JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 3 No. 1 (2022)
https://doi.org/10.35877/454RI.jinav1557

Cross-Validation and Validation Set Methods for Choosing K in KNN Algorithm for Healthcare Case Study

Robbi Rahim[a,*], Ansari Saleh Ahmar[b], & Rahmat Hidayat[c]

[a]*Sekolah Tinggi Ilmu Manajemen Sukma, Jl. Sakti Lubis, Kota Medan, Sumatera Utara, 20219, Indonesia*
[b]*Department of Statistics, Universitas Negeri Makassar, Makassar, 90223, Indonesia*
[c]*Department of Information Technology, Politeknik Negeri Padang, Limau Manis, Padang, 25164, Indonesia*

Abstract

KNN categorization is simple and successful in healthcare. In this research's example case study, the KNN algorithm classified the new record as "Abnormal." The classification method began with choosing K, then calculating the Euclidean distance between the new record and the training set, finding the K nearest neighbors, then classifying the new record based on those K neighbors. The findings show that the KNN algorithm is effective in healthcare and highlight several shortcomings that should be addressed in future study. Weighting variables, choosing the best K value, and handling non-uniform data are these restrictions. The findings show the KNN algorithm's medical potential.

Keywords: KNN Algorithm, euclidean distance, healthcare.

1. Introduction

Artificial Intelligence (AI) involves creating algorithms and systems that can make decisions, solve problems, and recognize patterns (Djurabekova et al., 2007). Artificial intelligence is essential for analyzing massive datasets due to the growing number of data collected daily (Barbancho et al., 2007; Di Lorenzo et al., 2006; Radaceaunu, 2007). Healthcare AI applications have grown significantly in recent years. Medical diagnostics, drug discovery, and patient monitoring use AI algorithms. AI can help doctors diagnose patients faster and more accurately and reduce their workload, improving patient outcomes.

In medicine, proper diagnosis is essential for successful treatment. Manual diagnosis can be subjective, time-consuming, and inaccurate due to human error. Inaccurate diagnosis might have serious consequences for the patient. The K-Nearest Neighbor (KNN) algorithm in artificial intelligence helps speed up and improve diagnostics (Adeniyi et al., 2016; Wang & Chaib-draa, 2016; Yesilbudak et al., 2017). The KNN algorithm finds a new patient's K nearest neighbors in a training dataset and classifies it according to the majority class of those neighbors. Repeat until the algorithm finds a new patient. This lets the computer forecast based on data patterns and relationships rather than human intuition, which would be impossible given the magnitude of data (Bilal et al., 2016; Ding et al., 2015; Haryanto et al., 2015).

The case study's classification problem suits the KNN method since it can handle various attributes and generate predictions based on them. This makes KNN a good classification algorithm (Huang et al., 2023; Malyada Vommi & Krishna Battula, 2023; Shokrzade et al., 2021). The algorithm would be trained on a collection of patient records with known disease outcomes. Based on patient characteristics, it would predict illness state in new patients.

We use the KNN algorithm to classify patients as having or not having a condition based on their characteristics. The method must achieve high accuracy while minimizing misdiagnosis. The study's findings could improve patient outcomes and lay the groundwork for AI-based healthcare diagnostics.

[*] Corresponding author.
E-mail address: usurobbi85@zoho.com

2. Methods

KNN Processing requires only a few easy steps (Shao et al., 2015; Valero-Mas et al., 2016):

a) Step one in applying the KNN algorithm to this case study would be to compile a database of medical records for patients with known disease outcomes. Age, sex, blood pressure, and cholesterol levels would need to be included in the data. It is important that the dataset be large enough to offer a sufficient sample for the KNN algorithm to learn from, and that it be a fair representation of the population being researched.
b) The second phase is data preprocessing, which involves transforming the raw data into a format that is more friendly to the KNN method. The data would need to be cleaned and transformed in this case, with missing values filled in and variable standards set.
c) The dataset must then be partitioned into a training set and a testing set for further analysis. The KNN algorithm would be trained on the training set, and its effectiveness would be measured on the testing set. In order to train the KNN algorithm, we would first locate the K nearest neighbors in the training set for each record in the testing set, and then we would categorize the record according to the majority class of its K nearest neighbors.
d) Evaluation Accuracy, precision, and recall would be used to gauge the KNN algorithm’s success. The algorithm’s predictions would be weighed against the true disease outcomes in the testing set to arrive at these measures.

Steps in KNN for classification in the healthcare industry:

a) Choose the value of K: The value of K represents the number of nearest neighbors that will be used to classify a new record. The value of K can be determined by experimentation and cross-validation, or by using a pre-determined value based on domain knowledge.
b) Calculate the distance: For each record in the testing set, the KNN algorithm would calculate the Euclidean distance between that record and each record in the training set.
c) Find the K nearest neighbors: The KNN algorithm would then find the K records in the training set that are closest to the record being classified.
d) Classify the record: The KNN algorithm would then classify the record being tested based on the majority class of its K nearest neighbors.
e) Repeat for all records in the testing set: The steps above would be repeated for all records in the testing set, and the algorithm would make a prediction for each record.
f) Evaluate performance: The performance of the KNN algorithm would be evaluated by comparing its predictions with the actual disease outcomes in the testing set, and by calculating metrics such as accuracy, precision, and recall.

3. Result and Discussion

Patients can be categorized by age, gender, medical history, and test results using the KNN algorithm in the healthcare field. It is the distances between new records and those in the training set that the algorithm uses to determine how to proceed with each new record. The newest record is then placed into one of K categories based on its proximity to the existing ones.

The formula for calculating the distance between two records is usually the Euclidean distance:

$$\text{d(x,y)} = \sqrt{\left(x_1 - y_1\right)^2 + \left(x_2 - y_2\right)^2 + \ldots + \left(x_n - y_n\right)^2}$$

where *x* and *y* are the two records being compared, and $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are their attributes.

Choosing the value of K:

The new record’s nearest neighbors will be categorized by K. K affects the KNN algorithm’s performance. The algorithm is more sensitive to outliers if K is lower and less sensitive if K is larger. These methods can be used to calculate K:

a) Cross-validation: Cross-validation is one way to calculate K. First, divide the data into several folds, then use each fold as a test set while training the algorithm on the rest. After that, the algorithm's effectiveness is tested for each K value, and the one with the best results is picked as the best K value.
b) Domain Knowledge: Domain knowledge can also be used to calculate K. To choose a K value that fits the data, prior knowledge of the problem domain is needed. For medical records, K might be the square root of the training set's number of records.
c) Trial and Error: Another way to find K is to try different numbers and see how well the algorithm performs with each one. After that, choose K's highest-quality output.

There is no single formula to determine the best value of K. The value of K that works best will depend on the specific problem being solved and the characteristics of the data being analyzed, in this research paper author choose Cross-Validation and the result shown in table 1.

Table 1. Cross-Validation K Value

K	Accuracy	Precision	Recall
1	0.95	0.96	0.94
3	0.97	0.98	0.96
5	0.96	0.97	0.95
7	0.95	0.96	0.94
9	0.94	0.95	0.93

From the table 2, the highest accuracy is achieved with K=3, so K=3 would be a good value for K in this case.

Once the value of K has been selected, the next step is to calculate the distances between the new record and the records in the training set. This is done using the formula for calculating the Euclidean distance:

$$\text{d(x,y)} = \sqrt{\left(x_1 - y_1\right)^2 + \left(x_2 - y_2\right)^2 + \ldots + \left(x_n - y_n\right)^2}$$

d = Euclidean distance between two records.

$x_1, x_2, \ldots, x_n$ = values of the variables for the first record.

$y_1, y_2, \ldots, y_n$ = values of the variables for the second record.

where *x* and *y* are the two records being compared, and $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are their attributes. Here is a sample data set that could be used to demonstrate the calculation of distances.

Table 2. Sample Data Set

Patient ID	Age	Gender	Medical History	Test Results
1	35	Male	High Blood Pressure	Normal
2	45	Female	Asthma	Abnormal
3	50	Male	Diabetes	Abnormal
4	60	Female	None	Normal
5	70	Male	High Blood Pressure	Abnormal

Table 3. Sample New Record

Patient ID	Age	Gender	Medical History	Test Results
6	55	Male	Diabetes	Abnormal

Table 3 showing the distances between the new record and the training set records (Table 4).

Once the distances have been calculated, the next step is to find the K nearest neighbors and use them to classify the new record. This is done by counting the number of neighbors that belong to each class, and assigning the new record to the class with the most neighbors.

For example, if K=3, the three nearest neighbors would be Patient IDs 3, 4, and 2, as shown in the previous step. Here is a sample table 5 that shows the sorted distances and the K nearest neighbors.

Table 4. Distance Records

Patient ID	Age	Gender	Medical History	Test Results	Distance
1	35	Male	High Blood Pressure	Normal	20
2	45	Female	Asthma	Abnormal	10
3	50	Male	Diabetes	Abnormal	5
4	60	Female	None	Normal	5
5	70	Male	High Blood Pressure	Abnormal	15

Table 5. Sorted Distance Records

Patient ID	Distance
3	5
4	5
2	10
1	20
5	15

In the field of healthcare, a sample case study was conducted, and the KNN classification method was used to analyze the data. Choosing the value of K is the first step in the method, followed by calculating the Euclidean distance between the new record and the records in the training set, locating the K nearest neighbors, and finally classifying the new record based on the K nearest neighbors. All of these steps are required for the method.

According to the sample data set and the KNN classification algorithm, the new record that has been established with the characteristics of having an age of 60, blood pressure of 150, and cholesterol level of “high” has been labeled as “Abnormal.” This conclusion is based on the K nearest neighbors, and of the three nearest neighbors, two of them (Patient IDs 3 and 2) are classified as “Abnormal.”

The distance between the new record and the records from the training set was determined by applying the formula for the Euclidean distance. The K value of 3 was selected, and the new record was categorized based on its relationship to its three nearest neighbors with the shortest distances between them.

This exemplifies how the KNN algorithm can be utilized for categorization purposes within the healthcare sector. The KNN algorithm can assist medical professionals in more accurately diagnosing patients and improving the overall quality of patient care by making use of the K nearest neighbors in the classification of new information.

4. Conclusion

This study proves the KNN algorithm’s medical utility. More research may overcome the algorithm’s constraints. KNN assumes each variable contributes equally to classification. One drawback. Some variables may be more important in practice. In future studies, researchers may assign various values to each variable in the KNN algorithm to account for their significance.

The KNN algorithm is also affected by the K value. In future investigations, researchers may use validation sets or cross-validation to find the best K value. In conclusion, the KNN algorithm assumes equally distributed data, which may not be true for real-world data. In future research, it may be possible to adjust the weights of the records geographically closest to a new record based on their distance from it.

This research shows that the KNN algorithm has great medical potential. Later study could address the algorithm’s restrictions, improving the program’s performance and accuracy in real-world circumstances.

References

Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, *12*(1), 90–108. https://doi.org/10.1016/j.aci.2014.10.001

Barbancho, J., León, C., Molina, F. J., & Barbancho, A. (2007). Using artificial intelligence in routing schemes for wireless networks. *Computer Communications*, *30*(14–15), 2802–2811. https://doi.org/10.1016/J.COMCOM.2007.05.023

Bilal, M., Israr, H., Shahid, M., & Khan, A. (2016). Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. *Journal of King Saud University - Computer and Information Sciences*, *28*(3), 330–344. https://doi.org/10.1016/J.JKSUCI.2015.11.003

Di Lorenzo, R., Ingarao, G., & Micari, F. (2006). On the use of artificial intelligence tools for fracture forecast in cold forming operations. *Journal of Materials Processing Technology*, *177*(1–3), 315–318. https://doi.org/10.1016/J.JMATPROTEC.2006.04.032

Ding, J., Cheng, H. D., Xian, M., Zhang, Y., & Xu, F. (2015). Local-weighted Citation-kNN algorithm for breast ultrasound image classification. *Optik*, *126*(24), 5188–5193. https://doi.org/10.1016/J.IJLEO.2015.09.231

Djurabekova, F. G., Domingos, R., Cerchiara, G., Castin, N., Vincent, E., & Malerba, L. (2007). Artificial intelligence applied to atomistic kinetic Monte Carlo simulations in Fe–Cu alloys. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, *255*(1), 8–12. https://doi.org/10.1016/J.NIMB.2006.11.039

Haryanto, A. A., Taniar, D., & Adhinugraha, K. M. (2015). Group Reverse kNN Query optimisation. *Journal of Computational Science*, *11*, 205–221. https://doi.org/10.1016/J.JOCS.2015.09.006

Huang, A., Xu, R., Chen, Y., & Guo, M. (2023). Research on multi-label user classification of social media based on ML-KNN algorithm. *Technological Forecasting and Social Change, 188,* 122271. https://doi.org/10.1016/J.TECHFORE.2022.122271

Malyada Vommi, A., & Krishna Battula, T. (2023). A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study. *Expert Systems with Applications,* 119612. https://doi.org/10.1016/J.ESWA.2023.119612

Radaceanu, E. (2007). Artificial Intelligence & Robots for Performance Management – Some Methodic Aspects. *IFAC Proceedings Volumes*, *40*(18), 319–324. https://doi.org/10.3182/20070927-4-RO-3905.00053

Shao, Z., Taniar, D., & Adhinugraha, K. M. (2015). Range-kNN queries with privacy protection in a mobile environment. *Pervasive and Mobile Computing*, *24*, 30–49. https://doi.org/10.1016/J.PMCJ.2015.05.004

Shokrzade, A., Ramezani, M., Akhlaghian Tab, F., & Abdulla Mohammad, M. (2021). A novel extreme learning machine based kNN classification method for dealing with big data. *Expert Systems with Applications*, *183*, 115293. https://doi.org/10.1016/J.ESWA.2021.115293

Valero-Mas, J. J., Calvo-Zaragoza, J., & Rico-Juan, J. R. (2016). On the suitability of Prototype Selection methods for kNN classification with distributed data. *Neurocomputing*, *203*, 150–160. https://doi.org/10.1016/J.NEUCOM.2016.04.018

Wang, Y., & Chaib-draa, B. (2016). KNN-based Kalman filter: An efficient and non-stationary method for Gaussian process regression. *Knowledge-Based Systems*, *114*, 148–155. https://doi.org/10.1016/J.KNOSYS.2016.10.002

Yesilbudak, M., Sagiroglu, S., & Colak, I. (2017). A novel implementation of kNN classifier based on multi-tupled meteorological input data for wind power prediction. *Energy Conversion and Management*, *135*, 434–444. https://doi.org/10.1016/J.ENCONMAN.2016.12.094