iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** VII **Month of publication:** July 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

Machine Learning Prediction and Classification

R.Subha [1], R. Tamilarasan [2]
[1]*Assistant professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India*
[2]*Research Scholar, Department of Computer Science, PSG College of Arts and Science, Coimbatore,*

Abstract: Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so perceive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI. Machine Learning (ML) is a vast interdisciplinary field which builds upon concepts from computer science, statistics, cognitive science, engineering, optimization theory and many other disciplines of mathematics and science. There are numerous applications for machine learning but data mining is most significant among all. Machine learning can mainly classified into two broad categories include supervised machine learning and unsupervised machine learning.
Keywords: Machine Learning, Classification, Prediction.

I. MACHINE LEARNING INTRODUCTION

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter sub field focuses more on exploratory data analysis and is known as unsupervised learning. Machine learning can also be unsupervised and be used to learn and establish baseline behavioural profiles for various entities and then used to find meaningful anomalies. A supervised learning algorithm supervises the training data and produces a general rule (function), which can be used for mapping new inputs.

A. Machine learning is categorized into three broad categories

1) Supervised learning: Supervised learning is inferring a function from a given set of data (inputs with their respective outputs).The training data consists of a set of examples with which the computer is trained. Each example, a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm supervises the training data and produces a general rule (function), which can be used for mapping new input

a) *Supervised Learning Comes in Two Different Flavours:* We consider the each training case consists of an input vector x and a target output t.

b) *Regression:* The target output is a real number or a whole vector of real numbers such a price of stock in 6 months time or the temperature at noon tomorrow.

c) *Classification:* The target output is a class label like in the simplest case choosing between positive and negative. We can also have multiple alternative levels.

2) Unsupervised learning: Unsupervised learning is much harder because here the computer have to learn to perform specified tasks without telling it how to perform..

1) *Reinforcement learning:* In Reinforcement learning (RL) the output is an action or a sequence of actions and the only supervisory signal is an occasional scalar reward. The basic reinforcement method consists of: a set of environment sets S; a set of actions A; rules of transitioning between states; rules that determine the scalar immediate reward of a transition; rules that describes what the agent observes.

B. How Machine Learning Works

Machine learning algorithms are often categorized as supervised or unsupervised. Supervised algorithms require a data scientist or data analyst with machine learning skills to provide both input and desired output.

Unsupervised algorithms do not need to be trained with desired outcome data. Instead, they use an iterative approach called deep learning to review data and arrive at conclusions. Unsupervised learning algorithms also called neural networks are used for more complex processing tasks than supervised learning systems, including image recognition, speech-to-text and natural language generation. These neural networks work by combining through millions of examples of training data and

automatically identifying often subtle correlations between many variables. Once trained, the algorithm can use its bank of associations to interpret new data.

C. Examples of Machine Learning

Machine learning is being used in a wide range of applications today. One of the most well-known examples is Face book's News Feed. The News Feed uses machine learning to personalize each member's feed. If a member frequently stops scrolling to read or like a particular friend's posts, the News Feed will start to show more of that friend's activity earlier in the feed. Behind the scenes, the software is simply using statistical analysis and predictive analytic to identify patterns in the user's data and use those patterns to populate the News Feed.

D. Types of machine learning algorithms

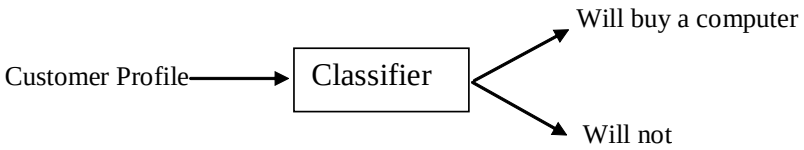
Just as there are nearly limitless uses of machine learning, there is no shortage of machine learning algorithms. They range from the fairly simple to the highly complex. Here are a few of the most commonly used models:

1) *Decision trees:* These models use observations about certain actions and identify an optimal path for arriving at a desired outcome.
2) *K-means clustering:* This model groups a specified number of data points into a specific number of groupings based on like characteristics.
3) *Neural networks:* These deep learning models utilize large amounts of training data to identify correlations between many variables to learn to process incoming data in the future.
4) *Reinforcement learning:* This area of deep learning involves models iterating over many attempts to complete a process. Steps that produce favourable outcomes are rewarded and steps that produce undesired outcomes are penalized until the algorithm learns the optimal process.

II. CLASSIFICATION

Classification is a process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood.

Classification is the process where computers group data together based on predetermined characteristics this is called supervised learning. There is an unsupervised version of classification, called clustering where computers find shared characteristics by which to group data when categories are not specified.



A. Classification

1) It makes things easier to find and recognise e.g. A fork is a piece of cutlery, so I will look in the cutlery draw
2) Classification can and should be used in sorting anything from documents to students

B. Hierarchies

1) To simplify classification hierarchies are set up, classifying into sub groups makes it much easier to find the object we are looking for.
2) Objects could fit into 2 categories in hierarchies if it is a different context.

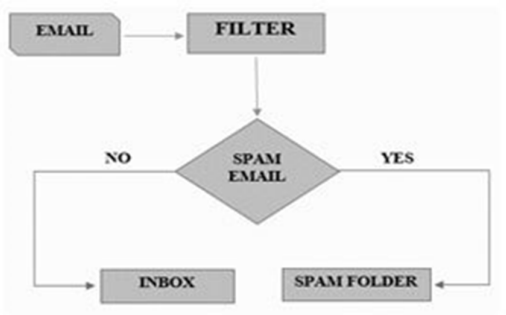
C. Classification Problems

Classification is an important tool in today's world, where big data is used to make all kinds of decisions in government, economics, medicine, and more. Researchers have access to huge amounts of data, and classification is one tool that helps them to make sense of the data and find patterns.

A common example of classification comes with detecting spam emails. To write a program to filter out spam emails, a computer programmer can train a machine learning algorithm with a set of spam-like emails labelled as spam and regular emails labelled as not-spam. The idea is to make an algorithm that can learn characteristics of spam emails from this training set so that it can filter out spam emails when it encounters new emails.

D. Few of the Terminologies Encountered in Machine Learning – Classification

1) *Classifier:* An algorithm that maps the input data to a specific category.
2) *Classification model:* A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.



3) *Feature:* A feature is an individual measurable property of a phenomenon being observed.
4) *Binary Classification:* Classification task with two possible outcomes. E.g.: Gender classification (Male / Female)
5) *Multi class classification:* Classification with more than two classes. In multi class classification each sample is assigned to one and only one target label. E.g.: An animal can be cat or dog but not both at the same time
6) *Multi label classification:* Classification task where each sample is mapped to a set of target labels (more than one class). E.g.: A news article can be about sports, a person, and location at the same time. The following are the steps involved in building a classification model:
8) *Initialize*: the classifier to be used.
9) *Train the classifier:* All classifiers in spirit-learn uses a fit(X, y) method to fit the model (training) for the given train data X and train label y.
10) *Predict the target:* Given an unlabeled observation X, the predict (X) returns the predicted label y.
11) *Evaluate* the classifier model.

E. Comparative Study

The similarities of the above algorithms are:

1) A machine learning algorithm learns from past experiences and produces an output based on the experiences.
2) The algorithms have strong relations to mathematical optimization.
3) The algorithms are related to statistical computation.

DISSIMILARITIES		
SUPERVISED LEARNING	UNSUPERVISED LEARNING	REINFORCEMENT LEARNING
The output is based on the training data set. Classification is used here.	The output is based on the clustering of data.	The output is based on the agent's interaction with the environment. It used deterministic or nondeterministic way of learning.
Priori is necessary.	Priori is not necessary.	Priori is required
It will always produce same output for a specific input.	It will produce different outputs on each run for a specific input.	The output changes if the environment does not remain same for a specific input.

F. Classification Problems in Real Life

Here are a few interesting examples to illustrate the widespread application of prediction algorithms.

G. Handwritten Digit Recognition

Goal is to identify images of single digits 0 - 9 correctly. The raw data comprises images that are scaled segments from five digit ZIP codes. In the diagram below every green box is one image. The original images are very small, containing only 16 × 16 pixels.

Figure 1.2: *Examples of handwritten digits from U.S. postal envelopes.*

Every image is to be identified as 0 or 1 or 2 ... or 9. Since the numbers are handwritten, the task is not trivial. For instance, a '5' sometimes can very much look like a '6', and '7' is sometimes confused with '1'.

To the computer, an image is a matrix, and every pixel in the image corresponds to one entry in the matrix. Every entry is an integer ranging from a pixel intensity of 0 (black) to 255 (white). Hence the raw data can be submitted to the computer directly without any feature extraction. The image matrix was scanned row by row and then arranged into a large 256 dimensional vector. This is used as the input to train the classifier. Note that this is also a supervised learning algorithm where Y, the response, is multi-level and can take 10 values.

III. WHAT IS PREDICTION?

(Numerical) prediction is similar to classification construct a model use model to predict continuous or ordered value for a given input Prediction is different from classification Classification refers to predict categorical class label Prediction models continuous-valued functions Major method for prediction: regression model the relationship between one or more independent or predictor variables and a dependent or response variable Regression analysis Linear and multiple regression Non-linear regression Other regression methods: generalized linear model, Poisson regression, log-linear models, regression trees.

A. Prediction

1) Models continuous-valued functions, i.e., predicts unknown or missing values

B. Example

1) A marketing manager would like to predict how much a given costumer will spend during a sale
2) Unlike classification, it provides ordered values
3) Regression analysis is used for prediction
4) Prediction is a short name for numeric prediction

C. The Prediction Problem

Prediction of future outcomes of systems must be couched in terms of probability Statements. For example, one can predict whether or not it will rain tomorrow in Point Pleasant, on the coast of New Jersey, as follows:

1) *First Prediction:* Given the historic data on rainy days in Point Pleasant for the past 20years, one can predict the probability of rain by dividing the number of rainy days by the total number of days. If the number of rainy days for the past 20 years (7300 days) was 730, then the probability of rain tomorrow (or any day for that matter) is 10%. This is a Statistical estimate based upon historic data.
2) *Second Prediction:* Given the number of rainy days in each month in Point Pleasant for the past 20 years, one can make separate predictions of the probability of rain for each month.

3) *Third Prediction:* Using knowledge of the weather patterns around the coast of New Jersey, one can generally rely upon the fronts moving from west to east. Given Knowledge about a rainstorm heading toward Point Pleasant from Pennsylvania, one can predict the probability of rain over the next 24 hours in 6 hour increments.

All of these predictions may contain valid probability statements based upon historic
Measurements. However, the accuracy of each is obviously different. The difference in Accuracy is determined by the conditioning of the probability statement. The first prediction is conditioned only upon the number of rainy days in a year, with no additional information. The second prediction is conditioned upon additional information, i.e., the number of rainy days in each month of the year. It will be a more accurate statement. The third prediction is conditioned upon a dynamic model of weather patterns. This model contains much more information than the other two, and is much more accurate.

B. The Future of Machine Learning

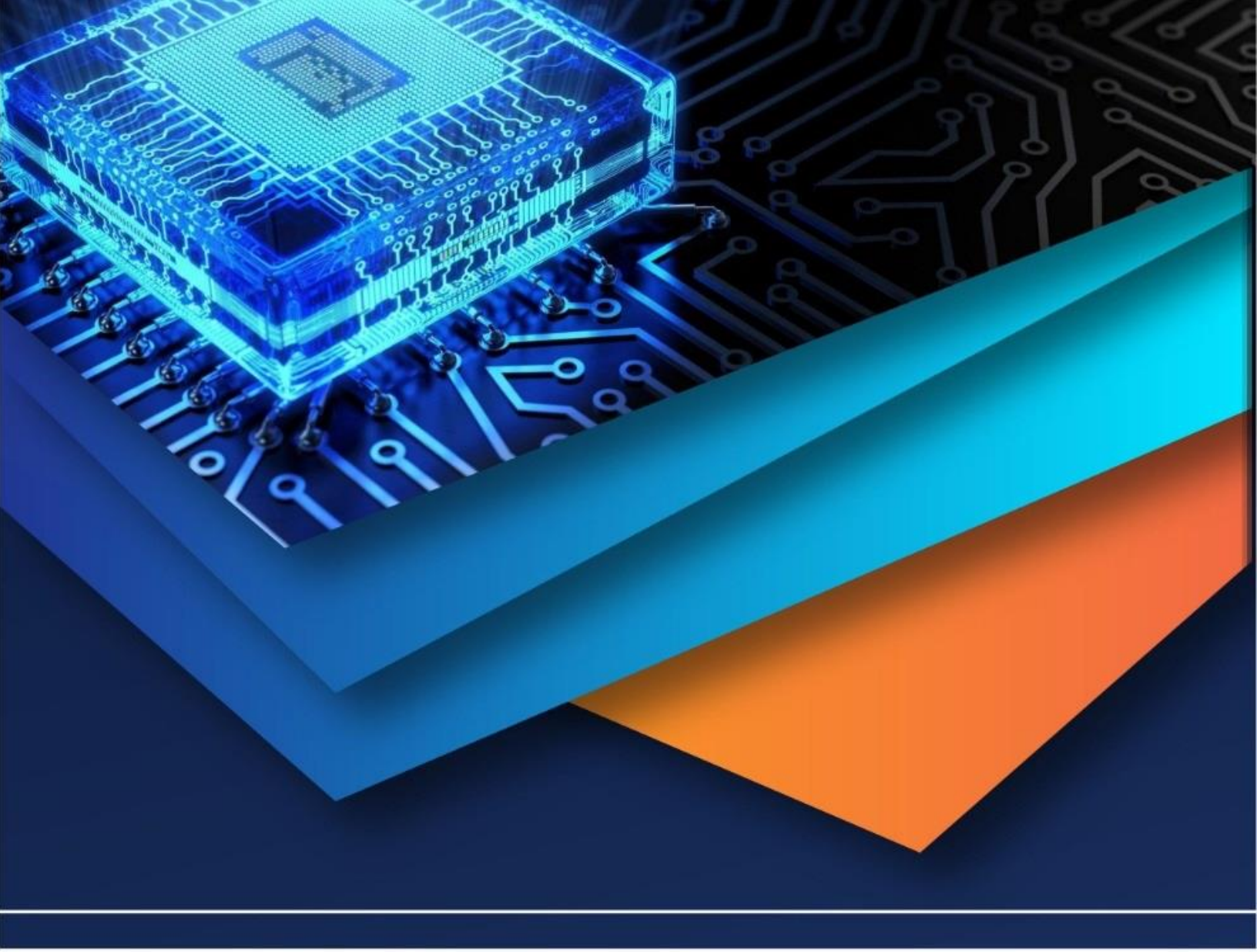
While machine learning algorithms have been around for decades, they've attained new popularity as artificial intelligence (AI) has grown in prominence. Deep learning models in particular power today's most advanced AI applications. Machine learning platforms are among enterprise technology's most competitive realms, with most major vendors, including Amazon, Google, Microsoft, IBM and others, racing to sign customers up for platform services that cover the spectrum of machine learning activities, including data collection, data preparation, model building, training and application deployment. As machine learning continues to increase in importance to business operations and AI becomes ever more practical in enterprise settings, the machine learning platform wars will only intensify. Continued research into deep learning and AI is increasingly focused on developing more general applications. Today's AI models require extensive training in order to produce an algorithm that is highly optimized to perform one task. But some researchers are exploring ways to make models more flexible and able to apply context learned from one task to future, different tasks.

IV. CONCLUSION

These days, machine learning techniques are being widely used to solve real-world problems by storing, manipulating, extracting and retrieving data from large sources. Supervised machine learning techniques have been widely adopted however these techniques prove to be very expensive when the systems are implemented over wide range of data. This is due to the fact that significant amount of effort and cost is involved because of obtaining large labelled data sets. Presents an evaluation of state-of-the-art machine learning algorithms on the basis of efficiency, for the task of classification. Machine learning algorithms perform more efficiently for a classification task when they are combined together. For the prediction of the correct output class, combined learner selects the class to which highest probability has been assigned among all the learners.

REFERENCES

[1] Kesavaraj G, Sukumaran S. A study on classification techniques in data mining. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, 2013; pp. 1-7.
[2] Sharma S, Agrawal J, Agarwal S. Machine learning techniques for data mining: A survey, in Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on, 2013; pp. 1-6.
[3] Rizwan M, Anderson DV. Using k-Nearest Neighbor and Speaker Ranking for Phoneme Prediction, in Machine Learning and Applications (ICMLA), 2014 13th International Conference on, 2014; pp. 383-387.
[4] Kasemsumran P, Auephanwiriyakul S, Theera-Umpon N. Face recognition using string grammar fuzzy K-nearest neighbor, in 2016 8th International Conference on Knowledge and Smart Technology (KST), 2016; pp. 55-59
[5] Viswanath P, Sarma TH. An improvement to k-nearest neighbor classifier, in Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE, 2011; pp. 227-231.
[6] Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moore "Reinforcement Learning: A Survey", Journal of Artificial Intelligence, Research 4 (1996) 237-285, May 1996.
[7] R. Sathya, Annamma Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013.
[8] R. Sathya and A. Abraham, "Unsupervised Control Paradigm for Performance Evaluation", International Journal of Computer Application, Vol 44, No. 20, pp. 27-31, 2012
[9] J.A. Hartigan and M.A. Wong, "A K-Means Clustering Algorithm", Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1 (1979), pp. 100-108
[10] Sharma S, Agrawal J, Agarwal S. Machine learning techniques for data mining: A survey, in Computational Intelligence and Computing Research (ICCIC), 2013 IEEE InternatioSnal Conference on, 2013; pp. 1-6.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY