Heart disease recognition AI

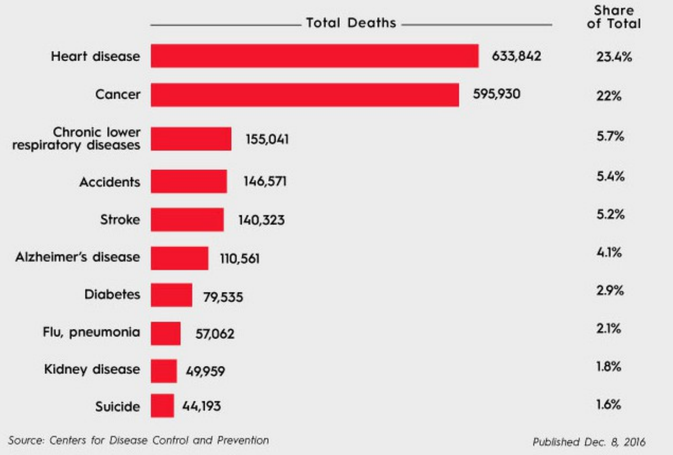
Kristina Grujić
Department of computing and control engineering
Faculty of technical sciences
Novi Sad, Serbia

***Abstract*—At current times, information technology is developing faster than ever, and becoming present in every industry and each field of human interest. At same time, heart diseases are amongst the deadliest and most common diseases affecting human kind, being the leading cause of death for both men and women, with people of all ages and backgrounds being able to get a condition. If the disease is recognized at early stage, fatal outcome mostly can be prevented. Even so, recognizing the problem isn't always easy. Informat ion technology comes here into play: we can now make an Artificial Intelligence (AI) software that'd be able to recognize the disease or risk of disease. This document's purpose is to analyze various algorithms that can be used for implementing this type of software, and compare algorithm's implementation results, with goal to understand what algorithm is the best for this problem.**

Keywords*—AI; machine learning; classification algorithms; algorithms; health;

I. Introduction

Heart disease is the leading cause of death in the United States [1], and is in top 10 causes of death of world population. High blood pressure, high cholesterol, smoking are some of the key risk factors for heart disease. As with some other diseases, recognition of heart disease presence or understanding the possibility of developing a heart disease can help preventing the fatal outcome and increase the chance of curing it. Still, doctors can't always recognize the potential patient at early stages, but only after the disease has developed, which decreases the chance for healing.



Source: Centers for Disease Control and Prevention

Published Dec. 8, 2016

With current progress of information technology development, it is possible to, based on data of previous patients, make a software able to understand the risk of getting a deadly condition, or be able to diagnose it at early stage, if having enough informations of the potential patient.

In this document, we're using the Heart disease dataset from UCI [2] to analyze various classification algorithms and their efficiency in recognizing the type of disease present in patient. In chapter [two] we will first analyze the dataset we're using. In chapter [three], we will briefly describe the algorithms used in the analysis. Finally, in chapter [four] we will take a look at the results of these algorithms and analyze them, deciding which one did the best job in predicting and classifying the diseases.

II. Dataset brief analysis

The dataset used contains 4 databases, all of them which were used in the project for algorithms evaluation:

1. Cleveland Clinic Foundation database
2. V.A. Clinical centre database
3. Swiss university hospital database
4. Hungarian institute of cardiology

All of these three databases have the same structure. Although they are originally made of 76 attributes, all numeric valued, only 14 that were used by all of the published experiments so far have been used in this analysis. Only Cleveland database has been used by published experiments so far. In this analysis all of the four databases listed above were used to evaluate algorithm efficiencies.

Attribute information:

1. age of patient
2. sex of patient
3. chest pain type
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar (>120mg/dl)
7. resting electrocardiographic results

8. maximum heart rate achieved
9. exercise included angina
10. ST depression induced by exercise relative to rest
11. the slope of peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal
14. diagnosis of heart disease

First 13 of the attributes in databases describe patient's data : it's age, sex, pressure rate, blood sugar levels, cholesterol levels etc. The last, 14th attribute describes the type of disease, with 0 describing that no heart disease is present, and values 1-4 describing different types of diseases. In the past experiments and researches, the goal was mostly only to make the AI which predicts if the disease is present, ignoring the value types of different diseases, so no prediction of disease type was done. In this document, we'll be analyzing algorithms efficiency also when distinguishing the type of disease, and not only when recognizing wether the patient is healthy or not. The best results were gotten when using first 9 attributes to analyze the data and make predictions.

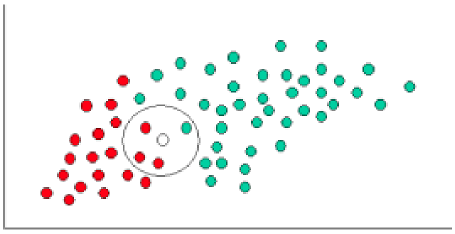
III. Algorithms used

A. *Support vector clustering (SVC)*

SVC is method similar to support vector machines learning model, used for classification and regression analysis. SVC is considered a fundamental method in data science, and is appropriate for unsupervised learning and data-mining. SVM and SVC can perform linear and non-linear classifications efficiently, operating with kernels (class algorithms for pattern analysis). The idea of SVC and SVM is to, supposing that some given data points belong to one of two classes, decide successfully what class a new data point will be in. Algorithm tries to find the hyperplanes so the distance from it to the nearest data point on each side is maximized, and then, based on the position of the new data point, classify it. In analysis of our dataset, we tested SVC using linear and radial basis function kernels (RBF).

B. *Naive Bayes*

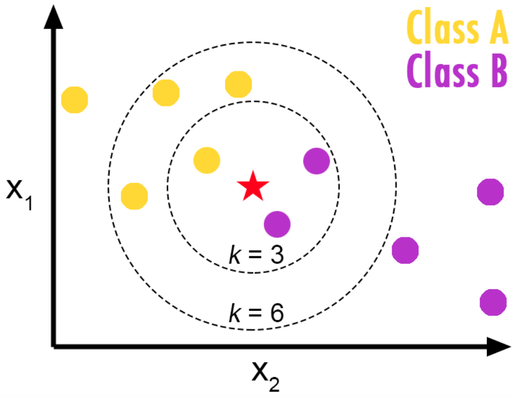
Naive Bayes classifiers are a family of probabilistic classifying algorithms. These classifiers are simple, yet proved to be very efficient in classifying the data.



The idea of Naive Bayes can be explained with green and red dots, and a new white dot not yet classified. In certain radius around the new dot, number of green and red ones is summed. Based on these numbers, and total numbers of red and green dots already classified, the white dot is classified as green or red one. For implementation, various distributions can be used. In this experiment, Gaussian, Bernoullian and multinomial distributions were used. These were applied both when just trying to predict the presence and the actual type of health disease.

C. *K-nearest neighbors*

K-nearest neighbors is a non-parametric algorithm used for classification and regression. The idea of algorithm is to, based on k closest data points, determine the class of the new one. For this algorithm, various metrics can be used in implementation. The metrics we used are Euclidean, Manhattan and Chebyshev, with trying out different values of k until finding the one working the best for each of these metrics. The best results on this certain dataset was given with k value being 3 for Euclidean and Chebyshev metrics, and k value that gave the best results using Manhattan metrics was 4.



D. *Logistic regression*

Logistic regression is a regression model with categorical dependent variable. It's used in many fields, and being very popular in medical fields, using to predict wether a patient has certain disease. Logistic regression can be binomial, ordinal or multinomial. It is analogous to linear regression, but the model is based on different assumptions. In this analysis, two different implementations were used : multinomial logistic regression based on Newton conjugate gradient algorithm, and using the linear classifier.

IV. Result analysis

After implementing the algorithms and their variants described above, we came to next results. Recognition if disease is present has way more accurate results than recognition of the exact type of disease. This big result difference was expected, as the attributes that the databases consist of don't have informations significant enough to differ between certain types and make a proper diagnose.

	Cleveland	Swiss	VA	Hungarian
Linear SVC	0.6	0.41	0.3	0.79
RBF SVC	0.61	0.54	0.26	0.64
Gaussian NB	0.36	0.12	0.12	0.77
Bernoulli NB	0.59	0.38	0.18	0.77
Multinomial NB	0.55	0.48	0.2	0.7
Euclidean KNN	0.56	0.48	0.34	0.64
Manhattan KNN	0.6	0.29	0.32	0.68
Chebyshev KNN	0.52	0.41	0.3	0.64
Linear regression	0.65	0.45	0.32	0.75
Newton-CG LR	0.57	0.45	0.3	0.75

In the table above, the results of algorithms used on recognizing the type of disease are displayed, for each of the dataset databases separately. Green font algorithm was the one that provided best result,the logistic regression implemented with linear algorithm. Second best result, highlighted with orange color, which is very similar to best one, was result of logistic regression implemented with Newton conjugate gradient algorithm.

	Cleveland	Swiss	VA	Hungarian
Linear SVC	0.78	1	0.74	0.75
RBF SVC	0.63	1	0.78	0.64
Gaussian NB	0.82	0.29	0.76	0.71
Bernoulli NB	0.72	0.9	0.76	0.74
Multinomial NB	0.73	0.93	0.66	0.62
Euclidean KNN	0.69	1	0.78	0.64
Manhattan KNN	0.73	0.96	0.68	0.7
Chebyshev KNN	0.68	1	0.78	0.64
Linear regression	0.77	0.93	0.76	0.74
Newton-CG LR	0.77	0.93	0.76	0.71

The worst result was the result of Naive Bayes implemented with Gaussian distribution, highlighted in red.

In table above, the results of disease presence recognition are shown. Here, once again, we can see, marked in green, algorithms that provided best results, which, in this case, are linear SVC implementation and linear logistic regression, which have really similar results. On second place is Newton conjugate gradient algorithm, with again pretty similar result as algorithms with best results. The worst prediction was made by multinomial Naive Bayes. The absolute worst result in prediction was given by Gaussian Naive Bayes algorithm when working on Switzerland's database.

First 10 columns of files were found to be the most important in properly predicting the condition of patient. Best recognition percentage of disease type is around 55% (average for all four databases), using logistic regression based on linear algorithm. Recognition of presence of disease is at average of 80% recognition success, with best recognition results we got from logical regression, and then k-nearest neighbors algorithms.

Conclusion

Based on results described in last chapter, we can see that logistic regression indeed works really well for both prediction of presence and diagnosing the type of heart diseases, as both linear and Newton-CG algorithms provided best results out of all tested algorithms. Even so, most of the algorithms did pretty well and provided pretty similar results, proving that even a bit less complex algorithms are proficient in classification. Based on results from prediction of type of disease, we can understand that the problem is a bit more complex here, and needs more detailed database structure for more accurate prediction.

References

[1] CDC, NCHS. Underlying Cause of Death 1999-2013 on CDC WONDER Online Database, released 2015.

[2] http://archive.ics.uci.edu/ml/index.php