Seminal Quality Prediction using Optimized Artificial Neural Network with Genetic Algorithm

Azam Asilian Bidgoli[1], Hossein Ebrahimpour Komleh[1], Seyed jalaleddin Mousavirad[1]

[1] Department of Computer and Electrical Engineering
University of Kashan
Kashan, Iran
Asilian.b@gmail.com, Ebrahimpour@kashanu.ac.ir, Jalalmoosavirad@gmail.com

Abstract

Infertility problem is an important issue in recent decades. Semen analysis is one of the principle tasks to evaluate male partner fertility potential. It has been seen in many researches that life habits and health status affect semen quality. Data mining as a decision support system can help to recognize this effect. The artificial neural network (ANN) is a powerful data mining tool that can be used for this goal. The performance of ANN depends heavily on network structure. It is a very difficult task to determine the appropriate structure and is a discussable matter. This paper utilizes a genetic algorithm to optimize the structure of artificial neural network to classify the semen samples. These samples usually suffer from unbalancing problem. Thus, this paper attempts to resolve it by using the bootstrap method. The performance of the proposed algorithm is significantly better than the previous works. We achieve accuracy equal to 93.86% in our experiments on a real fertility diagnosis dataset that is a good improvement compared with other classification methods.

1. Introduction

Infertility problem is an important issue between couples in the last two decades. Demand for reproduction treatment has encouraged researchers to study about the reason of reproductive diseases [1]. Semen quality analysis is one of the most important activities to evaluate the male partner. Different factors can effect on the parameters of quality of sperms [2], [3]. These parameters are concentrated, motility and morphology of spermatozoa. Life habits and health status are part of such factors. In each of these categories, several instances exist, for example accident or serious trauma, childish diseases are considered as health status and smoking habit or frequency of alcohol consumption are life habits. Semen analysis can be a good predictor of male fertility potential.

In recent years, Machine learning and data mining techniques have been considered as good tools for automated medical diagnosis systems [4], [5]. The goal of using these tools in this special case is classification data in two classes: normal or altered. There are several algorithms for classification data. All of them have common phases for this goal: training and test phases. In the training phase, a classifier tries to learn some feature vectors as training data. In fertility detection every feature vector belongs to a volunteer in the database. Features include life habits and health status. The classifier should achieve the highest performance for classifying unknown data that are presented at the test phase.

A general problem in medical diagnosis data is class imbalance. This issue occurs when the number of examples that represent one class is much lower than the ones of the other classes [6]. There are different methods to handle this problem before classification phase [6]-[8]. Resampling techniques are in this category that affect the performance of classification. This technique deals with imbalance dataset using changing class distribution. Fertility data set suffers from this trouble. Hence, the class of "altered" cases has the minor distribution rather than "norma"l class. Therefor classification task is more sophisticated for such samples.

There are some works for predicting seminal quality with artificial intelligence methods. David Gi etal. [9] asses the accuracy in the prediction of seminal quality using three classification methods included: decision trees(DT) [10], Multi-Layer Perceptron(MLP) [11], [12] and Support Vector Machines(SVM) [13], [14]. In [15], MLP is utilized to achieve better performance to study different parameters of semen quality separately. Wong et al. [16] proposed a Clustering-Based Decision Forests method to tackle unbalanced class learning problem in seminal quality prediction.

This paper proposed an optimized MLP for predicting seminal quality based on life habits and health status data. Training in artificial neural networks is based on finding weights of neurons. The neural network is widely employed for data mining applications, but to achieve the best performance, a user should determine several parameters for constructing the structure of the network. The number of neurons and learning rate should be optimized in learning phase to reach the best performance. So it is an optimization problem. In this work, genetic algorithm(GA) [17] is applied for finding the appropriate artificial neural network structure and parameters. Neural network combined with genetic algorithm [18] gives the better performance compared with the previous works. The proposed method is compared with several classification methods such as SVM, DT and NaÃ¯ve Bayesian. The reminder of this paper is divided into several sections. Section 2 presents an introduction to MLP. The genetic algorithm is explained in section 3. The proposed approach is discussed in section 4. The results and conclusion are provided in section 5 and 6 respectively.
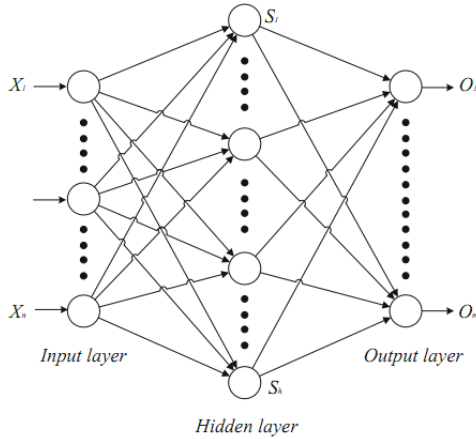
Figure 1. The structure of Neural Network [19]

2. Artificial neural networks

Neural Network is one of the most powerful tools for pattern recognition and data mining that inspired by human brains. The network structures made up of 2 parts: Neurons and weighted connections between neurons. Multilayer perceptron is a type of feed forward networks. These networks have an input layer for getting input data, an output layer for specifying the output of classification and one or more hidden layers between input and output layers for the learning process. A learning process in the network is based on finding the best connection weights and thresholds of neurons in hidden and output layers. The aim of determining weights is achievement of minimum error for classifying test data. The number of neurons in the first layer is equal to the dimensionality of feature vectors. The number of neurons in the last layer is based on the number of classes in classification problems. Finally, the number of hidden layers and the neurons in hidden layers depends on the specific problem. The number of neurons in the hidden layers can heavily affect the performance of the network. The process of learning is as follows.

The process of learning is as follows. At the first, the input data are fed to the ANN. Then weighted sum of input data by considering a threshold is calculated based on (1).

$$s_j = \sum_{i=1}^{n}(W_{ij}X_i) - \theta_j, \quad j = 1, 2, ..., h \qquad (1)$$

Where n is the number of dimensionality in feature vector equal to input neurons. W_{ij} is the connection weight from the ith node in the input layer to the jth node in the hidden layer, θ_j is the threshold of the jth hidden node, and X_i shows the future vector. The output of each hidden node is calculated by an activation function:

$$S_j = sigmoid(s_j) = \frac{1}{(1 + exp(-s_j))}, \quad j = 1, 2, ..., h \quad (2)$$

final outputs are defined based on the calculated outputs of the hidden nodes as follows:

$$o_k = \sum_{j=1}^{h}(w_{jk}.S_j) - \theta_k', \quad k = 1, 2, ..., m \qquad (3)$$

Where w_{jk} is the connection weight from the jth hidden node to the kth output node, and θ_k' is the threshold of the kth output node. Finally, the error is calculated by difference between the desirable output and obtained output as Mean Square Error(MSE).

$$E = \frac{1}{2}\sum_{p=1}^{n}(d_p - y_p)^2 \qquad (4)$$

where y_p is the actual output and d_p is the desired output for input pattern p. This calculation is a mathematical gradient-based algorithm that at each iteration of algorithm the error is minimized using the best weights and thresholds adjustment.

$$w(t + 1) = w(t) + \Delta w(t), \quad \Delta w(t) = -\eta \frac{\partial E(t)}{\partial(w)} \quad (5)$$

η is an independent parameter (a predetermined constant) known as the learning rate. A large value of η will lead to rapid learning but the weight may then oscillate , while low values imply slow learning [11].

3. Genetic algorithm

The genetic algorithm (GA) is based on Charles Darwin's theory of evolution. GA is an evolutionary algorithm that attempts to find the best solution in a search area for a determined problem. It is an iterative algorithm. The population is made up of a number of chromosomes as different potential solutions. Every time, a new population is formed using two operators named crossover and mutation. Initial population is chosen randomly. At each iteration of algorithm, an evaluation function named fitness is calculated for individuals in the population. All of the individuals are then ranked based on their fitness. After that, two best individuals are applied at the crossover stage to generate two offsprings. Mutation stage is applied on some offsprings to produce next generation. One or more gene values in a chromosome are altered from its initial state at mutation phase.. The population constantly improve and represent a better set of possible solutions after each iteration. The general structure of a GA is shown in Fig. 2.

4. The proposed method

In this section, the proposed method is explained. One of the most important problem about fertility detection is unbalanced data. The considered solution in this paper is bootstrap sampling. It uses sampling with replacement. In sampling with replacement, at every step all samples have equal probability of being selected. Once an example has been selected for sampling, it remains a candidate for selection and it can be selected
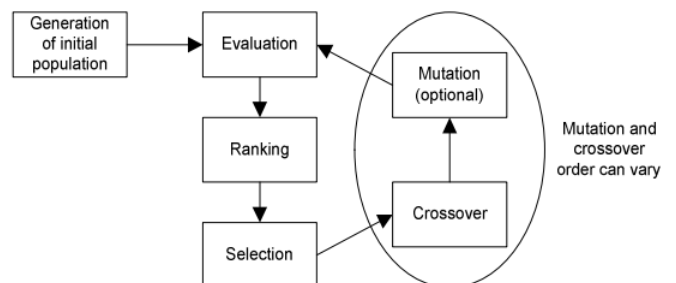

Figure 2. The structure of Genetic Algorithm [18]

Table 1. The description and value of fertility dataset features

Feature description	Values	Normalized
Season in which the analysis was performed	1) winter, 2) spring, 3) Summer, 4) fall	(-1, -0.33, 0.33, 1)
Age at the time of analysis	18-36	(0, 1)
Childish diseases	1) yes, 2) no	(0, 1)
Accident or serious trauma	1) yes, 2) no	(0, 1)
Surgical intervention	1) yes, 2) no	(0, 1)
High fevers in the last year	1) less than three months ago, 2) more than three months ago, 3) no	(-1, 0, 1)
Frequency of alcohol consumption	1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never	(0, 1)
Smoking habit	1) never, 2) occasional , 3) daily	(-1, 0, 1)
Number of hours spent sitting per day	0-16	(0, 1)

again in any other coming steps. Thus, a sample with replacement can represents multiple number of times in training data. It can compensate the lack of enough samples in minority class. The balanced class results better learning in classification using artificial neural network.

After resampling and constructing balanced classes, MLP should be applied to data to classify them. As mentioned before finding an optimal structure for MLP is an important topic. Optimized MLP with genetic algorithm tries to achieve appropriate structure of network. Optimized MLP is a simple algorithm for both learning rate and size adjustment of artificial neural networks during training. The algorithm applies genetic algorithms for optimizing the structure of MLP. It performs a number of networks that are trained with different rates and different numbers of hidden units to get the best choice.

The algorithm starts with some random chromosomes. Each chromosome represents an architecture for a neural network and the related learning rate. There are different approaches for encoding information about structure of ANN in the chromosome [20]. Since the used MLP has a single hidden layer, one parameter should be optimizd for the structure. A choromosome is composed of two parts: number of neurons of hidden layer and leraning rate.

The fitness function for the proposed algorithm is mean square error (MSE) as mentioned in the section 2. In each iteration, parents are selected from the population based on their fitness. Finally a search operator is applied to the parents for generating offspring which form the next generation. After all, the best architecture and learning rate is obtained by minimizing the MSE. The artificial neural network with this parameter is utilized for the training data set. Algorithm 16 represents a pseudocode of proposed method.

5. Experiments

5.1. Fertility dataset

The dataset [9] is obtained from the UCI Machine learning repository. 100 volunteers provide a semen sample analyzed according to the World Health Organization 2010 criteria [21]. Sperm concentration is related to socio-demographic data, environmental factors, health status, and life habits. The dataset includes two classes as semen quality: "normal"(N) and "altered"(O). There are 12 samples as altered and 88 as normal. It can be seen, the dataset is so unbalanced. Therefore, classification of altered data (as important data) is very hard. The

description of feaures in dataset is presented in Table 1.

5.2. Development of the method

For the experiments, the complete data set is divided into two subsets: training and test subsets. Because of the unbalancing of the data set, the minority class of training set including altered class (O) is upsampled using the bootstrap sampling method. So a balanced dataset is achieved and every class has 88 samples. 176 samples are as input vector for the first layer of MLP. The number of neurons in this layer is equal to the number of features that is 9 in this dataset. The last layers is responsible for specifying the result of classification so the number of neurons in this layer depends on the number of classes that is 2 in this case therefore one neuron is sufficient for classification. As stated before, genetic algorithms are applied for estimating the number of neurons in the hidden layer. The parameters that should be determined in experiments are the mutation rate equal to 0.1, the number of iterations and the number of generations of the algorithm that value 15 is considered for both of them.

After some iterations, the last structure of network is obtained for classification of the dataset. Thus the test subset is applied for evaluating the network. For a better generalization of network and real evaluation, the cross validation method is utilized. For this study 10-fold cross validation is applied.

In this method, model assessment is done for 10 times using test set. For this goal the whole dataset is divided into 10 subsets randomly. Every time the one part of the data is determined as test set and the remaining form training data. The errors of each model at every iteration are accumulated to provide the mean absolute test set error.

6. Results and disscusion

In this paper the performance of optimized MLP is compared with 4 different classiifiers include SVM, Decision Tree (DT) and Naive Bayesian based on accuracy and AUC and confusion matrix criteria. The defnition of these criteria is as bellows:

$$Accuracy = \frac{TP + TN}{P + N} \qquad (6)$$

True positives (TP): These refer to the positive tuples that were correctly labeled by the classier.
True negatives (TN): These are the negative tuples that were correctly labeled by the classier.

Table 2. The confusion matrix of optimized mlp, decision tree, svm, nave bayesian

	Optimized MLP			Decision Tree			SVM			Naive Bayesian		
	true O	true N	precision	true O	true N	precision	true O	true N	precision	true O	true N	precision
pred. O	85	6	93.41%	86	25	77.48%	80	24	76.92%	74	32	69.81%
pred. N	5	82	94.25%	4	63	94.03%	10	64	86.49%	16	56	77.78%
recall	94.44%	93.18%		95.56%	71.59%		88.89%	72.73%		82.22%	63.64%	

Algorithm 1: Pseudocode of Proposed Method.

1 **begin**
2   Partition the original samples into k equal sized subsamples.
3   **for** *i=1 to k* **do**
4
             Bootstrap sampling Phase
5   Consider ith part as test set and the remaining as training set.
6   Keep sampling the "altered" class of training data using bootstrap method to obtain balanced classes.

      MLP training phase with genetic algorithm

7   **repeat**
8   Decode each individual in the current generation into an architecture of MLP.
9   Train each MLP with the decoded architecture by a predefined learning rule.
10   Compute the fitness of each individual based on MSE function.
11   Select parents from the population based on their fitness.
12   Apply crossover and mutation operators and generate offspring which form the next generation.
13   **until** *termination criterion is satisfied*;
14   **end**
15   Average (or otherwise combine) The k results from the folds to produce a single estimation.
16 **end**

Table 3. Comparison with other balancing methods in terms of accuracy and auc

Method	Accuracy	AUC
Optimized MLP	**93.86%**	**0.933**
Decision Tree	83.82%	0.858
SVM	80.88%	0.882
Naive Bayesian	73.10%	0.81

the value of specificity, positive predictive value, and negative predictive of optimized ANN is the best in compared methods . Its sensitivity is approximately similar to DT. The value of accuracy and AUC of all algorithms are showed in Table 3. The best performance is presented by optimized MLP. The performance of SVM and DT in this paper are better than those are reached in previous works. The reason for this occurrence is attention to unbalancing data in this paper while previous research didn't consider this issue. For best generalization of the algorithm the cross k-fold validation is utilized. In the experiment, different k is employed. Table 4 shows the accuracy and AUC of the proposed method with different values of k. It is demonstrated that the performance of algorithm enhances when the value of k increases. It is clear that when k increases the number of training samples increase, therefore the better classfication accuracy is achieved. When k is 10 the data set is divided into 10 subsets and 90% of data is used for training and 10% for the test. In this condition the accuracy of optimized MLP is equal to 93.86%.

7. Conclution

In this paper we evaluated the optimized neural network for semen analysis in infertility diagnosis. Fertility dataset is an unbalanced dataset. Therefore, it was necessary to solve this problem before classifcation task. The bootstrap sampling is utilized to increase the number of minor class samples. Thus the number of altered samples in dataset reaches to normal samples as major class. For classification task the neural network combined

False positives (FP): These are the negative tuples that were incorrectly labeled as positive.
False negatives (FN): These are the positive tuples that were mislabeled as negative.

Accuracy is the recognition rate of a method. The accuracy measure works better when the data classes are fairly evenly distributed. Therefore, it is not enough for evaluating a method in unbalanced dataset such fertility dataset. In this case other measures are better suited. AUC is a good criterion in such cases. AUC means area under the receiver operating characteristic (ROC) curve. ROC cure is created by plotting the true positive rate (precision) against the false positive rate (recall). The maximum value of AUC is 1.

The confusion matrix is an another measure. This measure is a table that each element of it indicates one of the above terms (TP, TN, FP and FN).

Table 2 compares the confusion matrix of the proposed method and other classification algorihms. It can be seen that

Table 4. Performance of optimized mlp with defferent number of folds

Fold	accuracy	AUC
2	84.83%	0.784
4	90.52%	0.877
6	91.65%	0.913
8	93.26%	**0.941**
10	**93.86%**	0.933

with genetic algorithm is employed. Genetic algorithm tries to obtain the best number of neurons in hidden layer and learning rate for neural network by an opti- mization process. The appropriate structure of artificial neural network is obtained based on this method. The experiments are done on a real fertility diagnosis dataset. The method outperforms SVM, decision tree and nave Bayesian. The accuracy and AUC of method are 93.98% and 0.933 respectively that are higher than other methods.

8. References

[1] C. Splingart, C. Frapsauce, S. Veau, C. Barthelemy, D. RoyÃ¨re, and F. GuÃ©rif, "Semen variation in a population of fertile donors: evaluation in a French centre over a 34 year period," International journal of andrology, vol. 35, pp. 467-474, 2012.

[2] A. Giwercman and Y. L. Giwercman, "Environmental factors and testicular function," Best practice & research Clinical endocrinology & metabolism, vol. 25, pp. 391-402, 2011.

[3] A. C. Martini, R. I. Molina, D. EstofÃ¡n, D. Senestrari, M. F. de Cuneo, and R. D. Ruiz, "Effects of alcohol and cigarette consumption on human seminal quality," Fertility and sterility, vol. 82, pp. 374-377, 2004.

[4] I. Kononenko, I. Bratko, and M. Kukar, "Application of machine learning to medical diagnosis," Machine Learning and Data Mining: Methods and Applications, vol. 389, p. 408, 1997.

[5] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," Artificial Intelligence in medicine, vol. 23, pp. 89-109, 2001.

[6] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 42, pp. 463-484, 2012.

[7] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," International Journal of Emerging Technology and Advanced Engineering, vol. 2, pp. 42-47, 2012.

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

[9] D. Gil, J. L. Girela, J. De Juan, M. J. Gomez-Torres, and M. Johnsson, "Predicting seminal quality with artificial intelligence methods," Expert Systems with Applications, vol. 39, pp. 12564-12573, 2012.

[10] K. Polat and S. GÃ¼nes, "A novel hybrid intelligent method based on decision tree classifier and one against all approach for multiclass classification problems," Expert Systems with Applications, vol. 36, pp. 1587-1592, 2009.

[11] K. Mehrotra, C. K. Mohan, and S. Ranka, Elements of artificial neural networks: MIT press, 1997.

[12] B. D. Ripley, Pattern recognition and neural networks: Cambridge university press, 1996.

[13] D. Conforti and R. Guido, "Kernel based support vector machine via semidefinite programming: Application to medical diagnosis," Computers & Operations Research, vol. 37, pp. 1389-1394, 2010.

[14] V. N. Vapnik and S. Kotz, Estimation of dependences based on empirical data vol. 40: Springer-verlag New York, 1982.

[15] J. L. Girela, D. Gil, M. Johnsson, M. J. Gomez-Torres, and J. De Juan, "Semen parameters can be predicted from environmental factors and lifestyle using artificial intelligence methods," Biology of reproduction, vol. 88, p. 99, 2013.

[16] H. Wang, Q. Xu, and L. Zhou, "Seminal Quality Prediction Using Clustering-Based Decision Forests," Algorithms, vol. 7, pp. 405-417, 2014.

[17] D. Whitley, T. Starkweather, and C. Bogart, "Genetic algorithms and neural networks: Optimizing connections and connectivity," Parallel computing, vol. 14, pp. 347-361, 1990.

[18] A. Kopel, "NEURAL NETWORKS PERFORMANCE AND STRUCTURE OPTIMIZATION USING GENETIC ALGORITHMS," California Polytechnic State University, San Luis Obispo, 2012.

[19] S. Mirjalili, "How effective is the Grey Wolf optimizer in training multi-layer perceptrons," Applied Intelligence, vol. 43, pp. 150-161, 2015.

[20] X. Yao, "Evolving artificial neural networks," Proceedings of the IEEE, vol. 87, pp. 1423-1447, 1999.

[21] J. Lu, Y. Huang, and N. LÃ¼, "[WHO Laboratory Manual for the Examination and Processing of Human Semen: its applicability to andrology laboratories in China]," Zhonghua nan ke xue= National journal of andrology, vol. 16, pp. 867-871, 2010.