HEART DISEASE DATA SET CLASSIFICATIONS: COMPARISONS OF CORRELATION CO EFFICIENT BY APPLYING VARIOUS FUNCTIONS

Dr.G.Ayyappan,
Associate Professor, Bharath Institute of Higher Education and Research,Chennai
ayyappangmca@gmail.com
K.SivaKumar,
SIPS Technologies, Chennai, India.

Abstract: Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential used in various commercial applications including retail sales, e-commerce, remote sensing, bioinformatics etc. There are varieties of popular data mining task within the educational data mining e.g. classification, clustering, outlier detection, association rule, prediction etc. This paper focuses the comparisions of various correlation coeffient accuracies by applying the different paprameters pruning methods and analysis in weka tool.

Keywords: Linear Regression, Multi Layer Perception, Gaussian Processes, Simple Linear Regression, SMOreg.

1. INTRODUCTION

Now a days, large quantities of data is being accumulated. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. Data could be large in two senses: in terms of size & in terms of dimensionality. Also there is a huge gap from the stored data to the knowledge that could be construed from the data.

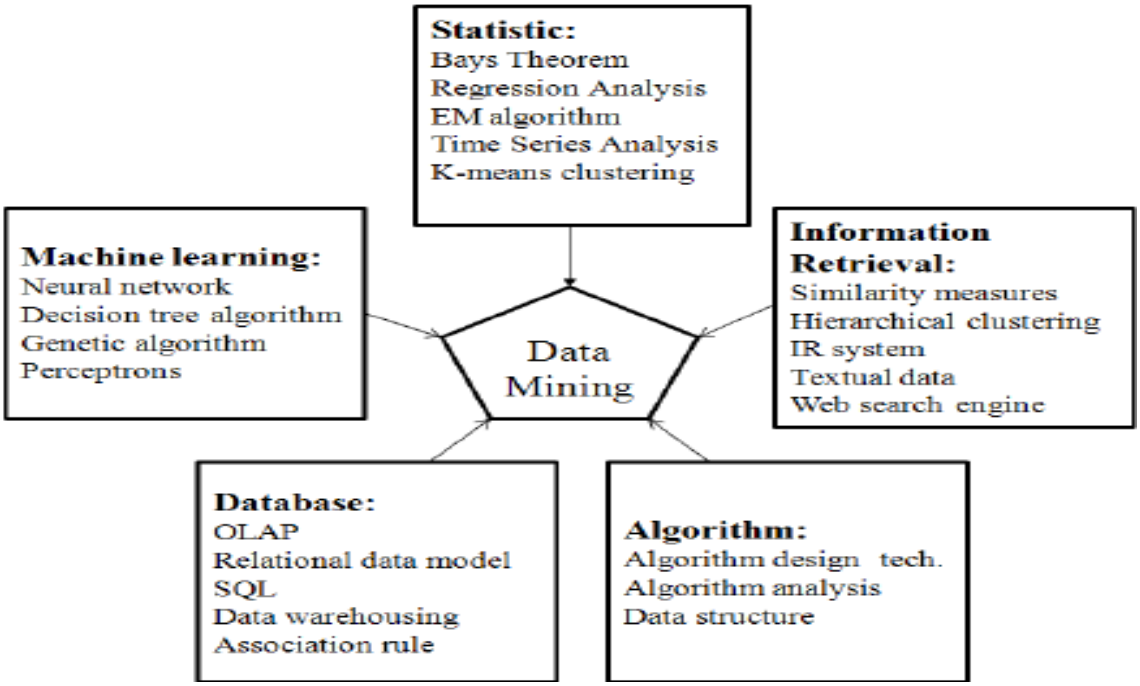

Fig 1:Data Mining

Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis. The transition won't occur automatically; in this case, we need the data mining. Data Mining could help in a more in-depth knowledge about the data.

Extreme learning machine (ELM) is a special single-hidden layer feed-forward neural network (SLFN). Due to its lower computational complexity and better generalization performance, ELM has recently attracted a lot of interests in research and industry and is used in a wide range of applications. ELM uses a random method to determine input weights/hidden layer biases and analytically computes the output weights. Therefore, it is extremely fast to train an ELM model. It has also been proved that ELM can guarantee the universal approximate capability of ELM .

In this paper section 1 focuses, introduction about the machine learning, Section 2 presents materials and methods, In section 3 presents results and discussions and finally Section 3 presents conclusion about this research work.

2. MATERIALS AND METHODS

In this section presents the materials and methods of this research work. Here it has implemented the weka tool for mining process The WEKA GUI chooser launches the WEKA's graphical environment which has five buttons: Simple CLI, Explorer, Experimenter, Knowledge Flow and Workbench.

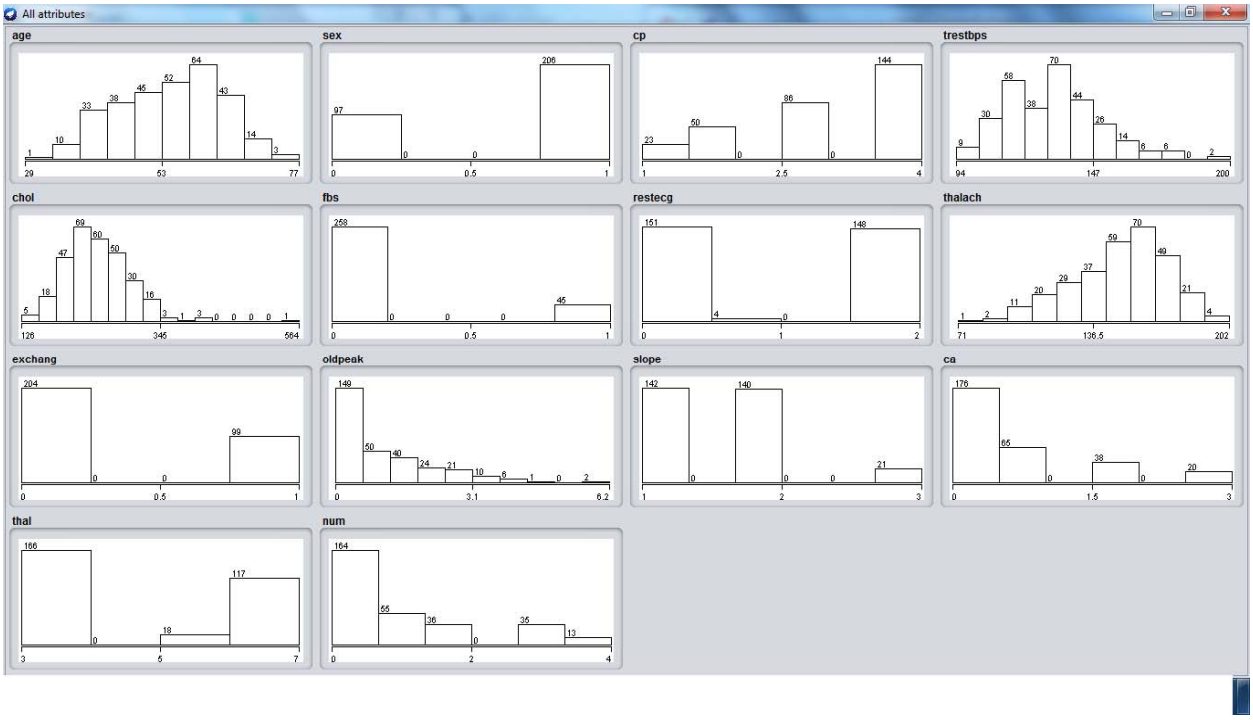

Fig 2:DataVisualization

Dataset Information

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

Table 1: Dataset Information

Name	Data type category	No of attributes	No of Instances	No of Attributes
Heart Disease	Multivaiate	Categorical, Integer, and Real	303	75

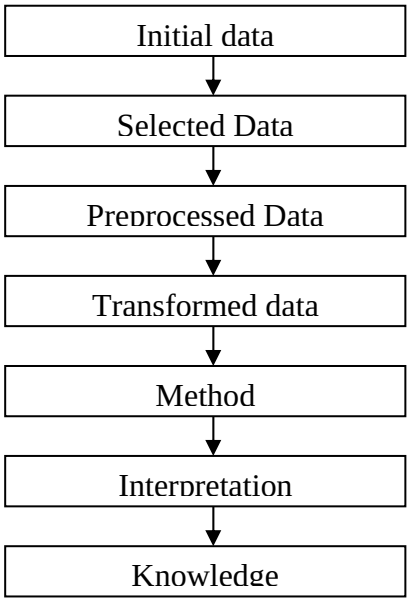

Fig 3:Flow process of Heart Diseas dataset classification

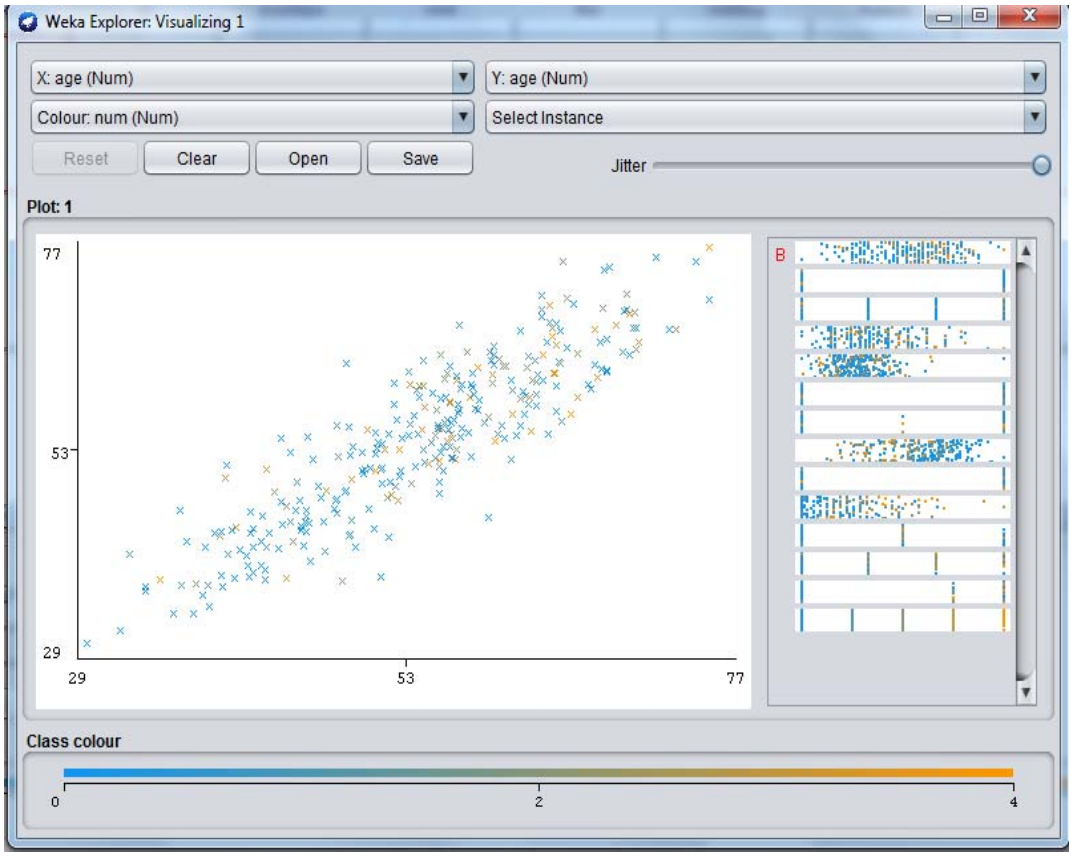

Fig 4:Data visualization based on age of Heart Diseas dataset classification

Table 2: List of Attributes

S.No	Label	Meaning and functions of Lebel
1	age	age in years
2	sex	Sex (1=male ; 0=female)
3	cp	chest pain type
4	trestbps	resting blood pressure(in mm Hg on admission to the hospital)
5	chol	serum cholestoral in mg/dl
6	fbs	(fasting blood sugar>120 mg/dl) (1=true; 0=false)
7	restecg	resting electrocar diographic results
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina(1=yes; 0=no)
10	oldpeak	ST depreve to restse relatission induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment
12	ca	number of major vessels(0-3)colored by flourosopy
13	thal	3=normal; 6=fixed defect; 7=reversable defect
14	num	diagnosis of heart disease (angiographic disease status)

The above 14 data attributes are processed and it has derived from 76 attributes.

In this research paper implements the various functions. They are

Functions Name	Functions
Linear Regression	Class for using linear regression for prediction. Uses the Akaike criterion for model selection, and is able to deal with weighted instances.
Multi Layer Perceptron	A classifier that uses back propagation to learn a multi-layer perceptron to classify instances. The network can be built by hand or set up using a simple heuristic. The network parameters can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric, in which case the output nodes become unthresholded linear units).
Gaussian Processes	Implements Gaussian processes for regression without hyperparameter-tuning. To make choosing an appropriate noise level easier, this implementation applies normalization/standardization to the target attribute as well as the other attributes (if normalization/standardizaton is turned on). Missing values are replaced by the global mean/mode. Nominal attributes are converted to binary ones. Note that kernel caching is turned off if the kernel used implements CachedKernel.
Simple Linear Regression	Learns a simple linear regression model. Picks the attribute that results in the lowest squared error. Can only deal with numeric attributes.
SMOreg	SMOreg implements the support vector machine for regression. The parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer. The most popular algorithm (RegSMOImproved) is due to Shevade, Keerthi et al and this is the default RegOptimizer.

3. RESULTS AND DISCUSSIONS

In this section clearly described the results of applying the various parameters in Gaussian processes and it produces the various co efficient accuracies and also it mentioned the time taken to build the models.

Table 2: Various Correlation coefficient accuracies of Gaussian processes

Functions	Accuracy of Correlation Coefficient	Time taken to build the model (In Seconds)
Linear Regression	0.5189	0.36
Multi Layer Perception	0.3071	1.61
Gaussian Processes	0.5145	0.36
Simple Linear Regression	0.3834	0.01
SMOreg	0.5156	0.22

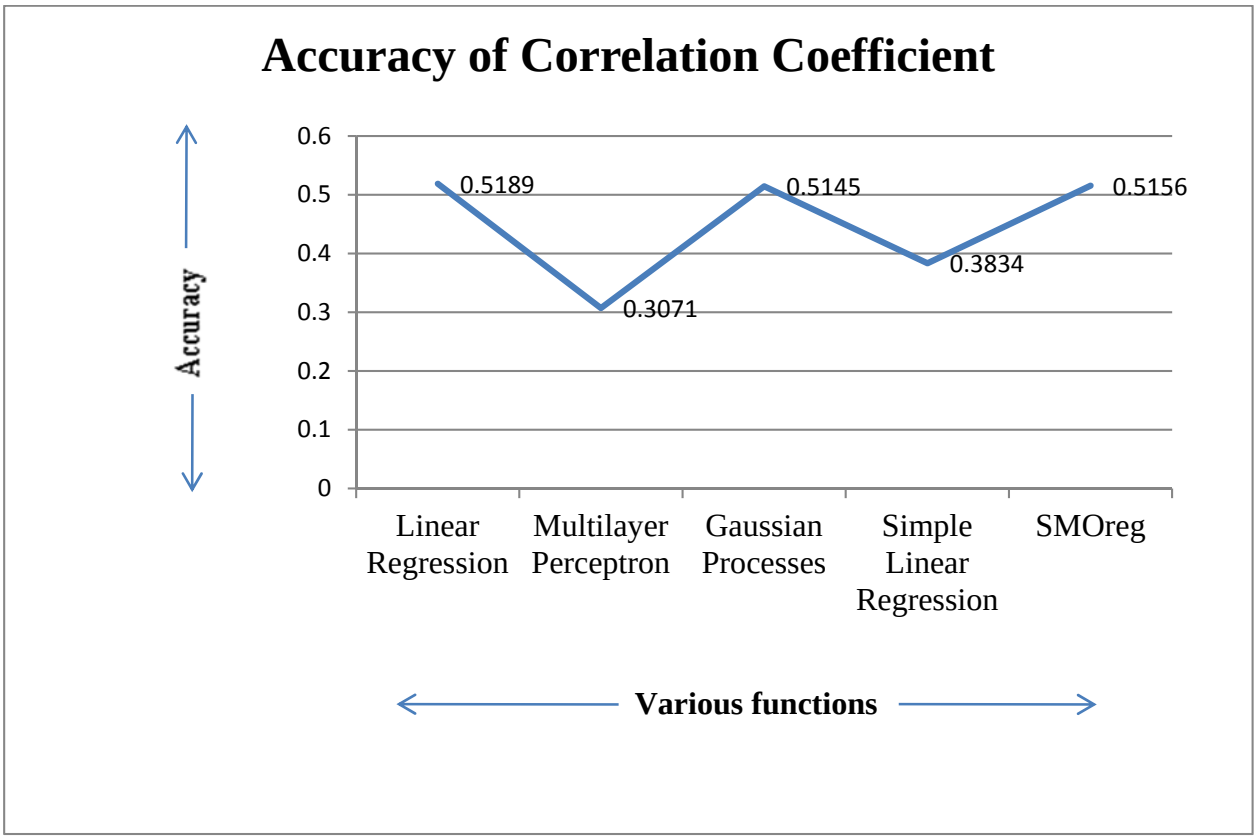

Fig 4: Various Co Efficient Accuracies

This above diagram represents the various kernel parameters and the models produces the various correlation coefficients in gaussians processes.The model produces the correlation coefficent accuracy level 0.5189 while applying linear regression in this data set and time taken to be build this model 0.36 seconds. The model produces the correlation coefficent accuracy level 0.3071 while applying Multilayer perceptron in this data set and time taken to be build this model 1.61 seconds. The model produces the correlation coefficent accuracy level 0.5145 while applying Gaussian Processes in this data set and time taken to be build this model 0.36 seconds.. The model produces the correlation coefficent accuracy level 0.3834 while applying Linear Regression in this data set and time taken to be build this model 0.01 seconds.. The model produces the correlation coefficent accuracy level 0.5156 while applying SMOreg in this data set and time taken to be build this model 0.22 seconds..

4. CONCLUSIONS

The results clearly demonstrate the various functions correlation coefficient accuracies and the model discovered the time has taken to build the models. The Linear Regression has the highest accuracy compare than other accuracy results. So the recommended model is linear regression model.

5. REFERENCES

[1] Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and Space Technologies July 2005 Volume 2 Number 2 (53-57)
[2] Romero C. and Ventura S.,"Educational data mining: A Survey from 1995 to 2005".Expert Systems with Applications (33) 135-146. 2007
[3] https://www.researchgate.net/publication/266602921 _Data_Mining_in_Educational_System_using_WEKA.
[4] https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[5] W. Huber, On the use of the correlation coefficient r for testing the linearity of calibration functions. Accreditation and Quality Assurance 9 (2004):726–727.
[6] http://www.dtic.mil/dtic/tr/fulltext/u2/1015871.pdf
[7] Wong, K.I., Vong, C.M., Wong, P.K., Luo, J.H.: Sparse Bayesian extreme learning machine and its application to biofuel engine performance prediction. Neurocomputing 149, 397–404 (2015)
[8] Matias, T., Souza, F., Ara´ujo, R., Antunes, C.H.: Learning of A Single-Hidden Layer Feedforward Neural Network Using An Optimized Extreme Learning Machine. Neurocomputing 129, 428–436 (2014)
[9] Huang, G.B., Chen, L., Siew, C.K.: Universal Approximation Using Incremental Constructive Feedforward Networks with Random Hidden Nodes. IEEE Trans. Neural Netw. 17(4), 879–892 (2006)
[10] Huang, G.B., Wang, D.H., Lan, Y.: Extreme Learning Machines: A Survey. Int. J. Mach. Learn. & Cybern. 2(2), 107–122 (2011)
[11] Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme Learning Machine: Theory and Applications. Neurocomputing 70(1), 489–501 (2006).