Machine Learning and its Applications: A Review

Sheena Angra [1], Sachin Ahuja [2]

[1] Ph.D. Scholar, Chitkara University,India

[2] Associate Director, CURIN, Chitkara University, India

sheena.angra@chitkara.edu.in [1], sachin.ahuja@chitkara.edu.in [2]

***Abstract* — Nowadays, large amount of data is available everywhere. Therefore, it is very important to analyze this data in order to extract some useful information and to develop an algorithm based on this analysis. This can be achieved through data mining and machine learning. Machine learning is an integral part of artificial intelligence, which is used to design algorithms based on the data trends and historical relationships between data. Machine learning is used in various fields such as bioinformatics, intrusion detection, Information retrieval, game playing, marketing, malware detection, image deconvolution and so on. This paper presents the work done by various authors in the field of machine learning in various application areas.**

Keywords- Machine Learning; SVM; clustering; feature selection; decision tress; classification; logistic regression.

I. INTRODUCTION

Machine learning has evolved from the study of computational learning theory and pattern recognition. It is the most effective method used in the field of data analytics in order to predict something by devising some models and algorithms. These analytical models allow researchers, engineers, data scientists and analysts to produce reliable and valid results and decisions. It also helps to discover some hidden patterns or features through historical learning's and trends in data [1]. Feature selection is the most important task of machine learning [5]. Model is created based on the results gathered from the training data that is why machine-learning algorithms are non-interactive. It studies the past observations to make precise predictions. It is a very difficult task to make an accurate prediction rule based on which algorithm can be developed [2]. For example, spammed and non-spammed emails have to be distinguished using machine learning, then this could be done by collecting examples of spammed and non-spammed emails. Then these examples are fed into the machine-learning algorithm to indicate whether the mails are spammed or not by generating an accurate prediction rule [2]. ML is suitable for dealing with those problems where theoretical knowledge is still insufficient but for which we have an ideal number of observations and results [16]. Section II and III discusses the literature survey and conclusion respectively.

II. LITERATURE SURVEY

Miroslav Kubat et al. in 1998 [7], described the machine learning's application to detect oil spills from the radar images. Some of the issues of machine learning were also discussed with the approach to solve these issues. Experimental studies for two main issues of machine learning were also discussed. These issues are batched and imbalanced training sets. Two algorithms SHRINK and one-sided selection were also discussed. False alarms rate could be controlled using SHRINK according to their study.

Asa Ben-Hur et al. in 2001 [4], presented a new method for clustering by using Support vector machines. Mapping of data points to high dimensional space were done by using Gaussian Kernel and when these points were mapped back to data space, they were separated into several clusters. To identify these clusters an algorithm is presented. The algorithm presented is SVC, which is based on SVM, and the proposed method is unbiased of the shape and number of clusters. The authors used two parameters p and q. Parameter p, also known as soft

978-1-5090-6399-4/17/$31.00 ©2017 IEEE

margin is used to control the amount of outliers and parameter q is used to determine the data probing scale and with the increase in this parameter clusters tends to split. The authors discussed the following advantages of the proposed algorithm: Cluster boundaries of any shape are generated using the proposed algorithm and unnecessary calculations are avoided which lead to high efficiency.

Robert E. Schapire in 2003 [2], overviewed the work done on boosting including analysis of AdaBoost's generalization and training error, the relationship between logistic regression and boosting, applicability of boosting on linear programming and game theory, incorporation of human knowledge into boosting. Author discussed the following advantages of AdaBoost: AdaBoost is to find outliers and it is used to reduce the error produced by committing some mistakes on the training set.Adaboost is simple, easy and fast to program.

Jose M. Jerez et al. in 2010 [3], evaluated the performance of machine learning and statistical imputation methods to identify the repetition in the patients in data set of breast cancer. Some of the Imputation methods based on machine learning techniques includes k-nearest neighbor, multi-layer perceptron, self-organization maps (SOM), and statistical techniques are multiple imputation, mean, and hot-deck, they were applied to the collected data, and the results of these techniques were then compared to list wise deletion imputation method. The database included information of 3679 women who are diagnosed with breast cancer in 32 different hospitals. The results showed that that the machine learning imputation methods gave better results than statistical imputation methods.

J.R. Otukei and T. Blaschke in 2010 [6], explored the use of support vector machines, decision trees and classification for land cover changes and mapping in rural areas. For this purpose three objectives were achieved which were exploration of possible data mining techniques for the recognition of suitable bands for classification, performance comparison of all the three techniques and identifying the changes in land cover. Before the analysis data, preprocessing was done using ERDAS IMAGINE 9.1 and ENVI 4.5. Decision trees achieved high performance and accuracy as compared to other two methods when applied to the data. Failure degradation was also estimated.

Wahyu Caesarendra et al. in 2010 [9], proposed the combination between logistic regression and relevance vector machine for assessing the failure degradation in order to predict the failures before they actually occur. Failure degradation is measured by logistic regression and some vectors measured the results obtained. These vectors were then trained using relevance vector machine. Failure simulation data was employed in the proposed method in order to experiment it on run-to-failure- data. For this, Kurtosis is calculated which is a one-dimensional feature and every unit of machine component is predicted by applying LR and RVM on it. Training performance was evaluated using correlation and root mean squared error.

Degang Chen et al. in 2010 [10], improved the hard margin SVM's by taking the membership of every tuple which is to be trained into consideration and this was done by using the technique of fuzzy rough sets. First of all, fuzzy transitive kernel was proposed which is based on fuzzy rough sets. Lower approximation operator was used for binary classification to calculate the membership of every training input. After this, the comparative analysis of the proposed method was done with fuzzy SVM's and soft margin SVM's. The experiments showed that the proposed method is valid, stable, and fuzzy theory and SVM's were interlinked to each other.

Dursun Delen in 2010 [12], developed the models to analyze and to predict the reasons behind the disintegration of students who are fresher's. For this the factors were analyzed which could affect their retention. The models were developed using some data mining techniques and taking data of five years of an institution. Performance of the models used for prediction was estimated using the 10-fold cross validation method. In this process, the whole dataset was divided into 10 subsets, which are mutually exclusive. These models predicted the students who would retain and who would drop out before their second year. The SVM gave better results than logistic regression, decision trees and neural networks.

Sajjad Ahmad et al. in 2010 [14], presented SVM which is a regression technique and applied this technique to estimate the soil moisture by using remote sensing data. This model was applied to 10 sites for estimating soil moisture in the western United States. 5 years data from 1998 to 2002 was trained using SVM model and was tested on data of three years from 2003 to 2005. To evaluate the performance of SVM two models were developed. In first model, data of 6 sites was first trained and then tested which resulted into 6 different models for 6 different sites. The second model combined the data of all the sites used in model one and then the single model was developed to test these sites again and then this model was tested on remaining 4 sites. Results showed that SVM performed better than MLR and ANN models.

Fan Min et al. in 2012 [5], proposed the feature selection with test constraint problem for the issue of test cost constraint due to unavailability or scarcity of resources. Selection of feature with the test constraint was defined with 4 facets: constraint, input, optimization, and output objective. Backtracking algorithm was developed for this problem for medium and small-sized datasets and a heuristic algorithm was developed for huge datasets. Performance of the proposed algorithm was tested on 4 datasets. Backtracking algorithm proved to be efficient for data with medium size but in general, heuristic algorithm is more efficient and stable than the backtracking algorithm.

Athanasios Tsanas and Angeliki Xifara in 2012 [13], studied the effect of two output parameters and eight input parameters by developing a machine learning framework. Input parameters includes: surface area, roof area, orientation, relative compactness, glazing area, overall height and glazing area distribution and the output parameters includes: cooling and heating load. Then the association between every input and output parameter is measured with the help of machine learning tools. After that, random forests and linear regression were compared to estimate the cooling and heating load. For every test repetition, the mean relative error (MRE),mean square error (MSE) and the mean absolute error(MAE) was recorded for both testing and training subsets. Correlation between all the input variables mentioned above was calculated using Spearman rank correlation which is described in the Figure 1[13].

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution
Relative Compactness	1	-1	-0.256	-0.871	0.869	0	0	0
Surface Area	-1	1	0.256	0.871	-0.869	0	0	0
Wall Area	-0.256	0.256	1	-0.193	0.221	0	0	0
Roof Area	-0.871	0.871	-0.193	1	-0.937	0	0	0
Overall Height	0.869	-0.869	0.221	-0.937	1	0	0	0
Orientation	0	0	0	0	0	1	0	0
Glazing Area	0	0	0	0	0	0	1	0.188
Glazing Area Distribution	0	0	0	0	0	0	0.188	1

Figure 1: Spearman Rank Correlation between all the input variables

Christian J. Schuler et al. in 2013 [15], dealt with non-blind deconvolution and to make the blurred images bright authors followed a two-step procedure. In the first step, noise is amplified and colored, the image information is corrupted, and in the second step, an algorithm is used for the removal of colored noise. Machine learning approach, which was used to sharpen the images, is neural networks. First of all, regularized inversion of the blurred image was done in Fourier domain and then denoising is done by using neural networks. Then the proposed method was compared with the existing methods. Results showed that the proposed method outperformed the existing methods.

Behshad Hosseinifard et al. in 2013 [11], differentiated the depression and normal patients by studying the non-linear analysis of EEG signal. Study was done on 45 normal patients and 45 depression patients. Some of the techniques were used to differentiate between these two groups. These techniques includes logistic regression, linear discriminant analysis and K-nearest neighbor. The method used to train the data sets was leave-one-out and based on the results, this method was applied on test data sets. According to the experiments, LR gave better results than KNN and LDA and highest accuracy is achieved and these results are best described in Figure 2[11].

Classifier	Feature			
	Delta(%)	Theta(%)	Alpha(%)	Beta (%)
KNN	66.6	70	70	66.6
LDA	66.6	70	73.3	70
LR	70	70	73.3	70

Figure 2:Classification Accuracy for power EEG Bands

Nouman Azam and JingTao Yao. in 2014 [8], described the issue of determining suitable threshold values for boundary, negative and positive regions. This issue could be resolved by identifying some sort of relationship between changes in the possible thresholds and their impact on all the three regions. They explored this relationship by investigating the use of the Game theoretic rough set model. Authors used this model to analyze and make intelligent decisions in the cases where multiple criteria are involved. A game was formulated between frequent and prolonged regions to configure the uncertainties in these regions by applying some techniques.

Eric J.Parish and KarthikDuraisamy in 2016 [17], proposed a modeling prototype known as field

inversion and machine learning for physics applications. Information is directly inferred from the data and then the inferred function in terms of different parameters and variables is reconstructed by applying it over a number of problems. This then aimed to create common modelling information from the concluded information. The rebuilt function was then embedded into an analytical solution process. This technique helped in identification of possible errors at the initial level rather than discovering them at the output level.

III. CONCLUSION

We have discussed the various machine learning techniques and approaches in various fields and application areas. Machine learning is similar to data mining but the difference is that based on observations and analysis new algorithm or model is developed in the former approach whereas only analysis is done in the latter approach. We have discussed the role of machine learning in different fields such as image deconvolution, student retention, detection of oil spills, land cover changes, and some of the other applications. This gave us a brief idea about the machine learning and the fields where it can be used.

REFERENCES

[1]."Machine Learning: What it is and why it matters". www.sas.com. Retrieved 2016-09-25.

[2].Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In Nonlinear estimation and classification. Springer New York, 149-171.

[3].Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine, 50(2), 105-115.

[4].Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. Journal of machine learning research, 2(Dec), 125-137.

[5].Min, F., Hu, Q., & Zhu, W. (2014). Feature selection with test cost constraint. International Journal of Approximate Reasoning, 55(1), 167-179.

[6].Otukei, J. R., & Blaschke, T. (2010). Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. International Journal of Applied Earth Observation and Geoinformation, 12, S27-S31.

[7].Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. Machine learning, 30(2-3), 195-215.

[8].Azam, N., & Yao, J. (2014). Analyzing uncertainties of probabilistic rough set regions with game-theoretic rough sets. International Journal of Approximate Reasoning, 55(1), 142-155.

[9].Caesarendra, W., Widodo, A., & Yang, B. S. (2010). Application of relevance vector machine and logistic regression for machine degradation assessment. Mechanical Systems and Signal Processing, 24(4), 1161-1171.

[10].Chen, D., He, Q., & Wang, X. (2010). FRSVMs: Fuzzy rough set based support vector machines. Fuzzy Sets and Systems, 161(4), 596-607.

[11].Hosseinifard, B., Moradi, M. H., & Rostami, R. (2013). Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. Computer
methods and programs in biomedicine, 109(3), 339-345.

[12].Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498-506.

[13].Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings, 49, 560-567.

[14].Ahmad, S., Kalra, A., & Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. Advances in Water Resources, 33(1), 69-80.

[15].Schuler, C. J., Christopher Burger, H., Harmeling, S., & Scholkopf, B. (2013). A machine learning approach for non-blind image deconvolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ,1067-1074.

[16]. Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1), 3-10.

[17]. Parish, E. J., & Duraisamy, K. (2016). A paradigm for data-driven predictive modeling using field inversion and machine learning. Journal of Computational Physics, 305, 758-774.