Data mining classification algorithms: An overview

Saeed Ngmaldin Bardab [1, *], Tarig Mohamed Ahmed [2, 3], Tarig Abdalkarim Abdalfadil Mohammed [1]

[1]Department of Computer Sciences, ALNeelain University, Khartoum, Sudan
[2]Department of IT, King Abdul-Aziz University, Jeddah, Saudi Arabia
[3]Department of Computer Sciences, University of Khartoum, Khartoum, Sudan

A B S T R A C T

Data mining is also defined as the process of analyzing a quantity of data (usually a large amount) to find a logical relationship that summarizes the data in a new way that is understandable and useful to the owner of the data. This paper examines the various types of classification algorithms in Data Mining, their applications and categorically states the strengths and limitations of each type. The weaknesses found in each algorithm demonstrate how tasks cannot be performed well when only one type of algorithm is applied. For this reason, it is the view of the writer that further research needs to be carried out to explore the potential of combining several of these algorithms to solve machine learning problems.

1. Introduction

Over the years, the use of technology has evolved from simple ideas to concrete realities. This has opened exciting vistas of opportunities for its application in varied fields of disciplines. The classification algorithm is being used today in important sectors, such as health and business. Awad and Khanna (2015) outlined the extensive application of machine learning to include web searches, stock market predictions, gene sequence analysis, weather forecasting, and drug development. These broad uses of machine learning are indebted to new advances in classification algorithms. Data Mining is a process of searching for knowledge of data without prior assumptions about what this knowledge can be. Data mining is also defined as the process of analyzing a quantity of data (usually a large amount) to find a logical relationship that summarizes the data in a new way that is understandable and useful to the owner of the data.

2. Definition of classification

A technical process that uses algorithms to analyze data from multiple perspectives and extract meaningful patterns that can be used to predict users 'future behavior. The market basket analysis system that Amazon.com uses recommends new products to its customers based on their past purchases and is a widely known example of how data mining can be used in marketing.

Han et al. (2011) defined classification as a method of identifying a form or pattern that illustrates as well as differentiates data types or notions. Such a model, according to them, is developed from the scrutiny of a group of data items with recognized group descriptions. This model is then utilized to forecast the group of items whose group labels we do not know. Neelamegam and Ramaraj (2013) described classification as a machine learning method that is applied to foretell the clusters that data instances belong.

3. Types of classification algorithms in machine learning

According to Samuel (2000), machine learning is the study of how computers are made to learn automatically. Over time, different types of machine learning processes have been developed. There are types of algorithms as follows:

1. Linear Regression
2. Logistic Regression
3. Decision Trees
4. SVM
5. Naive Bayes
6. KNN
7. K-Means
8. Random Forest

9. Dimensionality Reduction Algorithms
10. Gradient Boosting Algorithms

Machine learning systems can be grouped into different categories according to the expected results of the algorithm. In fact, classification involves two steps process. The first step is supervised learning for the sake of the predefined class label for the training data set. The second step involves classification accuracy evaluation (Deng et al**.,** 2010).

3.1. Supervised learning

Under supervised learning, the algorithm works with a set of examples whose labels are known. It is mostly applied in situations in which stored data forecasts possible impending events. Supervised learning applications fall into two key groups: Classification and regression. Deng et al. (2010) explained that the labels could be nominal values in the instance of the classification task or numerical values in the case of the regression task. In other words, when a computer is given inputs and anticipated outputs, the computer must make sense of the relationship, if any, or find out the general rule between the two. Nevala (2017) contended that some practical applications of supervised learning are detection of fraud, evaluation of risks, customer segmentation, image, speech, and text recognition.

3.2. Decision tree

A decision tree is commonly used for inductive inference. It involves approximating discrete-valued functions for noisy data and capable of learning inconsistent phrases. Kesavaraj and Sukumaran (2013) contended that it is a model that estimates the value of the expected variable due to multiple input variables.

Patil et al. (2010) mentioned that the decision tree occurs in two main stages: Tree growth, at a training set, and tree pruning, which occurs when the size of the tree is decreased for easy understanding. Some types of decision trees are ID3, C45, and ASSISTANT. The decision tree works through the classification of instances by arranging them from the root to some leaf node. Each node in the tree indicates a test of some characteristic of the instance, and every branch that inclines from the node matches a probable equivalent of the feature. Classification starts by starting a root node of the tree, then checking the characteristics for the specified feature before finally declining from the branch in line with a given attribute. This process goes on repeatedly for the subtree rooted at the new node.

ID3 is also known as Iterative Dichotomiser 3, and it is a common algorithm used in machine learning. One of its advantages is that it is easy to understand. This is admitted by Quinlan (1986), who posited that its popularity is due to its efficiency and simple use. On the contrary, it cannot be used to handle missing values and does not have a backtracking search and global optimization.

Similarly, C4.5, as an extended version of the ID3 algorithm, reduces the weaknesses found in ID3. Bhukya and Ramachandram (2010) agreed that unwanted branches are replaced with leaf nodes in the pruning stage by C4.5 and returned to the tree after generation. Adhatrao et al. (2013) felt that C4.5 has the advantage of handling the problem of missing feature values, providing both pre and post pruning capacity and handling of distinct and constant features. It also has a great processing time when compared to the other decision trees.

Generally, a decision tree can also be used in classification problems to categorize diseases, equipment malfunction, and also in loan applications. Quinlan (1993) recognized thatID3 or decision tree algorithm could be used to solve some problems in which the final results are changed to C4.5.

Some problems with decision trees are: deciding how intensely the decision tree is developed, selecting the best attribute, handling missing values, and managing differing costs. Another flaw is its unsuitability for small data set 18.

3.3. Bayesian networks

Heckerman et al. (1995) claimed that over the past years, the Bayesian network had become a common way of programming indefinite adept information in expert systems. A Bayesian Network is a visual representation of a model used for probability connections among a cluster of variable characteristics. Phyu (2009) reported that it also has graphical models for reasoning under uncertainty where the nodes represent variables (discrete or continuous), and arcs represent direct connections between them. Kotsiantis et al. (2007) established that the form of Bayesian Network is made of a controlled acyclic graph called DAG. The nodes are in parallel communication with the X items. The curved shapes demonstrate an unpredicted connection to the nodes, but the shortage of probable arcs in S codes show provisional autonomies. There are two subgroups: (1) Parameters determination and (2) Network DAG structure learning. Yang and Webb (2009) pointed out that one challenge faced by the Bayesian networks classifier is the fact that it needs discretization of its non-stop features. This is a procedure which comprises changing constant attributes to a distinct attribute or feature to create issues in a grouping. According to Friedman and Goldszmidt (1996), one other problem is when a constant attribute or feature is not changed to a distinct attribute. They explain that the valuation of the attribute's conditional density plays an active role in the process. The problem with conditional density estimation of attributes, according to Friedman and Goldszmidt (1996), Gaussian kernel function has steady constraints for the evaluation of attributes density. In their investigation, a constant attribute or feature gives enhanced classification

accuracy in relation to other methods by Gaussian kernel function in the Bayesian Network classifiers.

3.4. K-nearest neighbor (KNN)

This is a familiar classification technique that is used when there is little or no previous information about the dissemination of data. It is a potent non-parametric method that escapes the problem of probability densities. The K-Nearest neighbor rule categorized x by giving it the tag appearing regularly and is denoted among the K nearest samples. The technique was made to help in executing discriminant analysis in cases where dependable parametric estimates probability densities are difficult to define. Fig. 1 shows **the** KNN model.
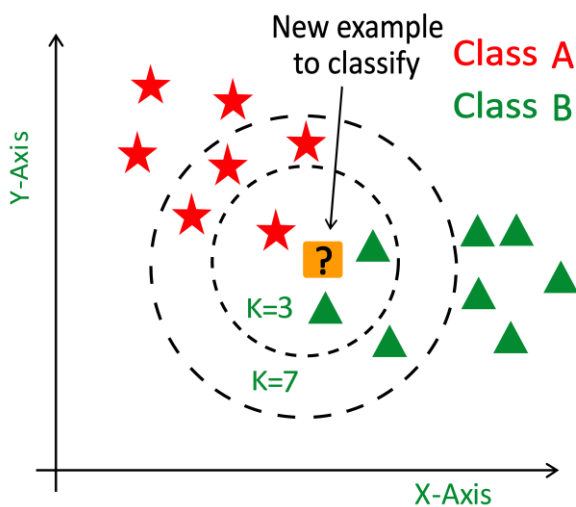

Fig. 1: KNN model

In the diagram, we need to determine which class the new example should be grouped with. K=3 means the new example will be identified as a Class B member. However, if K=7, the new example will belong to Class A (Fig. 1).

Covert and Hart (1967) contended that **t**he technique is also measured as K. This means the total nearest neighbors need observation in order to label the sample data point group. Wu et al. (2008) divided KNN into two main groups: The structure-based KNN and the structureless KNN. The former procedure occurs with the fundamental form, which has a lower system of working in the training of data samples. However, there is an organization of data into sample data point and training data resulting in the calculation of the sample points.

One of the advantages of k-NNMC is its high accuracy and efficiency. Similarly, it is transparent, simple to use, and has the ability to withstand noisy data training and adaptation.

3.5. Challenges of KNN

Viswanath and Sarma (2011) identified two main problems with the use of KNN: Space requirement and the classification time used. To resolve these difficulties, different procedures such as the K-Nearest Neighbor Mean Classifier can be used. This is done by searching for k nearest neighbors for each training form of class and computing the mean for all k-neighbors. This is shown through the application of different average data-sets. Bhatia (2010) also argued that computing difficulties, memory inadequacy, low run-time presentation for a huge training set of data, and interference by unrelated attributes are some drawbacks of KNN.

3.6. Linear classifiers (support vector machines)

These are similar to classical multilayer perception neural networks. In the Support Vector Machine, a predictor variable is called an attribute, and a transformed attribute defines the hyperplane, which we call a feature. Feature selection refers to the work involved in choosing the most appropriate representation. A vector refers to the features that describe a row of predictor values. Fig. 2 shows a support vector machine example.
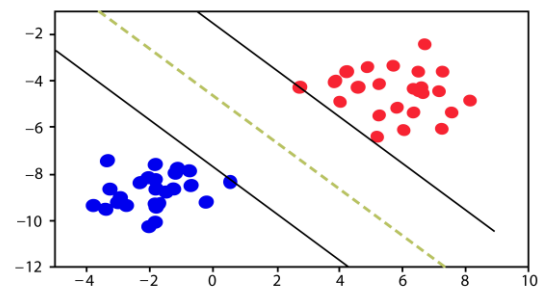

Fig. 2: Support vector machine example

The yellow broken-line is the ideal hyperplane which separates the two classes with the maximal margin. The other two lines only separate the classes with small margins (Fig. 2).

Nizar et al. (2008) suggested that a support vector machine (SVM) is the most popular and most suitable procedure for resolving issues in data classification, learning, and estimation. In addition, Wu et al. (2008) accepted that the origin of SVM is the maximal margin classifier, and it defines the least classification challenges of linear separable training with binary grouping. The main benefits of SVM are solving classification issues such as high dimensional and non-linear separable problems. However, one setback of these vector machines is that it needs important parameters to establish the expected results.

4. Application of classification algorithms

Currently, classification is used worldwide in different fields, such as data mining, science, industrial, and law. Machine Learning has become very important and useful for technology. Face recognition, Chabot for online customer service, radiologists looking for early signs of cancer, and self-driving cars are few examples of machine learning usage. According to Smola and Vishwanathan (2008), machine learning helps in the

classification and verification of faces. Many security applications, such as access control and face recognition, are some other examples of its application.

Nevala (2017) clarified that Machine learning could also be used in situations where:

1. Relevant rules are not simply labeled by straightforward logical rules.
2. Possible outputs are not determined earlier than the event.
3. Precision is more essential than interpretability.
4. The data is not sound and has the potential to cause problems for long-established analytic methods.

5. Problems of classification in machine learning

Classification is beneficial in many fields, but it also has many shortcomings: Handling missing data, data identification irrelevant at the time of entry, data separation due to difference with other known data, and equipment breakdown and misconception due to lack of entry of a record, which can be solved by swapping whole omitted values with a known global constant. Another way to solve this problem is by the removal of values with the feature mean for a particular class. Deng et al. (2010) posited that data miners could overlook the omitted data by observing examples that have omitted values and manually adding a probable value.

Some establishments using machine learning face problems such as biases in the data, the taintlessness of the data, loss of control, and the low quality of work. In the loss of control establishments, machine learning is not capable of controlling the pace of the results produced. For business, careful evaluation and readiness to use the process are important. Funding the process is also of utmost importance.

Another problem is when performing spam filtering. During the process, we concentrate on yes/no answer as to whether an e-mail contains the most important information or not. To combat these problems, we need to build a system that is able to learn how to classify new e-mails.

Krishna et al. (2017) observed that hackers could cause a change in algorithms, and they can also make a fake one look genuine. They give an example of the introduction of the iPhone X in which hackers with forged biometric data can pretend to be the rightful owners or users.

6. Conclusion

Classification is a method that results in sorting out data into groups based on certain important characteristics or features. This paper has outlined the meaning of classification algorithms, applications, types, weaknesses or challenges, and solutions to tackle the challenges. Classification is very helpful in both data mining and machine learning. The types of classification are also mentioned, and the merits and demerits of each

technique during its applications were also explained. In stressing the strengths and extensively resolving the weaknesses, the application of these methods can be enhanced. It is important to note that the methods chosen are determined by the problems at hand. For this reason, one method is not seen as better than the others but can complement each other.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Adhatrao K, Gaykar A, Dhawan A, Jha R, and Honrao V (2013). Predicting students' performance using ID3 and C4. 5 classification algorithms. https://doi.org/10.5121/ijdkp.2013.3504

Awad M and Khanna R (2015). Efficient learning machines: Theories, concepts, and applications for engineers and system designers. Springer Nature, Berlin, Germany. https://doi.org/10.1007/978-1-4302-5990-9

Bhatia N (2010). Survey of nearest neighbor techniques. International Journal of Computer Science and Information Security, 8(2): 302-305.

Bhukya DP and Ramachandram S (2010). Decision tree induction: An approach for data classification using AVL-tree. International Journal of Computer and Electrical Engineering, 2(4): 660-665. https://doi.org/10.7763/IJCEE.2010.V2.208

Cover T and Hart P (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1): 21-27. https://doi.org/10.1109/TIT.1967.1053964

Deng J, Berg AC, Li K, and Fei-Fei L (2010). What does classifying more than 10,000 image categories tell us? In: Daniilidis K, Maragos P, and Paragios N (Eds.), European conference on computer vision: 71-84. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-15555-0_6

Friedman N and Goldszmidt M (1996). Discretizing continuous attributes while learning Bayesian networks. In the International Conference on Machine Learning: 157-165.

Han J, Kamber M, and Pei J (2011). Data mining concepts and techniques third edition. In: Han J, Kamber M, and Pei J (Eds.), The Morgan Kaufmann series in data management systems: 83-124. Morgan Kaufmann/Elsevier, Burlington, USA. https://doi.org/10.1016/B978-0-12-381479-1.00003-4

Heckerman D, Geiger D, and Chickering DM (1995). Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 20(3): 197-243. https://doi.org/10.1007/BF00994016

Kesavaraj G and Sukumaran S (2013). A study on classification techniques in data mining. In the 4th International Conference on Computing, Communications and Networking Technologies, IEEE, Tiruchengode, India: 1-7. https://doi.org/10.1109/ICCCNT.2013.6726842

Kotsiantis SB, Zaharakis I, and Pintelas P (2007). Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160(1): 3-24.

Krishna D, Albinson N, and Chu Y (2017). Managing algorithmic risks. Deloitte, New York, USA.

Neelamegam S and Ramaraj E (2013). Classification algorithm in data mining: An overview. International Journal of P2P Network Trends and Technology, 4(8): 369-374.

Nevala K (2017). The machine learning primer. SAS Institute, North Carolina, USA.

Nizar AH, Dong ZY, and Wang Y (2008). Power utility nontechnical loss analysis with extreme learning machine method. IEEE Transactions on Power Systems, 23(3): 946-955. https://doi.org/10.1109/TPWRS.2008.926431

Patil DD, Wadhai VM, and Gokhale JA (2010). Evaluation of decision tree pruning algorithms for complexity and classification accuracy. International Journal of Computer Applications, 11(2): 23-30. https://doi.org/10.5120/1554-2074

Phyu TN (2009). Survey of classification techniques in data mining. In the International Multi Conference of Engineers and Computer Scientists, Hong Kong, China, 1: 1-5.

Quinlan JR (1986). Induction of decision trees. Machine Learning, 1(1): 81-106. https://doi.org/10.1007/BF00116251

Quinlan JR (1993). Combining instance-based and model-based learning. In the 10th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., Amherst, USA: 236-243. https://doi.org/10.1016/B978-1-55860-307-3.50037-X

Samuel AL (2000). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 44(1.2): 206-226. https://doi.org/10.1147/rd.441.0206

Smola A and Vishwanathan SVN (2008). Introduction to machine learning. Cambridge University Press, London, UK.

Viswanath P and Sarma TH (2011). An improvement to k-nearest neighbor classifier. In the IEEE Recent Advances in Intelligent Computational Systems, IEEE, Trivandrum, India: 227-231. https://doi.org/10.1109/RAICS.2011.6069307

Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, and Zhou ZH (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1): 1-37. https://doi.org/10.1007/s10115-007-0114-2

Yang Y and Webb GI (2009). Discretization for Naive-Bayes learning: Managing discretization bias and variance. Machine Learning, 74(1): 39-74. https://doi.org/10.1007/s10994-008-5083-5