All Sciences Proceedings
All Sciences Proceedings
http://as-proceeding.com/

3rd International Conference on Innovative Academic Studies
September 26-28, 2023 : Konya, Turkey

ICIAS
https://www.icias.net/

The Impact of the Distance Weighting Function on the Performance of K-Nearest Neighbor Algorithm in Medical Data sets

Elif Varol Altay

Department of Software Engineering, Manisa Celal Bayar University, Turkey

elif.altay@cbu.edu.tr

Abstract – The K nearest neighbor (KNN) technique is well recognized and extensively used in the field of machine learning classification algorithms. In this study, the performance of distance weighting functions, one of the most important factors affecting the performance of KNN, was compared. These functions are equal, inverse, and squared-inverse. The performance of the functions was examined in five different medical data sets. To evaluate the performances, a confusion matrix and five different metrics commonly used to evaluate classification problems were used. Out of the five data sets used in the research, the inverse KNN method yielded effective outcomes in four of them, while both the equal and squared-diverse methods achieved success in three data sets.

Keywords – K- Nearest Neighbor, Classification, Distance Weighting Function, Medical Data Sets, Machine Learning

I. Introduction

The K nearest neighbor (KNN) algorithm is a commonly used classification approach that remains popular owing to its simplicity in both conceptual understanding and practical implementation. The technique employs a classification approach that relies on the notion of using the majority vote of k neighboring instances to determine the class labels. This determination is made by evaluating the distance between the unlabeled sample and all the labeled samples in the training set using a range of distance formulae [1]. Therefore, it may be inferred that all characteristics provide an equal contribution to the categorization process. Nevertheless, this circumstance is not always considered favorable. The effect of the feature index on the classification result is significant, and the typical KNN method does not consider the weight of the classification feature index. Also, the ability of the KNN method to classify is affected by other factors, such as the number of neighborhoods (k) and the choice of a good distance function for defining neighbors [3]. So, the weighted k-nearest neighbors algorithm, which uses the feature index weights, may greatly improve the accuracy of classification if the optimal number of neighbors (k) and the right distance function are carefully chosen. Numerous experiments have been undertaken over the years to enhance the classification accuracy of the KNN algorithm via the use of various weighting strategies [3]. KNN is widely used in different fields in the literature to solve classification problems. Water quality prediction [4], English language readability [5], Covid 19 studies [6-7], fresh performance of steel fiber reinforced self-compacting concrete prediction [8], E-mail spam detection [9], Warfarin dosage prediction [10], fall detection [11] and stock movement prediction [12] can be given as examples of these fields.

This research investigated the impact of the weighting function parameter on the classification performance of the K-NN algorithm. In this study, three different versions of the KNN algorithm were applied on 5 different medical data sets. The weighting approach used in this study involves assigning weights based on the inverse of the distance and the inverse of the square of the distance. The classification performance of the

KNN algorithm was then compared with the results obtained using these weighting methods. The remaining part of the article is as follows: In the 2nd section, materials and methods are included. Section 3 contains experimental results. In the fourth chapter, the results are given and future studies are mentioned.

II. Materials and Method

A. Properties of the Data Sets

Five medical data sets with different numbers of features and different numbers of samples were selected from the UCI machine learning repository [13]. Details of these data sets are listed in Table 1.

Table 1. Details of data sets

Data sets	Number of samples	Number of feature	Train	Test	Number of class
Cervical cancer	72	19	50	22	2
Cleveland	303	13	212	91	2
Mammographic	830	5	581	249	2
Pima	768	8	537	231	2
Spectfheart	267	44	186	81	2

B. Data Pre-Processing

When there is a significant disparity in the data, reducing the data to a single order yields more accurate results. The attributes of the data sets utilized in the research were distributed between 0 and 1 for this purpose using min-max normalization. Eq. 1 describes min-max normalization.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

C. Performance Evaluation

When evaluating the performance of a classification algorithm, it is customary to use evaluation metrics such as accuracy, sensitivity, specificity, precision, and F1 score to gauge the algorithm's efficacy. The use of the confusion matrix facilitates the computation of these measures. A confusion matrix is a commonly used tabular representation that effectively demonstrates the efficacy of a classification model when evaluated against a predetermined set of test data. The confusion matrix has four factors, namely True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Table 2 illustrates the configuration of the confusion matrix. The performance measuring formulae based on the confusion matrix are defined in Eqs (2-6).

Table 2. Confusion matrix

		Predicted Class	
Actual Class		1	0
	1	TP	FN
	0	FP	TN

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Fmeasure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (6)$$

D. K Nearest Neighbor Algorithm

The KNN technique, first introduced by T. M. Cover and P. E. Hart, is widely recognized as a straightforward, efficient, and widely used classification approach within the field of machine learning [14]. The KNN technique computes the distance between a new sample and the existing data points in order to identify the class of the sample. Using the calculated distances, it is determined which class the new sample belongs to according to the number of k neighbors. During the classification process in the KNN method, the distance between the new data point and the existing data points is calculated. Subsequently, the classes of the k nearest neighbors, determined based on the calculated distances, are inspected. The KNN method employs a range of distance determination functions. Nevertheless, the lack of consideration for the feature's usefulness in addressing the classification assignment is a significant challenge. Hence, the equidistant contributions of all traits are considered in the selection of standard KNN. Different weighting methods are used to determine the contribution of neighbors to the class label according to their distance. When the literature is scanned, we can say that there are two types of weighting functions: inverse and squared-inverse.

Inverse: It is calculated by taking the inverse of the distance, as in Eq. 7.

$$w = 1/d \quad (7)$$

Squared-inverse: It is calculated by taking the inverse of the distance squared, as in Eq. 8.

$$w = 1/d^2 \quad (8)$$

III. EXPERIMENTAL RESULTS

The data sets included in the research were subjected to min-max normalization, resulting in the distribution of all data points within the range of 0 to 1. The data sets were randomly divided into 70 percent training and 30 percent test data. In order for the analysis of performances to be fair, the same training and test data were used in each different KNN model. The use of Euclidean distance was employed in the computation of the spatial separation between samples inside the KNN algorithms. The k value was taken as 5. Table 3 displays the confusion matrix for the given data sets.

Table 3. Confusion matrix of algorithms for data sets

	KNN(Equal)			KNN (Inverse)			KNN (Square-diverse)		
	Actual Class	Predicted Class 1	Predicted Class 0	Actual Class	Predicted Class 1	Predicted Class 0	Actual Class	Predicted Class 1	Predicted Class 0
Cervical cancer	1	5	0	1	5	0	1	5	0
	0	0	17	0	0	17	0	1	16
Cleveland	1	37	6	1	38	5	1	38	5
	0	4	44	0	4	44	0	4	44
Mammographic	1	95	20	1	86	29	1	85	30
	0	18	116	0	13	121	0	15	119
Spectfheart	1	60	9	1	60	9	1	60	9
	0	2	10	0	2	10	0	2	10
Pima	1	50	18	1	51	17	1	51	17
	0	27	136	0	26	137	0	26	137

Table 4. Experimental results

	Distance weighting function	Evaluation metrics				
		Accuracy	**Sensitivity**	**Specificty**	**Precision**	**F measure**
Cervical cancer	Equal	1.0000	1.0000	1.0000	1.0000	1.0000
	Inverse	1.0000	1.0000	1.0000	1.0000	1.0000
	Squared-inverse	0.9545	1.0000	0.9412	0.8333	0.9091
Cleveland	Equal	0.8901	0.8605	0.9167	0.9024	0.8810
	Inverse	0.9011	0.8837	0.9167	0.9048	0.8941
	Squared-inverse	0.9011	0.8837	0.9167	0.9048	0.8941
Mammographic	Equal	0.8474	0.8261	0.8657	0.8407	0.8333
	Inverse	0.8313	0.7478	0.9030	0.8687	0.8037
	Squared-inverse	0.8193	0.7391	0.8881	0.8500	0.7907
Spectfheart	Equal	0.8642	0.8696	0.8333	0.9677	0.9160
	Inverse	0.8642	0.8696	0.8333	0.9677	0.9160
	Squared-inverse	0.8642	0.8696	0.8333	0.9677	0.9160
Pima	Equal	0.8052	0.7353	0.8344	0.6494	0.6897
	Inverse	0.8139	0.7500	0.8405	0.6623	0.7034
	Squared-inverse	0.8139	0.7500	0.8405	0.6623	0.7034

Table 4 presents the derived values for accuracy, sensitivity, specificity, precision, and the F measure. When Table 4 is examined, while equal and inverse obtained the same results in the cervical cancer data set, squred-inverse obtained a worse result than these two methods. While inverse and squared-inverse obtained the same results in the cleveland and pima data sets, equal obtained a worse result than these two methods. While equal achieved the best result in the mammographic data set, the methods could not outperform each other in the spectfheart data set. It is seen that the inverse method gives relatively better results in five different data sets. It has been observed that the performance of the methods varies depending on the distribution of the data sets, the number of samples, and the number of features.

IV. CONCLUSION

In this study, the performance of distance weighting functions commonly used in KNN was examined. For this purpose, 3 different KNN versions were applied to 5 different medical data sets. Among the five data sets used in the study, it was observed that the inverse KNN approach demonstrated favorable results in four of them. Conversely, both the equal and squared-diverse methods exhibited successful outcomes in three of the data sets. In future work, the distance weighting functions of KNN can be examined in different data sets, depending on the number of samples and the number of features.

REFERENCES

[1] H. Yigit, “A weighting approach for KNN classifier”, in 2013 International Conference on Electronics, Computer and Computation (ICECCO), 7-9 Nov. 2013 2013, pp. 228-231.

[2] H. Zhang, K. Hou, and Z. Zhou, “A Weighted KNN Algorithm Based on Entropy Method”, in Intelligent Computing and Internet of Things, Pt Ii, vol. 924, (Communications in Computer and Information Science. Berlin: Springer-Verlag Berlin, 2018, pp. 443-451.

[3] N. Biswas, S. Chakraborty, S. S. Mullick, and S. Das, “A parameter independent fuzzy weighted k-Nearest neighbor classifier”, Pattern Recognition Letters, vol. 101, pp. 80-87, 2018.

[4] A. Juna, M. Umer, S. Sadiq, H. Karamti, A. A. Eshmawi, A. Mohamed, & I. Ashraf, “Water quality prediction using KNN imputer and multilayer perceptron”, Water, 14(17), 2592, 2022.

[5] O. Altay, “Performance of different KNN models in prediction english language readability”, In 2022 2nd International Conference on Computing and Machine Intelligence (ICMI) (pp. 1-5, IEEE, 2022.

[6] A. Almomany, W. R. Ayyad, A. Jarrah, “Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study”, Journal of King Saud University-Computer and Information Sciences, 34(6), 3815-3827, 2022.

[7] A. R. Lubis, S. Prayudani, M. Lubis, O. Nugroho, “Sentiment Analysis on Online Learning During the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method”, In 2022 1st International Conference on Information System & Information Technology (ICISIT) (pp. 106-111). IEEE, 2022.

[8] O. Altay, M. Ulas, K. E. Alyamac, “Prediction of the fresh performance of steel fiber reinforced self-compacting concrete using quadratic SVM and weighted KNN models”, IEEE Access, 8, 92647-92658, 2020.

[9] A. R. Yeruva, D. Kamboj, P. Shankar, U. S. Aswal, Aa. K. Rao, C. S. Somu, “E-mail Spam Detection Using Machine Learning–KNN”, In 2022 5th International Conference on Contemporary Computing and Informatics (IC3I) (pp. 1024-1028), IEEE, 2022.

[10] O. Altay, M. Ulas, M Ozer, & E. Genç, “An expert system to predict warfarin dosage in turkish patients depending on genetic and non-genetic factors”, In 2019 7th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6), IEEE, 2019.

[11] O. Altay, M. Ulas, “The use of kernel-based extreme learning machine and well-known classification algorithms for fall detection”, In Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2017, Volume 2 (pp. 147-155). Springer Singapore, 2019.

[12] R. S. Latha, G. R. Sreekanth, R. C. Suganthe, M. Geetha, R. E. Selvaraj, S. Balaji, ... &P. P. Ponnusamy, “Stock Movement Prediction using KNN Machine Learning Algorithm”, In 2022 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE, 2022.

[13] A. Frank and A. Asuncion (2010) UCI machine learning repository: data sets https://archive.ics.uci.edu/ Accessed 24 September 2023.

[14] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification”, IEEE Trans Inf Theory 13(1), 21–27, 1967.