INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 478-480

ASSESS THE RELATIVE EFFICACY OF THE K-NEAREST NEIGHBOR (KNN) ALGORITHM AND ITS VARIOUS ADAPTATIONS IN PREDICTING DISEASES THROUGH A COMPARATIVE ANALYSIS

P Sneka[1]

[1]Dept. of CS, Fatima College, India.

ABSTRACT

Predicting disease risk poses a growing challenge in the field of medicine, prompting researchers to extensively employ machine learning algorithms. Among the array of machine learning algorithms, the k-nearest neighbor (KNN) algorithm is particularly prevalent. This paper conducts a comprehensive study on various KNN variants, including the Classic one, Adaptive, Locally adaptive, k-means clustering, Fuzzy, Mutual, Ensemble, Hassanat, and Generalized mean distance, comparing their performance in disease prediction. Performance measures such as accuracy, precision, and recall are considered for the comparative analysis. The average accuracy values across these variants range from 64.22% to 83.62%. The Hassanaat KNN exhibits the highest average accuracy (83.62%), closely followed by the ensemble approach KNN (82.34%).The study extends its analysis to precision and recall measures, offering a relative comparison among KNN variants. Ultimately, the paper summarizes the most promising KNN variant based on three performance measures (accuracy, precision, and recall) for disease prediction. Healthcare researchers and stakeholders can leverage the insights from this study to make informed decisions regarding the selection of the most suitable KNN variant for predictive disease risk analytics.

1. INTRODUCTION

The k-nearest neighbor (KNN) algorithm is a supervised machine learning approach primarily employed for classification purposes, particularly in the context of disease prediction. Functioning as a supervised algorithm, KNN predicts the classification of unlabeled data by considering the features and labels of the training data. Essentially, the algorithm classifies datasets by utilizing a training model similar to the testing query, taking into account the k nearest training data points (neighbors) that are closest to the query being tested. Subsequently, the algorithm employs a majority voting rule to determine the final classification.Among the myriad machine learning algorithms. Consequently, this paper undertakes a study focused on the application of the KNN algorithm to classify medical datasets, given the real-world challenge of predicting diseases. The objective is to explore how the algorithm can adapt and contribute to addressing this critical issue. The algorithm operates with simplicity in its processes and computations, providing avenues for modification across different dimensions to mitigate limitations and challenges. The classic KNN algorithm faces several limitations that dampen its classification capabilities, including an inherent bias towards all classification-dependent neighbors, a lack of features for distance calculations between data points, and the consideration of unnecessary dataset features.However, owing to KNN's flexibility for numerous modifications, it gives rise to various forms or variants. These KNN variants vary in algorithmic aspects, such as optimizing the k parameter, refining distance calculations, introducing weighting for different data points.

K-nearest neighbour algorithm and its different variants:

The Classic KNN Algorithm:

The conventional KNN algorithm is a supervised machine learning approach primarily employed for classification tasks. This algorithm involves a variable parameter, denoted as k, representing the number of 'nearest neighbors.' The operation of the KNN algorithm entails identifying the nearest data point(s) or neighbor(s) from a training dataset in relation to a given query. Proximity is determined based on the closest distances from the query point. Once the k nearest data points are identified, the algorithm applies a majority voting rule to determine the class that appears most frequently. The class with the highest occurrence is designated as the final classification for the query.

With a value of k set to 3 for Query B, the algorithm searches for the three nearest neighbors and discovers that among these, two belong to class 1, while one belongs to class 0. Applying the majority voting rule, it classifies Query B as belonging to class 1. Likewise, for Query A with a value of k set to 5, it identifies a greater number of neighbors characterized as Class 0, leading to the classification of Query A as belonging to class 0.

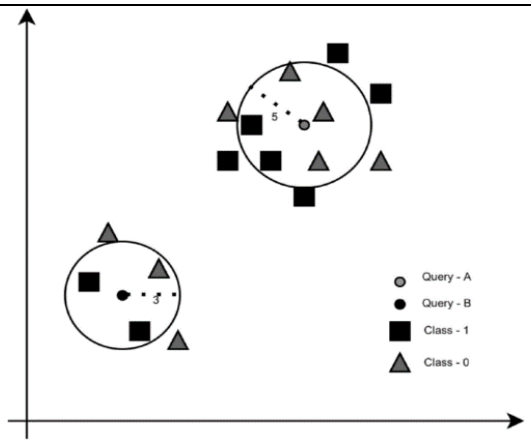
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 478-480



Locally adaptive KNN with Discrimination class (LA-KNN):

This specific variant incorporates information from discrimination classes to determine the optimal k value. Discrimination classes involve assessing the quantity and distribution of neighbors from both the majority class and the second majority class within the k-neighbourhood of a given testing data point. The algorithm follows a series of steps to define discrimination classes. After selecting a class, it proceeds to generate a ranking table that encompasses different k values, distances from centroids, and their ratios. Employing this table, the algorithm engages in a ranking process to identify and output the optimal k value.

Fuzzy KNN:

The fuzzy KNN algorithm is centered on the concept of membership assignment15. Similar to the classic KNN algorithm, this variant begins by identifying the k nearest neighbors of a testing dataset from the training dataset. Subsequently, it assigns "membership" values to each class present in the list of k nearest neighbors. The calculation of membership values is carried out using a fuzzy math algorithm that emphasizes the weight of each class. The class with the highest membership is then chosen as the classification result.

Weight adjusted KNN (W-KNN):

This iteration of the KNN algorithm emphasizes the utilization of attribute weighting. Initially, the algorithm assigns a weight to each training data point employing a function called the kernel function12. The purpose of this weight assignment is to attribute greater weight to closer points and lesser weight to more distant points. Any function that diminishes in value with increasing distance can serve as the kernel function. Subsequently, the frequency of all nearest neighbors is employed to predict the output class for a given testing data point. Notably, this KNN variant takes into account the classification significance of different attributes in formulating the kernel function for a multiattribute dataset.

Mutual KNN (M-KNN):

The mutual KNN algorithm is based on the notion of mutual neighbors 14. Initially, the method modifies the training dataset by removing sets that do not have k closest neighbors with other sets. This produces a reduced training dataset with fewer occurrences of noise and abnormalities. The method then uses the testing dataset to identify the k nearest neighbors from the training dataset, followed by determining the k nearest neighbors of the nearest neighbors in the testing dataset. This step allows the algorithm to find reciprocal nearest neighbors, which are then classified. The test datasets are classified using the majority voting rule.

2. METHODS

The research focuses largely on the medical domain, with secondary areas chosen at random to eliminate bias. Table 1 summarizes the datasets used in this work, along with their respective parameters such as the number of features, data size, and domain origin. The datasets were obtained from Kaggle19, the UCI Machine Learning Repository20, and OpenML21. They display a wide range of characteristics in terms of features, traits, and sizes, with a strong emphasis on the medical sector to assure usefulness in disease risk prediction.

Confusion matrix: This generates a matrix as output, providing a comprehensive overview of the model's performance. The evaluation of the analyses involves the use of four essential metrics: True Positive Ratio (TPR), True Negative Ratio (TNR), False Positive Ratio (FPR), and False Negative Ratio (FNR) rates. Specifically, these metrics are computed individually. True positive, true negative, false positive, and false negative represent the quantities associated with correctly predicted positive, correctly predicted negative, incorrectly predicted positive, and incorrectly predicted negative instances, respectively.

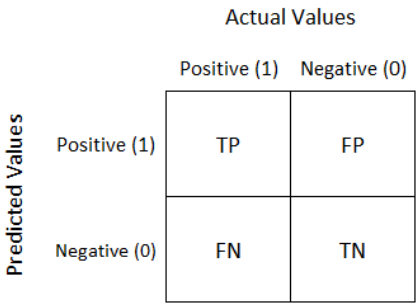
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 478-480



Accuracy is the ratio of the sum of true positive and true negative to the sum of all the predicted samples

Accuracy = TP+TN

TP+TN+FP+FN

Sensitivity which is also called recall is the measure of the ratio of true positive predictions to the sum of true positive and false negative.

Sensitivity (recall) = TP

TP+FN

Specificity is the measure of the ration of true negative to the sum of true negative and false positive

Specificity = TN

TN+FP

Precision is the ratio of the number of true positives to the sum of true positive and false positive. It can be said to be the measure of the quality of the positive feedback data.

Precision = TP

TP+FP

Relative performance index (RPI):

The relative performance index is a ground-breaking evaluator that combines data findings from numerous metrics (such as accuracy, precision, and recall) to produce a probabilistic conclusion for the final assessment. The unique performance metric developed in this work is modeled after another RPI (Relative Performance Index) measure established by Nagle29. Nagle's initial idea included a probabilistic computation aimed to reduce bias by considering the range of outcomes within a single field and evaluating how frequently those results outperformed those of other fields. The RPI for a given field is then determined using the retrieved values and the total number of existing fields.

3. CONCLUSION

In summary, Hassanat, the ensemble approach, and the generalized mean distance emerge as the most suitable KNN variants for disease prediction, demonstrating high accuracy, precision, and recall measures, respectively. These variants effectively address various limitations of the classic KNN and outperform others in overall performance. Among the top three performers, the ensemble approach KNN stands out, achieving the highest precision and performing well in both accuracy and recall measurements. Its distinctive design for overcoming multiple limitations positions it as the prime variant among the others, particularly excelling in disease risk prediction within the medical domain. Across the research datasets, most variants prove their effectiveness in achieving high-performance measures, showcasing adaptability and the capability to address general constraints prevalent in the medical domain

4. REFERENCES

[1] Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. Nat. Methods **15**, 5–6 (2018).

[2] Ogwueleka, F. N., (2011). Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science and Technology.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011).Data mining for credit card fraud: A comparative study. DecisionSupport Systems.

[4] Shen, A., Tong, R., & De ng, Y. (2007). Application of classificationmodels on credit card fraud detection. In Service Systems and ServiceManagement, 2007 .

[5] Data Analytics vs Data Science: Two Separate, but InterconnectedDisciplines, Data Scientist Insights, 28-Apr-2018.

[6] Lamba, A. & Kumar, D. Survey on KNN and its variants. Int. J. Adv. Res. Comput. Commun. Eng. **5**, 430–435 (2016).