Network Modeling Analysis in Health Informatics and Bioinformatics (2020) 9:34
https://doi.org/10.1007/s13721-020-00237-8

ORIGINAL ARTICLE

Breast cancer classification with reduced feature set using association rules and support vector machine

Abderrahmane Ed-daoudy[1] · Khalil Maalmi[1]

Received: 18 November 2019 / Revised: 13 April 2020 / Accepted: 16 April 2020 / Published online: 21 May 2020

Abstract
In the last few years, machine learning is one of the driving forces of science and industry, but increasing of data requires paradigm shifts in traditional methods in the application of machine learning techniques on this data especially in healthcare field. Furthermore, with the availability of different clinical technologies, tumor features have been collected for breast cancer classification. Therefore, feature selection and accuracy improvement have become a challenging and time-consuming task. In this paper, the proposed approach has two stages. In the first, Association Rules (AR) are used to eliminate insignificant features. In the second, several classifiers are applied to differentiate the incoming tumors. Feature space dimension is reduced from nine to eight and four attributes by using AR. In test stage, threefold cross-validation method was applied to the Wisconsin Breast Cancer Diagnostic (WBCD) dataset from the University of California Irvine machine learning repository to evaluate the proposed system performances. The correct classification rate obtained with Support Vector Machine (SVM) model with AR shows the highest classification accuracy (98.00%) for eight attributes and 96.14% for 4 attributes. The results show that the proposed approach can be used for feature space reduction and saving of time during the training phase leading to better accuracy and fast automatic classification systems.

Keywords Data mining · Association rules · Support vector machine · Feature selection · Breast cancer · Classification · Wisconsin breast cancer diagnostic

1 Introduction

Breast cancer has become a common disease around the world and is the most deadly and frequent cancer in women. It appears in women in the form of tumors in the breast. It can be diagnosed and detected by physical examination or image analysis (Nguyen et al. 1998). So, early detection of the stage of cancer allows treatment which could lead to high survival rate. Mammography is currently the most effective imaging modality for breast cancer screening. However, the traditional approach to cancer diagnosis depends highly on the experience of doctors and their visual inspections, which is limited due the human mistakes.

Knowledge and mining information from large database has been recognized by many researchers as a key research topic in database system and machine learning. The potential of data-mining techniques and the usefulness of knowledge detection from medical data for the disease diagnosis problems have been identified by World Health Organization (WHO) (Gulbinat 1997). Expert systems developed by data-mining techniques are valuable tools that have been successful for the disease diagnosis. Computer-aided diagnostic tools are intended to help physicians in order to improve the accuracy of the diagnosis (Tartar et al. 2013; Kilic et al. 2009). A study was carried out to demonstrate that the machine learning may improve the accuracy of diagnosis. In Brause's work, the results show that the most experienced physician diagnoses with an accuracy of 79.97%, while the correct diagnosis achieved using machine learning is 91.1%. This study demonstrated that machine learning could improve the accuracy of diagnosis (Brause 2001).

In order to improve the accuracy of breast cancer classification as benign and malignant, handle the dramatically increasing tumor feature data and information, a number

✉ Abderrahmane Ed-daoudy
a.eddaoudy@gmail.com

Khalil Maalmi
khalil.maalmi@usmba.ac.ma

[1] Laboratoire d'Intelligence Artificielle, Sciences de données et Systèmes Émergents (LIASSE) - ENSAF, Sidi Mohamed Ben Abdellah University, Fez, Morocco

of researchers have turned to data-mining technologies and machine learning approaches for predicting breast cancer. Data mining is the process of extracting hidden interesting patterns from massive database. Techniques of data mining help to process the data and turn them into useful information, it has been shown to be highly applicable in the real world. In the literature, data-mining techniques are widely used in diagnosing breast cancer based on tumor feature data. Due the number of descriptive tumor features increases, the computational time increases rapidly as well. Feature extraction is the key for pattern recognition and classification. Filtering and extracting feature vectors which contain most of the useful information from the original vector with high accuracy has become a new issue. Feature extraction is the most important phase in data-mining techniques, the best classifier will perform poorly if the features are not chosen well. As the characteristics of the tumor can be described in as much detail as possible, the information can be redundant but without significant contribution to the final classifier which leads to a longer computation time for tedious calculations. At this stage, breast cancer requires better accuracy and fast automatic diagnostic.

Healthcare is often the subject of researchers and doctors, both in diagnosis and treatment. Early detection of cancer is fundamental for a rapid response and better chances of cure. Unfortunately, because of the symptoms of the disease at the beginning are absent, early detection of cancer is often difficult. Mammography is a very complex process needing human resources and material resources that have not been effective. Based on available medical records data, early detection can be simplified by exploiting past cases to predict current situations based on data-mining techniques and machine learning.

This work demonstrates the application of association rules on breast cancer dataset for reducing the dimension of feature vector and support vector machine to obtain fast automatic diagnostic systems. The rest of this paper is organized as follows: in Sect. 2, we discuss about related works, Sect. 3 discusses about the proposed work and the dataset considered for the experiment, the proposed machine learning model is discussed also. Section 4 presents results of our method. Finally, in Sect. 5 we conclude the paper and present future work.

2 Related work

In recent decades, different types of techniques and methods of computer-aided diagnosis (CAD) systems have been proposed to improve the accuracy of breast cancer classification aiming at assisting doctors in making diagnostic decisions. Recently, machine learning is the modern science of discovering patterns and predicting from big data based on statistics, data drilling, pattern recognition and predictive analyses. Data mining (DM) is one of the steps of knowledge discovery for extracting implicit patterns from vast, incomplete and noisy data (Fayyad et al. 1996). Consequently, to date, several expert systems have been developed which are used in the medical field for the classification of breast cancer, the diagnosis is done using different methodology such as statistical, fundamental, support vector machine, neural network, probabilistic, fuzzy systems and hybrid of these techniques (Arya and Tiwari (2016).

There are many techniques to predict and classify breast cancer pattern. Albrecht et al. used a combination of perceptron algorithm with simulated annealing and reported accuracy of 98.8% (Albrecht et al. 2002). Karabatak and Ince (2009) combined an association rule with neural network. Association rules are used for reducing the dimension of breast cancer database and Neural Network(NN) is used for intelligent classification, they have proposed two different techniques to eliminate inputs. These are named as AR1 and AR2, respectively. AR1 feature selection removes the attributes if it depends on others with threshold value of confidence and support, AR2 uses all input parameters but not all their records, the correct classification rate of proposed system is 95.6%. Pauline and Santha kumaran Paulin and Santhakumaran (2011) used Feed Forward Artificial Neural Networks and back propagation algorithm to train the network. WBCD is used to evaluate the performance of the network for various training algorithms. Koyuncu and Ceylan implemented rotation forest artificial neural network (ANN) using 9 classifiers and achieved a classification accuracy of 98.05% (Koyuncu and Ceylan 2013). In Stoean and Stoean (2013), support vector machine (SVM) and evolutionary algorithm were used, and obtained accuracy was around 97%.

Zheng et al. (2014) proposed a hybrid method (K-means + SVM) and achieved a classification accuracy of 97.38% using tenfold cross-validation. The above authors used highly computational complex methods for classification, they used K-means to extract useful information and SVM to diagnose the tumor. Xue, et al. proposed a novel technique for initializing and updating in Particle Swarm Optimization (PSO) for feature selection to achieve an accuracy of 94.74% (Xue et al. 2014). A model based on Rough Set (RS) and SVM classifier (RSSVM) was developed by Chen et al. for breast cancer diagnosis. They used RS as a feature selection technique to select the best features of dataset. Further improvement for the accuracy of diagnostic system was obtained by SVM (Chen 2014). Dheeba et al. investigated a new classification approach for detection of breast abnormalities in digital mammograms using Particle Swarm Optimized Wavelet Neural Network (PSOWNN). The proposed abnormality detection algorithm is based on extracting laws texture energy measures from mammograms and classifying the suspicious regions by applying a pattern classifier. They achieved 93.67%, 92.10% and 94.16% for accuracy, specificity, and sensitivity, respectively (Dheeba et al. 2014).

Karabatak Karabatak (2015) developed a weighted NB classifier for the application of breast cancer detection. Using fivefold cross-validation, their method obtained 99.11%, 98.25%, and 98.54% for sensitivity, specificity and accuracy, respectively. Bhardwaj and Tiwari (2015) proposed a method based on NN technique for solving breast cancer classification problem. Nahato, Harichandran, and Arputharaj Nahato et al. (2015) used a Rough Set indiscernibility relation method with Back Propagation Neural Network (RS-BPNN). This model works in two stages. The first stage handles missing values to obtain a smooth dataset and to select appropriate attributes from the clinical dataset by indiscernibility relation method. In the second stage classification is done using backpropagation neural network. The accuracy obtained from the proposed method was 98.6% on breast cancer dataset. Mert, Kiliç, Bilgili, and Akan explored features reduction properties of independent component analysis (ICA) on breast cancer decision support system. They proved that a one-dimensional features vector obtained from (ICA) causes Radial Bases Function Neural Network (RBFNN) classifier to be more distinguishing with the increased accuracy from 87.17 to 90.49% (Mert et al. 2015).

Most of this work uses the classification for breast cancer diagnosis with a totality of dataset features. As the amount of available data has increased dramatically, feature extraction and selection has become a new issue for both the accurate result and information redundancy which is the key to simplifying the training part of the data-mining process and improving the performance without changing the main body of data-mining algorithms Optimizat (1999). In this paper, an approach combining AR and SVM has been proposed to perform breast cancer classification problem. The proposed method consists of two parts. Firstly, the dataset feature space dimension is reduced by using AR which consist of elimination of unnecessary and redundant data. Secondly, support vector machines classifier has been applied on the new reduced dataset for predicting breast cancer.

3 Materials and methods

3.1 Association rules

The association rules (AR) is a technique that allows the user to discover the correlation between a different object in databases (Agrawal et al. 1996). An AR has the following general logical form:

If *Conditions* **then** *Results* and presented in the form of antecedent and consequence $X_1 \wedge X_2 ... \rightarrow X_n$. An AR is an X to Y implication relationship between two sets of attributes *X* and *Y*. This rule indicates that transactions that contain the attributes of the set *X* tend to contain the attributes of the set *Y*. AR algorithms are used to find all frequent itemsets then generate significant rules that explain the presence of some attributes according to the presence of other attributes in database by satisfying some parameters like minimum support threshold and minimum confidence threshold, it was initiated by Agrawal et al. (1993) for the first time, to analyze transactional databases. It is a statement of the form $A \rightarrow B$, where $A, B \subset I$ such that, $A \neq \emptyset$, $B \neq \emptyset$ and $A \cap B = \emptyset$. The set A is called antecedent of the rule, the set B is called the consequent of the rule and I is an itemset.

Let $L = \{i_1, i_2, \cdots i_m\}$ be a set of attributes and D be a database of transaction, a transaction T is a set of attributes such that $T \subseteq L$, let X and Y be two set of attributes , the association rule, $X \Longrightarrow Y$ with $X \subseteq T, Y \subseteq T, X \cap Y = \emptyset$ is at least defined by two measures of quality: **support** and **confidence** given by Eqs. 1 and 2, respectively. The support defined as the proportion of transaction in the database, which contains the items A, while the confidence determines how frequently items in B appear in the transaction that contains A.

$$Supp(A \rightarrow B) = Supp(A \cup B) = \frac{|t(A \cup B)|}{t(A)} \tag{1}$$

$$Conf(A \rightarrow B) = \frac{Supp(A \cup B)}{Supp(A)} \tag{2}$$

Let D be a database, I an itemset and S_{min} a minimal support then I is considered to be frequent if and only if $Support(I) \geq S_{min}$. AR aims at discovering the patterns of co-occurrence of attributes in a database. For example, an AR in breast cancer dataset may be in 47% of transactions, 99% of the patients who have the value of 1st and 6th parameters as 1 also have the value of 9th input parameter equal to 1.

3.2 Apriori algorithm

AR generation is usually split up into two separate steps. First, minimum support is applied to find all frequent itemsets in a database. Second, these frequent itemsets and the minimum confidence constraint are used to form rules. Exploiting this property, Apriori algorithm can find all frequent itemsets.

The Apriori algorithm is a data-mining algorithm designed in 1994 by Agrawal et al. (1993). It is used to recognize properties that occur frequently in a dataset and to deduce a categorization. It is the key algorithm for the extraction of AR because it constituted the basis of the majority algorithms that are designed to extract the AR and works iteratively, it first generate itemsets and calculate their frequency, keep those whose frequency exceeds a certain threshold and generate rules from these sets, finally keep those whose reliability exceeds a certain threshold. Algorithm 1 shows the Apriori algorithm presented in Agrawal and Srikant Agrawal and Srikant (1994):

Algorithm 1 Apriori algorithm pseudocode.

```
procedure Apriori (T, minSupport) {
//T is the database
// minSupport is the minimum support
$C_k$ : Candidate itemset of size k
$L_k$ : frequent itemset of size k
$L_1$= {frequent items};
for (k= 2; $L_{k-1}$ !=$\emptyset$; k++) {
$C_k$= candidates_generated from $L_{k-1}$
for each transaction t in database do{
$L_k$ = candidates in $C_k$ with minSupport
}//end for each
}//end for
return $\cup L_k$ ;
}
```

Let D be a database of transaction. Firstly, we have to find out the frequent itemset using Apriori algorithm. Then, AR will be generated using minimum support and minimum confidence. In this step, the Apriori algorithm scans D for count of each candidate C_1 and compares candidate support count with minimum support count for generating 1-itemset frequent pattern. The set of frequent $1-itemsets$ consists of the candidate $1-itemsets$ satisfying minimum support. Generate C_2 candidates from L_1 for generating 2-itemset and so on. The algorithm uses a level-wise search, where k-itemsets are used to explore $(k+1)-itemsets$. Apriori determines frequent itemsets that can be used to determine AR which highlight general trends in the database. The candidate generation algorithm (Algorithm 2) is given as follows (Karabatak and Ince 2009):

Algorithm 2 Candidate generation algorithm.

```
candidates_generated($L_{k-1}$)
$C_k$=$\emptyset$
for all itemsets $X \in L_{k-1}$ and $Y \in L_{k-1}$ do
if $X_1 = Y_1 \wedge ... \wedge X_{k-2} = Y_{k-2} \wedge X_{k-1} < Y_{k-1}$ then
begin
C= $X_1 X_2 ... X_{k-1} Y_{k-1}$
add C to $C_k$
end
delete candidate itemsets in $C_k$ whose any subset is not in $L_{K-1}$.
```

3.3 Support vector machines

Support Vector Machines (SVM) proposed by Vapnik (2013) is a supervised machine learning algorithm which can be used for both classification and regression challenges. It is a linear discriminative classifier that attempts to maximize the margin between classes during training. However, it is mostly used in classification problems because of its advantage of accuracy. Support vectors are the points lying on the margins that are shown in Fig. 1.

The equation for the separating hyperplane is defined by the linear equation:

$$g(x) = w^T x + b, \tag{3}$$

where x describes data points, y is a coefficient vector, and b shows offset from the origin. In case of linear SVM, $g(x) \geq 0$ for the closest point on the one of the class, $g(x) \leq 0$ for the closest point belongs to another class. The generalized SVM model is revised for this problem to search the classifier as follows Cortes and Vapnik (1995):

$$maximize \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(X_i, Y_j) \right] \tag{4}$$

$$subject\ to \left[\sum_{i=1}^{n} \alpha_i y_j = 0, \ \ 0 \leq \forall \alpha_i \leq L \right], \tag{5}$$

where X is the training vector, Y is the label associated with the training vectors, a is the parameters vector of classifier hyperplane, $K(X_i, Y_j)$ is the kernel function equal to $\{\Phi(X_i), \Phi(Y_i)\}$, where $\Phi(X)$ represents the mapping of input vectors, onto the kernel space X for measuring the distance between the training vector X_i and Y_j, and L is a penalty parameter to control the number of misclassification. After the dimension of the feature space has been reduced, and the data set with new features has been rebuilt, different data-mining techniques can be applied in this step.

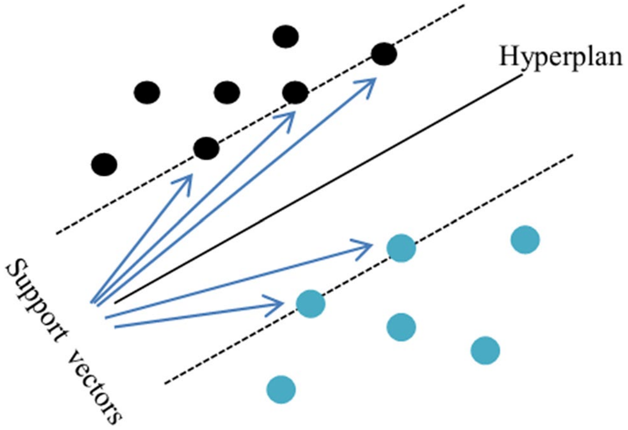

Fig. 1 The separating hyperplane with support vectors

3.4 Wisconsin breast cancer database

In this study, the Wisconsin breast cancer database (WBCD) was used and analyzed, it was used for training and testing the machine learning algorithm which predicts breast cancer (WBCD 1995). They have been collected by Wolberg and Mangasarian Wolberg and Mangasarian (1990) at the University of Wisconsin-Madison Hospitals, it was used in many machine learning research works. For each cancer observation, we have the following information detailed in Table 1. We have constructed a labeled dataset with attributes, where class label attribute is labeled with two classes, benign and malignant. The class label attribute values modified to just 0 and 1, where value 1 indicates malignant and value 0 indicates benign, turning it to a binary class dataset, the other features are detailed in Table 1. The dataset consist of 699 records, 241 (34.5%) malignant and 458 (65.5%) benign. Each record in the database has nine attributes. The nine attributes are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state. Table 2 shows the sample WBCD dataset.

In this study, the original nine features of WBCD data and reduced features using AR are deployed to evaluate the classifier performances on breast cancer decision. Thus, the proposed model is shown in Fig. 2.

4 Results and discussion

The proposed work is performed with two stages, first, WBCD pre-processing which consists of feature extraction using AR, and second, by applying machine learning algorithm especially SVM. Using the publically available open source machine learning software, called **Weka** (Waikato environment for knowledge analysis) with Apriori algorithm, AR have been extracted then the dimension of the feature space has been reduced from nine to eight and four attributes based on extracted AR. After that the different data-mining techniques are tested on the reduced datasets using threefold cross-validation and average values were calculated. The implementation was performed using python. The SVM were trained for different C values and different kernel functions until obtaining the best result. The best result was obtained for $C = 1$ with Poly kernel functions in the testing procedure.

4.1 Applying association rules to the WBCD

The selection of appropriate values depends on the data and the user preference. Different datasets can have different values. Generally, we set minimum support with a high value and then decrease them gradually until you find enough rules. The minimum support is not fixed. Apriori algorithm starts with an upper bound minimum support

Table 1 Attributes description of Wisconsin breast cancer data

Attribute number	Attribute description	Values of attributes	Mean	Standard deviation
1	Clump thickness	1–10	3.42	2.82
2	Uniformity of cell size	1–10	3.13	3.05
3	Uniformity of cell shape	1–10	3.20	2.97
4	Marginal adhesion	1-10	2.80	2.86
5	Single epithelial cell size	1–10	3.21	2.21
6	Bare nuclei	1–10	3.46	3.64
7	Bland chromatin	1–10	3.43	2.44
8	Normal nucleoli	1–10	2.87	3.05
9	Mitoses	1–10	1.59	1.71

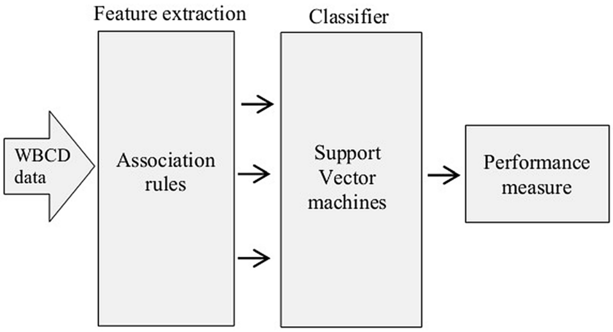

Fig. 2 The basic model involved in this study

Table 2 Sample WBCD dataset (10 attributes)

Attributes									
Clump thickness	Uniformity of cell size	Uniformity of cell shape	Marginal adhesion	Single epithelial cell size	Bare nuclei	Bland chromatin	Normal nucleoli	Mitoses	Label
5	1	1	1	2	1	3	1	1	0
5	4	4	5	7	10	3	2	1	0
8	10	10	8	7	10	9	7	1	1
5	3	3	3	2	3	4	4	1	1

(default 1.0 = 100%), then iteratively decreases it by delta (default 0.05 = 5%). The algorithm stops when lower bound minimum support is reached, or required number of rules has been generated.

In case of minimal confidence, this is obvious because it represents the confidence we want in the rules. The minimum confidence value also depends on domain of knowledge and data. So, usually, we choose values greater than 60%, because we are not interested in a rule that is really less than 60% of confidentiality. In general, the greater the value of the minimum confidence, the rules become more meaningful. On the other hand, with a less confidence value, the algorithm may not generate any association rules. Setting these parameters also depends on how many rules you want.

In our case study, since the dataset is not too large, only 699 transactions, we set the minimum threshold and the minimum confidence to 5% and 80%, respectively, which can potentially explain meaningful and relevant rules. Despite the minimum thresholds, the algorithm stops at a minimum support and confidence of 50% and 99%, respectively, because the number of rules is fixed at 4. We tried to extract 4 most relevant rules with a confidence of 100%. The adjustment of the parameters of the AR generates other rules which will be either significant or not significant, according to corresponding threshold values, in our case study we tried to generate rules which maximize the confidence.

In fact, applying Apriori algorithm on WBCD, we have extracted the following best rules:

1. Attribute8 =1, Attribute9 =1 ⇒ Attribute2 =1 (350 transactions)
2. Attribute2 =1, Attribute8 =1 ⇒ label = 0 (355 transactions)
3. Attribute2 =1, Attribute8 =1, Attribute9 =1 ⇒ label = 0 (349 transactions)
4. Attribute6 =1 ⇒ label = 0 (401 transactions)

We say that when the value of 8th and 9th attributes is 1, the value of 2nd attribute is 1, with a confidence of 100%. As a result from this rule, the 2nd attribute depends on others, 8th and 9th attributes. So we can eliminate it from dataset. We say also that when the value of 2nd and 8th attributes is 1, the class is benign, with a confidence of 100%. This rule presents large itemsets for benign class (78% of all dataset), and when the value of 2nd, 8th and 9th attributes are 1, the class is benign, with a confidence of 100%, this rule presents large itemsets for benign class (76% of all dataset). We say also that when the value of 6th attribute is 1, the class is benign, this rule presents large itemsets for benign class (88% of all dataset). According to these rules, we can reduce the WBCD from nine to eight attributes (all attributes except the second attribute) which are:

- Clump thickness
- Uniformity of cell shape
- Marginal adhesion
- Single epithelial cell size
- Bare nuclei
- Bland chromatin
- Normal nucleoli
- Mitoses

On the other hand, we can reduce the WBCD from nine to four attributes (2nd, 6th, 8th and 9th attribute) which are:

- Uniformity of cell size
- Bare nuclei
- Normal nucleoli
- Mitoses

4.2 Performance measures

There are several ways to evaluate the performance of classifiers. Confusion matrix keeps the correct and incorrect classification results to measure the quality of the classifier, where TP, TN, FP and FN denote true-positive, true-negative, false-positive and false-negative counts, respectively. The most common empirical measure to assess effectiveness is the accuracy for classifiers and it is calculated by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

F-measure is a measure of test accuracy. It considers both precision and the recall to compute. These are calculated by:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{10}$$

4.3 Result without association rules

The experimental results achieved for WBCD dataset are given in Table 3. We get an average accuracy of 94.85% for Decision Tree (DT) (J48), 96.42% for Random Forest (RF), 96.85% for Logistic Regression (LG), 97.42% for Bayes Net (BN), 97.00% for SVM, 95.57% for Multilayer Perceptron (MLP).

Table 3 Results of WBCD data analysis without using AR

Classifiers	Time to build a model (s)	Correctly classified instances	Incorrectly classified instances	TP	TN	FP	FN
BN	0.05	681	18	443	238	3	15
SVM	0.02	678	21	446	232	9	12
DT (J48)	0.01	663	36	436	227	14	22
LG	0.16	669	22	447	230	11	11
RF	0.14	674	25	442	232	9	16
MLP	0.73	668	31	442	226	15	16

Table 4 Performances evaluation metrics without using AR

Classifiers	Precision	Recall	F-measure	Specificity	Accuracy (%)
BN	99.33	96.72	98.01	98.75	97.42
SVM	98.02	97.38	97.70	96.26	97.00
DT (J48)	96.89	95.20	96.04	94.191	94.85
LG	97.60	97.60	97.60	95.43	96.85
RF	98.00	96.51	97.25	96.26	96.42
MLP	96.72	96.51	96.61	93.77	95.57

The performance evaluation metrics of other algorithms and SVM classifiers such as precision, recall, f-measures, specificity and accuracy are given in Table 4 and Fig. 3.

4.4 Result with association rules

In the second test, we first used AR-based feature selection to select the best attributes, and then, we used the same datamining techniques as in the previous section. Experimental results showed that highest classification performances are achieved when SVM is used as classifier. Therefore, the approach proposed in this study is an approach where AR is used for feature reduction and SVM used for classification as given in Fig. 2.

The experimental results achieved for WBCD dataset are given in Tables 5, 6 and Fig. 4. With 8 attributes we get the an accuracy of 95.42% for DT (J48), 96.28% for RF, 97.00% for LG , 97.14% for BN, 98.00% for SVM and 95.42% for MLP. With 4 attributes we get an accuracy of 94.99% for DT (J48), 95.57% for RF, 95.57% for LG , 96.14% for BN, 96.14% for SVM and 96.14% for MLP. From Table 6 results obtained with AR were better than those achieved by the other classifiers without AR. As shown in Table 6, the best classifier performance is maintained with AR with SVM in case of eight attributes with a classification accuracy of 98.00%. The performance measures of other algorithms and SVM classifiers such as time to build a model, correctly classified instances, incorrectly classified instances, TP, TN, FP, FN are given in Table 5 to compare the effect of AR on the classification.

The performance measures of other algorithms and SVM classifiers such precision, recall, f-measures, specificity and accuracy are given in Table 6 and Fig. 4. Table 7 presents AR with SVM performance.

Table 8 gives the confusion matrices showing the classification results of the different classifiers implemented for detection of breast cancer. From these matrices, one can tell the frequency with which a record is misclassified as another.

As shown in Table 9, the proposed method provides stable and high prediction quality in comparison with the

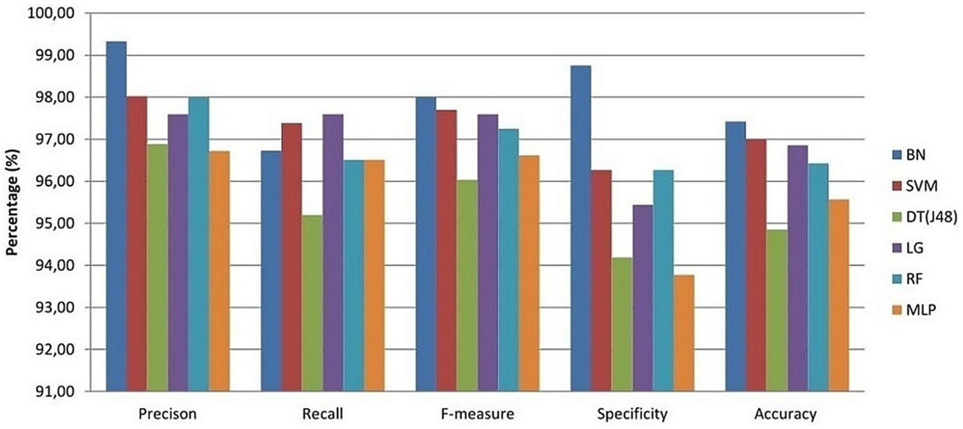

Fig. 3 Performance measures comparison

Table 5 Results of WBCD data analysis using AR

Classifiers	Features space dimension	Time to build a model (s)	Correctly classified instances	Incorrectly classified instances	TP	TN	FP	FN
BN	8	0,02	679	20	442	237	4	16
	4	0,01	672	27	441	231	10	17
SVM	8	0,02	685	14	448	237	5	9
	4	0,01	672	27	446	226	15	12
DT (J48)	8	0,01	667	32	442	225	16	16
	4	0,00	664	35	436	228	13	22
LG	8	0,12	672	21	447	231	10	11
	4	0,10	667	31	445	223	18	13
RF	8	0,10	673	26	444	229	12	14
	4	0,09	668	31	440	228	13	18
MLP	8	0,67	667	32	441	226	15	17
	4	0,38	672	27	439	233	8	19

Table 6 Performances evaluation metrics using AR

Classifiers	Features space	Precision (%)	Recall (%)	F-measure (%)	Specificity (%)	Accuracy (%)
BN	8	99.10	96.51	97.79	98.340	97.14
	4	97.78	96.29	97.03	95.85	96.14
SVM	**8**	**98.90**	**98.03**	**98.46**	**97,93**	**98,00**
	4	96.75	97.38	97.06	93.776	96.14
DT (J48)	8	96.51	96.51	96.51	93.361	95.42
	4	97.10	95.20	96.14	94.606	94.99
LG	8	97.81	97.60	97.70	95.85	97.00
	4	96.11	97.16	96.63	92.53	95.57
RF	8	97.37	96.94	97.16	95.02	96.28
	4	97.13	96.07	96,60	94,60	95,57
MLP	8	96.71	96.29	96.50	93.77	95.42
	4	98.21	95.85	97.02	96.680	96.14

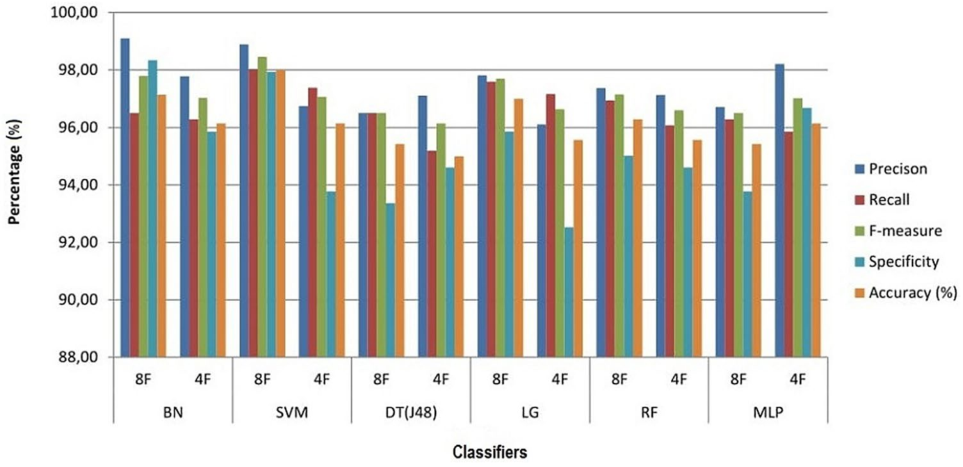

Fig. 4 Performance measures comparison

Table 7 AR with SVM performance comparison

	SVM	AR+SVM	
Feature space dimension	9	8	4
Accuracy (%)	96.99	98.00	96.14

previous experimental results especially those performed by Karabatak and Ince (2009). The performance demonstrated by the ensemble data-mining techniques for breast cancer diagnosis lies in input variable choice and classification method selection. The parameters, which are most appropriate for breast cancer diagnosis, must be utilized as the inputs of the model. For this reason, association rules are appropriate for features selection of the WBCD data in the breast cancer diagnosis. In the second test, where AR were applied, the highest obtained accuracy is 98.00% with the same classifier used without AR.

In the literature, several methods have been proposed for the diagnosis of breast cancer based on classification and clustering. However, in recent years, the amount of data available has increased considerably. Traditional methodologies show their drawbacks on a large-scale dataset. Although the use of these methods for classification needs a high computation time for different training. As the classification is based on relationship between input variables and the target variable, selecting input variables that have the strongest relationship with the target variable will significantly improve the classification accuracy. On the other hand, AR mainly focused on mining and finding interesting associations and relationships among large sets of attributes. It allows to extract the variables that contribute significantly to the target variable. Therefore, the results achieved show that using AR with an efficient classier like SVM is even significant over other methods.

Based on the results achieved, the proposed approach to classify breast cancer not only provides best accuracy in comparison with the previous experimental results with a minimal set of features, but also shows the features reduction in a simple way leading to time savings during the training phase, especially in big data context which has become a great challenge. On the other hand, physicians can also benefit from the representative tumor characteristics extracted by understanding the relationship between different characteristics.

5 Conclusion

Extensive research has been conducted in the medical area to study medical conditions and find precise diagnosis. Data-mining techniques are widely used to uncover patterns from the data, which can then be used for various purposes. Features reduction is an important step in the data preprocessing process. Indeed, for data belonging to a large space, some attributes do not provide any information, others are redundant or correlated. This makes the decision algorithms,

Table 8 Confusion matrices of the classifiers used for detection of breast cancer

Classifier	Feature space dimension	Desired result	Output result	
			Benign records	Malignant records
SVM	9	Benign records	446	12
		Malignant records	9	232
AR+SVM	8	Benign records	448	9
		Malignant records	5	237
AR+SVM	4	Benign records	446	12
		Malignant records	15	226

Table 9 Comparison of the methods and accuracy of previous studies and this study

References	Feature space dimension	Classifier	Accuracy (%)
Karabatak and Ince (2009)	8 (wbcd)	AR-NN	97.40
Zheng et al. Zheng et al. (2014)	6 (wdbc)	K-SVM	97.38
Stoean and Stoean (2013)	–	SVM-evoalgo.	97.23
Prasad and Biswas (2010)	17 (wdbc)	PSO-SVM	97.37
Sweilam et al. (2010)	30 (wdbc)	QPSO + SVM	93.06
Our study	**8** (wbcd)	**AR-SVM**	**98.00**
	4 (wbcd)	**AR-SVM**	**96.14**

ineffective and difficult to interpret. So, features extraction and reduction will significantly improve a learning algorithm's performance. In this work, an approach for detecting breast cancer based on AR and SVM has been proposed. The proposed AR + SVM approach performance is evaluated and compared with SVM model and previous work in this context. Using AR, eight and four important input features have been extracted from the nine original features. Based on threefold cross-validation method, the performances of the proposed approach were evaluated on WBCD. SVM and AR feature selection gave the highest accuracy of 98.00% for eight inputs, while 96.14% is achieved using four inputs. This research demonstrated that the AR can be used for reducing the dimension of feature vector and proposed SVM model can be used to obtain efficient automatic diagnostic systems.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors

References

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Acm sigmod record, volume 22, pages 207–216. ACM

Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast discovery of association rules. Adv Knowl Discov Data Min 12(1):307–328

Agrawal R, Srikant R et al (1994) Fast algorithms for mining association rules. Proc 20th Int Sonf Very Large Data Bases VLDB 1215:487–499

Albrecht A.A, Lappas G, Vinterbo S.A, Wong C, Ohno-Machado L (2002) Two applications of the lsa machine. In: Proceedings of the 9th international conference on neural information processing, 2002. ICONIP'02., volume 1, pages 184–189. IEEE

Arya C, Tiwari R (2016) Expert system for breast cancer diagnosis: a survey. In: 2016 international conference on computer communication and informatics (ICCCI), pages 1–9. IEEE

Bhardwaj A, Tiwari A (2015) Breast cancer diagnosis using genetically optimized neural network model. Expert Syst Appl 42(10):4611–4620

Brause RW (2001) Medical analysis and diagnosis by neural networks. In: International symposium on medical data analysis, pages 1–13. Springer

Chen C-H (2014) A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. Appl Soft Comput 20:4–14

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Dheeba J, Singh NA, Selvi ST (2014) Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach. J Biomed Inform 49:45–52

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Magazine 17(3):37–37

Gulbinat W (1997) What is the role of who as an intergovernmental organisation. In: The coordination of telematics in healthcare. World Health Organisation Geneva, Switzerland

Karabatak M (2015) A new classifier for breast cancer detection based on naïve bayesian. Measurement 72:32–36

Karabatak M, Ince MC (2009) An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl 36(2):3465–3469

Kilic N, Ucan ON, Osman O (2009) Colonic polyp detection in ct colonography with fuzzy rule based 3d template matching. J Med Syst 33(1):9

Koyuncu H, Ceylan R (2013) Artificial neural network based on rotation forest for biomedical pattern classification. In: 2013 36th international conference on telecommunications and signal processing (TSP), pages 581–585. IEEE

Mert A, Kiliç N, Bilgili E, Akan A (2015) Breast cancer detection with reduced feature set. Comput Math Methods Med 2015:1–11

Nahato KB, Harichandran KN, Arputharaj K (2015) Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Comput Math Methods Med 2015:1–13

Nguyen H, Hung W, Thornton B, Thornton E, Lee W (1998) Classification of microcalcifications in mammograms using artificial neural networks. In: Proceedings of the 20th annual international conference of the IEEE engineering in medicine and biology society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286), volume 2, pages 1006–1008. IEEE

Optimizat PJSM (1999) ion: A fast algorithm for training support vector machines. In: Press MIT (ed) Advances jn KerneI Metbods Support Vector I earrljng., Cambridge MA, pp 185–208

Paulin F, Santhakumaran A (2011) Classification of breast cancer by comparing back propagation training algorithms. Int J Comput Sci Eng 3(1):327–332

Prasad Y, Biswas KK (2010) Pso-svm based classifiers: a comparative approach. In: International conference on contemporary computing, pages 241–252. Springer

Stoean R, Stoean C (2013) Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. Expert Syst Appl 40(7):2677–2686

Sweilam NH, Tharwat A, Moniem NA (2010) Support vector machine for diagnosis cancer disease: a comparative study. Egypt Inform J 11(2):81–92

Tartar A, Kilic N, Akan A (2013) Classification of pulmonary nodules by using hybrid features. Comput Math Methods Med 2013

Vapnik V (2013) The nature of statistical learning theory. Springer science & business media, Berlin

WBCD (1995) https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). Accessed 23 July 2017

Wolberg WH, Mangasarian OL (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci 87(23):9193–9196

Xue B, Zhang M, Browne WN (2014) Particle swarm optimisation for feature selection in classification: novel initialisation and updating mechanisms. Appl Soft Comput 18:261–276

Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Syst Appl 41(4):1476–1482