IJCSEInternational Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Performance Assessment of Machine Learning Algorithms with Feature Selection Methods

A. Thakur[1]

[1]School of Computer Science & IT, Devi Ahilya Vishwavidyalaya, Indore, Madhya Pradesh, India

Author's Mail Id: archana227@gmail.com

Available online at: www.ijcseonline.org

Received: 22/Mar/2018, Revised: 28/Mar/2018, Accepted: 20/Apr/2018, Published: 30/Apr/2018

Abstract— Machine learning is a field of artificial intelligence in which computers learn from experience. The field of machine learning is a famous research area in computer science. Machine learning applications are helpful in various domains of computer science, chemical sciences, spatial technology, bioinformatics, agriculture, digital forensics and more. Machine learning algorithms are useful in the fields of pattern recognition, pattern classification, text classification, SMS classification, computer vision, mobile learning and more. In the present work performance assessment of three machine learning algorithms namely logistic regression, random forest and naïve bayes with three feature selection methods viz. correlation based, Information based and gain ratio is conducted on a mobile device. The above-mentioned machine learning algorithms along with feature selection methods are assessed for the performance metrics of accuracy, precision, F- Measure, recall and Receiver Operating Characteristics.

Keywords — Machine Learning; Logistic Regression; Naïve Bayes; Random Forest; Gain Ratio, Information Gain.

I. INTRODUCTION

The field of machine learning is a subfield of the field of *artificial intelligence*, in which machines are trained to imitate intelligent behavior of human beings. In other words, machine learning is a field of programming computers to augment a performance criterion with example data or some past experience [1]. Machine learning algorithms use the principles of statistics in designing mathematical models for inference mechanism. There are numerous applications of machine learning viz. image recognition, traffic prediction, malware detection, speech recognition, self-driving cars, digital forensics, mobile user location identification, biometrics, product recommendations, spam email filtering, virtual personal assistant, detecting online fraud and much more. There are four types of machine learning namely supervised, unsupervised, semi supervised and reinforcement learning. We train the machines by means of some "labelled" dataset in supervised learning and based upon the training, the machine predicts the outcome. Here, the labelled data signifies that some of the inputs are already mapped to some specified output. In other words, first we train the machine with the particular input and its corresponding output, and then we ask the machine to forecast on the basis of some test dataset. Some popular supervised learning algorithms are logistic regression, decision tree, random forest, support vector machine, naïve bayes, k-nearest neighbor and more. Supervised learning works with the labelled dataset so we can have an exact knowledge about the classes of objects. Supervised learning algorithms are helpful in predicting the output on the basis of some previous experience. But these algorithms are not able to solve complicated problems. Supervised learning algorithm may result in incorrect output if the test data is different from the training data. Lot of computational time is required to train the supervised learning algorithm. Some important applications of supervised learning are medical diagnosis, image segmentation, fraud and spam detection, speech recognition and more.

Unsupervised learning is different from supervised learning as there is no need for supervision. In unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine forecasts the output without any supervision. The major focus of the unsupervised learning algorithm is to categorize or group the unsorted dataset according to the patterns, similarities and differences. Machines are instructed to determine the hidden patterns from the input dataset. Clustering and association are two types of unsupervised learning. As unsupervised algorithms work on the unlabeled dataset hence, they can be used to perform complicated tasks as compared to the supervised algorithms. Unsupervised algorithms are better for various tasks as receiving the unlabeled dataset is easier as compared to the labelled dataset. But output from an unsupervised algorithm may be less accurate as the dataset is not labelled, and the algorithms are not trained with the exact output previously. As unlabeled dataset does not exactly map with the output, hence working with unsupervised learning is harder as compared to supervised learning algorithm. The applications of unsupervised learning are in network analysis, anomaly detection, recommendation systems and in single value decomposition.

In order to overcome the disadvantages of supervised learning and unsupervised learning algorithms, the concept of semi-supervised learning was presented. The major focus of semi-supervised learning is to efficiently use all the available data, rather than only labelled data as in case of supervised learning. Initially, comparable data is clustered using an unsupervised learning algorithm, and further, it assists to label the unlabeled data into a set of labelled data. Semi-supervised learning algorithms are relatively simple and easy to understand. These algorithms have good efficiency. These are used to solve the drawbacks of supervised and unsupervised learning algorithms. But these algorithms offer low accuracy as compared to supervised and unsupervised algorithms. Further, these algorithms cannot be applied on network-level data.

Reinforcement learning works on a feedback-based procedure, in which an AI agent (generally a software component) inevitably explores its surrounding by hit and trail, taking some action, learning from experiences, and thereby improving its performance. The agent receives reward for each good action and receives punishment for each bad action. Here the aim of reinforcement learning agent is to maximize the rewards. There are two types of reinforcement learning viz. positive reinforcement learning and negative reinforcement learning. Reinforcement learning helps in solving complex real-world problems. It produces accurate results but this learning is not used for solving simple problems. Reinforcement learning methods employ usage of huge data and high computation. Some real-world applications of reinforcement learning include video games, resource management with deep reinforcement learning, robotics and text mining.

II. REVIEW OF THE RELATED LITERATURE

Machine learning algorithms were used in domains of Wireless Sensor Networks and Mobile Ad-hoc Networks, design of Multi-Hop Broadcast Protocols for VANET, eHealth systems and learning based on user-specific touch input model [11, 12, 13, 14]. Machine learning algorithms were employed to signify the behavior of children suffering with autism communicating through a humanoid robot [2]. A comparison was conducted amongst a dynamic model and a static model through hand-coded features in [2]. A significant accuracy (above 80%) was attained in forecasting child vocalizations. Also, the future directions for modeling the children behavior suffering from autism were recommended in [2]. Artificial Neural Network (ANN) and Bayesian learning were employed for modeling the response time of service-oriented systems in [3]. It was an important observation that Bayesian learning outperformed ANN but had reduced sensitivity to limited sized dataset [3]. Bayesian learning was recommended more appropriate for varying environments and require recurrent response-time model constructions. ANN was recommended for attaining faster model estimation time. ANN assisted in appropriate management routines which claim exhaustive response-time predictions [3]. Machine learning algorithms were also helpful in the predicting the condition numbers of sparse matrices [4]. Condition number of a matrix provides a vital measure in linear algebra and numerical analysis [4]. Support Vector Machine (SVM) and Modified K-Nearest Neighbor (KNN) algorithms were employed to approximate the condition number of a specified sparse matrix. The experimental results showed that modified KNN outperformed SVM on the selected data set. SVM, ANN, KNN were used for assessing the function points of a software in [5]. The experiments conducted in [5] proved that ANN and SVM are efficient algorithms for predicting software function points. A framework for assessing machine learning based methods for call admission control was presented in [6]. A comparative performance investigation of two major machine learning algorithms namely ANN and Bayesian Network for QoS prediction was conducted in [6]. The size of training data set for ANN was relatively larger than Bayesian Network. A comparative performance examination of machine learning methods Decision tree, Flexible neural tree and Particle Swarm Optimization (PSO) for intrusion detection on network traffic was conducted in [7]. The experimental results revealed that Decision tree offered better accuracy as compared to the other methods. Random forest and lasso regularization algorithms were employed for predicting software anomalies in [8]. Machine learning algorithms are also fruitful in transportation. SVMs were employed for the short-term prediction of travelling time in [9]. A comparative investigation amongst ANN and SVM was conducted in [9]. It was observed that SVM performs better for the short-term prediction of travelling time. It was also observed that the effect of the amount of training data employed was more on ANN method than on the SVM method.

III. MATERIALS AND METHODS

The following feature selection methods and machine learning algorithms are used in the present work.

a) Feature selection methods

The main aim of any feature selection algorithm is to determine optimal set of features for classification. Feature extraction algorithms determine a novel set of dimensions that are groupings of original dimensions. Correlation based Feature Selection employs heuristic for estimating the value or excellence of the feature subset. The method determines the subset of features which contain features that are highly correlated with the class but are not correlated with each other [10, 15]. There are many feature selection methods. Information gain is a vital feature selection method employed for ranking of features. It measures the strength of information gained during classification provided that the feature is considered. It measures the amount of impurity in a set. A common measure of recognizing impurity in a set is the entropy.

Information gain is evaluated by feature's influence on overall reducing entropy [16,19]. Gain ratio feature selection method is somewhat comparable to information gain method. It approximates the gain in information for a classification related to entropy of a given feature. In other words, it estimates the merits of a feature by evaluating the gain ratio of that feature with the respective class.

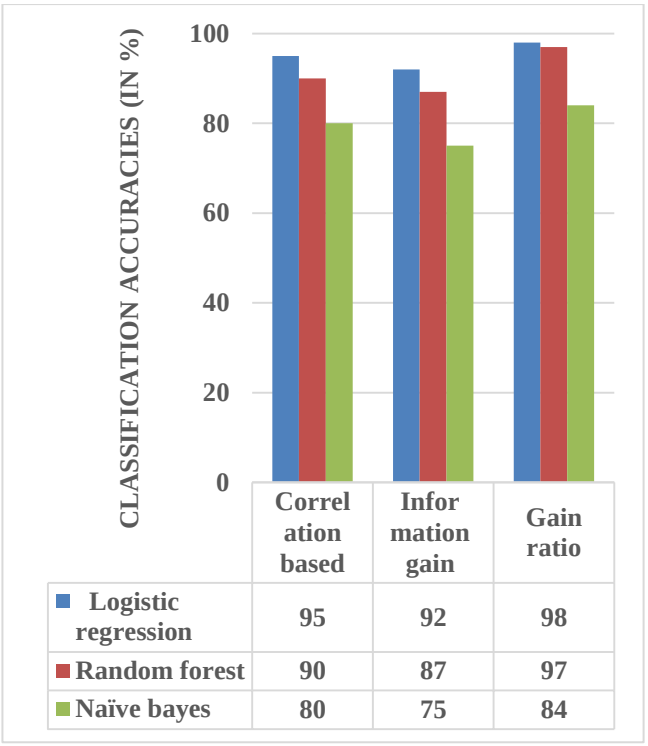
b) Machine learning algorithms

Logistic regression is one of the most famous supervised machine learning algorithms. It is employed for predicting a dependent categorical variable using a set of independent variables. Hence, in this case the outcome must be a categorical or discrete value.

Random forest is one of the most popular machine learning algorithms. It belongs to the category of supervised learning. It can be applied for classification as well as for regression problems in machine learning. It is based upon the principle of ensemble learning, which is a method of merging multiple decision tree algorithms to solve a complex problem. Generally, it contains a number of decision trees designed on different subsets for a given dataset. Instead of depending on a single decision tree, the random forest takes the prediction from every tree and based upon the majority votes of predictions, it predicts the final outcome. Naïve bayes is one of the simple and most famous machine learning algorithms. It assists in designing fast machine learning models that give quick predictions. It is a probabilistic classification algorithm. The algorithm predicts on the basis of the probability of an object. It works using Bayesian principle.

IV. RESULTS AND DISCUSSIONS

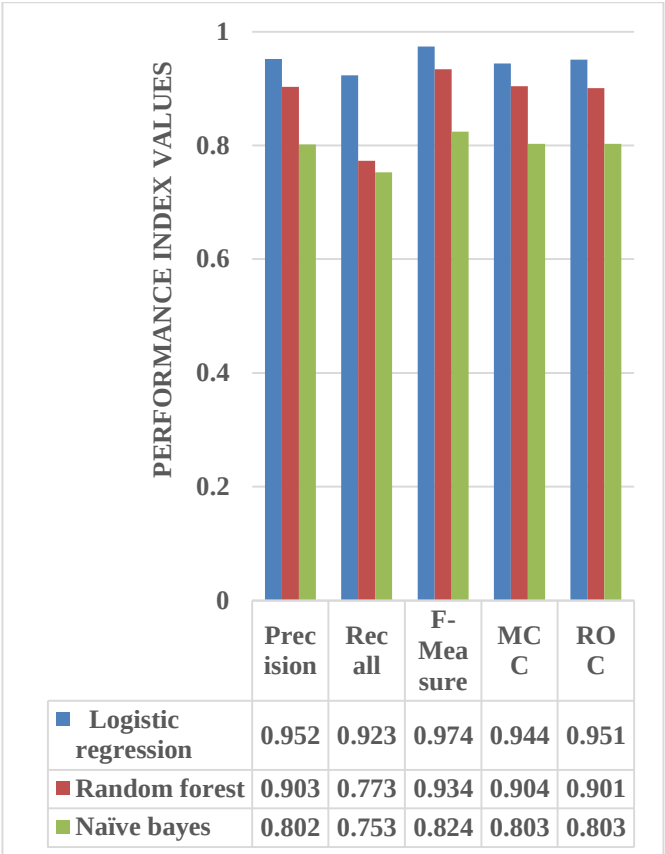
The performance of any machine learning algorithm is evaluated by some simple measures [17, 18]. In the present work the performance is assessed using the performance measures viz. classification accuracy, precision, recall and F-measure values. The experiments are conducted using 10-fold cross validation. The dataset used in the present work is the crop disease dataset having features related to crop and symptoms of various diseases of crop. It is evident from the experiments conducted as shown in Fig. 1 that logistic regression outperformed the other two algorithms. Random forest showed better performance than Naïve bayes. It is also clear from Fig. 2, the gain ratio feature selection method performed better than the other two feature selection methods. The correlation-based feature selection method performed better than the information gain feature selection method.



	Correlation based	Information gain	Gain ratio
Logistic regression	95	92	98
Random forest	90	87	97
Naïve bayes	80	75	84

Fig.1 Classification accuracy observations.

It is clear from Fig. 2 that logistic regression outperforms random forest and naïve bayes for the chosen performance measures viz. precision, recall and F-measure values.



	Precision	Recall	F-Measure	MCC	ROC
Logistic regression	0.952	0.923	0.974	0.944	0.951
Random forest	0.903	0.773	0.934	0.904	0.901
Naïve bayes	0.802	0.753	0.824	0.803	0.803

Fig. 2 Performance measure micro averaged values

V. CONCLUSIONS

The field of machine learning is a subfield of artificial intelligence. Machine learning is primarily concerned with the design of algorithms which permit a computer to learn from the data and previous experiences. In the present work the performance is assessed using the performance measures viz. classification accuracy, precision, recall, F-measure, Mathew's Correlation Coefficient (MCC) and Receiver's Operating Characteristics (ROC) values. It is evident from the experiments conducted that logistic regression outperformed the other two algorithms. Random forest algorithm showed better performance than Naïve bayes algorithm. It is also clear from the experiments conducted that the gain ratio feature selection method performed better than the other two feature selection methods. The correlation-based feature selection method performed better than the information gain feature selection method.

References

[1] Ethem Alpaydin. Introduction to Machine Learning, Second Edition, MIT press, Cambridge, London, **2009.**

[2] E. Short, D.F.Seifer, M. Matari, "A Comparison of Machine Learning Techniques for Modeling Human-Robot Interaction with Children with Autism", Human-Robot Interaction (HRI), 6th ACM/IEEE International conference, 251 – 252, Lausanne, **2011.**

[3] R. Zhang, A.Bivens, "Comparing the Use of Bayesian Networks and Neural Networks in Response Time Modeling for Service-oriented Systems", In Proceedings of the 2007 workshop on Service-oriented computing performance: aspects, issues, and approaches, pp.**67 – 74**, New York, USA, **2007.**

[4] D. Han, J. Zhang, " A Comparison of Two Algorithms for Predicting the Condition Number", Sixth International Conference on Machine Learning and Applications, 13-15 Dec. pp.**223 – 228**, Cincinnati, OH, **2007.**

[5] G. Sikka, A. Kaur, M. Uddin, "Estimating Function points: Using Machine Learning and Regression Models", 2nd International Conference on Education Technology and Computer (ICETC), Shanghai, **2010.**

[6] A. Bashar, G. Parr, S. McClean, B. Scotney, D. Nauck, "Machine Learning based Call Admission Control Approaches: A Comparative Study", IEEE International Conference on Network and Service Management (CNSM), pp.**431–434**, Niagara Falls, ON, **2010.**

[7] M. Bahrololum, E. Salahi, M. Khaleghi, "Machine Learning Techniques for feature Reduction in Intrusion Detection Systems: A Comparison", Fourth International Conference on Computer Sciences and Convergence Information Technology, pp.**1091–1095**, Seoul, **2009.**

[8] J. Alonso, L. Belanche, D.R. Avresky, "Predicting Software Anomalies using Machine Learning Techniques", IEEE International Symposium on Network Computing and Applications (NCA), pp.**163–170**, Cambridge, MA, **2011.**

[9] L. Vanajakshi, L. Rillet, "Support Vector Machine Technique for the Short Term Prediction of Travel Time", IEEE International Symposium on Intelligent Vehicles, pp.**600–605**, Istanbul, **2007.**

[10] R. Eisinger, R. Romero, R. Goularte, "Machine Learning Techniques Applied to Dynamic Video Adapting", IEEE Seventh International conference on Machine Learning and Applications (ICMLA '08), **2008.**

[11] A. Forster, "Machine Learning Techniques Applied to Wireless Ad-Hoc Networks: Guide and Survey", International conference on Intelligent Sensors, Sensor Networks and Information, pp. 365 – 370, Melbourne, Queensland, **2007.**

[12] M. Slavik, I. Mahgoub, "Applying Machine Learning to the Design of Multi-Hop Broadcast Protocols for VANET", Seventh International Conference on Wireless Communications and Mobile Computing (IWCMC), pp.**1742–1747, 2011.**

[13] M. Grajzer, M. Koziuk, P. Szczechowiak, A. Pescap, "A Multi-Classification Approach for the Detection and Identification of eHealth Applications", Twenty First International conference on Computer Communications and Networks (ICCCN), pp. 1 – 6, Munich, **2012.**

[14] D. Weir, S. Rogers, R. Murray-Smith, M. Lochtefeld, "A User-Specific Machine Learning Approach for Improving Touch Accuracy on Mobile Devices", UIST, Cambridge, Massachusetts, USA, **2012.**

[15] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning",University of Waikato, Hamilton, **1999.**

[16] A. Sharma, S. Dey, "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis features", Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computingand Communication Technologies for HPC Applications, **2012**.

[17] M. Sokolova, G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing and Management, Vol.**45**, pp.**427–437, 2009.**

[18] P. Liu, Y. Chen, W. Tang, Q. Yue, "Mobile WEKA as Data Mining Tool on Android**",** Advances in Intelligent and Soft Computing Vol.**139**, pp.**75-80, 2012.**

[19] L. Silva, M. Koga, C. Cugnasca, A. Cost Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedling", Computers and Electronics in Agriculture, Vol.**97**, pp.**47–55, 2013.**

AUTHOR'S PROFILE

Dr. Archana Chaudhary Thakur received M.Tech. and Ph.D. from School of Computer Science & IT, Devi Ahilya University, Indore. She is working as an Assistant Professor at School of Computer Science & IT, Devi Ahilya University, Indore. She is involved in coordinating postgraduate-level training program in computer science for the university. She is guiding many M.Tech. and Ph.D. research scholars. She has published many research papers in various reputed national and international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE whose manuscripts are also available online. She has also been esteemed author and reviewer for many Elsevier journals. Her research areas include Artificial Intelligence, Machine learning, Data Mining and Soft Computing.

