Prediction and Classification of Cardiovascular Diseases (CVDs) using Averaged One-Dependence Estimators (AODE) Classifier

Ayad R. Abbas
Department of Computer Science, University Of Technology, Baghdad, Iraq

Email: ayad_cs@yahoo.com, Mobile: 07715433873

Abstract

Cardiovascular diseases (CVDs) were the most well-known reason for death globally and the disease progresses are a sudden heart attack and stroke. Decrease the number of heart specialists and experts led to an increase the injury cases. Recently, many machine learning techniques are used to diagnose CVDs. However, most of those studies were not taken into consideration the data pre-processing before the diagnostic process, causing a variation in the accuracy results and the decision-making time. The proposed approach consists of two stages. First, data pre-processing (data cleaning, discretization and selection) are applied to the CVDs dataset taken from the UCI machine learning repository. In this stage, we modified the Correlation Feature Selection (CFS) with Best First Search (BFS) based on the Discriminant Index (DI) in order to enhance both time complexity and accuracy. Second, Averaged One-dependence Estimators (AODE) classifier is applied to predict CVDs and support non-expert doctors. The experimental results reveal that the proposed methods can implement with less time prediction of disease and more accuracy in order to support non-expert doctors in decision making without requiring the direct consultation with the specialists.

Keywords

Cardiovascular diseases (CVDs), Correlation Feature Selection (CFS), Best First Search (BFS), Averaged One-dependence Estimators (AODE), Discriminant Index (DI).

تنبؤ وتصنيف الأمراض القلبية الوعائية باستخدام AODE

د. أياد روضان عباس

[١] الجامعة التكنولوجية ، قسم علوم الحاسوب

ayad_cs@yahoo.com

المستخلص

ان الأمراض القلبية الوعائية تعد من الاسباب الاكثر شيوعا للموت المفاجئ خصوصا عند تقدم المرض مما يسبب نوبة قلبية وسكتة الدماغية. ان انخفاض عدد المختصين في هذا المجال أدى إلى زيادة حالات الإصابة. في الآونة الأخيرة، استخدمت العديد من تقنيات تعلم الماكنة لغرض تشخيص امراض القلب، الان أغلب تلك الدراسات لم تؤخذ بنظر الاعتبار معالجة البيانات قبل عملية التشخيص مما سبب تفاوت في دقة النتائج وسرعة اتخاذ القرار. لذلك، هذا البحث يقدم نموذج مقترح والذي يتكون من مرحلتين اعتمادا على البيانات التي تم الحصول عليها من مستودع تعليم الآلة وهو UCI. أولا، معالجة البيانات وهي (تنظيف البيانات وتفريدها واختيار الخصائص التي تؤثر في القرار). في تلك المرحلة تم تطوير نموذج اختيار الخصائص المرتبطة (CFS) مع استخدام خوارزمية البحث الستدلالي (BFS) استنادا إلى مؤشر التمايز (DI) من اجل تقليل الوقت وزيادة دقة التشخيص. المرحلة الثانية، ولاول مرة استخدمت طريقة Averaged One- Estimators (AODE) dependence وهي طريقة مطورة استخدمت لتشخيص امراض القلبية الوعائية. اظهرت النتائج التجريبية أن الطرق المقترحة يمكن أن تتنبأ بالاصابات باقل وقت وأكثر دقة لدعم صناعة القرار للأطباء المبتدئين دون الحاجة إلى التشاور المباشر مع المختصين.

1- Introduction

The World Health Organization (WHO) assessed 17 million people die of Cardiovascular diseases (CVDs), representing 31% of all global deaths. Of these deaths, an expected 7.4 million were because of CVDs and 6.7 million were because of stroke [1]. Most developing countries settings lacking the sufficient number of heart disease specialists causing the high mortality rate. Therefore, many researchers are encouraged by the machine learning tools and techniques to support non-expert doctors in decision making without requiring the direct consultation with the heart specialist[2].

Moreover, many researchers have focused on using machine learning algorithm in heart diagnosis with different accuracy, such as decision tree, rough set, support vector machine, etc. [3-5]

The proposed approach consists of two stages. First, Data cleaning is applied to remove noise data and full in missing value in the data. Then, the combination of CFS with BFS are modified and explored to robust heart disease dataset taken from UCI machine learning repository [6], with the objective of effectively identifying the several attributes, which best predict a selected target attribute. Second, AODE classifier is applied to significantly predict CVDs mortality.

2- Related Works

Many researches have applied different data mining technigues to diagnosis of heart disease over different dataset, these technigues are listed as following:

١. *Decision tree*:[7] has explored the applying a range of techniques to different types of J8.4 decision trees looking for better performance in heart disease diagnosis.[8] Has proposed a medical decision support system for coronary illness hazard classification utilizing C4.5 decision tree also datasets have classified and examined into 4 classes utilizing decision tree classifier.

٢. *Rough set and Neural Network*: the modified rough set is used to classify of three heart valve data sets. Four types of classification approaches were compared to evaluate the discriminatory power of the classification such as (Decision table, MultiLayer Perceptron (MLP), Back Propagation Network (BPN) and Navie Bayes)[9].

٣. *Fuzzy set*: A rough-fuzzy classifiers have been introduced by consolidating rough set with the fuzzy set. This classifier process is divided into two stages: First, generating classification rules using rough set theory. Second, predicating utilizing fuzzy classifier system [10]. In pre-process, rough set theory has been used to extract suitable rules, whereas the formal concept analysis in post-process has been used from these suitable rules to investigate critical elements influencing the decision making and better knowledge[3]. Fuzzy decision support system has used for the diagnosis of CVDs based on evidence. The knowledge base of decision support system is taken by utilizing rules extraction from rough set theory[5].

٤. *Support Vector Machine (SVM)*: an intelligent system based SVM along with a radial basis function network is presented for heart diagnosis. Expert system based on heart disease data sets is used to predicate what type of heart disease is possible to appear for a patient, whether it is heart attack or not[2].

3- The Proposed Materials and Methods

In this section, the proposed approaches for CVDs prediction of heart patients starting with set of training example called heart disease data sets then applying data mining techniques on these data sets: data preprocessing, Features Selection Algorithm (CFS based on BFS) and machine learning algorithm.

3-1 Cardiovascular Diseases Dataset[6]

Cardiovascular data sets from UCI are utilized as a part of this paper. The UCI Cardiovascular data sets of 303 patients are collected from Switzerland, Hungarian institute of cardiology, Switzerland, university hospital, V.A. Medical Center, Zurich, Basel, Long Beach and Cleveland Clinic Foundation. The 14 information of an attributes for these data sets are associated with physical examination, diagnostic laboratory, and stress tests as shown in Table 1.

Table 1: Complete Cardiovascular Diseases Dataset Description

Id	Attributes	Attribute Description
1	**age**	The patient age in years.
2	**Sexe**	Patient sex where 1 for male and 0 for female.
3	**cp**	Patient chest pain type.
4	**trestbps**	Patient resting blood pressure in mm Hg on admission to the hospital.
5	**chol**	Patient serum cholesterol in mg/dl.
6	**fbs**	Patient fasting blood where sugar > 120 mg/dl and 1 = true; 0 = false
7	**restecg**	Value 0:normal Value 1:having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05mV). Value 2:showing probable or definite left ventricular hypertrophy by Estes' criteria.
8	**thalach**	Patient maximum heart rate achieved.

9	**exang**	Patient exercise induced angina where 1 for yes and 0 for no.
10	**oldpeak**	Patient ST depression induced by exercise relative to rest.
11	**slope**	The slope of the peak exercise ST segment. Value 1:up sloping.
12	**ca**	Number of major vessels (0-3) colored by fluoroscopy.
13	**thal**	3 for normal, 6 for fixed defect and 7 for reversible defect.
14	**num**	Diagnosis of heart disease. Value 0:<50% diameter narrowing. Value 1:>50% diameter narrowing.

3-2 Proposed Data mining Model

In this section, the intelligent data mining model is proposed as shown in following diagram Figure 1. This model consists of two tasks: preprocessed data and machine learning tasks. The preprocessed data consists of three sub tasks: data cleaning and replacing missing values, discretization, and the features selection. The features selection task is proposed using CFS, rough set and Best first search.

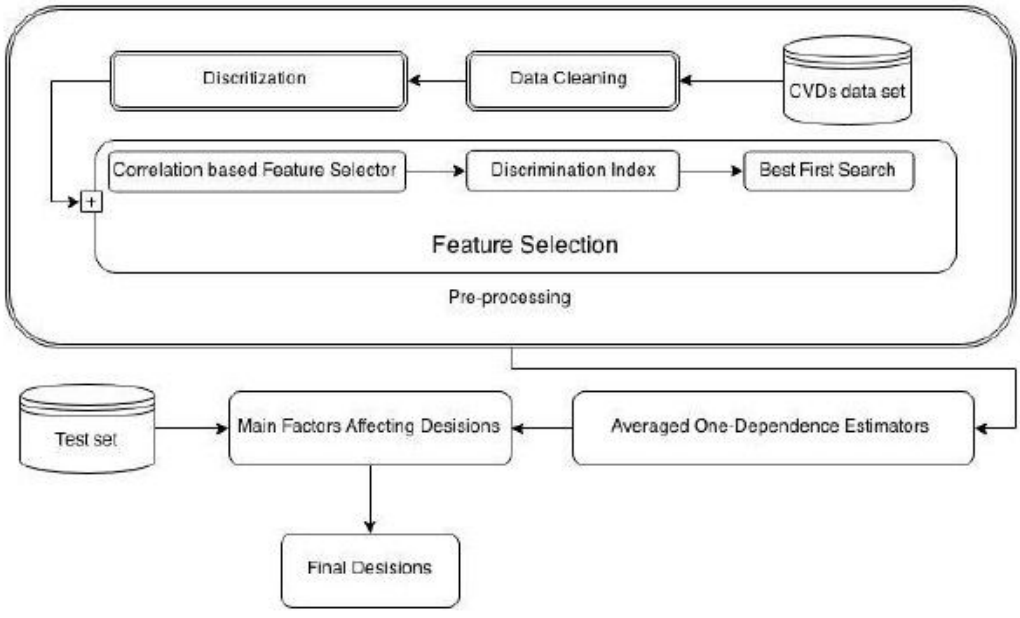

Figure 1. Proposed data mining model

3-2-1 Preprocessed CVDs

١. **Data Cleaning and Unsupervised Discretization**

Before the **AODE** process and to improved accuracy of data analysis and classification, two data preprossing have been applied. *Firstly*, the data cleaning task is applied for handling missing values on CVDs data using attribute mean to fill missing or unknown values.

Secondly, it is necessary to re-encode each continuous attribute using unsupervised discretization to discrete age, trestbps, thalach and oldpeak attributes constituted by a set of intervals using equal width interval binning.

1- The *Age* attribute can be transformed in 3 discrete values representing three intervals:

(<= 45, <45 <= 61, > 61) .

2- The *Trestbps* attribute can be transformed in 4 discrete values representing three intervals:

(<= 120.5, <120.5 <= 147, <147<= 173.5,> 173.5) .

3- The *Chol* attribute can be transformed in 3 discrete values representing three intervals:

*(**Desirable**: <=200, **Borderline high**: <200 <= 239, **High**: >239)*

4- The *Thalach* attribute can be transformed in 5 discrete values representing three intervals:

*(**1:** <= 97.2, **2:** <97.2<= 123.4,**3:** <123.4<= 149.6, **4:** <149.6<= 175.8 ,**5:** > 175.8) .*

5- The *Oldpeak* attribute can be transformed in 4 discrete values representing three intervals:

(<= 1.55, <1.55 <= 3.1, <3.1<= 4.65,> 4.65) .

3-2-2 Correlation Based Feature Selector CFS [11]

This section displays and selects relevant attributes using a CFS. It is a one of selector algorithms that sorts attribute subset considers a correlation degree using evaluation function. The influence of the evaluation or heuristic function is subsets that contain attributes that are exceptionally associated with the class and unassociated with one another. This algorithm reduces unrelated attributes from the information table because they will have low merit with the class. Whereas, it screens out redundant attributes as they will be high merit with one or more of the remaining attributes. CFS's feature subset merit Equation 1 is represented here:

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \qquad (1)$$

Where M_S is the merit of an attribute subsets S with k attributes, r_{cf} is the correlation degrees between the attributes and the class, and r_{ii} is the inter correlation degree between attributes using Discrimination Index (DI).

Because of all CVDs attributes are nominal values DI is applied to find the correlations between the attributes and the class where DI indicates the ratio of the number of positive instances to the number of all instances covered by the rule using Rough Set Methods in (Equation 2).

$$DI = 1 - \frac{The\ set\ all\ boundary\ region}{The\ set\ of\ all\ training\ examples} \qquad (2)$$

3-2-3 BFS Method[12]

BFS is one of the artificial intelligent search methods that work with backtracking through the search path. It uses the search space to move by making local search to the current attribute subset. If the path being investigated starts to look less encouraging, the BFS can backtrack to an additionally encouraging past subset and proceed the search from there. Sufficiently time, a BFS will investigate whole search space, so it is normal to utilize a stopping condition as shown in the following algorithm:

Algorithm 1: BFS Algorithm

1. CLOSED_list is empty and Add start state to OPEN_list and BEST_VAL=start_state.

2. Let state=argmax e(x) by getting the highest merit state from OPEN_list.

3. Delete state from OPEN_list and add state to CLOSED_list.

4. If e(state)>=e(BEST_ VAL), then BEST_ VAL = state.

5. For every child of state that is belong to OPEN_list or CLOSED_list, evaluate child and add it to OPEN_list.

6. If BEST_ VAL is changed, go to 2.

7. Return BEST_ VAL.

3-2-4 Averaged One-Dependence Estimators (AODE)[13]

The most popular problem in naive Bayes classifier is attribute-independence problem. Therefore, AODE is a probabilistic classification learning procedure. It was made to address this issue. It regularly modifies considerably more precise classifiers than naive Bayes at the expense of an unobtrusive increment in the measure of calculations. The likelihood of every class y a predefined arrangement of attributes x_1, ... x_n, $P(y \mid x_1, \ldots x_n)$ is evaluated by utilizing AODE as shown in Equation 3:

$$\widehat{P}(y|x_1, \ldots x_n) = \frac{\sum_{i:1\le i\le n \wedge F(x_i)\ge m} P(y,x_i) \prod_{j=1}^{n} \widehat{P}(x_j|y,x_i)}{\sum_{\acute{y}\in Y} \sum_{i:1\le i\le n \wedge F(x_i)\ge m} P(\acute{y},x_i) \prod_{j=1}^{n} \widehat{P}(x_j|\acute{y},x_i)} \qquad (3)$$

Where $\hat{P}(.)$ means an assessment of P(.), F(.) is the frequency with which the argument shows in the example and m is a user determined minimum frequency with which a term must show to be utilized as a part of the external summation. In recent practice m is generally situated at 1.

3-3 Experimental Results and Evaluation

The experimental results and evaluation in discovering main factors for heart attack are presented in here. The CSVs data set is mined here by eliminating identical records and replacing missing values using mean values method. After that, CFS is applied to select relevant CSVs attributes. Table 2 shows the feature correlation matrix for the CSVs data set using rough set method in order to calculate the correlation scors.

Table 2. Feature correlations calculated from Rough set

	Age	Sex	Cp	Trestbps	Chol	Fbs	Resteg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Num
Age	1.000	0.015	0.012	0.038	0.02	0.019	0.013	0.066	0.008	0.019	0.014	0.076	0.015	0.064
Sex		1.000	0.013	0.006	0.011	0.002	0.008	0.005	0.017	0.014	0.001	0.01	0.108	0.028
Cp			1.000	0.019	0.01	0.008	0.018	0.068	0.128	0.061	0.044	0.053	0.068	0.15
Trestbps				1.000	0.018	0.026	0.019	0.008	0.018	0.021	0.006	0.016	0.011	0.011
Chol					1.000	0.001	0.023	0.015	0.007	0.011	0.003	0.012	0.004	0.01
Fbs						1.000	0.008	0.004	0.001	0.003	0.007	0.014	0.008	0.001
Resteg							1.000	0.039	0.007	0.022	0.024	0.015	0.005	0.023
Thalach								1.000	0.077	0.063	0.1	0.038	0.049	0.088
Exang									1.000	0.05	0.056	0.024	0.072	0.149
Oldpeak										1.000	0.157	0.053	0.067	0.106
Slope											1.000	0.019	0.069	0.102
Ca												1.000	0.038	0.146
Thal													1.000	0.189
Num														1.000

Thus, forward BFS for features selection through the feature subset space along with the merit of each subset is applied using Equation 1. As mentioned above the calculation of selection high quality features depends on higher correlation between specific attribute and class, and lower inner correlation between attributes. The higher correlation degrees of an attributes *thal*, *ca*, *slop*, *oldpeak*, *exang* and *cp* with *num as a* class are 0.189, 0.146, 0.102, 0.106, 0.149 and 0.15 respectively. The BFS begins with the empty set of attributes, which has zero merit. When K=1 means Each single attribute addition to the empty set is evaluated; *thal* is added to the subset because it has the highest score. The next step when K=2 involves trying each of the remaining features with *thal* and choosing the best (*ca*). Because the score of attribute *ca* with *num* class is 0.146 (high score) and 0.038 (low score) with the *thal* attribute. As a results the relevant features selection are *thal*, *ca*, *slope*, *oldpeak*, *exang* and *cp* are shown in Table 3. Thuse, the merit of best subset found is 0.299.

Table 3: The Cardiovascular Diseases Dataset after Preprocessing

No.	cp	exang	oldpeak	slope	ca	thal	num
1	typ_angina	no	(1.55-3.1]	down	0	fixed_defect	<50
2	asympt	yes	(-inf-1.55]	flat	3	normal	>50_1
3	asympt	yes	(1.55-3.1]	flat	2	reversable_defect	>50_1
.							
.							

303	non_anginal	no	(-inf-1.55]	up	0	normal	<50

The AODE classifier is applied using equation 5.2 where Prior probability = 0.54 and 0.45 for Class <50 and Class >50_1 respectively. Correctly classified instances are 259 with 85.4785 % and incorrectly a classified instance is 44 with 14.5215 %. The evaluation of the proposed method is shown in tables 4 and 5.

Table 4: Tools test

Tool test	**Results**
Kappa statistic.	0.7051
Mean absolute error.	0.0908
Root means squared error.	0.2168
Relative absolute error.	45.2382 %
Root relative squared error.	68.8213 %
Total Number of Instances	303

Table 5: Related Classifier Methods

Classifier	**Correctly Classified Instances**	**Execution Time**
Proposed AODE	85.4785 %	0.00001 s
AODE without preprocessing	84.4785 %	0.01 s
Naïve Bayes	84.8185 %	0.01 s
J.48	85.3334 %	0.08 s
Net Bayes	84.4884 %	0.02 s

4- Conclusion

Many attempts have been made in this research to diagnose CVDs on UCI Cardiovascular data sets. The CFS model has been modified using ID and BFS as well as AODE has been used to develop a predictive model for diagnose and risk classification. The experimental results have been evaluated using many tests. By using Cohen's kappa measure and many absolute error tests as shown in Table 5 the correctly classified instances is 85.4785 % and classification execution time near to zero time (0.00001 s) that is the proposed methods can implement with less time prediction of disease and more accuracy in order to support non-expert doctors in decision making without requiring the direct consultation with the specialists.

5- References

[1] WHO. *Cardiovascular diseases(CVDs)*. Retrieved 5 2015, from Media centre Available:http://www.who.int/mediacentre/factsheets/fs317/en/. 2015.

[2] S. Ghumbre, C. Patil, and A. Ghatol. "Heart disease diagnosis using support vector machine." *International Conference on Computer Science and Information Technology (ICCSIT)*. 2011. 84-88.

[3] B. K. Tripathy, D. P. Acharjya and V. Cynthya. "A Framework for Intelligent Medical Diagnosis using Rough Set with Formal Concept Analysis." *International Journal of Artificial Intelligence & Applications*. 2011. Vol. 2 (2). 45– 66.

[4] Dr. Anooj P.K. "Clinical Decision Support System: Risk Level Prediction of Heart Disease Using Decision Tree Fuzzy Rules." *Asian Transactions on Computers*. 2012.

[5] SETIAWAN, NOOR AKHMAD. "*Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System." PhD thesis, Universiti Teknologi PETRONAS*. 2009.

[6] Robert Detrano & M.D & PhD, V.A. *Medical Center, Long Each and Cleveland Clinic Foundation*. Retrieved 5 2015, Available:

http://www.archive.ics.uci.edu/ml/datasets/Heart+Disease. 2015.

[7] Shouman, M., Turner, T. and Stocker. "Using Decision Tree for Diagnosing Heart Disease Patients." *Australasian Data Mining Conference (AusDM 11)*. 2011. 23-30.

[8] Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen. "A Heart Disease Prediction Model using Decision Tree.". *e - ISSN: 2278 - 0661, p - ISSN: 2278 – 8727*. 2013. Vol. 12. 83 -86.

[9] H. Hannah Inbarani, S. Senthil Kumar, A. E. Hassanien, and A. T. Azar, "Soft Rough Sets for Heart Valve Disease Diagnosis." *The 2nd International Conference on Advanced Machine Learning Technologies and Applications. Egypt*. 2014. 17-19.

[10] K. Srinivas, G. Raghavendra Rao, A. Govardhan, "Rough-Fuzzy Classifier: A System to Predict the Heart Disease by Blending Two Different Set Theories." *Arabian Journal for Science and Engineering*. 2014, Vol. 39 (4). 2857-2868.

[11] Mark A. Hall. "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning.". *ICML*. 2000. 359-366.

[12] E. Rich and K. Knight. *"Artificial Intelligence."* McGraw-Hill, 1991.

[13] Webb, G. I., J. Boughton, and Z. Wang "Not So Naive Bayes: Aggregating One-Dependence Estimators." *Machine Learning*. 2005. 58(1). 5–24.