac J Cancer Prev. 2008 Dec;9(4):703-7.

[3] Li, Yihao, and Theresa Beaubouef. "Data Mining: Concepts, Background and Methods of Integrating Uncertainty in Data Mining." CCSC SC Stud. EJ 3 (2010): 2-7.

[4] Slimani, Thabet, and Amor Lazzez. "Efficient analysis of pattern and association rule mining approaches." arXiv preprint arXiv:1402.2892 (2014).

[5] Dina, T. S., A. Amir, and S. Olfat. "Studying Combined Breast Cancer biomarkers using Machine Learning Techniques." (2016): 247-251.

[6] Seeja, K. R., M. A. Alam, and S. K. Jain. "An association rule mining Approach for co-regulated Signature genes identification in cancer." Journal of Circuits, Systems, and Computers 18.08 (2009): 1409-1423.

[7] Li, Fangfang, and Yu Duan. "An Analysis of the Survivability in SEER Breast Cancer Data Using Association Rule Mining." International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage. Springer, Cham, 2016.

[8] Roberto Cesar, Morales-Ortega, et al. "Method based on data mining techniques for breast cancer recurrence analysis." International Conference on Swarm Intelligence. Springer, Cham, 2020.

[9] Umesh, D. R., and Bharathkumar Ramachandra. "Association rule mining based predicting breast cancer recurrence on SEER breast cancer data." 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT). IEEE, 2015.

[10] Zand, Hamid Karim Khani. "A comparative survey on data mining techniques for breast cancer diagnosis and prediction." Indian Journal of Fundamental and Applied Life Sciences 5.s1 (2015): 4330-9.

[11] Liu, Ya-Qin, Cheng Wang, and Lu Zhang. "Decision tree based predictive models for breast cancer survivability on imbalanced data." 2009 3rd international conference on bioinformatics and biomedical engineering. IEEE, 2009.

[12] Fan, Qi, et al. "An application of apriori algorithm in SEER breast cancer data." 2010 International Conference on Artificial Intelligence and Computational Intelligence. Vol. 3. IEEE, 2010.

[12] Agrawal, Rakesh, et al. "Fast discovery of association rules." Advances in knowledge discovery and data mining 12.1 (1996): 307-328.

[13] Usama, Fayyad, Gregory Piatetsky-Shapiro, and R. Uthurusamy. "Advances in knowledge discovery and data mining." Computers & Mathematics with Applications. Vol. 32. No. 10. American Association for Artificial Intelligence, 1996. 128.

[14] Jahid, Md Jamiul, and Jianhua Ruan. “Identification of biomarkers in breast cancer metastasis by integrating protein-protein interaction network and gene expression data.” 2011 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS). IEEE, 2011

[15] Agarwal, Devika, et al. “A systems biology approach to identify proliferative biomarkers and pathways in breast cancer.” 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2014.

[16] He, D. Wang and Z. Guo, “Identification of Potential Non-invasive Biomarkers for Breast Cancer Prognosis and Treatment by Bioinformatics Analysis,” 2015 7th International Conference on Information Technology in Medicine and Education (ITME), Huangshan, China, 2015, pp. 117-120, doi: 10.1109/ITME.2015.28.

[17] Tseng, Yi-Ju, et al. “Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies.” International journal of medical informatics 128 (2019): 79-86.

[18] Marion T Weigel, Mitch Dowsett. “Current and emerging biomarkers in breast cancer: prognosis and prediction” 2010 Sep 23;17(4):R245-62. doi: 10.1677/ERC-10-0136. PMID: 20647302.

[19] Michael J Duffy, Enda W McDermott, John Crown. “Bloodbased biomarkers in breast cancer: From proteins to circulating tumor cells to circulating tumor DNA” Tumor Biology. May 2018. doi:10.1177/1010428318776169.

[20] Loke, Sau Yeen, and Ann Siew Gek Lee. “The future of blood-based biomarkers for the early detection of breast cancer.” European Journal of Cancer 92 (2018): 54-6

[21] Azadeh Soltani and M.-R. Akbarzadeh-T., “Confabulation-Inspired Association Rule Mining for Rare and Frequent Item sets” IEEE Transactions on Neural Networks And LearningSystems, VOL. 25, NO. 11, November 2014