A Brief Review of Nearest Neighbor Algorithm for Learning and Classification

Kashvi Taunk
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
1705409@kiit.ac.in

Sanjukta De
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
1705451@kiit.ac.in

Srishti Verma
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
1705829@kiit.ac.in

Aleena Swetapadma
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
aleena.swetapadma@gmail.com

Abstract—**k-Nearest Neighbor (kNN) algorithm is an effortless but productive machine learning algorithm. It is effective for classification as well as regression. However, it is more widely used for classification prediction. kNN groups the data into coherent clusters or subsets and classifies the newly inputted data based on its similarity with previously trained data. The input is assigned to the class with which it shares the most nearest neighbors. Though kNN is effective, it has many weaknesses. This paper highlights the kNN method and its modified versions available in previously done researches. These variants remove the weaknesses of kNN and provide a more efficient method.**

Keywords— K-Nearest Neighbors; Machine Learning; Lazy Learner; Euclidean distance; Confusion Matrix; Unlabeled data set; Kernel Matrix.

I. INTRODUCTION

The K-Nearest-Neighbors (KNN) is a non-parametric classification algorithm, i.e. it does not make any presumptions on the elementary dataset. It is known for it's simplicity and effectiveness. It is a supervised learning algorithm. A labeled training dataset is provided where the data points are categorized into various classes, so that the class of the unlabeled data can be predicted.

In Classification, different characteristics determine the class to which the unlabeled data belongs. KNN is mostly used as a classifier. It is used to classify data based on closest or neighboring training examples in a given region. This method is used for its simplicity of execution and low computation time. For continuous data, it uses the Euclidean distance to calculate its nearest neighbors.

For a new input, the K nearest neighbors are calculated and the majority among the neighboring data decides the classification for the new input. Even though this classifier is simple, the value of 'K' plays an important role in classifying the unlabeled data.

There are many ways to decide the values for 'K', but we can simply run the classifier multiple times with different values to see which value gives the most effective result. The computation cost is slightly high because all the calculations are made when the training data is being classified, not when it is encountered in the dataset. It is a lazy learning algorithm as not much is done when the dataset is being trained except storing the training data and memorizing the dataset instead.

It does not perform generalization on the training dataset. So the entire fundamental dataset being trained is required when in the testing stage. In regression, KNN predicts continuous values. This value is the average of the values of its K - nearest neighbors.

KNN is used in datasets where data is separated into different clusters so that the class of the new input can be determined. KNN is more significant for a study where there is no previous knowledge about the data being used.

II. K-NN

K-nearest-neighbor classification was developed to execute characteristic analysis when clear parametric approximations of probability densities were unknown or difficult to determine. In an unpublished US Air Force School of Aviation

978-1-5386-8113-8/19/$31.00 ©2019 IEEE

Medicine report in 1951, Fix and Hodges introduced a non-parametric algorithm for pattern classification that has since become known the K-nearest neighbor rule.

A. WORKING

k-NN is a classification algorithm. Mainly there are two steps in classification:

1. Learning Step: Using the training data a classifier is constructed.

2. Assessment of the classifier.

According to the nearest neighbor technique, the new unlabeled data is classified by determining which classes its neighbors belong to. KNN algorithm utilizes this concept in its calculation. In the case of KNN algorithm, a particular value of K is fixed which helps us in classifying the unknown tuple.

When a new unlabeled tuple is encountered in the dataset, KNN performs two operations. First, it analyzes the K points closest to the new data point, i.e. the K nearest neighbors. Second, using the neighbors' classes, KNN determines as to which class the new data should be classified into. Fig. 1 shows a simple K-NN structure.

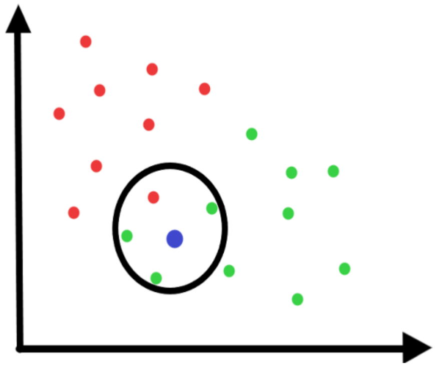
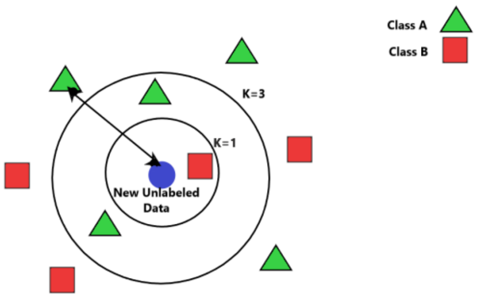


Fig. 1. A simple KNN

When some new data is added, it classifies the data accordingly. It is more useful in a dataset which is roughly divided into clusters and belongs to a specific region of the data plot.

Thus this algorithm brings more accuracy in dividing the data inputs into different classes in a clearer way.KNN figures out the class having the maximum number of points sharing the least distance from the data point that needs to be classified.

Hence, the Euclidean distance needs to be calculated between the test sample and the specified training samples.



As shown in Fig. 2, the new unlabeled data calculate it's the distance from each of it's neighbours according to the value of K. Then, it determines the class it belongs to, containing the maximum number of nearest neighbours.

After we gather K-Nearest Neighbours, we simply take the majority of them to predict the class of the training example. The factors that affect the performance of KNN are the value of K, the Euclidean distance and the normalisation of the parameters. To understand the detailed working of the algorithm, the steps are as follows:

Given the training dataset :

{ (x(1), y(1)) , (x(2), y(2)), …… , (x(m), y(m)) }

Step1: Store the training set

Step2: For each new unlabeled data,

Calculate Euclidean distance with all training data points using the formula:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Find the k- nearest neighbours

Assign class containing the maximum number of nearest neighbours.

After storing the training, set all parameters must be normalized, so that the calculations become easier. The result of the classification is sensitive to the value of 'K'.The input variable 'K' decides the number of neighbours that must be considered. The value of 'K' effects the algorithm as using the 'K' value we can build the boundaries of each class. The best value of K is chosen by first examining the data. Larger values of K are more precise as they reduce the net noise but this is not guaranteed. A good value of K can also be determined using cross-validation.

If K=1, then the data is simply allocated to the class of its nearest neighbour. At K=1, the error rate is consistently zero for the training data. This happens because the nearest point to any training data point is itself. Hence the best results are obtained if the value of K=1. But with K=1, the boundaries are over-fitted. In the case of very small values of 'k' the algorithm is too sensitive to noise.

To get a favourable value of K, the training and validation set must be segregated from the initial dataset. If the two Nearest neighbors (K=2) belong to two different classes, the outcome is unknown.

So, we increase the number of nearest neighbours to a larger value (say, 5-nearest neighbours). This will define the nearest neighbour region and will provide clarity. Larger values of 'K' make the class boundaries smoother, which might not be desirable as then the points of other classes may get included in the neighbourhood. When the training data points are present in a scattered manner, the value of K is difficult to determine.

B. KNN AND ITS VARIANTS

As discussed earlier, the efficiency of the algorithm can be improved by making changes in the factors that govern it. There are many variants of KNN that have been studied before to make this algorithm more effective. Some of them are:

1. Locally Adaptive KNN:

Locally adaptive kNN algorithms proposed by [1]. It chooses the value of k that should be used to classify an input by comparing the results of cross-validation computations in the local neighbourhood of the unlabeled data.

2. Weight Adjusted KNN:

The algorithm by [2] suggests that the distances, on which the search for the nearest neighbours is based in the first step, have to be transformed into similar measures, which can be used as weights. The assigned weights decide how much an attribute influences the classification operation. This classifier is particularly useful in the case where a dataset has many features, some of which can be considered unnecessary, but it has high computational cost.

3. Improved KNN for Text Categorization:

In [3], a refined KNN algorithm for text categorization, which constructs the classification model by merging KNN text categorization and restricted one pass clustering algorithm has been proposed. If a constant value of K is used for all the classes, then the class with a larger number of attributes will have an advantage. In improved KNN, a suitable number of nearest neighbours are used according to the distribution of data in the training set, to predict the class of unlabeled data.

4. Adaptive KNN:

KNN identifies the same number of nearest neighbours for each new input. Adaptive KNN by [4] finds out a fit value of K for each test sample. First, an optimal value of K is found. Then, to predict the classification of the unlabeled data, the value of K is set equal to the optimal value of K of it's nearest neighbour in the training dataset. The execution of the suggested algorithm is then tested on different datasets.

5. KNN with Shared Nearest Neighbours

An improved K-nearest neighbour algorithm is presented by [5] using shared nearest neighbour similarity which can compute the similarity between test samples with nearest neighbour samples. It uses a Similarity judgment algorithm and calculates the nearest neighbour similarity value for each training sample.Then it calculates the maximum between these values.

6. KNN with K-Means:

Another improvised approach to the algorithm is depicted by [6]. This algorithm tries to divide a set of points into K sets or clusters so that the points in each cluster are close to each other. The centres of these freshly made clusters are taken as the new training samples. To predict the classification of unlabeled data, its distance from the newly found training centre is calculated, and the centre which shares the minimum distance from the data is allocated to that class. Unlike standard KNN, there is the input parameter K is not passed. This accounts for being one of its advantages.

7. SVM KNN

Support Vector Machine (SVM) is a classification technique that can be applied on linear as well as non-linear data. It is a composite version of KNN mixed with SVM for visual category recognition and is augmented in [7]. In this algorithm, the training is done with the help of K nearest neighbours to the unlabelled data point. First, the K-nearest data points are determined. Then, the pairwise distance between these K data points is computed. Thus we obtain a distance matrix from the calculated distances. A Kernel matrix is then designed from the obtained distance matrix. This kernel matrix is fed as input to SVM classifier. The result obtained is the class of the unknown data point. Alternatively, one could use SVMs but time consumption is one of its drawbacks. Also, it involves the calculation of pairwise distances.

8. KNN with Mahalanobis Metric

The metric distance is of great importance in the classification of a new data point. Mahalanobis is a new distance metric, the approach of which is covered in [8]. The metric ensures that the K-nearest neighbours are contained in the same class and the samples belonging to different classes are separated by a large degree of difference.

9. Generalized KNN

KNN can also be used for continuous - valued class attributes. For this classification, the average values calculated among neighbours is allocated to the class attribute of the unlabeled data. [9] implements this algorithm to predict the continuous-valued class attribute.

10. Informative KNN

Usually, the value of K relies on the data, making it difficult to choose the parameter in accordance with different applications. [10] introduced a new metric that measures the informativeness of objects to be classified. Informativeness measures the importance of points. In this method, there are two input parameters K and I. The majority class of most informative training examples will be the class of the new test sample.

11. Bayesian KNN

The data values surrounding the target are generated by the same probability distribution, expanding outwards over the suitable number of neighbours. [11]recursively computed the probability of the last change-point and moved towards the target, and computed the posterior probability distribution over K.

III. ADVANTAGES AND DISADVANTAGES

A. *Advantages*

The KNN is known for its simplicity, comprehensibility, and scalability. It is easy to interpret. The calculation time is less. Also, the predictive power is very high which makes it effective and efficient. KNN is very effective for large training sets. The steps followed in the classification done by this algorithm are relatively less complex than those followed by other algorithms. The mathematical computations are easy to comprehend and understand. They do not involve calculations that seem difficult.

Basic concepts like that of Euclidean distance calculation are used which enhance the simplicity of the algorithm instead of opting for other composite methods like that of integration or differentiation. It is useful for non-linear data. KNN is effective for classification as well as regression.

B. *Disadvantages*

The KNN can be expensive in the determination of K if the dataset is large. It requires greater storage than an effective classifier. In KNN the prediction phase is slow for a larger dataset. Also, the computation of accurate distances plays a big role in the determination of the algorithm's accuracy. One of the major steps in KNN is determining the parameter K. Sometimes it is not clear which type of distance to use and which feature will give the best result. The computation cost is quite high as the distance of each training example is to be calculated. KNN is a lazy learning algorithm as it doesn't learn from the training data, it implies memorizes it and then uses that data to classify the new input.

IV. APPLICATIONS OF K-NN

A. *Medical Predictions*

In medicine, prediction plays an important role. Prediction of a second heart attack in a patient admitted due to a heart attack, or if a person with increased cholesterol may have a cardiac arrest or the chances of cancer in a patient. These predictions are based on several factors such as region, diet, hereditary factors and height, weight and other clinical measurements for that patient. We will analyse its use in predicting cardiac arrests. Heart disease is one of the most leading causes of death for the past 10 years. Health care professionals store huge amounts of patients' data that can be used for statistical analysis and data mining techniques for the diagnosis of heart disease. The risk factors associated with heart disease were found out to be:

Age: The risk of heart disease increases with age. Around 4 out of every 5 heart-related deaths transpire in people older than 65.

Blood pressure and cholesterol: High blood pressure and cholesterol can increase the chances of heart disease significantly.

Smoking: It increases heart rate, leads to the dysfunctionality of major arteries, and can create contortion in heartbeats and is seen to increase the chances of having a heart disease majorly.

Heredity: Having a relative with a heart disease significantly increases the chances of having one.

Obesity: It also contributes majorly to the risk of having a heart disease.

Having analyzed the factors that are needed to be studied in order to make a proper database for operation, the KNN classification algorithm can be applied to it. The difference between the attributes is calculated using the Euclidean distance. However, one drawback is that the frequency of the large values has an advantage over the smaller ones. For example, in heart disease records, the pressure level of the patients is recorded between ranges between 80 and 150 while the age has ranged between 40 and 90. So the impact of the blood pressure measurements will be higher than the age which is not correct. In such cases, normalization is done on the continuous attributes so that they have the same influence on the distance measure between instances. In the case of distinct attributes, if the attribute values for the two instances are equal, the difference between them is taken as zero otherwise one. The accuracy of knn is higher than other algorithms in the prediction of heart diseases.

B. Data Mining and Financial Modeling

Data mining is extracting useful information from a huge set of Data. This has become really important as real life data increases enormously. Data mining as a discovers useful patterns and associations which has its applications in financial modeling. Stock market forecasting is one of the most important financial tasks of KNN. It includes exposing the latest market trends, analysing market conditions, planning investment strategies, which stock to purchase and identifying the best time to purchase the stocks. The price of the stock depends on numerous factors that help in predicting stock prices that include a deep analysis of the company's business and financial data. They include:

Changes in market pattern and trends

company's performance

Demographics

currency exchange rates

Inflation

credit ratings

These factors can be expressed in numeric forms and analyzed to create a database. In classification, the elementary dataset is divided into two sets - a training set and a testing set. KNN uses compare an unlabeled data with the training dataset as discussed in [12]. Each tuple represents a record with n features. In order to predict a value for an unknown data. In KNN, K records of the training set are chosen that are nearest to the unknown data. Setting a proper value of K minimizes the error and gives a nearly accurate result.

V. COMPARATIVE STUDY

The main difference between KNN and other models is that KNN needs more computational time than other algorithms. Neural Networks, for instance, needs more training data to achieve precision. Also, if the training data is much larger than the number of features, KNN is considered better than SVM. But, SVM is considered better if the number of features is much more than the training data. KNN supports non-linear solutions whereas, logistic regression only supports linear solutions. Although, Logistic regression is relatively faster than KNN. KNN is a non - parametric model, whereas Linear Regression is a parametric model. KNN is slow in real time as it has to keep account of all training data and find the nearest neighbours, whereas Linear Regression can easily derive solutions from the tuned coefficients. Naive Bayes, as another example, is much faster than KNN due to KNN's real-time execution. Naive Bayes is parametric whereas KNN is non-parametric. The decision tree, another classification technique, supports automatic feature interaction, whereas KNN does not. Hence, KNN is a simple machine learning algorithm with local approximation. Since there is no training involved in KNN, it is faster and has fewer parameters to tune. Proper scaling must be done for a proper result. KNN, with its computational simplicities, serves a number of purposes that other classification methods cannot incorporate. But this simplicity also fails to challenge the fact that it may not be the most efficient algorithm in designing solutions that demand more complex models for accurate results. KNN manages to cover commendable domains of data prediction, but it still has a long way to go to stand out as the most efficient algorithm serving the purpose of Machine Learning as compared to other designed algorithms.

VI. CONCLUSION

Machine learning algorithms have improved with the increase in research and data mining tools. K- nearest neighbour algorithm is a simple but high accuracy algorithm that has proven effective in several cases. The above shows two well-known applications of this algorithm i.e healthcare and Stock Market Forecasting. The nearest neighbour algorithm works by classifying the new unlabeled data by examining the classes of it's nearest neighbours. In KNN algorithm, a constant number of nearest neighbours determine the classification of an unlabeled data which is assigned by K, where K is a positive integer.

The value of K is important as it determines the accuracy and effectiveness of the algorithm. There are other proposed extensions for KNN algorithm which are Locally adaptive KNN classifier, K-means KNN classifier, weighted KNN classifier, Shared Nearest Neighbour KNN classifier, SVM KNN classifier. These save the time of execution and improve accuracy. It makes predictions more effective by adding in variables other than k. KNN is highly effective in determination of disease like diabetes and heart risks among others. KNN also finds application in stock market predictions and financial modeling.

Further research is needed to improve the classification accuracy of marginal data which falls outside the regions of typical data. The reliability and the scaling properties of the classifier are being investigated. Another area of research is the size of the dataset. All machine learning algorithms need huge datasets for training. Sometimes we do not have a huge dataset to operate upon like in human gene data. Then, we need to find the most effective algorithm to find maximum accuracy. There is ongoing research on the different kinds of knn other than standard knn. All found until now provide higher speed but lower accuracies. It is also being used in the surveillance system, face recognition, fault detections, etc. this uses knn along with svm to increase efficiency and recognition rates. hence, advancements are being made every day in technology with exponential growth, artificial intelligence being an important area. The improvement in Knn classifier can to increase speed and accuracy and can improve machine learning algorithms to great extends.

References

[1] D. Wettschereck and D. Thomas G., "Locally adaptive nearest neighbour algorithms," Adv. Neural Inf. Process. Syst., pg. 184–186, 1994.

[2] Han EH.., Karypis G., Kumar V. (2001) "Text Categorization Using Weight Adjusted k-Nearest Neighbour Classification". In: Cheung D., Williams G.J., Li Q. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2001. Lecture Notes in Computer Science, vol 2035. Springer, Berlin, Heidelberg.

[3] Shengyi Jiang,Guansong Pang,Meiling Wu,Limin Kuang, "An improved K-nearest-neighbour algorithm for text categorization",Expert Systems with Applications,Elsevier(2012)

[4] Shiliang Sun ; Rongqing Huang," An adaptive k-nearest neighbour algorithm", in 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery

[5] Wei Zheng,HaiDong Wang,Lin Ma,RuoYi Wang,"An Improved k-Nearest Neighbour Classification Algorithm Using Shared Nearest Neighbour Similarity" ,Metallurgical & Mining Industry . (2015), Issue 10, pg. 133-137. 5p.

[6] P. WiraBuana, S. Jannet D.R.M., and I. Ketut Gede Darma Putra, "Combination of K-Nearest Neighbour and K-Means based on Term Re-weighting for Classify Indonesian News," Int. J. Comput. Appl., vol. 50, no. 11, pp. 37–42, Jul. 2012.

[7] Hao Zhang ; A.C. Berg ; M. Maire ; J., Malik"SVM-KNN: Discriminative Nearest Neighbour Classification for Visual Category Recognition", 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)

[8] "Mahalanobis distance," Wikipedia, the free encyclopedia. 04-Mar-2016.

[9] M. Sharma and S. Sharma, "Generalized K-Nearest Neighbour Algorithm- A Predicting Tool," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 11,(Nov. 2013).

[10] Song Yang, Jian Huang, Ding Zhou, Hongyuan Zha, and C. Lee Giles, "Iknn: Informative k-nearest neighbour pattern classification," in Knowledge Discovery in Databases, (2007), pg. 248–264.

[11] Giuseppe Nuti,"An Efficient Algorithm for Bayesian Nearest Neighbours",Cornell University, (2017)

[12] Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm Khalid Alkhatib, Hassan Najadat Ismail Hmeidi , Mohammed K. Ali Shatnawi, International Journal of Business, Humanities and Technology,Vol. 3 No. 3; March 2013