ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

PREDICTIVE ANALYTICS IN MACHINE LEARNING

Kiranjit Kaur

Assistant Professor, Department of Computer Science, Guru Nanak College, Moga

Parminder Kaur

Associate Professor, Department of Computer Science, Khalsa College for Women, Civil Lines, Ludhiana

ABSTRACT

Machine learning algorithms detect hidden patterns from the data and predict the output based on these patterns. These algorithms accumulate knowledge from experiences and undergo a continual process of learning and self-improvement. This results in an improved and refined performance over time. The identified patterns and rules are mathematical in nature, and they can be easily defined and processed by a computer. This paper focuses on different classification techniques of supervised machine learning with emphasis on predictive analytics. Predictive analytics involve leveraging machine learning to predict future outcomes.

KEYWORDS: Artificial Neural Networks, K-Nearest Neighbour (KNN), Support Vector

Machine (SVM), Naive-Bayes Classification (NBM), Random Forest Classification (RFC)

1. INTRODUCTION

With the wide usage of computers and internet, huge amount of data is generated from social sites and web traffic. This big data can be a problem and an opportunity. The problem is that it is difficult for us to analyze such large data. The opportunity is that this type of data can be processed by computers in much fast and accurate way by storing them digitally Concept of machine learning is something which comes out of this environment. The computer can then use those rules to meaningfully characterize new data. The creation of rules from data is an automatic process, and it is something that continuously improves with newly presented data.

Applications of machine learning cover a wide range of areas. Search engines use machine learning to better construct relations between search phrases and web pages. By analyzing the content of the websites, search engines can define which words and phrases are the most important in defining a certain web page, and they can use this information to return the most relevant results for a given search phrase (D. Kalles et al ,2006). Image recognition technologies also use machine learning to identify particular objects in an image, such as faces (Ethem Alpaydin, 2004). First, the machine learning algorithm analyzes images that contain a certain object. If given enough images to process, the algorithm is able to determine whether an image contains that object or not (Ian H. Witten et al , 2016). In addition, machine learning can be used to understand the kind of products a customer might be interested in. By analyzing the past products that a user has bought, the computer can make suggestions

about the new products that the customer might want to buy (D. Kalles et al, 2006). Increase in data has made these applications more effective and thus more common in use. Depending on the type of input data, machine learning algorithms can be divided into supervised and unsupervised learning.

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

1.1 SUPERVISED LEARNING

In this system, it is assumed that the desired response of the system is obtained at each moment of application of the input. It tries to predict the results of known examples. Such a system compares its predictions with known results and learns from them. In supervised learning, input data comes with a known class structure (Jeremy Watt et al, 2016). This input data is known as training data. The algorithm is usually tasked with creating a model that can predict one of the properties by using other properties. After a model is created, it is used to process data that has the same class structure as input data.

1.2 UNSUPERVISED LEARNING

In this mode, the required response is unknown, so clear error information cannot be used to improve network behavior. There is no information available for the correctness or inaccuracy of the response. To some extent, learning can be done on the basis of observation of the response to the input. In unsupervised learning, input data does not have a known class structure, and the task of the algorithm is to reveal a structure in the data (M. J. Islam et al, 2007).

This paper focuses on classification techniques with supervised machine learning, which is the process of using machine learning to predict future outcomes. Predictive analytics has a wide range of applications, such as fraud detection, analyzing population trends, or understanding user behavior (M. Mohri et al, 2012). Classification is a method that organizes the unstructured data into structured classes or groups, which in-turn helps the customers in knowledge discovery and furthermore provides intellectual decision making. There are two stages in classification, i.e. training and testing. Training is a learning process stage in which huge training data sets are provided and investigation takes place then instructions and designs are generated. Afterwards, execution of the second phase initiates i.e. assessment or testing of data sets and archives the precision of a classification patterns.

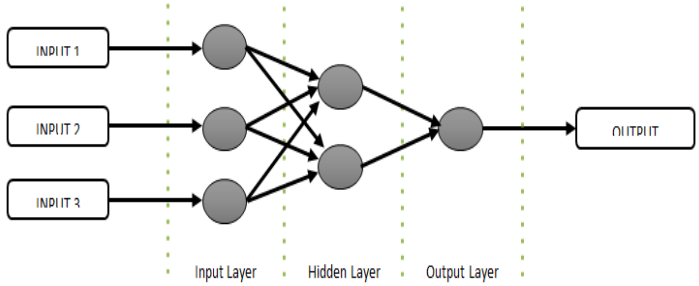
2. TECHNIQUES OF CLASSIFICATION

2.1 ARTIFICIAL NEURAL NETWORK (ANN)

Artificial Neural Network (ANN) is basically a network model which is proficient of recognizing complex nonlinear connections among the input and output data clusters. The ANN models are proved to be helpful and perform efficiently where some particular problems are challenging to be answered by means of physical equations.ANN is primarily based upon neural structure of the brain. Human brain usually learns everything by experience. Some of the problems that are beyond the reach of current computers are certainly deciphered by trivial energy proficient packages. ANN ultimately ensures a less technical approach to develop machine solutions (Nitish Srivastava et al, 2014). Artificial Neural Networks are statistical learning algorithms which is one of the most effective ways for performing pattern recognition and data classification. These biologically inspired methods of computing are thought to be the next major advancement in the computing industry. Human brain is capable of functions that are currently impossible for computers. Whereas, computers only do things efficiently that are based on memorization by repetition like keeping records and executing multidimensional arithmetic functions.

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

Fig.1 - Layered Architecture of Artificial Neural Network (ANN)



An Artificial Neural Network is an analytical model which usually attempts to pretend the composition and purpose of the biological neural networks (Nitish Srivastava et al, 2014). The elementary structure of the artificial neural network is normally comprised of three different layers that are broadly known as input layer, hidden layer and output layer respectively. The primary elements of all artificial neural networks are closely interlinked artificial neuron imitated by multiple nodes where each node accepts an input, implements a function to it and after that transfers the output. At every node the output is referred as its activation or it is also known as node value. Each link is associated with weight. ANNs are capable of learning, which takes place by altering the weight values. In the input layer, the inputs values are weighted i.e. each input value is multiplied with individual weight. In the hidden layer, neurons implement sum function which aggregates all the weighted inputs. In the output layer of ANN, the aggregation of previously weighted inputs is carried through activation function in order to generate the output value. The methodology of Artificial Neural Network is defined below:

The initialization of the artificial neural network is the main step of accepting three parameters, namely training data, the number of each element belonging to the data group and the hidden neurons in the hidden layer. Finally, specify the number of iterations.

There are two parameters of validations and to stop the artificial neural network training.

The first is the validation checks Either the validation gets complete or the Gradient becomes equal to the decided gradient of the neural network that would be sufficient enough to stop the training of the artificial neural network.

After that, data is passed to artificial neural network for training purpose of data which create the category according to the properties of data. ANN toolbox has a range of architectures of supervised and unsupervised network With the toolbox's modular approach to build a network, custom network architectures can be developed for specific problem. The network architecture can be developed including all inputs, layers, outputs, and interconnections. After that, numbers of iteration are given that will be sufficient for complete training.

Later, the best gradient value, mutation value and validation checks is being find out. The gradient magnitude, mutaion and the validation checks number are utilized for terminating the training.

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

Gradient: The gradient would become very little if the training accomplish a performance minimum. If the gradient magnitude is very less, than the training will stop. The maximum value of mean square error is called gradient. It falls with increase in number of iterations.

Mutation (mu): It is the control parameter for the algorithm used to train the neural network.

Validation: The number of times the verification check represents the number of consecutive iterations that cannot be reduced by the verification performance. If the number reaches 6 (the default), the training will stop. It specifies the number of times each iteration is verified

The next step in verifying the network is to create a regression graph that shows the relationship between the network output and the target. If the training is perfect, the network output and the target will be exactly the same, but in fact the relationship is rarely perfect.

2.2 KNN ALGORITHM (KNN)

The K-nearest neighbour procedure (KNN) is a way for classification of entities on the basis of the adjoining training specimens in feature space. The prime intention of the k Nearest Neighbours (KNN) process is to use the database wherein the data are divided into a number of isolated classes to prognosticate the classification of a new sample point (Sas, 2017). KNN classification distributes the data into test set and training sets. Then the K nearest training set objects are originated for every single row of the test set, and the process or task of classification is performed by predominance vote with connections which can be broken at any moment.

The K-nearest neighbour algorithm (KNN) can be summarized as:

A positive number k is stated, with a new sample

The k items are selected from the database that are next to new sample

The utmost mutual classification of selected entries is determined.

Resulted Classification is offered to the new sample.

In KNN classification, the output is a class membership. An object is classified through the bulk vote from the nearby neighbours, with entity being allocated to class most mutual among the entities k adjoining neighbours.

If k = 1, the object is assigned to class of that sole nearest neighbour. A peculiarity of KNN algorithm is that its sensitivity to local structure of data.

Assume, training set D

1. Object to be tested x = (x_, y_),

2. After that algorithm calculates the similarity between z and all training objects to conclude its nearest-neighbour list i.e. Dz.

3. Training objects =(x, y) ∈ D

4. x = data of a training object,

y= is its class.

Similarly, x_ = data of the test object

y_= is its class

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

Once the list of nearest-neighbour is acquired. The classification of test object is done on the basis of majority class of its nearest neighbours which is describes in the equation below

$$MajorityVoting: y' = argmaxv \sum \quad I(v = y_i), (x_i, y_i) \epsilon D_z$$

In the above equation;

v = class label

y_i=class label for ith nearest neighbours

$I(\cdot)$ = indicator function which returns the value 1

If its argument = true and otherwise 0 is returned as a value.

An Example of the *k*-NN classification has been explained briefly along with the figure representation as follows:

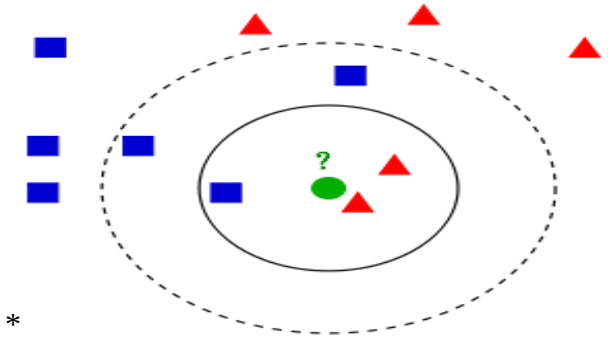


Fig. 2 - KNN classification paradigm

In the above figure, it is demonstrated that the test model (i.e. green colored circle) can be classified either to first class of the blue colored squares or to the other class of red colored triangles. If $k = 3$, (considering solid line circle) then the test model is allocated to the second class as there are 2 triangles inside the inner circle and only 1 square. Whereas, if $k = 5$, (considering the dashed line circle), the test model is allotted to the first class since there are 3 squares inside the outer circle and only 2 triangles. The allocation is based on the majority vote of its neighbour.

Euclidean Distance can be calculated by using:

$$D(x, y) = \sqrt{\sum_{i=1}^{n} \quad (x_i - y_i)^2}$$

K-Nearest Neighbour can be predicted by employing the following equation:

$$y = \frac{1}{k} \sum_{i=1}^{x} \quad y_i$$

In the above equation,

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

y_i = i^{th} case of test model

y = outcome of the query point.

In classification problems, on a voting scheme the KNN predictions are based and the winner is used to label the query.

Pseudo code for *K*NN Algorithm

K ⟵number of nearest neighbors

For each object *X* in the test set **do**

Calculate the distance D(*X*, *Y*) between *X* and every object *Y* in the training set

Neighborhood ⟵the *k* neighbours in the training set closest to *X*

X.class⟵SelectClass (*neighborhood*)

End for

2.3 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is an algorithm that was developed for pattern classification but has recently been adapted for other uses, such as finding regression and distribution estimation. It has been used in many fields such as bioinformatics, and is currently a very active research area in many universities and research institutes. Although the SVM can be applied to various optimization problems such as regression, the classic problem is that of data classification. In machine learning, task of deducing a category from supervised training data is known as Supervised Learning. In supervised learning the training data consist of a set of training examples, where each example is a pair consisting of an input and an anticipated output value. A supervised learning algorithm examines the training data and then predicts the correct output categorization for given data-set input. Support Vector Machine (SVM) was introduces by Vapnik in 1979, and it is primarily defined as a supervised learning approach that utilizes a subcategory of training point which is also acknowledged as support vectors in order to categorize diverse entities. SVM basically finds the optimal linear decision surface concerning the binary classes. The biased arrangement of the support vectors is generally known as decision surface. In other words, the nature of the margin among the two classes is decided by the support vectors.

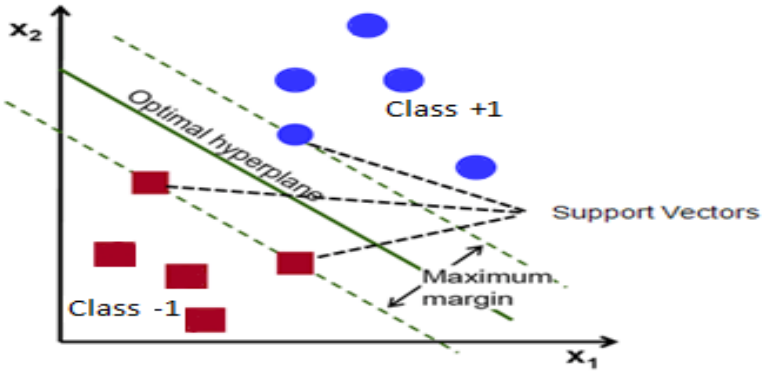


Fig. 3- Support Vector Machines

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

The above figure only shows the 2-dimensional case where the data points are linearly separable. The red boxes and blue circle are the support vectors that are primarily the observations which supports hyper plane on both sides. The hyper plane is simply defined as a line in 2D, whereas in 3D it defined as plane. In greater dimensions (i.e. more than 3D), it is entitled as hyper plane. SVM aid to discover a hyper plane or unscrambling boundary that can isolate two classes. Margin is well-defined as the space among the hyper plane and the nearest data point. If that space is doubled up, it would be identical to the margin.

Internally, SVMs are quadratic optimizers that search for a hyper plane that best separates the classes. A maximum margin SVMs strives to maximize the distance of this hyper plane to the support vectors; i.e. examples from different classes that fall either side of the hyper plane.

Support Vector Machines are versatile, for different decision function. SVMs are very effective when a very high dimensional spaces are concerned. Also, when the number of dimensions turn out to be bigger than the current number of samples, then the SVM's appeared to be very effective.

2.4 NAIVE-BAYES CLASSIFICATION

This classification type is primarily based on Bayes' Theorem, named after Thomas Bayes. The Bayesian classifiers mainly represents supervised learning methods and are statistical in nature which can predict the possibilities of specific class memberships. Several researchers tried to construct a function explicitly from a mutual set of values of certain aspects to the class labels. The Bayesian classification delivers a valuable view for indulging and assessing certain learning algorithms. This classification method can be used to resolve the diagnostic and prognostic problems. It normally determines explicit possibilities for assumption and it is robust to noise in input data.

Due to the simplicity of the Naïve Bayes technique in permitting each attributes to assist equally and individually concerning the final decision, it has become very popular for machine learning applications. Such simplicity also makes this technique eye-catching and appropriate for several fields. One of the best example of Naive Bayesian classification usage is Spam Filtering. Naive Bayes Classifiers are employed in order to classify the spam emails. In spite of its impractical independence theory, the Naive Bayes classifier is remarkably effective in practice since its classification decision may frequently be precise even if its probability approximations are imprecise.

Naive Bayes classifiers are hardly naïve (Sas, 2017). In fact, they offer a range of important services such as learning from very large data sets, incremental learning, anomaly detection, row pruning and feature pruning—all in near linear time (i.e. very fast). The reason these classifiers are called naive is that they assume that within one class, all features are statistically independent. That is, knowledge about the value of one feature does not tell us anything about the value of any other. So, a Naive Bayes classifier can never look at table of medical diagnosis to infer that *pulse=0* is associated with *temperature=cold.*

2.5 RANDOM FOREST CLASSIFICATION

Random forests (RF), also known as random decision forests is a supervised machine learning approach, which was first proposed by Tin Kam Ho. Random forests is a popular technique, which can be utilized for several purposes i.e. for classification, prediction, learning variable

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

importance, variable selection, and outlier discovery. It is a combination of the random sub space method that is proposed by Tin Kam Ho and bagging.

The basic idea of Random Forest is to generate various small decision trees from the random subsets of the data. As only the subsets of the data are considered, therefore each decision tree delivers a biased classifier which acquire varied developments in the data. This represents a group of the several experts, all having trivial information about the entire topic but complete in their region of expertise. Now, the majority vote is considered in case of classification in order to classify the class. As soon as the forest is put up to classify a new instance, it run across all the trees that are grown-up in the forest. Each tree provides classification for the new instance which is noted as a vote. The votes received from all the trees are pooled and the class having maximum number of votes is established as classification of the new instance (Tom M. Mitchell, 1997).

Once the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This oob (out-of-bag) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

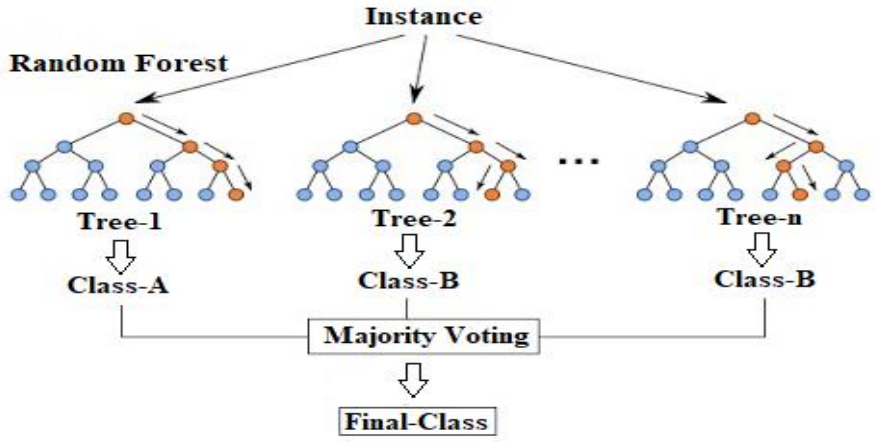


Fig. 4 - Random Forest Classification

3. CONCLUSION

In this paper, various Machine Learning techniques have been discussed for classification and predictive analysis. With the help of these techniques, data analysis is no more a difficult job. It replaces traditional methods of data analysis. In future we can use them in real environment by combining them with IoT. Combining IoT and machine learning techniques can lead to higher productivity in data analysis.

REFERENCES

D. Kalles and C. Pierrakeas. 2006. Analyzing student performance in distance learning with genetic algorithms and decision trees. Applied Artificial Intelligence 20(8), 655- 674

Ethem Alpaydin. 2004. Introduction to Machine Learning. Cambridge, MA.

Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

Jeremy Watt, Reza Borhani, and AggelosKatsaggelos. 2016. Machine Learning Refined: Foundations, Algorithms, and Applications. Cambridge University Press.

M. J. Islam, Q. J. Wu, M. Ahmadi, and M. A. Sid-Ahmed. 2007. Investigating the performance of naive-Bayes classifiers and k-nearest neighbor classifiers. In: International Conference on Convergence Information Technology. IEEE, 1541-1546.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. 2012. Foundations of Machine Learning (Adaptive Computation and Machine Learning Series). MIT Press.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, IlyaSutskever, and RuslanSalakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1), 1929-1958.

Sas. 2017. Predictive Analytics: What it is and why it matters, SAS. https://www.sas.com/en_us/insights/analytics/predictive-analytics.html. Retrieved April 24, 2017.

Tom M. Mitchell. 1997. Machine Learning. McGraw-Hill.