IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Special Issue || Dec. 2019 || PP 58-64
International Conference on Management Practices, Innovations & Research 2019

Analytical Study Of Various Machine Learning Algorithms For Cancer Diagnosis

Devashri Raich[1] ,Bireshwar Ganguly[2]
[1]*Department of Information Technology, Rajiv Gandhi College of Engineering Research & Technology*
[2]*Department of Information Technology, Rajiv Gandhi College of Engineering Research & Technology*

Abstract—*Machine learning algorithms are computer programs that attempt to predict something (behavior, type of cancer, image, stock market fluctuations, etc.) according to the situations that led to the past results. The ultimate goal of machine learning in cancer diagnosis is a machine learning algorithm that can accurately predict the type and severity of cancer, using gene expression data or other patient data and help the practitioner in the treatment. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and made available to the medical research community. However, accurate prediction of the course of the disease is one of the most interesting and challenging tasks of physicians. As a result, BC methods have become a popular tool for medical researchers. These techniques can assess and identify patterns and relationships between them based on complex datasets, while effectively predicting the future outcomes of a cancer type.*
Keywords—*Machine Learning, prediction, diagnosis, cancer*

I. Introduction

Cancer is not a single disease, but many related diseases that involve uncontrolled cell growth and reproduction. Cancer is the leading cause of death in the developed world and the second in the developing world, killing almost 8 million people a year. Since cancer is many diseases, treating an individual cancer requires knowing what abnormal behavior occurs inside the cells. This information can also help to understand the mechanisms underlying cancer, which can lead to new treatments. Cellular gene expression levels are a very useful data set.

This paper is a comparison of several machine learning algorithms, comparing its result on the diagnosis of cancer based on data on the level of gene expression. Both data sets that have been used were a set of data on breast cancer, classifying cancers in basal and luminal cancer, and a set of data on colorectal cancer, it was decisive if a cancer has a move in the p53 gene. This article compares four different automatic learning algorithms: Decision Tree, Majority, the closest neighbors and best Z-Score (an algorithm of my own conception that is a slight variant of the Bayes de Naıve algorithm).

II. Related Work

Machine learning is a branch of the search on artificial intelligence that uses various statistical and probabilistic and optimization tools to "learn" past examples and then use this previous training to classify new data, identify new models or predict new trends (Mitchell, 1997). Machine learning, like statistics, is used to analyze and interpret data. Contrary to statistics, machine learning methods can use Boolean logic (and, or, not), absolute conditionality (if, then, otherwise), conditional probabilities (the probability of X being given Y) and unconventional optimization strategies to model data or classify models. The latter methods are in fact similar to the approaches that humans routinely use to learn and classify. Machine learning is still widely founded on statistics and probabilities, but it is more powerful fundamentally because it allows conclusions to be drawn or decisions that otherwise could not be taken through conventional statistical methods (Mitchell, 1997; Doubt and coll., 1997). 2001). For example, numerous statistical methods are based on a regression or multivariate correlation analysis. Although they are generally very powerful, these approaches assume that the variables are independent and that the data can be modeled with the help of linear combinations of these variables. When relationships are nonlinear and that the variables are interdependent (or conditionally dependent), conventional statistics are generally vague. It is in situations that machine learning tends to shine. Numerous biological systems are fundamentally non-linear and their parameters depend conditionally. Numerous physical systems are linear and their parameters are essentially independent. The success of machine learning is not always guaranteed. As with every method, it is important to understand the problem well and understand the limits of the data. It is also to understand well the hypotheses and the limits of the applied algorithms. If a machine learning experience is correctly conceived, that the apprenants are correctly implemented and that the results are validated reliably, we generally have a good chance of success.

III. Machine Learning Algorithms

Machine learning (ML) may be defined as a subset of Artificial Intelligence that inculcates the ability of learning into a system on the basis of a data set used for the purpose of training in contrast to the normal approach of coding all possible outcomes beforehand. Multiple approaches and techniques are present to making systems which can learn. Some of them are neural networks, decision trees and clustering. ML is to be broadly categorized under three categories namely - reinforcement learning, supervised learning and unsupervised learning.

1) Supervised learning: generates a prediction function based on input observations. The function is generated from the training data and guides the system to produce useful epiphanies for the new data sets introduced into the system. In supervised learning algorithms a "premonitory" provider or teacher gives the learning algorithm marked a set of training data or examples. These (as) tagged examples are the set of training that the program tries to learn or learn to mapper the input data for the desired output.

2) Unsupervised learning: in this technique, the machine is forced to train from an unlabeled set of data and then differentiate it on a basis of certain characters and allow the algorithm to act on this information without external guidance. In unsupervised learning, a set of examples is given, but no label is provided. He belongs rather than learns to find the model or to discover the groups. This situation is a bit analogous to the process by which most graduate students learn. Unsupervised learning algorithms include methods such as self-organizing function charts (SOMs), hierarchical regrouping and K-means regrouping algorithms. These approaches create clusters from raw, unmarked or unclassified data. These groups can be used later to develop classification schemes or classifiers. The SOM approach (Kohonen 1982) is a specialized form of neural network or ANN. It is based on the use of an artificial neuron fence whose weights are adapted to correspond to the input vectors to a training set. In fact, SOM has been conceived at the beginning to model the biological function of the brain (Kohonen 1982).

A. SOM begins with a set of artificial neurons, each having its own physical location on the exit card, participating in a winner-take-all process process (a competitive network) where a knot with its weight vector closer to the vector of entries is declared the winner and their weights are adjusted that make them closer to the entry vector. Each knot has a set of neighbors. When this knot wins a competition, the weights of the neighbors are also changed, although to a lesser extent. The more the neighbor is removed to the winner, the more his weight changes. This process is then repeated for each input vector for a large number of cycles. Different entries produce different winners. The clear result is SOM that is capable of associating output nodes with specific groups or motifs in input data set. Interesting fact, almost all machine learning algorithms used in cancer prediction and prognosis use supervised learning. In addition, most of these supervised learning algorithms belong to a particular category of classifiers classified on the basis of conditional probabilities or conditional decisions.

3) Reinforcement Learning: This learning process is continued iteratively from the environment. All possible states of the system are eventually learned by the system over a prolonged period.

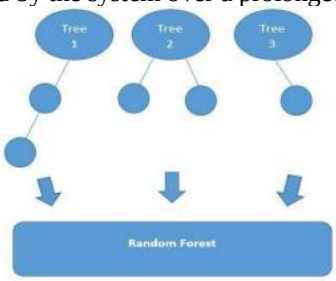

Figure 1: How Random Forest Works

B. Random Forest: It is a supervised learning algorithm. A set of decision trees is created. The method on the ground on which this technique is based is recurrence. A random sample of cut N is selected from the data set in each instance of iteration.

The dataset has been divided into training and testing sets, there are 398 observations for training set and 171 observations for testing. The number of estimators is set to 72 thus it is ensured that every observation is predicted at least a few times. It is obvious that diagnosis, radius_mean, texture_mean, perimeter_mean are influential variables, the other variables are of moderate influence but none of them can be neglected to increase the model accuracy. The confusion matrix of random forest is quite promising. There are only five observations that are misclassified as Benign and four observations are misclassified as Malignant and the accuracy equals 94.74%.

		Predicted	
		Benign	Malignant
Actual	Benign	103	5
	Malignant	4	59

Table 1: Random Forest Confusion Matrix C.

C. K-Nearest-Neighbor (kNN): K may be seen as the representation of the data points for training in close proximity to the test data point which we are going to use to find the class. A k-nearest-neighbor may be defined as the algorithm used to determine where a data set belongs to on the basis of the other data sets present around it. The technique is a supervised learning approach used for regression and classification. To process a new data point, KNN gathers all the data points close by to it. Attributes which have a large degree of variation are key factors in determining the distance. Given N training vectors in the Figure 2, kNN algorithm identifies the k nearest neighbors of regardless of labels.

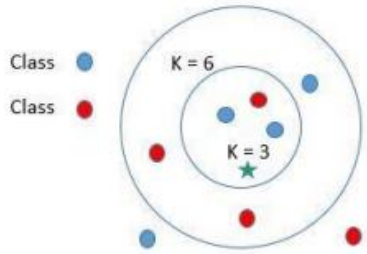

Figure 2: kNN Illustration

The accuracy of kNN is found to be 95.90% , there is only one observation that is misclassified as Benign and four observations are misclassified as Malignant as represented in Table 2. The results are comparatively better than Random Forest algorithm.

		Predicted	
		Benign	Malignant
Actual	Benign	107	1
	Malignant	6	57

Table 2: kNN Confusion Matrix

IV. Machine Learning Applications In Cancer Prediction

To prepare this review, several electronic databases have been consulted, particularly PubMed (biomedical literature), Science Citation Index (biomedical literature, engineering, computer science and physicochemistry), CiteSeer (computer literature), Google and Google Scholar (scientific literature accessible on Web). The Interrogation Terms included 'cancer and machine learning', 'cancer prediction and machine learning', 'cancer prognosis and machine learning', 'cancer risk assessment and machine learning' as well as multiple sub-interrogations with specific types of machine learning algorithms [21].

The relevance of the individual papers was assessed by reading the titles and abstracts and identifying papers that used recognizable machine learning methods as well as molecular, clinical, histological, physiological or epidemiological data in carrying out a cancer prognosis or prediction. Papers that focused on diagnoses or simple tumor classifications were excluded as were papers that had coincidental appearances of the words “machine” or “learning” in their abstracts.[15]. A PubMed search of “cancer and machine learning” yielded 1585 results, while searches of “cancer prediction and machine learning” and “cancer prognosis and machine learning” yielded 174 and 240 hits respectively. A detailed review of these abstracts led to the identification of 103 relevant papers of which 71 could be accessed through various library holdings. Using CiteSeer, a search with the terms “cancer and machine learning” yielded 349 results, of which 12 (3.4%) were deemed relevant to cancer prognosis. Using Google Scholar, a search using “cancer prognosis and ‘machine learning’” yielded 996 results, of which 49 (4.9%) were judged relevant to cancer prognosis. Many of these papers were previously identified in the PubMed searches as were the vast majority of the hits in the Science Citation Index searches. From the initial group of papers identified from these electronic searches, their

reference lists were further consulted to identify additional papers of interest or relevance. In the end more than 120 relevant papers were identified. Of these, more than 79 papers could be accessed from existing library holdings and were selected for more detailed analysis (Table 2). While it is impossible to be certain that we achieved complete coverage of all literature on machine learning and cancer prediction/prognosis, we believe that a significant portion of the relevant literature has been assessed for this review.

V. Proposed Methodology

The main objective of Machine Learning techniques is to produce a model that can perform different functions such as classification, prediction, estimation or very different similar task. The most common task in the learning process is classification. As mentioned above, this learning function classifies the data element into one of the many pre-defined classes. When a classification model is developed, training and generalization errors can be produced through ML techniques. The first technique refers to false classification errors on the training data, while the other refers to the expected errors on the test data. A good classification model should adapt well to the training set and accurately classify all instances. If the trial error rates of a model begin to increase, even if the training error rates decrease, then the phenomenon of surrégime of the model occurs. This situation is tied to the complexity of the model, which means that the errors in the formation of a model can be reduced if the complexity of the model increases. Clearly, the ideal complexity of a model that is not likely to be overestimated is the one that produces the weakest generalization error. A formal method to analyze the expected generalization error of a learning algorithm is the bias – variance decomposition. The rodeo of the component of a learning algorithm that measures the error rate of the algorithm.Additionally, a second source of error over all possible training sets of given size and all possible test sets is called variance of the learning method. The overall expected error of a classification model is constituted of the sum of bias and variance, namely the bias–variance decomposition.

When the data is preprocessed and we have defined the type of learning task, a list of machine learning methods is available, which includes (i) ANN, (ii) DT, (iii) SVM and (iv) BN. In this paper, we will refer only to these ML techniques that have been widely applied in the literature for the case study of cancer prediction and prognosis [28]. We identify trends in the types of ML methods used, the types of integrated data and the evaluation methods used to assess the overall performance of the methods used for cancer prediction or disease outcomes.

Naïve Bayes Classifiers which are probabilistic in nature, based on the application of Bayes theorem may be defined as Naive Bayes classifiers. It is naïve because it assumes that all features are independent from each other, this is generally not the case in real life scenarios, but still Naïve Bayes proves to be efficient for wide variety of machine learning problems. There are sixteen misclassified observations, seven of them being benign and nine of them are malignant. The same 398 observations are used for training set and 171 observations for testing and the accuracy equals to 94.47%.

		Predicted	
		Benign	Malignant
Actual	Benign	101	7
	Malignant	9	54

Table 3: Naïve Bayes Confusion Matrix E.

A. Comparison Among Proposed Algorithms:

Each one of the three algorithm's – kNN, Naïve Bayes and Random Forest have their advantage and disadvantage over each other in terms of performance, the type of problem they handle etc. As shown in Table 4: kNN test time is O(1) without preprocessing of training set [8], in the case of Naïve Bayes: N is the number of training examples and d is the dimensionality of the features whereas for Random Forest [9]: N is the number of samples and K is the number of variables randomly drawn at each node. Naïve Bayes algorithm deal only with classification problems whereas both kNN and Random Forest can deal with classification as well as regression problems. In terms of accuracy both kNN and Random Forest can deliver high accuracy but Naïve Bayes algorithm need large number of records in order to yield a better accuracy. Algorithms that simplify the function to a known form are called parametric machine learning algorithms, Naïve Bayes algorithm can be expressed as parametric as well as non-parametric model.

Parameter	KNN	Naïve Bayes	RandomForest
Time complexity (training phase)	O(1)	O(Nd)	Θ(MKNlog2N)
Problem Type	Classification & Regression	Classification	Classification & Regression
Accuracy	High	High only for large records	High
Model Parameter	Non Parametric	Non /Parametric	Non Parametric

B. **Dataset Description:**

The paper is based on Wisconsin Diagnosis Breast Cancer data set. The data set has been obtained from the 'UCI ML' repo, it has 570 instances and 35 attributes and there are no missing values. The output variable is either benign (358 observations) or malignant (212 observations). The most influential variables are diagnosis, radius_mean, texture_mean, perimeter_mean, area_mean etc. The positive class is used to for benin cases and the negative class is used in malignant cases. The k-fold cross-validation is utilised in which the presented data is divided into k equally sized bits.

C. **Performance Metrics**

.Data set	Number of attributes	Number of instances	Number of classes
WDBC	35	570	2

This section describes the parameters that are used for measuring performance of machine learning techniques. A confusion matrix for actual and predicted class is derived comprising of the standard five values namely True Positive, False Positive, True Negative and False Negative to evaluate the performance.

1. Accuracy: Accuracy is a good predictor for the degree of correctness in the training of the model and how it may perform generally. It may be defined as the measure of the correct prediction in correspondence to the wrong ones. Thus the equation presented can be used to calculate the value of accuracy:

Accuracy= (True Positive + True Negative)
(True Positive + False Positive+ True Negative + False Negative)

2. Recall: Recall known as sensitivity in general terms, may be defined as the ratio of rightfully determined positive instances to the all observations. Recall may be seen as a measure for the effectiveness of the system in predicting positives and determining costs.

Recall = True Positive
(True Positive + False Negetive)

3. Precision: The degree of correctness in determining the positive outcomes may be defined as precision. It is basically the ratio between true positives and the overall set of positives. This depicts the handling capacity of the system for positive values but does not provide insight into the negative values.

Precision = TP (TP + FP)

4. F1 Score: It is the weighted average of Precision and Recall. This measure hence, considers both type of false values. F1 score is considered perfect when at 1 and is a total failure when at 0.
F1 Score = 2*(Precision*Recall) (Precision + Recall)

VI. Implementation And Result Analysis

A comparative study using Random Forest, kNN (kNearest-Neighbor) and Naïve Bayes algorithm which are implemented in a computer having configuration as Intel Core i7 with 16GigaBits RAM has been proposed. We have used numpy, pandas and Scikit-learn which are open source machine learning libraries in Python. An open source web application named as Jupyter Notebook is used to run the program. The classifier was tested using the k − fold cross validation method. We have utilized the 10 fold technique that is the data set segregated in ten different chunks. Nine out of the folds used in the system are used for training and the last set is used for the purposes of testing and analysis. We have utilized 399 observations for training set and 171 observations for testing out of 570 observations. The graphical representations of the performance metrics for the three illustrated algorithms are shown in Figure 4. The results presented in Table 6 shows that Random Forest's has the best recall performance measure but kNN has the best accuracy, precision and F1 Score over Naïve Bayes and Random Forest.

Testing Phase			
%	RFF	KNN	NB
Accuracy	94.74	95.90	94.47
Recall	92.18	98.27	88.52
Precision	93.65	90.47	85.71
F1 Score	92.70	94.20	87.09

Table 6: Performance Measure Indices

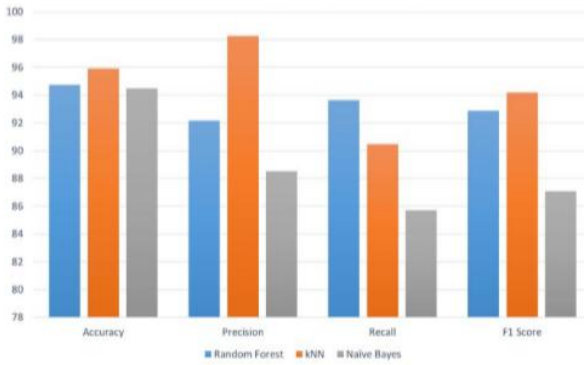

Figure 3: Graphical representation of Performance Measure Indices

VII. Conclusion

The most frequently occurring type of across cancer is breast cancer. There is a chance of twelve percent for a women picked randomly to be diagnosed with the disease [10]. Thus, early detection of breast cancer can save a lot of valuable life. The proposed model in this paper presents a comparative study of different machine learning algorithms, for the detection of breast cancer. Performance comparison of the machine learning algorithms techniques has been carried out using the Wisconsin Diagnosis Breast Cancer data set. It has been observed that each of the algorithm had an accuracy of more than 94%, to determine benign tumor or malignant tumor. From Table 6, it is found that kNN is the most effective in detection of the breast cancer as it had the best accuracy, precision and F1 score over the other algorithms. Thus supervised machine learning techniques will be very supportive in early diagnosis and prognosis of a cancer type in cancer research. As has been frequently noted before, the size of a given training set has several implications pertaining to robustness, reproducibility and accuracy.

The first implication is that for a smaller sample, almost all models are subject to overtraining. Overtraining can lead to informed details that can be misleading or inaccurate. For example, a first study reported only a classification error in training and RNA tests to predict the survival of hepatectomized patients using 9 different characteristics (Hamamoto et al., 1995). However, all data (training and tests) included only 58 patients. This particular study then used an external data set to validate the model in which the authors predicted the survival outcome prospectively with 100% accuracy. However, the set of external tests included only 11 patients. The fact that 100% accuracy is achieved for a prospective prediction is impressive, but given the size of the validation set and the small sample-function relationship, some doubts about the strength of the predictor can be expressed. Certainly, a larger validation set to strengthen the claim with 100% accuracy would be desirable. In another example, only 28 cases were used to construct RNA to predict throat cancer recurrence using expression levels of 60 genes from microarray data (Kan et al., 2004). The accuracy of the model has been estimated at 86%, but this is particularly suspicious given the very small sample size. In fact, it is very likely that this ANN has received excessive training.

The size of a given data set also significantly affects the sample-per-feature ratio. As a rule, the sampleper-feature ratio should be at least 5-10 (Somorjai et al. 2003). Small sample-per-feature ratios are a particularly big problem for microarray studies, which often have thousands of genes (ie features), but only hundreds of samples. The study by Ohira et al. (2005) provides one such example of the problems one may encounter trying to process too much microarray data. These authors created a probabilistic output statistical classifier to predict prognosis of neuroblastoma patients using microarray data from 136 tumor samples.

Each microarray had 5340 qualities, which brought about an example trademark proportion of around 0.025. Such a little example work relationship is touchy to overtraining issues. Furthermore, with an example trademark relationship of this size, it is likewise conceivable to grow exceptionally excess arrangement models that work similarly well in spite of preparing in various subsets of qualities. The issue with repetitive models is that the quality of any model can't be ensured as more experiments are accessible. Information size isn't the main restriction for compelling AI. The nature of the informational index and the cautious determination of highlights are additionally significant. For huge informational collections, information catch and confirmation is of fundamental significance. Frequently, careless information passage can prompt straightforward, once mistakes in which all qualities of a specific variable go up or down a line in a table. That is the reason a free review led by a subsequent caretaker or information verifier is constantly gainful. Far reaching confirmation or convenient check of information trustworthiness by an equipped master, not only an information section worker, is additionally an important exercise.

References

[1]. National Institute of Cancer Prevention and Research, cancer statistics [Online], Available: http://cancerindia.org.in/statistics/
[2]. WHO breast cancer statistics [Online]. Available: http://www.who.int/cancer/prevention/diagnosis-screening/breastcancer/en/
[3]. B.M. Gayathri, Dr. C.P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer" 2016
[4]. S Kharya and S Soni,"Weighted Naïve Bayes classifier –Predictive model for breast cancer detection", January 2016
[5]. Sivakami, "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model" 2015
[6]. B.M.Gayathri and C.P.Sumathi,"Mamdani fuzzy inference system for breast cancer risk detection", 2015.
[7]. Mohd,F.,Thomas,M, "Comparison of different classification techniques using WEKA for Breast cancer" 2007.
[8]. Time complexity and optimality of kNN [Online] Available: https://nlp.stanford.edu/IR-book/html/htmledition/time-complexityand-optimality-of-knn-1.html
[9]. Gilles Louppe, "Understanding Random Forests from theory to practice" 2015.
[10]. U.S. Breast Cancer Statistics [Online] Available: https://www.breastcancer.org/symptoms/understand_bc/statistics
[11]. T Choudhury, V Kumar, D Nigam ,An Innovative Smart Soft Computing Methodology towards Disease (Cancer, Heart Disease, Arthritis) Detection in an Earlier Stage and in a Smarter Way- International Journal of Computer Science and Mobile Communication (IJCSMC) 2014.
[12]. T Choudhury, V Kumar, D Nigam, B Mandal ,Intelligent classification of lung & oral cancer through diverse data mining algorithms, International Conference on Micro-Electronics and Telecommunication Engineering 2016
[13]. T Choudhury, V Kumar, D Nigam,Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm
[14]. [14] Aha D. 1992. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. International Journal of ManMachine Studies, 36:267-287.
[15]. Ando T, Suguro M, Hanai T, et al. 2002. Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma. Jpn J Cancer Res, 93:120712.
[16]. Ando T, Suguro M, Kobayashi T, et al. 2003. Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling. Cancer Sci, 94:906-13.
[17]. Atlas L, Cole R, Connor J, et al. 1990. Performance comparisons between backpropagation networks and classification trees on three real-world applications. Advances in Neural Inf. Process. Systems, 2:622-629.
[18]. Bach PB, Kattan MW, Thornquist MD, et al. 2003. Variations in lung cancer risk among smokers. J Natl Cancer Inst, 95:470-8.
[19]. Baldus SE, Engelmann K, Hanisch FG. 2004. MUC1 and the MUCs: a family of human mucins with impact in cancer biology. Crit Rev Clin Lab Sci, 41:189-231.
[20]. Bellman R. 1961. Adaptive Control Processes: A Guided Tour, Princeton University Press.
[21]. Bocchi L, Coppini G, Nori J, Valli G. 2004. Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks. Med Eng Phys, 26:303-12.
[22]. Bollschweiler EH, Monig SP, Hensler K, et al. 2004. Artificial neural network for prediction of lymph node metastases in gastric cancer: a phase II diagnostic study. Ann Surg Oncol, 11:506-11.
[23]. Bottaci L, Drew PJ, Hartley JE, et al. 1997. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. Lancet, 350:469-72.
[24]. Bryce TJ, Dewhirst MW, Floyd CE Jr, et al. 1998. Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck. Int J Radiat Oncol Biol Phys, 41:239-45.
[25]. Burke HB, Bostwick DG, Meiers I, et al. 2005. Prostate cancer outcome: epidemiology and biostatistics. Anal Quant Cytol Histol, 27:211-7.
[26]. Burke HB, Goodman PH, Rosen DB, et al. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, 79:857-62.
[27]. Catto JW, Linkens DA, Abbod MF, et al. 2003. Artificial intelligence in predicting bladder cancer outcome: a comparison of neurofuzzy modeling and artificial neural networks. Clin Cancer Res, 9:4172-7.
[28]. Cicchetti DV. 1992. Neural networks and diagnosis in the clinical laboratory: state of the art. Clin Chem, 38:9-10.