Heart disease prediction using neighborhood component analysis and support vector machines

No Author Given

No Institute Given

Abstract. Nowadays, one of the main reasons for disability and mortality premature in the world is the heart disease, which make its prediction is a critical challenge in the area of healthcare systems. In this paper, we propose a heart disease prediction system based on Neighborhood Component Analysis (NCA) and Support Vector Machine (SVM). In fact, NCA is used for selecting the most relevant parameters to make a good decision. This can seriously reduce the time, materials, and labor to get the final decision while increasing the prediction performance. Besides, the binary SVM is used for predicting the selected parameters in order to identify the presence/absence of heart disease. The conducted experiments on real heart disease dataset show that the proposed system achieved 85.43% of prediction accuracy. This performance is 1.99% higher than the accuracy obtained with the whole parameters. Also, the proposed system outperforms the state-of-the-art heart disease prediction.

Keywords: Heart disease · prediction · Neighborhood Component Analysis · Support Vector Machines · feature selection.

1 Introduction

Heart disease is one of the main reasons for disability and premature death of people in the world. According to the World Health Organization an estimation of 17.9 million deaths have occurred worldwide due to Heart diseases in 2016 [2]. However, some key factors help us reduce the risk of heart disease as the blood pressure and cholesterol controlling [13]. In fact, heart disease should not be diagnosed when a heart attack, angina, stroke or heart failure occur. Therefore, the prediction of heart disease is a delicate, risky, and very important factor [4]. If done properly it can be used by the medical staff to save life. This process can be realized by exploring the registered patient data. Usually, the existing healthcare systems use electronic health records to store this data. Advances in computer and information technologies can deal with this routine data to make critical medical decisions [6].

Actually, machine learning techniques have been widely explored in healthcare systems to make decisions building on clinical data. In this context, many researchers have used these techniques on heart diseases prediction. In [3], the authors developed two systems based on Artificial Neural Network (ANN) and

Neuro-Fuzzy approaches in order to develop an automatic heart disease diagnosis system. According to Shafenoor et al. [4], a proper evaluation and comparison to test the different data mining techniques that can improve the accuracy of predicting cardiovascular disease. In [6], a dynamical ECG recognition framework for human identification and cardiovascular diseases classification based on radial basis function (RBF) has been explored. In [5] many machine learning techniques, such as: Baysian Network, Decision tree, Artificial Neural Network, Fuzzy pattern tree and Support Vector Machine (SVM), have been used to classify the Cleavland heart disease data set using 10-fold-cross validation. SVM achieved the highest prediction accuracy compared to other classifiers. Otoom et al. [10] have presented a system for Coronary artery disease detection and monitoring where three machine learning techniques are performed such as: Bayes Net, SVM, and Functional Trees. The authors have used WEKA tool for detection and by applying feature selection, SVM has provided the best accuracy with 85.1% correctly.

Practically, in heart disease prediction, there is a huge number of samples corresponding to several installed sensors and/or physico-chemical analysis to identify the presence/absence of the disease. Because of the increasing number of patients day after day, the preparation of these samples becomes very costly in time, materials, and labor. To address this problem, a selection of relevant parameters can efficiently reduce the data dimension and minimize the number of treatment conducted which minimize consequently the medical equipment and the corresponding health staff.

In this paper, we propose a heart disease prediction system based on Neighborhood Component Analysis (NCA) for selecting the most relevant parameters to make a good decision. In fact, the parameter selection in such a medical application can seriously reduce: (1) the number of medical equipments, (2) the labor required, and (3) the computing time to get the final decision while increasing the prediction performance. Besides, the binary support vector machine (SVM) is used for predicting the selected parameters in order to identify the presence/absence of heart disease. In fact, SVM can perfectly classify data in binary problem by finding the optimal hyper-plane that separates the data points of first class from those of second class. This makes it very adequate for heart disease prediction system which contains data of patient records with binary target, i.e. referring to the presence or absence of heart disease.

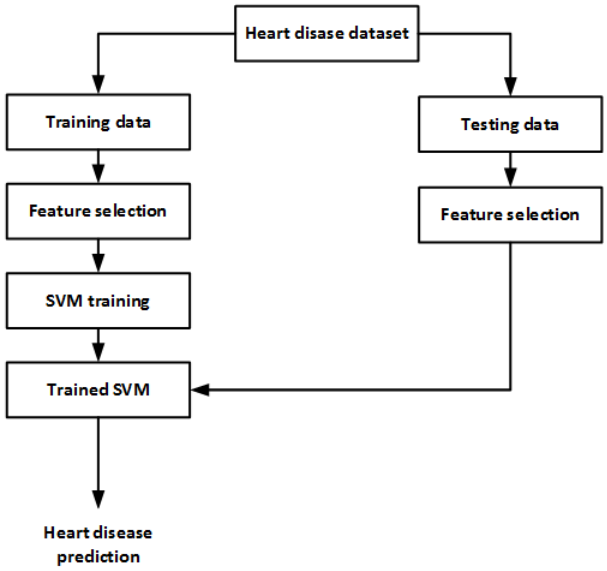
The rest of the paper is organized as follows: Section 2 presents an overview of the proposed system including data description, NCA, and SVM. In section 3, the experimental results are presented with details. In Section 4, a comparison of the state-of-the-art methods is given. Finally, the conclusions drawn from this work are in Section 5.

2 Proposed System

The proposed heart disease prediction system is compound of several stages that work together to achieve the desired results. The first stage consists of dividing

equally the used data into two subsets, one for training and the rest for testing. The second task consists in applying the NCA feature selection algorithm on the training subset in order select the most relevant parameters. Then, SVM classifier performs the training phase with the best parameters. Finally, the trained SVM model can easily predict the testing samples. Figure 1 illustrates the overall flowchart of the proposed system.

Fig. 1. The proposed system flowchart.



2.1 Dataset Description

The performance evaluation of this work is conducted on the famous cardiovascular dataset called the Heart UCI disease (University of California, Irvine, CA, USA). It has been collected from UCI machine learning repository [1]. This dataset contains in total 303 patient records with 76 attributes for each one, but only 14 of them are used for our evaluation to make our scores comparable to previous works. Table 1 provides a brief description about the selected attributes and their proprieties. The last attribute serves as the prediction target that indicates the absence or presence of heart disease in a patient (0 or 1 value, respectively). Of the 303 records, 165 records with Target 1 and the rest for patients with Target 0. To build this dataset, patients from age between 29 and 79 have been selected. Male patients are represented by 1 and female ones are represented by 0. Also, four types of chest pain 'Cp' have been considered such as: a typical type Angina, atypical type Angina, non-angina pain,

Table 1. Dataset parameter description

No.	Attributes Name	Type	Description	Range of values
1	(age)	Continuous	patient age in years	29 to 79
2	(sex)	Discrete	patient gender	0=female,1=male
3	(cp)	Discrete	Chest pain type	1,2,3,4
4	(trestbps)	Continuous	Resting blood pressure (mm/Hg)	94 to 200
5	(chol)	Continuous	Serum cholesterol (mg/dl)	126 to 564
6	(fbs)	Discrete	fasting blood sugar (mg/dL)	0,1
7	(restecg)	Discrete	resting Electrocardiograph results	0,1,2
8	(thalach)	Continuous	Maximum Heart Rate Achieved	71 to 202
9	(exang)	Discrete	Exercise induced angina	0,1
10	(old peak)	Continuous	ST depression induced by exercise relative to rest	1 to 3
11	(slope)	Discrete	The slope of peak Exercise ST segment	1,2,3
12	(ca)	Discrete	Number of major vessels colored by fluoroscopy	0 to 3
13	(thal)	Discrete	Represents heart rate of the patient	3,6,7
14	(num)	Discrete	Presence or absence of heart disease	0,1

and the Asymptomatic types. Each type of them is described by a value from 1 to 4, respectively. The next attribute '*trestbps*' is the resting blood pressure measured at the hospital admission. '*Chol*' is the blood cholesterol level. The parameter '*Fbs*' is the fasting blood sugar levels which is represented by 0 if the fasting blood sugar is less than 120 mg/dl and 1 if it is higher. '*Restecg*' is the resting electrocardiographic results classified in three levels 0, 1, and 2. Besides, the attribute '*thalach*' describes the maximum heart rate achieved. '*exang*' is the Exercise induced angina which which is recorded as 1 if there is pain and 0 else. The attribute '*oldpeak*' represents the ST depression induced by exercise relative to rest. Furthermre, the slope of the peak exercise ST segment has been recorded with the values 0, 1, and 2. The attribute '*ca*' is the number of major vessels colored by fluoroscopy varying from 0 to 4. Finally, the attribute '*thal*' represents the nature of defect which can take an integer value from 0 to 3.

2.2 Feature selection with NCA

NCA is a non-parametric and embedded method, based on the k-Nearest neighbor (KNN) algorithm [7], that handles the tasks of dimensionality reduction in a unified manner. NCA is used to learn feature weighting vector by maximizing the Leave-One-Out (LOO) accuracy of classification with an optimized regulation parameter [16] and [18].

Let $T = (x_1, y_1), ..., (x_i, y_i), ..., (x_N, y_N)$ the set of training samples, where x_i is a feature vector of dimension $d, y \in 1, ..., H$ is its corresponding class label and N is a number of samples. We denote the weighted distance between two samples x_i and x_j by:

$$d_W(x_i, x_j) = \sum_{r=1}^{d} W_r^2 \, |x_{ir} - x_{jr}| \tag{1}$$

where w_r is a weight associated with r^{th} feature.

However, due to the non differentiable function of LOO accuracy that identify a classification reference point by the nearest neighbor, the reference point can be determined by a probability of x_i selects x_j as its reference point as follows:

$$p_{ij} = \{ \quad \frac{K(d_W(x_i, x_j))}{\sum_{k \neq i} K(d_W(x_i, x_j))}, if \ i \neq j \ and \ 0, if \ i = j, \tag{2}$$

where $K(.)$ is a kernel function. Hence, the probability of x_i being accurately classified in the correct label is given in Eqs. 3 and 4:

$$p_i = \sum_j y_{ij} p_{ij} \tag{3}$$

$$H_i = \{j \mid y_j = y_i\} \tag{4}$$

where $y_{ij} = \{ \quad 1, if \ y_i = y_j \ and \ 0, otherwise.$

The final optimization criterion $f(A)$ can then be defined by:

$$f(A) = \sum_i p_i. \tag{5}$$

In order to optimize the matrix A, we use the gradient rule as follows:

$$\frac{\partial f}{\partial W} = 2W \sum_i \left(p_i \sum_k p_{ik} x_{ik} x_{ik}^T - \sum_{j \in H_i} p_{ik} x_{ik} x_{ik}^T \right) \tag{6}$$

where x_{ij} designs the short-hand for $x_i - x_j$. This function can be optimized using the gradient methods.

2.3 Prediction with SVM

The SVM technique, which was initially proposed by Vapnik [17], has been widely used for classification and prediction [14]. These technique is based on a set of powerful learning methods that perform statistical learning theory [17]. Firstly, SVMs were designed to solve binary classification. They can function for multiclass classification problem by combining several binary SVM classifiers for each pair of classes. Furthermore, SVM can be adapted to work as a nonlinear classifier by using nonlinear kernels. The SVM basic form which classifies an input vector $x \in \Re^n$ is defined as:

$$f(x) = w\phi(x) + b \tag{7}$$

where w and b are two parameters that have to be estimated from inputs. $\phi(x)$ is the non-linear mapping function in the feature space. Unlike other classifiers

which are based on the empirical error for minimizing the classification errors on training dataset, the SVM is based on the structural risk minimization for finding a compromise between the confidence measure corresponding to generalization and the empirical error. Consequently, The goal is to find the parameters w and b such that $f(x)$ can be determined by optimizing the following constrained problem:

$$min_{w,b,\xi_i} \qquad \frac{1}{2}w^2 + C\frac{1}{N}\sum_{i=1}^{N}\xi_i \tag{8}$$

subject to

$$y_i(w.\phi(x_i) + b_i) \leq 1 - \xi_i \tag{9}$$

$$\xi_i \geq 1, for\ i = 1, ..., N \tag{10}$$

where $x \in \Re^n$, $i = 1, , N$ represents the training vectors with the corresponding class labels, and the parameter C controls the smoothness between the two measures. In our case, SVM output identify the presence/absence of the heart disease. The solution of the optimization problem can give the following:

$$w = \sum_{i=1}^{N}\alpha_i y_i \phi(x_i), \tag{11}$$

$$b = \sum_{j=1}^{N}\alpha_j y_j \phi(x_i)\phi(x_j) + y_i,\ \forall i \tag{12}$$

where α_i, $i = 1, , N$ are the multipliers of Lagrange corresponding to the training input x_i. Here, the support vectors are the training vectors with non-zero coefficients α_i which they contribute to determine the parameter w. The computation in input space can be mapped in another feature space using kernel function as follows:

$$K(x_i, x_j) = \phi(x_i)\phi(x_j) \tag{13}$$

It is worth noticing that any function that satisfies Mercers theorem [12] can be used as a kernel function. Finally, the estimating function can be expressed as:

$$f(x) = \sum_{j=1}^{N}\alpha_j y_j K(x_i, x_j) + b \tag{14}$$

The SVM classifier can use some kernel functions to mapping an original input space into a higher dimensional feature space, where the optimal hyperplane is determined to maximize the generalization ability of the classifier through some mapping functions chosen a priori. Table **??** presents four types of SVM kernels that we use in this work including, linear, polynomial, radial basis function (RBF), and sigmoid kernel.

Fig. 2. The SVM principle.

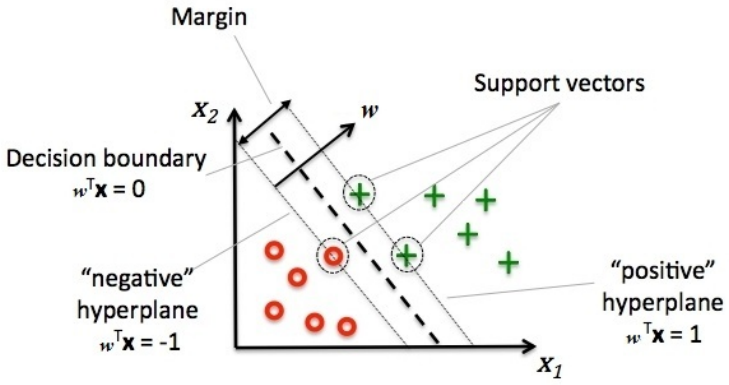


Table 2. Kernel functions used in SVM training.

Kernel name	**Mathematical function**
Linear	$K(x_i, x_j) = x(i)x(j)$
Polynomial	$K(x_i, x_j) = (\gamma x(i)x(j) + C)^d$
RBF	$K(x_i, x_j) = \exp(-\gamma \lvert x(i)x(j)\rvert^2)$
Sigmoid	$K(x_i, x_j) = \tanh(\gamma x(i)x(j) + C)$

3 Experimental results

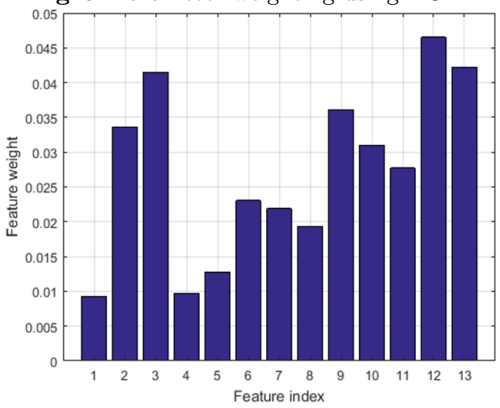
In this section, we carried out several experiments to show the effectiveness of the proposed method. Firstly, SVM classifier is evaluated with four different kernels above-mentioned in Table 2 for selecting the best one among them. Then, we perform NCA algorithm on the training subset in order to reduce the number of parameters required for the treatment. It should be noted that the materials used to perform our experiments are as follows, we have used an Intel Core TM i5-4300U cores and 1.9-GHz CPU processor with 8-GB RAM memory.

3.1 Features selection results

The parameter selection consists in reordering them according to theirs weights W_i obtained from NCA algorithm. This order describes the feature importance to the heart disease, such that the most relevant parameter takes the higher weight. Figure 3 shows the weight computed by NCA for each parameter and Table 3 reports the new order of parameters according to their relevance.

As we can see in Figure. 3, the parameter 'ca' has the higher weight value compared to others. It's worth saying that this parameter is the most representative in heart disease diagnosis. Also, the parameters '*cp*', '*thal*', and '*exang*' have a considerable importance for the treatment.

Fig. 3. Parameter weighting using NCA.



3.2 Prediction results

Now, to check how many parameters can be involved in the disease prediction while keeping a higher performance, we evaluated the SVM classifier with different input combinations according to the NCA ordering. We recall that our evaluation scenario consists of adding at each experiment a parameter from the higher weight to the least. Finally, the considered parameter combination is one that provides a performance greater than or equal to prediction system accuracy with all parameters. Table 4 reports the obtained performance with different experiments.

It is worth noticing that the the training accuracy (atr), training time (ttr), testing accuracy (ats), and testing time (tts) have been considered as performance assessment criteria. Besides, the parameter adjustments of SVM and its kernels are set by trial and error.

For the prediction results without feature selection, we found that the best training accuracy is achieved by SVM-RBF with 97,35 %. While the SVM-linear provides the best testing accuracy with 84.10%. Concerning the execution time, SVM-RBF is the fastest classifier in the training and testing phases compared to other SVM variants.

Now, by applying the NCA feature selection, we note that the prediction performance has been significantly improved with all SVM variants. For example, NCA-SVM-RBF achieved 79.74% in term of testing accuracy using only the first five parameters the most relevant. This performance is 3.98% higher than the accuracy obtained with the whole parameters. However, the best performance is obtained with the combination NCA-SVM-Polynomial by reducing three parameters, namely '*trestbps* ', '*chol*', and '*age*'. This combination obviously outperforms SVM-Polynomial without feature selection by 1.99% testing accuracy margin and more than two times fast.

Table 3. Parameter ordering before and after NCA application.

Parameter order before NCA	**Parameter order after NCA**
1: age	12: ca
2: sex	3: cp
3: cp	13: thal
4:trestbps	9: exang
5:chol	2:sex
6:fbs	10:oldpeak
7:restecg	11: Slope
8:thalach	6:fbs
9:exang	7: restecg
10: oldpeak	8:thalach
11: Slope	1:age
12: ca	4:trestbps
13: thal	5:chol

Table 4. The obtained performance with and without feature selection.

Variants	**# of parameters**	**atr (%)**	**ttr (ms)**	**ats (%)**	**tts (ms)**
SVM-Linear	13	88.07	15	84.1	1.3
NCA-SVM-Linear	11	86.09	4	84.76	0.1
SVM-Polynomial	13	86.75	27	83.44	5
NCA-SVM-Polynomial	10	86.09	9	**85.43**	2
SVM-RBF	13	97.35	7	75.49	2
NCA-SVM-RBF	5	92.05	3	79.47	0.1
SVM-Sigmoid	13	53.64	23	55.62	3
NCA-SVM-Sigmoid	2	72.84	3	77.48	0.1

4 comparison of the state-of-the-art methods

In order to give an idea on where our heart disease prediction system ranks performance-wise, we compare with works that used the same experimental protocol, the same performance measures, and the same datasets.Also, we note that the execution time dont be considered in the comparison due to the lack of this information in the works that we compare with. The heart disease prediction accuracy computed in our proposed system as well as in other similar works have been reported in Table 5.

From Table 5, it is clear that the NCA-SVM-based heart disease prediction outperforms the similar state-of-the-art systems. The advantages of our proposed system can be summarized as follows: (1) it reduces significantly the number of required parameters by increasing the importance of parameter that help in heart disease prediction and reducing the weight of those which do not. (2)

Table 5. Comparison of state-of-the-art methods.

Authors	**Method**	**Accuracy (%)**
Shouman et al. [15]	Decision tree	81.4
Peter et al. [11]	Multilayer perceptron	82.22
Nahar et al. [9]	Naive Bayes	69.11
Ismaeel et al. [8]	Extreme learning machine	80
Amin et al. [4]	Logistic regression	78.03
Our work	NCA-SVM	**85.43**

Our proposed system can reduce the decision-making time by eliminating the irrelevant samples as well as the material required. (3) SVM can perfectly classify the heart disease data by finding the optimal hyper-plane that separates the first class from those of second class. This makes it very adequate for heart disease prediction system which contains data of patient records with binary target.

5 Conclusion

In this work, the combination of NCA and SVM is used to predict heart disease from real datasets provided by the UCI machine learning repository. The NCA feature selection can effectively select the most relevant parameters to make a good decision. Thus, the parameter selection in such a clinical application can seriously reduce the number of medical equipment as well as the labor and the time required to get the final decision while increasing the prediction performance. However, SVM was used for predicting the selected parameters in order to identify the presence/absence of heart disease. The obtained results showed that the NCA-SVM improve considerably the prediction system performance. Applying this model can have a direct impact and economic savings on the design and development of heart disease prediction systems in healthcare.

References

1. Uci data homepage, https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. World health organization (who), http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
3. Abushariah, M.A., Alqudah, A.A., Adwan, O.Y., Yousef, R.M.: Automatic heart disease diagnosis system based on artificial neural network (ann) and adaptive neuro-fuzzy inference systems (anfis) approaches. Journal of software engineering and applications **7**(12), 1055 (2014)
4. Amin, M.S., Chiam, Y.K., Varathan, K.D.: Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics **36**, 82–93 (2019)
5. Bouali, H., Akaichi, J.: Comparative study of different classification techniques: heart disease use case. In: 2014 13th International Conference on Machine Learning and Applications. pp. 482–486. IEEE (2014)

6. Deng, M., Wang, C., Tang, M., Zheng, T.: Extracting cardiac dynamics within ecg signal for human identification and cardiovascular diseases classification. Neural Networks **100**, 70–83 (2018)
7. Goldberger, J., Hinton, G.E., Roweis, S.T., Salakhutdinov, R.R.: Neighbourhood components analysis. In: Advances in neural information processing systems. pp. 513–520 (2005)
8. Ismaeel, S., Miri, A., Chourishi, D.: Using the extreme learning machine technique for heart disease diagnosis. In: International Humanitarian Technology Conference (IHTC2015). pp. 1–3. IEEE (2015)
9. Nahar, J., Imam, T., Tickle, K.S., Chen, Y.P.P.: Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications **40**(1), 96–104 (2013)
10. Otoom, A.F., Abdallah, E.E., Kilani, Y., Kefaye, A., Ashour, M.: Effective diagnosis and monitoring of heart disease. International Journal of Software Engineering and Its Applications **9**(1), 143–156 (2015)
11. Peter, T.J., Somasundaram, K.: An empirical study on prediction of heart disease using classification data mining techniques. In: International conference on advances in engineering, science and management (ICAESM-2012). pp. 514–518. IEEE (2012)
12. Radhika, Y., Shashi, M.: Atmospheric temperature prediction using support vector machines. International journal of computer theory and engineering **1**(1), 55 (2009)
13. Raza, K.: Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In: U-Healthcare Monitoring Systems, pp. 179–196. Elsevier (2019)
14. Schölkopf, B., Smola, A.J., Bach, F., et al.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2002)
15. Shouman, M., Turner, T., Stocker, R.: Using decision tree for diagnosing heart disease patients. In: 9th Australasian Data Mining Conference. pp. 23–30 (2011)
16. Singh-Miller, N., Collins, M., Hazen, T.J.: Dimensionality reduction for speech recognition using neighborhood components analysis. In: 8th International Conference of InterSpeech. pp. 1158–1161 (2007)
17. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2000)
18. Yang, W., Wang, K., Zuo, W.: Neighborhood component feature selection for high-dimensional data. Journal of Computers **7**(1), 161–168 (2012)