IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

K -NEAREST NEIGHBORS MACHINE LEARNING ALGORITHM

SIDDHARTH NANDAKUMAR CHIKALKAR

Bachelor of Computer Applications, Vivekananda collage, Kolhapur, Maharashtra, India.

***Abstract**: KNN algorithm is the simplest supervised machine learning algorithm it is mostly used to classification. It the method where KNN algorithm save the data and classifies new cases on a similarity measure according to that data. K value in KNN is most impactable and the high influence on the performance of the KNN algorithm. Other terms of data mining are knowledge extraction, knowledge mining from data base, data or pattern analysis, classification of data. Classification is one the data mining technique to predict the unknown data class with their training set with known class labels data sets. KNN is used to classify future data and it's easy to implementation and it has higher accuracy.*

***Keyword:** machine learning, supervised learning algorithm, data mining, classification algorithm*

Introduction : The KNN algorithm means K-NEREST NEIGHBOURS. This algorithm often used in classification when we have some classified data and we have new data item, but we not sure which is the class of that new data, then we use KNN machine learning algorithm. KNN is supervised and pattern classification learning algorithm. KNN can be used in classification as well as regression. The KNN algorithm is the most accurate model because it makes highly accurate prediction, so we can use the KNN algorithm where we want highly accuracy. This algorithm has some drawback which is the outcomes accuracy is depend on the quality of the available data. So, if we have good quality of data then outcomes accuracy is higher else, we won't get the higher accuracy. The KNN algorithm is easy to implement there are two parameters is required to implement. first is the value of the K, and second one is the distance function. KNN is one of the most use data mining and classification algorithms. And KNN is used in cancer diagnosis, pattern recognition, text classification, email spam detection, fraud detection, and in regression it used to risk assessment, score prediction etc. in this paper we will see how KNN machine learning algorithm actually work and its challenges as well as advantages. KNN algorithm store all the available cases and classify new data of K in similarity measure. It suggests that if you are similar to your neighbour then you are one of them. For example, apple is more similar to orange, banana, and mango rather than dog , monkey, lion then most likely apple is belonging to group of fruits. KNN used is search application when you want to similar items then you called the search as an a KNN search. What is the actually k in KNN ? K is the number of neighbours near to the new object which we have to assign. If k=3 then the most common three nearest neighbours are checked and the most common neighbours' class are assigning to the testing data item. So, this is the K in KNN algorithm. The biggest use of KNN algorithm is the recommendation system. The recommended system is like the shop counter when you asking for a product it not shows only that product, it displayed you and also suggest your relevant sets of products and related to the item you are already too interested to buying it. The KNN algorithm is used in recommended product like in amazon and recommended media in case of NETFLIX. More them 35% of amazon revenue is generated by its recommendation engine. For this kind of purpose, we use KNN algorithm and more advanced example may like handwriting detection, image recognition or even video recognition, and it is used to get missing value, used in pattern recognition, and it used in gene expression this is also the example of KNN machine learning algorithm

HOW DOSE KNN MACHINE LEARNING ALGORITHM WORK

The k nearest neighbour algorithm find the nearest neighbour of new data item, if k= 3,(k is the nearest neighbour) then 3 closest neighbours has been checked and most common of cases data item class has been assign to new data item. This is about KNN algorithm. So, the question arises how we calculate the distance between k and new data item? We can measure distance between k and new data point through Euclidean distance. we can also calculate distance through hamming distance, Manhattan distance formula for KNN algorithm.

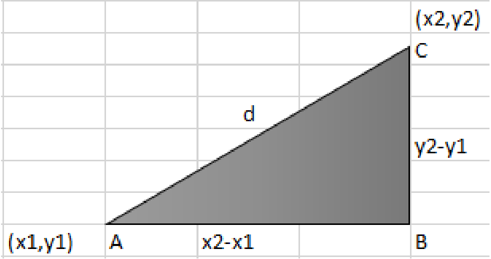
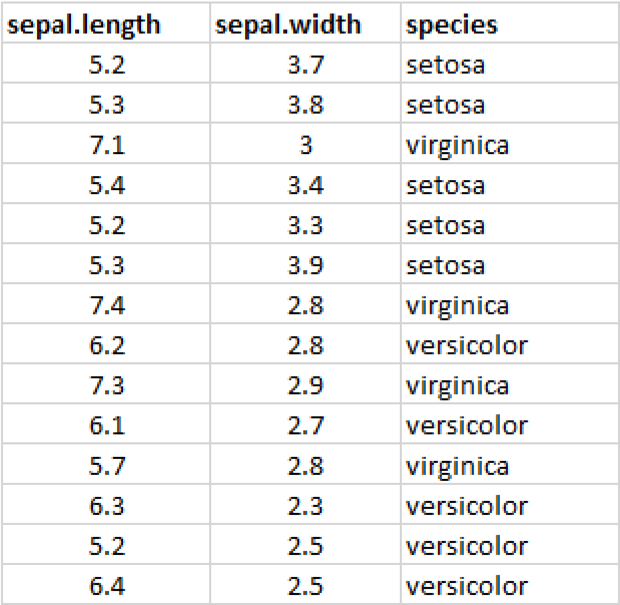


FIGURE 1. digram of paythagorous distance calculation

The euclidean distance is based on the paythgorus therom if we want to to calculate the distance of AC^2(AC SQUAR) then in paythagorus we calculate the AB^2 + BC^2(AB SQUAR + BC AQUAR) this method can be used in paythagoros. Likely if we want to calculate the distance between two point (x1,x2) and (y1,y2) in two dimensional space then the euclidean distance between two points is

D = √(x2−x1)^2 + (y2−y1)^2.(^2 is denoted the square) And in three-dimensional space, for points (x1,y1,z1) and (x2,y2,z2) then in this case the Euclidean distance of this point is D = √(x2−x1)^2 + (y2−y1)^2+(z2-z1)^2. This is how Euclidean distance is calculated.

Now will see how dose KNN algorithm work. We see how the Euclidean distance was calculated. now we apply this distance calculating formula in KNN and see how its work. So, in KNN we have to have data set first, before applying KNN algorithm. KNN will study this data and then predict the new data item class that's why KNN called as a lazy algorithm because it does not have specialized training phase it uses all data for training while classification. We have some flower sepal length and sepal width and their spices. So, if new flower sample is found but we don't know the new flower spices. We apply KNN algorithm to find the classification in flower spices. KNN algorithm use facture similarities to predict the value so the first step is storing the data set during the first step of KNN algorithm.



sepal.length	sepal.width	species
5.2	3.7	setosa
5.3	3.8	setosa
7.1	3	virginica
5.4	3.4	setosa
5.2	3.3	setosa
5.3	3.9	setosa
7.4	2.8	virginica
6.2	2.8	versicolor
7.3	2.9	virginica
6.1	2.7	versicolor
5.7	2.8	virginica
6.3	2.3	versicolor
5.2	2.5	versicolor
6.4	2.5	versicolor

This is the data of the flower sepal length and width. Now we have found a new flower which is look similar like setose, virginica, and versicolor but we could not guess the new flower spices so we will predict the new flower spices using KNN machine learning algorithm. The new flower sepal width is 5.2 and sepal width is 3.2 this is the sample of new flower.

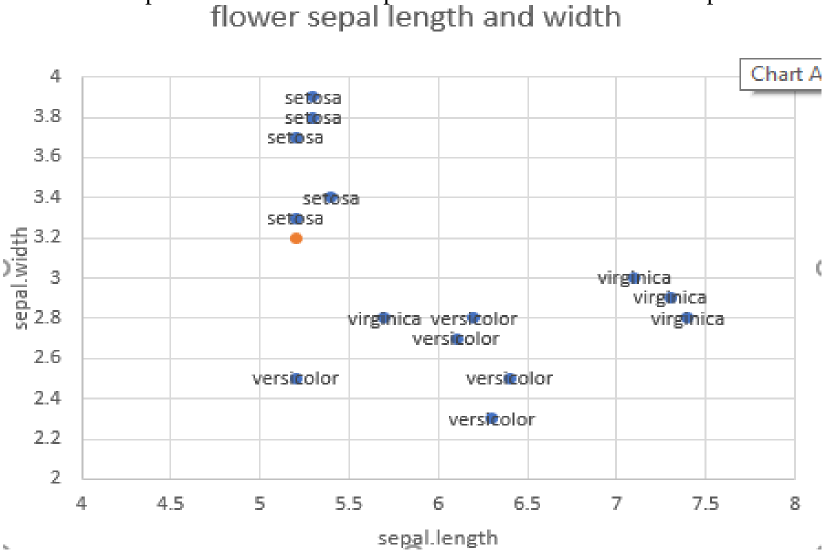
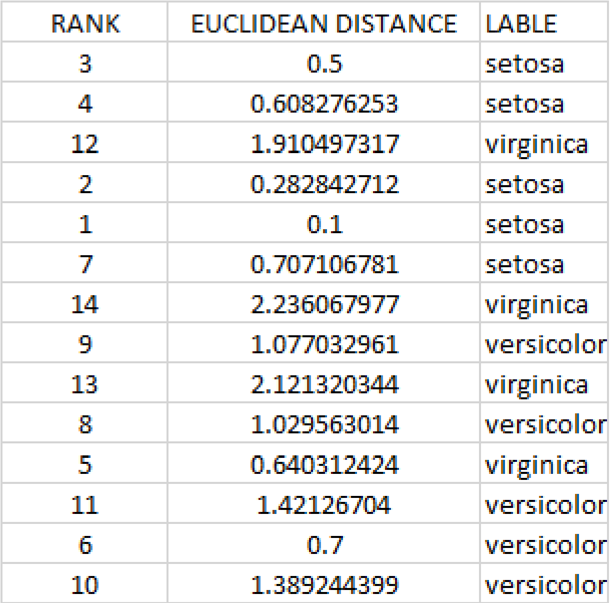


Figure 2. flower sepal length and width

In above scatter chart on x and y axis present the length and width respectively. And new flower sample is in orange dot at (5.2,3.2). so, to apply KNN we have to assign k value which is nearest neighbour, k can be any integer suppose we decide k=3. Then second step is calculating distance between testing point data and each row of data means sepal length and width. Here the formula of the Euclidean distance is D = √(observed length−sample length)^2 + (observed width−sample width)^2. And it could be like that D = √(5.2-5.2)^2 + (3.7-3.2)^2 = 0.509901951 this is the distance between first row data item and new sample data item. Now Calculate distance between all rows and new sample data item. And based on distance, sort them into ascending order. Below is the Euclidean distances and their ranks.



RANK	EUCLIDEAN DISTANCE	LABLE
3	0.5	setosa
4	0.608276253	setosa
12	1.910497317	virginica
2	0.282842712	setosa
1	0.1	setosa
7	0.707106781	setosa
14	2.236067977	virginica
9	1.077032961	versicolor
13	2.121320344	virginica
8	1.029563014	versicolor
5	0.640312424	virginica
11	1.42126704	versicolor
6	0.7	versicolor
10	1.389244399	versicolor

Now we choose k=3, so see first three rank and assign new sample item class to most frequent class of these rows. So, we have first second and third is setose. So, the predicted class of new sample flower is setose. If we choose k=5 then the 5th rank belongs to virginica flower so most frequent case in k =5, is setose then new sample predicted class is setose. KNN have higher accuracy in prediction while available data is in good quality.

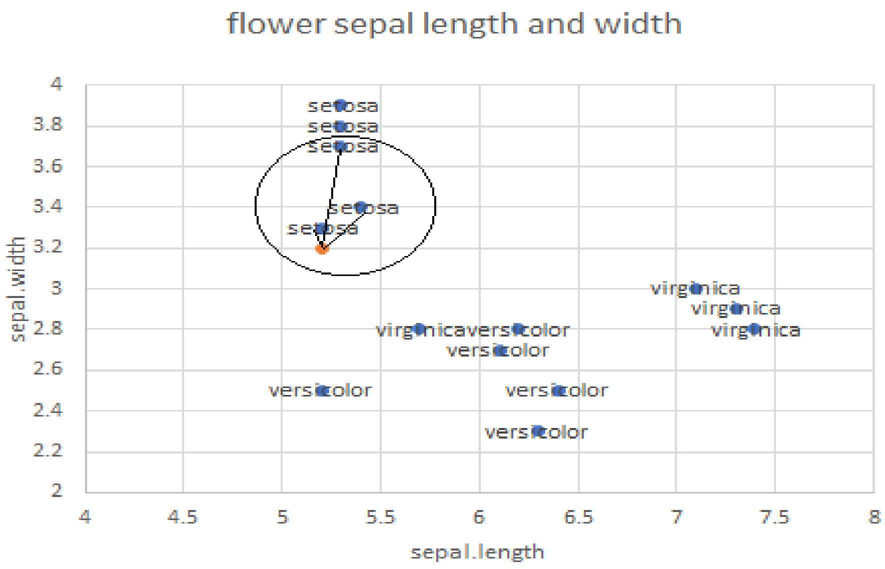


FIGURE 3. AFTER CLASSIFICATION

We have got the sample flower predicted class when we apply KNN machine learning algorithm. And this is accurate if we increase k value. after increasing k value its work efficiently. This is the all how KNN algorithm is work. Its work efficiently in two-dimensional space . in low quality of data set it won’t work efficiently. It is very simple algorithm to implement and understand and we can use this for regression. We have to store first all training data and its make computationally a bit expensive algorithm. In KNN algorithm we need to have high memory storage as compered others supervised algorithm.

Conclusion : KNN, k-nearest neighbour machine learning algorithm is higher accurate with good quality of data and KNN is widely use in different kinds of learning because of its uncomplicated and easy to apply nature. And we have to provide only two metrics in the algorithm first is value of K and second is distance metric. We can add easily to new data to algorithm. It constantly evolves. KNN is instant base learning memory-based approach if we collect new data then we can apply and it allows algorithm to respond quickly to change in the input during real-time use. it’s easy to implement to multi classes, most of classification algorithm easy to implement for binary problem and need efforts to implement for multi classes but KNN algorithm adjust with multi classes without any extra efforts. And KNN have variety of distance criteria to be choose. We can choose Euclidean distance, hamming distance, Manhattan distance for distance calculate between two data points.

The KNN has some drawback which is cost of predicting k nearest neighbour is very high and its does not work expected when we working with large amount of data or with big numbers features/parameters. It is very slow algorithm it's easy to implement but when data set grows the efficiency of speed of algorithm become slow. And one of the biggest issues of KNN algorithm is to choose optimal number of neighbours while classifying new data entry. Imbalanced data problem if we have two classes A and B and majority of the training data is belonging to class A then KNN ultimately gives lot of preferences to class A. and new data entry might be wrongly classified. KNN has no capability to dealing with missing value. This is the all about KNN machine learning algorithm work and its advantages and drawbacks.

References

1. W. Baobao, M. Jinsheng, and S. Minru, "An Enhancement of KNearest Neighbor Algorithm Using Information Gain and Extension Relativity," Proc. International Conference on Condition Monitoring and Diagnosis (CMD 2008), Apr. 2008, pp. 1314-1317, doi:10.1109/CMD.2008.4580218.
2. H. Hong, G. Juan and W. Ben, "An Improved KNN Algorithm Based on Adaptive Cluster Distance Bounding for High Dimensional Indexing," Proc. 3rd Global Congress on Intelligent Systems (GCIS), Conference Publishing Services, Nov. 2012, pp. 213-217, doi:10.1109/GCIS.2012.86.
3. X. Xiao, and H. Ding, "Enhancement of K-nearest Neighbor Algorithm Based on Weighted Entropy of Attribute Value," Proc. 5th International Conference on BioMedical Engineering and Informatics (BMEI 2012), IEEE Press, Oct. 2012, pp. 1261-1264, doi:10.1109/BMEI.2012.6513101.
4. S. Sun and R. Huan, "An Adaptive k-Nearest Neighbor Algorithm," Proc. 7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), vol. 1, IEEE Press and IEEE Circuits and System Society, Aug. 2010, pp. 91-94, doi:10.1109/FSKD.2010.5569740.
5. S. Bo, D. Junping, and G. Tian, "Study on the Improvement of KNearest-Neighbor Algorithm," Proc. International Conference on Artificial Intelligence and Computational Intelligence (AICI 09), vol. 4, IEEE Computer Society, Nov. 2009, pp. 390-393, doi:10.1109/AICI.2009.312
6. J. Gou, L. Du, Y. Zhang and T. Xiong, "A New Distance-weighted k-nearest Neighbor Classifier," Proc. Journal of Information & Computational Science, June 2012, pp. 1429-1436.
7. M. A. Amal and B. A. Riadh, "Survey of Nearest Neighbor Condensing Techniques," Proc. International Journal of Advanced Computer Science and Applications (IJACSA), vol. 2, No. 11, 2011, pp. 59-64.
8. N. Bhatia and Vandana, "Survey of Nearest Neighbor Techniques," Proc. International Journal of Computer Science and Information Security (IJCSIS), vol. 8, 2010, pp. 302-305.
9. S. Taneja, C. Gupta, D. Gureja and K. Goyal, "K Nearest-Neighbor Techniques for Data Classification-AReview," Proc. International Conference on Computing, Informatic and Network (ICCIN2K14), Jan. 2014, pp. 69-73.
10. S. Chen and H. Hsiao, "A New Approach for Fuzzy Query Processing Based on Automatic Clustering Techniques," Information and Management Sciences, vol. 18, 2007, pp. 223-240.
11. Cover T, Hart T P. Nearest neighbor pattern classifi cation [J]. IEEE, 1967 (1): 21 - 27.
12. Cover T. Rates of convergence for nearest neighbor procedures [J]. Systems Sciences, 1968.
13. . Stone C J. Consistent Nonparametric Regression [J]. Institute of Mathematical Statistics, 1977 (7), 5 (4): 595-620
14. Todeschini R, Ballabio D, Consonni V (2006) Distances and other dissimilarity measures in chemometrics. In: Meyer RA (ed) Encyclopedia of analytical chemistry: applications, theory and instrumentation. Wiley, New York, pp 1–34
15. Lopes N, Ribeiro B (2016) On the impact of distance metrics in instance-based learning algorithms. In: Iberian conference on pattern recognition and image analysis. Springer, Berlin, pp 48–56
16. Liu C, Cao L, Philip SY (2014) Coupled fuzzy k-nearest neighbors classification of imbalanced non-IID categorical data. In: 2014 international joint conference on neural networks (IJCNN). IEEE, pp 1122–1129
17. Singh A, Halgamuge MN, Lakshmiganthan R (2017) Impact of diferent data types on classifer performance of random forest, naive Bayes, and k-nearest neighbors algorithms. Int J Adv Comput Sci Appl 8:1
18. Chomboon K, Chujai P, Teerarassamee P, Kerdprasop K, Kerdprasop N (2015) An empirical study of distance metrics for k-nearest neighbor algorithm. In: Proceedings of the 3rd international conference on industrial application engineering, pp 1–6
19. Chen S-M, Chang C-H (2015) A novel similarity measure between atanassov's intuitionistic fuzzy sets based on transformation techniques with applications to pattern recognition. Inf Sci 291:96–114
20. Chen S-M, Cheng S-H, Lan T-C (2016) A novel similarity measure between intuitionistic fuzzy sets based on the centroid points of transformed fuzzy numbers with applications to pattern recognition. Inf Sci 343:15–40.
21. Liu C, Cao L, Philip SY (2014) Coupled fuzzy k-nearest neighbors classification of imbalanced non-IID categorical data. In: 2014 international joint conference on neural networks (IJCNN). IEEE, pp 1122–1129
22. . Ali N, Rado O, Sani HM, Idris A, Neagu D (2019) Performance analysis of feature selection methods for classification of healthcare datasets. In: Intelligent computing-proceedings of the computing conference. Springer, Berlin, pp 929–938
23. Pereira CL, Cavalcanti GDC, Ren TI (2010) A new heterogeneous dissimilarity measure for data classification. In: 2010 22nd IEEE international conference on tools with artificial intelligence, vol 2. IEEE, pp 373–374

24. Deekshatulu BL, Chandra P (2013) Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technol. 10:85–94
25. Walters-Williams J, Li Y (2010) Comparative study of distance functions for nearest neighbors. In: Elleithy K (ed) Advanced techniques in computing sciences and software engineering. Springer, Berlin, pp 79–84