**Abstract**

K-Nearest Neighbour Algorithm for Learning And Classification

K-Nearest Neighbour (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 neighbours. 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 Neighbours; Machine Learning; Lazy Learner; Euclidean distance; Confusion Matrix; Unlabelled data set; Kernel Matrix.

1. **INTRODUCTION**

The K-Nearest-Neighbours (KNN) is a nonparametric 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 labelled training dataset is provided where the data points are categorized into various classes, so that the class of the unlabelled data can be predicted. In Classification, different characteristics determine the class to which the unlabelled data belongs. KNN is mostly used as a classifier. It is used to classify data based on closest or neighbouring 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 neighbours. For a new input, the K nearest neighbours are calculated and the majority among the neighbouring 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 unlabelled 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 neighbours.

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.

1. **K-NN**

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

1. **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 neighbour technique, the new unlabelled data is classified by determining which classes its neighbours 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 unlabelled 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 neighbours. Second, using the neighbours’ classes, KNN determines as to which class the new data should be classified into. Fig. 1 shows a simple K-NN structure.

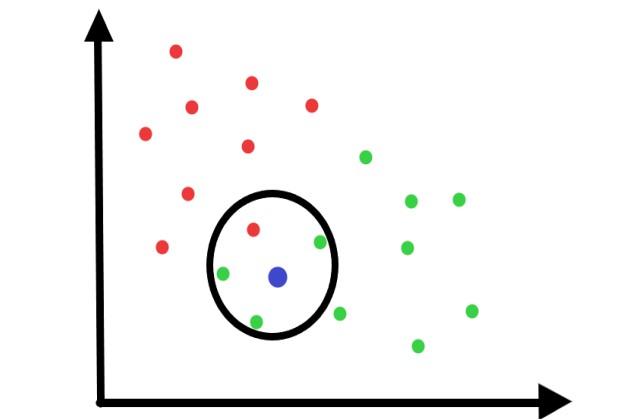


Fig. 1. A simple KNN

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.

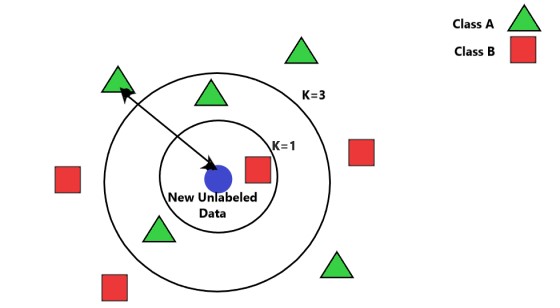


Fig. 2. K-NN with two classes A and B. As shown in Fig. 2, the new unlabelled data calculate it’s the distance from each of it’s neighbours according to the value of K. Then, it determines the class it belongs to, containing the maximum number of nearest neighbours.

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 unlabelled data,

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



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 neighbours (K=2) belong to two different classes, the outcome is unknown.

1. **ADVANTAGES AND DISADVANTAGES**
2. **Advantages**

The KNN is known for its simplicity, comprehensibility, and scalability. It is easy to interpret. The calculation time is less. Also, the predictive

**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 analysed. 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

**B. Data Mining and Financial** Modelling 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 modelling. 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 analysed 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 unlabelled 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

**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 unlabelled data by examining the classes of it’s nearest neighbours. In KNN algorithm, a constant number of nearest neighbours determine the classification of an unlabelled 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 modelling. Further research is needed to improve the classification accuracy of marginal data which falls

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