National Conference on Recent Trends in VLSI, Communication and Networks 2018 (RTVCN2K18) ISSN: 2454-4248
Volume: 4 Issue: 5 53 – 56

A Review on Machine Learning Algorithms

Sreehary P L[1], Manasi Sandeep[2], Sreelakshmi Sreekumar[3], Vignesh V[4], Divya V Chandran[5]
[1,2,3,4]UG Scholar, [5]Assistant Professor
Department of Electronics and Communication Engineering
AdiShankara Institute of Engineering and Technology
Kalady, Kerala

Abstract—In machine learning a computer learns to perform a task by studying a training set of examples. The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown. This paper describes various machine learning classification techniques and a comparison of the same.

Keywords— *Machine learning, Classification, Neural networks, Clustering,Supervised learning*

*****

I. INTRODUCTION

II.

Over the past two decades Machine Learning has become one of the pillars of information technology and a rather central part of our life. With the ever increasing amounts of data becoming available there is good reason to believe that smart data analysis will become even more pervasive as a necessary ingredient for technological progress. Machine learning research strives to open possibility of instructing computers in such new ways, and thereby promises to ease the burden of hand-programming growing volumes of increasingly complex information into computers of tomorrow [1].The purpose of this paper is to provide the reader with an overview of different classification algorithms which have vast range of applications in different fields of science and technology. Finally, we have outlined a set of basic effective algorithms to solve an important problem, namely that of classification.

III. LITERATURE SURVEY : MACHINE LEARNING

Machine learning is the science of getting computers to learn and act without being explicitly programmed [2]. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithm that can learn from make predictions on data through building a model from sample inputs [3]. Machinelearning tasks are classified into two broad categories,depending on whether there is a learning "signal" or "feedback" available to a learning system [4]:

- Supervised learning
- Unsupervised learning

A. Supervised Learning

B.

Supervised learning is the machine learning task of inferring a function from labelled training data. The training data consist of a set of training examples. In supervised machine learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (supervisory signal). A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new sets of data. A popular example of supervised learning outside the field of high content screening is the classification of handwritten digits in an image for automated sorting of postal zip codes. In this example the input is the pixel intensities in the image, and the output is the classification of a number (0, 1, etc) [4]

C. Classification

Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations or instances whose category membership is known. In classification the inputs are divided into two or more classes, and a learner must produce a model that assigns unseen inputs to one or more of these classes [5]. A classification problem is when the output variable is a category, such as “Red” or “blue” or “disease” and “no disease”. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of

the patient (gender, blood pressure, presence or absence of certain symptoms, etc.)[6].

D. Regression

Regression analysis is a set of statistical processes for calculating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or predictors).More specifically, this analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables are varied, while the other independent variables are held fixed [7].Regression analysis is used for prediction and forecasting. A regression problem happens when the output variable is a real or continuous value, such as "salary" or "weight". Many different models can be used; the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points. Logistic regression takes some inputs and calculates the probability of some outcome. For example, if a child has a temperature of 104F (40C) and they have a rash and nausea then the probability that they have chickenpox might be 80%. A rule of thumb in logistic regression is if the probability is > 50% then the decision is true. So in this case, the determination is made that the child has chickenpox [8].

E. Unsupervised Learning

Unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from "unlabelled" data (a classification or categorization is not included in the observations). Since the examples given to the learner are unlabelled, there is no evaluation of the accuracy of the structure that is output by the relevant algorithm—which is one way of distinguishing unsupervised learning from supervised learning and reinforcement learning. A central case of unsupervised learning is the problem of density estimation in statistics, though unsupervised learning encompasses many other problems (and solutions) involving summarizing an explaining key features of the data [9].Unsupervised learning classified into two categories of algorithms: clustering and association.

F. Clustering

Clustering is grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a common technique for statistical analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression and computer graphics. Cluster analysis as such is not an automatic multi-objective optimization that involves trial and failure [10].

IV. CLASSIFICATION OF MACHINE LEARNING ALGORITHMS

A. Decision Tree Algorithm

A decision tree constructs a tree like structure involving of possible solutions to a problem based on certain constraints. It is so named for it begins with a single simple decision or root, which then forks off into a number of branches until a decision or prediction is made, forming a tree [11]. The three basic algorithms are widely used that are ID3, C4 .5 and CART [12].

1) ID3:

It used to generate a decision tree from a dataset.

Steps:

- Calculate the entropy of every attribute using the data set S.
- Split the set into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or, equivalently, information gain is maximum).

- Make a decision tree node containing that attribute.
- Recurse on subsets using remaining attributes.
-

2) C4.5:

C4.5 builds decision trees form a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S=S1, S2, S3… of already classified samples. Each sample Si consists of a p-dimensional vector (X1, i, X2,i,……. Xp, i.), where the Xj represent attribute values or features of the sample, as well as the class in which Si falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other and this gives the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists [12].

3) CART:

The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.

B. Bayesian Networka Bayesian Network (BN):

It is a graphical model for relationships among a set of various variable features [12]. A directed acyclic graph (DAG)

is a finite directed graph with no directed cycles. That is, it consists of finitely many vertices and edges, with each edge directed from one vertex to another, such that there is no way to start at any vertex v and follow a consistently-directed sequence of edges that eventually loops back to v again. Equivalently, a DAG is a directed graph that has a topological ordering, a sequence of the vertices such that every edge is directed from earlier to later in the sequence.

1) Naive Bayes Classification:

Naive Bayes classifiers are simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Bayes theorem is shown in equation (1).

$$P(H|X) = P(X|H)\,P(H)\,/P(X) \quad (1)$$

X - Some evidence, describe by measure on a set of attributes.

P (H|X) - posterior probability that the hypothesis H holds given the avid.

X P (H) - prior probability of H, independent of X.

P(X|H) - posterior probability that of X conditioned on H.

Advantages

- Neural networks are able to handle noisy data, classify patterns untrained data on which they are not being trained.
- Well suited for continuous feature valued inputs and outputs.

Disadvantages

- Training time will be large.
- Poor interpretability [12].

C .k-Nearest Neighbor (kNN):

It is based on the principle that the instances within a dataset will generally exist in close proximity to other instances that have similar properties. If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbors. The kNN locates the k nearest instances to the query instance and determines its class by identifying the single most frequent class label [13]. Its role implicitly computes the decision boundary and it is also possible to compute the decision explicitly. So, the computational complexity of NN is the function of the boundary complexity [12].

Advantages:

- Easy to understand and implement classification technique.

Disadvantage:

- Computational costs are expensive, when sample is large.

D. Support Vector Machines (SVMs):

Support Vector Machines are the newest supervised machine learning technique [13].Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. SVMs are well suited to deal with learning tasks where the number of features is large with respect to the number of training instances [13].

Advantages:

- Accurate methods among all machine learning algorithms. It finds the best classification function of training data.
- SVM prevents over fitting than other methods

Disadvantages:

- It is computationally expensive.

An algorithm which implements classification especially in a concrete implementation is known as a classifier. Therefore choosing an apt classifier is important. If your training data is small, NB can be selected whereas if dataset is high kNN would yield high accuracy.

IV. RESULT

Our aim is to find the best classifier suitable for classification of driving character. The classifier must be accurate as well as time required for classification must me minimum. The features selected for training the classifiers where Throttle position, RPM, Engine Load and speed. To test the accuracy and speed of classification of the above-mentioned classifiers, classifiers where created using sci-kit learn in python. Each of the four classifiers where trained using the same training data set. A training dataset is a dataset of examples used for learning/training that is to fit the parameters.

Data for classification where collected doing real drives with drivers having different driving character. The data set contained data of normal, slow and rash drives. After the training of the classifier testing of the classifier was done. For this a test data set was used.

TABLE I

Classifier	Decision Tree	KNN	Naive Bayesian	SVM
Person 1	75.6 %	81.5 %	84.5%	63.1%
Person 2	72.6 %	70.2 %	89.1%	53.1%
Person 3	74.2 %	61.7 %	86.2%	54%
Person 4	86%	79.5 %	87%	70%
Person 5	83.1 %	77.3 %	86%	65.5%
Speed of classifier	High	High	High	Average

Testing data set contained 450 samples, 150 samples each from slow, normal and rash drive data sets. The testing data set was given to the classifier to generate its predictions about the type of driving style. The predictions from the classifier and the correct output for the corresponding training data set samples, where compared to get the accuracy of the classifier. The total time taken to classify 450 sets of value was also noted. To further check the accuracy drives were conducted
using five different drivers. The parameter values where recorded and saved in csv file. These data where fed to the classifier to check the accuracy. The results obtained are shown in the Table-I.

I. CONCLUSION AND FUTURE DIRECTIONS

In this paper we discussed about how driving pattern is recognized using machine learning. Primarily the car parameters are collected from real drives. Different classifiers are trained using the data collected. Accuracy of these classifiers was obtained. The accuracy and speed of classification will vary according to the type of features selected. From all these observations we came to the conclusion that the best suited classifier for our particular application is Naïve Bayes classifier, which showed the highest accuracy.

REFERENCES

[1] R.S. Michalski, J.G. Carbonell, T.M. Mitchell, “Machine Learning: An Artificial Intelligence Approach”, *2013.*

[2] Samuel, Arthur, "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal of Research and Development,1959.*

[3] Nasser M.Nasrabadi, “ Pattern Recognition and Machine Learning”, *October 1,2007.*

[4] Steven A.Haney,DouglasBowman,ArjithChakravarty,“ An Introduction to High Content Screenign”,November 24,2014.

[5] R. Kohavi and F. Provost,Glossary of terms," Machine Learning”, *vol. 30, no. 2-3, pp. 271-274, 1998*

[6] Alpaydin, Ethem , “Introduction to Machine Learning*”*,2010.

[7] Armstrong,J.Scott, "Illusion in Regression Analysis”,*International Journal of Forecasting (forthcoming). 28 (3): 689,2012.*

[8] MehryarMohri, AfshinRostamizadeh, AmeetTalwalkar , “ Foundations of Machine Learning”, 2012.

[9] Hastie, Trevor, Robert Tibshirani, Friedman, Jerome ,“The Elements of Statistical Learning”: *Data mining, Inference, and Prediction. New York*,2009.

[10] Estivill-Castro, Vladimir,"Why so many clustering algorithms — A Position Paper". *ACM SIGKDD Explorations Newsletter. 4 (1):pp. 65–75,June 20,2002.*

[11] Kajaree Das, Rabi Narayan Behera, “A Survey on Machine Learning: Concept, Algorithms and Applications,” *International Journal of Innovative Research in Computer and Communication Engineering. Vol. 5, Issue 2, February 2017.*

[12] Seema Sharma, JitendraAgrawal, ShikhaAgarwal, Sanjeev Sharma, “Machine Learning Techniques for Data Mining: A Survey,” *Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on. 27 January 2014.*

[13] S. B. Kotsiantis, I. D. Zaharakis, P. E. Pintelas, “Machine learning: a review of classification and combining techniques,” *Springer Science + Business Media B.V. 2007. Published online: 10 November 2007*