



A NOVEL THRICE FILTERED INFORMATION ENERGY BASED PARTICLE SWARM FEATURE SELECTION FOR THE HEART DISEASE DIAGNOSIS

P. Umasankar

Assistant Professor, Department of Computer Science
M.I.E.T College of Arts and Science
Tiruchirappalli, India
pus.msu2013@gmail.com

V. Thiagarasu

Associate Professor, Department of Computer Science
Gobi Arts and Science College
Erode, India

Abstract

Cardiovascular disease is one of the most prevalent causes of death around the world and has deemed as a vital illness in older and Middle ages. Coronary artery disease, in appropriate, is a general cardiovascular disease involving high death rates. Angiography is, more frequently than not, regarded as the best system for the examination of coronary artery disease; on the other hand, it was connected with significant side effects and high costs. Much investigation has, consequently, be conveyed using data mining and machine learning to attempt alternative modalities. In this paper, a novel Thrice Filtered Information Energy based Particle Swarm Feature Selection method for identifying the relevant features in the classification of heart disease.

Keywords: Heart Disease, Feature Selection, Symmetrical Uncertainty, Chi-Square Analysis, Information Gain

1. Introduction

Data mining introduces to the extraction of relationships, hidden information and valid patterns from huge datasets [1]. This interdisciplinary subfield has applied in insurance, marketing and banking for the quality improvement and cost reduction [2]. Recent years have observed a wave of interest in employing data mining and machine learning in the field of medication for early investigation [3][4][5][6][7][8][9][10]. Cardiovascular disease is the most general cause of death the world over. Mainly, coronary artery disease (CAD) [11] is the usual cardiovascular condition. CAD happens when at least one of the left anterior descending (LAD) [12], left circumflex (LCX), and right coronary (RCA) arteries are stenotic [13].

In order to identify CAD, practitioners currently apply various techniques, among which angiography is widely regarded as the most well-defined method. It is, however, associated with major side effects and high costs; hence researchers have long attempted to devise specific diagnostic modalities. In fact, a large number of studies with the UCI datasets [14] being the most often used sets. These datasets, however, are not up-to-date.

2. Feature Selection Methods

2.1 Symmetrical Uncertainty

The symmetrical uncertainty (SU) [15] between target concept and features are applied to obtain the best features for classification. The elements with higher SU values have the higher weight. SU measures the relationship among A, B variables based on the information theory. It was calculated as follows

$$SU(A, B) = 2 \frac{I(A, B)}{H(A) + H(B)}$$

Computing $I(A, B)$ as the MI among A, B. $H(\cdot)$ as an entropy function for A, B features. The SU shows the normalized range value [0,1] as correction factor value is 2. If SU value is 1, then the information of one feature is predictable. If SU value is 0, then A, B are not associated.

2.2 Information Gain

Entropy is generally utilized in the information theory measure, which defines the purity of an absolute collection of examples. It is in the foundation of Gain Ratio, Information Gain and Similarity Uncertainty (SU) [15]. The entropy measure is considered a measure of the system's unpredictability. The entropy of Y is

$$H(Y) = \sum_{y \in Y} p(y) \log_2(p(y)) \quad (3.1)$$

Where $p(y)$ is the marginal probability density function for the random variable Y . If the observed values of Y in the training data set S have partitioned according to the values of a second feature X , and the entropy of Y for the partitions induced by X is less than the entropy of Y before partitioning, then there is a relationship between features Y and X . The entropy of Y after observing X is then:

$$H(Y|X) = \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2(p(y|x)) \quad (3.2)$$

where $p(y|x)$ is the conditional probability of y given x .

Given the entropy is a criterion of impurity in a training set S , we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases. This measure is known as IG. It is given by

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \quad (3.3)$$

IG [9] is a symmetrical measure and it is given by equation (3.3). The information gained about Y after observing X is equal to the information gained about X after observing Y . A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.

2.3 Chi-Square Analysis

Feature Selection via chi-square χ^2 test [17] is another, very commonly used method. Chi-squared method estimates the worth of a feature by calculating the value of the chi-squared statistic with respect to the class. The initial hypothesis H_0 is the assumption that the two features are unrelated, and it is tested by chi-squared formula:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \left(\frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2 \quad (3.6)$$

Where O_{ij} is the observed frequency, and E_{ij} is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of χ^2 , the greater the evidence against the hypothesis H_0 .

3. Proposed Thrice Filtered Information Energy based Particle Swarm Feature Selection (TFIE-PSFS) method

In this proposed method, the three filtered feature selection techniques like Symmetrical Uncertainty, Information Gain and Chi-Square analysis are combined to find the information energy of the features in the heart disease dataset. Particle Swarm Optimization [18] has used as the search space to get the optimal dataset for the prediction of heart disease.

PSO has used as the search space, and it is combined with the calculation of information energy using SU, IG, CS to develop the Thrice Filtered Information Energy based Particle Swarm Feature Selection method. The classification accuracy of the particle has compared with the decision feature. If the accuracy of the feature has a value greater than the Decision Feature, then the feature is stored in T and passed to empty set R. Now, the R has a feature with higher classification accuracy. Then the velocity is calculated and updated by using the equation

$$U_{ud} = U_{ud} + C1i1(p_{jd} - y_{jd}) + C2i2(p_{gd} - y_{jd}) \quad (1)$$

$$Y_{jd} = Y_{jd} + U_{ud} \quad (2)$$

This procedure has repeated the classification accuracy of the Similarity Uncertainty SU based on decision feature Df is equal to the classification accuracy of conditional feature based on Df. The final output is the optimal data set.

Table 1: Algorithm terms and its description

Term	Description
------	-------------

ω	Inertia
Df	Decision Features
Cf	Conditional Feature
c1 & c2	Social and Cognitive Components
SS	Swarm Size
DV	Data Fitness Value
pt_{best}	Particle Best Position
Gl_{best}	Global Best Position
R	Null Set
T	Feature Set
U_{ud}	A velocity of the Particle
y_{jd}	Individual Best position
j1 & j2	Random Values
m	This is a counter which counts from i to d
d	dimension of the dataset: This variable shows number of features in the dataset

Pseudo Code: Proposed Thrice Filtered Information Energy based Particle Swarm Feature Selection (TFIE-PSFS)

Input: Heart Disease Dataset

Step 1: Initialize the for loop for the given number of features in heart disease dataset

$for\ m \leftarrow d$

Step 2: Calculating the information energy of SU, IG, and CS for m^{th} feature and assigning them as

$E \leftarrow \{IG(m), GR(m), CS(m)\}$

Step 3: Calculating the information energies and saving its value in an array

$$IE = \frac{1}{3} \sum_{j=1}^3 E_{\sigma(j)}^2 F(m)$$

Step 4: end for

Step 5: Initialize Cf, Df from IE calculation

Step 6: Initialize the Swarm Size as 25, c1 and C2 as 2.0. Random values j1 & j2 as 0.2. Inertia value as 0.33.

Step 7: For each particle j in SS do

Step 8: Initialize the null variable as R

Step 9: Set the value of R in T.

Step 10: Initialize $\forall x \in (Cn - R)$ after storing the empty data set in T, the conditional features have checked with an empty set

Step 11: Use Griewangk function for calculation of Data Fitness Value, and it has represented DV

Step 12: Compare if $DV > Pt_{best}$

Step 13: Then set $Pt_{best} = DV$

Step 14: Compare If $Pt_{best} > Gl_{best}$

Step 15: Then Set $Gl_{best} = Pt_{best}$

Step 16: Compare if $R_U(X) > \gamma_T(D)$

Step 17: Then Set $T \leftarrow R \cup (X)$

Step 18: Finally set $R \leftarrow T$

Step 19: Calculate and update the velocity by using the equation 1

Step 20: Repeat the process from step 7 until $\gamma_R(Df) == \gamma_C(Df)$ is satisfied.

Step 21: If $\gamma_R(Df) == \gamma_C(Df)$

Step 22: Return R .

Output: Optimal Dataset

4. Result and Discussion

4.1 Description of the Dataset

The description of the Heart Disease dataset [7] has given below in the table 2.

Table 2: Description about the Dataset

No	Name	Description
1	Age	Age in Years
2	Sex	1=male, 0=female
3	cp	chest pain type(1 = typical angina, 2 =atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4	trestbps	Resting blood sugar(in mm Hg on admission to hospital)
5	chol	Serum cholesterol in mg/dl
6	fbs	Fasting blood sugar>120 mg/dl(1= true, 0=false)
7	restecg	Resting electrocardiographic results(0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy)
8	thalach	Maximum heart rate
9	exang	Exercise-induced angina
10	old peak	ST depression induced by exercise relative to rest

11	slope	Slope of the peak exercise ST segment (1=upsloping, 2=flat, 3= downsloping)
12	Major Vessels Colored (ca)	Number of major vessels colored by fluoroscopy
13	thal	3= normal, 6=fixed defect, 7= reversible defect
14	num	Class(0=healthy, 1=have heart disease)

4.2 Experimental Result and Discussion

The above algorithm has implemented in WEKA and MATLAB. The following tables represented the result obtained by existing feature selection and proposed feature selection method. The features with the rank 0 has removed to improve the classification accuracy. Table 3 depicts the result obtained by implementing the Symmetrical Uncertainty feature selection method on Heart Disease dataset. In the table 3, the features like thal, thalach, restecg, fbs have removed since its rank is 0.

Table 3: Result obtained by Symmetrical Uncertainty Feature Selection Method

Average Merit	Average Rank	Attribute name
0.193 +- 0.015	1 +- 0	13 thal
0.177 +- 0.004	2.2 +- 0.4	5 chol
0.162 +- 0.015	3.3 +- 0.9	3 cp
0.148 +- 0.013	4.2 +- 1.25	9 exang
0.149 +- 0.016	4.7 +- 0.9	12 ca
0.13 +- 0.005	6 +- 0	8 thalach
0.119 +- 0.007	6.7 +- 0.9	11 slope
0.108 +- 0.005	7.9 +- 0.3	10 oldpeak
0.077 +- 0.01	9.5 +- 0.81	2 Sex
0.072+-0.004	9.7+-0.46	1 age

0.068+-0.006	10.8 +- 0.4	4 trestbps
0.035 +- 0.005	12 +- 0	7 restecg
0.004 +- 0.001	13 +- 0	6 fbs

Table 4 gives the result obtained by using Information Gain feature selection method. The features chol, thalach, old peak, sex, restecg, fbs with rank 0 are removed to improve the prediction accuracy of the heart disease.

Table 4: Result obtained by Information Gain Feature Selection Method

Average Merit	Average Rank	Attribute name
0.697 +- 0.02	1 +- 0	5 chol
0.464 +- 0.019	2 +- 0	8 thalach
0.278 +- 0.014	3 +- 0	10 oldpeak
0.226 +- 0.02	4.9 +- 1.04	3 cp
0.218 +- 0.013	5 +- 0.77	1 age
0.216 +- 0.017	5.5 +- 1.12	13 thal
.0.195 +- 0.022	7.1 +- 0.83	12 ca
0.19 +- 0.017	7.5 +- 0.67	4 trestbps
0.147 +- 0.014	9.3 +- 0.46	9 exang
0.141 +- 0.008	9.7 +- 0.46	11 slope
0.075 +- 0.01	11 +- 0	2 sex
0.036 +- 0.005	12 +- 0	7 restecg
0.003 +- 0.001	13 +- 0	6 fbs

Table 5 presents the result obtained from Chi-Square Feature selection method. The features with rank 0, i.e., chol, thalach, sex, restecg, fbs have removed.

Table 5: Result obtained by Chi-Square Feature Selection Method

Average Merit	Average Rank	Attribute name
459.807 +-84.083	1 +- 0	5 chol
284.386 +-55.823	2 +- 0	8 thalach
89.128 +-10.471	3.4 +- 0.49	1 age
86.108 +- 7.402	3.6 +- 0.49	10 oldpeak
66.103 +- 7.559	5.9 +- 0.94	4 trestbps
65.876 +- 5.532	6.4 +- 1.28	3 cp
65.249 +- 7.466	6.4 +- 0.92	12 ca
63.823 +- 4.639	7.3 +- 0.78	13 thal
44.131 +- 3.899	9.2 +- 0.4	9 exang
41.778 +- 2.366	9.8 +- 0.4	11 slope
22.094 +- 2.801	11 +- 0	2 sex
10.235 +- 1.501	12 +- 0	7 restecg
0.692 +- 0.278	13 +- 0	6 fbs

Table 6 gives the result obtained by proposed TFIE-PSFS method. This method also removes the features with rank 0. The features like sex, restecg, thalach, age, fbs, trestbps have removed for obtaining the optimal dataset.

Table 6: Result obtained by proposed TFIE-PSFS Method

Average Merit	Average Rank	Attribute name
0.025 +- 0.007	9.5 +- 1.36	3 cp
0.014 +- 0.009	11.5 +- 1.43	12 ca
0.126 +- 0.017	3.3 +- 0.46	13 thal
0.107 +- 0.014	4.1 +- 0	2 sex
0.101 +- 0.012	5.1 +- 1.04	11 slope
0.082 +- 0.011	5.8 +- 0.6	9 exang
0.068 +- 0.011	6.7 +- 0	7 restecg

0.025 +- 0.007	8.8 +- 0	8 thalach
0.19 +- 0.022	1.1 +- 0.3	10 oldpeak
0.025 +- 0.008	9.6 +-0	1 age
0.015 +- 0.006	11.2 +- 0	6 fbs
0.171 +- 0.016	1.9 +- 0	4 trestbps
0.01 +- 0.003	12.4 +- 0.66	5 chol

Table 7 represents the number of features that are selected by Symmetrical Uncertainty, Information Gain, Chi-Square and Proposed Thrice Filtered Information Energy based Particle Swarm Feature Selection methods. From the table 7, the proposed TFIE-PSFS method gives the less number of features than the existing feature selection methods.

Table 7: Number of Features selected by SU, IG, CS and proposed TFIE-PSFS method

Feature Index	<i>Number of Features selected by existing feature selection methods and proposed TFIE-PSFS method</i>			
	SU	IG	CS	TFIE-PSFS
1	chol	cp	age	cp
2	cp	age	old peak	ca
3	exang	thal	trestbps	thal
4	ca	ca	cp	slope
5	slope	trestbps	ca	exang
6	old peak	exang	thal	chol
7	sex	slope	exang	
8	age		slope	
9	trestbps			

5. Conclusion

The proposed TFIE-PSFS method gives only six features among 13 features of heart disease dataset. Symmetrical Uncertainty feature selection method gives nine features, Information Gain feature selection method gives seven features, Chi-Square feature selection method gives eight features among 13 features. Analyzing the result obtained by these methods, Proposed TFIE-PSFS gives the least number of features than the existing feature selection methods in the prediction of heart disease. In future work, the various classification algorithms can be used to evaluate the proposed TFIE-PSFS method regarding Classification Accuracy, Error rates, Kappa Statistics, Precision, Recall and F-Measure.

References

- [1] Broder, Michael S., et al. "An Approach to Using Data Mining to Support Early Identification of Acromegaly." *Endocrine Practice* 23.4 (2017): 422-431.
- [2] Kaffash, Sepideh, and Marianna Marra. "Data envelopment analysis in financial services: a citations network analysis of banks, insurance companies and money market funds." *Annals of Operations Research* 253.1 (2017): 307-344.
- [3] Platis, Charalampos, and Emmanouil Zoulias. "Leadership Capabilities and their Effect on Job Performance, an Approach in Healthcare Sector." *Strategic Innovative Marketing*. Springer, Cham, 2017. 365-371.
- [4] Kavakiotis, Ioannis, et al. "Machine learning and data mining methods in diabetes research." *Computational and structural biotechnology journal* 15 (2017): 104-116.
- [5] Mohr, David C., Mi Zhang, and Stephen M. Schueller. "Personal sensing: understanding mental health using ubiquitous sensors and machine learning." *Annual review of clinical psychology* 13 (2017): 23-47.
- [6] Liem, David A., et al. "Phrase Mining and Machine Learning in Textual Data to Uncover Distinct Protein Patterns in Cardiovascular Disease." *Journal of Molecular and Cellular Cardiology* 112 (2017): 168-169.

- [7] Zheng, Tao, et al. "A machine learning-based framework to identify type 2 diabetes through electronic health records." *International journal of medical informatics* 97 (2017): 120-127.
- [8] Zheng, Tao, and Ya Zhang. "A Big Data Application of Machine Learning-Based Framework to Identify Type 2 Diabetes Through Electronic Health Records." *International Conference on Knowledge Management in Organizations*. Springer, Cham, 2017.
- [9] Feldman, Keith, et al. "Beyond Volume: The Impact of Complex Healthcare Data on the Machine Learning Pipeline." *Towards Integrative Machine Learning and Knowledge Extraction*. Springer, Cham, 2017. 150-169.
- [10] Manogaran, Gunasekaran, and Daphne Lopez. "A survey of big data architectures and machine learning algorithms in healthcare." *International Journal of Biomedical Engineering and Technology* 25.2-4 (2017): 182-211.
- [11] Tayefi, Maryam, et al. "hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm." *Computer methods and programs in biomedicine* 141 (2017): 105-109.
- [12] Gefen, David, et al. "Identifying patterns in medical records through latent semantic analysis." *Communications of the ACM* 61.6 (2018): 72-77.
- [13] Alizadehsani, Roohallah, et al. "Non-invasive Detection of Coronary Artery Disease in High-Risk Patients Based on the Stenosis Prediction of Separate Coronary Arteries." *Computer Methods and Programs in Biomedicine* (2018).
- [14] <https://archive.ics.uci.edu/ml/datasets.html>
- [15] Trambaiolli, L. R., et al. "Feature selection before EEG classification supports the diagnosis of Alzheimer's disease." *Clinical Neurophysiology* 128.10 (2017): 2058-2067.
- [16] Yuvaraj, N., and K. R. SriPreethaa. "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster." *Cluster Computing* (2017): 1-9.

- [17] Kuang, Sicong, and Brian D. Davison. "Learning Word Embeddings with Chi-Square Weights for Healthcare Tweet Classification." *Applied Sciences* 7.8 (2017): 846.
- [18] Jabbar, M. A. "Prediction of heart disease using k-nearest neighbor and particle swarm optimization." *Biomedical Research* 28.9 (2017).

