Journal of AI and Data Mining
Vol 7, No 1, 2019, 59-68
DOI: 10.22044/JADM.2018.6489.1763

A New Knowledge-based System for Diagnosis of Breast Cancer by a combination of Affinity Propagation Clustering and Firefly Algorithm

N. Emami[1*] and A. Pakzad[2]

1. Department of Computer Science, Kosar University of Bojnord, Bojnord, Iran.
2. Department of Industrial Engineering, Kosar University of Bojnord, Bojnord, Iran.

Received 04 December 2017; Revised 18 April 2018; Accepted 12 June 2018
**Corresponding author: nasibeh.emami@kub.ac.ir (N. Emami).*

Abstract
Breast cancer has become a widespread disease around the world in young women. Expert systems, developed by data mining techniques, are valuable tools in the diagnosis of breast cancer, and can help physicians for decision-making processes. This paper presents a new hybrid data mining approach to classify two groups of breast cancer patients, malignant and benign. The proposed approach, AP-AMBFA, consists of two phases. In the first phase, the Affinity Propagation (AP) clustering method is used as an instance reduction technique, which can find noisy instances and eliminate them. In the second phase, feature selection and classification are conducted by the Adaptive Modified Binary Firefly Algorithm (AMBFA) for selection of the most related predictor variables to target variables and the Support Vectors Machine (SVM) technique as classifier. It can reduce the computational complexity and speed up the data mining process.
The experimental results on the Wisconsin Diagnostic Breast Cancer (WDBC) datasets show a higher predictive accuracy. The classification accuracy obtained was 98.606%, a very promising result compared to the current state-of-the-art classification techniques applied to the same database. Hence, this method will help physicians in a more accurate diagnosis of breast cancer.

Keywords: *Breast Cancer, Affinity Propagation Clustering, Feature Selection, Binary Firefly Algorithm, Support Vector Machine.*

1. Introduction

Nowadays, the knowledge discovery process has been comprehensively used in medicine to identify and exploit the hidden patterns among a large number of the patients' historical data stored within datasets [1, 2]. A major class of problems in medical science involves the correct diagnosis of disease based upon various tests performed upon the patient [3].
Breast cancer is the second largest cause of cancer deaths and the most frequently diagnosed cancer in young women [4, 5]. This type of cancer happens when cells in the breast tissue divide and grow without their normal control [6].
The correct and early detection of breast cancer can ensure a long survival of the patients [7]. The early detection of breast cancer requires an accurate and reliable diagnosis procedure that allows physicians to classify two groups of breast cancer patients, malignant and benign [8, 9]. There are three common methods available for diagnosing breast cancer: mammography, Fine Needle Aspiration (FNA) biopsy with visual interpretation, and surgical biopsy. The reported sensitivity of mammography varies from 68% to 79% [10], of FNA with visual interpretation from 65% to 98% [11], and of surgical biopsy close to 100%. Therefore, mammography lacks sensitivity, FNA sensitivity varies widely, and surgical biopsy, although accurate, is invasive, time-consuming, and costly [12]. Thus we can say that the quickest and simplest diagnostic tool for breast cancer without any negative aspects of surgical biopsy is an FNA test.
The cytological characteristics extracted from the FNA biopsy and Machine Learning (ML) techniques can be used for the diagnosis of breast cancer [13]. Data mining is one of the analytic processes commonly used to identify, validate,

and prediction of data [14]. The different artificial intelligence techniques for classification also help experts a great deal. Classification systems minimize the possible errors that might be made due to fatigued or inexperienced experts and provide more detailed medical data for examination in a shorter time [15]. Expert systems, developed by data mining techniques, are valuable tools to improve medical decision and assist the physician.

One of the main objectives of an ML algorithm is to build reliable classifiers [16] with good classification accuracy rates. Classification accuracy can decrease in the presence of noise in the training data [17]. Therefore, a data pre-processing phase of the ML algorithm, in which noise can be detected and handled through an appropriate cleaning or correction procedure, is recommended [18, 17]. Noise filtering has been implemented in different forms with different types of classifiers [16, 19] and has been proven to be effective, to some extent, in improving the classification accuracy [20].

In addition to noise filtering, it should be noted that there are many cases that some of the attributes of a dataset are irrelevant or redundant in making a certain decision [21]. Thus selecting an appropriate set of features to represent the main information of original datasets is an important factor that influences the accuracy of classification methods [22].

Feature Selection using Feature Similarity (FSFS) has been proposed by Mitra et al. [23]. They introduced a new similarity measure known as Maximal Information Compression Index (MICI), which was used to iteratively remove some number of features. In [24], FSFS has been used, and the classification accuracies reached with Support Vectors Machine (SVM) on Wisconsin Diagnostic Breast Cancer (WDBC) dataset were 94.41%. He et al. have presented Laplacian Score for Feature Selection (LSFS) [25]. LSFS selects some top-ranking features that have maximum locality preserving power computed in terms of Laplacian score. In WDBC dataset LSFS algorithm for SVM, the classifier reached a 96.87% accuracy [24]. Multi-Cluster Feature Selection (MCFS) is another feature selection algorithm presented by Cai et al [25]. The main motivation of using spectral analysis technique here is to efficiently compute the correlations among different features of a candidate feature subset in an unsupervised manner. Thus identifying the multi-cluster data structure is a major advantage of this approach [26]. An accuracy of 96.68% was obtained with the application of MCFS with SVM classification technique on WDBC [24]. In [27], diagnosis of breast cancer tumor has been conducted based on manifold learning and SVM, and the reported accuracy was 97.3%.

Recursive Feature Elimination (RFE) and Feature Selection Concave (FSC) that have come in [28] report 95.25% and 95.23% accuracies for SVM classification on WDBC, respectively.

Maldonado et al. [28] have proposed an embedded method Kernel-Penalized Support Vectors Machine (KP-SVM) that simultaneously selects relevant features during classifier construction by penalizing each feature's use in the dual formulation of SVM. The classification accuracy obtained by KP-SVM was 97.55% on WDBC. In [29], a feature selection method that combines Data Envelopment Analysis (DEA) and entropy technique has been presented. The classification accuracies of feature selection method with SVM, C5.0, and Logistic Regression (LR) on WDBC dataset were 89.86%, 93.92%, and 95.95%, respectively. Also in [29], the results of the presented model were compared with two other filter feature selection methods (Correlation based Feature Selection (CFS) and Filter). Feature selection is inherently a combinatorial search problem, and there is no polynomial algorithm known to solve this problem. Thus stochastic searching is preferable to achieve the approximate global best solution in polynomial time [30]. In [31], a Particle Swarm Optimization-Kernel Density Estimation (PSO-KDE) model has been presented that hybridizes the PSO and non-parametric KDE-based classifier for diagnosis of breast cancer and is compared with Genetic Algorithm-Kernel Density Estimation (GA-KDE) [31]. Although both reached a 98.45% accuracy, PSO-KDE is better than GA-KDE because of less selected features. Ensemble feature selection based on bi-objective genetic algorithm can be found in [32]. For WDBC dataset, Ensemble-FSGA reached an 82.2% accuracy rate. A SVM-based ensemble algorithm has been presented in [33]. SVM was aggregated into the proposed Weighted Area under the Receiver Operating Characteristic Curve Ensemble (WAUCE) approach. The best average accuracy achieved by the WAUCE model is 97.68% on the WDBC dataset [33].

Although many researchers have carried out diagnosis of breast cancer using machine learning techniques, it is absolutely important to offer a reliable diagnosis model. Thus this paper proposed the new hybrid model including Affinity Propagation (AP) clustering method [34] as

instance reduction technique for the first time and Adaptive Modified Binary Firefly Algorithm (AMBFA) for extracting the optimal feature subset. AMBFA has two parts: the adaptive part that makes balance in exploration and convergence [35] and the modified part that reaches a qualified and fast solution [36]. We used the accuracy rate of SVM as the cost function for the proposed AP-AMBFA algorithm. In the proposed algorithm, the concept of clustering and classification is conjoined and shows the practical application of the AP clustering to reduce noise data for the first time. The main goal of AP-AMBFA is to construct a powerful intelligent model for increasing the predictive accuracy of breast cancer disease classification. The proposed hybrid model is applied on the WDBC dataset in UCI to diagnose whether the tumor of the patient is malignant or benign.

This paper is organized as what follows. In Section 2, the research methodology is presented; this section consists of the prerequisites for the research work. Our proposed adaptive modified binary firefly algorithm is presented in Section 3. In Section 4, the application of this hybrid model is demonstrated and comparisons of the results are presented. Finally, in Section 5, we conclude the paper and present future work.

2. Methodology

In this paper, the researchers have presented a new hybrid model to solve the breast cancer classification problem. This model basically consists of the affinity propagation clustering model, adaptive modified binary firefly algorithm, and support vector machine technique. The framework of the proposed knowledge-based system comes in figure 1.

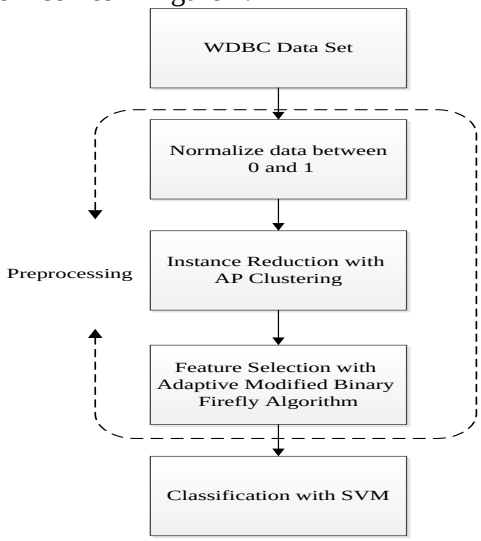

Figure 1. The proposed hybrid approach for breast cancer classification.

In this work, a study on the WDBC dataset is carried out. The researchers agree that the data mining tools perform more effectively when preprocessing is applied on the input datasets [37]. Thus in the first step, after normalization of the WDBC dataset, the AP clustering process is performed to cluster the data. AP is used to find noisy instances. For carrying out the next step, at first, all noisy instances are removed from the dataset, and then for the problem of feature selection, a new wrapper method is used. The wrapper method searching for the best feature subset is conducted using a classifier. For this reason, an AMBFA is applied to select the optimum subset of the feature that maximizes the accuracy of SVM classifier. In this case, the breast cancer classification process will be faster and more accurate if a less number of features are used.

The details of the WDBC dataset and the mentioned tools that construct the proposed approach are introduced in the following sub-sections.

2.1. Wisconsin Diagnostic Breast Cancer Dataset (WDBC)

In this work, we carried out the experiment on WDBC. It is publicly available at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). This dataset involves the measurements taken according to the FNA test [12]. This test involves fluid extraction from a breast mass using a small-gauge needle, and then a visual inspection of the fluid under a microscope [12].

The WDBC dataset includes 569 and 32 attributes. All the features represent the characteristics of cell nuclei present in the image, and are recorded with four significant digits. Table 1 depicts the dataset attribute information.

Table 1. WDBC cell nuclei characteristic attributes.

1) ID_ Number
2) Class_ Label (M=malignant, B=benign)
3-32) Ten features are computed for each Nucleus
a) radius [mean of distances from center to points on the perimeter]
b) texture [standard deviation of grey-scale values]
c) perimeter
d) area
e) smoothness [local variation in radius lengths]
f) compactness [((perimeter)2/ area) −1]
g) concavity [severity of concave portions of the contour]
h) concave points [number of concave portions of the contour]
i) symmetry
j) fractal dimension [“coastline approximation” −1]

In table 1, the first two attributes correspond to a unique identification number and the diagnosis

status (benign/malignant). The rest of the 30 features are computations for ten real-valued features along with their mean, standard error, and the mean of the three largest values ("worst" value) for each cell nucleus, respectively. For instance, field 3 is the mean radius, field 13 is the Radius SE, and field 23 is the worst radius.

2.2. Affinity Propagation (AP)

AP is a new clustering method proposed by Frey and Dueck in 2007 [34], which has been shown to produce clusters in a much less time and with a much less error than the previous techniques (such as the mixtures of Gaussians [38], K-Means algorithm [39, 40, 41], K-Medoids algorithm [42, 43], spectral clustering [43, 44], and hierarchical clustering [42, 43]). Another useful feature of the algorithm is that AP does not require the number of clusters beforehand, which is a major distinction to many other clustering algorithms. Furthermore, AP does not randomly choose some data points as cluster representatives initially, which give an advantage in such a case that the initial choices do not conclude with a good solution [34]. Due to all of these advantages, the AP algorithm has become an attractive clustering method and has been used in various domains.

AP is a data-based clustering algorithm. In other words, it uses data to learn a set of centers such that the sum of squared errors between the data points and their nearest centers is small. The centers are selected from the actual data points, and are called "exemplars." This distinctive clustering algorithm does not require the number of clusters to be pre-determined like other clustering algorithms do; instead, it considers all data points as the potential exemplars and transmits two types of messages between them until it finds the optimal ones through continuous iteration. The clusters are gradually generated during the message-passing procedure. Two types of messages are responsibility and availability. The responsibility messages are sent from the data points to their candidate exemplars. A responsibility message contains information about how well the data point serves to that candidate exemplar. The availability messages are sent from the candidate exemplars to the data points. An availability message represents how appropriate the candidate exemplar is for that data point.

This algorithm takes an input function of similarities, s(i, j), which reflects how well-suited the data point j is to be the exemplar of the data point i. AP aims to maximize the sum of similarities between the data points and their exemplars; therefore, an application requiring a minimization (e.g. Euclidean distance) should have a negative similarity function. There is a special parameter, s(i, i), which indicates how likely the relevant data point i is to be chosen as an exemplar. s(i, i) is named as the i^{th} element preference pK. The data point with a larger pK value is more likely to be chosen as an exemplar [34, 40]. Using the result of the similarity function, the algorithm updates the responsibility and the availability values of a data point, as shown in (1-3).

$$r(i,j) = s(i,j) - \max_{j' s.t\ j' \neq j} \left\{ a(i,j') + s(i,j') \right\} \quad (1)$$

$$a(i,j)_{i \neq j} = \min \left\{ 0, r(i,j) + \sum_{\forall i' \notin \{i,j\}} \max \left\{ 0, r(i',j) \right\} \right\} \quad (2)$$

$$a(i,j) = \sum_{i' s.t\ i' \neq j} \max \left\{ 0, r(i',j) \right\} \quad (3)$$

In Equations 1 to 3, the self-responsibility, r(i, i), and the self-availability, a(i, i), both reflect the accumulated evidence that i is an exemplar. Finally, the exemplar of each node i is found as (4).

$$Exemplar_i = \arg\max_j \left\{ a(i,j) + r(i,j) \right\} \quad (4)$$

In (4), the exemplar for node i is defined to be the node with the maximum collective availability and responsibility for node i. When a data point's self-responsibility plus self-availability becomes positive, that data point becomes the exemplar [34].

The R package "apcluster" is available via CRAN—The Comprehensive R Archive Network: http://cran.rproject.org/web/packages/apcluster.

2.3. Firefly Algorithm

Xin-She Yang introduced the Firefly algorithm in 2008, which is inspired by the social behavior of fireflies [45]. It is a nature-inspired meta-heuristics algorithm that can solve an NP-hard problem such as feature selection problem [46, 47]. Fireflies produce flashes to attract other fireflies. There are three rules to formulate FA by idealizing some of the flashing characteristics of fireflies [36]:

a) All fireflies are attracted to other fireflies regardless of their sex, which means that they are unisex. b) Fireflies have their own attractiveness, which is proportion to their brightness. Every two fireflies are attracted to a brighter one. A more brightness firefly means the less distance between two fireflies. The brightest firefly moves randomly. c) The fitness function appoints the brightness of each firefly. Three important strategies embedded in the firefly algorithm

include distance, attractiveness, and movement, which come as what follow.

2.3.1. Distance

The distance of two fireflies i and j at positions Xi and Xj, $r_{i,j}$, can be determined in the Euclidean distance as (5) [45]:

$$r_{i,j} = \|x_i - x_j\| \sqrt{\sum_{k=1}^{d} \left(x_{i,k} - x_{j,k}\right)^2} \quad (5)$$

where, $X_{i,k}$ and $X_{j,k}$ are the k[th] component of the i[th] and j[th] firefly, respectively, and d is the total number of dimensions.

2.3.2. Attractiveness

Measuring the attractiveness function β(r) can perform any monotonically decreasing function such as (6).

$$\beta = \beta_0 \times e^{-\gamma r^2} \quad (6)$$

where, r is the distance between two fireflies, β_0 is the attractiveness parameter, and γ is the light absorption coefficient.

2.3.3. Movement

The movement of firefly i toward firefly j as the more attractive firefly is determined by formula (7).

$$x_i(t+1) = x_i(t) + \beta(r)(x_i(t) - x_j(t)) + \alpha(rand - 0.5) \quad (7)$$

The first term is the current position of firefly i, the second one refers to attractiveness, and the third one is the randomized movement of the i[th] firefly within the search space with the randomized parameter α. Rand is a random number generator uniformly distributed in [0, 1] [35].

A Binary Firefly Algorithm (BFA) is used to solve discrete problems [48]. In this model, the position of each firefly is characterized by two values of 0 and 1 in each dimension. When firefly i moves in the direction of firefly j, the position of firefly i changes from the binary- to the real-coded. Thus for converting the binary form, at first, the position of firefly i is mapped to 0 and 1 interval using a sigmoid function as (8).

$$s(x_{i,k}(t+1)) = 1 \Big/ 1 + \exp(-x_{i,k}(t+1)) \quad k = 1,2,...,d \quad (8)$$

where, $X_{i,k}$ is the k[th] component of the i[th] firefly and d is the dimension. Then the new position of firefly i is calculated by (9).

$$x_{i,k} = \begin{cases} 1, & if\ rand\ \pi\ s(x_{i,k}) \\ 0, & otherwise \end{cases} \quad (9)$$

Rand is a random number in [0, 1].

2.4. Support Vector Machine (SVM)

SVM was applied in this work as the classifier to classify the two groups of breast cancer patients (malignant and benign). SVM, first introduced by Vapnik [49], performs classification by constructing a hyperplane that optimally separates the data points into two categories. It has been recently proposed as a very effective method for regression, classification, and general pattern recognition [50]. It is considered a good classifier due to its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high.

3. Proposed Combined Affinity Propagation-Adaptive Modified Binary Firefly Algorithm (AP-AMBFA)

The proposed AP-AMBFA has two phases:

In the first phase, the Affinity Propagation (AP) clustering method and in the second one, AMBFA are implemented.

We utilized AP clustering as instance reduction because data imperfection impairs classification accuracy and it can harm the classifier performance when a high amount of noise is present. The main idea of instance reduction by AP clustering is to eliminate clusters with only one instance. In fact, since the exemplar of these clusters isn't similar to each data point, they get in a separate cluster. After the data reduction phase, AMBFA is called to find the optimum subset of the feature that maximizes cost function (accuracy of SVM classifier).

AMBFA consists of two parts:

- modification of α step
- modification of binary step

They come in the following sub-sections.

3.1. α step

In standard BFA, the method of setting α step is static. It cannot really reflect the searching process. In general, it is useful for fireflies to explore a new search space with a large step but it is not helpful to the convergence of global optimum. If the step has a small value, the result is contrary. Therefore, step α has a great effect on the exploration and convergence of the algorithm. It would be beneficial to balance the ability of global exploration and local exploitation, and it should also be concerned with its current situation. For this reason, we designed an adaptive adjusting scheme of step α that can be controlled. In this paper, this parameter is modified according to (10) and (11), as shown as follow [35]:

$$\alpha(t+1) = (1 - \Delta) \times \alpha(t) \quad (10)$$

$$\Delta = 1 - (10^{-4}/0.9)^{1/\max iter} \quad (11)$$

where, Δ determines the step size when changing$(t+1)$. Note that this parameter decreases with increase in the generation counter t. This is given in figure 2.

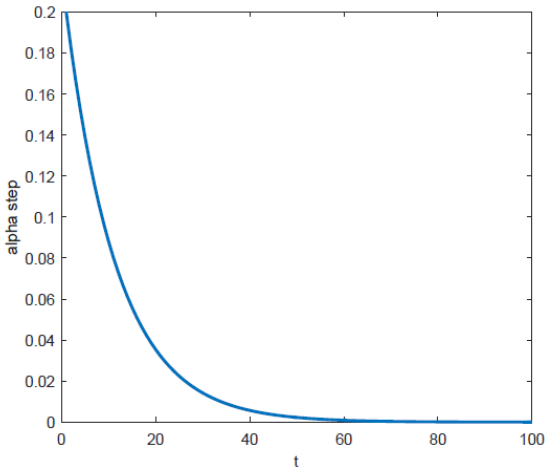

Figure 2. The status of α values in different generations.

3.2. Binary step

In order to make an improvement in BFA, another function is offered to utilize. It is called "tanh", which is described in (12) [36].

$$f(x_{i,k}) = \tanh(|x_{i,k}|) = \exp(2*|x_{i,k}|)-1 / \exp(2*|x_{i,k}|)+1 \quad (12)$$

where, $X_{i,k}$ is the kth component of the ith firefly and d is the dimension. Then the new position of firefly i is calculated by (13).

$$x_{i,k} = \begin{cases} 1, & if\ rand\ \pi\ \tanh(x_{i,k}) \\ 0, & otherwise \end{cases} \quad (13)$$

Rand is a random number in [0, 1]. Both functions (sigmoid and tanh), scale the $x_{i,k}$ value in the [0, 1] range, as shown in figure 3.

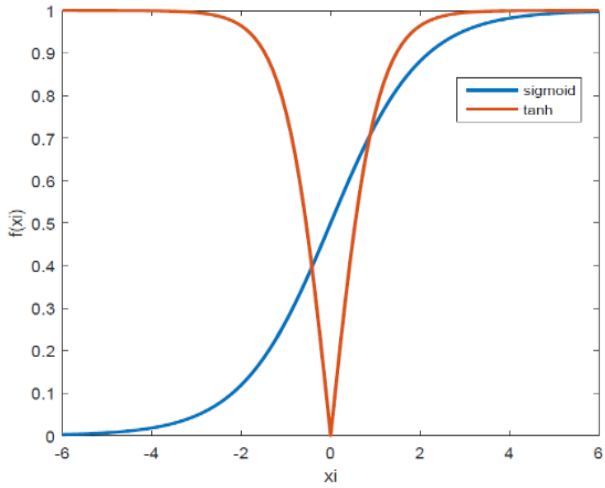

Figure 3. Mapping functions of $x_{i,k}$ values.

It has been shown that after performing the specified trails, performance of the tanh function on reaching a quality solution is fast in comparison with the sigmoid function. The pseudo-code of the proposed AP-AMBFA is illustrated in algorithm 1.

Algorithm 1. The Pseudo-code of AP-AMBFA.

%%%AP clustering
Begin
WDBC dataset as input.
Initialize availabilities $a(i,j)$ to zero $\forall\ i,j$.
while (the exemplars have not changed)
Update using equation (1), all the responsibilities given the availabilities.
Update using equation (2), all the availabilities given the responsibilities.
Combine availabilities and responsibilities to obtain the exemplar decisions.
End while
If (Number of data in clusters==1)
Remove data on WDBC dataset(Reduced WDBC dataset)
End If
End
%%%AMBFA
Begin
Reduced WDBC dataset.
setting Binary Firefly Algorithm Parameters .
define the objective function: $f(X_i)$=accuracy of SVM.
randomly generate Initial population of binary Fireflies.
while (t <Max_iteration)
for i=1:n all n binary fireflies
for j=1:n all n binary fireflies
if$(f(X_j) > f(X_i))$
move binary firefly i towards j and then move randomly by equation(7).
Else
move firefly i randomly.
End if
position of firefly i is mapped to 0 and 1 by equation(12).
calculate brightness of firefly i by objective function.
Update best solution
End for j
End for i
Update step α by equation (10, 11).
End while
maximum $f(X_i)$ is output, The Best firefly that makes maximum $f(X_i)$.
End

4. Experimental results

The proposed hybrid model was implemented in R and MATLAB software and on a computer using Intel core i7. The WDBC datasets, as discussed in Section 2.1, were used to illustrate the performance of the proposed method. To avoid feature values in greater numeric ranges from dominating those in smaller numeric ranges, the values for the features were normalized between 0 and 1 by (14).

$$x = x - x_{\min} / x_{\max} - x_{\min} \quad (14)$$

In (14), X is the value of feature, and X_{MIN} is the minimum and X_{MAX} is the maximum value for each feature.

As mentioned earlier, firstly, AP clustering was carried out to determine the single member clusters and remove them to construct the reduced

dataset. Then this new reduced dataset was entered as the input of AMBFA to find the optimum subset of the feature that maximizes the accuracy of SVM classifier. In this algorithm, each firefly represents one subset of features to the breast cancer classification problem. Representation of each firefly is illustrated in figure 4. In this kind of representation, each element of array stands for a feature whether a feature is selected (1) or not (0).

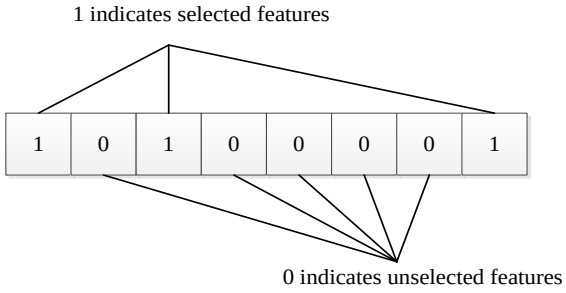

Figure 4. Representation of a firefly.

AMBFA searches in the space of this new dataset and sets its parameters to find the optimal property subset. It should be noted that the appropriateness of parameters is an important issue. Thus parameter settings of AMBFA are conducted based on the nature and complexity of the problem domain. The parameters for AMBFA were set as given in table 2.

Finally, the algorithm will stop if it reaches a pre-determined maximum iteration with the maximum classification accuracy. For calculation of the accuracy, the 10-fold cross-validation method was utilized.

Table 2. Parameter setting of firefly algorithm.

Parameters	Value
Population size	40
light absorption coefficient(γ)	1
Attractiveness(β_0)	2
mutation rate (α)	0.2
Maximum iteration	100

In the following sub-sections, we provide the performance results of the techniques incorporated in the proposed knowledge-based system.

4.1. Results of AP clustering

By applying AP clustering on the normalized WDBC dataset in R software, 43 clusters emerged. The information for the 43 clusters comes in table 3. The numbers in boldface denote the clusters with only one instance that must be eliminated. In other words, these clusters contain a noisy instance.

Pay attention to table 3; clusters number 1, 7, 10, 12, 15, 16, 21, 22 and, 36 have only one instance that correspond to the instance numbers 4, 43, 69, 79, 123, 153, 213, 214, and 462, respectively. In fact, these 9 instances were detected as noisy instances by AP, and were removed to create the reduced WDBC dataset.

Table 3. Results of AP clustering on WDBC.

No. of clusters	No. of exemplars	No. of instances in cluster
1	**4**	**1**
2	13	2
3	15	8
4	23	5
5	24	9
6	29	10
7	**43**	**1**
8	44	16
9	65	10
10	**69**	**1**
11	75	40
12	**79**	**1**
13	105	22
14	118	12
15	**123**	**1**
16	**153**	**1**
17	168	19
18	177	4
19	178	9
20	205	38
21	**213**	**1**
22	**214**	**1**
23	272	30
24	302	15
25	318	22
26	321	20
27	341	18
28	362	28
29	394	10
30	406	14
31	417	7
32	430	33
33	434	27
34	435	34
35	453	16
36	**462**	**1**
37	474	4
38	486	7
39	488	18
40	505	2
41	515	19
42	522	9
43	549	23

4.2. Evaluation of proposed hybrid model

In this section, the performance evaluation of the proposed hybrid model on the WDBC dataset is presented. Due to the random nature of heuristic algorithms, the average rate of accuracy in 10 separate runs over 100 iterations is reported. Table 4 shows comparisons between the results of our hybrid model and previous models.

According to the results tabulated in this table, our hybrid model has the best classification accuracy compared with the other models reported in the literature in diagnosis of breast cancer on the WDBC dataset.

Table 4. Comparison results between our hybrid model and previous models on WDBC dataset.

Ref.	Year	Feature selection method	Classifier	Accuracy
[27]	2008	Manifold	SVM	97.3%
[28]	2011	KP	SVM	97.55%
[28]	2011	RFE	SVM	95.25%
[28]	2011	FSV	SVM	95.23%
[28]	2011	Fisher	SVM	94.70%
[24]	2014	MCFS	SVM	96.68%
[24]	2014	FSFS	SVM	94.41%
[24]	2014	LSFS	SVM	96.87%
[29]	2014	DEA & Entropy	SVM	89.86%
			C.5	93.92%
			LR	95.95%
[29]	2014	CFS	SVM	87.84%
			C.5	92.75%
			LR	95.95%
[29]	2014	Filtered	SVM	87.84%
			C.5	91.22%
			LR	96.62%
[31]	2016	PSO-KDE	SVM	98.45%
		GA-KDE	SVM	98.45%
[32]	2017	Ensemble-FSGA	SVM	82.2%
[33]	2018	WAUCE	SVM	97.68%
Our hybrid model	**2018**	**AP-AMBFA**	**SVM**	**98.606%**

Additionally, the performance of the proposed method with/without both the AP and AMBFA methods was investigated based on the accuracy, precision, and recall measures. The results obtained are shown in table 5.

Table 5. Comparison of the results of three measures for four different models.

Model	Accuracy	Precision	Recall
AP- AMBFA	98.606	95.11	97.32
AMBFA	98.21	95.03	97.2
AP- BFA	98.54	94.24	97.15
BFA	98.17	94.55	96.93

Table 5 shows the classification accuracy, precision, and recall rates in 10-fold cross-validation schemes for 100 (random) repetitions. The classification accuracy rates for BFA, AP-BFA, and AMBFA were measured to be 98.17%, 98.54%, and 98.21%, respectively. The proposed hybrid model surpassed all of them by a slight difference, achieving a rate of 98.606%. Also the AP-AMBFA model surpassed the BFA, AP-BFA, and AMBFA models in terms of precision and recall rate.

We also provided the precision-recall (PR) curves as a useful tool to represent the superiority of the AP-AMBFA model on the other models. The corresponding PR curves for the four models AP-AMBFA, BFA, AP-BFA and AMBFA are shown in figure 5. For each model, we reported the PR curves based on ten independent replications of each algorithm.

It can be seen again in figure 5 that the presented AP-AMBFA model outperforms the BFA, AP-BFA, and AMBFA models.

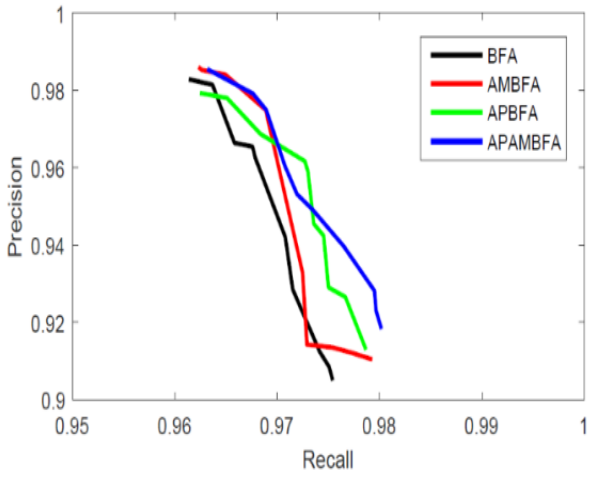

Figure 5. PR curves for four different models.

5. Conclusion and future work

The main goal of this article is to introduce an efficient prediction model to aid physicians for diagnosis of breast cancer. Thus in this work, a new hybrid model of AP clustering method and AMBFA was presented and successfully applied to the classification of breast cancer on the WDBC dataset. According to the experimental results, the proposed hybrid model can improve the accuracy to 98.606%. These results are very promising compared to the previously reported classification techniques and the three models BFA, AP-BFA, and AMBFA for mining breast cancer data.

Furthermore, the advantage of using the AP clustering method is to eliminate the noisy instance that can prevent decrease in the accuracy rate. Besides, AMBFA can improve the constraints of binary firefly algorithm based on making a balance between the abilities of global and local searches and also reaching the most

qualified and fast solutions. Due to the modifications on the original BFA and combined with the AP clustering, the accuracy, precision, and recall measures were improved. The high classification accuracy from our proposed algorithm can be used as the reference for decision-making in a hospital and the researchers. In the future, the main aim is to propose a multi-objective method for a feature selection problem, and also it is recommended to combine feature selection with feature construction using other ML algorithms.

References

[1] Lavrac, N. (1999). Selected techniques for data mining in medicine. Artificial Intelligence in Medicine, vol. 16, no. 1, pp. 3-23.

[2] Richards, G. et al. (2001). Data mining for indicators of early mortality in a database of clinical records. Artificial Intelligence in Medicine, vol. 22, no. 3, pp. 215-231.

[3] Marcano-Cedeño, A., Quintanilla-Domínguez, J. & Andina, D. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications, vol. 38, no. 8, pp. 9573–9579.

[4] McCarthy, A.M., Yang, J. & Armstrong, K. (2015). Increasing disparities in breast cancer mortality from 1979 to 2010 for US black women aged 20 to 49 years. American Journal of Public Health, vol. 105, no. 3, pp. 446–448.

[5] Kharazmi, E. et al. (2016). Survival in familial and non-familial breast cancer by age and stage at diagnosis. European Journal of Cancer, vol. 52, pp. 10–18.

[6] Jerez-Aragone´s, J. M. et al. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence in Medicine, vol. 27, no. 1, pp. 45-63.

[7] Sizilio, G. R. et al. (2012). Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis. Biomedical engineering online, vol. 11, no. 1, pp. 83-83.

[8] Paulin, F. & Santhakumaran, A. (2011). Classification of breast cancer by comparing back propagation training algorithms. International Journal on Computer Science and Engineering, vol. 3, no. 1, pp. 327-332.

[9] Sahan, S. et al. (2007). A new hybrid method based on fuzzy artificial immune system and k-nn algorithm for breast cancer diagnosis, Computers in Biology and Medicine, vol. 37, no. 3, pp. 415-423.

[10] Fletcher, S. W. et al. (1993). Report of the international workshop on screening for breast cancer. Journal of the National Cancer Institute, vol. 85, no. 20, pp. 1644-1656.

[11] Giard, R. W. M. & Hermans, J. (1992). The value of aspiration cytologic examination of the breast a statistical review of the medical literature. Cancer, vol. 69, no. 8, pp. 2104-2110.

[12] Mangasarian, O., Nick Street, W. & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, vol. 43, no. 4, pp. 570-577.

[13] Chen, H. L., Yang, B., Liu, J., Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications, vol. 38, no. 7, pp. 9014–9022.

[14] Sousa, T., Silva, A. & Neves, A. (2004). Particle swarm based data mining algorithms for classification tasks. Parallel Computing, vol. 30, no. 5-6, pp. 767–783.

[15] Subashini, T.S., Ramalingam, V. & Palanivel, S. (2009). Breast mass classification based on cytological patterns using RBFNN and SVM. Expert Systems with Applications, vol. 36, no. 3, pp. 5284–5290.

[16] Muhlenbach, F., Ephane, S. T. & Zighed, D. A. (2004). Identifying and Handling Mislabelled Instances. Journal of Intelligent Information Systems, vol. 22, no. 1, pp. 89–109.

[17] Yin, H. & Dong, H. (2011). The problem of noise in classification: past, current and future work. Proceedings of the Communication Software and Networks (ICCSN), Xi'an, China, 2011.

[18] Hamidzadeh, J. (2015). IRDDS: Instance reduction based on Distance-based decision surface. Journal of AI and Data Mining, vol. 3, no. 2, pp. 121-130.

[19] SegataN. & Blanzieri, E. (2010). Noise reduction for instance-based learning with a local maximal margin approach. Journal of Intelligent Information Systems, vol. 35, no. 2, pp. 301-331.

[20] Brodley, C. E. & Friedl, M. A. (1999). Identifying mislabeled training data. Journal of Artificial Intelligence Research, vol. 11, no. 1, pp. 131–167.

[21] Blum, A. L. & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, vol. 97, no. 1, pp. 245–271.

[22] De Stefano, C., et al. (2014). A GA-based feature selection approach with an application to handwritten character recognition, Pattern Recognition Letters, vol. 35, pp. 130-141.

[23] Mitra, P., Murthy, C. A. & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301–312.

[24] Bandyopadhyay S. et al. (2014). Integration of dense subgraph finding with feature clustering for

unsupervised feature selection. Pattern Recognition Letters, vol. 40, pp. 104–112.

[25] He, X. Cai, D. & Niyogi, P. (2005). Laplacian score for feature selection. Proceedings of the 18th International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2005.

[26] Cai, D, Zhang, C. & He, X. (2010). Unsupervised feature selection for multi-cluster data. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, USA, 2010.

[27] Zhaohui, L. et al. (2008). Diagnosis of breast cancer tumor based on manifold learning and support vector machine. IEEE International Conference on Information and Automation, Changsha, China, 2008.

[28] Maldonado, S., Weber, R. & Basak J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. Information Sciences, vol. 181, no. 1, pp. 115-128.

[29] Bamakan, S. M. H. & Gholami, P. (2014). A Novel Feature Selection Method based on an Integrated Data Envelopment Analysis and Entropy Model. Procedia Computer Science, vol. 31, pp. 632-638.

[30] Liu, H. & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4. pp. 491- 502.

[31] Sheikhpour, R., Agha Sarram, M. & Sheikhpour, R. (2016). Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. Applied Soft Computing, vol. 40, pp. 113-131.

[32] Das, A. K., Das, S. & Ghosh, A. (2017). Ensemble feature selection using bi-objective genetic algorithm. knowledge-Based Systems, vol. 123, pp. 116-127.

[33] Wang, H., Zheng, B., Yoon, S. W. & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research, vol. 267, no. 2, pp. 687- 699.

[34] Frey, B.J. & Dueck, D. (2007). Clustering by passing messages between data points. Science, vol. 315 pp. 972–976.

[35] Fister, I. et al. (2013). Memetic Self-Adaptive Firefly Algorithm. Swarm Intelligence and Bio-Inspired Computation, pp. 73-102.

[36] Chandrasekaran, K. & P. Simon, Sishaj (2012). Network and reliability constrained unit commitment problem using binary real coded firefly algorithm. Electrical Power and Energy Systems, vol, 43, no. 1, pp. 921–932.

[37] Liu, H. & Motoda, H. (2012). Feature selection for knowledge discovery and data mining. Springer Science & Business Media, Boston.

[38] Ding, S., Ma, G. & Shi, Z. (2013). A novel self-adaptive extreme learning machine based on affinity propagation for radial basis function neural network. Neural Computing and Applications, vol. 24, no. 7-8, pp. 1487–1495.

[39] Chehdi, K. & Soltani, M. (2008). Pixel classification of large-size hyperspectral images by affinity propagation. Genetics and Molecular Biology, vol. 31, no. 1, pp. 64–67.

[40] Guan, R., et al. (2011). Text clustering with seeds affinity propagation. IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 4, pp. 627–637.

[41] Yang, C., et al. (2013). Incremental and decremental affinity propagation for semisupervised clustering in multispectral images. IEEE Transactions on Geoscience and Remote Sensing, vol. 51, no. 3, pp. 1666–1679.

[42] Dueck, D. (2009). Affinity Propagation: Clustering Data by Passing Messages. University of Toronto, Toronto.

[43] Frey, B. J. & Dueck, D. (2005). Mixture modeling by affinity propagation. 18th International Conference on Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada, 2005.

[44] Shang, F. H., et al. (2012). Fast affinity propagation clustering: a multilevel approach. Pattern Recognition, vol. 45, no. 1, pp. 474–486.

[45] Yang, X. S. (2008). Nature-inspired Metaheuristic Algorithm. University of Cambridge. United Kingdom: Luniver Press.

[46] Yang, X. S. (2010). Engineering optimization: an introduction with metaheuristic applications. Wiley.

[47] Liu, H. & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491–502.

[48] Palit, S. et al. (2011). A Cryptanalytic attack on the Knapsack Cryptosystem using Binary Firefly Algorithm, 2thInternational Conference on Computer and Communication Technology (ICCCT), Allahabad, India, 2011.

[49] Vapnik, V. N. (1998). Statistical learning theory. New York: Wiley.

[50] Cristianini, N. & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press New York.

نشریه هوش مصنوعی و داده‌کاوی

ارائه یک سیستم جدید مبتنی بر دانش برای تشخیص سرطان پستان با ترکیبی از الگوریتم‌های خوشه-بندی انتشار وابستگی و کرم شب‌تاب

نسیبه امامی[1*] و آیلین پاکزاد[2]

[1] گروه علوم کامپیوتر، دانشگاه کوثر بجنورد، بجنورد، ایران.

[2] گروه مهندسی صنایع، دانشگاه کوثر بجنورد، بجنورد، ایران.

ارسال ۲۰۱۷/۱۲/۰۴؛ بازنگری ۲۰۱۸/۰۴/۱۸؛ پذیرش ۲۰۱۸/۰۶/۱۲

چکیده:

سرطان پستان در زنان به رایج‌ترین نوع سرطان در سراسر جهان تبدیل شده‌است. سیستم‌های خبره مبتنی بر تکنیک‌های داده‌کاوی ابزاری ارزشمند در تشخیص سرطان پستان می‌باشند و می‌توانند برای پزشکان جهت کمک در تصمیم‌گیری موثر باشند. این مقاله رویکرد جدیدی شامل ترکیب تکنیک-های داده‌کاوی برای تشخیص خوش‌خیم و بدخیم بودن تومورهای پستان ارائه می‌کند. مدل ارائه شده، AP-AMBFA، شامل دو فاز می‌باشد. در فاز اول، روش خوشه‌بندی انتشار وابستگی به عنوان تکنیک کاهش رکورد استفاده شد؛ که می‌تواند رکوردهای پرت را شناسایی نموده و حذف کند. در فاز دوم، الگوریتم‌های انتخاب ویژگی و دسته‌بندی اجرا شد. الگوریتم کرم شب‌تاب باینری تطبیقی تغییر یافته جهت انتخاب متغیرهای موثر در پیش‌بینی متغیر هدف و تکنیک ماشین بردار پشتیبان به عنوان کلاسه‌بند استفاده گردید. مدل پیشنهادی می‌تواند پیچیدگی محاسباتی را کاهش داده و سرعت پردازش اطلاعات را افزایش دهد. نتایج تجربی در مجموعه داده‌های سرطان پستان WDBC، پیش‌بینی دقیق‌تری را ارائه می‌دهد. نرخ صحت کلاس-بندی ۹۸/۶۰۶٪ به دست آمده است؛ که در مقایسه با سایر روش‌های کلاس‌بندی اجرا شده بر روی مجموعه داده WDBC، بالاتر می‌باشد. از این رو این روش پزشکان را در تشخیص دقیق‌تر سرطان پستان کمک خواهد کرد.

کلمات کلیدی: سرطان پستان، خوشه‌بندی انتشار وابستگی، انتخاب ویژگی، الگوریتم کرم‌شب‌تاب باینری، ماشین بردار پشتیبان.