Intelligent Diagnosis of Asthma Using Machine Learning Algorithms

Taha Samad Soltani Heris¹, Mostafa Langarizadeh², Zahra Mahmoodvand³, Maryam Zolnoori⁴

1. PhD student of medical informatics, the College of Paramedics, Tehran University of Medical Science, Tehran, Iran
2. Assistant professor in medical informatics, the College of Paramedics, Tehran University of Medical Sciences, Tehran, Iran
3. M.Sc. student of health information technology, the College of Management and Medical Information, Tehran University of Medical Sciences, Tehran, Iran
4. Post-doctoral student in health informatics, the College of Informatics, the State University of Indiana, The United States of America

*Corresponding Author email: t-ssoltany@razi.tums.ac.ir

ABSTRACT: Data mining in healthcare is a very important field in diagnosis and in deeper understanding of medical data. Health data mining intends to solve real-world problems in diagnosing and treating diseases. One of the most important applications of data mining in the domain of machine learning is diagnosis, and this type of diagnosis of the disease asthma is a notable challenge due to the lack of sufficient knowledge of physicians concerning this disease and because of the complexity of asthma. The purpose of this research is the skillful diagnosis of asthma using efficient algorithms of machine learning. This study was conducted on a dataset consisting of 169 asthmatics and 85 non-asthmatics visiting the Imam Khomeini and Masseeh Daneshvari Hospitals of Tehran. The algorithms of k-nearest neighbors, random forest, and support vector machine, together with pre-processing and efficient training were implemented on this dataset, and the degrees of accuracy and specificity of the system used in our study were calculated compared with each other and with those of previous research. From among the different values for neighborhood, the highest degree of specificity was achieved with five neighbors. Our method was investigated together with other methods of machine learning and similar research, and the ROC curve was plotted, too. Other methods achieved suitable results as well, and they can be relied on. Therefore, we propose our approach based on the k-nearest algorithm together with pre-processing based on the Relief-F strategy and the Cross Fold data sampling as an efficient method in artificial intelligence with the purpose of data mining for the classification and differential diagnosis of diseases.

Keywords: asthma, Data mining, Diagnosis, intelligent systems, machine learning.

INTRODUCTION

Data mining is an important step in discovering and extracting knowledge and implies exploring large datasets with the purpose of extracting unknown patterns among (Han 2005). Since discovering relations among data is very difficult in traditional statistical methods (Lee, Liao et al. 2000), this is not a new approach and it is widely in financial institutions in areas such as recognizing fraud and pricing. Salespersons and retailers use this approach in dividing their markets and in distributing their products among warehouses, and manufacturers employ it for controlling quality and for scheduling maintenance and repairs (Koh and Tan 2011). Data mining applications have rapidly spread in large sections of organizations offering medical care, in financial predictions, and in weather forecasting (Obenshain 2004). There is a high potential for data mining applications in healthcare. These applications can be grouped, in general, into evaluation of the effects of treatments, management of healthcare, management of relations with patients, and recognition of fraud and misuse. More specialized data mining includes medical diagnosis and analysis of DNA micro-arrays (Koh and Tan 2011).
Data mining is a very important field in diagnosis and in deeper understanding of medical data. Health data mining aims to solve real-world problems in diagnosis and in disease treatment (Liao and Lee 2002); and also, in healthcare, it is also considered to be an important area for research in predicting diseases and in gaining a more profound understanding of health data. The purpose of health data mining is the resolution of real-world health problems in diagnosis and disease treatment (Liao and Lee 2002). Researchers use this method to diagnose different diseases and, to this end, employ various approaches and algorithms with different specificity and accuracy (Shouman, Turner et al. 2012).

Methods and algorithms of data mining have been used in diagnosing diseases. The most important of these studies include the automatic diagnosis of the skin disease erythema-squamous in which these methods and algorithms are employed (Güvenir, Demiröz et al. 1998; Übeyli and Güler 2005; Übeyli and Doğdu 2010; Xie and Wang 2011; Abdi and Gíveki 2013). Data classification is carried out by using a variety of methods including k-nearest neighbor, support vector machine, and random forest, which are popular methods for multi-class diagnosis in the area of pattern recognition (Athitsos 2005). These are the most commonly used methods in data mining and classification (Moreno-Seco, Micó et al. 2003). In this article, we use the above-mentioned algorithms for the diagnosis of the disease asthma in order to achieve a suitable efficiency in its diagnosis.

**METHODS AND MATERIALS**

The process of solving the problem was modeled in steps to obtain the desired output. In these steps, which are shown in Figure 1, the processes of organizing the input, pre-processing, data sampling, adjusting the parameters of the algorithms, executing the algorithms, and analyzing the outputs were carried out.

**Organizing input data**

In this research, the dataset regarding patients visiting the Imam Khomeini and the MasseehDaneshvari Hospitals of Tehran was used (Zolnoori, Zarandi et al. 2007). The dataset contains 250 records consisting of 24 input features and one discrete output feature shown in Table 1. Each of these features has a range of numerical values. For example, coughing, depending on its severity, may have values from zero to 50. Output classes also include the two classes of zero and one, which denote those diagnosed with and those not diagnosed with asthma.

<table>
<thead>
<tr>
<th>Asthma disease (number of patients)</th>
<th>Features</th>
</tr>
</thead>
</table>

**Pre-processing**

The pre-processing of data is an important step in the process of data mining. In data mining, the methods of data collection are not very controlled and outlier, repetitive, and wrong values may result in obtaining invalid output. Therefore, it is necessary to take a set of actions before utilizing data to ensure good quality (Pyle 1999). If the information used in the processes of extracting knowledge and educating contains redundant or irrelevant data, it will be very difficult to carry out these processes. The pre-processing step may be time-consuming and include processes such as the cleaning, normalization, conversion, and extraction of features. The output of this step will be the training set (Kotsiantis, Kanellopoulos et al. 2006). In this study, the pre-processing of data was conducted in two steps. In the first step, incomplete data was eliminated from the dataset in order to achieve better agreement.
in later stages. The second step consisted of the selection of the features based on the strategy and algorithm of Relief–F. In the strategy of Relief–F, which is one of the most successful methods of selecting features (Liu and Motoda 2008), samples are randomly chosen, and the weight of each feature is calculated based on the nearest neighbor. After experimenting with the different values of the features and calculating the outputs, 15 superior features were selected and applied as the input for the algorithm.

Data sampling
In the data-sampling step, a subset of data is extracted from the input dataset. This is performed for sampling and for dividing the data into the two classes of training and testing data. In this research, the cross validation method was used. This method is employed for estimating common errors based on numerous sampling, and the result of such estimation is usually used for selecting from among various models including network architecture. In the k-fold CV method, the data and the samples are divided into k almost equal subsets. The network is then tested k times and each time one of the subsets is omitted (Sarle 2012). In our research, we divided the data and the samples into 10 subsets and calculated the accuracy and sensitivity of the outputs to make the necessary. After investigating the best accuracies and sensitivities, 26 records were selected for modeling and 228 records remained.

Adjusting the algorithm
In this step, the parameters of the algorithms are tested and evaluated by using different values in order to find the best possible result in the system output. The values of neighborhood, the resolution type, and the number of trees are the parameters in the k – nearest neighbor, the support vector machine, and the random forest algorithms, respectively.

Executing the algorithm
The k – nearest neighbor algorithm
The KNN algorithm, also called the memory – based classification method, is one of easiest and most direct techniques of data mining (Alpaydin 1997). In this method, differences between features are calculated by using Euclidean distances in order to work with continuous and discrete features. For example, if the first sample is \((a_1, a_2, a_3, \ldots, a_n)\) and the second sample \((b_1, b_2, b_3, \ldots, b_n)\), then the distance between them is calculated using formula 1:
\[
\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2} 
\] (1)

The most difficult problem in employing the formula of Euclidean distances is that larger values often flood smaller ones. The KNN algorithm usually works with continuous data but it is efficient for discrete data as well. If the discrete values in the hypothetical samples \(a_i\) and \(b_i\) differ from each other, this difference will be considered one; and if they do not differ from each other, the difference will be considered zero (Shouman, Turner et al. 2012). The algorithm was run at different neighborhood values until the highest levels of sensitivity and specificity were achieved (which will be referred to in the section on results).

The support- vector machine (SVM) algorithm
Vapnik et al. (Cortes and Vapnik 1995) introduced the support - vector machine (SVM) algorithm as a very efficient method for the purpose of pattern recognition. The SVM is a powerful tool for data classification. This algorithm allocates semi – discrete spaces, together with the space of the original data, to two sets of data in linear classification and in high- level dimensions in non-linear classification (Fung and Mangasarian 2005). In classification, data is usually separated into the two sets of training and testing data. Each sample in the training set includes the target value (output). The purpose in SVM is to produce a model based on training data in such a way that the model can predict the target values of the training set (Hsu, Chang et al. 2003). A support- vector machine produces a hyperplane, or a set of hyperplanes, in a high – dimensional space. These hyperplanes can be used for classification, regression or for other applications. Good separations are performed by hyperplanes that have the longest distances to the nearest points of the training data in each class (Deepa and Thangaraj 2011).

The random forest algorithm
The random forest (s) algorithm is a general classification method consisting of a large number of decision trees and class outputs, each output being implemented by specific trees. Leo Breiman and Adele Cutler introduced the random - forest deduction algorithm (Ho 1995). Random forests are combinations of computational trees each of which is related to values of random vectors that are independently sampled; and the same
distribution is applied for all the trees in the forest. This classification algorithm consists of a set of tree classifications which has the form of \{h(x,\theta,k = 1}\}, in which \{\theta \}

Evaluating the output

Output evaluation was performed on the basis of sensitivity analysis, accuracy, specificity, the plotting of the confusion matrix, and the plotting of the ROC curve (Zhu, Zeng et al. 2010). Relations 2, 3, and 4 were used to calculate sensitivity, specificity, and accuracy.

Sensitivity is the ratio of correct positive diagnoses that are correctly revealed by a specific test. Based on this definition, one hundred percent sensitivity implies correct diagnoses of the disease in all patients:

\[ \text{Sensitivity} = \frac{\text{The number of correct positive diagnoses}}{\text{The number of correct positive diagnoses} + \text{the number of correct negative diagnoses}} \]  

(2)

The statistical measure determining the accuracy of the test for the purpose of identifying the negative diagnoses is called specificity.

\[ \text{Accuracy} = \frac{\text{The number of correct positive diagnoses}}{\text{The number of correct positive diagnoses} + \text{the number of false positive diagnoses} + \text{the number of correct negative diagnoses}} \]  

(3)

Finally, the following formula is used to calculate sensitivity:

\[ \text{Accuracy} = \frac{\text{The number of correct positive diagnoses} + \text{the number of correct negative diagnoses}}{\text{The number of false negative diagnoses} + \text{the number of correct positive diagnoses} + \text{the number of false positive diagnoses} + \text{the number of correct negative diagnoses}} \]  

(4)

RESULTS AND DISCUSSION

This research was conducted in the area of clinical diagnosis in the Canvas Orange data mining environment and implemented in the Python language. The input features for the application program prepared are listed in Table 1. The suitable performance of the algorithms depends on the sizes of the training and testing datasets. A set of identical samples of the same size was employed for evaluating and comparing the implemented algorithms.

We obtained an optimal output for training the SVM from the RBF kernel functions by adjusting the cost parameters and the numerical accuracy of 2, 001, and zero, respectively. Euclidean distances were used in the nearest – neighbor algorithm, weighting was performed based on these distances, and the continuous values were normalized. In the random forest method, twenty trees were used.

Table 2 in the section on results shows sensitivity, specificity, and accuracy of the results of the algorithms in the diagnosis of asthma. The results of our study were compared with the outputs of the fuzzy expert system designed by Zarandi et al. based on this dataset. Note that every time the program is run these values undergo change because the samples had been randomly taken. These values change and so we stopped the iteration process after obtaining 100 percent specificity, sensitivity, and accuracy from one of the methods.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Description</th>
<th>Class</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>K=5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SVM</td>
<td>(\exp(-g</td>
<td>x-y</td>
<td>^2))</td>
<td>1</td>
<td>0.9870</td>
</tr>
<tr>
<td>RF</td>
<td>Number of trees=20</td>
<td>1</td>
<td>0.9652</td>
<td>0.9868</td>
<td>0.9211</td>
</tr>
<tr>
<td>Fuzzy</td>
<td>Fuzzy Expert system(Zarandi, Zolnoori et al. 2010)</td>
<td>1</td>
<td>-</td>
<td>0.94</td>
<td>1</td>
</tr>
<tr>
<td>Decision Rule</td>
<td>Computer-Assisted, Symptom-Based Diagnosis(Choi, Yoo et al. 2007)</td>
<td>1</td>
<td>-</td>
<td>0.562</td>
<td>0.935</td>
</tr>
<tr>
<td>Decision Rule</td>
<td>CAIDSA(Chakraborty, Mitra et al. 2009)</td>
<td>1</td>
<td>0.90</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In the nearest – neighbor method, the best results were obtained when there were five neighbors, because this k value corresponds to the specificity, accuracy, and sensitivity values of one. In the SVM method, the best results were achieved, i.e., specificity, sensitivity, and accuracy were 0.9934, 0.9737, and 0.9870, respectively,
when the radial basis function, which is a non-linear method, was used (Chang, Hsieh et al. 2010). In the random forest method, in which the number of trees was 20, the values for specificity, sensitivity, and accuracy were 0.9868, 0.9211, and 0.9652, respectively. For all values, both output classes of zero and one, which stand for people diagnosed and not diagnosed with asthma, were calculated. Of course, class 1 has more applications. The confusion matrices for all three algorithms are shown in Figure 1. They indicate the numbers of false and correct positive diagnoses, and the numbers of false and correct negative diagnoses.

The ROC curve was plotted to reveal and evaluate the efficiency of the system. The nearer the points are to the top left part of the curve, the more suitable they will be and the closer the prediction model will be to its ideal state (Metz 1978). The curves for all three algorithms are shown in Figure 2.

The results obtained from our research, the findings of studies carried out in which the dataset we used was employed, and the conclusions drawn in similar research on diagnosing asthma are compared, and important points are revealed and listed, in Table 2.

In 2010, FazelZarandi et al. introduced a fuzzy method for diagnosing the disease asthma in the dataset used in our study (29). The general specificity and sensitivity of the classification of their model were 100 and 94 percent, respectively. The algorithms we used in our study yielded better results than the fuzzy method with respect to the sensitivity of the diagnosis.

In another research titled “Easy diagnosis of asthma: computer – assisted, symptom – based diagnosis”, the specificity and sensitivity obtained were 93.5 and 56.2 percent, respectively (30). This study was conducted using questionnaires and the general diagnosis of the disease was offered by calculating the sum of the scores given. The evaluation indices in our research are much more efficient and are better, compared to those employed in this study.
Chakraborty et al. (2009) also proposed a system for the intelligent diagnosis of asthma with an accuracy of 90.03 percent. In this research, a neural network and decision rules were used (31). This system has a weaker performance than ours (in which we obtained an accuracy of 96.52 to 100 percent).

In this study, three approaches of the K – nearest neighbor, the random forest, and the support – vector machine algorithms together with pre – processing based on the Relief – F strategy and the Cross Fold data sampling are proposed for data mining the dataset on asthma for diagnosing this disease.

The highest efficiency for the k-nearest neighbor algorithm was achieved when there were five neighbors, because in this situation all cases of the disease were correctly diagnosed. The other algorithms also performed better than the study based on the fuzzy expert system. Efficient pre – processing omits incomplete and problematic data, discretizes values through the use of discretization techniques, adopts a suitable training algorithm , selects and estimates appropriate algorithm parameters, and , thereby, greatly affects the output and performance of classification methods.

REFERENCES


Moghazy A, Gargour A. et al.2012. Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. 2012 International Conference on Knowledge Discovery (ICKD 2012) IPCSIT.


Zolnoori M, Zarandi M.HF. et al.2007. Designing fuzzy expert system for diagnosing and evaluating childhood asthma. Department of information technology management, Tarbiat Modares University, Master of Science in Information Technology Management: 199.