Performance Analysis of K-Nearest Neighbor (KNN) Classifier for Optimum K Value

Laxmi Pawar[1] Mr. Manoj Soni[2]
[1]PG Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]JIT Borawan Khargone, India

Abstract— The K-nearest neighbor (KNN) is one of the simplest and accurate techniques for classification. K-nearest neighbors algorithm is an example of supervised learning. K-nearest neighbor can be used for predictive of classification and regression problems. In K-nearest neighbor, nearest neighbor is measured with respect to value of k. The value of K defines how many nearest neighbors are need to examine. The main advantages of K-nearest neighbors technique is its effectiveness for large training data. It is also robust to noisy training data. K-nearest neighbors is a lazy learning algorithm. In K-nearest neighbors there is no assumption for underlying data distribution. There are several algorithms and methods have been developed to solve the problem of classification. But for finding new algorithm is a process for improving accuracy and efficiency. In this paper we try to determine the best value for K, by continues trying a few values before settling final values.

Keywords: K-Nearest Neighbor (KNN)

I. INTRODUCTION

KNN is a lazy learning algorithm. In KNN there is no assumption for underlying data distribution. KNN will be very helpful in practice for most of the real world datasets do not follow mathematical assumptions. KNN does not need any training data points for model construction. In KNN, K is the number of nearest neighbors. The number of neighbors is the main deciding factor.

To make you understand how KNN algorithm works, let's consider the following scenario

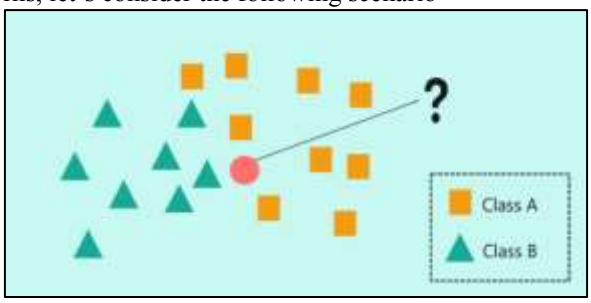

Fig. 1: Two classes with one unknown objects

In figure 1 there two classes of data, namely class A and Class B .The problem statement is to assign the new input data point to one of the two classes. First we need s to define the value of K. K stands for the number of Nearest Neighbors

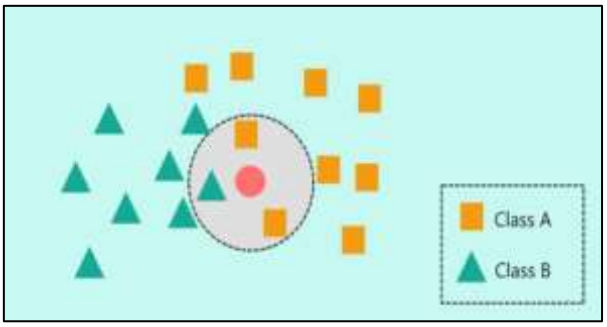

Fig. 2: K=3 (Three nearest Neighbors)

In figure 2 we decide the value of K as 3. We need to consider the three neighbors that are the closest to unknown data point to decide the class of this data point. For K = 3, the neighbors include two object of class A and one object of class B. So, unknown data point classifies and would be assigned to Class A.

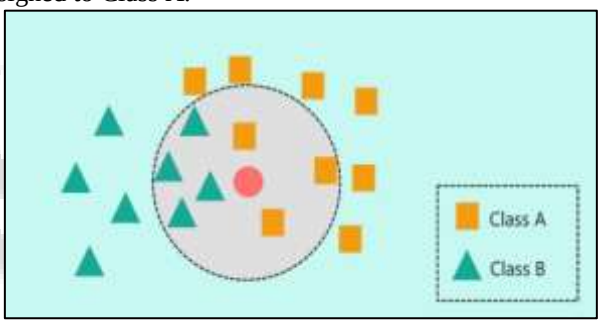

Fig. 3: K=7 (Seven nearest neighbors)

In figure 3 we decide the value of K as 7. We need to consider the three neighbors that are the closest to unknown data point to decide the class of this data point. For K = 7, the neighbors include three object of class A and four object of class B. So, unknown data point classify and would be assigned to Class B .

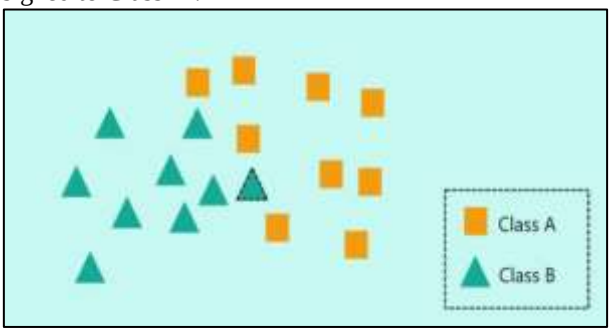

Fig. 4: Unknown object belongs to class B

II. FEATURES OF KNN

The KNN algorithm has the following features:

- KNN is a Supervised Learning algorithm.
- It is one of the most simple algorithms and it can be easily implemented

- It is mainly based on feature similarity.
- KNN does not make any assumptions about the data set.
- It memorizes the training data set instead of learning a discriminative function
- KNN can be used for solving both classification and regression problems

III. LITERATURE SURVEY

In 2010 Shiliang Sun e al proposed "An Adaptive *k*-Nearest Neighbor Algorithm". Performance of the KNN algorithm, and proposed KNN algorithm focuses on finding out the suitable K for each test example. The proposed algorithm finds out the optimal K, the number of the fewest nearest neighbors that every training example can use to get its correct class label. They proposed an adaptive KNN algorithm is proposed for classification. Making use of the traditional KNN algorithm and the rule that the nearest neighbors have similar attributes. They showed the feasibility of this algorithm. Experimental results also show that the proposed algorithm is superior to the traditional KNN algorithms in most cases [1].

In 2011 P. Viswanath et al proposed "An Improvement to k-Nearest Neighbor Classifier". They presented a novel improvement to the k-NNC called k-Nearest Neighbor Mean Classifier (k-NNMC). K-NNMC finds k nearest neighbors for each class of training patterns separately, and finds means for each of these k neighbors (class-wise). They proposed method does not have a design phase as the Hamamoto's method, and this is suitable for parallel implementations which can be coupled with any indexing and space reduction methods easily. It is a suitable method to be used in data mining applications. They presented a novel fundamental improvement to the well-known k-nearest neighbor classifier[2].

In 2012 Loai Abed All al proposed "*k* Nearest Neighbor using Ensemble Clustering". They found that objects belonging to the same cluster usually share some common traits even though their geo-metric distance might be large. They decided to define a metric based on clustering. As there is no optimal clustering algorithm with optimal parameter values, several clustering runs are performed yielding an ensemble of clustering (EC) results. They run result of the algorithm on equivalence classes instead of single objects. By the experiments they showed the number of equivalence classes is usually one tenth to one fourth of the number of objects[3].

In 2013 Vasile Lucian et al proposed "Improving Recall of k-Nearest Neighbor Algorithm for Classes of Uneven Size". They describe a method of weighting the prototypes for each class of the k nearest neighbors to cope with the uneven distribution of data. The proposed method increases the classification rate in terms of recall measure To cope with the unbalanced category distribution in a classification task using the KNN method, they proposed and tested two approaches based on averaging the similarity functions (named M3 and M4). One conclusion is that it is better to weight the sum of similarities to the whole number of prototypes from a class (like in M4) than to the number of prototypes per class from the k selections [4].

In 2014 Shichao Zhang and Xuelong Li proposed "Efficient KNN Classification with Different Numbers of Nearest Neighbors". They proposed a k-Tree method to learn different optimal *k* values for different test/new samples, by involving a training stage in the KNN classification. They proposed K Tree method has a similar running cost but higher classification accuracy, compared with traditional KNN methods, which assign a fixed *k* value to all test samples. The proposed K Tree method needs less running cost but achieves similar classification accuracy, compared with the newly KNN methods, which assign different *k* values to different test samples[5].

In 2015 Omisore, O. M.and Azeez N. A. proposed "Predicting Academic Performance of Students in Higher Institutions with k-NN Classifier". They adopted k-Nearest Neighborhood (k-NN) classifier for predicting academic performance of students in higher institution. They proposed a study for neighborhood classification approach (k-NN) is applied to predict the academic performance of students in institutions of higher learning. The use of data mining in educational field is being adopted to enhance humans' understanding of learning process [6].

In 2016 Alka Lamba et al proposed "Survey on KNN and Its Variants". They discussed about KNN algorithm and its various modified versions available in formerly done studies. Despite the simplicity of KNN algorithm there are a number of shortcomings in it i.e. high computational cost, large memory requirement, equal-weighted features and in last deciding appropriate value of the input parameter k. The proposed modified KNN algorithms can be broadly classified into two categories. The first category encompasses those variants of KNN that tried to improve the performance of KNN by making alterations to the result influencing factors for KNN i.e. deciding appropriate value of KNN, distance metric to be used to measure similarity between data tuples, weights assigned to the attributes of data set[7].

In 2017 Okfalisa et al proposed "Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification". They analyze the comparison of KNN and MKNN algorithms to classify the data of Conditional Cash Transfer Implementation Unit which consist of 7395 records. The results of proposed KFold Cross Validation modeling will be the model for training data samples and testing data to test KNN and MKNN for classification. Classification result produced accuracy based on the rules of confusion matrix. Almost all testing from the first test up to the tenth, MKNN algorithm is superior and has better accuracy value than KNN so it can be analyzed that the ability of MKNN algorithm in accuracy is better than KNN[8].

In 2018 Jingwen Sun,et al proposed "A Survey of kNN Algorithm". They summarize the kNN algorithm and related literature, introduce the idea, principle, implementation steps and implementation code of kNN algorithm in detail, and analyze the advantages and disadvantages of the algorithm and its various improvement schemes. They introduced the development of kNN algorithm and its importance. They showed the application of kNN algorithm is introduced, and its implementation in text classification is emphasized[9].

In 2018 Md Rafiqul Islam et al proposed "Detecting Depression Using K-Nearest Neighbors (KNN) Classification Technique". They considered depression problems among various Face book users. Already, a number of researchers have studied and applied many techniques to detect depression, but still need to detect accurately from social network data. They investigate the possibility to utilize Face book data and apply KNN (k-nearest neighbors) classification technique for detecting depressive emotions. They studied four types of factors (emotional process, temporal process, linguistic style, and all features) and trained a model to utilize each type independently and jointly[10].

In 2018 Muhammad Azam et al proposed "Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm" .They shown better performances of classification algorithms such as Naïve Bayes (NB) and K-Nearest Neighbor (K-NN) improvement using Bayesian boost and bagging. They showed that boosting and Bagging remarkably increased the overall accuracy using Naïve Bayes Classifier, whereas, performance of k-NN remained unchanged and better than Naïve Bayes classifier. Additionally, They shown that Bagging and Boosting brought an impact in the performances using k-NN and concluded that k-NN with Bagging and Boosting revealed better classifier for classifying the scientific publications[11].

In 2019 Wenchao Xing et al proposed "Medical Health Big Data Classification Based on KNN Classification Algorithm". They proposed an improved KNN algorithm and compares it with the traditional KNN algorithm. They proposed an improved KNN algorithm based on cluster denoising and density cropping. By the experimental results they showed that the proposed algorithm can effectively improve the classification efficiency of KNN algorithm in processing large data sets, and maintain the classification accuracy of KNN algorithm. They modified the algorithm for the latest classifier on the dataset. The data sets used have different minority category percentages. The advantages and disadvantages of traditional KNN classifiers are analyzed[12].

In 2019 Harikumar Rajaguru et al proposed "Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer". They proposed a model employs two machine learning (ML) algorithms for the categorization of breast tumour; Decision Tree and K-Nearest Neighbour (KNN) algorithm is used for the breast tumour classification. Result: They proposed model using decision tree and KNN algorithm is implemented after feature selection which shows that the lazy learning algorithm is more effective than the decision-tree algorithm[13].

IV. PROBLEM STATEMENT

K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. To evaluate any technique we generally look at 3 important aspects:

1) Ease to interpret output
2) Calculation time
3) Predictive Power

V. KNN ALGORITHM

1) Calculate D(x, x_i) i =1, 2, ….., **n**; where **D** denotes the Euclidean distance between the points.
2) Arrange the calculated **n** Euclidean distances in non-decreasing order.
3) Let **k** be a +ve integer, take the first **k** distances from this sorted list.
4) Find those **k**-points corresponding to these **k**-distances.
5) Let $\mathbf{k_i}$ denotes the number of points belonging to the i^{th} class among **k** points i.e. $k \geq 0$
6) If $k_i > k_j \; \forall \; i \neq j$ then put x in class i.

VI. IMPLEMENTATION AND RESULT ANALYSIS

Value of K	Number of Records tested	Number of record and correctly classified
9	20	11
13	25	21
17	30	19

TABLE 1 Different value of K and number of records

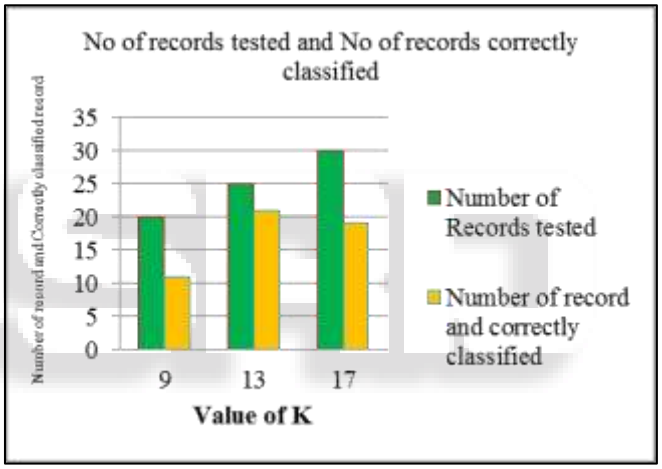

Fig. 5: Graph with Different value of K and Number of Records Correctly classified

VII. CONCLUSION AND FUTURE WORK

There are several algorithms and methods have been developed for classification. The most popular classification methods are artificial neural networks, Decision Tree, Support Vector Machine Naïve Bayes Classifier and K-Nearest Neighbor Classifier. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency. In the proposed work we used K-Nearest Neighbor Classifier Algorithm and analysis this algorithm for different value of K. From experiment it clear that value of should always take odd number. The value of parameter K is taken in such a way so that it is not very small and not very big. Deciding value of k is a continue process until number of records are correctly classified. K-NN Algorithm is also simple to understand and calculation is easy.

REFERENCES

[1] Shiliang Sun Rongqing Huang "An Adaptive *k*-Nearest Neighbor Algorithm" 2010 Seventh International

Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010) 978-1-4244-5933-9/10/$26.00 ©2010 IEEE.

[2] P. Viswanath and T. Hitendra Sarma "An Improvement to k-Nearest Neighbor Classifier" Conference Paper · September 2011 DOI: 10.1109/RAICS.2011.6069307

[3] Loai AbedAllah and Ilan Shimshoni ***k*** Nearest Neighbor using Ensemble Clustering Conference Paper · September 2012 DOI: 10.1007/978-3-642-32584-7_22 at:

[4] Vasile Lucian Boiculese, Gabriel Dimitriu "Improving Recall of k-Nearest Neighbor Algorithm for Classes of Uneven Size" The 4th IEEE International Conference on E-Health and Bioengineering - EHB 2013.

[5] Shichao Zhang, and Xuelong Li, "Efficient kNN Classification with Different Numbers of Nearest Neighbors" IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 5, MAY 2018.

[6] Omisore, O. M and Azeez, N. A "Predicting Academic Performance of Students in Higher Institutions with k-NN Classifier" Conference'15, August 20–22, 2015, Ibadan, Ondo State, Nigeria. Copyright 2015 ACM 1-58113-000-0/00/0010.

[7] Alka Lamba, Dharmender Kumar " Survey on KNN and Its Variants" International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 5, May 2016

[8] Okfalisa and Mustakim "Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification" Conference Paper · November 2017 https://www.researchgate.net/publication/323056633. ICITISEE.2017.8285514.

[9] Jingwen Sun,Weixing Du,Niancai Shi "A Survey of kNN Algorithm" Information Engineering College, Panzhihua University of Technology, Sichuan, China Information Engineering and Applied Computing (2018)

[10] Md Rafiqul Islam1, Abu Raihan M.Kamal "Detecting Depression Using K-Nearest Neighbors (KNN) Classification Technique" Conference Paper · February 2018 DOI: 10.1109/IC4ME2.2018.8465641.

[11] Muhammad Azam, Tanvir Ahmed "Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm" IJCSNS International Journal of Computer Science and Network Security, VOL.18 No.12, December 2018

[12] Wenchao Xing 1 and Yilin Bei "Medical Health Big Data Classification Based on KNN Classification Algorithm" Received November 4, 2019, accepted November 14, 2019, date of publication November 25, 2019, date of current version February 17, 2020.

[13] Harikumar Rajaguru, Sannasi Chakravarthy " Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer" *DOI:10.31557/APJCP.2019.20.12.3777 Breast Cancer Classification*

