International Journal of Science, Technology and Management (IJSTM) ISSN (online): 2321-774X

Volume 9, Issue 1, 2022

Review on Different Models of Classification in Data Mining

Abhishek

abhishekmahla22@gmail.com

M.Tech Scholar, Department of CSE, BRCM CET, Bahal, (Haryana), India

Dr. Dinesh Kumar

dinesh.muwal@gmail.com

Professor, Department of CSE, BRCM CET Bahal, (Haryana), India

ABSTRACT

A classification is a technique for estimating comparable information based on the value of a classified target or classified class variable. Any kind of statistical data can benefit from this strategy. Such techniques are used for a variety of tasks, including image processing, predictive analysis, data mining, etc. Building a straightforward and unambiguous model of how class labels are distributed in relation to predictor features is the primary goal of supervised learning. The classifiers are then used to compare the value of the unknowable class label to the class labels of the testing cases where the values of the predictor characteristics are known. In this article, we present examples of several categorization methods used in supervised machine learning.

Keywords: Classification, supervised, machine learning, pattern recognition.

INTRODUCTION

A machine learning strategy is a method for teaching computers to handle data in a better way and give precise results. Sometimes, even after seeing the data, we are unable to recognize patterns or draw conclusions from the data. In this scenario, machine learning techniques are used to predict the data [1]. Machine learning is needed since there are numerous datasets accessible from various sources. Machine learning is being used by a wide range of industries, including the military and healthcare, to extract pertinent data from the datasets that are accessible. Learning from the data already available is the fundamental goal of machine learning. Numerous techniques have been developed to enable computers to learn on their own [2] [3]. Numerous mathematicians and programmers use a variety of techniques to solve this problem. Figure. 1 shows some of them in action. every guided learning session Section 2 provides an explanation of machine learning classification strategies. This essay is concluded in Section 3.

TYPES OF LEARNING

An artificial intelligence (AI) system uses lessons from the past to enhance the functionality of intelligent application programmers. Two types of machine learning systems exist.

- Supervised Learning
- Unsupervised Learning

The learning model that "learns" how and where to predict from training data of a particular example is created through supervised learning. Without any prior understanding of the data, unsupervised learning creates a model from "unlabeled" data in order to define and estimate its important aspects.

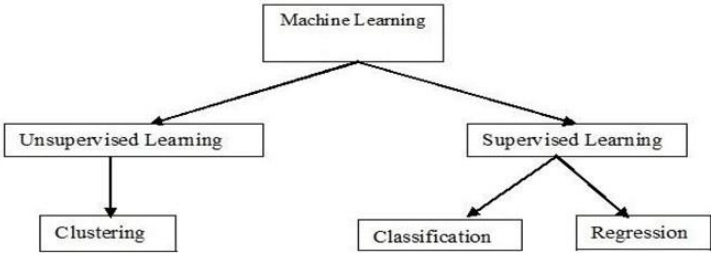


Figure-1: Types of Machine Learning

Supervised Learning

In this article, we discuss several supervised learning classification algorithms. Under supervised learning, the entire dataset is split into two sections: one for training, where the classifier learns from the data, and the other for assessing the classifier's accuracy. When it is finished, we can use these supervised learning classifiers to evaluate fresh data and predict future information. Classification techniques for supervised learning classifiers are divided into five primary categories based on frequency tables, covariance matrices, similarity indices, vectors and margins, and neural networks. We have several classification methods from this area of classification.

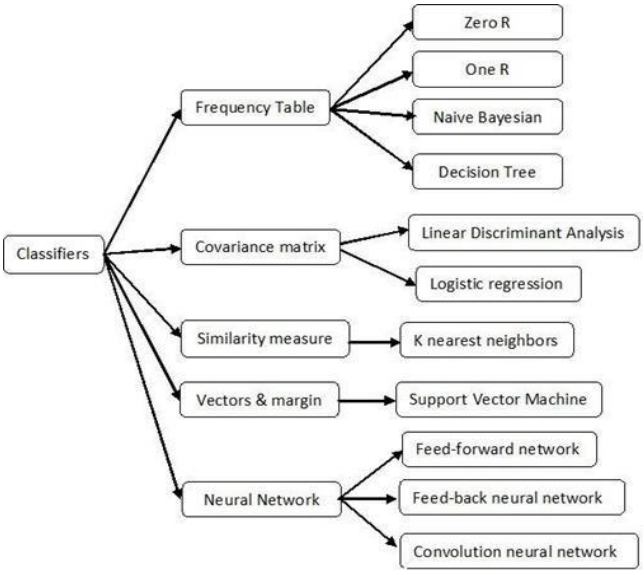


Figure 2: Different classification algorithms in Supervised Learning

ZeroR

The most basic classification technique, ZeroR, just considers the target data and ignores all other predictors. The bulk of category labels are simply predicted by the ZeroR classifier. ZeroR does not have predictive power, but it may be used to provide a baseline performance for other categorization approaches [4].

Algorithm

Create a frequency table for the goal and pick the value that appears most often.

OneR

OneR, or One Rule, is a straightforward classification technique that produces one rule for each predictor in the data but isn't very accurate. Using the OneR method, it only chooses the best predictor with the minimum overall error from the frequency table to forecast the target. Additionally, it was somewhat less accurate than cutting-edge categorization systems [4].

Algorithm

For each predictor,

For each value of that predictor, make a rule as follows; Count how often each value of target (class) appears Find the most frequent table

Make the rule assign that class to this value of the predictor Calculate the total error of the rules of each predictor Choose the predictor with the smallest total error.

Naive Bayesian

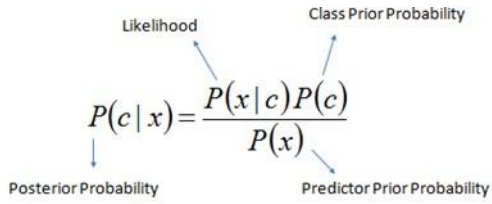
The Naive Bayesian classifier used by the Bayes theorem makes predictions about the predictors' independence. A Naive Bayesian model is simple to construct and does not require time-consuming iterative parameter estimation, making it especially beneficial for very large datasets. Despite being straightforward, the Naive Bayesian classifier frequently beats more complex classification

techniques, frequently performs shockingly well, and is extensively used.

Algorithm

The posterior probability may be calculated using the Bayes theorem., P(c|x), from P(c), P(x), and P(x|c).

The Naive Bayes classifier makes the assumption that the impact of a predictor's value (x) on a particular class (c) is unrelated to the values of other predictors. The term "class conditional independence" refers to this presumption.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Decision Tree

A hierarchical structure is created for classification models via decision trees. The process of incrementally segmenting the dataset into smaller and smaller pieces results in the development of the decision tree. A tree containing decision nodes and leaf nodes is produced throughout the final procedure. There are two or more branches on a decision node. A judgement or categorization is represented by a leaf node. the root node of a tree that represents the ideal prediction from the available datasets.

Both category and numerical data may be classified using a decision tree classifier [6].

Algorithm

1. Using the idea of knowledge gain, the root of the tree is chosen from a dataset attribute.
2. Subset the training dataset, as necessary. And each of these subsets was created so that it contained data with the same value for each attribute.
3. Repeat steps 1 and 2 on each subset until you discover leaf nodes in every branch of the tree.

Entropy

A decision tree is built from the root node up using a top-down technique. Partitioning the data into subsets that contain instances with comparable values are then placed on leaf nodes. The ID3 algorithm, which takes a top-down approach and is the primary algorithm for decision tree generation, uses entropy to determine how similar the samples are. If all of the samples are identical, then the entropy is zero, and if any samples are dissimilar, then the entropy is divided by one.

Using frequency tables, we must identify two categories of entropy in order to create a decision tree.

- Entropy with a probability table that contains just one attribute.
- Entropy with a recurrence table that contains two characteristics.

Information Gain

After a dataset is split up according to an attribute, the information gain is based on the decrease in entropy. The key
to creating a decision tree is identifying the attribute that yields the most information gain, or homogenous
data.

Linear Discriminant Analysis

Linear Discriminant Analysis utilises the Covariance Matrix approach (LDA). Accuracy can be generated using more sophisticated procedures, such as mathematical ones that frequently develop models [5].

Algorithm

LDA, which is focused on looking for a that best distinguishes two classes, uses the notion of a linear combination of

variables (predictors) (targets).

Logistic Regression

The chances of an outcome, which can only have Boolean values, is predicted via logistic regression. Both numerical and categorical data are used to make the prediction. For two reasons, a linear regression cannot be used to predict the value of a binary variable.

The values cannot be predicted using a linear regression with an acceptable range. The residuals for dichotomous experiments won't be regularly distributed around the expected line since there is only one of two potential values for each experiment. However, a logistic curve that is restricted to values between 0 and 1 is produced via logistic regressions. Similar to a linear regression, a logistic regression builds its curve using the natural logarithm rather than the probability. Additionally, each group will have identical variances or the predictors will have an odd distribution [5].

Boosting (Meta-algorithm)

Boosting is a machine learning meta-algorithm for reducing bias in supervised learning. Boosting is based on the question posed by Kearns: [1] Can a set of weak learners create a single strong learner? A weak learner is defined to be a classifier which is only slightly correlated with the true classification (it can label examples better than random guessing). In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification.

K Nearest Neighbors

K nearest neighbors is a straightforward method that categorizes new instances based on a similarity metric and saves all of the existing examples (e.g., distance functions). Based on their closest neighbours and an odd number, KNN has been employed in statistical estimation and pattern identification. It makes use of distance metrics like Euclidean, Manhattan, Minkowski, and others [6].

Algorithm

A distance function such the Euclidean, Manhattan, Minkowski, etc. is used to calculate the K nearest neighbours. K is often seen as an odd number exclusively when making decisions.

If K = 1, then simply assigned to the class of its nearest neighbor. Or

If K is odd number, then assigned to the class of maximum votes of its nearest neighbor.

Support Vector Machine

A Support Vector Machine (SVM) categorizes data by locating the hyperplane that is maximized and utilized as the border in between 2 classes [6].

Algorithm

1. Create many hyperplanes, then choose the appropriate one.
2. Increase the margin between the classes while optimizing the hyperplane.

Feed-forward neural network

A network that goes in one direction and is non-repetitive is called a feed-forward network. It has hidden layers, input layers, and output layers. Input layer elements are supplied to computations

to process data. Each input will have some weight, and there will be links between the input, hidden, and output layers. These weights go through a processing procedure and compute using the input weights. Before reaching the output, it will be computed and forwarded together with input to additional hidden layers and counties. The output of a neuron in the output layer is measured using a threshold function [7][8].

Feed-back neural network

Through the use of repeating loops, a feed-back network may propagate feed-back pathways backwards in both directions. Neurons can communicate with each other in any way. This kind of network has repeated elements, which transforms it into a non-linear dynamic system that changes continually until it achieves an equilibrium state. The expected and actual outputs of the feed-back networks of the neural network are compared. When solving optimization issues, the parameters are modified based on the mistake and then sent back into the neural network to create the optimal configuration of linked neurons [7][8].

Convolutional neural network

Regular neural networks and convolutional neural networks are extremely similar. This network's neurons have biases and weights that can be learned. Each neuron processes a few inputs, conducts a dot product, and may optionally do a non-linearity as a follow-up. The act of receiving an input image and producing an output class (a cat, dog, etc.) or a probability of classes that best fit the image is known as image classification [9].

Name of the Classification	Advantage	Disadvantage
Decision Tree	selection of features or variable screening. Making data is simple. Identifies all feasible possibilities and tracks each possibility.	They are prone to sampling mistakes because they are excessively fitted.
ZeroR	standard for additional categorization techniques provided	Exclusively on the target data
K Nearest Neighbors	noisy training data robust. No period of training can readily handle complicated models	Difficult to involve situations with bigger dimensions. the kind of distance measurement to choose. High price
Naive Bayesian	Simple to implement reduced training data categorization issues using binary and many classes.	extreme assumption data shortage ongoing characteristics.

OneR	Contemporary classification	It chose one of the better predictors from a frequency table, although it wasn't very accurate.
Linear Discriminant Analysis	One of the top face recognition algorithms speedy and portable. Useful for starting a project.	Antiquated algorithm Other algorithms are far more accurate predictors than this.
Logistic Regression	Prediction of nonlinear effects is handled using both numerical and categorical methods.	Just Boolean values. Unsuitable for predicting a binary variable's value
Feed-forward neural network	Complex functions can be solved quite quickly. In order to simulate nonlinear relationships simple to manage	Not applicable to accessible smaller data. Not appropriate for mathematical computations or exact calculations
Support Vector Machine	It applicable to greater dimensions. different decision functions require different kernel functions.	larger features than samples. Probabilities are not guessed at directly.
Meta-algorithm	Often provides predictive accuracy that cannot be trumped. Lots of flexibility	It is sensitive to outliers
Convolutional neural network	Error is far lower than in the past. Identification of object classification issue.	More veiled facets Space and time
Feed-back neural network	Reverse propagation Both directions of travel. Dynamic system that is always changing.	Sluggish and ineffective May become stranded at a nearby minimum.

Table-1 Advantages and Disadvantages of different classifications

CONCLUSION

Several classification methods that are employed in machine learning algorithms are presented in this paper. In category or numerical datasets, a classification is a technique for predicting comparable information. For categorization challenges, machine learning methods are becoming more and more common. In this article, the majority of the well-known machine learning methods for pattern recognition and classification are introduced.

REFERENCES

[1] Dalibor Bužić, Jasminka Dobša ―Lyrics Classification using Naive Bayes―, in MIPRO 2019, 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, At Opatija, Croatia, DOI: 10.23919/MIPRO, 2019

[2] Zhongheng Zhang ―Naive Bayes classification in R‖ Ann Trans Med; 4(12):241Annals of Translational Medicine,Vol 4, 12 June 2018, DOI: 10.21037/atm.2018.0

[3] LiangZhao, Deng-Feng Chen, Sheng-Jun Xu and Jun Lu, "The Research of Data Mining Classification Algorithm that Based on SJEP", International Journal of Database Theory and Application, Vol.8, No.2, pp.223-234, 2019.

[4] Anita Ganpati, "A Performance Comparison Of End, Bagging and Dagging Meta Classification Algorithms", Proceedings of Academics World 24th International Conference, 2018

[5] Ali Selamat Faculty and Nurulhuda Zainuddin "Sentiment Analysis Using Support Vector Machine" 2019 IEEE 2019 International Conference on Computer, Communication, and Control Technology (I4CT 2019), September 2 - 4, 2019 - Langkawi, Kedah, Malaysia978.

[6] W. Richert, L. P. Coelho, "Building Machine Learning Systems with Python", Packt Publishing Ltd., ISBN 978-1-78216-140-0

[7] M. Welling, "A First Encounter with Machine Learning"

[8] M. Bowles, "Machine Learning in Python: Essential Techniques for Predictive Analytics", John Wiley & Sons Inc., ISBN: 978-1-118-96174-2.

[9] Chitra Nasa, Suman "Evaluation of Different Classification Techniques for WEB Data " International Journal of Computer Applications (0975 – 8887) Volume 52– No.9, August 2012.

[10] Sandhya N. dhage, Sandhya N. dhage "A review on Machine Learning Techniques" I nternational Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 3

[11] Ayon Dey "Machine Learning Algorithms: A Review" International Journal of Computer Science and Information Technologies, Vol. 7 (3) , 2016, 1174-1179.

[12] S. B. Kotsiantis "Supervised Machine Learning: A Review of Classification Techniques" Informatica 31(2007) 249-268

[13] V. Sharma, S. Rai, A. Dev, "A Comprehensive Study of Artificial Neural Networks", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN 2277128X,Volume 2, Issue 10, October 2012.

[14] Rajesh Wankhede and Vikrant Chole (2016), Intrusion Detection System using Classification technique, International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, pp. 25-28.

[15] H.F. Eid, A. Darwish A. H. Ella and A. Abraham, ―Principle components analysis and Support Vector Machine based Intrusion Detection System,‖ 2010, 10th International Conference on Intelligent Systems Design and Applications (ISDA), 2010.

[16] Mrutyunjaya Panda and Manas Ranjan Patra, ―A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection‖, First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008.