Applied Soft Computing 62 (2018) 536–549

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Full Length Article

WCBA: Weighted classification based on association rules algorithm for breast cancer disease

Jaber Alwidian*, Bassam H. Hammo, Nadim Obeid

King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan

ARTICLE INFO

Article history:
Received 23 February 2017
Received in revised form 2 November 2017
Accepted 5 November 2017
Available online 15 November 2017

Keywords:
Data mining
Association classification
Association rules
Apriori
Breast cancer

ABSTRACT

Breast cancer is the second most frequent human neoplasm that accounts for one quarter of all cancers in females. Among the other types of cancers, it is considered to be the main cause of death in women in most countries. An efficient classifier for accurately helping physicians to predict this chronic disease is in high demand. One approach for solving this problem has been tackled by many scholars using Association Classification (AC) techniques to enhance the classification process through applying association rules. However, most AC algorithms are suffering from the estimated measures used in the rule evaluation process and the prioritization techniques used at the attributes level, which could play a critical role in the rule generation process. In this article we attempt to solve this problem through an efficient weighted classification based on association rules algorithm, named WCBA. We also present a new pruning and prediction technique based on statistical measures to generate more accurate association rules to enhance the accuracy level of the AC classifiers. As a case study, we used WCBA to classify breast cancer instances with the help of subject matter experts from King Hussein Cancer Center (KHCC) located in Amman, Jordan. We compare WCBA with five well-known AC algorithms: CBA, CMAR, MCAR, FACA and ECBA running on two breast cancer datasets from UCI machine learning data repository. Experimental results show that WCBA, in most cases, outperformed the other AC algorithms for this case study. In addition, WCBA generates more accurate rules that contain the most efficient attributes for predicting breast cancer. WCBA algorithm aims to predict breast cancer in a patient. It serves all breast cancer patients by reducing the fear of the possibility of the recurrence of the disease and takes the necessary measures to prevent the progression of the disease and to predict breast cancer in a patient. The algorithm can be generalized to work on different domains with the help of subject matter experts.

1. Introduction

Breast cancer is a very critical disease in women across the world and is growing annually in an unpredictable rate. The lack of awareness regarding this type of disease motivated the researchers to use new techniques to overcome this issue. Some of these techniques are data mining techniques such as classification, clustering and Association Classification (AC). Breast cancer is the second most frequent human neoplasm and represents around one quarter of all cancers in females [1–5]. Furthermore, breast cancer is considered to be the main cause of death in women in most countries so a lot of effort should be made to reduce this type of chronic disease [2,3].

Data mining is the main phase in Knowledge Discovery from Databases (KDD) which includes a set of approaches for different aims, such as classification, clustering, association rules and association classification (AC) [6–8]. In this article, we will focus on the AC technique and show how it can be used to enhance the decision-making process in the breast cancer field.

Classification techniques aim to predict the class label of any given instance after the learning phase completed from a large dataset. Meanwhile, association techniques discover the relationships between items in a large database to represent hidden patterns and new knowledge. To enhance the accuracy measure of the classification process that can serve many critical fields like breast cancer, the researchers have, in the last few years, employed association rule techniques in the classification process to explore a new technique called the Association Classification technique (AC) [9,6,10].

The AC technique employs the association rules in the classification process in order to enhance hidden pattern that can

* Corresponding author.
E-mail addresses: j.alwidian@gmail.com (J. Alwidian), b.hammo@ju.edu.jo (B.H. Hammo), nadim@ju.edu.jo (N. Obeid).

https://doi.org/10.1016/j.asoc.2017.11.013
1568-4946/

improve the accuracy of the classification process which plays the main role in the decision-making process in many applications and domains. Thus, the AC technique is the second generation of the association rules technique. It has been implemented to find the correlation between items and classes. For example, as the rule *R:* $It_1, It_2 \rightarrow Class_1$, is interpreted as follows: if the Item values (It_1 *and* It_2) occur together for a particular object *o* in any instance, then *o* can be classified as $Class_1$, which represents the class value [9,6,10].

Generally speaking, most AC algorithms use two estimated measures to generate association rules: support and confidence without focusing on the application domain. In addition, they do not follow any prioritization technique to differentiate between the significances of the attributes for a given dataset. To avoid the shortcoming of the current AC methods, we proposed a new AC approach. The basic idea of this new classification approach is to replace the traditional support-confidence structure of association rule mining model by the weighted model. Attributes that have major influence in determining the breast cancer disease were assigned the highest priority by subject matter experts. We also use a statistical harmonic mean (HM) measure to prioritize the association rules at the pruning and generation phases. Hence, the rules with weak attributes were eliminated from the top rules generated in the rule pruning stage. Experimental results showed that weights of the attributes in the dataset have much impact on the classification rules.

To test our approach, we showed how WCBA can be applied to different domains through a running example in Section 5.2. A more sophisticated and thorough experimental case study has been carried out using the breast cancer datasets available on the UCI machine learning repository [11]. In these experiments, we compared the WCBA technique against a group of well-known AC algorithms. In most cases, as we will show later in Section 6, the WCBA algorithm outperform the AC algorithms in terms of accuracy. Furthermore, the help of experts in each field of study is mandatory before applying our proposed algorithm. To be generalized, other domains of study must be investigated and compared with the AC algorithms in the future. Finally, we used the WEKA tool to explore and visualize the results.

The rest of the article is organized as follows: In Section 2 we provide a background on AC. Section 3 presents an overview of breast cancer. In Section 4 we show some of the related work. Section 5 presents a detailed description of the WCBA technique. In Section 6 we discuss the experimental results on breast cancer. Finally, Section 7 concludes our work and shed some lights on future work avenues.

2. Association classification background

The AC technique is a combination of the association rules and classification techniques. The association rules technique aims to discover a correlation or association between attributes, while the classification technique is responsible for predicting the class label. For example, in a rule such as $At_1, At_2 \rightarrow C_1$, C_1 is the class attribute, while At_1 and At_2 are attribute values. As mentioned above, this rule can be interpreted as follows: if At_1 and At_2 attribute values occur together for any object; this object can be classified as C_1 [9,6,10].

The formal description of the AC problem can be found in Thabtah et al. [35]. We use the dataset (T) shown in Table 1 to explain the AC concepts and definitions.

In the AC problem, the association rules are employed in the classification process. If a rule states that $At_1 \rightarrow C_1$, then C_1 has to be a class attribute. The training data set T has *m* distinct attributes (At_1, $At_2 \ldots At_m$), and C is a list of classes. Attributes could be categorical or continuous. In the case of categorical attributes, all possible values are mapped to a set of positive integers, while continuous attributes use any discretization method. A row or a training object in T can be described as a combination of attribute names At_i and values v_i, plus a class denoted by C_j, and the item can be described as an attribute name At_i and value v_i. As shown in Table 1, (At_1, v_1) is an item; an itemset is a set of items contained in a training object, for example, (At_1, v_1) (At_2, v_3) is an itemset.

Table 1
Dataset sample (T) with four training objects.

Training object	Attribute 1 (At_1)	Attribute 2 (At_2)	Attribute 3 (At_3)	Class (C)
1	v_1	v_3	v_5	C_1
2	v_1	v_3	v_6	C_1
3	v_1	v_3	v_6	C_2
4	v_2	v_4	v_7	C_3

An itemset rule *r* is of the form < itemset, C_i>, where C_i is the class. In Table 1, training object 1 has <(At_1, v_1) (At_2, v_3), C_1> as an itemset rule. The actual occurrence (actoccr) of the itemset rule *r* in T is the number of rows in T that match the itemsets defined in *r*; thus, for (At_1, v_1) (At_2, v_3) as itemset, the actoccr = 3. Based on that, the support count (suppcount) of itemset rule *r* is the number of rows in T that matches *r*'s itemsets, and belong to a class C_i for *r*, as given in Eq. (1).

$$\textbf{Suppconut} = \mathbf{r} \cup \mathbf{c_i} \tag{1}$$

The suppcount for <(At_1, v_1) (At_2, v_3), C_1> itemset rule is 2, which means there are two occurrences of this itemset rule.

An itemset rule *r* passes the minsupp threshold if (suppcount(*r*)/|T|) ≥ minsupp, where |T| is the number of instances in T, as given in Eq. (2).

$$\textbf{Support} = \frac{\text{suppcount}(r)}{|\mathbf{T}|} \tag{2}$$

From Table 1, the number of training objects in the dataset T is 4; since the suppcount of the itemset rule <(At_1, v_1) (At_2, v_3), C_1> is 2, the support equals 2/4 = 0.5.

An itemset rule *r* passes the minconf threshold if (suppcount(*r*)/actoccr(*r*)) ≥ minconf, as shown in Eq. (3).

$$\textbf{Confidence} = \frac{\text{suppcount}(r)}{\text{actoccr}(r)} \tag{3}$$

For <(At_1, v_1) (At_2, v_3), C_1> rule item, the suppcount = 2 and actoccr = 3 so, confidence = 2/3.

Any itemset rule *r* that passes the minsupp threshold is said to be a frequent itemset rule, and an actual class association rule is represented in the form: $(At_{i1}, v_{i1})\hat{}(At_{i2}, v_{i2})\hat{} \ldots \hat{}(At_{1m}, v_{im}) \rightarrow C_j$, where the antecedent of the rule is an itemset and the consequent is a class.

3. Breast cancer overview

The most critical issue in the medical field is to diagnose diseases as early as possible. One of these diseases is cancer where early diagnosis can decrease the death ratio in cancer patients. There are commonly two types of cancers: benign cancer and malignant cancer.

Breast cancer is one of the most dangerous cancers for women in most countries. Worldwide, it is the main form of cancer in females that is affecting 10% of all women at some phase of their lives. It is the second most common reason of cancer death in women. The malignant tumor grows when cells in the breast tissue split and produce with no control on cells death and cells splitting up. Although experts do not know the precise reasons of most breast cancers, they know some of the risk features that escalate the likelihood of a woman developing breast cancer ([12,36].

Table 2
Comparisons of CBA, CMAR, MCAR, FACA, ECBA and WCBA algorithms.

Algorithm	Data Layout	Rule Discovery	Ranking Strategy	Pruning	Prediction Method	Classifier
CBA	Horizontal	Apriori candidate generation	Confidence, support, rules generated first	Database coverage (M1 method)	Maximum likelihood	Primary
CMAR	Horizontal	FP-growth approach	Confidence, support, rules cardinality	Database coverage (M1 method), redundant rule	CMAR multiple label	Primary
MCAR	Vertical	Tid-list intersections	Confidence, support, cardinality, class distribution frequency	Database coverage (M1 method)	Maximum likelihood	Primary
FACA	Vertical	Diffset	Least number of features, confidence, support, first occurrence	Database coverage (M1 method)	All exact match prediction method	Primary
ECBA	Horizontal	Optimized Apriori	HM measure, confidence, support, rules generated first	Database coverage (M1 method)	Maximum likelihood	Primary
WCBA	Horizontal	Weights by experts, Apriori candidate generation based on weight	HM value (weighted support), confidence, support, rules generated first	Database coverage (M1 method)	Multiple rules based on HM measure	Strong rules & spare rules

Breast cancer research is usually biological and/or clinical in nature. Statistical data of the patients has become a common complement. Predicting the recurrence of a disease is one of the most challenging and interesting tasks to develop data mining applications [2,13]. Enhancing the prediction process for the breast cancer recurrence will decrease the fear of the patient and it plays a key role in his pathological case. Data mining techniques can be employed to predict recurrence of breast cancer in a patient using various indicators data from statistical results [14].

In this research, to test the accuracy of our proposed algorithm and compare it with the other common AC algorithms, we used two breast cancer datasets from the UCI repository: the breast cancer recurrences dataset and the breast cancer diagnosis dataset. More details on the datasets and the experiments will be given in Section 6.

4. Related work

Many AC algorithms have been developed based on rule generation, pruning and prediction techniques. In this section, we will focus on the most common algorithms which we believe are useful to be compared with WCBA while testing it on the breast cancer datasets. We shall present a summary of those algorithms and show how they are different from WCBA.

Liu et al. [9] developed an algorithm based on the AC technique. They called it Classification Based on Association Rules (CBA). The CBA algorithm was implemented based on three stages: the rule generation stage, which used the Apriori algorithm to discover the frequent patterns that represent the class association rules (CARs); the pruning stage, which is responsible for selecting the best rules from the generated rules, and the third stage, which is the prediction stage used to predict unknown instances. This algorithm was applied through many experiments running on many datasets from the UCI machine learning repository.

Classification Based on Multiple Association Rules (CMAR) was proposed by Li et al. [15]. Like the other AC algorithms, it is mainly based on the integration of the association rules and the classification techniques. This algorithm implemented a new technique for the rule-generation and classification phases. At the rule-generation phase, CR-tree and FP-tree algorithms were used to discover patterns instead of using the Apriori algorithm. The classification stage in this algorithm depends on finding the label class for its unknown instances by finding all patterns that can be matched with these instances and then analyzing all of these patterns to predict the class. CMAR was compared with CBA and C4.5 regarding the accuracy measure using UCI datasets. The experiments showed that CMAR performed better than the other AC algorithms.

Thabtah et al. [16] developed the Multi-class Classification Based on Association Rule (MCAR) algorithm to solve the multi-scanning problem of the CBA algorithm to generate association rules. In this algorithm, a single itemset was generated using the Tid-list method which stored the occurrence positions for each item to help the next phase during the generation process without making an extra scanning of the dataset.

The CBA and the MCAR algorithms both have a common step towards finding the frequent patterns. In addition, minimum support and confidence play a key role in their rule evaluation process.

Hadi et al. [17] proposed a new Fast Associative Classification Algorithm (FACA). The Diffset method has been used in the rule generation process to enhance the speed of building the model. Furthermore, it sorts the discovered rules according to the least number of attributes in the left hand side, with confidence, support and rule generated first respectively. FACA also proposed a multiple rules method for the prediction phase to increase the accuracy of the classification process. In particular, this algorithm divides the matched rules to set clusters based on their labels and then selects the label that has a maximum number of rules. To evaluate the performance of their algorithm, the authors compared their work with CBA, CMAR, MCAR and the Enhanced Class Association Rule (ECAR) described in Hadi [37].

The Enhanced CBA (ECBA) algorithm is based on Apriori Optimization and the Statistical Ranking Measure proposed by Alwidian et al. [18]. This algorithm was compared with CBA, CMAR and MCAR algorithms to evaluate its accuracy and speed. Experimental results showed better performance for ECBA compared with the other algorithms in terms of accuracy.

Table 2 compares WCBA with the previous five AC algorithms. The main differences between WCBA and the other AC algorithms include the use of attributes weights assigned by the subject matter experts and the statistical harmonic mean measure to prioritize the generated rules. While the other AC techniques might withdraw early some good association rules due to the problematic estimated measures that are given by the user, WCBA kept these prioritized rules up till the end and hence improved its accuracy.

Breast cancer prediction using association rules mining was also tackled by many researchers. Shrivastava et al. [3] presented an overview of data mining techniques used on breast cancer datasets. In their article, the authors applied a set of classifiers on a breast cancer dataset contains 699 instances with two classes. Experimental results showed a good performance in the prediction process

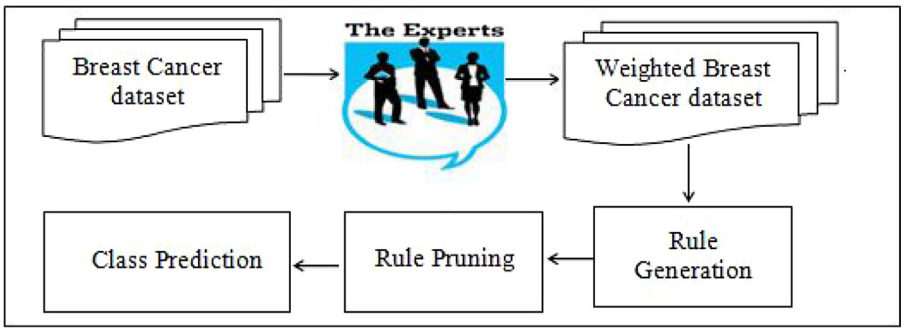

Fig. 1. The main stages of the WCBA algorithm.

for unknown instances. It also showed the potential of using data mining techniques in this area of research.

Majali et al. [5] deployed an association classification technique for the diagnosis and prognosis of cancer. The Frequent Pattern Growth (FP-Growth) algorithm was used to generate association rules to discover hidden patterns for benign and malignant patients to enhance the classification process. This was in addition to using the Decision Tree algorithm to predict if patients' cancers were based on the age values. Kulkarni and Bhagwat [19] investigated the behavior of many classifiers to predict if a patient might face a recurrence or not. Experiments showed the potential of these algorithms regarding the accuracy measure for the classification process. Some of these algorithms were Bayes Net, Naïve Bayes, Kstar, IBK Part and Decision table.

One criticism of the algorithms developed by Shrivastava et al. [3] and Majali et al. [5] is that they both used the support and confidence threshold values which were estimated by the users and hence the ranking process is always affected by these values. In addition, the investigations of Kulkarni and Bhagwat [19] used black box classifiers which did not generate rules that could be interpreted and understandable by users.

To avoid these drawbacks, we proposed WCBA and applied it to the problem of breast cancer prediction as a study case. The basic idea of WCBA is to replace the traditional support-confidence structure of the association rule mining model by the weighted model. Attributes that have major influence in determining the breast cancer disease were assigned the highest priority by subject matter experts. We also use a statistical harmonic mean (HM) measure to prioritize the association rules at the pruning and generation phases. Hence, the rules with weak attributes were eliminated from the top rules generated in the rule pruning stage.

In addition to the above AC algorithms, the literature has many more to list. Few to mention are the following algorithms: Classification based on Predictive Association Rules (CPAR) [20], Negative-Rules [21], Live and Let Live (L^3) [22], Multi-class, Multi-label Associative Classification (MMAC) [38], 2-PS [23], Class Based Associative Classification Approach (CACA) [24], Boosting Association Rules (BCAR) [25], a Multiclass Associative Classification algorithm (MAC) [26], a novel associative classification model based on a fuzzy frequent pattern mining algorithm [27], Enhanced Associative Classification based on Incremental Mining algorithm (E-ACIM) [28] and Multiple Minimum Supports (MMSCBA) [29].

5. Weighted association rule mining

In this section, we describe WCBA, a weighted classification based on association rules. First, we look at the motivation behind the weighted association rule mining (WARM) then we discuss WCBA in more details.

5.1. Introduction to weighted association rule mining

In traditional mining rules, whether a rule is important or not depends on the count of its itemsets in a database. Traditional mining rules consider support and confident measures to find out frequent itemsets. They also assume all items are having equal significance. On the other hand, social science and business market researchers have a different point of view. Whether a rule is interesting or not can depend on quantitative aspects (i.e. the number of appearances of an item in a database) and on qualitative aspects (i.e. the human interpretation), rather than totally based on the database. Weights can represent the knowledge for the items in a dataset.

Wang et al. [30] proposed the Weighted Association Rule Mining (WARM) approach. It generalizes the traditional association rule mining by assigning weights to the items in a database. Weights are used to reflect the importance of an item in the database. In this approach a dataset goes through a weight generation algorithm which uses domain knowledge to assign weights to the items within the dataset. The weights are then used as input into the WARM algorithm, which applies the weights to the items, and uses weighted support to generate a list of interesting rules. The goal is to guide the mining process to those significant associations involving items with significant weights and uses weighted support to generate a list of interesting rules [31].

Most data items do not come with pre-assigned weights [32]. Weights are subject to adjustment by human experts in the domain. This is interesting as experts can assign different weights and hence generate different rules. In some cases, using domain knowledge to determine accurate weights for all items may be impractical especially when the dataset has huge number of items. In this case, a semi-automated or automated approach can be used [33,34].

Based on these reasons, we enhanced CBA and proposed a new weighting algorithm named WCBA, in this paper.

5.2. The proposed WCBA algorithm

The WCBA technique aims to enhance the accuracy of association classifiers based on an efficient rule weighting technique. It has been developed to solve the problem of the estimated support and confidence values that are used to generate class association rules. Our technique was developed based on the following assumption:

Assumption. we assume that the WCBA technique will differentiate between the attributes in any dataset based on two factors: 1) the importance of the attribute in the application domain and 2) the weight assigned for each attribute.

First, we explain our assumption through the following scenario then we explain how WCBA works. Suppose that we have two attributes to decide whether or not one can buy a car and let these attributes being the job title and the gender of the buyer. In

this particular scenario, it is more obvious that the job title of the buyer is more important than the gender and hence, we should assign a higher weight for the job title attribute compared with the gender attribute. Now, the possible values assigned to the gender attribute are only two values: male and female, whereas the job title attribute might have five possible values or more.

Furthermore, the values associated with the gender attribute will exceed the estimated minimum support and could be generated as rules, while most values of the job title attribute may not exceed the estimated measures: support and confidence. Thus, a very strong rule could be eliminated and weak rules could be generated, which at the end might effect on the accuracy measure.

Given the above assumption and the example, we can now explain the mechanism of WCBA in more details. Like the other AC algorithms, WCBA functions in three general stages as depicted in Fig. 1. These stages include: rule generation, rule pruning and prediction. However, WCBA has some differences from the other AC algorithms at these stages (cf. Table 2).

5.3. Detailed description of the WCBA algorithm

Algorithm1 shows the workflow of the WCBA algorithm which involves three stages: (1) rule generation, (2) rule pruning and (3) class prediction.

Algorithm1: WCBA

```
1  Dataset T with n training objects
2  WCBA (T, n)
3  {
4    Divide (T) → [Training-data, Test-data] // T is divide into a training data and testing data
5    S = empty set
6    S = Generate-Rule (Training-data, n, minimum-support)
7    Rule-Pruning (S, minimum-confidence) → [strong-rules, spare-rules]
8    Prediction (strong-rules, spare-rules, Test-data)
9  }
```

Stage 1: Rule generation

Unlike the other AC algorithms and before generating the association rules, weights need to be assigned to dataset attributes using subjective measures identified by experts in their field. Attribute weights can be assigned using a scale from 1 to 10 based on three different measures: high, medium and low. A high measure denotes a scale value from 8 to 10 points, medium denotes a value from 4 to 7 points and a low measure denotes a value from 1 to 3 points. We will show how these measures were used to prioritize the breast cancer recurrences and diagnosis datasets. Algorithm2 shows how the WCBA algorithm generates the rules from a weighted datasets.

Based on Algorithm2, the main input to the rule generation algorithm is (Training-data, n, minimum-support), where n is the size of the training-data. In the first step, WCBA generates all candidate single-items from the training-data (line6). The support and the weighted support values for each item are then calculated (line8–10). Multiplying the support of each item with its weight generates its weighted support value (line9). The computed values promote the algorithm to assign high priority for the most critical features so they appear among the top generated rules.

In the next step, WCBA finds the frequent single-itemset rules. Those are the rules that have weighted support greater than or equal the given minimum support and hence to be inserted in S', while the other rules will be eliminated (line11). This step will be repeated to find the candidate and frequent two-itemset rules from S_k, where k is the number of items in the rule.

Algorithm2: Rule Generation

```
1  Rule-Generation (Training-data, n, minimum-support)
2  {
3    S'= empty set
4    k=1
5    Do{
6      Sk = generate all candidate k-itemset
7      For each item r in Sk
8        Support (r) = suppcount(r) / n
9        Weight(r) = (Weight(item1)+···+Weight(itemk)) / k
10       Weighted support (r) = Weight(r) * Support (r)
11       If Weighted support (r) >= minimum-support
12         S'= S'+ r
13       End if
14     End for
15     k=k+1
16   }
17   while (Sk-1 != empty)
18   Return S'
19 }
```

Stage 2: Rule Pruning

Algorithm3 shows how WCBA prunes the rules from the set of the generated rules. In the first step, the confidence for each frequent rule is calculated (line4 and 5). Then it **e**liminates every rule that has a confidence value less than the minimum-confidence. The rest of the successful rules will be added to the Class Association Rules (CARs) (line6 and 7). Next, it finds the harmonic mean (HM) measure for each frequent rule (line10 and 11).

A good argument for using the HM measure is to overcome the problem of the estimated minimum-support and minimum-confidence usually used by the other AC algorithms. In these algorithms rules that have support or confidence less than the estimated minimum-support and minimum-confidence even with very slight values will not be generated in the final rules set. For instance, if the minimum-confidence = 0.6 and the minimum-support = 0.2, then if there is a rule with confidence = 0.59 and support = 0.6 this rule will not be generated. Therefore, the HM measure which produces a value between the support and confidence is used by the WCBA algorithm.

Finally, to get the strongest rules, all generated rules are sorted based on their HM measure values. For the rules that have the same HM values, they will be sorted based on their confidence, support

and their first occurrence respectively (line13). First occurrence in this context refers to the rule that has been generated first. The final step in this stage (line14) is to apply the M1 method for data coverage to separate the rules into two sets: strong-rules and spare-rules [9].

```
1   Rule-Pruning (S, minimum-confidence)
2   {
3     CAR= empty set
4     For each r in S
5       Confidence (r) = suppcount(r) / actoccr(r)
6       If Confidence (r) >= minimum-confidence
7         CAR= CAR+ r
8       End If
9     End for
10    For each r in CAR
11      HM = 2*(Weighted Support (r)*Confidence (r)) / Weighted support (r)+Confidence(r)
12    End for
13    Sort (CAR) // Sort all generated rules based on their HM values in descending order
14    M1 (CAR) → [strong-rules, spare-rules] // (Liu et al., 1998)
15    Return [strong-rules, spare-rules]
16  }
```

Stage 3: Class Prediction

Algorithm4 explains how the WCBA algorithm predicts the class of a given instance. For any given instance *i*, WCBA checks the set of strong-rules for a rule that can classify *i* (line4-7). Next, it splits the rules into groups based on the class labels (line11). Then it computes the average HM value for each class (line12) and selects the maximum one (line13).

If WCBA fails to find a matching rule in the strong rules set, then it continues searching the set of the spare rules (line15-26). Otherwise, the given instance will be predicted as the default class where the default class is the class that has the maximum frequency (line29). If more than one class has the same frequency, then the class that has the first occurrence will be selected. The first class occurrence in this context refers to the class that occurred first in the training dataset. WCBA ignores any missing value for any tested instance.

Algorithm4: Prediction

```
1   Prediction (strong-rules, spare-rules, Test-data)
2   {
3     ST-rules= empty set
4     For each i in Test-data
5       For each r in strong-rules
6         If r matches i Then
7           ST-rules= ST-rules +r
8         End if
9       End for
10      If ST-rules != empty
11        Split the ST-rules into groups based on the class values.
12        Compute the average of the HM value for each class.
13        Select the maximum HM value
14      End If
15      Else
16      {
17        For each r in spare-rules
18          If r matches i Then
19            ST-rules= ST-rules +r
20          End if
21        End for
22        If ST-rules != empty
23          Split the ST-rules into groups based on the class values.
24          Compute the average of the HM value for each class.
25          Select the maximum HM value
26        End If
27      }
28      Else
29        Select default class
30    End for
31  }
```

5.4. A running example

The following example illustrates how the WCBA algorithm works. The same scenario can be applied to any domain under the condition of assigning attributes weights by the subject matter experts. To begin, let's assume we have the dataset (T) as shown in Table 3 with a minimum support = 4 and a minimum confidence = 0.5.

We would like to predict the class value for instance *i* which has the attribute values *Age*='senior', *Income*='middle' and *Has a car*='n'. For this instance, Table 3 shows a 'no' label assigned to the class *Buy*. Now, we would like to see whether the WCBA algorithm is capable to predict the same result for this instance or not?

In the automotive industry, for instance, analysts might want to mine the rules based on the importance of the attributes in the related dataset. So they might be more interested in the rules that contain "Income" attribute more than the "Age" of a person

Table 3
Dataset sample (*T*) with seven training objects.

Age	Income	Has a car	Buy/Class
senior	middle	N	yes
youth	low	Y	no
junior	high	Y	yes
youth	middle	Y	yes
senior	high	N	yes
junior	low	N	no
senior	middle	N	no

Table 4
Assigned weights to all attributes of the dataset *T*.

Attribute name	Importance	Assigned weight
Age	low	3
Income	high	10
Has a car	medium	7

when both are associated with buying more cars. It is more probable that *high* Income → buying more cars has more support than *youth* Age → buying more cars. Hence, in this dataset the "Income" attribute should receive a higher weight than the "Age" attribute to reflect its importance. Therefore, weights can provide the users with a convenient way to indicate the importance of the attributes, and obtain more interesting rules. Based on this argument, Table 4 shows the suggested weights assigned to all attributes of the dataset according to their importance to the dataset.

The following is a step-by-step illustration on how the WBCA algorithm works:

S1. Generate all candidate single-itemset rules by computing the support (i.e. instance count) and the weighted support (i.e. instance count * attribute's weight) for each rule as shown in Table 5. The following example shows how to calculate the weighted support for a particular rule *r*.

Example on calculating the weighted support (r): let's take the first row of Table 5;

Rule *r* = (senior → yes), k = 1 (single-itemset rule), support = 2 & the assigned weight to attribute "Age"=3 (c.f. Table 4), then from Algorithm2, line#9 and 10:

$$\text{Weight}(r) = \frac{\text{Weight}(\text{item1}) + \ldots + \text{Weight}(\text{item}k)}{k}$$

$$\text{Weightedsupport}(r) = \text{Weight}(r) * \text{Support}(r)$$

$$\text{Weight (senior} \rightarrow \text{yes)} = 3/1 = 3$$

$$\text{Weighted support (senior} \rightarrow \text{yes)} = (3 * 2) = 6$$

S2. Find the frequent single-itemset rules where a rule in this matter is any rule that has a weighted support greater than or equal to the given minimum support as shown in Table 6.

S3. Find the candidate two-itemset rules with their support and weighted support as shown in Table 7.

S4. Find the frequent two-itemset rules where a rule in this matter is any rule that has weighted support greater than or equal to the given minimum support as shown in Table 8.

S5. Find the candidate three-itemset rules using the same procedure explained in the previous steps. The candidate three-itemset rules are shown in Table 9. To create candidate three-itemset rules, only look at rules that have the same first item (in candidate k, the first k-2 items must match).

S6. Find the frequent three-itemset rules as explained before. Table 10 shows the results of this step.

S7. To create candidate four-itemset rules, only look at rules that have the same first two items. From Table 10 there are no rules which have the same first two items. Hence, the algorithm concludes and stops generating rules.

S8. Keep all rules that satisfy the minimum confidence (≥0.4) in CAR and eliminate the others as shown in Table 11. The fol-

Table 5
Candidate single-itemset rules with support and weighted support.

Candidate single- itemset rules	Support	Weighted support
senior → yes	2	3*2 = 6
senior → no	1	3*1 = 3
youth → yes	1	3*1 = 3
youth → no	1	3*1 = 3
junior → yes	1	3*1 = 3
junior → no	1	3*1 = 3
middle → yes	2	10*2 = 20
middle → no	1	10*1 = 10
low → yes	0	10*0 = 0
low → no	2	10*2 = 20
high → yes	2	10*2 = 20
high → no	0	10*0 = 0
n → yes	2	7*2 = 14
n → no	2	7*2 = 14
y → yes	2	7*2 = 14
y → no	1	7*1 = 7

Table 6
The frequent single-itemset rules with weighted support.

Frequent single-itemset rules	Weighted support
senior → yes	6
middle → yes	20
middle → no	10
low → no	20
high → yes	20
n → yes	14
n → no	14
y → yes	14
y → no	7

Table 7
The candidate two-itemset rules with support and weighted support.

Candidate two-itemset rules	Support	Weighted support
senior, middle → yes	1	6.5*1 = 6.5
senior, high → yes	1	6.5*1 = 6.5
senior, n → yes	2	5*2 = 10
senior, y → yes	0	5*0 = 0
middle, high → yes	0	*ignore same attribute*
middle, n → yes	1	8.5*1 = 8.5
middle, y → yes	1	8.5*1 = 8.5
middle, low → no	0	*ignore same attribute*
middle, n → no	1	8.5*1 = 8.5
middle, y → no	0	8.5*0 = 0
low, n → no	1	8.5*1 = 8.5
low, y → no	1	8.5*1 = 8.5
high, n → yes	1	8.5*1 = 8.5
high, y → yes	1	8.5*1 = 8.5
n, y → yes	0	*ignore same attribute*
n, y → no	0	*ignore same attribute*

Table 8
The frequent two-itemset rules with weighted support.

Frequent two-itemset rules	Weighted support
senior, middle → yes	6.5
senior, high → yes	6.5
senior, n → yes	10
middle, n → yes	8.5
middle, y → yes	8.5
middle, n → no	8.5
low, n → no	8.5
low, y → no	8.5
high, n → yes	8.5
high, y → yes	8.5

Table 9
Candidate three-itemset rules with support and weighted support.

Candidate three-itemset rules	Support	Weighted support
senior, middle, high → yes	0	*ignore same attribute*
senior, middle, n → yes	1	6.7*1 = 6.7
senior, high, n → yes	1	6.7*1 = 6.7
middle, n, y → yes	0	*ignore same attribute*
low, n, y → no	0	*ignore same attribute*
high, n, y → yes	0	*ignore same attribute*

Table 10
The frequent three-itemset rules with support and weighted support.

Frequent three-itemset rules	Weighted support
senior, middle, n → yes	6.7
senior, high, n → yes	6.7

Table 11
The rules in CAR which satisfy the minimum confidence.

Rules	Support	Weighted support	Confidence
senior → yes	2	6	0.67
middle → yes	2	20	0.67
middle → no	1	10	0.33 (*eliminated*)
low → no	2	20	1
high → yes	2	20	1
n → yes	2	14	0.5
n → no	2	14	0.5
y → yes	2	14	0.67
y → no	1	7	0.33 (*eliminated*)
senior, middle → yes	1	6.5	0.5
senior, high → yes	1	6.5	1
senior, n → yes	2	10	0.67
middle, n → yes	1	8.5	0.5
middle, y → yes	1	8.5	1
middle, n → no	1	8.5	0.5
low, n → no	1	8.5	1
low, y → no	1	8.5	1
high, n → yes	1	8.5	1
high, y → yes	1	8.5	1
senior, middle, n → yes	1	6.7	0.5
senior, high, n → yes	1	6.7	1

lowing example shows how to calculate the confidence value for a particular rule *r*.

Example on calculating the confidence (*r*): let's take the first row of Table 11 and with reference to Table 3;

Rule *r* = (senior → yes), suppcount (senior → yes) = 2, actoccr (senior) = 3, then from Algorithm3, line# 5:

$$Confidence\,(r) = \frac{\text{suppcount}\,(r)}{\text{actoccr}\,(r)}$$

$$\text{Confidence (senior} \rightarrow \text{yes)} = 2/3 = 0.67$$

S9. Sort the rules in CAR based on HM value as shown in Table 12. If more than one rule has the same HM measure value, the rules will be sorted based on confidence, support and the rule first occurrence respectively. The following example shows how to calculate the harmonic mean value for a particular rule *r*.

Example on calculating the harmonic mean measure for a particular rule: let's take the first row of Table 12;

Rule *r* = (low → no), Weighted support (*r*) = 20, Confidence (*r*) = 1, then from Algorithm3, line# 11:

$$HM\,(r) = \frac{2 * (\text{WeightedSupport}\,(r) * \text{Confidence}\,(r))}{\text{Weightedsupport}\,(r) + \text{Confidence}\,(r)}$$

$$\text{HM (low} \rightarrow \text{no)} = 2 * 20 * 1/(20 + 1) = 40/21 = 1.90$$

Table 12
Rules sorted on the HM values.

Order	Rules	Support	Weighted support	Confidence	HM values
1	low → no	2	20	1	1.90
2	high → yes	2	20	1	1.90
3	middle, y → yes	1	8.5	1	1.79
4	low, n → no	1	8.5	1	1.79
5	low, y → no	1	8.5	1	1.79
6	high, n → yes	1	8.5	1	1.79
7	high, y → yes	1	8.5	1	1.79
8	senior, high, n → yes	1	6.7	1	1.74
9	senior, high → yes	1	6.5	1	1.73
10	middle → yes	2	20	0.67	1.30
11	y → yes	2	14	0.67	1.28
12	senior, n → yes	2	10	0.67	1.26
13	senior → yes	2	6	0.67	1.21
14	n → yes	2	14	0.5	0.97
15	n → no	2	14	0.5	0.97
16	middle, n → yes	1	8.5	0.5	0.94
17	middle, n → no	1	8.5	0.5	0.94
18	senior, middle, n → yes	1	6.7	0.5	0.93
19	senior, middle → yes	1	6.5	0.5	0.93

Table 13
The set of the strong rules.

Strong rules	Coverage	Rank
low → no	*row 2 and 6*	1
high → yes	*row 3 and 5*	2
middle, y → yes	*row 4*	3
middle → yes	*row 1*	4
n → no	*row 7*	5

Table 14
The set of the spare rules.

Spare Rule	Rank
low, n → no	1
low, y → no	2
high, n → yes	3
high, y → yes	4
senior, high, n → yes	5
senior, high → yes	6
y → yes	8
senior, n → yes	9
senior → yes	10
n → yes	11
middle, n → yes	12
middle, n → no	13
senior, middle, n → yes	14
senior, middle → yes	15

S10. Apply the M1 method for data coverage to split the rules into two sets, strong rules and spare rules as shown in Tables 13 and 14.

S11. As for now, assume one needs to predict the class of an unknown instance with values [senior, middle, n]. The algorithm proceeds to check the potential strong rules set that can match this instance. Finally, it found the following two rules: middle → yes, and n → no. Then, the two rules are clustered into two groups based on the type of the target class '*Buy*' as follows:

Cluster (yes) has the rule: middle → yes, and

Cluster (no) has the rule: n → no

Computing the average of the HM values of cluster Buy = 'yes' returned (1.3), while the same value for cluster Buy = 'no' returned (0.97). Hence, this instance is assigned a class value = 'yes'. Accordingly the computed value complied with the actual value from Table 3.

To show the usefulness of the spare rules set, let's explain with the following example: assume we need to predict the class of an unknown instance with values [junior, ?, y], where '?' denotes a

Table 15
Breast cancer recurrences dataset features.

No.	Name of attribute	Number of Values	Possible Values
1	Age	9	10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99
2	Menopause	3	lt40, ge40, premeno
3	Node-caps	2	yes, no
4	Deg-malig	3	1, 2, 3
5	Tumor size	12	0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59
6	Inv-nodes	13	0–2, 3–5, 6–8, 9–11, 12–14, 15–17, 18–20, 21–23, 24–26, 27–29, 30–32, 33–35, 36–39
7	Breast	2	left, right
8	Breast-quad	5	left-up, left-low, right-up, right-low, central
9	Irradiat	2	yes, no
10	Class	2	no-recurrence-events, recurrence-events

Table 16
Breast cancer diagnosis dataset features.

No.	Name of attribute	Domain
1	Clump Thickness	1–10
2	Cell Size Uniformity	1–10
3	Cell Shape Uniformity	1–10
4	Marginal Adhesion	1–10
5	Single Epithelial Cell Size	1–10
6	Bare Nuclei	1–10
7	Bland Chromatin	1–10
8	Normal Nucleoli	1–10
9	Mitoses	1–10
10	Class	benign and malignant

missing value. First, our algorithm ignores the missing value and proceeds to check the set of the strong rules to find some matching rules. As it failed to find any rule in the strong rules set, it then tried to find some matching rules in the spare set. Finally, it found only one matching rule: y → yes. Hence, this instance is assigned a class value Buy = 'yes'. Accordingly the predicted value complied with the actual value from Table 3.

6. Experimental results

We performed an extensive analysis to assess the accuracy of the WCBA algorithm. The WCBA algorithm was compared with five AC algorithms, namely, CBA, MCAR, CMAR, FACA and ECBA. The comparison was based on a set of experimental results in terms of model accuracy.

We conducted our experiments on a 3 GHz i3 PC with a 4GB main memory. WCBA was developed using Java within the WEKA tool [39]. All AC algorithms compared with WCBA were solely developed by their own developers. The minimum support and minimum confidence parameters were used for testing all algorithms.

6.1. The breast cancer datasets

To test our technique, we used two breast cancer datasets from the UCI repository: breast cancer recurrences and breast cancer diagnosis. Breast cancer recurrences dataset contains 10 attributes and 286 instances as shown in Table 15. Where, Breast cancer diagnosis dataset contains 10 attributes and 699 instances as shown in Table 16. Figs. 2 and 3 visualize the distribution of the breast cancer datasets attributes.

The visualization process in Fig. 2 shows that there are some attributes that will not appear in the final generated rules. This was expected because of the large number of values associated with these attributes such as age, tumor size and inv-nodes. On the other hand, some attributes will be repeated too many times because of the small number of values associated with these attributes which lead to an increase in the support value for these attributes such as: irradiat, breast and node-caps. In Fig. 3, all attributes have the same number of values but the level of importance for each attribute is different.

Table 17
Weights assigned to breast cancer recurrence attributes by KHCC experts.

No.	Name of attribute	Importance	Attribute Weight
1	Age	medium	5
2	Menopause	medium	4
3	Node-caps	high	10
4	Deg-malig	high	9
5	Tumor size	high	8
6	Inv-nodes	high	8
7	Breast	low	2
8	Breast- quad	low	2
9	Irradiat	high	8

Table 18
Weights assigned to breast cancer diagnosis attributes by KHCC experts.

No.	Name of attribute	Importance	Attribute Weight
1	Clump Thickness	high	9
2	Cell Size Uniformity	medium	7
3	Cell Shape Uniformity	medium	5
4	Marginal Adhesion	high	8
5	Single Epithelial Cell Size	medium	6
6	Bare Nuclei	high	10
7	Bland Chromatin	high	9
8	Normal Nucleoli	low	3
9	Mitoses	medium	6

To pinpoint the importance of the attributes of the breast cancer datasets and to help us assign accurate weights for each attribute, we asked for the help of the subject matter experts from King Hussein Cancer Center (KHCC) located in Amman Jordan. KHCC is considered one of the advanced centers in the Middle East and has experts graduated from the best universities in the world. We discussed the UCI datasets with the experts and we asked them to classify the attributes based on their importance in predicting breast cancers recurrences and diagnosis using the three measures: low, medium and high as explained earlier. The results are depicted in Tables 17 and 18.

6.2. Analysis of the results

6.2.1. Experiment I: comparing WCBA with the most common AC algorithms

In this experiment we investigated the accuracy of WCBA compared with five AC algorithms: CBA, CMAR, MCAR, FACA and ECBA running on the same breast cancer UCI datasets. Figs. 4–6 show the performance of the six algorithms in three different runs where we applied different values for the minimum support while fixing the minimum confidence for each run. The minimum-support percentage was set at 10% to 30% based on recommendations of previous studies [16,6,17] and the confidence was fixed at 50% as recommended in the same studies.

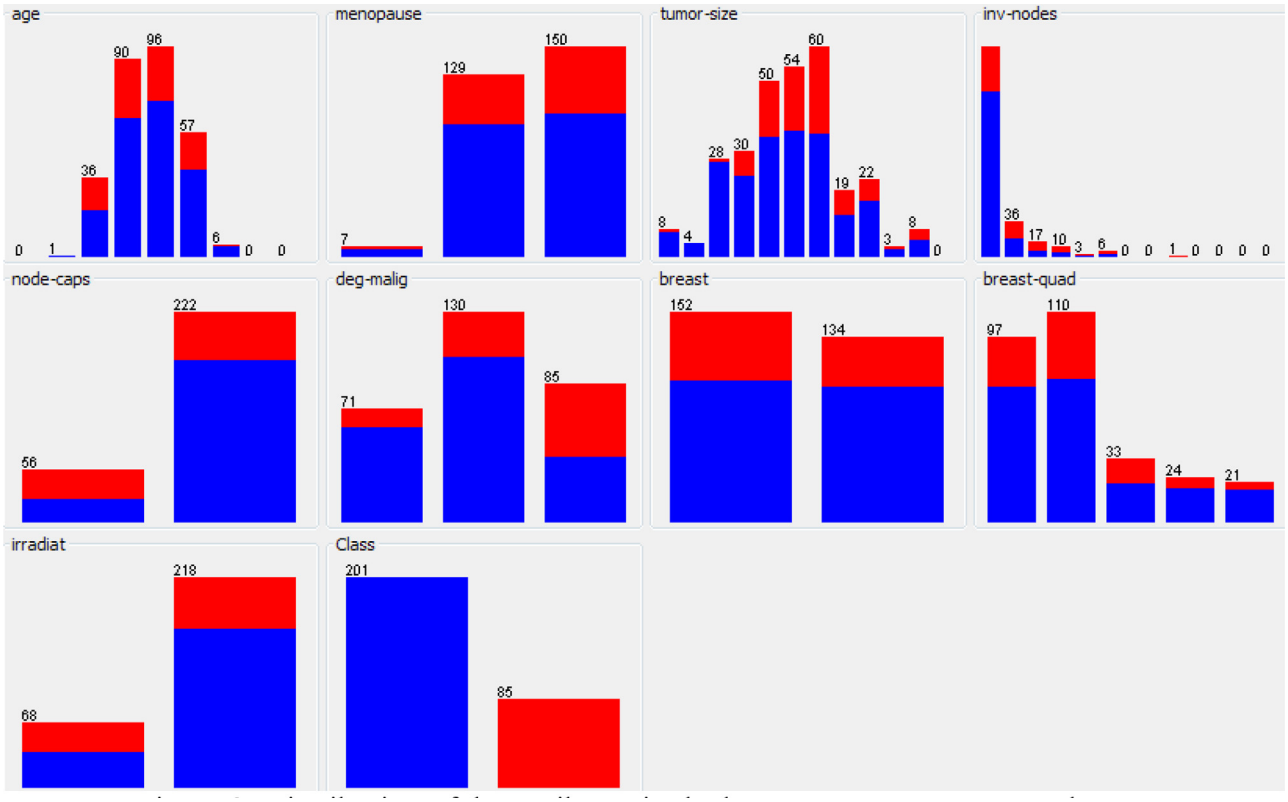

Fig. 2. Distribution of the attributes in the breast cancer recurrences dataset.

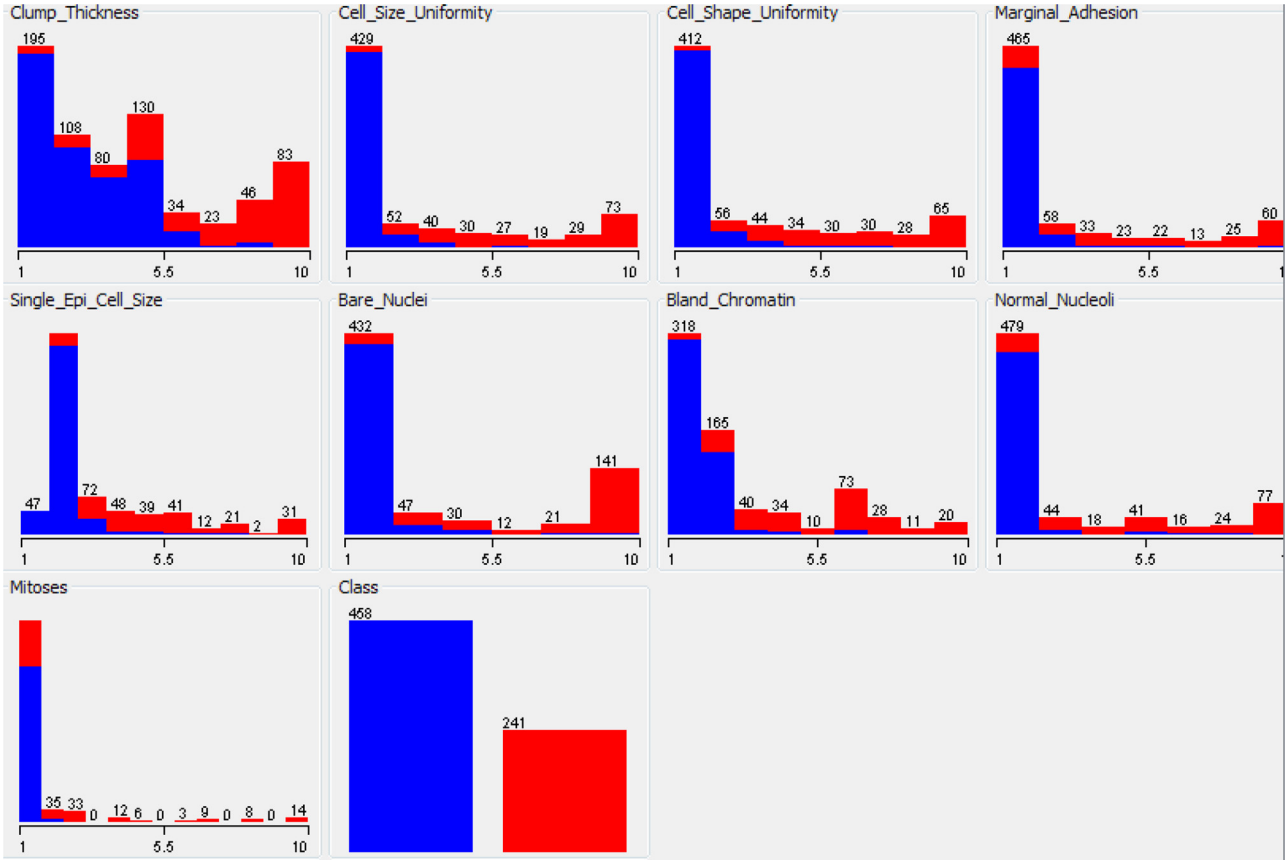

Fig. 3. Distribution of the breast cancer diagnosis dataset.

Table 19
Results of the three runs of all the algorithms running on two breast cancer datasets.

Algorithm	First Run: Fig. 4 Support = 0.1 Confidence = 0.5		Second Run: Fig. 5 Support = 0.2 Confidence = 0.5		Third Run: Fig. 6 Support = 0.3 Confidence = 0.5	
	Recurrence	Diagnosis	Recurrence	Diagnosis	Recurrence	Diagnosis
CBA	0.6279	0.948	0.6279	0.947	0.6279	0.795
CMAR	0.6279	0.912	0.6279	0.887	0.6279	0.814
MCAR	0.6395	0.938	0.6744	0.941	0.6628	0.937
FACA	0.6744	0.951	0.6861	0.941	0.6395	0.940
ECBA	0.6744	0.924	0.6279	0.936	0.6279	0.927
WCBA	**0.6977**	**0.974**	**0.7326**	**0.974**	**0.7093**	**0.968**

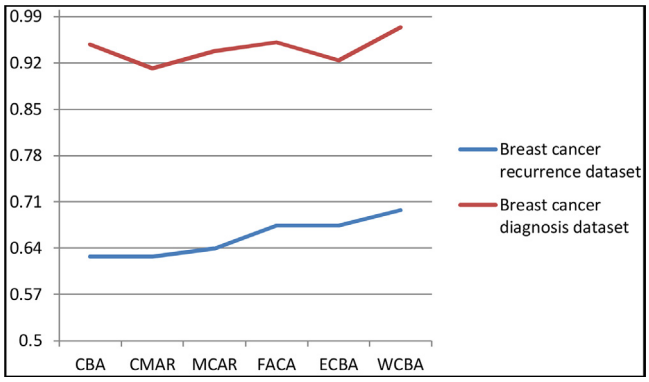

Fig. 4. Accuracy of CBA, CMAR, MCAR, FACA, ECBA and WCBA with minimum support = 0.1 and minimum confidence = 0.5.

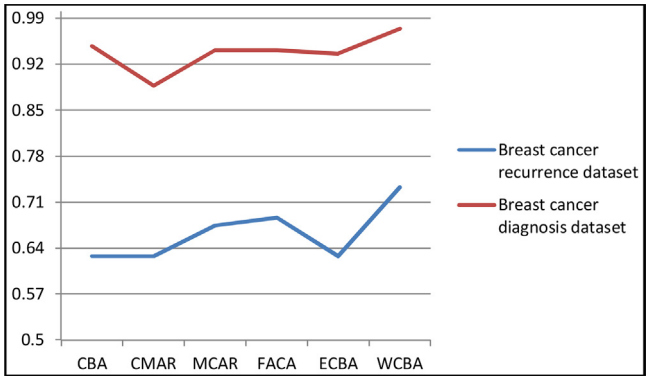

Fig. 5. Accuracy of CBA, CMAR, MCAR, FACA, ECBA and WCBA with minimum support = 0.2 and minimum confidence = 0.5.

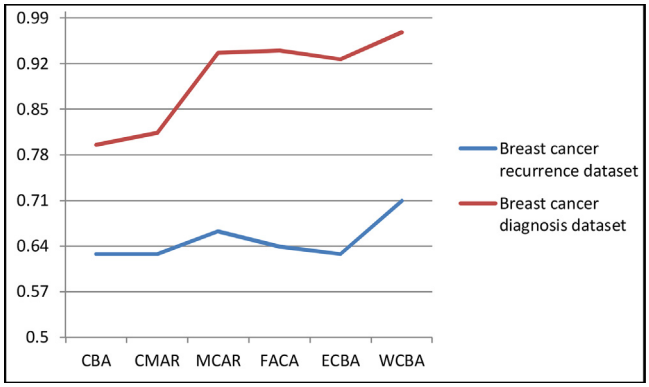

Fig. 6. Accuracy of CBA, CMAR, MCAR, FACA, ECBA and WCBA with minimum support = 0.3 and minimum confidence = 0.5.

From Table 19, the experiments show that for the three runs the WCBA algorithm outperformed the other five AC algorithms with an accuracy = 69.77%, 73.26 and 70.93% respectively while running on the breast cancer recurrences dataset.

Meanwhile, for the breast cancer diagnosis dataset the WCBA algorithm outperformed the other five AC algorithms with an accuracy = 97.4%, 97.4% and 96.8% respectively. Furthermore, while experimenting with the breast cancer recurrences dataset, the FACA and ECBA algorithms came in the second place for the first run with an accuracy = 67.44%. The MCAR and FACA algorithms came in the second place for the second run with an accuracy = 67.44% and 68.61% respectively. In the third run the MCAR algorithm came in the second place with an accuracy of 66.28%.

Additionally, while experimenting with the breast cancer diagnosis dataset, the FACA algorithm came in the second place for the first run with an accuracy of 95.1%. The CBA algorithm came in the second place for the second run with an accuracy of 94.7%. In the third run the FACA algorithm came in the second place with an accuracy of 94.0%.

The reason behind why WCBA outperformed all the AC algorithms in terms of accuracy was due to the new weighting technique which gave the highest priority to the attributes that have the highest impact on determining the breast cancer disease while eliminated the weak attributes from the top rules list generated at the rule pruning process as shown in Figs. 7 and 8. Another noticeable issue in favor of WCBA was that its performance was not affected by increasing the minimum support value.

It is worthwhile noticing, in Fig. 7, it is obvious that the top five rules generated by the WBCA algorithm contained the attributes with the highest impact on the breast cancer as they were identified by the subject matter experts from KHCC. These filtered rules have a major role in the accurate class prediction and hence enhancing the accuracy of the classifier. Fig. 7 also shows an agreement with our expectations where the weak attributes such as “breast” and “breast quad” were eliminated from the rules as decided by the KHCC experts.

Unlike WCBA, the other AC algorithms generated top rules that were associated with weak attributes such as “breast” and “breast quad” while they were running on the breast cancer recurrences dataset as shown in Figs. 9 and 10. We believe that these rules were the reason for decreasing the classification accuracy of these algorithms.

Furthermore, we may notice in Fig. 7, that the top two rules generated by WBCA were neglected by the other AC algorithms although they had some critical attributes according to the experts. This was mainly because of the drawback estimated minimum-support and minimum-confidence values used by the AC algorithms. Furthermore, our classifier used the HM measure as a statistical measure to order the generated rules to enhance the pruning process. This measure assigns a new weight for each rule by using the support and confidence values. Finally, the WCBA algorithm did not eliminate the generated rules that did not pass the database coverage method. These rules were stored in a spare rule

```
Classification Rules (ordered):
===========================

1.      deg-malig=1 5 0 menopause=ge40 1 1 irradiat=no 6 1  ==> Class=no-recurrence-events
2.      deg-malig=1 5 0 menopause=ge40 1 1  ==> Class=no-recurrence-events     conf:(0.91),
3.      age=50-59 0 4 menopause=ge40 1 1 inv-nodes=0-2 3 0  ==> Class=no-recurrence-events
4.      deg-malig=1 5 0 irradiat=no 6 1 node-caps=no 4 1  ==> Class=no-recurrence-events
5.      menopause=ge40 1 1 inv-nodes=0-2 3 0 irradiat=no 6 1  ==> Class=no-recurrence-events
```

Fig. 7. The first five rules generated by the WCBA algorithm with minimum support = 0.2 and minimum confidence = 0.5 running on the breast cancer recurrences dataset.

```
Classification Rules (ordered):
===========================

1.      Bare_Nuclei='(-inf-1.5]' 5 0 Single_Epi_Cell_Size='(-inf-2.5]' 4 0  ==> Class=benign     conf:(1),
2.      Cell_Size_Uniformity='(-inf-1.5]' 1 0 Bare_Nuclei='(-inf-1.5]' 5 0  ==> Class=benign     conf:(1),
3.      Cell_Shape_Uniformity='(-inf-1.5]' 2 0 Bare_Nuclei='(-inf-1.5]' 5 0  ==> Class=benign     conf:(1)
4.      Clump_Thickness='(-inf-4.5]' 0 0 Marginal_Adhesion='(-inf-1.5]' 3 0  ==> Class=benign     conf:(1)
5.      Bland_Chromatin='(-inf-2.5]' 6 0 Marginal_Adhesion='(-inf-1.5]' 3 0 Single_Epi_Cell_Size='(-inf-2.
```

Fig. 8. The first five rules generated by the WCBA algorithm with minimum support = 0.2 and minimum confidence = 0.5 running on the breast cancer diagnosis dataset.

```
Generated Rules :
*       0-2     no      1       left    *       no      recurrence-events:2     no-recurrence-events:30
*       0-2     no      1       left    recurrence-events:1     no-recurrence-events:1
*       *       no      1       *       *       no      recurrence-events:5     no-recurrence-events:25
*       0-2     *       1       left    *       no      recurrence-events:2
*       0-2     no      *       left    left_low        no      recurrence-events:5     no-recurrence-even
*       *       no      1       *       recurrence-events:1     no-recurrence-events:1
50-59   0-2     no      *       *       *       no      recurrence-events:6     no-recurrence-events:24
*       0-2     no      *       left    left_low        recurrence-events:1     no-recurrence-events:4
*       0-2     *       2       right   *       no      recurrence-events:5     no-recurrence-events:22
```

Fig. 9. The top rules generated by the MCAR algorithm with minimum support = 0.2 and minimum confidence = 0.5 on breast cancer recurrences dataset.

```
Classification Rules (ordered):
===========================

1.      deg-malig=1 2 0 breast=left 3 0 irradiat=no 5 1  ==> Class=no-recurrence-events     conf:(0.88),
2.      deg-malig=1 2 0 irradiat=no 5 1  ==> Class=no-recurrence-events     conf:(0.86),  (55),
3.      deg-malig=1 2 0 breast=left 3 0  ==> Class=no-recurrence-events     conf:(0.84),  (31),
4.      breast-quad=left_up 4 0 deg-malig=2 2 1  ==> Class=no-recurrence-events     conf:(0.84),  (36),
5.      age=50-59 0 4 breast=right 3 1 irradiat=no 5 1  ==> Class=no-recurrence-events     conf:(0.83),
6.      deg-malig=1 2 0  ==> Class=no-recurrence-events     conf:(0.83),  (59),
```

Fig. 10. The top rules generated by the FACA algorithm with minimum support = 0.2 and minimum confidence = 0.5 on breast cancer recurrences dataset.

set to be used in the classification process when the strong rules fail to predict new instances.

6.2.2. Experiment II: modifying the most common AC algorithms using the WBCA weighting model

In this experiment, we aim to prove that the attribute weighting approach used by WBCA is effective and has an impact on the accuracy of class prediction. Therefore, we repeated Experiment I running on the same datasets after we applied the WBCA weighting model to all the five AC algorithms. The experimental results showed better performance in most cases for the modified AC algorithms as shown in Figs. 11–13 and the summary in Table 20. The new experiment provides evidence that the WCBA approach works well and is very promising.

Experiment II reveals that for the first two runs on the breast cancer diagnosis dataset, the modified CBA algorithm outper-

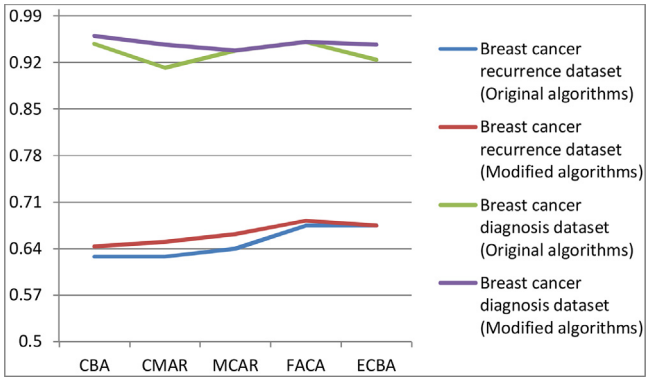

Fig. 11. Accuracy of original and modified algorithms with minimum support = 0.1 and minimum confidence = 0.5.

Table 20
Results of the modified three runs of all five algorithms running on two breast cancer datasets.

Algorithm	First Run: Fig. 11 Support = 0.1 Confidence = 0.5				Second Run: Fig. 12 Support = 0.2 Confidence = 0.5				Third Run: Fig. 13 Support = 0.3 Confidence = 0.5			
	Recurrence		Diagnosis		Recurrence		Diagnosis		Recurrence		Diagnosis	
	old	new	old	new	old	new	old	new	old	new	old	New
CBA	0.6279	0.9430	0.948	0.960	0.6279	0.6400	0.947	0.947	0.6279	0.6279	0.795	0.854
CMAR	0.6279	0.6500	0.912	0.947	0.6279	0.6400	0.887	0.918	0.6279	0.6500	0.814	0.892
MCAR	0.6395	0.6620	0.938	0.938	0.6744	0.6744	0.941	0.952	0.6628	0.6924	0.937	0.937
FACA	0.6744	0.6820	0.951	0.951	0.6861	0.7326	0.941	0.952	0.6395	0.6740	0.940	0.948
ECBA	0.6744	0.6744	0.924	0.947	0.6279	0.6540	0.936	0.940	0.6279	0.6630	0.927	0.940
WCBA	**0.6977**		**0.974**		**0.7326**		**0.974**		**0.7093**		**0.968**	

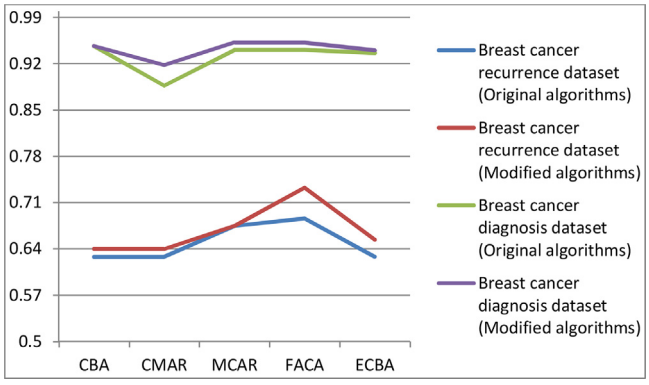

Fig. 12. Accuracy of original and modified algorithms with minimum support = 0.2 and minimum confidence = 0.5.

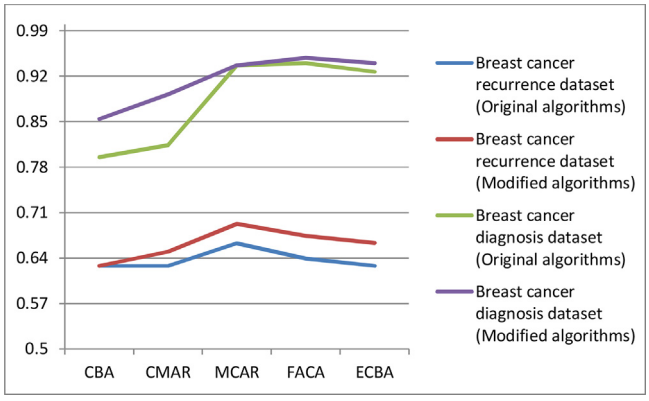

Fig. 13. Accuracy of original and modified algorithms with minimum support = 0.3 and minimum confidence = 0.5.

formed the other modified AC algorithms with an accuracy = 96% and 94.7%. In addition, the first two runs on the breast cancer recurrences dataset showed that the modified FACA algorithm outperformed the other modified AC algorithms with an accuracy = 68.20% and 73.26% respectively.

7. Conclusion and future works

In this article we investigated a new association rule based algorithm named WBCA. The WCBA technique aimed to enhance the accuracy of association classifiers based on an efficient rule weighting technique. It has been developed to solve the shortcomings of the estimated support and confidence values that are used in all AC algorithms to generate class association rules. Our new technique was developed based on a weighted method to select more useful association rules and a statistical measure for pruning rules. All these features contributed towards improving the performance of the WCBA algorithm in terms of accuracy. We have tested the WCBA algorithm against five common AC algorithms running on breast cancer datasets as a case study. In all experiments, WCBA outperformed the other AC algorithms. From the experimental results, the following conclusions can be drawn:

(1) WCBA applied an attribute weighting scheme to prioritize the important attributes from the least important ones. The decision on assigning weights is made by subject matter experts in the domain. It has been shown by the experiments that the weighting scheme could positively improve the accuracy of the AC algorithms.
(2) The WCBA algorithm applied two classifiers; one running against the strong rules list containing the most important attributes and the other running against the spare rules list containing the least important attributes. This approach limited the use of the default class rule which normally has an unacceptable error rate.
(3) WCBA used a statistical HM measure to solve the estimated minimum-support and minimum-confidence measure problem which allowed WCBA to use multiple rules to predict unseen instances.

For future work, we will test WCBA on different domains starting with a real breast cancer dataset form KHCC. Furthermore, we plan to investigate the use of different weighting, pruning and prediction techniques and we aim to examine their impact on different fields. Finally, we are in the process of developing a user interface for the WEKA tool to facilitate the weighting mechanism for datasets.

Acknowledgments

The authors would like to thank KHCC experts for their help and support in all phases of this research. Also they would like to thank the editor and the reviewers for their valuable comments and suggestions to improve the manuscript.

References

[1] S. Sarvestani, A. Safavi, M. Parandeh, M. Salehi, Predicting breast cancer survivability using data mining techniques, 2010 2nd International Conference in Software Technology and Engineering (ICSTE) vol. 2 (2010) V2–V227, IEEE.
[2] S. Gupta, D. Kumar, A. Sharma, Data mining classification techniques applied for breast cancer diagnosis and prognosis, Indian J. Comp. Sci. Eng. (IJCSE) 2 (2) (2011) 188–195.
[3] S. Shrivastava, A. Sant, R. Aharwal, An overview on data mining approach on breast cancer data, Int. J. Adv. Comp. Res. 3 (13) (2013) 256–262.
[4] S. Shajahaan, S. Shanthi, V. Chitra, Application of Data Mining techniques to model breast cancer data, Int. J. Emerg. Technol. Adv. Eng. 3 (11) (2013) 362–369.
[5] J. Majali, R. Niranjan, V. Phatak, O. Tadakhe, Data mining techniques for diagnosis and prognosis of cancer, Int. J. Adv. Res. Comp. Commun. Eng. 4 (3) (2015) 613–616.

[6] N. Abdelhamid, A. Ayesh, W. Hadi, Multi-label rules algorithm based associative classification, Parallel Process. Lett. 24 (01) (2014) 1450001–14500021.
[7] B. Ma, H. Zhang, G. Chen, Y. Zhao, B. Baesens, Investigating associative classification for software fault prediction: an experimental perspective, Int. J. Softw. Eng. Knowl. Eng. 24 (1) (2014) 61–90.
[8] S. Taware, C. Ghorpade, P. Shah, N. Lonkar, M. Bk, Phish detect: detection of phishing websites based on associative classification (AC), Int. J. Adv. Res. Comp. Sci. Eng. Inf. Technol. 4 (3) (2015) 384–395.
[9] B. Liu, W. Hsu, Y. Ma, Integrating classification and association rule mining, in: Proceedings 4th International Conference on Knowledge Discovery and Data Mining, August 1998, New York, NY, 1998, pp. 80–86.
[10] N. Abdelhamid, A. Ayesh, F. Thabtah, Emerging trends in associative classification data mining, Int. J. Electron. Electr. Eng. 3 (1) (2015) 50–53.
[11] M. Lichman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2013 http://archive.ics.uci.edu/ml.
[12] V. Krishnaiah, D. Narsimha, D. Chandra, Diagnosis of lung cancer prediction system using data mining classification techniques, Int. J. Comp. Sci. Inf. Technol. 4 (1) (2013) 39–45.
[13] W. Huang, W. Chen, C. Lin, W. Ke, F. Tsai, SVM and SVM ensembles in breast cancer prediction, PLoS One 12 (1) (2017) e0161501.
[14] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, A knowledge-based system for breast cancer classification using fuzzy logic method, Telemat. Inf. 34 (4) (2017) 133–144.
[15] W. Li, J. Han, J. Pei, CMAR: Accurate and efficient classification based on multiple class-association rules, in: Data Mining, ICDM 2001, Proceedings IEEE International Conference, November 2001, 2001, pp. 369–376, IEEE.
[16] F. Thabtah, P. Cowling, Y. Peng, MCAR: multi-class classification based on association rule, in: Computer Systems and Applications, 3rd ACS/IEEE International Conference, January 2005, 2005, pp. 33–40, IEEE.
[17] W.E. Hadi, F. Aburub, S. Alhawari, A new fast associative classification algorithm for detecting phishing websites, Appl. Soft Comput. 48 (2016) 729–734.
[18] J. Alwidian, B. Hammo, N. Obeid, Enhanced CBA algorithm Based on Apriori Optimization and Statistical Ranking Measure. Proceeding of 28th International Business Information Management Association (IBIMA) conference on Vision 2020: Innovation Management, Development Sustainability, and Competitive Economic Growth, Seville, Spain, 2016, pp. 4291–4306.
[19] S. Kulkarni, M. Bhagwat, Predicting breast cancer recurrence using data mining techniques, Int. J. Comp. Appl. 122 (23) (2015).
[20] X. Yin, J. Han, CPAR: classification based on predictive association rules, in: SDM, May 2003, 2003, pp. 331–335.
[21] M.L. Antonie, O.R. Zaïane, An associative classifier based on positive and negative rules, in: Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, June 2004, 2004, pp. 64–69, ACM.
[22] E. Baralis, S. Chiusano, P. Garza, On support thresholds in associative classification, in: Proceedings of the 2004 ACM Symposium on Applied Computing, March 2004, 2004, pp. 553–558, ACM.
[23] T. Qian, Y. Wang, H. Long, J. Feng, 2-ps based associative text classification, in: Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery, August 2005, 2005, pp. 378–387, Springer, Berlin, Heidelberg.
[24] Z. Tang, Q. Liao, A new class based associative classification algorithm, IMECS 36 (2) (2007) 685–689.
[25] Y. Yoon, G.G. Lee, Text categorization based on boosting association rules, Semantic Computing, 2008 IEEE International Conference (2008) 136–143, IEEE.
[26] N. Abdelhamid, A. Ayesh, F. Thabtah, S. Ahmadi, W. Hadi, MAC: a multiclass associative classification algorithm, J. Inf. Knowl. Manage. 11 (2) (2012) 1250011.
[27] M. Antonelli, P. Ducange, F. Marcelloni, A. Segatori, A novel associative classification model based on a fuzzy frequent pattern mining algorithm, Expert Syst. Appl. 42 (4) (2015) 2086–2097.
[28] M.A. Al-Fayoumi, Enhanced associative classification based on incremental mining algorithm (E-ACIM), Int. J. Comp. Sci. Issues (IJCSI) 12 (1) (2015) 124.
[29] L.Y. Hu, Y.H. Hu, C.F. Tsai, J.S. Wang, M.W. Huang, Building an associative classifier with multiple minimum supports, SpringerPlus 5 (1) (2016) 1–19.
[30] W. Wang, J. Yang, P.S. Yu, Efficient mining of weighted association rules (WAR), Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2000) 270–274.
[31] F. Tao, F. Murtagh, M. Farid, Weighted association rule mining using weighted support and significance framework, Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003) 661–666.
[32] K. Sun, F. Bai, Mining weighted association rules without preassigned weights, Proc. IEEE Trans. Knowl. Data Eng. 20 (4) (2008) 489–495.
[33] R. Pears, Y.S. Koh, Weighted association rule mining using particle swarm optimization, Pacific-Asia Conference on Knowledge Discovery and Data Mining (2011) 327–338, Springer, Berlin, Heidelberg.
[34] Y.S. Koh, R. Pears, G. Dobbie, Automatic item weight generation for pattern mining and its application, Developments in Data Extraction, Management, and Analysis (2013) 187–207, IGI Global.
[35] F. Thabtah, P. Cowling, S. Hammoud, Improving rule sorting, predictive accuracy and training time in associative classification, Expert Syst. Appl. 31 (2) (2006) 414–426.
[36] R.R. Szvarça, S.O. Ioshii, D.R. Carvalho, W.F. Sokoloski, Temporal association rules in breast cancer, Iberoam.J. Appl. Comput. 4 (3) (2016).
[37] W. Hadi, W. ECAR: a new enhanced class association rule, Adv. Comput. Sci. Technol. 8 (1) (2015) 43–52.
[38] F.A. Thabtah, P. Cowling, Y. Peng, MMAC: A new multi-class, multi-label associative classification approach, in: In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference, November, 2004, 217–224. IEEE, 2004.
[39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11 (1), (2009) 10–18.