International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 02 | February - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Review and Applications of Machine Learning

KONAKALLA TEJA, B UDAY JASWANTH REDDY

KONAKALLA TEJA, DEPARMENT:COM COLLEGE: PRESIDENCY UNIVERSITY

B UDAY JASWANTH REDDY, DEPARMENT:COM COLLEGE: PRESIDENCY UNIVERSITY

ABSTRACT:

With the vast amount of data available nowadays, it is crucial to analyze it to uncover valuable insights and develop algorithms. Data mining and machine learning are essential tools in achieving this goal, where machine learning utilizes historical data relationships and trends to create algorithms. Machine learning has a broad range of applications such as bioinformatics, intrusion detection, marketing, and image deconvolution. This paper provides a summary of the research conducted by multiple authors in the field of machine learning across various industries.

KEYWORDS : Machine learning, SVM, clustering,feature selection: decision tress;classification ;logistics;regression

1. INTRODUCTION:

Machine learning is an advanced method for analyzing data, which has evolved from computational learning theory and pattern recognition. It is widely used in the field of data analytics to develop algorithms and models that can predict outcomes. These models help researchers, engineers, data scientists, and analysts to make reliable and valid decisions by identifying hidden patterns and features in the data.

In machine learning, feature selection is a critical task that determines the usefulness of the data. The model is created based on training data, which is used to generate an accurate prediction rule. Machine learning algorithms are non-interactive and rely on past observations to make precise predictions. However, developing an accurate prediction rule is a challenging task.

Machine learning can be applied in various fields, such as spam detection, bioinformatics, intrusion detection, information retrieval, and game playing, to name a few. For example, spam detection can be achieved by collecting examples of spammed and non-spammed emails and using them to train the machine-learning algorithm to predict whether an email is spam or not.

Machine learning is suitable for dealing with problems for which theoretical knowledge is still insufficient, but an adequate number of observations and results are available. This paper reviews the literature on machine learning in various application areas and concludes with a discussion of the findings.

2.LITERATURE SURVEY

In 1998, Miroslav Kubat and his colleagues conducted a study on the use of machine learning to detect oil spills from radar images. The study highlighted some challenges associated with machine learning, such as batched and imbalanced training sets, and proposed two algorithms to address them: SHRINK and one-sided selection. The study found that using SHRINK could help to control false alarm rates.

In 2001, Asa Ben-Hur and his team presented a new clustering method that used Support Vector Machines (SVM) with a Gaussian Kernel to map data points to high dimensional space. The algorithm, called SC, was designed to identify clusters based on the mapped points, and was unbiased towards the shape and number of clusters. The authors used two parameters, p and q, to control the number of outliers and data probing scale respectively. The proposed method was found to be highly efficient due to the avoidance of unnecessary calculations and the ability to generate cluster boundaries of any shape.

Overall, both studies demonstrated the potential of machine learning in various applications, including the detection of oil spills and clustering of data points. The authors proposed novel algorithms to address some of the challenges associated with machine learning and highlighted the advantages of these approaches. These findings are important in advancing the development of machine learning algorithms and their applications in various fields.

In 2003, Robert E. Schapire provided an overview of boosting, including the analysis of AdaBoost's generalization and training error, the relationship between logistic regression and boosting, and the applicability of boosting on linear programming and game theory. The author also discussed incorporating human knowledge into boosting. According to Schapire, AdaBoost is useful for detecting outliers and reducing errors from training set mistakes. It is a simple and fast programming method.

In 2010, Jose M. Jerez et al. conducted a study to compare the performance of machine learning and statistical imputation methods for identifying repetition in breast cancer patient data. The study utilized a database containing information on 3,679 women who were diagnosed with breast cancer in 32 different hospitals. The authors applied several imputation methods based on machine learning techniques, including k-nearest neighbor, multi-layer perceptron, and self-organizing maps, as well as statistical techniques such as multiple imputation, mean, and hot-deck, and compared their results to the list-wise deletion imputation method. The study found that the machine learning imputation methods outperformed the statistical imputation methods in terms of accuracy and precision in identifying repeated patient data.

J.R. Otukei and T. Blaschke in 2010 [6], investigated the use of support vector machines, decision trees, and classification for mapping and detecting land cover changes in rural areas. The study aimed to explore suitable data mining techniques for identifying appropriate bands for classification, compare the performance of the three methods, and detect changes in land cover. Data preprocessing was done using ERDAS IMAGINE 9.1 and ENVI 4.5 before the analysis. The results showed that decision trees achieved better performance and accuracy than the other two methods when applied to the data. The study also estimated failure degradation.

In 2010, Wahyu Caesarendra et al. proposed a method to predict failures before they occur by combining logistic regression and relevance vector machine. The method measures failure degradation using logistic regression, and the obtained results are used as vectors to train the relevance vector machine. The proposed method was tested on run-to-failure data by employing failure simulation data. To predict the unit of machine components, Kurtosis is calculated, which is a one-dimensional feature. The training performance was evaluated using correlation and root mean squared error. The study aimed to develop a reliable and accurate method to predict failures and prevent potential losses due to unexpected failures.

In 2010, Degang Chen and colleagues presented an approach to improve hard margin support vector machines (SVMs) using fuzzy rough sets. They proposed a fuzzy transitive kernel based on fuzzy rough sets to consider the membership of each training tuple. The membership of every training input was calculated using the lower approximation operator for binary classification. The performance of the proposed method was compared to fuzzy SVMs and soft margin SVMs. Experimental results indicated that the proposed method was effective, stable, and successfully combined fuzzy theory and SVMs. In summary, the authors showed that incorporating fuzzy rough sets into SVMs could improve their performance.

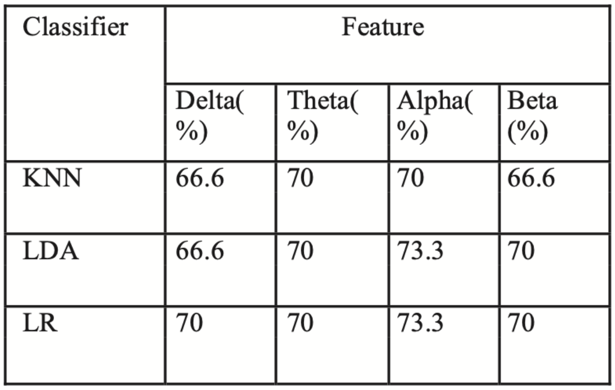
In 2010, Dursun Delen conducted a study to investigate the factors that could lead to the disintegration of first-year students and developed models to analyze and predict their retention. Five years of data from an educational institution were collected, and data mining techniques were applied to develop the models. The performance of the models was evaluated using the 10-fold cross-validation method, which involved dividing the dataset into 10 mutually exclusive subsets. The models were used to predict which students would retain and which ones would drop out before their second year. The results showed that support vector machines (SVM) outperformed logistic regression, decision trees, and neural networks in predicting student retention.

In 2010, Sajjad Ahmad and colleagues presented a regression technique using Support Vector Machines (SVM) to estimate soil moisture from remote sensing data. The SVM model was applied to 10 sites in the western United States using five years of training data from 1998 to 2002 and three years of testing data from 2003 to 2005. To evaluate the performance of the SVM, two models were developed. The first model involved training and testing six different models for six different sites. The second model combined the data from all six sites to create a single model which was then tested on the remaining four sites. The results showed that SVM outperformed other models such as MLR and ANN.

In 2012, Fan Min et al. proposed a solution for the problem of test cost constraint due to limited resources. They developed a feature selection method that considered four factors: constraint, input, optimization, and output objective. To address this problem, a backtracking algorithm was developed for small and medium-sized datasets and a heuristic algorithm for larger datasets. The algorithm was evaluated on four datasets, and the heuristic algorithm outperformed the backtracking algorithm in terms of efficiency and stability for datasets of all sizes.

Christian J. Schuler et al. in 2013 [15] developed a method for non-blind deconvolution to improve the brightness of blurred images. The proposed approach consisted of a two-step procedure where noise is amplified and colored to corrupt image information in the first step, and then an algorithm is used to remove colored noise in the second step. A neural network was used as a machine learning approach to sharpen the images. The proposed method was compared to existing methods, and the results demonstrated that it outperformed the other methods. The first step was performed by regularized inversion in the Fourier domain, followed by denoising using the neural network. The proposed method showed a significant improvement in image quality over existing techniques.

Behshad Hosseinifard et al. in 2013 [11], aimed to distinguish between depression and normal patients by analyzing non-linear features of EEG signals. They conducted their study on 45 normal patients and 45 depression patients and used several techniques to differentiate between the two groups, including logistic regression, linear discriminant analysis, and K-nearest neighbor. The data was trained using leave-one-out cross-validation method and tested on new datasets. Based on the experimental results, logistic regression achieved the highest accuracy and outperformed KNN and LDA. These findings suggest that non-linear analysis of EEG signals can be used to differentiate between depression and normal patients with high accuracy. In **figure 1**



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

In 2014, Nouman Azam and JingTao Yao addressed the problem of determining appropriate threshold values for the boundary, negative, and positive regions by examining the relationship between potential threshold changes and their impact on these regions. To explore this relationship, they employed the game theoretic rough set model, which can make intelligent decisions in situations with multiple criteria. They developed a game between frequent and prolonged regions and applied techniques to configure uncertainties in these areas. This allowed for the analysis of the model and decision-making based on the outcomes of the game.

In 2016, Eric J. Parish and Karthik Duraisamy introduced a modeling prototype called "field inversion and machine learning for physics applications" which directly infers information from data and reconstructs the inferred function using different parameters and variables for various problems. This approach aimed to generate general modeling information from the extracted information and embed the rebuilt function into an analytical solution process. By using this technique, possible errors can be identified at the initial level, instead of discovering them at the output level.

3.CONCLUSION

In this discussion, we have explored the concept of machine learning and its various applications in different fields. Machine learning is a process of developing new algorithms or models based on observations and analysis of existing data, and it differs from data mining, which involves only analyzing data without developing new models. We have seen how machine learning has been applied in diverse areas such as image deconvolution, student retention, detection of oil spills, and land cover changes.

The use of machine learning in these areas has shown promising results, with improved accuracy and efficiency compared to traditional methods. For instance, machine learning has been used to develop models for predicting student retention rates, which has helped institutions identify at-risk students and intervene before they drop out. Similarly, machine learning has been applied to detect oil spills in oceans and to monitor land cover changes, which has important implications for environmental protection.

Overall, our discussion has provided a glimpse into the vast potential of machine learning in various fields and the benefits it offers in terms of improved accuracy, efficiency, and decision-making.

REFERENCE

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-Lancina, P. J., Alba, E., Ribelles, N., Martin, 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, S$27-$31.

[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.

[91. 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 learing 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 IBEE 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.