IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Special Issue-17, May 2019 E-ISSN: 2347-2693

Artificial Intelligence in Machine Learning Techniques for Clustering and Classification

M. Angelin Rosy[1*], M. Chaya[2], M. Felix Xavier Muthu[3]

[1,2,]Dept. of MCA, Er.Perumal Manimekalai College of Engineering, Anna University, Hosur, India
[3]Dept. of Mechanical Engineering, St.Xavier's Catholic College of Engineering, Anna University, Nagercoil, India

Corresponding Author: angel_rosym@yahoo.co.in, Tel- 9944579754

Available online at: www.ijcseonline.org

Abstract— Data mining is the search for hidden relationships in data set. Machine learning is implementing some of artificial learning. Machine learning is the ability to alter an existing model based on new information .Machine learning is mainly used for business learning to identify the information. The paper evaluates the performance of clustering and classification.Clustering analysis is one of the main analytical methods in machine learning. Machine learninmg is one of the leading fields where clustering is one of the significant task. Classification methods to improve business opportunity and to improve the quality os services. The machine learning in the computer system use to effectively perform a specific task without using explicit instruction.

Keywords: Data mining , Machine learning ,Clustering ,Classification ,Artificial learning, Relationship

I. INTRODUCTION

Machine learning(ML) is coming into its own, with growing recognition.ML can play a key role in a wide range of critical applications, such as datamining,natural language,image recognition and expert system.ML provides potential solution in all these domains and more.ML set to be a pillar of our future civilization.ML are classified into six types Supervised learning, UnSupervised learning, Semi-supervised learning, Reinforcement learning, Evolutionary learning, Deep learning .In supervised learning we are given a data set and already know what our correct output is.Unsupervised learning allows us to approach problems with little or no idea what our results should look like.Reinforcement is when exposed to an environment, how the machine trained itself using trial and error.ML learns from past experience semi-supervised learning is a class of machine learning tasks and technique. Evaluation is the process of determining value, worth, or meaning. Deep learning is a subset of machine learning in Artificial Intelligence (AI) to have network capable of learning unsupervised from information that is unstructured.

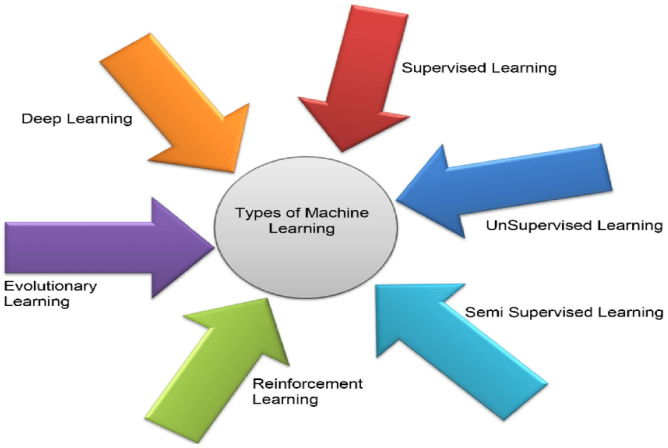


Figure 1.Types Of Machine Learning

II.LITERATURE REVIEW

X.Peng, K.Rafferty, S.Ferguson, "Building support vector machines in the context of regularized least squares" they defining a set of indicator variables of the errors, the solution to the RLS problem is represented as an equation that describe the error vector to the indicator variables.

Amancio D.R,Comin C.H,Casanova .D, Travieso .G,Bruno O.M,Rodrigues F.A, Da Fontoura Costa.L, " A systemetic comparision of supervised classifier" they projected a different model making use of interval-valued fuzzy

membership values for compose decision trees.Presently the fuzzy decision tree don't take into account the involved related to their membership values.

III.MACHINE LEARNING TECHNIQUES

1.Regression
2.Classification
3.Clustering
4.Anomaly

A.Regression:

Regression is basically a numerical approach to find the relationship between variables. In machine learning, this is used to guess the outcome of an event based on the association between variables obtained from the data-set. Linear regression is one type regression used inMachine Learning.

B.Classification:

In machine learning and statistics, classification is the problem of identify towhich of a set of categories a new observation belongs, on the basis of a preparation set of data containing annotations whose category membership is known.

C.Clustering:

Clustering is the assignment of a set of observations into subsets, so that observations in the same clusterare similar in some sense. Clustering is a method of unsupervised learning, and a common technique for numerical data study used in many fields.

D.Anomaly:

In Data Science, Anomalies are referred to as data points which do not conform to an expected pattern of the other items in the data set. The interestingness or real existence significance of anomalies is a key characteristic of anomaly discovery.

IV. ADVANTAGES OF MACHINE LEARNING IN ARTIFICIAL INTELLIGENCE

*****More powerful and more useful computer
*Solving new problems
*Better handling of information olving problems
*Conversion of information into knowledge

V. DISADVANTAGES OF MACHINE LEARNING IN ARTIFICIAL INTELLIGENCE

*Slow and expensive
*Increased cost
*Difficulty with software development

VI.CLUSTERING METHODS IN MACHINE LEARNING

A.CLUSTERING

It is essentially type of unsupervised machine learning . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled response. It is used as a process to find significant structure, descriptive original process, generative features, and grouping natural in a set of examples. Clustering is the task of separating the population or data points into a number of groups such that data points in the related groups are more similar to other data points in thesame group and different to the data points in other groups. It is basically a group of objects onthe basis of parallel and difference between them.

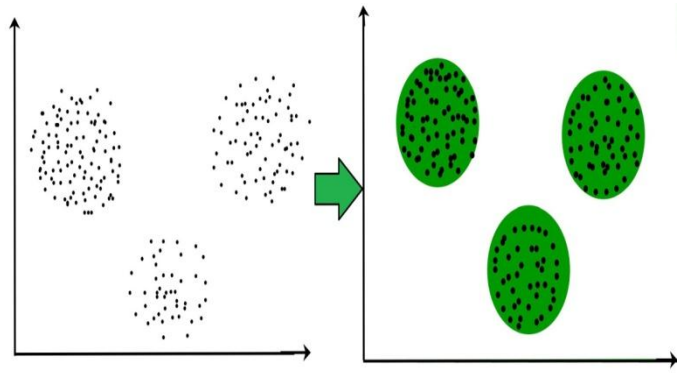


Figure 2. Merging Of Three Clusters

VII. TAXONOMYS OF CLUSTERING APPROACHES

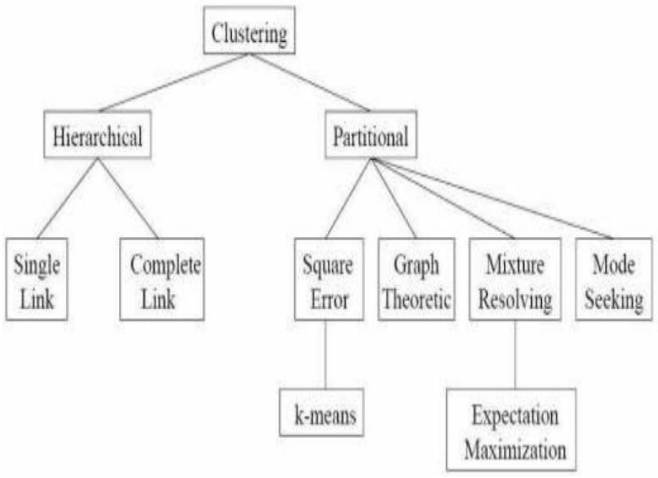


Figure 3.A Taxonomy Of Clustering Approaches

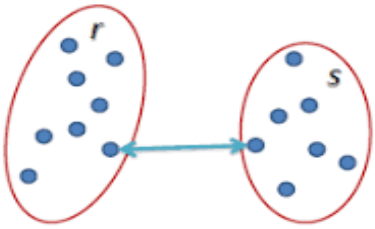
1. HIERARCHICAL CLUSTERING

Hierarchical clustering Technique is one of the popular Clustering techniques in Machine Learning. Clustering is basically a method which groups the parallel

data points such that the points in the same group are more parallel to each other than the points in the other groups.

A. SINGLELINK HIERARCHICAL

This type of clustering is often called as the connectedness, the minimum method or the nearest neighbour method. In single-linkage clustering, the link between two clusters is made by a single part pair, namely those two elements (one in each cluster) that are nearby to each other. In this clustering, the distance between two clusters is determined by nearest distance from any member of one cluster to any member of the other cluster, this also defines similarity. If the data is equipped with similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster

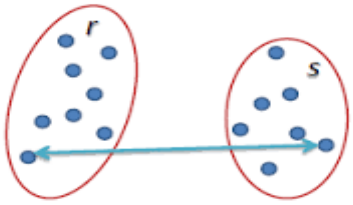


$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Figure 4. Singlelink Clustering

B. *COMPLETELINK HIERARCHICAL*

In complete-linkage clustering also called the diameter, the maximum method or the furthest neighbour method; the distance between two clusters is determined by longest distance from any member of one cluster to any member of the other cluster.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Figure 5.Completelink Clustering

2.PARTITIONAL CLUSTERING

Partitional clustering decomposes a data set into a set of disjoint clusters. Given a data set of *N* points, a partitioning method constructs *K* ($N \geq K$) partitions of the data, with each partition representing a cluster.

A.SQURE ERROR PARTITIONAL

To obtain a partition which, for a fixed number of clusters, minimizes the square-error where square-error is the sum of the Euclidean distances between each pattern and its cluster center.

- ***K-MEANS PARTITIONAL***

k-means clustering is a method of vector quantization originally fromsignal processing that is popular for cluster analysis in data mining. *k*-means clustering aims to separation *n* observations into *k* clusters in which each observation belongs to the cluster with the near mean, helping as a prototype of the cluster. This results in a partition of the data space into Voronoi cells.

B.GRAPH THEORETIC PARTITIONAL

A graph G is a set of vertex (nodes) v connected by edges (links).It is also called community, it refers to a group of nodes having a relations with each other than with the rest of the system. A large range of methods are used to make known clusters in a network.

C.MIXTURE RESOLVING PARTITIONAL

Model-based clustering consists of accurate a combination model to data and identify each cluster with one of its mechanism. Multivariate normal distributions are typically used. The number of clusters is usually determined from the data.

- ***EXPACTATION MAXIMIZATION PARTITIONAL***

The EM algorithm find highest probability estimates of parameters in probabilistic models. EM is an iterative method which alternates between two steps, expectation (*E*) and maximization (*M*). For clustering, EM makes use of the finite Gaussian mixtures model and estimates a set ofparameters iteratively until a desired meeting value is achieved. The mixture is defined as a set of *K*probability distributions and each distribution corresponds to one cluster. An instance is assigned with a relationship possibility for each cluster.

D.MODE SEEKING PARTITIONAL

Mean shift is a non parametric feature space study technique for locating the maxima of a density function , it is also called mode seeking algorithm. Application domain include cluster analysis in computer.

VIII.CLASSIFICATION METHODS IN MACHINE LEARNING

A.CLASSIFICATION

classification is the problem of identifying to which of a set of categories.A new study belongs, on the basis of a training

set of data containing explanation whose group membership is known.

classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses thislearning to classify new observation.

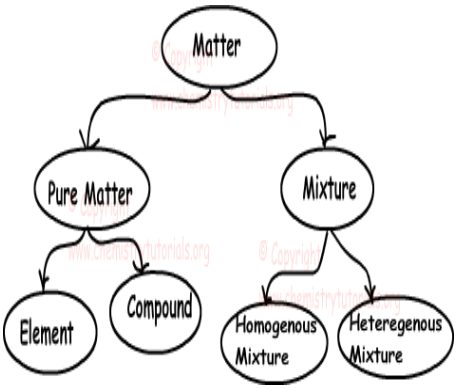


Figure 6.Classification Of Matters

IX.TYPES OF CLASSIFICATION ALGORITHMS IN MACHINE LEARNING

1. Random Forest
2. Nearest Neighbor
3. Support Vector Machines
4. Boosted Trees

A.Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at preparation time and outputting the class that is the mode of the classes (classification) or mean calculation (regression) of the character trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

In the training phase to adapt a neural network to the particular problem at hand.

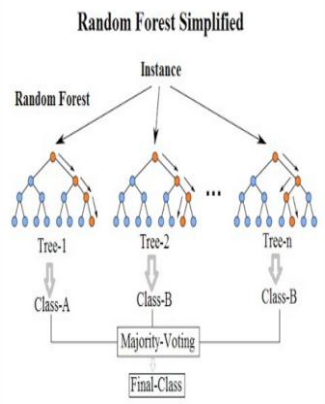


Figure1.1 Random Forest Simplified

B.Nearest Neighbor:

The k-nearest-neighbors algorithm is a sorting algorithm, and it is supervised: it takes a group of labelled points and uses them to learn how to label other points. To label a new point, it looks at the labelled points neighboring to that new point (those are its nearest neighbors), and has those neighbors choose, so either label the most of the neighbors have is the label for the new point (the "k" is the number of neighbors it checks).

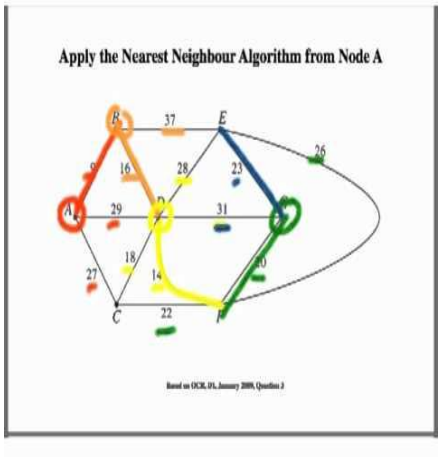


Figure 1.2 Nearest Neighbour Algorithm For Nodes

C.Support Vector Machines

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be working for both classification and regression purposes. SVMs are more frequently used in classification problems and as such, this is what we will focus on in this post.

SVMs are based on the idea of finding a hyperplane that best divides a dataset into two modules.

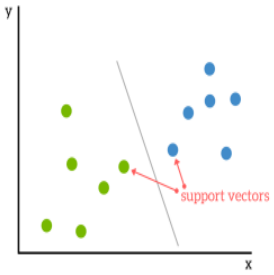


Figure 1.3 Support Vector Machine

Support vectors are the data points nearby to the hyperplane, the points of a data set that, if indifferent, would alter the position of the dividing hyperplane. Because of this, they can be measured the critical elements of a data set.

D.Boosted Tree

Boosting is a method of converting weak learners into strong learners.The gradient boosting algorithm (gbm) can be

most easily explained by first introducing the AdaBoost Algorithm.The AdaBoost Algorithm begins by preparation a decision tree in which each study is assigned an equal weight.

X. CONCLUSION

These days, machine learning techniques are being generally used to solve realworld problems by storing, manipulating, extracting and retrieving data from large sources. Supervised machine learning techniques have been generally 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 labeled data sets. Thus active learning provides a way to reduce the labeling costs by labeling only the most useful instances for learning. Unsupervised learning is a powerful tool that can make sense out of conceptual data set using model recognition. With enough preparation these algorithms can predict insights, decisions, and results across a huge number of data sets allowing automation of many business tasks.

REFERENCE

[1] M.Praveena,V.Jaiganesh "A Literature review on Supervised machine learning algorithms and boosting process", IJCA(0975-8887), Vol.169, No.8, July 2017.

[2] UmairAhmed,T.manorajithm "Literature survey on comparison of supervised learning clasification algorithms", IJPAM, Vol.118,No.20 ,815-823, 2018.

[3] Pedro Domingos, "A Few useful things to know about machine learning", WA98195-2350,USA.

[4]S.KumarJasra,J.Gauci,AMuscat,G.Valentino, DZammit-Mangion,R.Camilleri "Literature review of machine learning techiques to analyse flight data", AEGAT, No.19, October2018.

[5] SumitDas, AritraDey, AkashPal, NabamitaRoy "Applications of artificial intelligence in machine learning:Review and prospect", International Journal of Computer Applications(0975-8887), Vol.115, No.9, April 2015.

