Journal of Advanced Zoology

ISSN: 0253-7214

Volume ***45*** *Issue* ***-3 Year 2024*** *Page* ***923-928***

Distance Measures Insights For Breast Cancer Analysis Using K-NN Algorithm

Dr. S. Bharathi[1*], Krithika.L[2*]

[1]Assistant Professor, Department of Mathematics, Bharathiar University PG Extension and Research Center, Perundurai, Erode, Tamilnadu, India. Email:bharathikamesh6@gmail.com
*[2]*Research Scholar, Department of Mathematics, Bharathiar University PG Extension and Research Center, Perundurai, Erode, Tamilnadu, India. Email: keerthilakshminarayanan@gmail.com*

***Corresponding Author:** Krithika.L*
**Research Scholar, Department of Mathematics, Bharathiar University PG Extension and Research Center, Perundurai, Erode, Tamilnadu, India. Email: keerthilakshminarayanan@gmail.com*

ABSTRACT:

Breast cancer stands as a prevalent disease among women, ranking high in terms of frequency. It is treatable if caught early enough. The greatest technique for predicting breast cancer is what this paper seeks to deliver. Mammograms can detect abnormal growths, although they are not always 100% accurate in identifying breast cancer. This article provides a superior way of prediction without biopsy, as it is currently not possible to confirm the presence of breast cancer without a biopsy. This study proposes the k-Nearest Neighbor (k-NN) technique, which is commonly used in machine learning for regression and classification. This study requires a number of steps, such as importing the dataset, pre-processing the data, and choosing the characteristics that need to be classified. The k-NN method additionally employed a variety of distance metrics to distinguish between benign and malignant tumours. Additionally, to demonstrate the effectiveness of the suggested strategy, the produced anser is contrasted with other outcomes. With improved distance measurements made possible by the k-NN algorithm, the study's findings advance our understanding of breast cancer prediction. Topsoe, Lorentzian distance, and Average(L_1, L_∞)approaches produced the most reliable overall results. Comparisons are made between the outcomes and established techniques such as the Euclidean, Clark, and Bray-Curtis distances.

Keywords: Breast Cancer, k-Nearest Neighbor, Performance Measures, Topsoe, Average (L_1, L_∞) , Lorentzian.

1.INTRODUCTION:

Malignant cells in the breast tissue give rise to the development of breast cancer. Numerous variables, including age, genetics, family history, and lifestyle choices, raise the chance of contracting this illness. Even though a mammography can reveal abnormal cells, it cannot provide a 100% accurate result without a biopsy. This procedure, known as a breast biopsy, involves removing tissue from the afflicted area and using it for diagnostic purposes. This yields a satisfactory outcome for determining the cell abnormalities and for stage analysis. Though it has certain shortcomings, such as

1. The afflicted area may show indications of swelling and bruises.
2. The area where the biopsy was taken has the potential to get infected.
3. They can feel pain where the injection was made.

4. Because to the tissue excision, the breasts may vary in size.

Due to the aforementioned disadvantages, having a backup prediction strategy that doesn't include biopsy is a smart idea. Since biopsies are expensive, developing novel methods for more accurately predicting breast cancer without using a biopsy has become a top priority for research. The results are compared with previously published techniques such as the Clark, Euclidean, and Bray-Curtis distances in this publication. The most reliable overall results were attained by utilizing Average(L_1, L_∞), Topsoe and Lorentzian distance methods.

2.PRELIMINARIES:

The k-Nearest Neighbor(k-NN) was suggested by Evelyn Fix and Joseph Hodges in 1951[15.] This proposed model was used in classification and regression. It is also used in Pattern Recognition[19], ranking models[7],categorization of text[12], recognition of objects[3] and in medicines[9,10,14]. It is often referred to as “lazy learning” since the model merely remembers the training data set rather than developing a unique function from the training set. In order to determine the value (regression) or locate the k-Nearest data points in the input sample (classification), it uses the values or labels of the nearby points as a basis. Classification selects the data to yield a classification as an output, whereas regression yields an object value. In this work, the k-NN technique is applied to classify all objects in the dataset as either benign or malignant, yielding important results. In order to find an object's closest neighbors, the k-NN technique employs various distance functions. Converging features including the patient's age, mass, form, margin, and density are taken into account when establishing and computing this closet object.

3.MEASURES OF DISTANCE:

To execute the k-NN algorithm, it is essential for determining the distance between the testing and training data. This article, provides the mathematical equations that calculate the distance between two vectors x and y, that have numerical attributes. The distance measure $d_m(x, y)$ measures the distance between x and y based on the chosen metric m. The formulations and terminologies are defined from Abu Alfeilat [1].

3.1 CLARK DISTANCE:

This distance[5] is also referred as the coefficient of divergence which is calculated as the square root of half of the total divergence distance.

$$d_{clark} = \sqrt{\sum_{i=1}^{n} \left(\frac{x_i - y_i}{|x_i| + |y_i|} \right)^2} \qquad (1)$$

3.2 BRAY-CURTIS DISTANCE:

The measurement which is known as Bray-Curtis is frequently used in the fields such as ecology and environmental science [16] to characterize relationships. It’s like a modified version of Manhattan distance wherein the total sum of value is utilized to normalize the difference between the vectors x and y.The distance metric will range from 0 to where the vector vales are positive.

$$d_{Bray-Curtis} = \sum_{i=1}^{n} \frac{|x_i - y_i|}{(x_i + y_i)} \qquad (2)$$

3.3 EUCLIDEAN DISTANCE:

To calculate this distance[6] it is supposed to take the square root of the sum of the squares of the differences between the coordinates of each point.

$$d_{Euclidean} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (3)$$

4.PREDICTION OF BREAST CANCER USING AVERAGE(L_1, L_∞)DISTANCE, TOPSOE DISTANCE, LORENTZIAN DISTANCE

4.1 AVERAGE (L_1, L_∞) DISTANCE:

Average (L_1, L_∞) represents the mean of the Manhattan and Chebyshev distances.

$$d_{Avg} = \frac{\sum_{i=1}^{n}|x_i - y_i| + \max_i |x_i - y_i|}{2} \qquad (4)$$

4.2 TOPSOE DISTANCE:

The Topsoe distance[18], also known as information statistics, which is the symmetrical version of the Kullback-Leibler distance. Although the Topsoe distance itself is not a metric, its square root can be considered as metric.

$$d_{Topsoe} = \sum_{i=1}^{n} x_i \ln\left(\frac{2x_i}{x_i + y_i}\right) + \sum_{i=1}^{n} y_i \ln\left(\frac{2y_i}{x_i + y_i}\right) \qquad (5)$$

4.3 LORENTZIAN DISTANCE:

Lorentzian distance is defined by the logarithm of the absolute variance between two vectors. This measure is highly adaptable to minor alterations as the logarithmic scale magnifies the lower range and contrasts it with the higher range. In order to maintain the non-negativity characteristic and prevent taking the logarithmic of zero, the value of 1 is included.

$$d_{Lorentzian} = \sum_{i=1}^{n} ln(1 + |x_i - y_i|) \qquad (6)$$

5.STRUCTURE OF OUTCOMES:

For each classifier, 4 complementing measures to assess its performance is employed : Accuracy, Precision, Recall and F1 score. These calculations can be derived from the classification outcomes provided, where a certain subset of patterns is assigned to a particular class, while the remaining patterns are not assigned to that class.

5.1 TRUE POSITIVE(TP):

The number of positive patterns which is classified correctly belongs to the positive set.

5.2 TRUE NEGATIVE(TN):

The quantity of patterns from the outside of the positive set which are identified accurately doesn't belong to the positive set.

5.3 FALSE POSITIVE(FP):

The number of negative patterns which is classified incorrectly belongs to the positive set.

5.4 FALSE NEGATIVE(FN):

The number of positive patterns which is incorrectly classified doesn't belongs to the positive set. The appropriate metrices for evaluating performance can be subsequently established as;

5.5 PRECISION:

Precision[2] is defined as the proportion of correct positive predictions to all positive observation

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (7)$$

5.6 RECALL:

Recall[11] is the proportion of accurate positive predictions compared to the total positive outcomes.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (8)$$

5.7 F1 SCORE:

F1 Score [4] is defined as the arithmetic mean of Precision and Recall with equal weightage given to both metrics.

$$\text{F1 Score} = \frac{2(\text{precision}\times\text{Recall})}{\text{Precision}+\text{Recall}} \qquad (9)$$

5.8 ACCURACY:

Accuracy [2] is the ratio of correctly identified results.

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \qquad (10)$$

6.COMPARISON OF RESULTS WITH ALREADY EXISTING RESULTS

Table 1 displays the scores for the breast cancer data. The findings presented in this paper demonstrates superior performance when compared to the results obtained from existing papers that utilized various other distance metrics. Based on every performance metric, the Topsoe distance performed significantly better than the other distances.

Distance	Precision	Recall	F1Score	Accuracy
Clark	0.967	0.972	0.969	0.971
Euclidean	0.966	0.962	0.963	0.967
Bray Curtis	0.969	0.969	0.968	0.971
Topsoe	**0.9726**	**1.0**	**0.9861**	**0.9825**
Average (L_1, L_∞)	0.9718	0.9718	0.9718	0.9649
Lorentzian	0.9722	0.9859	0.9790	0.9737

Table 1: The scores among all evaluated k values for breast cancer data.

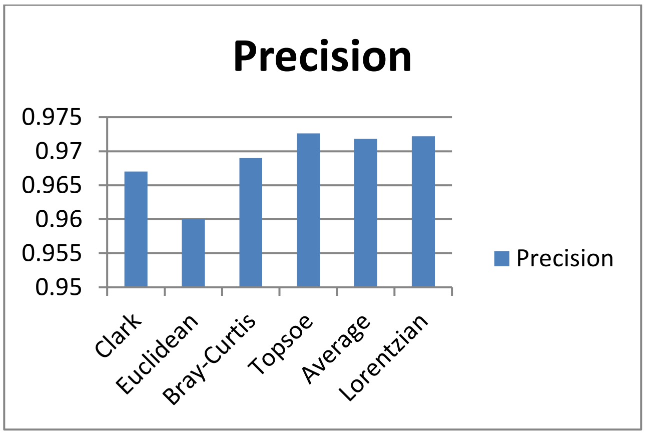

Figure 1: The Performance of Precision for Topsoe, Average(L_1, L_∞), Lorentzian yields the best scores when compared with existing metrices

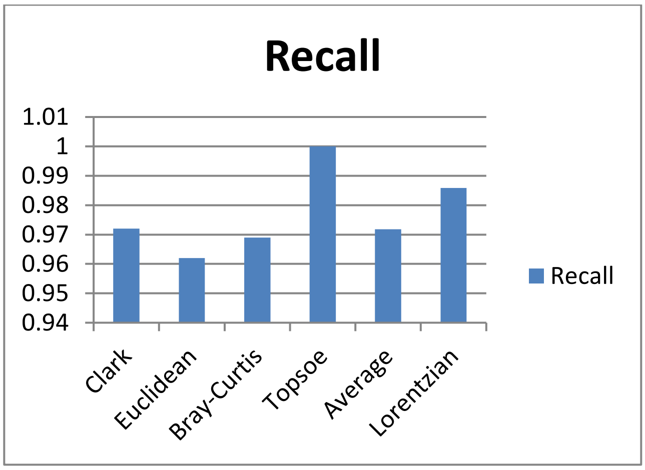

Figure 2: The Topsoe distance outperformed the other distance for Recall.

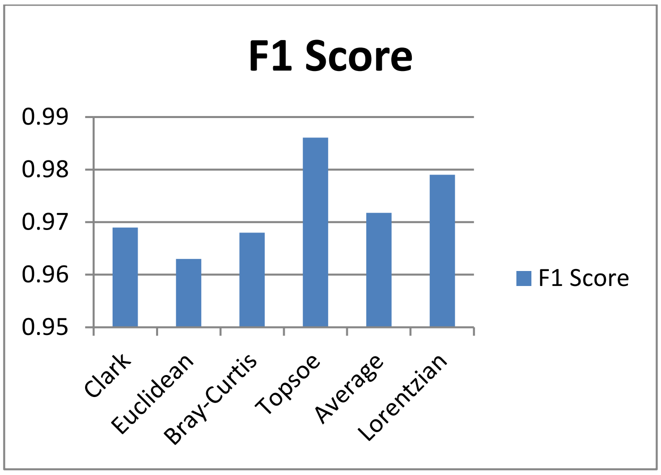

Figure3: The F1 score for Topsoe and Lorentzian distance shows the higher variance.

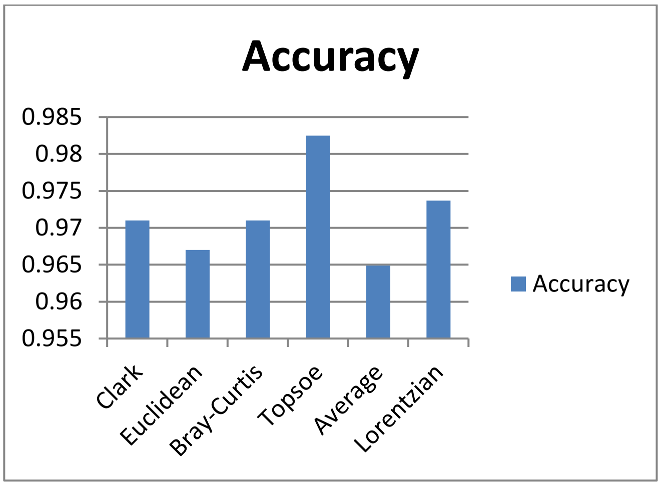

Figure 4: The Accuracy for Topsoe distance gives the better results.

7.ACCESS TO DATA AND MATERIALS:

This Paper uses the breast cancer dataset which is sourced by UCI-Machine Learning Repository. “Breast Cancer Wisconsin (Original) Data Set (https:// archive.ics.uci.edu/ml/datasets/Breast+ Cancer+ Wisconsin +%28Original%29).

8.SOFTWARE UTLIZED:

The Python Programming language (version 3.7.1) was utilized for scripting purposes. To calculate different distance, we employed the k-NN algorithm and made use of libraries from the scikit-learn package (version 0.20.1)

9.CONCLUSION:

The Performance examination of the k-NN classification of cancer data using various distance measures reveals significant variations in the data sets and distance measurements. It is important to note that no single measure is perfect for all data sets. Therefore it is recommended to test multiple measures on similar reference data before choosing the best one for a specific data.

REFERENCES:

1. Abu Alfeilat HA, Hassanat ABA, Lasassmeh O, et al. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big Data. 2019;7:221-248
2. Aghdam, H. H., & Heravi, E. J. (2017). Guide to convolutional neural networks: a practical application to traffic-sign detection and classification. Springer

3. Bajramovic F, Mattern F, Butko N, Denzler J. A Comparison of Nearest Neighbor Search Algorithms for Generic Object Recognition. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 4179. Berlin, Germany: Springer
4. Bramer, M. (2013). Principles of data mining, second edition. London: Springer
5. Clark PJ. An extension of the coefficient of divergence for use with multiple characters. Copeia. 1952;1952:61-64.
6. Euclid. (1956). The Thirteen Books of Euclid's Elements. Courier Corporation.
7. Geng X, Liu T-Y, Qin T, Arnold A, Li H, Shum H-Y. Query dependent ranking using K-nearest neighbor. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Singapore). New York, NY: Association for Computing Machinery; 2008:115-122.
8. Karan Sharma, Victoria Rodriguez et.al, Dana Walker(2018). Breast Cancer Prediction with K-Nearest Neighbor Algorithm using Different Distance Measurements
9. Khamis HS, Cheruiyot KW, Kimani S. Application of k-nearest neighbour classification in medical data mining. Int J Inform Commun Technol Res. 2014;4:121-128.
10. Kusmirek W, Szmurlo A, Wiewiorka M, Nowak R, Gambin T. Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers performance. BMC Bioinform. 2019;20:266.
11. Larose, D. T., & Larose, C. D. (2015). Data mining and predictive analytics. John Wiley & Sons.
12. Manne S, Kotha SK, Sameen Fatima S. Text categorization with k-nearest neighbor approach. In: Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012), Visakhapatnam, India; Berlin, Germany; Heidelberg, Germany: Springer; 2012:413-420
13. Rezvan Ehsani and Finn Drablos(2020) . Robust Distance Measures for k-NN Classification of Cancer data
14. Roder J, Oliveira C, Net L, Tsypin M, Linstid B, Roder H. A dropout-regularized classifier development approach optimized for precision medicine test discovery from omics data. BMC Bioinform. 2019;20:325.
15. Silverman BW, Jones MC, Fix E, Hodges JL. An important contribution to nonparametric discriminant analysis and density estimation: commentary on Fix and Hodges (1951). Int Stat Rev. 1989;57:233-238.
16. Sørensen T. A method of establishing groups of equal amplitudes in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskab, Biologiske Skrifter. 1948;5:1-34.
17. Szmidt E. Distances and Similarities in Intuitionistic Fuzzy Sets. Berlin, Germany: Springer; 2013.
18. Topsoe, F. (2000). Some inequalities for information divergence and related measures of discrimination. IEEE Transactions on information theory, 46 (4), 1602–1609.
19. Xu S, Wu Y. An algorithm for remote sensing image classification based on artificial immune B-cell network. In: Jun C, Jie J, Cho K, eds. Xxist ISPRS Congress, Youth Forum, Vol. 37. Beijing, China: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; 2008:107-112.