Journal of AI and Data Mining
Vol 6, No 2, 2018, 263-276
DOI: 10.22044/JADM.2017.5061.1609

Extracting Predictor Variables to Construct Breast Cancer Survivability Model with Class Imbalance Problem

S. Miri Rostami* and M. Ahmadzadeh

Faculty of computer and IT Engineering, Shiraz University of Technology, Shiraz, Iran.

Received 19 November 2016; Revised 14 June 2017; Accepted 02 September 2017
**Corresponding author: S.Miri@sutech.ac.ir (S. Miri Rostami).*

Abstract
Application of data mining methods as a decision support system has a great benefit to predict survival of new patients. It also has a great potential for health researchers to investigate the relationship between risk factors and cancer survival. However, due to the imbalanced nature of the datasets associated with breast cancer survival, the accuracy of survival prognosis models is a challenging issue for researchers. This work aimed to develop a predictive model for 5-year survivability of breast cancer patients and discover the relationships between certain predictive variables and survival. The dataset was obtained from the SEER database. First, the effectiveness of two synthetic over-sampling methods Borderline-Synthetic Minority Over-sampling Technique (Borderline-SMOTE) and Density-based Synthetic Oversampling (DSO) method is investigated to solve the class imbalance problem. Then a combination of Particle Swarm Optimization (PSO) and Correlation-based Feature Selection (CFS) is used to identify the most important predictive variables. Finally, in order to build a predictive model, the three classifiers decision tree (C4.5), Bayesian Network (BN), and Logistic Regression (LR) are applied to the datasets. Some assessment metrics such as accuracy, sensitivity, specificity, and G-mean are used to evaluate the performance of the proposed hybrid approach. Also the area under ROC curve (AUC) is used to evaluate the performance of the feature selection method. The results obtained show that among all combinations, DSO + PSO_CFS + C4.5 presents the best efficiency in terms of accuracy, sensitivity, G-mean, and AUC with the values of 94.33%, 0.930, 0.939, and 0.939, respectively.

Keywords: *Breast Cancer, Survival, Class Imbalance Problem, Over-Sampling Technique, Feature Selection.*

1. Introduction

Breast cancer is the most common cancer in women worldwide. In 2011, about 230,480 US women were diagnosed with breast cancer, and approximately 39,520 of them died due to this disease [1]. According to Iran's National Cancer Registry center, breast cancer incidence has increased dramatically from 2001; based on its report in 2011, 23% of all cancers diagnosed in women were breast cancer cases [2]. Knowledge about cancer prognosis helps physicians to estimate outcome of disease and determine chance of survival. Survival analysis is a part of medical prognosis, which predicts a patient survival in a specific time period [3]. Prognostic factors are important for detecting patients who may, or may not, benefit from treatment [4]. Data mining methods, as a decision support system, have a great benefit to predict survival of new patients.

Recently, advances in diagnosis and treatment of breast cancer have caused decrease in the death rate, i.e. the number of patients who survive is more than the number of patients who die. Thus this issue leads to imbalanced datasets that are a special case for classification problems [5]. In such datasets, the class distribution is not uniform among the classes. Typically, they are composed of two classes: the majority (negative) class and the minority (positive) class. In supervised learning, particularly in classification, the goal is to accurately predict the target class from the attribute values for each sample in the dataset, where the dataset consists of many samples in the

form of (attribute values, class Label). Classification of imbalanced dataset is a challenging issue for researchers. Thus in order to have a more accurate prognostic model for breast cancer survival, the class imbalance problem should be solved. Most of the standard data mining techniques assume the balanced dataset, and when they work with the imbalanced one, the results obtained are biased toward numerous majority class samples. Therefore, in spite of the high accuracy, the detection rate of the minority samples is very low [6]. In some fields such as medical, detecting minority samples are more important than overall accuracy. A large number of predictor variables is another issue in the field of constructing breast cancer survival models. Predictor variables are a set of attributes which their values are used to predict class label. Additionally, the presence of a correlation between the predictor variables can decrease the performance of models, so applying feature selection methods to the dataset by considering correlation between variables can improve the performance of models and provide a faster and more cost-effective prognosis model [7].

This work proposes a hybrid approach to predict a 5-year survivability of breast cancer patients. First, the two different methods Borderline-SMOTE (Borderline-Synthetic Minority Over-sampling Technique) and DSO (Density-based Synthetic Oversampling) are used to solve the class imbalance problem. Then a combination of PSO (Particle Swarm Optimization) and CFS (Coloration based Feature Selection) is used to extract the most important and relevant predictor variables. Finally, by applying the three classifiers decision tree (C4.5) [8], Bayesian Network (BN), and Logistic Regression (LR) to the datasets, the predictive model is built.

The rest of the paper is organized as what follows. In Section 2, the related works are reviewed. The methods used in the proposed approach are presented in Section 3. Section 4 describes the proposed approach. Section 5 provides the experiments, and results obtained are analyzed in this section. Finally, Section 6 concludes the paper.

2. Related works

Sampling methods, cost-sensitive learning methods, and techniques in algorithm level are solutions to handle the class imbalance problem [9]. Initial studies in the field of breast cancer survival prediction did not consider the class imbalance problem [4,10,11]. However, recently, researchers have investigated the effectiveness of some techniques to handle this problem. In [12], Liu, Wang, and Zhang have used the under-sampling method to deal with the class imbalance problem. Under-sampling refers to the process of decreasing the number of records in the majority class. Afshar, Ahmadi, Roudbari, and Sadoughi in [13] have used the data mining techniques to build a predictive model for breast cancer. In this work, the class imbalance problem was solved using an over-sampling method which refers to the process of increasing the number of records in the minority class. However, random under-sampling might eliminate valuable information. On the other hand, random over-sampling by generating too many repeated samples for minority class increases the likelihood of over-fitting. When over-fitting occurs, a model loses its generalization power, which leads to poor performance on new data [14]. Wang, Makond, Chen, and Wang [15] have used the SMOTE method to balance a dataset. After data balancing, a combination of PSO and classifiers were used to extract the most relevant features for classification. A survey on predicting breast cancer survivability and its challenges can be found in [16].

SMOTE method [17] is a well-known over-sampling method, which has been employed in several data preprocessing studies [18,19]. It gives equal weight to each minority sample to generate new synthetic samples. It does not matter if the data sample is noise data or centroid data. A centroid is a data point (imaginary or real) at the center of a cluster; in fact, it is a vector containing a value for each variable, where each value is the mean of a variable for the observations in that cluster. Therefore, SMOTE is sensitive to the noise data. Wang and Liang [20] have proposed DSO to solve drawbacks of SMOTE. This method assigns different weight to each minority samples. Based upon the idea behind DSO, the samples that are positioned on the boundary of positive and negative classes are not as useful as samples located around centroid data. On the other hand, the authors in [21] believe that increasing border samples can be useful for predicting the minority samples because these boundary samples have more potential for mis-classification.

Ren et al. [22] have proposed a new ensemble-based adaptive over-sampling method for imbalanced data learning to improve microaneurysm detection. Their approach is integration of adaptive SMOTE and ensembles. They utilized an adaptive scheme to assign weights to the minority class instead of uniform sampling weight. It can automatically decide the

number of synthetic samples that need to be generated for each minority instance. They used the three ensemble learning methods Boosting, Bagging, and Random sunspace to build a model. Lim et al. [23] have introduced a novel evolutionary cluster-based over-sampling ensemble framework named ECO-Ensemble. In order to deal with the class imbalance problem, i.e. to generate new synthetic samples, they tried to find oversampling regions using clusters. Thus the three methods mini-batch k-means [24], hierarchical agglomerative clustering, and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) are applied for clustering. Then an evolutionary algorithm (genetic algorithm [25]) is used to optimize the parameters of cluster-based synthetic oversampling (CSO). Their proposed method was evaluated on a set of 40 imbalance datasets, and the results obtained showed that it could tackle the class imbalance problems very well. Douzas and Bacao [26] have proposed a new method named Self-Organizing Map Over-sampling (SOMO) for imbalanced dataset learning. The self-organizing map algorithm allows an effective generation of artificial data points by producing a 2D representation of the input space. In this method, SOM is used to map closed data points in the input space into nearby map units to create a neighborhood structure that connects the map units to the adjacent ones. In fact, this topology helps for a better data clustering by identifying adjacent clusters. Then based on the instances of minority class belonging to the same cluster and adjacent clusters in the 2D space, the intra-cluster and inter-cluster synthetic samples are produced, respectively. For the evaluation of the proposed method, two algorithms Logistic Regression (LR) and Gradient Boosting Machine (GBM) were used. The results of this work indicate that using artificial data generated by SOMO improves the performance of these algorithms.

Moreover, the previous studies in the field of breast cancer survival prediction used similar predictor variables to build prognostic model [4,10,13]. However, [15] has proposed a feature selection method based on PSO that used classifiers accuracy rate as its fitness function. But feature selection based on learning algorithms has a large computational cost in terms of time and resource [27-29]. Furthermore, the correlation between predictor variables must be analyzed because it decreases the performance of models. This work aimed to consider these issues and used hybrid PSO_CFS method for feature selection, PSO for feature subset generation and CFS as a fitness function.

3. Preliminaries

3.1. Borderline-SMOTE

Han, Wang, and Mao have proposed a new over-sampling method named Borderline-SMOTE. They suggested that samples far from the borderline of minority class may contribute little to classification. Thus this method only over-samples or strengthens the borderline minority samples. It operates as follows [21]:

First, the borderline minority samples are found out; then the synthetic samples are generated from them and added to the original training set. Suppose that the original training set is T, the minority class is P, and the majority class is N,

$$P = \{p_1, p_2, ..., p_{num}\},\ N = \{n_1, n_2, ..., n_{num}\}$$

where, *pnum* and *nnum* are the number of minority and majority samples, respectively.

Step 1. For each p_i, $m = 5$ nearest neighbors from T are calculated. The number of majority samples existing among the m nearest neighbors is denoted by m'. Based on the relationship between m and m', three subsets are created as NOISE, DANGER, and SAFE.

Step 2. If $m = m'$, i.e. all the m nearest neighbors of p_i are majority samples, this sample is considered as NOISE. If $m/2 \le m' \le m$, p_i is a border sample of P and easily mis-classified, so it stays in DANGER. Finally, if $0 \le m' \le m/2$, it means that p_i is SAFE, and there is no need to generate new samples for this set.

Step 3. Samples in the DANGER set are the borderline data of minority class P,

$$DANGER = \{p'_1, p'_2, ..., p'_{num}\},$$
$$0 \le dnum \le pnum$$

where, dnum indicates the number of DANGER members. For each p'_i, its k nearest neighbors from P are calculated.

Step 4. In this step, $s \times dnum$ synthetic positive samples for members of DANGER set are generated, where s is an integer between 1 and k. For each p'_i, s nearest neighbors from its k nearest neighbors in P are selected randomly. The new synthetic samples are created as follow:

$$Synthetic_j = p'_i + r_j \times diff_{j'}, \quad (1)$$
$$i = 1, 2, ..., dnum;\ j = 1, 2, ..., s$$

where, r_j is a random number between 0 and 1 and $diff_j$ is a distance between p_i and its s

nearest neighbors from P. In this work, the parameters k and s are both set at 5 (based on the SMOTE method).

3.2. Density-based synthetic over-sampling

The DSO method is a kind of minority over-sampling technique that operates based on the density distribution of samples. This method assigns different weights to each minority sample. In order to generate new synthetic samples, it is needed to calculate the two parameters w_i and g_i, where the former is the sampling weight of instance i and the latter is the number of samples that must be generated from instance i. The DSO method works as follows [20]:

First, for each sample in the minority class P, the average distance (Dist) of it to all the other samples in minority class is calculated. Then the two parameters N_P and N_N should be calculated, where N_P is the number of positive neighbors (minority samples) in the specified area Dist and N_N is the number of negative neighbors (majority samples) in the same area Dist. w_i and g_i are calculated according to (2) and (3), respectively:

$$w_i = \left(\frac{N_P}{N_p + N_N}\right)^P \tag{2}$$

where, P is a parameter representing how much punishment will be put on the impure of instance i.

$$g_i = \left(\frac{w_i}{\sum_{i=1}^{m_p} w_i}\right) \times N_{syn} \tag{3}$$

where, $N_{syn} = k_s \times m_p$ is the number of synthetic minority samples required to be generated and k_s is the parameter used to control the number of the generated samples. Finally, for each sample i, in the minority class $(1 \le i \le m_p)$ k new synthetic samples are generated according to their g_i value. New synthetic samples X_k from sample X_i are generated as follow:

$$X_k = X_i + \delta(X_{point} - X_i) \tag{4}$$

where, X_{point} is one of the nearest neighbors of X_i, which is selected randomly, and δ is a random number between 0 and 1.

3.3. Particle swarm optimization (PSO)

PSO is a meta-heuristic algorithm that has been modeled based on the movement of organisms in a bird flock or fish school. In this algorithm, each candidate solution is considered as a particle that has a position and a fitness value. PSO finds the best solution based on a population in an iterative manner. Each population consists of many particles. Candidate solutions move in search space to find optimum position, so they have a velocity. In each iteration, each particle records its best position. The movement of particles depends on three factors: particle's current position (X_i), best position of the particle in the current iteration (P_{best}), and the best position among all particles in the neighborhood so far (G_{best}). The iterations are repeated until either the fitness value of the given particle equals the defined value or the number of iterations is achieved by the default value [30].

The velocity and position of the ith particle for a problem that has a search space with D dimensions are updated according (5) and (6), respectively, as follow:

$$V_{i,d}^{new} = \omega \times V_{i,d}^{old} + c_1 r_1 (P_{best\ i,d}^{old} - X_{i,d}^{old}) .. \\ + c_2 r_2 (G_{best\ d}^{old} - X_{i,d}^{old}),\ d = 1,2,...,D. \tag{5}$$

$$X_{i,d}^{new} = X_{i,d}^{old} + V_{i,d}^{new}),\ d = 1,2,...,D;\ i = 1,2,...,N. \tag{6}$$

where, $V_i = [v_{11}, v_{12}, ..., v_{1n}]$ is the velocity vector of the ith particle, ω is inertia weight between 0.4 and 0.9, c_1 and c_2 are positive constant values between 0 and 4, indicating the cognitive learning factor and the social learning factor, respectively, r_1 and r_2 are random numbers between 0 and 1, and N is the size of the swarms.

In order to apply PSO to a feature selection problem, the method proposed by Chen et al [31] is used. This method uses a binary digit to represent a feature; "0" indicates a feature that is non-selected and "1" indicates a feature that is selected. For this case, the position of dimension d of the ith particle was updated as follows:

$$X_{i,d}^{new} = \begin{cases} 1, & if\ \ sigmoid(V_{i,d}^{new}) > rand() \\ 0, & otherwise \end{cases} \tag{7}$$

where, $sigmoid(V_{i,d}^{new})$ is $\frac{1}{1+e^{-V_{i,d}^{new}}}$ and $rand()$ is a random number drawn from $U(0,1)$.

The local best position of a particle (P_{best}) and global best position (G_{best}) in each iteration are updated according to (8) and (9), respectively.

$$P_{best\ i,d}^{new} = \begin{cases} X_{i,d}^{new}, & if: f(X_{i,d}^{new}) \le f(P_{best\ i,d}^{old}) \\ else, & P_{best\ i,d}^{old} \end{cases} \tag{8}$$

$$G_{best}^{new} = \{P_{best\,d}^{new} \mid f(P_{best\,i}^{new}) \le f(G_{best}^{old})\} \quad (9)$$

In this work, the fitness value (f: fitness function) of a particle is calculated based on the CFS method.

3.4. Coloration-based feature selection

CFS is a feature selection method that uses a correlation-based heuristic to evaluate the merit of features [32]. Based on CFS, the high merit is given to a subset whose features are highly correlated to the class attribute but have a low correlation to each other. CFS calculates the heuristic merit of a feature subset S that consists of k features, as follow:

$$Merit_s(k) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (10)$$

where, $\overline{r}_{cf}$ is the average of the correlations between the subset features and the class variable and $\overline{r}_{ff}$ is the average inter-correlation between the subset features. Symmetrical Uncertainty is used to calculate the correlation between features according to (11).

$$SU(X,Y) = 2 \times \frac{IG(X \mid Y)}{H(X) + H(Y)} \quad (11)$$

where, $IG(X \mid Y)$ is information gain that is calculated based on Equation (12), $H(X)$ is the entropy of a variable X and $H(Y)$ is the entropy of a class variable.

$$IG(X \mid Y) = H(X) - H(X \mid Y) \quad (12)$$

where, $H(X \mid Y)$ is the entropy of a variable X conditioned on a variable Y.

4. Proposed approach

This work aimed to develop a predictive model for 5-year survivability of breast cancer patients and discovers the relationships between certain predictive variables and survival. Three challenges of the previous studies are the class imbalance problem, the large number of variables in original dataset and the lack of attention to the correlation between predictor variables. The proposed method is a hybrid approach that tries to overcome these limitations. Figure 1 shows the flowchart of the proposed method. As shown in this figure, the main contribution of this work is related to two parts: first dealing with the class imbalance problem and then using a combinational feature selection method (PSO_CFS) to extract the important predictor variables; in fact, this combination of methods is new hybrid approach for building predictive model for breast cancer survival. In the pre-processing step, the original dataset was cleaned. After studying the data dictionary of dataset, the unrelated variables to breast cancer were removed and the records containing the missing values were removed from the dataset as well. Additionally, data transformation including feature normalization was provided to scale attribute values. Then the effectiveness of the two over-sampling methods Borderline-SMOTE and DSO was investigated. As mentioned in Section 2, these two methods generate new synthetic samples based on minority samples in two different regions; therefore, they were selected to deal with the class imbalance problem. After applying them to the original dataset, two new datasets, dataset_1 and dataset_2, were generated for Borderline-SMOTE and DSO, respectively.

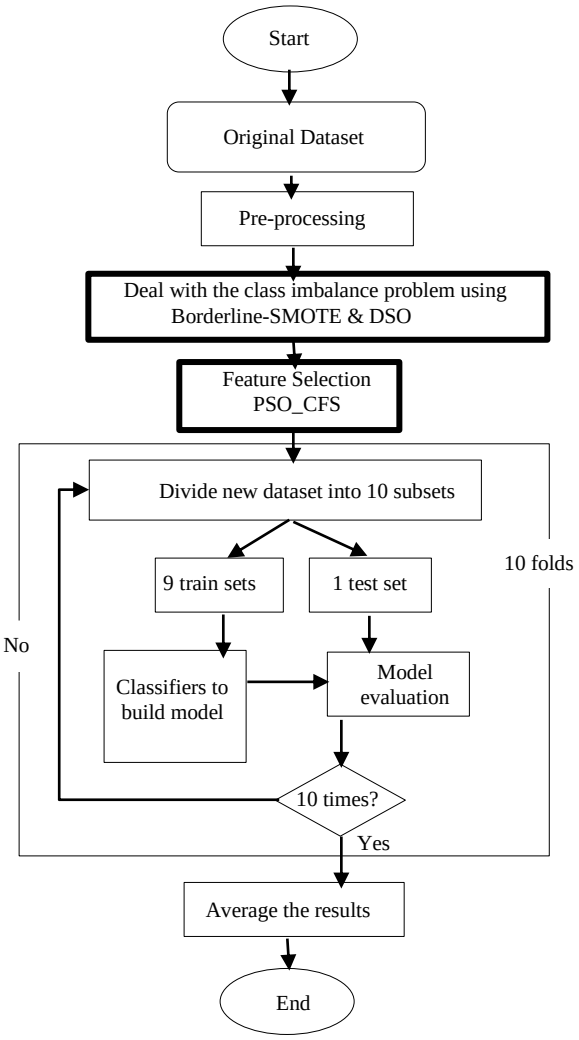

Figure 1. Flowchart of proposed approach.

Then a combination of the PSO method and the CFS algorithm is used as a feature selection approach. Since the correlation between predictor variables decreases the performance of a model, this measure is used as a fitness function in the feature selection method, and PSO is used as a feature subset generation. In this work, 10-fold cross-validation is used in order to evaluate predictive models to reduce the bias and variance of classification results. In this method, a full dataset is divided into 10 independent folds (subsets); each fold is approximately one-tenth of the full dataset (with approximately one-tenth of survival and one-tenth of non-survival). Nine out of the ten subsets are combined and used as the training set, and the remaining subset is used as the testing set. Each of the 10 subsets is used once as a testing set to evaluate the performance of a classifier, which is built from the combination of the other remaining subsets. The three classifiers decision tree (C4.5), BN, and LR are applied to the training set to build a predictive model.

5. Experiment and results

5.1. Data source

In this work, information of breast cancer patients was obtained from SEER (Surveillance, Epidemiology, and End Results) [33] with the SEERStat software. The data was recorded in 2004–2007 with 270,989 records and 151 variables, as shown in Appendix A. In the pre-processing step, the original dataset was cleaned. In fact, after studying the data dictionary of dataset, the unrelated variables to breast cancer and the records containing the missing values were removed from the dataset. Additionally, data transformation including feature normalization was provided to scale attribute values. One important thing that should be considered in the process of removing records with missing values is that data distributions should not be changed significantly. Here the records containing the missing values were removed from the dataset. Statistical analysis confirms that after removing missing values, there are not considerable changes in the distribution of variables. For instance, in terms of mean and standard deviation, there is no significant change in the normal distribution of age variable before (μ=59.18, σ=13.1) and after (μ=59.08, σ=12.92) the deletion of these records.

After data pre-processing, the final dataset contains 117,561 records and 58 variables that are selected based on studying data dictionary attentively. The final variables are shown with symbol "*" in Appendix A as well. The dependent variable is classified into two categories based on the method introduced in [10]: "Survival" class with a distribution of 91.76% and "Non Survival" class with 8.24%. After applying Borderline-SMOTE and DSO to original dataset to deal with the class imbalanced problem, the two different datasets dataset_1 and dataset_2 were generated, respectively. The distribution of the dependent variables for each dataset is shown in table 1.

Then PSO_CFS method was applied to these two datasets to extract the most important predictor variables. Summaries of selected predictor variables from each dataset are shown in tables 2 and 3.

This work uses 20% of the dataset with random sampling. Sampling was performed 10 times by changing the seed of the random number generator. Since samples are random, the accuracy of one execution of the algorithm on one set cannot be an indicator of its accuracy on the entire data. Therefore, sampling was repeated 10 times, and 10-fold cross-validation was used to evaluate each execution of the algorithm. As seen in tables 2 and 3, two different sets of features are selected for each dataset that have some common variables such as Grade, Tumor Size, and Num_ positive _nodes. Two types of variables categorical and numerical exist in the selected variables. For categorical variables, distinct values of each variable is presented, and for numerical variables, some statistical metrics such as Mean, Standard Deviation (S.D.), Maximum number (Max), and Minimum number (Min) are measured. Brief descriptions of predictor variables are presented in Appendix B.

Table 1. Distribution of class variable. Dataset_1 and Dataset_2 are new generated datasets by applying Borderline-SMOTE and DSO to original dataset, respectively.

Dataset	Class	Number of records	Percentage
Original Dataset	Survival=1	107,880	91.76%
	Non Survival=0	9,681	8.24%
	Total	117,561	100%
Dataset_1	Survival=1	107,880	83.23%
	Non Survival=0	21,733	16.77%
	Total	129,613	100%
Dataset_2	Survival=1	107,880	69.03%
	Non Survival=0	48,405	30.97%
	Total	156,285	100%

Table 2. Summaries of predictor variables selected by applying PSO_CFS to dataset_1.

Feature number	Categorical variable		Distinct value		
A1_V3	Marital status at DX		5		
A2_V20	Grade		4		
A3_V37	Extension		26		
A4_V38	Lymph Nodes Involvement		32		
A5_V39	Met at Dx		7		
A6_V100	Stage of cancer		4		
A7_V121	ER Status Recode Breast Cancer		3		
Feature number	**Numerical Variable**	**Mean**	**S.D.**	**Min**	**Max**
A8_V36	Tumor Size	38.67	128.38	0	998
A9_V27	Num_ positive _nodes	1.33	3.79	0	97

Table 3. Summaries of predictor variables selected by applying PSO_CFS to dataset_2.

Feature number	Categorical variable		Distinct value		
A1_V4	Race/ ethnicity		29		
A2_V15	Laterality		5		
A3_V18	Histology		86		
A4_V20	Grade		4		
A5_V65	Radiation		8		
Feature number	**Numerical Variable**	**Mean**	**S.D.**	**Min**	**Max**
A6_V8	Age at diagnosis	58.74	13.79	18	100
A7_V28	Num_ Nodes	9.37	7.81	1	90
A8_V27	Num_ positive _nodes	2.88	5.62	0	97
A9_V36	Tumor Size	50.94	144.54	0	998
A10_V105	Number of primaries	1.11	0.32	1	6

5.2. Assessment metrics

Accuracy is the most common measure for classification task. However, it can be a misleading metric in the presence of the class imbalance problem. Thus other metrics such as G-mean and sensitivity must be calculated in this case because these metrics specify the performance of classification for minority class. Also specificity indicates the proportion of majority samples that are correctly identified. In this work, these four metrics (Equations 13-16) were used to evaluate the performance of models [34,35]. Also the area under a ROC curve (AUC) (17) was used to evaluate the performance of the feature selection method. These metrics were calculated based on the confusion matrix shown in table 4.

Table 4. Confusion matrix.

Predicted class→ / ↓Actual class	Non-Survival	Survival
Non-Survival	TP	FN
Survival	FP	TN

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{16}$$

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{17}$$

where, TP_{rate} refers to the proportion of positive samples that are correctly considered as positive, with respect to all positive samples, and FP_{rate} refers to the proportion of negative samples that are mistakenly considered as positive, with respect to all negative samples. In facts, TP_{rate} is equal to sensitivity and FP_{rate} is calculated as follows:

$$FP_{rate} = \frac{FP}{FP + TN} \tag{18}$$

Table 5. Average performance of models.

Model	Accuracy	Sensitivity	Specificity	G-mean
A 1: C4.5	92.50	0.215	0.989	0.460
B 1: [15] + C4.5	85.87	0.856	0.865	0.860
C 1: [23] + C4.5	80.88	0.780	0.813	0.634
D 1: Borderline-SMOTE + PSO_CFS + C4.5	90.14	0.867	0.911	0.888
E 1: DSO + PSO_CFS + C4.5	94.33	0.930	0.948	0.939
A 2: BN	87.05	0.611	0.883	0.734
B 2: [15] + BN	85.87	0.856	0.865	0.860
C 2: [23] + BN	78.88	0.800	0.787	0.630
D 2: Borderline-SMOTE + PSO_CFS + BN	92.29	0.839	0.946	0.891
E 2: DSO + PSO_CFS + BN	91.64	0.876	0.948	0.911
A 3: LR	92.50	0.200	0.989	0.433
B 3: [15] + LR	74.52	0.752	0.782	0.766
C 3: [23] + LR	85.66	0.880	0.853	0.751
D 3: Borderline-SMOTE + PSO_CFS + LR	90.19	0.829	0.922	0.874
E 3: DSO + PSO_CFS + LR	89.75	0.917	0.888	0.902

5.3. Results

Table 5 provides the performance of the proposed approach compared to the conventional classifiers, methods in [15] and [23]. In this table, each combinational method is represented by an alphabetic (A: conventional classifier, B: [15] +conventional classifier, C: [23] + conventional classifier, D and E: proposed method) and numbers (numbers 1 to 3 refer to three classifiers C4.5, BN, and LR, respectively). For example, C2 means [23] + BN. The comparison between the methods is based on the averaged results of accuracy, sensitivity, specificity, and G-mean. Also, the evaluation result of PSO_CFS method in terms of AUC is presented in table 6, in which DSO + classifiers are applied to four various datasets with different features including original 58 features, 16 features in [10], 20 features in [15] and 10 selected features by PSO_CFS (proposed method).

According to table 6, the feature set 4 (selected features using proposed approach) outperforms among others using all three classifiers. Further analysis about whether these differences are statistically significant or not will be discussed later.

The differences in the performances of the algorithms are detected by statistical tests using SPSS. This work used analysis of variance (ANOVA) test, in which the model is treated as a factor. ANOVA test is used to determine whether there are any statistically significant differences between the means of three or more independent groups. However, it cannot tell you which specific groups were significantly different from each other. To determine it, a post hoc test was used. In this work, Tukey's HSD post hoc test are used to identify the distinctive models. In this test, first the differences between the means of all groups are found. Then this difference score is compared to HSD (honestly significant difference), which is calculated as follows:

$$HSD = q\sqrt{\frac{MS_{within}}{n}} \tag{19}$$

where, MS_{within} is Mean Square value from the ANOVA test that is already computed, q or the Studentized Range Statistic is a table value, and n is the number of values in each group. If the difference is larger than the HSD value, it means the difference is significant. In this work, the significant level (α) is defined at 0.05. The results of ANOVA test in table 7 detect a significant difference (p values <0.05) among these algorithms for all indices. Thus post hoc test is required to determine which specific groups were significantly different. Figures 2-6 are outputs of this test for all metrics. Tukey's HSD test identifies the differences between methods, it lists the different algorithms in different columns while the indifferent algorithms are listed in the same column. In our test, there are 15 different combinational methods, so there should be 15 columns if there are significant differences between all methods. As seen in figure 2, there are 13 columns because there are not significant differences between C1-C2 and B1-B2.

Table 6. AUC results of DSO + Classifiers with different set of features.

Features	C4.5	BN	LR
1- 58 original features	0.889	0.893	0.876
2- 16 features in [10]	0.921	0.901	0.891
3- 20 features in [15]	0.930	0.891	0.868
4- 10 features PSO_CFS (proposed approach)	0.939	0.905	0.903

Table 7. F Statistic and P value of ANOVA test for all indices.

Index	F statistic	P Value
G-mean	1.036×10^4	0.000
Sensitivity	2.137×10^4	0.000
Specificity	1.516×10^3	0.000
Accuracy	1.356×10^3	0.000
AUC	1.356×10^4	0.000

5.4. Discussion

As mentioned in the previous sections, predicting "Non-Survival' samples (minority samples) is a challenging issue for conventional classifiers. According to table 5, when these conventional classifiers are applied to the imbalanced dataset in spite of high accuracy, detection rate of "Non-Survival" samples is low in term of sensitivity and G-mean.

Sensitivity and G-mean are two important metrics for analysing problems with imbalanced nature. Experimental results indicate that the proposed method in both cases using Borderline-SMOTE (method D in the Table 5) and DSO (method E in Table 5) are effective and could detect "Non-survival' samples with all three classifiers.

Furthermore, the results in table 5 show that the combination of DSO + PSO_CFS + classifiers is more effective that Borderline-SMOTE + PSO_CFS + classifiers. Thus it can be concluded that for the SEER dataset, increasing samples around centroid data is more effective than boundary samples.

Tukey's HSD test results figure 2-5 also indicate that the results of these two methods are significantly different in all metrics except for specificity metric figure 4 (the difference between E1, E2, and D2 is not significant). Among three classifiers, C4.5 with DSO offers the best performance with sensitivity 0.930, and the second and the third ones are DSO + LR and method in [23] + LR with values of 0.917 and 0.880, respectively.

Furthermore, the performance of the proposed method in [15] is not effective enough and even results in two metrics accuracy and specificity are worse than conventional classifiers.

The evaluation result of feature selection method in table 6 shows that the proposed method provides the best performance compared to the others, and the result of Tukey's HSD test for AUC in figure 6 confirms it as well.

Model	α=0.05												
	1	2	3	4	5	6	7	8	9	10	11	12	13
A3	0.433												
A1		0.460											
C2			0.630										
C1			0.634										
A2				0.734									
C3					0.751								
B3						0.766							
B1							0.860						
B2							0.860						
D3								0.874					
D1									0.888				
D2										0.891			
E3											0.902		
E2												0.911	
E1													0.939
Sig.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure 2. Tukey's HSD test for G-mean.

Model	α=0.05													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A3	0.200													
A1		0.215												
A2			0.611											
B3				0.752										
C1					0.780									
C2						0.800								
D3							0.829							
D2								0.839						
B1									0.856					
B2									0.856					
D1										0.867				
E2											0.876			
C3												0.880		
E3													0.917	
E1														0.930
Sig.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure 3. Tukey's HSD test for Sensitivity.

Model	α=0.05								
	1	2	3	4	5	6	7	8	9
B3	0.782								
C2	0.787								
C1		0.813							
C3			0.853						
B1				0.865					
B2				0.865					
A2					0.883				
E3					0.888				
D1						0.911			
D3							0.922		
D2								0.946	
E1								0.948	
E2								0.948	
A1									0.989
A3									0.989
Sig.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure 4. Tukey's HSD test for Specificity.

Model	α=0.05											
	1	2	3	4	5	6	7	8	9	10	11	12
B3	74.52											
C2		78.88										
C1			80.88									
C3				85.66								
B1					85.87							
B2					85.87							
A2						87.05						
E3							89.75					
D1								90.14				
D3								90.19				
E2									91.64			
D2										92.29		
A1											92.50	
A3											92.50	
E1												94.33
Sig.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure 5. Tukey's HSD test for Accuracy.

Model	α=0.05							
	1	2	3	4	5	6	7	8
LR+3	0.868							
LR+1		0.876						
C4.5+1			0.889					
LR+2				0.891				
BN+3				0.891				
BN+1				0.893				
BN+2					0.901			
LR+4					0.902			
BN+4					0.905			
C4.5+2						0.921		
C4.5+3							0.930	
C4.5+4								0.939
Sig.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure 6. Tukey's HSD test for AUC.

6. Conclusion

Imbalanced problem is a challenging issue for predicting a 5-year survivability of breast cancer patients. This work investigated the effectiveness of the two methods Borderline-SMOTE and DSO on imbalanced breast cancer dataset obtained from SEER to build a more accurate survival predictive model. After balancing dataset, a combination of PSO and CFS was used for feature selection since the large number of predictor variables and also the presence of correlation between them decreased the performance of the model. Finally, the three classifiers C4.5, BN, and LR were used to evaluate the proposed hybrid approach. A

combination of PSO and CFS selected 10 predictor variables that is fewer than two other studies [10,15] and AUC results showed that the feature selection approach (PSO_CFS) was effective for this dataset. Additionally, between the methods for solving the class imbalance problem, DSO was better than Borderline-SMOTE. In overall, DSO+ PSO_CFS + C4.5 presents best efficiency in criteria of accuracy, sensitivity, and G-mean with values 94.33%, 0.930 and 0.939, respectively.

References

[1] Khan, U., et al. (2008). WFDT - Weighted Fuzzy Decision Trees for Prognosis of Breast Cancer Survivability, presented at the AusDM.

[2] Ministry of Health and Medical Education, (2013). Center for Disease Management, Cancer Department. National Registration Cancer Cases Reported in 2010. Iran: Center for Disease Management, P.344-9.

[3] Gupta, D. K. S., & Sharma, A. (2001). Data mining classification techniques applied for breast cancer diagnosis and prognosis, Indian Journal of Computer Science and Engineering, 188-193.

[4] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods, Artif Intell Med, vol. 34, 113-27.

[5] Wang, K.-J. B., et al. (2014). A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients, Applied Soft Computing, vol. 20, pp. 15-24.

[6] Chen, Y. (2009). Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets, Department of Computer Science Iowa State University, pp. 1-5.

[7] Wang, K-J., Makond, B., & Wang, K-M. (2013). An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data, BMC Medical Informatics and Decision Making, vol. 13.

[8] Ardakani, A., & Kohestani, V. R. (2015). Evaluation of liquefaction potential based on CPT results using C4. 5 decision tree. Journal of AI and Data Mining, vol. 3, no. 1, pp. 85-92.

[9] He, H., & Garcia, E., (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, vol. 21, pp. 1263-1284.

[10] Bellaachia, A., & Guven, E., (2006). Predicting Breast Cancer Survivability using Data Mining Techniques, presented at the Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining, 2006.

[11] Endo, A., Takeo, S., & Tanaka, H., (2007). Predicting Breast Cancer Survivability: Comparison of Five Data Mining Techniques. Journal of Korean Society of Medical Informatics, vol. 13, pp. 177-180.

[12] Liu, Y-Q., Wang, Ch., & Zhang, L., (2009). Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data, 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009.

[13] Lotfnezhad Afshar, H., et al. (2015). Prediction of breast cancer survival through knowledge discovery in databases, Glob J Health Sci, vol. 7, pp. 392-8.

[14] Holte, R.C., Acker, L., & Porter, B. (1989). Concept Learning and the Problem of Small Disjuncts, Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89), vol. 1, pp. 813-818.

[15] K.-J. Wang, et al. (2014). A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients, Applied Soft Computing, vol. 20, pp. 15-24.

[16] Miri Rostami, S., Parsaei, M. R., & Ahmadzadeh, M. (2016). A survey on predicting breast cancer survivability and its challenges, UCT Journal of Research in Science, Engineering and Technology, vol. 4, no. 1, pp. 37-42.

[17] Chawla, N., et al. (2002). SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. Vol. 16, pp. 321-357.

[18] Zhao, X. M., Li, X., Chen, L., & Aihara, K., (2007). Protein classification with imbalanced data, Proteins vol. 70, no. 4, pp. 1125–1132.

[19] Parsaei, M. R., Miri Rostami, S., & Javidan, R. (2016). A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset. International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 6, pp. 20-25. Doi: 10.14569/IJACSA.2016.070603.

[20] Chawla, N. V., et al. (2003). SMOTEBoost: improving prediction of the minority class in boosting, in: Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Database, 2003, pp. 107-119.

[21] Han, H., Wang, W.-Y., & Mao, B.-H. (2005) Borderline-SMOTE: a new oversampling method in imbalanced data sets learning," Lect. Notes Comput. Sci, vol. 36, pp. 878-887.

[22] Ren, F., et al. (2016). Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm, Computerized Medical Imaging and Graphics, doi: http://dx.doi.org/10.1016/j.compmedimag.2016.07.011.

[23] Lim, P., Goh, C. K. & Tan, K. C. (2016). Evolutionary Cluster-Based Synthetic Oversampling Ensemble (ECO-Ensemble) for Imbalance Learning,

IEEE Transactions on Cybernetics, Doi: 10.1109/TCYB.2016.2579658.

[24] Parsaei, M.R., & Salehi, M. (2015). E-mail spam detection based on part of speech tagging. 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), pp. 1010-1013. Doi: 10.1109/KBEI.2015.7436182.

[25] Parsaei, M. R., Javidan, R., & Sobouti, M. J. (2016). Optimization of Fuzzy Rules for Online Fraud Detection with the Use of Developed Genetic Algorithm and Fuzzy Operators. Asian Journal of Information Technology, vol. 15, no. 11, pp. 1856-1864. Doi: 10.3923/ajit.2016.1856.1864.

[26] Douzas, G. & Bacao, F. (2017). Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning, Expert Systems with Applications, doi: 10.1016/j.eswa.2017.03.073.

[27] Parsa, S. S., Sourizaei, M., Dehshibi, M. M., Shateri, R. E., & Parsaei, M. R. (2017). Coarse-grained correspondence-based ancient Sasanian coin classification by fusion of local features and sparse representation-based classifier. Multimedia Tools and Applications, vol. 76, no. 14, pp. 15535-15560.

[28] Gao, W., Sarlak, V., Parsaei, M. R., & Ferdosi, M. (2017). Combination of fuzzy based on a meta-heuristic algorithm to predict electricity price in an electricity markets. Chemical Engineering Research and Design, pp. 1-13. Doi: 10.1016/j.cherd.2017.09.021.

[29] Komijani, H., Parsaei, M. R., Khajeh, E., Golkar, M. J., & Zarrabi, H. (2017). EEG classification using recurrent adaptive neuro-fuzzy network based on time-series prediction. Neural Computing and Applications, pp. 1-12. Doi: 10.1007/s00521-017-3213-3.

[30] Nabaei, A., Hamian, M., Parsaei, M. R., Safdari, R., Samad-Soltani, T., Zarrabi, H. & Ghassemi, A. (2016). Topologies and performance of intelligent algorithms: a comprehensive review. Artificial Intelligence Review, pp. 1-25. Doi: 10.1007/s10462-016-9517-3.

[31] Chen, L. F., et al. (2011). Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis, Neural Computing and Applications, pp. 1–10.

[32] Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages pp. 359–366.

[33] SEER (2015) Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2012). National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2015, based on the November 2014 submission.

[34] Parsaei, M. R., Taheri, R. & Javidan, R. (2016). Perusing the effect of discretization of data on accuracy of predicting naive bayes algorithm. Journal of Current Research in Science, no. 1, pp. 457-462.

[35] Parsaei, M. R., Javidan, R., Kargar, N. S., & Nik, H. S. (2017). On the global stability of an epidemic model of computer viruses. Theory in Biosciences, pp. 1-10. Doi: 10.1007/s12064-017-0253-2.

Appendix

Appendix A. All variables from SEER database (151 variables).

Number	Feature	Number	Feature
1	Patient ID number	20	Grade *
2	Registry ID	21	Diagnostic confirmation
3	Marital status at DX*	22	Type of reporting source
4	Race/ethnicity*	23	EOD—Tumor size
5	Spanish/hispanic origin	24	EOD—Extension
6	NHIA derived hispanic origin	25	EOD—Extension prost path
7	Sex	26	EOD—Lymph node involv
8	Age at diagnosis*	27	Regional nodes positive*
9	Year of birth	28	Regional nodes examined*
10	Birth place	29	EOD—Old 13 digit
11	Sequence Number–Central *	30	EOD—Old 2 digit
12	Month of diagnosis*	31	EOD—Old 4 Digit
13	Year of diagnosis*	32	Coding System for EOD
14	Primary site*	33	Tumor Marker 1
15	Laterality*	34	Tumor Marker 2
16	Histology (92-00) ICD-O-2	35	Tumor Marker 3
17	Behavior (92-00) ICD-O-2	36	CS Tumor Size*
18	Histologic type ICD-O-3*	37	CS Extension*
19	Behavior code ICD-O-3*	38	CS Lymph Nodes*
Number	**Feature**	**Number**	**Feature**
39	CS Mets at Dx*	96	CS Schema v0204*

40	CS Site-Specific Factor 1*	97	Race recode (White, Black, Other)
41	CS Site-Specific Factor 2*	98	Race recode (W, B, AI, API)
42	CS Site-Specific Factor 3*	99	Origin recode NHIA (Hispanic, Non-Hisp)
43	CS Site-Specific Factor 4*	100	SEER historic stage A*
44	CS Site-Specific Factor 5*	101	AJCC stage 3rd edition (1988-2003)
45	CS Site-Specific Factor 6*	102	SEER modified AJCC Stage 3rd ed (1988–2003)
46	CS Site-Specific Factor 25*	103	SEER Summary Stage 1977 (1995–2000)
47	Derived AJCC T	104	SEER Summary Stage 2000 (2001–2003)
48	Derived AJCC N	105	Number of primaries*
49	Derived AJCC M	106	First malignant primary indicator*
50	Derived AJCC Stage Group	107	State-county recode
51	Derived SS1977	108	Cause of Death to SEER site recode
52	Derived SS2000	109	COD to site rec KM
53	Derived AJCC—Flag	110	Vital Status recode*
54	Derived SS1977—Flag	111	IHS Link
55	Derived SS2000—Flag	112	Summary stage 2000 (1998+)*
56	CS Version Input Original	113	AYA site recode
57	CS Version Derived	114	Lymphoma subtype recode
58	CS Version Input Current	115	SEER Cause-Specific Death Classification*
59	RX Summ—Surg Prim Site*	116	SEER Other Cause of Death Classification
60	RX Summ—Scope Reg LN Sur*	117	CS Tumor Size/Ext Eval*
61	RX Summ—Surg Oth Reg/Dis*	118	CS Lymph Nodes Eval*
62	RX Summ—Reg LN Examined	119	CS Mets Eval*
63	RX Summ—Reconstruct 1st	120	Primary by international rules
64	Reason for no surgery*	121	ER Status Recode Breast Cancer (1990+)*
65	RX Summ—Radiation*	122	PR Status Recode Breast Cancer (1990+)*
66	RX Summ—Rad to CNS	123	CS Schema -AJCC 6th ed (previously calledv1)
67	RX Summ—Surg/Rad Seq	124	S Site-Specific Factor 8*
68	RX Summ—Surgery type	125	CS Site-Specific Factor 10*
69	RX Summ—Surg site 98-02	126	CS Site-Specific Factor 11*
70	RX Summ—Scope Reg	127	CS Site-Specific Factor 13*
71	RX Summ—Surg Oth 98-02	128	CS Site-Specific Factor 15*
72	SEER record number	129	CS Site-Specific Factor 16*
73	Over-ride age/site/morph	130	Lymph vascular invasion*
74	Over-ride seqno/dxconf	131	Survival months*
75	Over-ride site/lat/seqno	132	Survival months flag
76	Over-ride surg/dxconf	133	Survival months – presumed alive
77	Over-ride site/type	134	Survival months flag – presumed alive
78	Over-ride histology	135	Insurance recode (2007+)
79	Over-ride report source	136	Derived AJCC-7 T
80	Over-ride ill-define site	137	Derived AJCC-7 N
81	Over-ride leuk, lymph	138	Derived AJCC-7 M
82	Over-ride site/behavior	139	Derived AJCC-7 Stage Grp
83	Over-ride site/eod/dx dt	140	Breast Adjusted AJCC 6th T (1988+)*
84	Over-ride site/lat/eod	141	Breast Adjusted AJCC 6th N (1988+)*
85	Over-ride site/lat/morph	142	Breast Adjusted AJCC 6th M (1988+)*
86	SEER type of follow-up	143	Breast Adjusted AJCC 6th Stage (1988+)*
87	Age recode <1 year olds*	144	CS Site-Specific Factor 7*
88	Site recode ICD-O-3/WHO 2008*	145	CS Site-Specific Factor 9*
89	Recode ICD-O-2 to 9	146	CS Site-Specific Factor 12*
90	Recode ICD-O-2 to 10	147	Derived HER2 Recode (2010+)
91	ICCC site recode ICD-O-3/WHO 2008	148	Breast Subtype (2010+)
92	ICCC site rec extended ICD-O-3/WHO 2008	149	Birthplace – country
93	Behavior Recode for Analysis	150	Birthplace – state
94	Histology Recode—Broad Groupings	151	Lymphomas: Ann Arbor Staging (1983+)
95	Histology Recode—Brain Groupings		

Appendix B. Description of selected variables using proposed method (Predictor variables)

Number	Predictor variables	Description
1	Race/ethnicity	It indicates ethnicity and race of patients.
2	Laterality	It describes the side of a paired organ or side of the body on which the reportable tumor originated.
3	Histology	It describes the microscopic composition of cells and/or tissue for a specific primary. The histology is a basis for staging and determination of treatment options.
4	Grade	It describes a tumor in terms of how abnormal the tumor cells are compared to normal cells.
5	Radiation	This variable indicates the method of radiation therapy performed as part of the first course of treatment.
6	Age at diagnosis	It represents the age of the patient at diagnosis for this cancer.
7	Num_Nodes	It records the total number of regional lymph nodes that were removed and examined by the pathologist.
8	Num_ positive_nodes	It records the exact number of regional lymph nodes examined by the pathologist that were found to contain metastases.
9	Tumor Size	Information on tumor size. It records the largest dimension of the primary tumor in millimeters.
10	Number of primaries	It describes the number of all reportable malignant, in situ, benign, and borderline primary tumors, which occur over the lifetime of a patient.

نشریه هوش مصنوعی و داده کاوی

استخراج متغیرهای پیش‌بینی‌کننده برای ساخت مدل بقای بیماران دارای سرطان پستان با حضور مشکل کلاس نامتوازن

سمانه السادات میری رستمی* و مرضیه احمدزاده

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی شیراز، شیراز، ایران.

ارسال ۲۰۱۶/۱۱/۱۹؛ بازنگری ۲۰۱۷/۰۶/۱۴؛ پذیرش ۲۰۱۷/۰۹/۰۲

چکیده:

کاربرد روش‌های داده‌کاوی به عنوان سیستم پشتیبان تصمیم برای پیش‌بینی بقای بیماران، مزیتی بزرگ و موضوعی جدید برای پژوهشگران حوزه سلامت می باشد که به دنبال یافتن رابطه بین عوامل خطرزا و بقا سرطان هستند. با این وجود به دلیل ماهیت نامتوازن مجموعه داده‌های مرتبط با بقای سرطان پستان، دقت مدل پیش‌بینی‌کننده برای این نوع از سرطان چالش برانگیز شده است. هدف این تحقیق ساخت مدل پیش‌بینی‌کننده برای بقای ۵ سال بیماران دارای سرطان پستان و یافتن رابطه بین متغیرهای پیش‌بینی‌کننده و بقا است. برای این کار از مجموعه داده SEER استفاده شده است. ابتدا برای حل مساله داده نامتوازن اثربخشی روش SMOTE مرزی و روش تولید نمونه مصنوعی بر اساس چگالی (DSO) مورد بررسی قرار گرفت. سپس برای یافتن مهمترین متغیرهای پیش‌بینی‌کننده از ترکیبی از الگوریتم بهینه‌سازی ازدحام ذرات (PSO) و روش انتخاب ویژگی مبتنی بر همبستگی (CFS) استفاده شده است. در نهایت برای ساخت مدل، سه دسته‌بند درخت تصمیم (C4.5)، شبکه بیزی و رگرسیون لجستیک به مجموعه داده نهایی اعمال شدند. معیارهای ارزیابی شامل Accuracy، Sensitivity، Specificity و G-mean برای ارزیابی عملکرد روش ترکیبی پیشنهادی بکار گرفته شدند. همچنین معیار AUC برای ارزیابی عملکرد روش انتخاب ویژگی پیشنهادی مورد استفاده قرار گرفت. نتایج حاصل نشان داد روش ترکیبیDSO+PSO_CFS+C4.5 از بین تمام ترکیب ها، بهترین عملکرد را بر اساس معیارهای Accuracy، Sensitivity، G-mean و AUC به ترتیب با مقادیر ۹۴/۳۳ درصد، ۰/۹۳۰، ۰/۹۳۹ و ۰/۹۳۹ ارائه داد.

کلمات کلیدی: سرطان پستان، بقا، مشکل کلاس نامتوازن، تکنیک بیش نمونه‌گیری، انتخاب ویژگی.