Predicting heart failure class using a sequence prediction algorithm

Carine Bou Rjeily
Nanomedecine Lab, Imagery & Therapeutics UBFC, Belfort, France

Georges Badr
TICKET Lab
Antonine University
Hadat-Baabda, Lebanon

Amir Hajjam Al Hassani
Nanomedecine Lab, Imagery & Therapeutics, UBFC
Belfort, France

Emmanuel Andres
Université de Strasbourg,
Centre Hospitalier Universitaire
Strasbourg, France

***Abstract*—One of the major causes of death in the world is Heart Failure. This disease affects directly the heart's pumping job. Because of this perturbation, nutriments and oxygen are not well circulated and distributed. The New York Heart Association has classified this disease into four different classes based on patient symptoms. In this paper, we are using a data mining technique, more precisely a sequential prediction algorithm (CPT+) to predict to which of the 4 classes a patient belongs. The algorithm was run on a dataset containing 14 attributes representing patients' vital signs, including the class of the disease. Category prediction yielded to an average accuracy of 90.5%.**

Keywords—Heart Failure; Classification; Data Mining; Sequence prediction.

I. INTRODUCTION

The human body functions normally when all its cells are nourished properly. This is established by the heart's pumping action, which results in delivering the oxygen and nutrient-rich blood, needed for the cells. When the heart fails in establishing properly this action, or cannot keep up with this workload, this is known as "Heart Failure". In definition, Heart Failure (HF) is a chronic, progressive condition in which the heart muscle is unable to pump enough blood to meet the body's needs for blood and oxygen causing its death [2]. HF should not be mistaken with Heart Attack (HA) (known as Myocardial Infraction). HA happens when one of the heart's own blood vessels becomes blocked by a blood clot reducing the heart's power to squeeze out blood as it should. HA is one of the major causes of HF. Coronary Artery Disease and Hypertension are the most common causes of heart failure. Past heart attack, abnormal heart valves, diabetes, obesity, myocarditis, alcohol etc. [3] are also conditions that lead to heart failure.

The New York Heart Association (NYHA) [1] categorizes heart failure into 4 types depending on its severity. Symptoms differ from one patient to another depending on the class. At early stages, symptoms are unlikely to be noticed. Accumulated fluid or congestion is the most common symptom. According to [4] heart failure has more symptoms like shortness of breath, coughing or wheezing, weight gain, swollen ankles, and poor blood flow to the body (tiredness / fatigue, dizziness, rapid heartbeat, loss of appetite, the need to urine at night, confusion, impaired thinking).

In this paper, we are presenting the use of a data-mining algorithm. We used a sequence prediction algorithm, namely the Compact Prediction Tree plus, CPT+, on a medical dataset. Our purpose is to predict one of four class of HF. The remainder of the paper is organized as follows. The next section deals with related and previous works on HF and its classes. Section 3 presents the adopted method and algorithm. Experiments and results are discussed in section 4 before ending with a conclusion and future works.

II. RELATED WORKS

Many studies have immerged to study heart diseases and especially heart failure. Researchers try to predict HF or its types.

Pandey et al. [5] used the medical dataset in [6] to test their prediction model for heart disease. The model is a J48 decision tree classifier. The data was divided into a training set and a testing set with a 60%-40% ratio. The obtained accuracy reached 75.73%.

Same dataset was processed by Bashir et al. [7] who used a framework with Naïve Bayes, Decision Tree and Support Vector Machine. An accuracy of 81.82% was obtained with two output classes.

Another framework was used in [8] containing CARD, ID3 and DT decision trees. The framework applies a 10-fold cross validation on the dataset. The highest accuracy (83.49%) was presented by the CARD decision tree, followed by DT (82.50%) and ID3 (72.93%).

The dataset in [6] contains 13 attributes consisting of patients' vital signs. Uppin et al. [9] have used only 7 out of the 13 attributes in a C4.5 decision tree classifier. The main purpose to reduce the number of attributes is to avoid the redundant features. Results showed an accuracy of 85.96%.

A reduced C4.5 decision tree with a new pruning method was used in [10]. Pruning was done by combining pre and post pruning methods. Testing the method on the dataset resulted in an accuracy of 76.51%.

Authors of [11] used the K-means cluster in order to diagnose heart disease. They obtained an accuracy of 83.9% by applying the inlier method with two clusters.

An alternative for conventional decision tree was proposed by Bohacik et al [12]. In this method, each part of the decision

978-1-5386-1642-0/17/$31.00 ©2017 IEEE

tree can be split multiple times while in classic methods, only lead nodes can be split. The authors used a dataset from Hull LifeLab containing 9 parameters and 2 prediction classes. A 10-fold cross validation was used resulting a 77.65% of accuracy.

CART decision tree was implemented on ECG recordings in Melillo et al [13]. The method was able to classify patient according to their risk factors and achieved 85.4% of accuracy.

In [14], the C4.5 decision tree was used to predict and classify heart failure into 5 risk levels. The dataset presented in [6] was used. The model resulted in 86.5% sensitivity and 86.53% accuracy.

III. ADOPTED ALGORITHM

A. Important terms and definitions

Before presenting the adopted method, it is important to define some basic terms used in sequential pattern mining.

a) An item is an entity that can have multiple attributes: date, size, color, etc.

b) $I = \{i_1, \ldots, i_n\}$ is a non empty set of items. A k-item set is an item set with k items.

c) A sequence "α" is an ordered list of item sets.

d) A sequential Database (SDB) is a list of sequences with a sequence ID (SID).

e) A sequential rule, denoted $X \rightarrow Y$, is a relationship between two unordered item sets. This means that if items of X appear in a sequence, items of Y will also occur in the same sequence.

f) The support of a rule r in a sequence database SDB is defined as the number of sequences that contains X∪Y divided by the number of sequences in the database.

g) A rule r is a frequent sequential rule iff $supSDB(r) \geq minsup$, with $minsup \in [0, 1]$ is a threshold set by the user.

h) The confidence of a rule r in a sequence database SDB is defined as the number of sequences that contains X∪Y, divided by the number of sequences that contains X.

i) A rule r is a valid sequential rule iff it is frequent and $confSDB(r) \geq minconf$, with $minconf \in [0, 1]$ is a threshold set by the user.

j) Sequence prediction is the fact of predicting the next element in given sequence. This prediction is preceded by a learning phase on a set of sequences, called training sequences.

B. CPT+

CPT or Compact Prediction Tree is a sequence prediction algorithm used to guess the next symbol in a given sequence [15]. CPT+ is an enhanced version of CPT and is up to 98 times more compact and 4.5 times faster than CPT [16]. The choice of CPT+ comes after its performance and its low resources consumption. It also stores the training sequences in a compressed form with no or a minimal loss of information. Its prediction is noise-tolerant after the use of a similarity measure in order to identify predictable sequences.

The CPT model is defined by two processes: a training process and a prediction process.

1) Training

CPT+ needs to be trained in order to guess the next symbol in a given sequence. A training set is given as input, and the algorithm processes the sequences one by one to incrementally build its three structures: the Prediction Tree (PT), the Inverted Index (II), and the Lookup Table (LT).

Prediction Tree: It is a type of prefix tree composed of recursively-defined node with all training sequences. Each symbol is a node with a pointer to its parent node. A sequence is defined by the path from a direct child of the root node to an inner node or leaf.

Inverted Index: It is a Hash Table, where each unique item is a key. If an it-th element in the i-th sequence exists, value is set to one, otherwise, it is set to zero. Sequences are then rapidly found.

Lookup Table: It is an associative array updated after each insertion in the PT. To guarantee the lossless of the data, the LT is used to get the sequences form the PT using their ID.

Fig. 1 below illustrates the creation of the CPT structures by successive insertion of the sequences S1 = (A, B, C), s2 = (A, B), s3 = (A, B, D, C), where the alphabet Z = {A, B, C, D, E} is used.

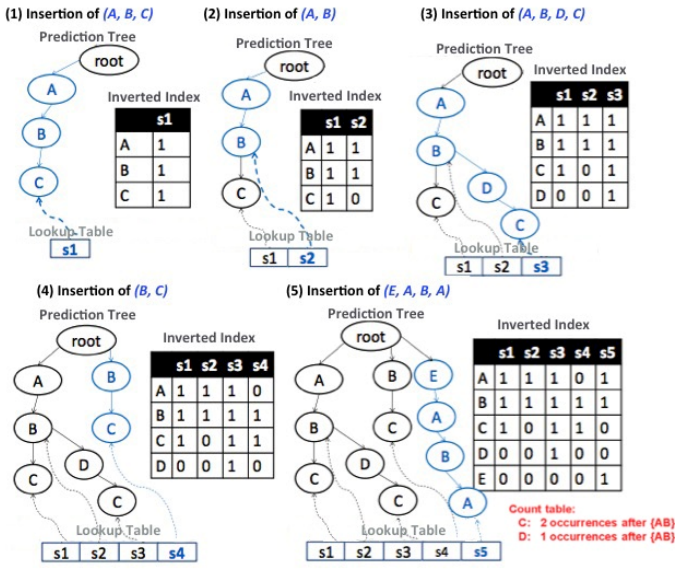

Fig. 1. Example ilustrating the structures used in CPT+

2) Predicting

The three described structures are used in the prediction process. The prediction is done by finding sequences similar to given sequence. This means that the algorithm tries to find X number of elements of the sequence that appears together in any order and any position. X is an integer called "prefix length". Then, for each similar sequence, the algorithm considers its consequent. It is the subsequence of the sequence

containing at least X elements. A hash table called Count Table (CT) stores each element of theses consequents. The element having the largest number of occurrences is the one to be predicted. The sequences are found using the II and accessed in the PT using the LT.

3) Optimization

An item is considered as noise if it has a low frequency (support). CPT+ algorithm removes noise items from subsequences and updates the CT. It presents a Prediction with Improved Noise Reduction (PNR) using two attributes:

- noiseRatio: % of items in a sequence considered as noise
- minimumPredictionRatio: minimum number of updates

IV. EXPERIMENTS AND RESULTS

CPT+ was used in early experiments to predict the presence or absence of heart disease considering the vital signs of a user. We proved that a sequence prediction algorithm can be useful to predict if a patient has or not a heart disease with an accuracy of 87.03% [17]. Then we conducted other experiments on another dataset, in order to verify if that kind of algorithms could also be used for classification.

Below, a detailed description of the dataset, the experiments will be presented. Then results will be discussed and interpreted.

A. Dataset

In this study, we used the Cleveland Clinic Foundation heart disease dataset. The dataset is freely available on the UCI Machine Learning repository at [6]. It contains 297 records without any missing value. Each record is formed of 13 attributes representing patient's vital signs, followed by the output class. The predicted class indicates to which of the 5 classes a patient may belong. 53.87% of the instances are for patients with no risk of developing a heart failure, while the remaining 46.13% are for patients with different risk levels. Table 1 below presents the 13 attributes.

TABLE I. ATTRIBUTES CONSIDERED IN HEART FAILURE CLASSIFICATION

Attributes	Type	Description
Age	Continuous	Age in years
Sex	Discrete	Gender (1: male and 0: female)
CP	Discrete	Chest pain 1: typical angina, 2: atypical angina 3: non-anginal pain, 4: asymptomatic
Chol	Continuous	Serum cholesterol in mg/dl
Trestbps	Continuous	Resting blood pressure in mm Hg
Fbs	Discrete	Fasting blood sugar > 120 mg/dl 1: true, 0: false
Restecg	Discrete	Resting electrocardiography results 0: normal, 1: having ST-T wave abnormality, 2: left ventricular hypertrophy
Thalach	Continuous	Maximum heart rate
Exang	Discrete	Exercise induced angina (1: yes, and 0: no)
Oldpeak	Continuous	Depression made by exercise relative to rest
Slope	Discrete	The slope of the peak exercise ST segment 1: up sloping, 2: flat, 3: down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy (from 0 to 3)
Thal	Discrete	3: normal, 6: fixed defect, 7: reversible defect
Class	Discrete	The predicted attributes 0: No risk, 1: Low risk 2: Moderate risk, 3: High risk 4: Extremely high risk

80% of sequences in the datasets were used as training set, and the remaining 20% were used for testing and validation.

B. Results

Confusion Matrix was used to calculate the results and the overall performance of the predictive model for heart failure risk. This includes sensitivity, specificity and accuracy. The idea is to compare the predicted class with the actual class. Thus, True Positive (TP), and True Negative (TN) are both correct classifications. Whereas, False Positive (FP) and False Negative (FN), occur respectively when the class is incorrectly predicted as positive when it is actually negative, and incorrectly predicted as negative when it is actually positive.

Let's consider that (0, I, II, III, IV) matches the risk level to be predicted (no risk, low level, moderate, high and extremely high). Results are then arranged in a confusion table shown in Table II.

TABLE II. THE CONFUSION MATRIX OF THE PREDECTIVE MODEL

		Predicted Class								
		0	**I**	**II**	**III**	**IV**	**TP**	**TN**	**FP**	**FN**
Actual Class	**0**	31	1	0	0	0	31	28	10	1
	I	6	6	1	1	0	6	43	3	8
	II	2	2	3	0	0	3	52	1	4
	III	2	0	0	4	0	4	53	1	2
	IV	0	0	0	0	1	1	59	0	0

Sensitivity, specificity, precision and accuracy can be calculated as follow.

$$Sensitivity = \frac{\sum TP}{\sum(TP + FN)}$$

$$Specificity = \frac{\sum TN}{\sum(TN + FP)}$$

$$Precision = \frac{\sum TP}{\sum(TP + FP)}$$

$$Accuracy = \frac{\sum(TP + TN)}{\sum(TP + FN + FP + TN)}$$

Table III shows the details of CPT+ performance of each class.

TABLE III. DETAILED PERFORMANCE

Output class	Sensitivity	Specificity	Precision	Accuracy
0	0.968	0.736	0.756	0.8423
I	0.428	0.934	0.666	0.8167
II	0.428	0.981	0.75	0.9167
III	0.666	0.981	0.8	0.95
IV	1	1	1	1
Average	0.698	0.926	0.794	0.905

TABLE III above shows that the maximum accuracy goes for class IV that is considered the most severe class and the most threatening level for the patient. The minimum accuracy is when detecting class I disease.

Those results were achieved for:

- minPredictionRatio = 0.3
- noiseRatio = 0.2.

C. Comparison with other algorithms

This paper shows the capability of a sequence prediction algorithm, namely the CPT+, to classify data especially heart disease data with 90.5% of average accuracy. The obtained results were compared to some other results that were published earlier in the literature. TABLE IV below illustrates this comparison.

TABLE IV. COMPARISON WITH OTHER EXISTING METHODS

Algorithm	Output classes	Accuracy
CPT+	5	90.5%
C4.5 Decision Tree [14]	5	86.53%
Decision Tree with Reduced Error Pruning Method [5]	2	75.73%
A combination of Naïve Bayes, Decision Tree and Support Vector Machine [7]	2	81.82%
CARD Decision Tree [8]	2	83.49%
C4.5 Decision Tree [9]	2	85.86%
C4.5 Decision Tree with New Pruning Method [10]	2	76.51%
K-means Clustering and Decision Tree [11]	2	83.9%
Alternating decision tree [12]	2	77.65%
CARD Decision Tree [13]	2	85.4%

Our adopted method is also capable of predicting 5 out of 5 of the disease class unlike other algorithms that can predict only two classes.

V. CONCLUSION

Disease detection or prevention is a primordial task researchers are trying to solve in order to help medical personals in their decisions. Those tasks could be achieved with data mining.

In this paper we have implemented a sequence prediction algorithm, namely the CPT+ to classify the risk level of a patient to have a heart failure. This algorithm predicts the next symbol from a sequence in any order or position.

To validate our proposition, we carried out some experiments on the Cleveland heart dataset which in available online [6].

Results showed an average accuracy of 90.5%, making CPT+ on the top of other well-known algorithms.

As for future works, the algorithm will be tested on real data and histories collection from hospitals.

REFERENCES

[1] New York Heart Association, 2003. New York Heart Association (NYHA) classification.

[2] Cannon, D. E., 2009. *What's the difference between heart attack and heart failure?*

[3] Aha, *Causes Of Heart Failure*, viewed 21 June 2016, <http://www.heart.org/HEARTORG/Conditions/HeartFailure/CausesAndRisksForHeartFailure/Causes-of-Heart-Failure_UCM_477643_Article.jsp#.V2kq2ut97IU>, 2016.

[4] Dickstein, K., *Symptoms Of Heart Failure*, viewed 29 June 2016, http://www.heartfailurematters.org/en_GB/Understanding-heart-failure/Symptoms-of-heart-failure

[5] A.K. Pandey, P. Pandey, K.L. Jaiswal, and A.K. Sen, "A heart disease prediction model using decision tree", IOSR Journal of Computer Engineering (IOSR-JCE), vol. 12, Issue 6, PP 83-86, Aug. 2013.

[6] Cleveland Clinic Foundation, "Heart Disease Data Set ", Available at: http://archive.ics.uci.edu/ml/datasets/Heart+Disease, Accessed in: 7 March 2015.

[7] S. Bashir, U. Qamar, and M.Y. Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis", International Conference on Information Society (i-Society 2014), London, IEEE, 2014.

[8] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques" Caribbean Journal of Science and Technology, vol.1, 208-217, 2013.

[9] S.K. Uppin, and M.A. Anusuya, "Expert system design to predict heart and diabetes diseases", International Journal of Scientific Engineering and Technology, vol. 3, no.8, pp: 1054-1059, 2014.

[10] A. M. Mahmood and M. R. Kuppa, "Early detection of clinical parameters in heart disease by improved decision tree algorithm", 2010 Second Vaagdevi International Conference on Information Technology for Real World Problems, Warangal, IEEE, 2010.

[11] M. Shouman, T. Turner and R. Stocker, "Integrating decision tree and kmeans clustering with different initial centroid selection methods in the diagnosis of heart disease patients", Proceedings of the International Conference on Data Mining, 2012.

[12] J. Bohacik, C. Kambhampati, D.N. Davis and G.F. John, "Alternating decision tree applied to risk assessment of heart failure patients", Journal of Information Technologies, vol. 6, no. 2, 2013.

[13] P. Melillo, N.D. Luca, M. Bracale and L, Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability", IEEE Journal of Biomedical and Health Informatics, vol. 17, issue 3, 2013.

[14] Aljaaf, A.J., Al-Jumeily, D., Hussain, A.J., Dawson, T., Fergus, P. and Al-Jumaily, M., 2015, April. Predicting the likelihood of heart failure with a multi level risk assessment using decision tree. In *Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2015 Third International Conference on* (pp. 101-106). IEEE.

[15] Gueniche T. et al., 2013, Compact prediction tree: A lossless model for accurate sequence prediction, *International Conference on Advanced Data Mining and Applications*, Springer, Berlin Heidelberg.

[16] Gueniche T. et al. 2015, CPT+: Decreasing the time/space complexity of the Compact Prediction Tree, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer International Publishing.

[17] Bourjeily C., Badr G., Hajjam Al Hassani A., Andres E., *Heart Failure Prediction with CPT+, In 9th International Conference on e-health, 20-22 July 2017*.