**Shouman: Prototype Development of a Novel Heart Disease Risk Evaluation Tool Using Data Mining Analysis**

<http://unsworks.unsw.edu.au/fapi/datastream/unsworks:12635/SOURCE02?view=true>

Heart disease leading cause of death, but preventable / controllable, over 7MM deaths in 2013, 80% could be prevented (WHO in 2013)

Coronary heart failure, hardened / narrow due to plaque (US DHHS 2005)

Heart failure doesn't mean stops, means body need for blood / oxygen not met (American Heart Association 2011), occurs when excess fluid collect in body as results of heart weakness, leads to build-up of fluid in lungs, sw3elling feet, tired, weak, breathing difficulty (US DHHS 2005)

Stroke, blood clot leaving hear and lodging in brain (American Heart Association 2013)

Early detection for health recovery / decreased mortality (per CDC 2013)

Growth in disease-related deaths coming from developing world

Need acrruate tools for early detection (Paladugu and Shyu 2010)

ECG / stress test / cartdiac angiogram expenseive and invasive, not good community -level screening (i.e. in pharmacies / public health clinics)

Farmingham Risk Eval Tool / Australian Abs Cardio Risk Calculator two common heart disease risk eval screens, but need prior blood samples, invasive / costly

Motivated by the increasing mortality rates of heart disease, researchers have been using data mining techniques to help healthcare professionals in the diagnosis of heart disease patients (Parthiban and Subramanian 2007, Polat , Sahan et al. 2007, Tu, Shin et al. 2009, Rajkumar and Reena 2010, Lakshmi, Krishna et al. 2013)

benchmark dataset and new larger dataset attributes

reliability of non-invasive attributes for diagnosis

enhancement of hybrid data mining of non-invasive attributes (K-means clustering, different initial centroid selection methods, with decision tree mining)

risk eval tool / expert system

Different results from researchers using different datasets and classification techniques: Kernel Density, Neural Network Automatically Defined Groups, Bagging Algorithm, Sequential Minimal Optimization, Direct Kernel Self-Organizing Map, and Support Vector Machine. Ranging from high 50 to mid-90 accuracy.

Exhaustive literature review between 2000 and 2016

Most common approaches are Decision Tree (Cleveland 70s based on type of tree / discretization method), Naïve Bayes (Cleveland 80s based on discretization method) and KNN

|  |  |  |
| --- | --- | --- |
| Author/Year | Technique | Accuracy |
| (Hall 2000) | Naïve Bayes | 83.24% |
|  | K Nearest Neighbour | 82.12% |
|  | Decision Tree | 75.32% |
| (Yan, Zheng et al. 2003) | Multilayer Perceptron | 63.6% |
| (Herron 2004) | Support Vector Machine | 83.6% |
|  | J4.8 Decision Tree | 77.56% |
|  | Naïve Bayes | 83.37% |
| (Andreeva 2006) | Naïve Bayes | 78.56% |
|  | Decision Tree |  |
|  | Neural network | 82.77% |
|  | Sequential minimal optimization | 84.07% |
|  | Kernel density | 84.44% |
| (Polat , Sahan et al. 2007) | Fuzzy-AIRS–k-nearest neighbour | 87% |
| (Palaniappan and Awang 2007) | Naïve Bayes | 95% |
|  | Decision Trees | 94.93% |
|  | Neural Network | 93.54% |
| (De Beule, Maesa et al. 2007) | Artificial neural network | 82% |
| (Tantimongcolwat, Naenna et al. 2008) | Direct kernel self-organizing map | 80.4% |
|  | Multilayer Perceptron | 74.5% |
| (Hara and Ichimura 2008) | Automatically Defined Groups | 67.8% |
|  | Immune Multi-agent Neural Network | 82.3% |
| (Sitar-Taut, Zdrenghea et al. 2009) | Naïve Bayes | 62.03% |
|  | Decision Trees | 60.40% |
| (Tu, Shin et al. 2009) | Bagging algorithm | 81.41% |
| (Das, Turkoglu et al. 2009) | Neural network ensembles | 89.01% |
| (Rajkumar and Reena 2010) | Naive Bayes | 52.33% |
|  | K nearest neighbour | 45.67% |
|  | Decision list | 52% |
| (Srinivas, Rani et al. 2010) | Naïve Bayes | 84.14% |
|  | One Dependency Augmented Naïve Bayes | 80.46% |
| (Kangwanariyakul, Nantasenamat et al. 2010) | Back-propagation neural network | 78.43% |
|  | Bayesian neural network | 78.43% |
|  | Probabilistic neural network | 70.59% |
|  | Linear support vector machine | 74.51% |
|  | Polynomial support vector machine | 70.59% |
|  | Radial basis function support vector machine | 60.78% |
| (Kumari and Godara 2011) | RIPPER | 81.08% |
|  | Decision Tree | 79.05% |
|  | Artificial Neural Network | 80.06% |
|  | Support Vector Machine | 84.12% |
| (Soni, Ansari et al. 2011) | Weighted Associative Classifier | 57.75% |
|  | Classification based on Association Rule (CBA) | 58.28% |
|  | Classification based on Multiple ClassAssociation Rules (CMAR) | 53.64% |
|  | Classification based on Predictive Association Rules (CPAR) | 52.32% |
| (Abdullah and Rajalaxmi 2012) | Decision Tree | 50.67 % |
|  | Random Forest | 63.33 % |
| (Rajeswari, Vaithiyanathan et al. 2013) | Neural Network | 80.53% |
|  | J4.8 Decision Tree | 77.89% |
|  | Support Vector Machine | 84.16% |
|  | Feature Selection with Neural Network | 84.49% |
|  | Feature Selection with Decision Tree | 84.16% |
|  | Feature Selection with Support Vector Machine | 87.46% |
| (Lakshmi, Krishna et al. 2013) | Support Vector Machine | 78.10 % |
|  | Decision Tree | 84.68 % |
|  | K Nearest Neighbour | 83.95 % |
|  | K mean | 80.29 % |
| (Pandey, Pandey et al. 2013) | COBWEB | 1.98% |
|  | EM | 81.51% |
|  | Farthest First | 73.59% |
|  | Make Density Based Clusters | 81.51% |
|  | Simple K-Means | 80.85% |
|  |  |  |

Just based on Cleveland Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Author/Year | Technique | Accuracy |
| Single | (Hall 2000) | K-Nearest Neighbour | 82.12% |
|  | (Cheung 2001) | Decision Tree | 81.11% |
|  |  | Naïve Bayes | 81.48% |
|  | (Tu, Shin et al. 2009) | J4.8 Decision Tree | 78.9% |
|  |  | Bagging algorithm | 81.41% |
| Hybrid | (Polat , Sahan et al. 2007) | Fuzzy-AIRS – K-Nearest  Neighbour | 87% |
|  | (Parthiban and Subramanian 2007) | Coactive Neuro-Fuzzy Inference System | Mean Square Error is  0.000842 |
|  | (Das, Turkoglu et al. 2009) | Neural Network Ensembles | 89.01% |
|  | (Ozsen and Gunes 2009 ) | Genetic Algorithms with  Artificial Immune System | 87% |
|  | (Rajeswari, Vaithiyanathan et al. 2013) | Feature Selection with Neural Network | 84.49% |
|  |  | Feature Selection with Decision Tree | 84.16% |
|  |  | Feature Selection with Support Vector Machine | 87.46% |
|  |  |  |  |
|  |  |  |  |

Takeaways: single data mining techniques implemented over different datasets show different levels of accuracies, and integrating more than one technique helps diagnostic accuracy

**Assari et al: Heart Disease Diagnosis Using Data Mining Techniques**

<https://www.omicsonline.org/open-access/heart-disease-diagnosis-using-data-mining-techniques-2162-6359-1000415.pdf>

Selected data mining techniques were applied on the studied dataset after dataset discretization. 10-fold cross-validation method was used to validate the results. This technique classifies dataset into 10 portions. 9 portions were used for training the algorithm and 1 portion was used for evaluation in each run-time. The process was repeated 10 times. This procedure is helpful specifically in datasets with small number of samples by prevention of over-fitting. Finally, the sensitivity, specificity and accuracy of each method were calculated.

Decision tree, Naïve Bayes, K-Nearest Neighbor and Support Vector Machine were applied to the studied dataset. Table 2 shows the sensitivity, specificity and accuracy of these data mining techniques. Their accuracy ranges from 79% to 84.33%. According to Table 2, SVM achieved the highest accuracy (84.33%).

SVM and Naïve Bayes achieved the highest accuracy, followed by KNN (k=7 resulted in the best accuracy as compared to other values) and decision tree, respectively. Weka and IBM SPSS Modeler were used to implement data mining techniques.

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Accuracy | Specificity | Sensitivity |
| Decision Tree | 0.79 | 0.8148 | 0.7608 |
| Naïve Bayes | 0.8366 | 0.8641 | 0.8043 |
| KNN K=7 | 0.81 | 0.8395 | 0.7753 |
| SVM | 0.8433 | 0.895 | 0.7826 |
|  |  |  |  |

Found that

chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal, 4: asymptomatic)

thalium stress test (3: normal, 6: fixed defect, 7: reversable defect)

coronary artery disease (number of major vessels colored by fluoroscopy, 1-3)

Naïve Bayes: Ca, Thal, Cp, and Oldpeak most important

KJNN K=7: no standouts (all tight)

SVM: Ca, Thal, Cp most important

Decision Tree: Thal, Cp, Ca

Avg: Thal, Ca, Cp

Sabay et al: Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data

<https://scholar.smu.edu/cgi/viewcontent.cgi?article=1038&context=datasciencereview>

Medical datasets often small, use logistic regression / decision trees

Train / test partitions need bigger data for models like NN, else not generalizable

Create synthetic 50k ob dataset based on characteristics of cleveland data using Synthpop

Test / train NN model using Keras API for Python

Best classification prediction accuracy / stability, mirrors Cleveland characteristics, 16% more accurate than logistic

Focus not just on ML algos for prediction, consider techniques that could be applied to dataset used by model to improve accuracy / stability of prediction

Can augment size, given cost / complexity of patient data

Mitigate risk of exposure of PHI

A novel solution for satisfying data volume requirements for machine learning models was developed by Torgyn Shaikhina and Natalia A. Khovanova in a 2016 publication titled Handling Limited Datasets with Neural Networks in Medical Applications: A Small-Data Approach [3]. In this research, a framework for generating surrogate data from small data sets (as small as ten observations) was developed and validated using neural network techniques. This technique utilizes multiple runs of 2000 neural networks in order to generate robust data sets that mimic the characteristics of the real data set and provides adequate data volumes to satisfy modern machine learning based prediction. The number of neurons required for this technique requires large computing resources therefore, in this experiment, we chose an alternative solution for surrogate data generation.

Synthpop has comparison tools against seed data that Simpop and boot packages do not

projecting and fitting linear/logistic modules using the original data and by implementing CART

Table 3. UCI Repository for Machine Learning Heart Disease Databases Database Donor Author Instances Cleveland Cleveland Clinic Foundation Robert Detrano, M.D., Ph.D. 303 Hungarian Hungarian Institute of Cardiology, Budapest Andras Janosi, M.D. 294 Switzerland University Hospital, Zurich, Switzerland William Steinbrunn, M.D. 123 Long Beach V.A. Medical Center, Long Beach Robert Detrano, M.D., Ph.D. 200

The four Heart Disease data sets are available in raw and processed formats2 . The raw format is space delimited text and where there is missing data, a ”-9” was inserted. The Cleveland raw data set contains 76 attributes, but half (38) of the attributes are not usable due to missing or undefined data. The number of final usable records from the raw data set after the data cleanup was 282 records with 38 attributes. The processed data (the Cleveland 14 data set) is a comma delimited text file and has been reduced to 14 attributes as a result of past research [7]. In the processed databases, the missing data encoding (-9) still exists. Of the four processed databases, the Cleveland 14 data is the most complete with only 6 records of missing data making the total usable size 297 records with 14 attributes.

stage 1

best accuracy performance from logistic (83-88), random forest(77-81), then decision tree (74-79)

recall for logistic has 7 poiunt variation, ranomd forest has 11 (important for false negative)

Table 5. Stage 1 Machine Learning Metric Results Machine Learning Model Accuracy Precision Recall Logistic Regression 83 - 88% 82 - 92% 75 - 82% Decision Tree 74 - 79% 71 - 75% 69 - 77% Random Forest 77 - 81% 74 - 84% 70 - 81%

stage 2

50k surrogate w/ logistic

logistic has similar performance

slightly lower accuracy / prevision, but similar recall

much narrower ranges due to bigger data range

stage 3

60k surrogate w/ keras deep learning

ANN perceptron

accuracy / prevision / recall around 96, 1 percent variation

El-Bialy et al: Feature Analysis of Coronary Artery Heart Disease Data Sets

Data sets dealing with the same medical problems like Coronary artery disease (CAD) may show different results when applying the same machine learning technique. The classification accuracy results and the selected important features are based mainly on the efficiency of the medical diagnosis and analysis. The aim of this work is to apply an integration of the results

of the machine learning analysis applied on different data sets targeting the CAD disease. This will avoid the missing, incorrect, and inconsistent data problems that may appear in the data collection. Fast decision tree and pruned C4.5 tree are applied where the resulted trees are extracted from different data sets and compared. Common features among these data sets are extracted and used in the later analysis for the same disease in any data set. The results show that the classification accuracy of the collected dataset is 78.06% higher than the average of the classification accuracy of all separate datasets which is 75.48%.