Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Comparative study between decision tree and knn of data mining classification technique

To cite this article: M Mohanapriya and J Lekha Mrs 2018 *J. Phys.: Conf. Ser.* **1142** 012011

Recent citations

- Z-Sequence: photometric redshift predictions for galaxy clusters with sequential random k-nearest neighbours
  Matthew C Chan and John P Stott

Comparative study between decision tree and knn of data mining classification technique

M Mohanapriya[1], **Mrs J Lekha**[2]
[1]Research Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India. *mohanamanimaran24@gmail.com*
[2]Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India. *saran.lekha@gmail.com*

Abstract. Data mining is used to discover hidden information using some process, techniques, and its algorithm. Data mining is a very beneficial method to analyze critical data. Many Researchers and organizations use data mining to extract useful knowledge regarding their need. Data mining has many techniques. For example, Classification, Clustering, Regression, Association, Summarization, Time- series etc. Each technique has some algorithms like classification has a decision tree, Naïve Bayes, Neural Networks and so on and Clustering has K-means and so on. The comparative study between Decision tree Algorithm and K- Nearest Neighbor Algorithm of Classification techniques is present in this paper.

1. Introduction

Data mining is the best way to find the information which is hidden from the data using some data mining process, techniques and its algorithm. From data, Data mining is extracting, unknown and unstated. There are various different meaning of Data mining are knowledge mining from Databases, Knowledge extraction, Data/Pattern analysis [1]. Classification is used to classify each item in a data set into one of a predefined set of classes or groups [2]. Classification is the chore of identifying a model or function. The function is mined based on the analysis of a set of training data [3]. The model is used to predict the class label of objects for which the class label is unknown [4]. This paper have done a comparative study in classification techniques algorithm of data mining and the comparative study between decision tree and KNN. Classification is the task of detecting a model or function. The function is extracted based on the analysis of a set of training data. In classification, it is essentially interested in modeling the boundaries between classes. There are a large number of different classification techniques, providing different ways to model decision boundaries.

2. Classification

Classification is used to allocate a set of attributes to suitable predefined classes. Generally, classification accepts categorical variables.

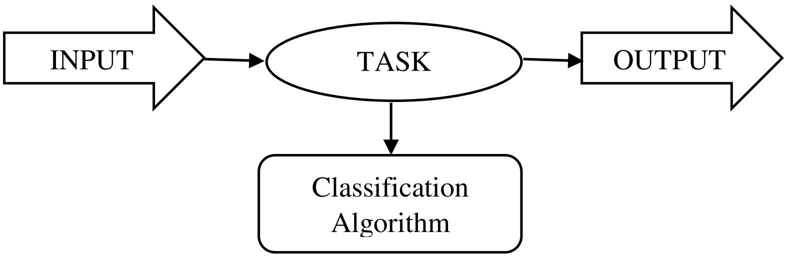

Figure 1. Simplest Representation of the Classification Process

Let's see Figure 1, In Classification, User gives Input to get classified output by performing a specific task as classification algorithm. The classification has Training set, Test set, Evaluation, and Accuracy. The training set is class value for learning. The test set is class value for evaluating. In Evaluation, hypotheses are used to infer classification of the test set and inferred classification is compared to known classification. Finally, Accuracy gives correctly classified test set with a percentage. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects [5]. Using Classifier, here formulate set of attributes of new cases based on the value of other attributes. Classification is one of the most common supervised learning models in the application of data mining. Using some common classification tools, user can get results. In many applications, the classification model is used to predict the effectiveness of the dataset. Model construction and Model Usage are two process of classification. Model construction describes a set of predefined classes and Model usage classifies future and unknown objects. Business is often interested in a simple Yes or No response to a proposal: medical experts want to know whether a patient is healthy or sick likewise. Each one of the classification problem has binary outputs – either success or failure. In this multiple outputs are also common, i.e: output can be yes, no or impartial. The applications which are all popular of classification algorithms include Medical Science, Image or Pattern Recognition, Fraud Detection, Financial sector, and Marketing sector. Classification algorithms output is usually called classifying attribute which means that either it is a discrete or categorical attribute. The goal of classification is to accurately predict the target class for each case in the data [6]. In Classification, an objects are nominated to predefined classes. In classification, each and every technique used to support a learning algorithm. Use learning algorithm to find the best fits relationship between the attribute set and class label of the input data and get a model from that identified set. There are several kinds of classification technique algorithms including Statistical Procedure Based Approach, Machine Learning Based Approach, Neural Network, ID3 Algorithm, Artificial Neural Network algorithm, C4.5Algorithm, K-Nearest Neighbor Algorithm, Naïve Bayes Algorithm, Support Vector Machine Algorithm, Decision Trees, Fuzzy Systems. The goal of this paper is to provide a comparative study between decision tree algorithm and the KNN algorithm.

2.1. Decision Tree

Decision trees just looks like trees, which shows like a hierarchal model. It classifies objects by sorting all objects based on attribute value. Normally trees have node – root node, leaf node, and branch. Each node in the tree represents the object based on attributes. Branches denotes objects value. From the root node the objects are classified. After that based on the attribute value the objects are sorted. A decision tree is a classifier which provides some rules in a tree structure. Here rules denote some sequence of test questions and conditions.

Here take an example Students efforts to win or lose in their career.

Table 1. Dataset of Students efforts.

S.no	Percentage scored in exams	Career chose	Effort	Result
1	50%	Sports	10%	Lose
2	60%	Education	60%	Won
3	70%	Sports	5%	Lose
4	80%	Education	25%	Lose
5	90%	Education	75%	Won

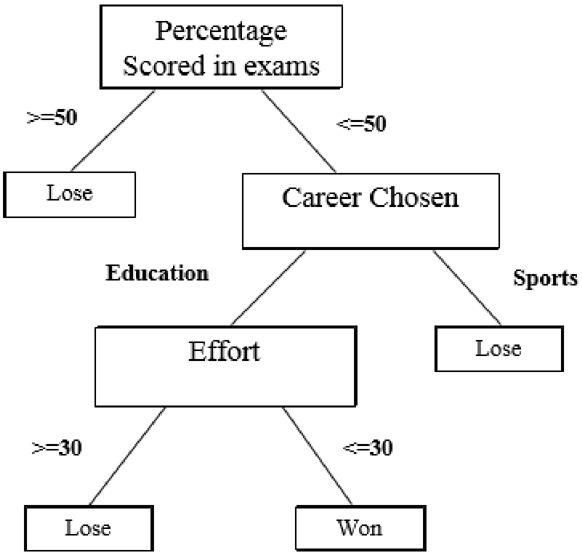

Figure 2. Sample Decision tree

Let's see Figure 2. Root node is Percentage scored in exams, it splitting into two leaf nodes which is the percentage is >=50 means the students may LOSE in their career. If it is <=50 means they have two choice of career [Career Chosen] which is Education and Sports. This node is a decision node. In education, if their effort is >=30 then they will LOSE, if it is <=30 then they will WON. When the node is LOSE then the path is terminated so it is called the terminal node.

After the decision tree is constructed, classifying a test record is uncomplicated. Here, user use some rules to test the record and splitting attributes as a branch and internal node by the output of the rules. Based on the outcomes, it leads us to another internal node or to a leaf node. To apply some new rule, user can go to the internal node. When user reach a leaf node as labeled LOSE then the path is terminated which is called the terminal node and the label which is associated with the leaf node is connected in the record.

A decision tree is constructed using the set of attributes which is given in the dataset. Ofcourse, in decision tree classifier, a very big difficult problem is to raise a best decision tree.The greedy strategy is used to grow a decision tree by using best decisions in which used to find the attribute to partitioning the data. Example- ID3 algorithm, Hunt Algorithm, C4.5 etc. These algorithms are used to find which attribute is best to partition the data to get suitable accuracy.

2.1.1. Hunt's Algorithm. Now everyone are having doubt that why here using Hunt algorithm? Let's see. Decision tree contains many algorithms. To find suitable accuracy in decision tree use Hunt algorithm. Hunt's algorithm grows a decision tree recursively by partitioning a training data set into smaller, purer subsets [6]. Each path taken in the decision tree is must end with the class already chosen. Let's see the process of the Hunts Algorithm. The process of a recursive algorithm is termination. It examines that every record in a node is of the same class. If training set holds records that to be held by the same class, then the node is a leaf node labeled as particular class name. If the training set is an empty set means then the node is a leaf node labeled by the default class name. Now user used to check until no more training record left. If the training set holds records that to be held by more than one class means then use rules (test condition) to split the data into smaller subsets. Then recursively perform in each subset until all the records held by the same class.

The View of the best split is to deciding the attribute rule which should be used. Choose the rule that results in subsets that are purer. So here user need to compare the impurity level of the parent and child node using the formulas that are entropy, Gini index, and miss-classification error.

2.2. KNN

K Nearest Neighbor is an Instance-based algorithm. KNN describes the numeric attributes. It converts the categorical attributes into numerical. Let's take gender as an example, where gender = male and female, now give value to both male and female as male = 0, female = 1. So it carries male and female as 0 and 1. In KNN, training samples are in n-dimensional attributes. If user need to find the value of unknown sample means, in KNN user need to find the closest sample from the record and evaluate the values between all samples from data and unknown sample. KNN assigns equal weight to each attribute. It is also used for prediction. It will predict the real value for an unknown sample.

Let's see with an example, an Unknown sample – If a person name is Rishi, gender is Male and age is 8. Now user have to find his favorite sportsman which may be is Mahendra Singh Dhoni, Sachin Ramesh Tendulkar or neither Virat Kohli. Here, user have to assume that he can't be both Sachin Ramesh Tendulkar and Virat Kholi fan. For this, user have to compute the distance between the unknown record and references to the data records. Predict the class using age and gender of the unknown record.

Table 2. Dataset of Favorite Sports Person

Name	Age	Gender	Favorite Sports Man
Micky	10	Male	Mahendra Singh Dhoni
Angel	15	Female	Virat Kholi
John	17	Male	Sachin Ramesh Tendulkar
Sheela	13	Female	Mahendra Singh Dhoni
Shilpa	18	Female	Sachin Ramesh Tendulkar

Now user want KNN to predict what class of fan Rishi is. Rishi is male and age is 8. Here User sets a value for K, which must be a positive integer. User can set the value of K to 3. From this, user can take 3 closest neighbors of the unknown record. As user already said, some of the data are numeric like age, but others are categorical or discrete like gender. So convert those discrete data to numeric data like Male= 0 and Female= 1. Now everything is in number.

Apply Distance equation, Square root $((x1\text{-}x2)\ \hat{}2 + (y1\text{-}y2)\ \hat{}2)$

Where,

x1 = Age of unknown record,

x2 = Age of Known record,

y1 = Gender of unknown record and

y2 = Gender of known record.

Evaluate the male and age values of Rishi with all reference attributes in the known data record and compare with them.

Let's start

First, take Mickey's age and gender and evaluate them with Rishi's age and Gender, Square root ((*Rishi's age – Mickey's age*) ^2 + (*Rishi's gender – Mickey's gender*) ^2))

Square root ((8 – 10) ^2 + (0 – 0) ^2)

Square root (4) (1)

Second, take Angel's age and gender and evaluate them with Rishi's age and Gender,

Square root ((*Rishi's age –Angel's age*) ^2 + (*Rishi's gender – Angel's gender*) ^2))

Square root ((8 – 15) ^2 + (0 – 1) ^2)

Square root (50) (2)

Third, take John's age and gender and evaluate them with Rishi's age and Gender,

Square root ((*Rishi's age – John's age*) ^2 + (*Rishi's gender – John's gender*) ^2))

Square root ((8 – 17) ^2 + (0 – 0) ^2)

Square root (81) (3)

Fourth, take Sheela's age and gender and evaluate them with Rishi's age and Gender,

Square root ((*Rishi's age – Sheela's age*) ^2 + (*Rishi's gender – Sheela's gender*) ^2))

Square root ((8 – 13) ^2 + (0 – 1) ^2)

Square root (11) (4)

Fifth, take Shilpa's age and gender and evaluate them with Rishi's age and Gender,

Square root ((*Rishi's age – Shilpa's age*) ^2 + (*Rishi's gender – Shilpa's gender*) ^2))

Square root ((8 – 18) ^2 + (0 – 1) ^2)

Square root (21) (5)

Table 3. Dataset of Favorite Sports Person with Calculated Distance

Name	Age	Gender	Distance	Favorite Sports Man
Micky	10	Male	4	Mahendra Singh Dhoni
Angel	15	Female	50	Virat Kohli
John	17	Male	81	Sachin Ramesh Tendulkar
Sheela	13	Female	11	Mahendra Singh Dhoni
Shilpa	18	Female	21	Sachin Ramesh Tendulkar

As already sets K = 3,

K= number of nearest neighbors.

Select 3 closest records compared to Rishi, in which Mickey's distance is 4, Sheela's distance is 11 and Shilpa's distance is 21. These are the closest records when compared to Rishi.

Thereafter, class of the Mahendra Singh Dhoni is the most common, then KNN will predict that Rishi is a Mahendra Singh Dhoni fan.

3. Compare between Decision Tree and Knn Algorithm

In previous topics, this paper discuss Decision tree and KNN and got some ideas about DT and KNN. Now compare DT and KNN.

Table 4. Comparison of DT and KNN

Decision Tree (DT)	K- Nearest Neighbor (KNN)
1. A decision tree is an Eagar Classification.	1. KNN is a Lazy Classification.
2. It is Supervised Learning.	2. It is Unsupervised Learning.
3. DT is used to classify the records using some rules.	3. There is a distance metric in KNN while it decides neighbors.
4. It accepts both numerical and categorical attributes.	4. But in KNN it accepts only numerical attributes if the attributes are categorical while evaluation it will convert them into numerical attributes.
5. DT speed is slower for a large amount of data.	5. KNN speed is faster for all types of data.
6. DT is "White boxes", which means that the received knowledge can be expressed in a readable form.	6. KNN is "Black boxes", which means that the received knowledge cannot be in a readable form.
7. DT is to predict a class for a given attribute.	7. KNN used to find similar value from the record.
8. DT is Deterministic because it is called "compute" and the maximum number of queries happens before a leaf is reached and a result is obtained.	8. KNN is Nondeterministic because it cannot return the same result in all time.
9. DT has an effectiveness on large data.	9. KNN has an effectiveness on small data.
10. DT can compact with noisy data.	10. KNN cannot compact with noisy data.
11. DT needs some algorithm like Hunt algorithm to find best attribute to partition the data.	11. KNN have its own way to get good result.

4. Conclusion

Now it's time to conclude. This paper have discussed the Decision tree and KNN algorithm of classification techniques. From the last topic, here discussed the comparison between DT and KNN. It shows us that the DT algorithm is an easier algorithm when compared to KNN and it is more accurate also. DT is used to partition the data to find accurate result but in KNN it used to find similar values from the data. Each and every algorithm has its own merits and demerits and never ever all algorithms have satisfied all criteria and requirements. Each algorithm has its own specification, so algorithms should be chosen according to the requirements.

5. Reference

[1] Available in the website : http://www.rroij.com/open-access/an-overview-of-knowledge-discovery-databaseand-data-mining-techniques.php?aid=48833

[2] Available in the website : http://www.zentut.com/data-mining/data-mining-techniques/

[3] Available in the website : https://www.tutorialspoint.com/data_mining/dm_quick_guide.htm

[4] Dorina Kabakchieva, November 2012 *Student Performance Prediction by Using Data Mining Classification Algorithms*, International Journal of Computer Science and Management Research, Vol 1 Issue 4, ISSN 2278- 733X.

[5] Krishnaiah V, Dr Narsimha G, Dr. Subhash Chandra N, 2013, *Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques*, International Journal of Computer Science and Information Technologies, Vol. 4 (1), 39 – 45, ISSN: 0975- 9646.

[6] Nagaparameshwara Chary S, Dr Rama B, *Analysis of Classification Technique Algorithms in Data Mining- A Review*, International Journal of Computer Science and Engineering, Volume- 4, Issue- 6, E-ISSN: 2347- 2693.

[7] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang, 21 Dec 2005, *A Framework for Authorship Identification of Online Messages: Writing- Style Features and Classification Techniques*, Journal of the American Society for Information Science and Technology, 57(3): 378- 393, 2006.

[8] Shelly Gupta, Dharminder kumar, Anand Sharma, Apr-May 2011, *Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis*, Indian Journal of Computer Science and Engineering, Vol. 2 No.2, ISSN: 0976- 5166.

[9] Ngai E W T, LiXiu, Chan D C K, 2008 , *Application of Data Mining Techniques in Customer Relationship Management: A Literature Review & Classification*, Elsevier.

[10] Ngai E W T, Yong Hu, Wong Y H, Yijun Chen, Xin Sun, 2011, *The Application of Data Mining Techniques in Financial Fraud Detection: A Classification framework and an academic review of literature*, Elsevier.

[11] Sayali D. Jadhav, Channe H P, *Comparative Study of KNN, Naïve Bayes and Decision Tree Classification Techniques*, International Journal of Science and Research, ISSN (Online): 2319- 7064.

[12] Chaitrali S. Dangare, Sulabha S. Apte, June 2012, *Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques*, International Journal of Computer Applications, Volume 47- No. 10, (0975- 888).

[13] Umesh Kumar Pandey S. Pal, 2011, *Data Mining: A Prediction of Performer or UnderPerformer Using Classification*, International Journal of Computer Science and Information Technologies, Vol. 2(2), 686-690, ISSN: 0975- 9646.

[14] Kesavaraj G, Dr Sukumaran S, July 4-6, 2013 *A Study on Classification Techniques in Data Mining*, IEEE- 31661, 4th ICCCNT – 2013.

[15] BhaveshPatankar and Dr. Vijay Chavda, December 2014, *A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining*, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 12.

[16] Thair Nu Phyu, Mar 18- 20, 2009, *Survey of Classification Techniques in Data Mining*, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, ISBN: 978- 988- 17012.

[17] NIHA S A R, Nov– 2017, *Study of Data Mining Methods and Its Applications*, International Research Journal of Engineering and Technology, Volume: 04 Issue: 11, e-ISSN: 2395- 0056, p-ISSN: 2395- 0072.

[18] Han J and Kamber M, Data Mining Concepts and Techniques, Elsevier.