International Journal of Advanced Science and Technology
Vol.96 (2016), pp.1-10
http://dx.doi.org/10.14257/ijast.2016.96.01

Conformal Prediction Technique to Predict Breast Cancer Survivability

Loai M. Alnemer, Lama Rajab and Ibrahim Aljarah

The University of Jordan, Amman, Jordan
l.nemer@ju.edu.jo, lama.rajab@ju.edu.jo, i.aljarah@ju.edu.jo

Abstract

Breast cancer is a common disease around the world. Much work has been done to predict the survivability of breast cancer patients. While the prediction is an information that is important in many applications, the reliability of an individual classification result, is of great interest in many application areas including bioinformatics. In this paper, we apply the Conformal Prediction algorithm to the classification results of four data mining algorithms in order to eliminate the non-reliable predictions and enhance the overall classification results. The proposed technique shows an enhancement in the effectiveness of the classification results.

Keywords: *Breast Cancer, Classification, Reliability, Conformal Prediction*

1. Introduction

Tumor is formed by an out of control grow in the cells. This tumor is produced by a group of diseases called cancer. Breast cancer is a tumor that is named so because it happens in the breast tissues [1]. In USA alone, approximately 17% of women and 19% of men that will be diagnosed with breast cancer will die [1]. Knowing the survival rates of any type of cancer is very important for both patients and doctors. Survivability prediction of any disease is possible by extracting the features that are related to that disease. In this paper, A SEER breast cancer database [2] (Surveillance Epidemiology and End Results) is used, this dataset is an important, reliable and essential resource for many type of cancers. It contains a lot of information including: tumor pathology, cancer stage, and the cause of death [3, 4].

Much work has been done in predicting the survivability of breast cancer [5-7]. Such approaches ignore the reliability of bare prediction where the prediction of an individual record is only a part of the information that is possibly relevant to a user.

The reliability of individual classification results is studied comprehensively in many statistics and machine learning techniques such as discriminant analysis, distance to hyperplane in support vector machines, and the Bayesian classifiers. Discriminant analysis is a technique to differentiate between groups based on several variables [8], and to classify samples into those groups. The input data for discriminant analysis are assumed to be normally distributed. This assumption is not satisfied in many data sets of binary input data. The distance to the hyperplane in support vector machines is also used directly or indirectly for measuring the reliability of bare prediction [9]. Another example for measuring the reliability of individual classification results is to use the Bayesian classifiers [10]. The probability of each prediction can be used to assess the classifier output as introduced in [11].

In this paper, a conformal prediction technique is introduced [12] which is applied to the traditional classification algorithm to enhance the prediction results by eliminating the non-reliable predictions. Conformal prediction algorithm is used to calculate the level of confidence of each prediction using the previous predictions [12]. It was originally designed for an online setting where the prediction of the record depends on the records that have

ISSN: 2005-4238 IJAST

been already predicted. According to this setting, the comparison step will be done between the current record and the records that had been already predicted. In this paper, reliability of the new prediction is evaluated assuming that all other records in the training set are already predicted.

For traditional prediction, four different data mining algorithms are used; Artificial Neural Networks (ANN) [13], Support Vector Machine [9], K-Nearest Neighbors and Decision Trees. The four mentioned algorithms are highly used in many classification techniques, and they prove their effectiveness in the prediction.

The rest of this paper is organized as follows: Section 2 discusses the related work. In Section 3 we will discuss the dataset and the used features and implemented algorithm. The compassion algorithms and the experimental results are described and discussed in Section 4. Finally, Section 5 concludes the paper.

2. Related Work

There are several studies on the survivability prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and survivability using data mining approaches like decision trees [5-7].

In [5], Delen *et. al.*, used two popular data mining algorithms (artificial neural networks and decision trees) along with the most commonly used statistical method (logistic regression) and developed the prediction models using a large SEER dataset. The SEER dataset was preprocessed to remove redundancies and missing information. The resulting dataset had more than 200,000 cases which were classified into two groups Survived and not Survived depending on the Survival Time Record (STR). In order to measure the unbiased estimate of the three proposed prediction models, a 10 Fold cross validation method was used. The results indicated that the decision tree induction method C5 is the best with 93.6% accuracy followed by the artificial neural networks with accuracy 91.2% and the worst was the logistic regression model with accuracy 89.2%.

In [14], the research has outlined and discussed the problem of breast cancer survivability prediction in SEER database. The authors applied some predication algorithms, and techniques such as NaiveBayes (NB), Back-Propagation neural network (BP), and C4.5 decision tree algorithms to SEER database. The pre-processed data consists of 151,886 records, which have all the available 16 fields from the SEER database. The proposed approach takes into consideration, the Survival Time Recode (STR), the Vital Status Recode (VSR) and Cause of Death (COD). Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, they found out that C4.5.algorithm has a much better performance than the other two techniques.

In [15], Saleema *et. al.*, compared three sampling techniques: random, stratified, and balanced stratified. The authors proposed an ideal sampling method based on the outcome of the experiments to enhance the results of traditional sampling methods.

In the context of the individual classification results reliability, much work has been presented in the literature using different techniques. In [16], the authors implement an algorithm to measure the significance of individual classification results. The algorithm was applied to three bioinformatics applications. In this work, the authors show the importance of the reliability of bare prediction in determining the position of gene on wheat chromosomes.

The conformal prediction approach was used in many classification problems to determine the reliability of individual classification results [17, 18].

3. Materials and Method

3.1 Dataset Preparation

SEER breast cancer database is used to evaluate the proposed method. The SEER data is requested through the SEER website (http://www.seer.cancer.gov) [2]. This database is widely used for research purposes. The survivability prediction of cancer patients were studied comprehensively. Each year, according to [4], SEER sends out approximately 1500 copies of their data files.

In data mining applications preparing and analyzing dataset is considered one of the most important steps. Dataset preparation consists of three main steps:

1. Data understanding: this step includes mainly; data exploration and data quality verification. This is done by viewing some of data statistics and identifying explicit patterns in the data.
2. Feature selection: this step determines the important features for the specific research objectives.
3. Missing value and redundant records elimination and data normalization.

Breast cancer data consists of 740,507 records and over 100 variables. These variables are not unique and need further processing to combine or eliminate some of them. For example the tumor size variable for cases before 2004 is called ”EOD−Tumor Size”, while it is called ”CS−Tumor Size” for cases after 2004. After eliminating the repeated records and the records with large number of missing data, the number of records is approximately 425,000 records. In this research, we used 15 attributes that we think they are related to the survivability of patients. The cause of death is used as a class label. If the patient dies because of the cancer the class label c_i=1, while if he still alive or died by other reason the class label c_i =0. The selected attributes are shown in Table 1, hence, the last attribute (CAUSE OF DEATH) is used as the class label. For this research, a random sample is selected, it contains 4000 records to evaluate the effectiveness of the proposed methodology. The ratio between positives and negatives example is preserved in the selected data.

Table 1. The Selected Attributes and the Class Label that used in this Research

Attribute number	Attribute name
1	MARITAL STATUS
2	RACE
3	SEX
4	AGE AT DIAGNOSIS
5	PRIMARY SITE
6	LATERALITY
7	HISTOLOGY
8	HISTOLOGIC TYPE
9	BEHAVIOR CODE
10	GRADE
11	TUMOR SIZE
12	EOD EXTENSION
13	LYMPH INVOLVEMENT
14	RX SUMMRADIATION
15	NUMBER OF PRIMARIES
16	CAUSE OF DEATH

3.2 Methodology

The proposed methodology is shown in Figure 1. The main idea is to use traditional classifiers to predict the class labels in the test set. Then for each classifier, the conformal prediction algorithm is applied to calculate the non-conformality score for each prediction and use it to calculate the confidence. The conformal prediction algorithm is fully described in sub-section 3.2.2. Then, we eliminate all the predictions that are below our confidence threshold.

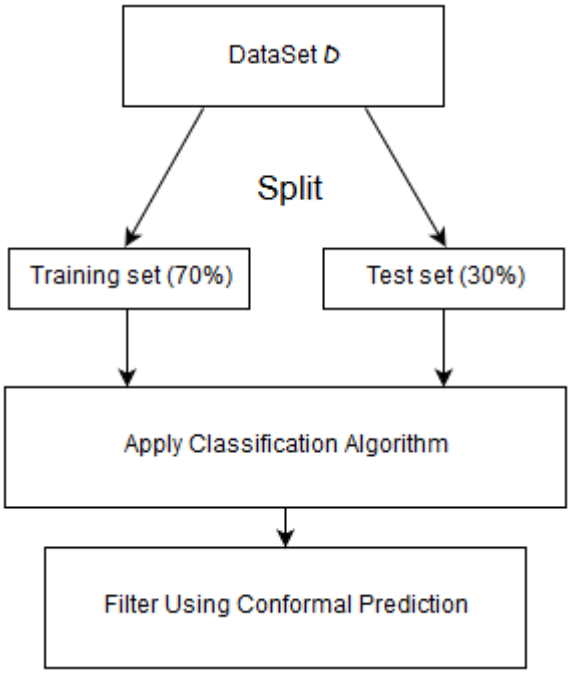


Figure 1. The Proposed Method Flowchart

3.2.1 Classification algorithms: Four well known classification algorithms were used to evaluate the effectiveness of the proposed method.

1. Support Vector Machines (SVM): which is one of the widely used techniques for data classification. In this paper, we used LIBSVM implementation [19] that is implemented in the WEKA data mining tools [20].

2. Decision Tree (DT): the main idea behind Decision Tree models is dividing data into groups based on empirically derived associations between the class label and the predictive attributes. In this paper, we use the J48 algorithm which is also implemented in the WEKA tools.

3. K-Nearest Neighbors (KNN): which is a lazy classifier that is widely used in data mining applications. In this work, we implement the KNN algorithm using Euclidian distance as a similarity measure with K=7.

4. Artificial Neural Networks (ANNs): which are non-parametric mathematical models inspired by the biological nervous system. In this work, we use a special type of the ANNs called Feedforward neural networks (FFNN) that are used to solve supervised problems such as classification problems [13]. In this paper, we use the Neural Network Toolbox to run the experiments [21].

3.2.2 Conformal Prediction: The conformal prediction algorithm is used to calculate the confidence level of each prediction using the previous predictions and it can be used with any classification technique [17]. Conformal prediction was originally designed for an

online learning, where the prediction of the record depends on the records that were already predicted. For comparison purposes, we calculate the confidence of the new record using the non-conformality score for all records in the training set. The non-conformality score measures the degree to which the relation between the record and all the records in the training set unusual and constructs a prediction region from the result. In this work, we use the cosine similarity measure to calculate the non-conformality score with the goal of eliminating the effect of distance measures in our comparison. The confidence of the prediction is calculated based on the following steps:

1. **Determining the non-conformality score:** Consider TR $=\{tr_1,tr_2,...,tr_n\}$ is the training dataset, where each tr_i contains the attribute set $att_i=\{att_1,att_2,...,att_m\}$ and the class label c_i, and consider that TS is the test set where ts_j is a record in TS, and $ts_j=\{att_j,c_j\}$ is a record where we know its attribute set att_j and we have a prediction of its class label c_i. Then, we measure the non-conformality score of ts_j by comparing its distance to all records in the training set that have the same label to its distance to all records in the training set with different label. Then, the non-conformality score is calculated by using the following Equation:

$$NC(TR,ts_i)=\frac{\min(1NN(ts_i,tr_i),\ 1\le j\le n\ \&\ c_i\neq c_j)}{\min(1NN(ts_i,tr_i),\ 1\le j\le n\ \&\ c_i=c_j)} \quad (1)$$

Where NC(TR,tsi) is the non-conformality score between the test record tsi, and all the records in the training set for the predicted class label ci.

2. **Calculating confidence of the prediction:** the confidence of the prediction is calculated by dividing the number of all records in the training set that have a non-conformality score higher than the test record over the number of all records in the training set.

4. Experiments and Results

The experiments in this research are conducted on a personal computer with a 2.6 GHz core i5 CPU and 4 GB RAM under Windows 8.1 Pro edition and the algorithm is implemented in Python 2.7 language.

As mentioned above, the proposed method is evaluated on a sample of SEER Breast cancer dataset, and the cause of death is considered as a class label. In order to evaluate the effectiveness of the proposed method, the Accuracy, Sensitivity, Specificity and Precision are calculated for each classifier results. Accuracy, Sensitivity, Specificity, Precision are calculated using Equations 2, 3, 4, and 5, respectively:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Sensitivity=\frac{TP}{TP+FN} \quad (3)$$

$$Specificity=\frac{TN}{TN+FP} \quad (4)$$

$$Precision=\frac{TP}{TP+FP} \quad (5)$$

After that, the classification results are compared with the results after applying the Conformal Prediction algorithm. Table 2 shows the number of True Positive, True Negative, False Positive, and False Negative predictions for each classifier before applying the Conformal Prediction algorithm.

Table 2. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) Predictions for Each Classifier

Algorithm	TP	FP	FN	TN
ANN	230	19	114	987
Decision Tree	246	31	98	974
KNN	227	32	117	973
SVM	246	25	98	980

The Accuracy of the classification results is shown in Figure 2. The Figure shows that the Accuracy after applying the Conformal Prediction algorithm outperforms the Accuracy of the original algorithms.

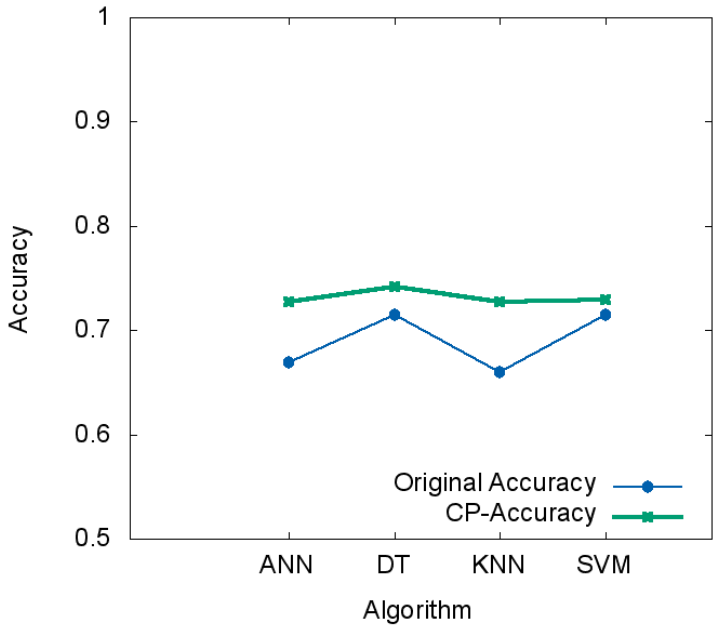


Figure 2. The Enhancement of the Accuracy of Each Algorithm after Applying the Conformal Prediction Algorithm

The Sensitivity of each classifier compared with the sensitivity of Conformal Prediction (CP) algorithm is shown in Figure 3. The Figure shows that the sensitivity, after applying the Conformal Prediction algorithm to each classifier results outperforms the sensitivity of the original algorithm for all used algorithms except the SVM where it is almost the same which means that the conformal prediction algorithm eliminates some of the false negative cases.

The Specificity of each classifier compared with the Specificity of Conformal Prediction algorithm is shown in Figure 4. The Figure shows that the Specificity after applying the Conformal Prediction algorithm to each classifier results is higher than the sensitivity of the original algorithm which means that the conformal prediction algorithm eliminates some of the false positive cases.

The Precision of each classifier compared with the Precision of Conformal Prediction algorithm is shown in Figure 5. The Figure shows that the Precision after we apply the Conformal Prediction algorithm to each classifier results is higher than the Precision of the original algorithm.

In general, the Conformal Prediction technique enhanced the results of all tested classifiers except for the SVM. This result can be justified since the distance to the decision boundaries in Support Vector Machines (SVM) can be considered as a reliability measure of a classifier. As discussed in [16] the distance to the decision boundaries is positively correlated with the confidence of conformal prediction algorithm.

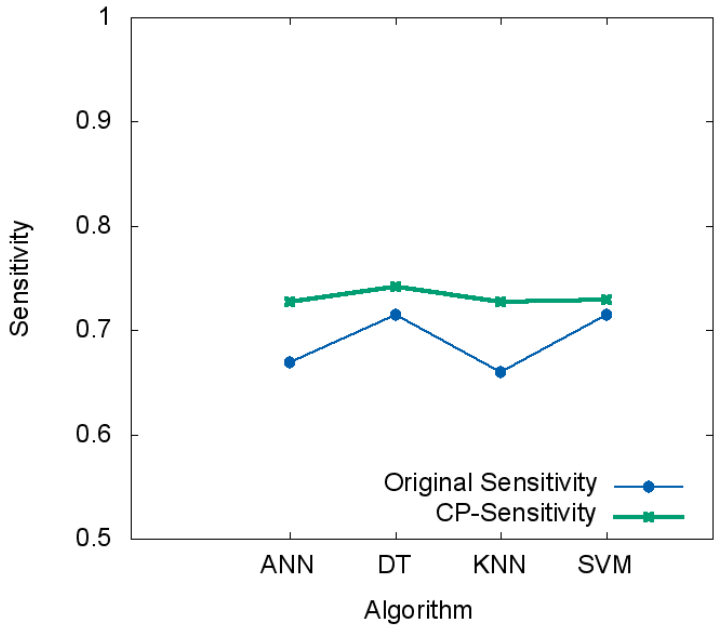


Figure 3. The Enhancement of the Sensitivity of Each Algorithm after Applying the Conformal Prediction Algorithm

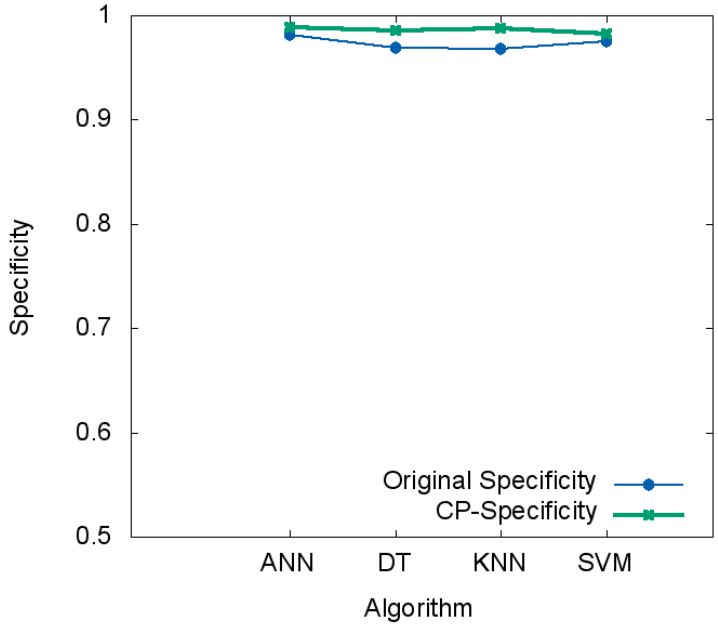


Figure 4. The Enhancement of the Specificity of Each Algorithm after Applying the Conformal Prediction Algorithm

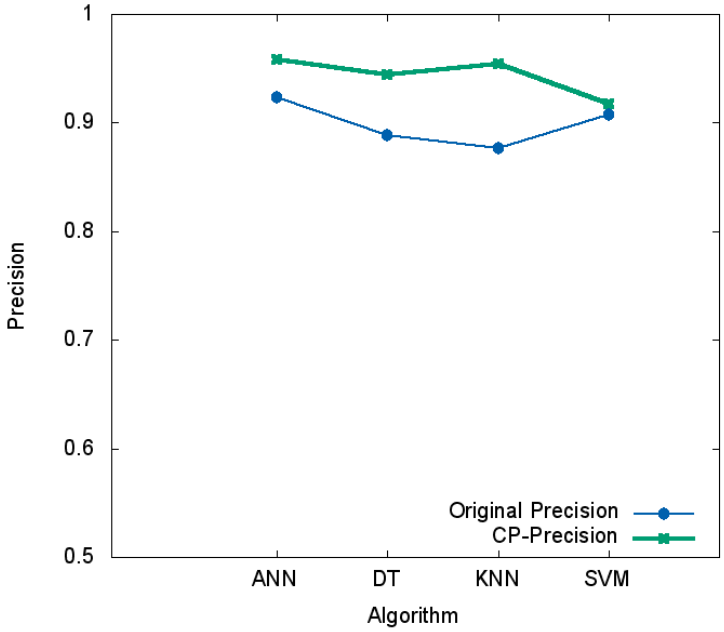


Figure 5. The Enhancement of the Precision of Each Algorithm after Applying the Conformal Prediction Algorithm

5. Conclusions

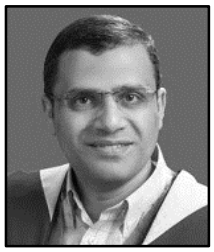
In this paper, a method to predict the survivability of breast cancer patients using four different machine learning algorithm was presented. Then the confidence of each individual results using Conformal Prediction algorithm was calculated to eliminate the non-reliable prediction. The proposed method showed that applying a confidence measure for each classification results reduces the number of false predictions and enhances the results.

References

[1] American Cancer Society, www.cancer.org.
[2] National Cancer Institute, Surveillance Research Program, C. S. B., 2003. Seer cancer statistics review. surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) publicuse data (19732012). Based on the November 2002 submission. Diagnosis period 19732012.
[3] D. R. Cox, D. O., 1984. Analysis of Survival Data. Chapman and Hall/CRC.
[4] B. F. Hankey, L. A. Ries and B. K. Edwards, "The surveillance, epidemiology, and end results program: A national resource", Cancer Epidemiology Biomarkers & Prevention, vol. 8, no. 12, **(1999)**, pp. 1117-1121.
[5] D. Delen, G. Walker and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods". Artif. Intell. Med., vol. 34, no. 2, **(2005)** June, pp. 113-127.
[6] Z.-H. Zhou and Y. Jiang, "Medical diagnosis with c4.5 rule preceded by artificial neural network ensemble", IEEE Transactions on Information Technology in Biomedicine, vol. 7, no. 1, **(2003)** March, pp. 37-42.
[7] M. Lundin, J. Lundin, H. Burke, S. Toikkanen, L. Pylkkanen and H. Joensuu, "Artificial neural networks applied to survival prediction in breast cancer". Oncology, vol. 57, no. 4, **(1999)**, pp. 281-286.
[8] W. R. Klecka, "Discriminant analysis", Wiley-Interscience Publication, **(1980)**.
[9] M. Fauvel, J. Chanussot and J.B enediktsson, "A combined support vector machines classification based on decision fusion", In 2006 IEEE International Symposium on Geoscience and Remote Sensing, **(2006)**, pp. 2494-2497.
[10] F. Ruggeri, R. Kenett and F. W. Faltin, "Encyclopedia of statistics in quality and reliability", **(2007)**.
[11] J. W. Krzanowski, C. T. Bailey, D. Partridge, E. J. Fieldsend, M. R. Everson and V. Schetinin, "Confidence in classification: A bayesian approach", Journal of Classification, vol. 23, no. 2, pp. 199-220.
[12] G. Shafer and V. Vovk, "A tutorial on conformal prediction", Journal of Mach. Learn. Res., vol. 9, **(2008)** June, pp. 371-421.

[13] I. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application", Journal of microbiological methods, vol. 43, no. 1, **(2000)**, pp. 3-31.
[14] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques", Age, vol. 58, no. 13, **(2006)**, pp. 10-110.
[15] J. Saleema, N. Bhagawathi, S. Monica, P. D. Shenoy, K. Venugopal and L. M. Patnaik, "Cancer prognosis prediction using balanced stratified sampling". arXiv preprint arXiv:1403.2950, **(2014)**.
[16] L. M. Al-Nimer, "Significance of individual classification results in bioinformatics applications", PhD thesis, Fargo, ND, USA. AAI3491589, **(2011)**.
[17] A. Shabbir, G. Verdoolaege, J. Vega and A. Murari, "Elm regime classification by conformal prediction on an information manifold", IEEE Transactions on Plasma Science, vol. 43, no. 12, **(2015)** December, pp. 4190-4199.
[18] S. Bhattacharyya, "Confidence in predictions from random tree ensembles", Knowledge and information systems, vol. 35, no. 2, **(2013)**, pp. 391-410.
[19] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines", ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, **(2011)** May, pp. 27:1-27:27.
[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The weka data mining software: an update", ACM SIGKDD explorations newsletter, vol. 11, no. 1, **(2009)**, pp. 10-18.
[21] Math works- Neural Network Toolbox, www.mathworks.com/products/neural-network/.

Authors



Loai M. Alnemer, he received the B.S. and M.S. degrees in Computer Science from The University of Jordan, Jordan, in 2001 and 2006, respectively, and Ph.D. degree in Computer Science from North Dakota State University, USA in 2012. He is currently an assistant professor at the Computer Information Systems Department in The University of Jordan and has held his position since 2012. His research interests include Bioinformatics, Data mining and Machine Learning.

Lama Rajab, she received the B.S. and M.S. degrees in Computer Science from The University of Jordan, Jordan, in 2002 and 2005. She is currently a Teacher at the Computer Information Systems Department at The University of Jordan and has held the position since 2006. Her research interests include Digital Image watermarking, Biomedical Image Processing and Data mining.

Ibrahim Aljarah, is an Assistant Professor of Computer Science at the University of Jordan - Department of Business Information Technology, Jordan. He obtained his bachelor degree in Computer Science from Yarmouk University - Jordan, 2003. Ibrahim also obtained his master degree in Computer Science and Information Systems from the Jordan University of Science and Technology - Jordan in 2006. He also obtained his Ph.D. in Computer Science from the North Dakota State University (NDSU), USA, in May 2014. He has published more than 25 papers in refereed international conferences and journals. His research focuses on Data mining, Machine Learning, Big Data, MapReduce, Hadoop, Swarm intelligence, Evolutionary Computation, Social Network Analysis (SNA), and large scale distributed algorithms. In addition, his research aims to make use of the nature-inspired approaches in data mining applications to be efficient and powerful for big data. Furthermore, His research benefits from the MapReduce methodology as a big data processing model to build scalable data mining algorithms.

