IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11706

A similarity measure to find Nearest Neighbours for heart disease to improve prediction accuracy

Mathura Bai B.[1], Mangathayaru N.[2], Padmaja Rani B.[3]

Associate Professor, Department of Information Technology, VNR VJIET, Hyderabad, India[1]

Professor, Department of Information Technology, VNR VJIET, Hyderabad, India[2]

Professor, Department of CSE, JNTUH, Hyderabad, India[3]

Abstract: k Nearest Neighbour Classifier (kNN) is a widely used non parametric machine learning model. This classifier can model complex data distributions and can achieve generalization. kNN algorithm coherently groups data into subsets and labels the test instance based on the similar or nearest training instances. The optimal selection of the nearest neighbours has to be done for accurate classification. Our work implements kNN algorithm with a similarity measure in identifying the optimal nearest neighbours for test instances and deciding their class label as the majority class label among the nearest neighbours. The proposed similarity measure considers the data distribution and thus helps in selecting the optimal nearest neighbours. The effectiveness of the proposed work is evaluated on several datasets with different classifiers like J48, Naive Bayes (NB) classifier. The proposed method outperforms in comparison with other ensemble learning techniques like Multilayer Perceptron (MLP) and Random Forest (RF) with high classification accuracy.

Keywords: similarity measure, nearest neighbour, k fold cross validation, classification, accuracy.

I. INTRODUCTION

Data Mining achieved success in handling and analyzing simple and small size datasets. But as the complexity of the data increased a number of problems were encountered by the researchers to handle huge volumes of data from different domains like business, health sector, banking sector, social networks, e-commerce etc. These challenges and issues are summarized by authors Yang & Wu, 2006, Hirji, 2001. The analysis of such complex and high dimensional data can be done efficiently using machine learning algorithms. Machine learning algorithms are one among the Computational Intelligence techniques widely used for solving real world problems as discussed by Mahesh, 2020. The data mining and machine learning algorithms when applied on real world datasets face a number of challenges and issues as discussed by the authors in Hirji, 2001, Yang & Wu, 2006, Bai et.al, 2015. The researchers have identified the various challenges such as scaling of high dimensional dataset, handling different types of data like unstructured, imbalance dataset, handling missing values.

The k Nearest Neighbours (kNN) is considered as a simple and flexible machine learning model popularly used in various domains for prediction and decision making. The kNN algorithm is very efficient and effective in determining risks of diseases like diabetes and heart attacks as discussed by Taunk et.al, 2019. The kNN algorithm further needs extensions to improve the learning ability and prediction accuracy. The model needs to adjust the hyper parameters like k value, distance or similarity measures to improve the prediction accuracy as indicated by the author in Zhang, 2016. The commonly used distance measures in kNN are Euclidean and Manhattan distance. As suggested by the author in Cunningham & Delany, 2021, the efficiency of kNN is based on the metrics learned from the data. The metrics should be data specific for better performance of the model like Euclidean distance for real data and minkowski distance and cosine similarity for vectorized data. The metrics are decided based on the dataset type. Hence, there is a need for a distance or similarity measure which can handle any type of data. The measure is independent of data type.

Main objective of the current work is to propose a nearest neighbour classifier (kNN) for better prediction. It can be considered as a modified kNN using a proposed similarity measure to identify k nearest neighbours. The paper organized with the Background and Motivation section details the survey on research contributions which can be considered as a basis for motivation. The Proposed Method section describes the methodology in detail. Experimentation & Results section figures out the results obtained using existing and proposed methods. Conclusions & Future Scope section concludes the important findings and briefs the future scope of the proposed method.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022
DOI: 10.17148/IJARCCE.2022.11706

II. BACKGROUND AND MOTIVATION

kNN is non-parametric as it does not assume any data distribution. The popularity of kNN is because of its easy implementation, flexibility, and robustness. The error rate of kNN is twice that of Bayes. kNN is widely used in medical diagnosis, image detection, fraud analysis, video recognition, stock market and weather forecasting, protein function prediction, image recognition, natural language processing etc. Even though kNN is popular it suffers from limitations like high computational cost, choice of distance or similarity measure, size of the dataset, sensitivity to irrelevant and noisy features Ray, 2019. kNN is transparent and can be extendable to any data by providing a suitable distance or similarity measure which is the objective of our work in this paper. kNN can be made effective by properly analysing the nearest neighbours. The similarity measure proposed is motivated from several research contributions which are widely used in medical data classification, pattern recognition etc.

Jiang et.al, 2010 proposed fuzzy feature Gaussian similarity function in feature clustering. A Gaussian text similarity function was designed by the authors in Jiang et.al, 2011 for text classification. Lin et.al, 2013 extended the text similarity function to clustering problems. Authors in Bai et.al, 2015 and Aljawarneh et.al, 2018 proposed MANTRA distance function used as an imputation measure to find similarity between instances of medical datasets. These research contributions motivated the proposed method. There is a scope for designing a similarity measure based on feature similarity to achieve better classification accuracies. The current work motivated by authors in Lin et.al, 2013, Bai et.al, 2021 by considering the imputation measure and proposing a similarity measure for identifying the nearest neighbors based on the feature distribution. The next section describes the nearest neighbour classifier with a proposed similarity measure as part of the learning process whose basic idea is to improve the classification accuracy and prediction rates.

III.PROPOSED METHOD

The authors in Dhanabai & Chandramathi, 2011 explored k nearest neighbor techniques for structures and unstructured data. The insights explored by them are more execution time for structured algorithms and limited memory allocation for both structured and unstructured algorithms. The biased k value has an impact on the performance of the kNN classifier. As outlined by the researchers in Dhanabai & Chandramathi, 2011 in order to reduce both time and computational complexity different kNN algorithms were proposed by researchers in the literature. This motivated the current work. The distance or similarity measure used by kNN is linear in nature, that is as the number of features increases the computational complexity also increases. The distance or similarity measure is heavily dependent on the number of features and the range of feature values. Such distance or similarity measure misleads in identifying the nearest neighbors as feature distribution is not considered. The need of such a measure which considers the feature distribution is proposed in our work. The similarity measure is based on the gaussian similarity function by Lin et.al, 2013 which measures the degree of similarity of the instances from each other in terms of the feature similarity. This improves the selection of most appropriate similar instances and is not sensitive to the number of features in the instance.

Algorithm: Proposed Nearest Neighbour Classifier

Input: Complete Dataset with no missing values. If missing values exists need to perform data pre-processing and impute the missing values

Output: Classification accuracy

Start

1. Consider the dataset has ‘m’ instances or observations, ‘n’ features and ‘l’ class labels. The proposed method reads the dataset in matrix form
2. Apply data pre-processing techniques like data imputation for handling missing values, data discretization for categorical data if required
3. Consider k-fold cross validation for dataset
    a. for each fold perform
        i. apply nearest neighbour proposed classifier with the proposed distance measure to identify the k nearest neighbours using Equation (1)

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11706

$$sim(x_i, x_j) = 0.5 + 0.5 * \frac{\sum_{k=1}^{n} e^{-\left(\frac{x_{ik} - x_{jk}}{\sigma_k}\right)^2}}{n} \quad (1)$$

where x_i and x_j are two instances for which similarity is to be calculated

x_{ik} and x_{jk} are corresponding k^{th} attribute of x_i and x_j instances

σ_k is the standard deviation of k^{th} attribute

n is the number of features in the dataset

ii. predict the class label

iii. calculate the prediction accuracy

b. calculate the mean accuracy for k folds

4. Apply other machine learning classification algorithms like J48, Naive Bayes (NB) on the matrix form of dataset to find the prediction accuracy
5. Evaluate the classification parameter accuracy for machine learning algorithms along with the proposed method

Stop

kNN is a lazy learning algorithm where learning is done at runtime and the performance of the training reduces as the size of the training dataset increases which provides a scope to use any dimensionality reduction techniques. kNN uses exhaustive brute force search strategy which can be altered with any other optimistic search strategy to speed up the searching process in identifying the k nearest neighbours.

IV. EXPERIMENTATION & RESULTS

The proposed method along with J48 and NB classifier is experimented on datasets like iris, wdbc, heart, cleveland, australian, german, yeast, glass, wine, pima, new thyroid. All these complete datasets are briefly described in TABLE I. and are used from the keel repository.

TABLE I DATASET DESCRIPTION

Datasets	No. of observations	No. of features	Classes
Iris	150	4 Real	3
Wdbc	569	30 Real	2
Heart	270	13 Numeric	2
Cleveland	303 (1.98% MVs)	13 Real	5
Australian	690	14 Real	2
German	1000	20 Numeric	2
Yeast	1484	8 Real	10
Glass	214	9 Real	7
Wine	178	13 Real	3
Pima	768	8 Real	2
New Thyroid	215	5 Real	3

The proposed classifier method has outperformed J48 and Naive Bayes algorithms by improving the accuracy as in Latha & Jeeva, 2019. The classifier results are clearly explained in TABLE II. The proposed classifier method has improved accuracy for datasets like iris by 1%, wdbc by 3%, heart by 1%, Cleveland by 4%, Australian by 2%, german by 1% approximately and minor increase in the other datasets. The accuracies are visualized in Fig. 1.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11706

TABLE III ACCURACIES OF DIFFERENT DATASETS FOR J48, NB AND PROPOSED METHOD

Datasets	Classification Algorithms		
	Proposed Method	J48	Naive Bayes
Iris	95.333	94.666	94
Wdbc	97.014	94.024	92.970
Heart	84.444	77.407	83.703
Cleveland	67.632	56.565	63.299
Australian	86.377	84.347	76.956
German	74.2	70.3	73.2
Yeast	56.803	56.671	57.884
Glass	64.762	69.158	48.598
Wine	94.967	93.820	97.752
Pima	74.735	74.349	75.520
New Thyroid	93.377	92.093	96.744

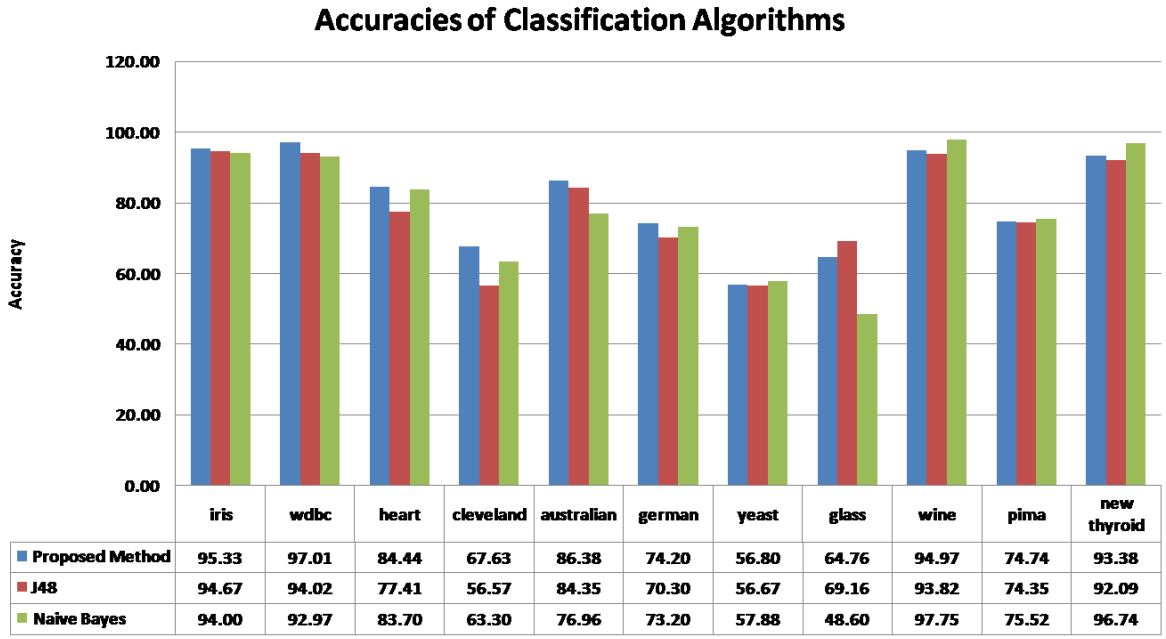


Fig. 1 Accuracies of different datasets for J48, NB and proposed method

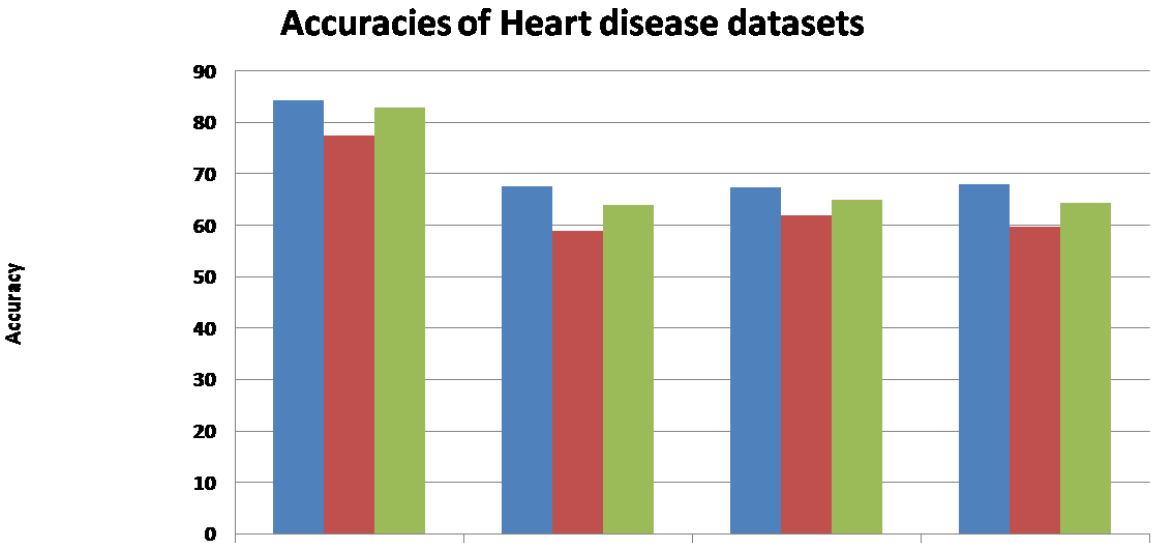
Authors in Latha & Jeeva, 2019 analysed performance accuracies of heart dataset using ensemble learning techniques like bagging and boosting as 80.53% for Random Forest and 81.52% for Multilayer Perceptron with feature selection. The proposed method for heart disease has achieved 84.44% accuracy which is higher when compared to Multilayer Perceptron (MLP) having 77.40% and Random Forest (RF) having 82.96% without feature selection as specified in TABLE III.

TABLE IIIII ACCURACIES OF HEART DISEASE DATASETS FOR MLP, RF AND PROPOSED METHOD

Datasets	Machine Learning Algorithms		
	Proposed Method	MLP	RF
Heart	84.444	77.407	82.963
Cleveland with ignore	67.632	58.922	63.973
Cleveland with mean imputation	67.312	62.046	65.016
Cleveland with MBI imputation	67.935	59.736	64.356

DOI: 10.17148/IJARCCE.2022.11706

The proposed classifier when applied for Cleveland datasets with missing values are handled with different imputation methods. Different imputation methods used for Cleveland dataset are ignoring the missing values, with mean imputation and MBI imputation method suggested by the author in Bai et.al, 2021. The results of the Cleveland dataset with all imputation methods outperformed the proposed method when compared to traditional machine learning algorithms like J48, NB algorithm and ensemble learning methods like Multilayer Perceptron (MLP) and Random Forest (RF). The proposed method has improved accuracy of heart dataset by 2%, Cleveland dataset by 4% by ignoring missing values, Cleveland dataset by 2% by performing mean imputation, Cleveland dataset by 3% by applying imputation method proposed in []. The accuracies of heart disease datasets for Multilayer Perceptron, Random Forest and the proposed method can be visualized in Fig. 2.



	heart	cleveland with ignore	cleveland with mean imputation	cleveland with MBI imputation
Proposed Method	84.444	67.632	67.312	67.935
Multilayer Perceptron	77.407	58.922	62.046	59.736
Random Forest	82.963	63.973	65.016	64.356

Fig. 2 Accuracies of heart disease datasets for MLP, RF and proposed method

V. Conclusions and Future Scope

The proposed method where nearest neighbors identified using similarity measure has outperformed the ensemble techniques like Random Forest and Multilayer Perceptron on heart disease and Cleveland data. The proposed method in comparison with J48, NB has also achieved high accuracy for complete datasets like iris, wdbc, australian, german, yeast, glass, wine, pima and new thyroid. The observed point is that the proposed method when applied to heart disease datasets has achieved higher accuracy than ensembling methods. kNN techniques are not suited for a multidimensional environment because of the large volume of data involved in it which can be the future scope of our work

References

[1]. Z. Zhang, “Introduction to machine learning: k-nearest neighbors,” Annals of the translational medicine, vol. 4, no. 11, Jun. 2016.
[2]. B. Mahesh, “Machine learning algorithms-a review,” International Journal of Science and Research (IJSR), [Internet], vol. 9, pp. 381-386, Oct. 2020.
[3]. K. K. Hirji, “Exploring data mining implementation,” Communications of the ACM, vol. 44, no. 7, Jul. 2001.
[4]. J. Kaiser, “Dealing with Missing Values in Data,” Journal of Systems Integration (1804-2724), vol. 5, no. 1, Jan. 2014.
[5]. Bai, B. Mathura, N. Mangathayaru, and B. Padmaja Rani, “Exploring research issues in mining medical datasets,” in Proc. ICEMIS’15, 2015, p. 1-8.
[6]. Taunk, Kashvi, Sanjukta De, Srishti Verma, and Aleena Swetapadma, “A brief review of nearest neighbor algorithm for learning and classification,” in Proc. ICCS, IEEE, 2019, p. 1255-1260.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022
DOI: 10.17148/IJARCCE.2022.11706

[7]. Yang, Qiang, and Xindong Wu, "10 challenging problems in data mining research," International Journal of Information Technology & Decision Making, vol. 5, no. 4, pp. 597-604, Dec. 2006.
[8]. Cunningham, Padraig, and Sarah Jane Delany, "K-nearest neighbour classifiers-a tutorial," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1-25, Jul. 2021.
[9]. Ray, Susmita, "A quick review of machine learning algorithms," in 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), p. 35-39, IEEE, Feb. 2019.
[10]. Jiang, Jung-Yi, Ren-Jia Liou, and Shie-Jue Lee, "A fuzzy self-constructing feature clustering algorithm for text classification," IEEE transactions on knowledge and data engineering, vol. 23, no. 3, pp. 335-349, Jul. 2010.
[11]. Jiang, Jung-Yi, Wen-Hao Cheng, Yu-Shu Chiou, and Shie-Jue Lee, "A similarity measure for text processing," in 2011 International Conference on Machine Learning and Cybernetics, vol. 4, pp. 1460-1465, IEEE, 2011.
[12]. Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee, "A similarity measure for text classification and clustering," IEEE transactions on knowledge and data engineering, vol. 26, no. 7, pp. 1575-1590, 2013.
[13]. Bai, B. Mathura, Nimmala Mangathayaru, and B. Padmaja Rani, "An approach to find missing values in medical datasets," in Proc. of the The International Conference on Engineering & MIS 2015, pp. 1-7, 2015.
[14]. Aljawarneh, Shadi, Vangipuram Radhakrishna, and Gali Suresh Reddy, "Mantra: a novel imputation measure for disease classification and prediction," in Proc. of the first international conference on data science, E-learning and information systems, pp. 1-5, 2018.
[15]. Bai, B. Mathura, N. Mangathayaru, B.Padmaja Rani, and Shadi Aljawarneh, "Mathura (MBI)-A novel imputation measure for imputation of missing values in medical datasets," Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), vol. 14, no. 5, pp. 1358-1369, 2021.
[16]. Dhanabal, S., and S. J. I. J. C. A. Chandramathi, "A review of various k-nearest neighbor query processing techniques," International Journal of Computer Applications, vol. 31, no. 7, pp. 14-22, 2011.
[17]. Latha, C. Beulah Christalin, and S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked 16, 2019: 100203.

