ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Evaluation and Performance Analysis of Machine Learning Algorithms

Mr. Shridhar Kamble[*1], Mr. Aaditya Desai[2], Ms. Priya Vartak[3]

[*1]M.E.IT (Pursuing), Thakur College of Engineering, Mumbai-400101, India

[2]Assistant Professor, IT Department, Thakur College of Engineering, Mumbai -400101, India

[3]M.E.IT (Pursuing), Thakur College of Engineering, Mumbai -400101, India

shridharkamble1@gmail.com

Abstract

Prediction is widely researched area in data mining domain due to its applications. There are many traditional quantitative forecasting techniques, such as ARIMA, exponential smoothing, etc. which achieved higher success rate in the forecasting but it would be useful to study the performance of alternative models such as machine learning methods. This paper gives performance measures of various machine learning algorithms used for prediction. The goal is to find how different machine learning algorithms gives performance when applied to different types of datasets.

Keywords: Machine Learning, J48, ZeroR, Random Forest, Naïve Bayes, SVM, MLP, RBF, MAE, RMSE, WEKA.

Introduction

Machine learning refers to a system that has the capability to automatically learn knowledge from experience and other ways. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends [3].

Performance analysis of machine learning algorithms is done in this paper, including Naïve Bayes, SVM, RBF neural networks, Decision trees and Multilayer Perceptron. These algorithms are used for classifying the Diabetes, Credit-g, Supermarket and Breast-Cancer dataset from UCI Machine learning repository [16]. Experiments are conducted using WEKA tool. Many researchers studied these algorithms and found efficient in some aspects. The goal of this research is to find the best classifier which outperforms other classifiers in all the aspects.

Data Mining Algorithms

All Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help us to provide with a better understanding of the large data. Classification predicts categorical (discrete, unordered) labels, while prediction models continuous valued functions. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity.

Classification is also called supervised Learning, as the instances are given with known labels, contrasts to unsupervised learning in which labels are not known. Each instance in the dataset used by supervised or unsupervised learning method is represented by set of features or attributes which may be categorical or continuous [1] [2].

Classification is the process of building the model from the training set made up of database instances and associated class label. The resulting model is then used to predict the class label of the testing instances where the values of the predictor features are known. Supervised classification is one of the tasks most frequently carried out by intelligent techniques. The large numbers of techniques have been developed.

Decision Trees - J48 & Random Forest

Decision trees are supervised algorithms which recursively partition the data based on its attributes; until some stopping condition is reached Decision Tree Classifier is one of the possible approaches to multistage decision-making.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

J48

J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further it provides an option for pruning trees after creation.

Random Forest

Random Forests is an ensemble learning method for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests are a wonderful tool for making predictions considering they do not over fit because of the law of large numbers.

Rule Based - ZeroR

The rule behind this algorithm is the consideration of the majority or common class of training data set to be taken as the real Zero R prediction. So, it relies on the target prediction and ignores all predictors. There is no predictability power of Zero R algorithm; however it is used to determine a baseline performance that acts as a benchmark for the other classification methods [1].

Bayesian - Naïve Bayes

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naïve Bayes model without believing in Bayesian probability or using any Bayesian methods [2][3].

Neural Networks - RBF and MLP

RBF

A radial basis function network (RBF) is an artificial neural network that uses radial basis functions as activation functions. By using RBF networks, the training of networks is relatively fast due to the simple structure of RBF networks. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation time series prediction, classification, and system control [1].

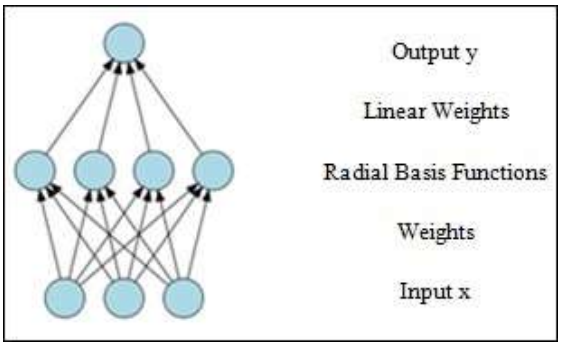

Figure 1: Architecture of a radial basis function network [18]

MLP

Artificial Neural Network (ANN) is a Machine learning techniques which largely used in forecasting, assists multivariate analysis [7]. Multi Layer Perceptron (MLP) is a feed forward neural network with one or more layers between input and output layer. Feed forward means that data flows in one direction from input to output layer (forward). This type of network is trained with the backpropagation learning algorithm. MLPs are widely used for pattern classification, recognition, prediction and approximation. Multi Layer Perceptron can solve problems which are not linearly separable [4].

Neural architecture consisted of three or more layers, i.e. input layer, output layer and hidden layer as shown in Figure 2. The function of this network was described as follows,

$$Y_j = f(\sum_i w_{ij} X_{ij}) \quad (4)$$

Where, Yj is the output of node j, f (.) is the transfer function, wij the connection weight between node j and node i in the lower layer and Xij is the input signal

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

from the node i in the lower layer to node j.

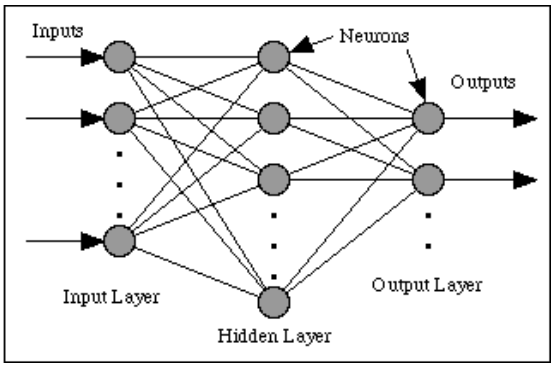

Figure 2: Artificial Neural Network Architecture [12]

Kernel Based - SVM

SVM

Support Vector Machine (SVM) is a Machine learning techniques comes under classification method which was based on the construction of hyperplanes in a multidimensional space [7]. As a result, it was allowed different class labels to be differentiated. Normally, SVM was utilized for both classification and regression tasks and it was able to handle multiple continuous and categorical variables. The purpose of the regression task of SVM was to find a function f (such that y = f(x) + noise) which was able to predict new cases. This was achieved by training the SVM model on a sample set, i.e., training set, a process that involved the sequential optimization of an error function[6][10].

Dataset Description

Experiments were conducted on the four datasets namely Diabetes [17], Credit-g [17], Super Market [16], Breast Cancer [16]. Machine with windows vista operating system and 2 GB of RAM is used. All experiments were rerun to ensure that the results are comparable.

Table 1: Dataset Description

Dataset	Data Types	#Att.	Attribute Types	#Inst.
Diabetes	Multivariate, Time-series	20	Categorical & Integer	786
Credit-g	Multivariate	20	Categorical & Integer	1000
Super-market	Multivariate	217	Integer & Real	4627
Breast-Cancer	Multivariate	10	Categorical	286

Experimental Results

Experiments were conducted in WEKA with 10 fold cross validation. Ten fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier [1]. To analyze the performance criterion for the various classifiers accuracy, precision, recall and F-Measure have been computed for all datasets. Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted. Recall is the percentage of positive labeled instances that were predicted as positive. Evaluations of time taken to build the model for different datasets are as follows,

Diabetes Dataset

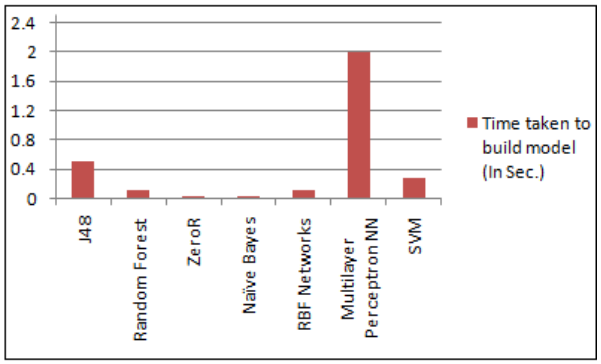

Figure 3: Analysis for Diabetes Dataset

Credit-g Dataset

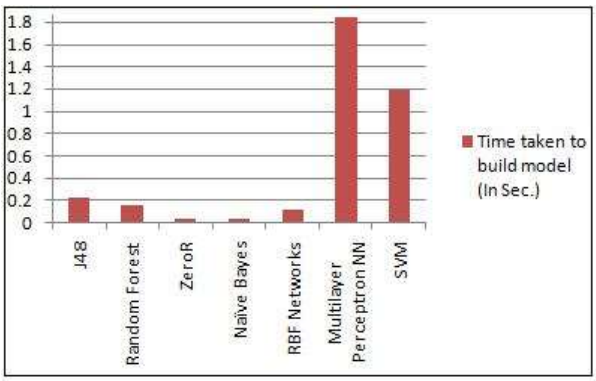

Figure 4: Analysis for Credit-g Dataset

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Super-Market Dataset

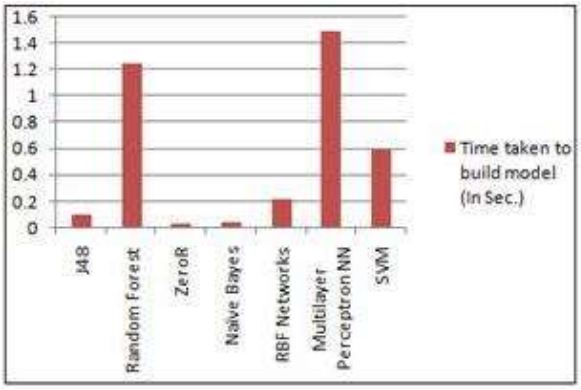

Figure 5: Analysis for Super-Market Dataset

Breast-Cancer Dataset

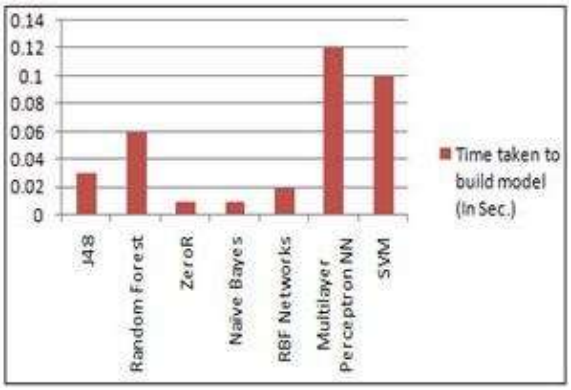

Figure 6: Analysis for Breast-Cancer Dataset

Table 2: Evaluation for Diabetes Dataset

Techniques → / Evaluation Parameters ↓	Decision tree		Rule Based	Bayesian	Neural Network		Kernel Based
	J48	Random Forest	ZeroR	Naïve Bayes	RBF Network	Multilayer Perceptron NN	SVM
Correctly Classified Instances	73.82%	73.43 %	65.10%	76.30%	72.50%	75.39%	77.34%
Incorrectly Classified Instances	26.17%	26.56%	34.89%	23.69%	27.50%	24.60%	22.65%
Kappa Statistics	0.416	0.387	0	0.466	0.423	0.448	0.468
Mean Absolute Error	0.315	0.315	0.454	0.284	0.274	0.295	0.226
RMS Error	0.446	0.428	0.476	0.416	0.429	0.421	0.476
Precision	0.735	0.726	0.424	0.759	0.735	0.750	0.769
Recall	0.738	0.734	0.651	0.763	0.740	0.754	0.773
F-Measure	0.736	0.727	0.513	0.760	0.748	0.751	0.763

Table 3: Evaluation for Credit-G Dataset

Techniques → / Evaluation Parameters ↓	Decision tree		Rule Based	Bayesian	Neural Network		Kernel Based
	J48	Random Forest	ZeroR	Naïve Bayes	RBF Network	Multilayer Perceptron	SVM
Correctly Classified Instances	70.50%	74.30 %	70%	75.40%	65.50%	71.50%	75.10%
Incorrectly Classified Instances	29.50%	25.70%	30%	24.60%	34.50%	28.50%	24.90%
Kappa Statistics	0.246	0.320	0	0.381	0.389	0.316	0.3654
Mean Absolute Error	0.346	0.336	0.420	0.293	0.385	0.288	0.249
RMS Error	0.479	0.419	0.458	0. 420	0.656	0.497	0.499
Precision	0.687	0.726	0.490	0.743	0.673	0.713	0.738
Recall	0.705	0.743	0.700	0.754	0.656	0.715	0.751
F-Measure	0.692	0.725	0.576	0.746	0.656	0.714	0.741

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Table 4: Evaluation for Supermarket Dataset

Techniques → / Evaluation Parameters ↓	Decision tree		Rule Based	Bayesian	Neural Network		Kernel Based
	J48	Random Forest	ZeroR	Naïve Bayes	RBF Network	Multilayer Perceptron	SVM
Correctly Classified Instances	63.713 %	63.713 %	66.22%	63.71%	60.5%	61.5%	63.71%
Incorrectly Classified Instances	36.287%	36.287%	33.78%	36.28%	39.5%	38.5%	36.28%
Kappa Statistics	0	0	0	0	0	0	0
Mean Absolute Error	0.462	0.462	0.432	0.462	0.474	0.462	0.362
RMS Error	0.480	0.480	0.450	0.480	0.485	0.515	0.602
Precision	0.406	0.406	0.396	0.406	0.456	0.692	0.406
Recall	0.637	0.637	0.626	0.637	0.673	0.685	0.637
F-Measure	0.496	0.496	0.476	0.496	0.688	0.688	0.496

Table 5: Evaluation for Brest-Cancer Dataset

Techniques → / Evaluation Parameters ↓	Decision tree		Rule Based	Bayesian	Neural Network		Kernel Based
	J48	Random Forest	ZeroR	Naïve Bayes	RBF Network	Multilayer Perceptron	SVM
Correctly Classified Instances	70.27%	69.930 %	71.279%	71.67%	65.5%	64.58%	69.58%
Incorrectly Classified Instances	29.72%	30.069%	28.720%	28.32%	34.5%	35.31%	30.42%
Kappa Statistics	0	0.204	0	0.285	0.389	0.157	0.198
Mean Absolute Error	0.418	0.365	0.428	0.327	0.385	0.355	0.304
RMS Error	0.457	0.468	0.477	0.453	0.473	0.542	0.551
Precision	0.494	0.674	0.484	0.704	0.656	0.648	0.671
Recall	0.703	0.699	0.723	0.717	0.673	0.647	0.696
F-Measure	0.580	0.679	0.590	0.708	0.656	0.647	0.677

Conclusion

Different machine learning algorithms are applied to various real world datasets and study is carried out to find out the classifier which can perform well on the real world data sets. The experiments were conducted in WEKA environment. After obtaining results it is observed that SVM, Naïve Bayes gives excellent performance rather than other classifiers with respect to accuracy, sensitivity, specificity and precision for both binary and multiclass datasets. Although other classifiers perform well in classification the behavior varies differently for each dataset. SVM, Naïve Bayes always outperforms other classifiers for all datasets. Bayes classification is outperformed by approaches such as boosted trees or random forests.

By observing the results Naïve Bayes Classifier and Random Forest gives highest percent of correctly classified Instances. For F-measure also, Naïve Bayes Classifier and Random Forest gives the highest value of all. Considering these evaluation measures it is observed that naïve Bayes Classifier is the best Classifier for many dataset. But it may not be same case for all the datasets. More generalized Classifier model needs to be built which would be adaptable to the different types of the datasets.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

References

[1] *lan H. Witten, Eibe Frank, "Data Mining – Practical Mahine Learning Tools and Techniques," 2nd Edition, Elsevier, 2005.*

[2] *Efraim Turban, Linda Volonino, Information Technology for Management: Wiley Publication, 8th Edition 2009.*

[3] *Chopra, Sunil and Peter Meindl. Supply Chain Management. 2 ed. Upper Saddle River: Pearson Prentice Hall, 2004.*

[4] *John Geweke and Charles Whiteman, Bayesian Forecasting," Bayesian Forecasting", Chapter 1 in Handbook of Economic Forecasting, vol. 1, pp. 3-80 from Elsevier 2006.*

[5] *Réal Carbonneau, Rustam Vahidov, Kevin Laframboise, "Forecasting Supply Chain Demand Using Machine Learning Algorithms", Chapter 6.9.*

[6] *Yang LanQin ; Xu Xin, "Research on the Price Prediction in Supply Chain based on Data Mining Technology", Published in International Symposium on Instrumentation & Measurement, Sensor Network and Automation (Volume:2), IEEE, 2012.*

[7] *Karpagavalli,Jamuna K and Vijaya MS, "Machine Learning Approach for Preoperative Anaesthetic Risk Prediction", International Journal of Recent Trends in Engineering, Volume No. 1, No. 2, 2009.*

[8] *Sanjeev Kumar Aggarwal, Lalit Mohan Saini, Ashwani Kumar, "Electricity price forecasting in deregulated markets: A review and evaluation, International journal of production economics, Volume 31, Issue 1, Pages 13-22, Elsevier B.V. 2009.*

[9] *J. Shahrabi, S. S. Mousavi and M. Heydar, "Supply Chain Demand Forecasting- A Comparison of Machine Learning Techniques and Traditional Methods", Journal of Applied Sciences, Volume 9,Issue 3, pp.521-527, 2009.*

[10] *Hamid R. S. Mojaveri, Seyed S. Mousavi, Mojtaba Heydar, and Ahmad Aminian, "Validation and Selection between Machine Learning Technique and Traditional Methods to Reduce Bullwhip Effects: a Data Mining Approach", World Academy of Science, Engineering and Technology, Volume 25, 2009.*

[11] *Neagu C.D., Guo G., Trundle P.R.,Cronin M.T.D.,"A comparative study of machine learning algorithms applied to predictive toxicology data mining", Alternatives to laboratory animals : ATLA 35:1, pp. 25-32, 2007.*

[12] *Karin Kandananond, "Consumer Product Demand Forecasting based on Artificial Neural Network and Support Vector Machine", World Academy of Science, Engineering and Technology 63, 2012.*

[13] *Liljana Ferbar, DavidC˘reslovnik, Blaz˘ Mojs "Demand forecasting methods in a supply chain: Smoothing and denoising", International journal of production economics, pp. 49-54 Elsevier B.V. 2009.*

[14] *John B. Guerard, Jr., Introduction to Financial Forecasting in Investment Analysis: Springer New York, Online ISBN : 978-1-4614-5239-3, 2013.*

[15] *Sanjeev Kumar Aggarwal, Lalit Mohan Saini, Ashwani Kumar, "Electricity price forecasting in deregulated markets: A review and evaluation", International journal of production economics, Volume 31, Issue 1, Pages 13-22, Elsevier B.V. 2009.*

[16] *Weka Dataset Repository, ww.cs.waikato.ac.nz/ml*

[17] *UC Irvine Machine Learning Repository, www.archive.ics.uci.edu/ml/*