RESEARCH ARTICLE OPEN ACCESS

# Classifying Brain Dataset Using Classification Based Association Rules (CBA)

Bahera H. Nayef
Computer Techniques Engineering
Al Turath University College
Iraq

## ABSTRACT

Classification Based Association (CBA) approach is one of the important techniques in data mining field. It performance is depends on finding rules which are related in high confidence and support in the training data split to classify the test data set by searching the relationships between the instances. The classifier performance depends on the quality of these rules. In this paper, many experiments are conducted to test the CBA performance with different number of attributes of the brain dataset with and without randomizing the subsets. Then using two methods of mining which are Apriori and predictive Apriori miners with the brain data. Another test is applied to compare the performance of CBA with Weighted and PART classifiers. Different datasets sizes with different number of attributes are used to measure the CBA performance. The result showed that CBA classifier performs better with small size and less number of attributes datasets. In addition large number of rules generated by predictive Apriori miner does not improve the classifier performance it increases the mining time.

***Keywords***:- Classification Based Association (CBA), Apriori miner, predictive Apriori miner, data mining, Weighted classifier, PART classifier.

## I. INTRODUCTION

An accurate and fast Classification of large size data sets with big number of attributes is a very crucial task. Recently many studies and researches conducted to construct an efficient and reliable classification system. The system is compound of classification and association rules mining approaches.

Data mining is concern in finding and extracting useful knowledge from a big heap of collected data. There many different techniques and approaches are produced for this purpose like statistical, “divide-and-conquer” and covering approaches. There are many classification algorithms are constructed from these approaches such as PART, C4.5, NAÏVE Bayes. These algorithms produced small subset of rules which means neglecting some information might be very essential for gaining good classification accuracy [1]. In this paper, a three classification and association rules algorithms are compared.

The main difference between association rule discovery and classification is that the first can predict any attribute in a data set but the second works by predicting one attribute which is the class [1].

Association rules proposed by Agrawal [2]. It is one of the most vital data mining research. The base of the approach is to search the relationship between item sets in a database and illustrate the interconnection between attributes. The association rules is appropriate approach for large datasets which increase in time commercial business and industries [3].

## II. LITERATURE REVIEW

In this section, a brief review related to some studies conducted over the past years related to classification based association techniques. These researches proposed the use of association rules and classification based on association rules. In Yin, X. and J. Han [4] study, the researchers tested Classification based on Predictive Association Rules (CPAR) approach with 26 datasets from UCI ML Repository with 10 fold cross validation. The main objective of this paper I to generate a large number of rules from the training data by using a greedy algorithm. The researchers P. Rajendran and M. Madheswaran [5] discussed the classification of CT scan of brain images. First they enhanced the images using image processing techniques and extract image features. Second step was classifying the resulted data set using association rule mining (decision tree) and hybrid classifier. The results of this study showed a significant classification accuracy rate 95% and 97% sensitivity . The paper [6] published by Thabtah etal (2005) proposed multi class, multi label associative classification technique to classify 28 datasets with different characteristics. The results showed accurate and effective classification performance in compare to the traditional techniques. D. Janssens [7, 8] applied the classification Based On Associations (CBA) algorithm to test the authentication of financial data set for bankruptcy prediction. They used C4.5, the original CBA and CART algorithms. The adaptive CBA showed better results than the above algorithms with 5%. in [1] proposed a classification association techniques to classify 13 datasets from UCI. The used the accuracy rate, numbers of generated rules , rules features and processing time as the base for their comparison between RIPPER, PART, MMAC, CBA, and C4.5. The gained results showed stable accuracy rate and reduced processing time when pruning the generated rules.

In this paper, CBA algorithm, PART, liner and equal Weighted Classifier are used to classify brain MRI images as normal and abnormal.

## III. CLASSIFICATION BASED ON ASSOCIATION RULE

In this section, a simple presentation is introduced to explain the process of item classification depending rules generation.

### *A. Frequent Items, Support and Confidence*

The data set used in the classification process is divided into two parts in a certain percentage. The first part is the training data set and the second is the testing data set. The training data set (T) is used to build the classification model and generate rules depending on the used techniques. The training data consists of (n) number of attributes refer to as $A_1$, $A_2$,$A_3$….,$A_n$ and C refer to the class. For each attribute $A_i$ there is a value $a_i$ and a class name $c_j$. Each item is represented by a group of different attributes from 1 to n . Each attribute is defined by the attribute name and value such as ($A_1$,$a_1$), ($A_2$,$a_2$), $A_3$,$a_3$) … so on. Many rules are generated from the training data and each rule is (r) for multi label classification is defined as the following form:

$$(A_{i1},a_{i1})\wedge \quad (A_{i2},a_{i2})\wedge \quad (A_{i3},a_{i3}) \quad \ldots(A_{im},a_{im}) \rightarrow c_{i1} \vee c_{i2} \vee \ldots . \vee c_{im}.$$

The left side of the rule represents the item condition and the right side is the resulting consequence in a form of ordered class labels. The relation between the rule ( r) and the actual occurrence (ActOccr), support count (suppCount) and minimum support threshold (Minsupp) is illustrated as follow [6]:

1- The number of times that the generated rule (r's) condition matches number of instances in the T is called the actual occurrence (ActOccr) and the number of instances |T| in T that matches r's condition is called the support count (suppCount) of (r). These instances |T| are belong to a class $c_j$. If an item is supported by more than one class then each one of them has (suppCount).

2- When the result of dividing the (suppcount) by the number of instances |T| is greater or equal to the (Minsupp) ,then the rule ( r) is passes the minimum support threshold (Minsupp). . As shown in equations (1)

3- If the suppcount of a rule ( r) is divided by the actual occurrence minimum and the result is greater than MinConf, then the rule passes the confidence threshold. As shown in equations (2)

$$\text{Suppcount ( r) / |T|} \geq \text{Minsupp} \qquad (1)$$

$$\text{Suppcount (r) /Actoccr (r)} \geq \text{MinConf} \qquad (2)$$

When an item is passes the Minsupp is called a frequent item.

### *B. Classification techniques based Association Rules*

Using association rules for classifying data was introduced by [8] by designing of CBA algorithm. The base of CBA algorithm performance depends on passing on the training data multiple to locate and discover the frequent items. This process is performed by the Apriority algorithm. The next step, the CBA generates a classifier rules from the frequent items that pass the minconf [2].

### *C. CBA classifier Algorithm*

And association rule consists of two parts the first part on the left side is the condition and the second part on the right side is called the consequent. The main advantage of using CBA algorithm in a classification task is that CBA generates rules by using the modified Apriority algorithm with a consequent that closely related to the class label $c_i$. This class is called Class Association Rule (CAR). In CBA algorithm there are two methods of pruning rules to decrease the total number of the generated rules by CAR, the first is the pessimistic error rate and the second method is database coverage pruning [9]. CAR arranges and ranks the generated rules according to the following criteria:

Lets consider $r_i$ have higher rank than $r_j$ , this means it matches one of the following conditions:

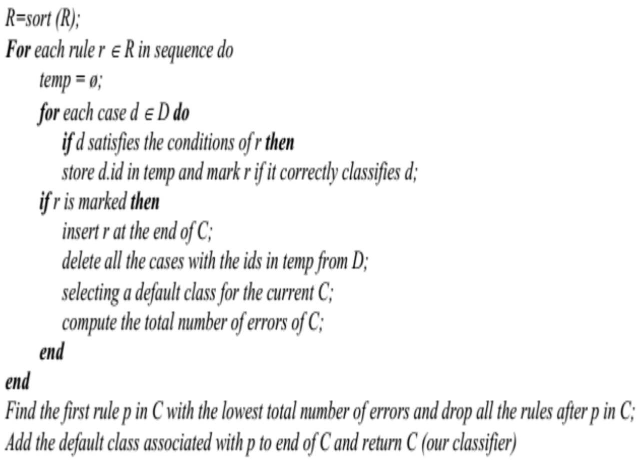
i- The confidence of $r_i$ > the confidence of $r_j$ .

ii- The confidence of $r_i$ = the confidence of $r_j$ , but minsupp of $r_i$ > the minsupp of $r_j$.

iii- The confidence of $r_i$ = the confidence of $r_j$ , but minsupp of $r_i$ = the minsupp of $r_j$ , but the generation of $r_i$ is before the generation of $r_j$ .

The rule is considered and inserted to the classifier when at least one case covered by it among many cases. After inserting the rule to the classifier the cases matches it will be removed from the database. The rules insertion process terminated when the database is empty which means no more cases left. If there are cases do not match any rule, then they will be considered as default class.

The next step the algorithm calculates the total number of errors caused by the generated rules and by the default class, The rule with minimum error rate is called the cutoff rule. All the other rules will be neglected because they reduce the classifier performance accuracy [8]. Figure 1 explains building a classifier in CBA algorithm.



```
R=sort (R);
For each rule $r \in R$ in sequence do
    temp = ø;
    for each case $d \in D$ do
        if d satisfies the conditions of r then
        store d.id in temp and mark r if it correctly classifies d;
    if r is marked then
        insert r at the end of C;
        delete all the cases with the ids in temp from D;
        selecting a default class for the current C;
        compute the total number of errors of C;
    end
end
Find the first rule p in C with the lowest total number of errors and drop all the rules after p in C;
Add the default class associated with p to end of C and return C (our classifier)
```

Figure 1 - CBA classification steps [8]

### *D. Weighted Associative Classifiers (WAC)*

For more accurate generating of association classification rules, weighted associative classifiers are introduced. This new technique base id using weighted support and confidence to mine association rules from datasets. This technique generates only the most significant rules and neglects the least significant rules. And that means reduce the number of generated rules. This method works by assigning a weight to each attribute between 1 and 0.1. After assigning the weights the weighted Association rule Mining algorithm is applied to generate the requested pattern and rules. The rules generated in the step is called classification Association Rule (CAR) [10]. In the study [10] the researchers explains how to weight and rank the attributes.

## IV. EXPERIMENTS AND RESULTS

In this paper many experiments have been conducted with the brain data set divided into splits 50% of training and testing data sets. The used data set is collected from some privet radiology centers with unseen identities of patients. The dataset contains 200 MRI images with two cases Normal and Abnormal. Image processing techniques are applied to extract the image features to be used for classification. The attribute resampling technique is applied. The following subsections discuss the experiments individually.

*1)* Experiment 1

This experiment is in two parts. The first studies the effect of attribute reduction and applying random subset filter on the performance of the CBA classifier with association Apriori algorithm for rules generation. In the second part the predictive Apriori association rules technique is used as rules miner with the CBA classifier.

The experiment also discusses time of building the classification model and the time of rules generation as explained in Table 1.

| no. of attributes | 21 | 21 | 11 | 6 |
|---|---|---|---|---|
| Time taken to build model | 1125.9 | 8692.22 | 0.1 | 0.01 |
| no. of classification rules | no | No | no | No |
| rules generation mining time in sec | 1125.9 | 8692.214 | 0.091 | 0.003 |
| Mean absolute error | 0.25 | 0.27 | 0.27 | 0.27 |
| Root mean squared error | 0.5 | 0.5196 | 0.5196 | 0.596 |
| Fmeasure | 0.75 | 0.616 | 0.616 | 0.616 |
| Pruning Time in sec. | 0 | 0 | 0.006 | 0.007 |
| accuracy % | 75 | 73 | 72 | 73 |
| resembling subsets | no | Yes | yes | Yes |

*TABLE 1 SHOWS the EFFECT of ATTRIBUTE REDUCTION BRAIN DATA on CBA PERFORMANCE*

As shown in Figure 2 reducing the number of attributes does not affect the performance accuracy rate so much.

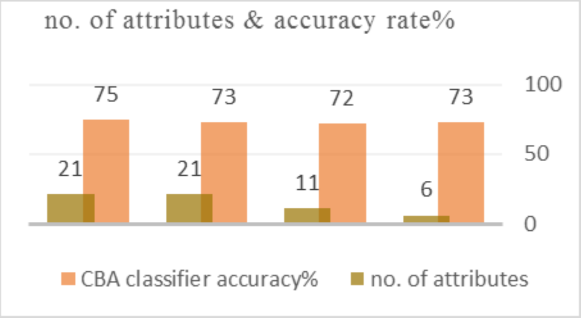


Figure 2 Accuracy rate and no. of attributes

Figure 3 explains the necessary time to build the CBA classification model with reducing the no. of attributes. The time required is reduced with reducing the no. of attributes enormously. For the 21 attributes, measures are taken for both cases with (8692.22 sec) and without resampling (1125.86 sec). Resampling increases the time for building model in compared to the time needed for 6 attributes which is 0.003 sec with resampling.

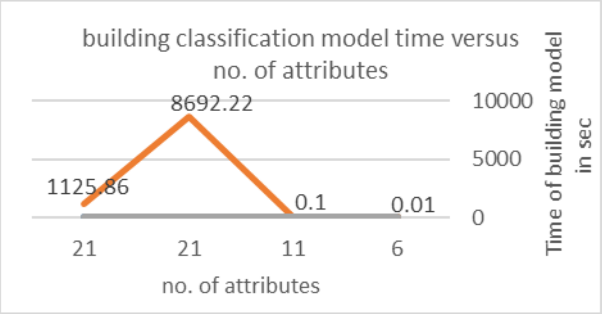


Figure 3 building model time with no. of attributes

Figure 4 explains the results of reducing no. of attributes and the time required to search the subsets for similarity to generate rules. As shown in the figure 4 the time necessary to search and generate rules with dataset of 21 attributes is 1125.852 sec before subset resampling, 8692.214 sec after subset resampling. Resampling increases the time for building model in compared to the time needed for 6 attributes which is 0.003 sec with resampling.

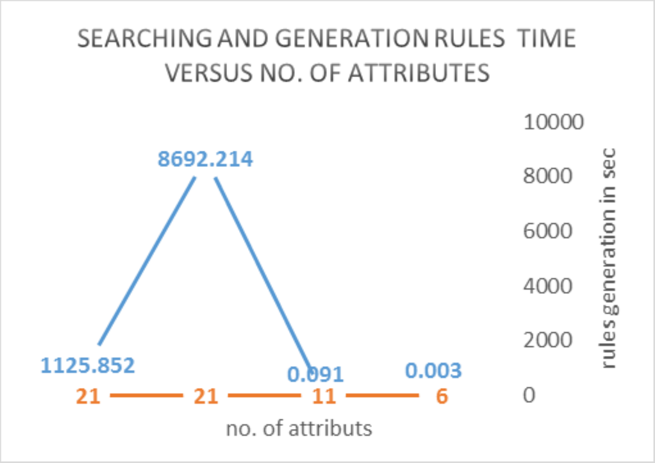


Figure 4 rules searching and generation time

According to the second part of this experiment the predictive Apriori association rules technique is used as rules miner with the CBA classifier for searching and generating rules. In this experiment we applied Bestfirst technique for attribute selection. The results are shown in Table 2.

| | CBA | | PART | |
|---|---|---|---|---|
| Dataset | collected brain set | | collected brain set | |
| no. attributes | 21 | 7 | 21 | 7 |
| Accuracy rate% | 73 | 73 | 73 | 75 |
| Classification Rules (ordered): | 8 | 7 | 4 | 5 |
| Default Class | Normal | normal | - | - |
| no Class. Assoc .Rules generated by Rule Miner | 99 | 37 | - | - |
| Mining Time in sec | 22718.22 | 0.936 | 0 | 0 |
| Time taken to build model in sec | 22718.23 | 1 | 0.03 | 0 |
| pruning time | 0.01 | 0.047 | - | - |

TABLE 2 CBA PERFORMANCE with PREDICTIVE APRIORI

| Comparison data | CBA classifier | | Weighted classifier (linear) | | Weighted classifier (equal) | |
|---|---|---|---|---|---|---|
| | Apr | predi | Apri | predi | Apri | predi |
| Accuracy % | 73 | 73 | 73 | 73 | 73 | 73 |
| Rules generated Mining Time in sec | 0.296 | 4.119 | 0.312 | 3.837 | 0.249 | 3.853 |
| Number of classification rules | no | 50 | 10 | 50 | 10 | 50 |
| No. of CA Rules generated by Miner | - | 8 | 10 | 50 | 10 | 50 |
| Pruning Time in sec | 0.047 | 0.015 | 0 | 0 | 0 | 0 |
| Building model time in sec | 0.31 | 4.13 | 0.31 | 3.85 | 0.25 | 3.85 |
| Time taken to test model on test split | 0 | 0 | 0 | 0.02 | 0 | 0.01 |

*Table 3 shows the performance of weighted classifier in comparison with CBA classifier*

From Table 2, number of generated rules increased to 99 but only 8 of them with high rank and 99% accuracy are used for classification for CBA classifier with 21 attributes and 37 rules generated but only 7 of them with the highest rank and 99% accuracy are used for building the classifier model with 7 attributes (*see Appendix 1 and 2*). Same thing for the PART, 4 rules are generated when using 21 attributes and 5 with 7 attributes as shown in Appendix 3. The mining time of CBA with predictive Apriori and 21 attributes to generate 99 rules is 22718.22 sec. which is higher than using 7 attributes which needs only 0.936 sec and the number of generated rules is 37. The number of rules in the late case is much less than the first. The pruning time to reduce the number of rules to the highest ranked only is 0.01 for CBA with 21 attributes but it takes a little longer (0.047sec) for CBA with 7 attributes. PART classifier did not take time to build and prune rules. It generated only 4 rules with 21 attributes and 5 rules with 7 attributes. That means PART classifier with predicate Apriori techniques generates more accurate and less rules in almost 0 pruning time. The accuracy rate of CBA is the same in both case but PART showed better accuracy rate (75%) with 7 attributes than 21(73%).

### *2) Experiment 2*

In this experiment introduces a comparison between the performance of CBA classifier and the weighted classifier. Both of them are used to classify the brain dataset with 13 attributes and Remove filter. The weighted classifier is applied with Apriori and predictive Apriori mining algorithms. The results are demonstrated in Table 3. The weighted classifier is used with both linear and equally (1) weighting scheme.

Actually the accuracy rate did not show any difference between CBA and Weighted classifiers. The main difference in rules mining time. In case of using CBA and weighted classifiers with Apriori miner, the mining time in both of them is almost the same. But with using predictive miner and CBA classifier needs a little bit longer time than with weighted classifier. No generated rules in CBA with Apriori miner. On the other hand with predictive it generated 50 rules but only 8 of them used in the classification phase. The weighted classifier in both schemes linear an equal shows no difference in the number of generated rules with both Apriori and predictive miners. These rules are not pruned also in both miners. Because in equal scheme all rules have the same weight (1). All the results are showed in table 4.

3) Experiment 3

The main aim of this experiment is to test CBA with both Apriori and predictive Apriori miners with a bunch of data sets from the UCI Machine Learning Repository. Different datasets sizes are chosen with different number of attributes as shown in Table 4

TABLE 4 SHOW the PERFORMANCE of CBA with VARIOUS DATASETS

| Data sets | Iris | | supermarket | | image segmentation | | segment test | | breast cancer | | Soybean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Apri | Pred | Apri | Pred | Apr | Pred | Apri | Pred | Apri | Pred | Apri | Pred |
| Data set size | 150 | 150 | 4627 | 4627 | 1500 | 1500 | 810 | 810 | 286 | 286 | 683 | 683 |
| Number of attributes | 2 | 2 | 109 | 109 | 11 | 11 | 11 | 11 | 6 | 6 | 19 | 19 |
| No. of classification rules | 2 | 3 | - | 79 | 1 | 28 | 2 | 22 | - | 68 | 2 | 10 |
| No. of generated rules by miner | 5 | 4 | - | 100 | 10 | 100 | 10 | 100 | - | 100 | 10 | 100 |
| mining Time in sec | 0.016 | 0.05 | 3.488 | 2234.853 | 0.514 | 364.144 | 0.281 | 48.5 | 0.015 | 0.843 | 14.787 | 12474.18 |
| Pruning time in sec | 0.015 | 0.002 | 0.065 | 24.73 | 0.016 | 0.148 | 0.015 | 0.078 | 0 | 0.14 | 0.063 | 0.169 |
| Building model time in sec | 0.05 | 0.05 | 3.55 | 2259.58 | 0.56 | 364.32 | 0.31 | 48.53 | 0.01 | 0.98 | 14.85 | 12474.35 |
| test model on test split 50% | 0.02 | 0 | 0.09 | 0.21 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| Accuracy% | 97.3 | 94.7 | 63.4 | 75.8 | 42.4 | 80.9 | 41.5 | 84 | 67.13 | 69.23 | 25.8 | 54.54 |

Figure 5 shows the performance of CBA classifier with Ariori and Apriori predictive miners using different data sets. It is obvious that the performance of CBA with predictive Apriori is significantly better than Apriori miner with all data sets.

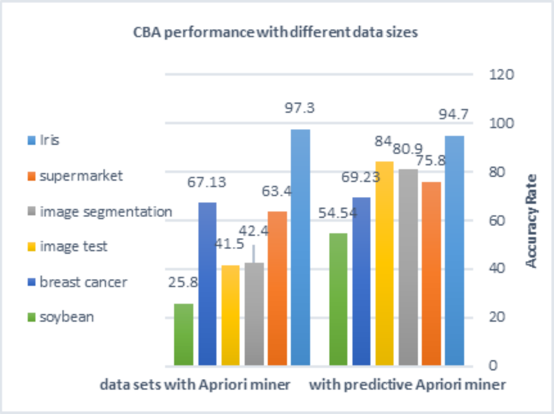
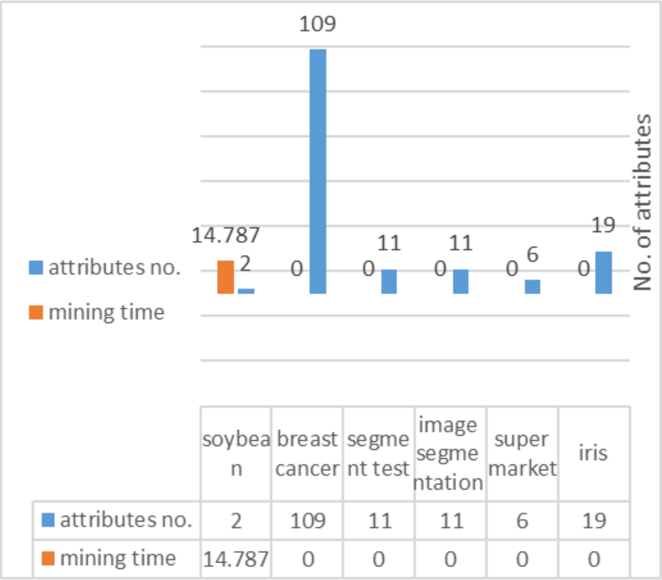


Figure 5 CBA classifier accuracy with various sizes of data sets

Figure 6 shows the mining time for classification rules of CBA with datasets with different number of attributes. As it is clear from Table 5 and Figure 6. Small size dataset with small number of attributes generates small number of classification rules in short mining time(sec) . Like with Iris data set it contains 2 attributes and 150 instances, 5 generated rules in 0.016 sec and after pruning only 2 of them is used for classification. The required pruning time is 0.015 sec. some datasets did not generate any rules with Apriori miner like supermarket and breast cancer. But they generate the maximum number of rules (100,10) with predictive apriori miner, only 79 and 68 rules in turn are used for classification process. Generating more rules using predictive miner improve the classifier performance rate as shown in Table 5.



| | soybean | breast cancer | segment test | image segmentation | super market | iris |
|---|---|---|---|---|---|---|
| attributes no. | 2 | 109 | 11 | 11 | 6 | 19 |
| mining time | 14.787 | 0 | 0 | 0 | 0 | 0 |

Figure 6 no. of attributes versus mining time in sec using Apriori miner

Figure 7 explains the relation between the accuracy rate and the number of classification rules generated by the predictive miner with CBA classifier for each dataset. From the figure we can notice that less rules gives better performance to the classifier. With iris (the smallest data set) only 3 rules are used to classify the test data which is 50% of dataset. It gives 94.7 % accuracy rate which is higher than the other data sets. The lowest is for the soybean data (54.5%) which generates 10 classification rules (12474.18 sec), and apparently that these rules do not satisfy the whole instances of the test data.

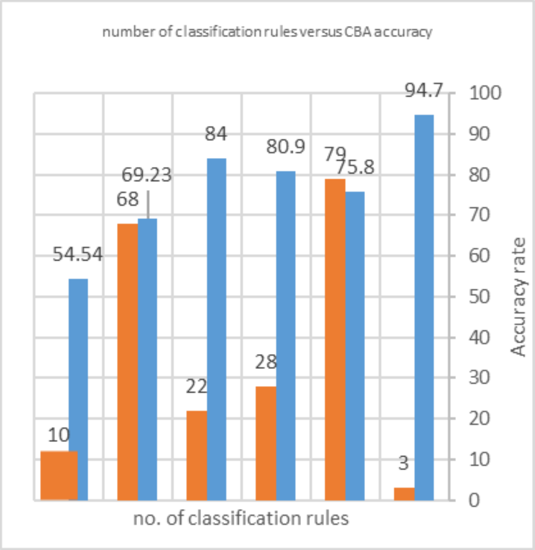


Figure 7 The effect of increasing no. of the classification rules on CBA performance

## V. CONCLUSION

CBA classifier performance depends on the rules generated from the training data split. The quality of these rules depended on high confidence and support. This paper tested the performance of CBA in different cases such as with large number of attributes and attributes reduction for the same dataset, using different miners (Apriori and predictive Apriori), using different sizes of datasets with different number attributes and comparing the performance of CBA with another rules association classifier which is the weighted classifier and rules dependent classifier which the PART classifier.

In all these experiments the classification accuracy rate of CBA kept settled specially with brain dataset. Even when changing the number of attributes (21,13,11,7,6) but the time required to mine them is reduced without random subset but with random subset the mining time is increased.

Number of rules generated with predictive miner was high but just few of them with high accuracy and support were used in classifying the test data. From experiment 4 we noticed that more rules do not give better classifier performance. With small size datasets the CBA performance is better than large dataset and needs less time. Mostly due to the CBA structure and the way it process data in form of binary.

## REFERENCES

[1] Bani-Ahmad, A.A.D.M.N.a.S., *CLASSIFICATION BASED ON ASSOCIATION-RULE MINING.* Ubiquitous Computing and Communication Journal. **5**(3): p. 9-17.

[2]. R. Agrawal, T.I.a.A.S., *Mining association*, in *ACM SIGMOD Conference on Management of Data*1993: Washington, D.C

[3]. Li Su, H.-y.L.a.Z.-H.S., *A New Classification Algorithm for Data Stream.* I.J.Modern Education and Computer Science, 2011. **4**: p. 32-39.

[4]. Yin, X. and J. Han. *CPAR: Classification based on predictive association rules*. in *Proceedings of the 2003 SIAM International Conference on Data Mining*. 2003. SIAM.

[5]. Rajendran, P. and M. Madheswaran, *Hybrid medical image classification using association rule mining with decision tree algorithm.* arXiv preprint arXiv:1001.3503, 2010.

[6]. Thabtah, F.A., P. Cowling, and Y. Peng. *MMAC: A new multi-class, multi-label associative classification approach*. in *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. 2004. IEEE.

[7]. Janssens, D., et al. *Integrating Classification and Association Rules by proposing adaptations to the CBA Algorithm*. in *Proceedings of the 10th International Conference on Recent Advances in Retailing and Services Science. Portland Oregon*. 2003.

[8]. Liu, B., Hsu, W., and Ma, Y., *Integrating Classification and association rule mining.* In KDD '98, New York. 1998., 1998.

[9]. Li, W., Han, J., Pei, J. . *Accurate and Efficient Classification Based on Multiple Class-Association Rules*. in *The 1st IEEE International Conference on Data Mining (ICDM 2001*. 2001. San Jose, California.

[10]. Soni, J., et al., *Intelligent and effective heart disease prediction system using weighted associative classifiers.* International Journal on Computer Science and Engineering, 2011. **3**(6): p. 2385-2392.