Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/jptcp.v31i7.6945

BREAST CANCER RECURRENCE ANALYSIS USING DISTANCE MEASURES IN K-NN ALGORITHM

Dr. S. Bharathi, Krithika.L*

Assistant Professor, Department of Mathematics, Bharathiar University PG Extension and Research Center, Perundurai, Erode, Tamilnadu, India. Email: bharathikamesh6@gmail.com
*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 is a prevalent type of cancer that primarily affects females. Extensive research has been conducted in the field of breast cancer, and with the advancements in technology, the early detection of this disease has become possible through the utilization of artificial intelligence or machine learning techniques. The objective of this study is to assess the accuracy of predicting the recurrence of breast cancer by employing the k-Nearest Neighbor (k-NN) algorithm. The k-NN classifier is a straightforward and versatile approach to classification, which often demonstrates comparable performance to more intricate machine-learning algorithms. The effectiveness of k-NN classifiers is closely associated with the selection of a suitable distance or similarity measure. Therefore, it is crucial to investigate the impact of employing various distance measures when analyzing biomedical data. The findings of this study indicate that the k-NN algorithm, utilizing diverse distance measurements, yields the most favorable outcomes in terms of accurately predicting the recurrence of breast cancer.

Keywords:- Breast Cancer-KNN-Distance Measurements-Recurrence Prediction

1.Introduction

Breast cancer has emerged as the most common carcinogenic disease globally, posing a threat of causing a million deaths. Early detection stands out as the most effective strategy to mitigate breast cancer fatalities. If left untreated, the cancer can affect almost all vital organs in the body, leading to fatal consequences in some cases. Diagnostic tests such as MRI, mammogram, ultrasound, and biopsy play a crucial role in the early detection of breast cancer. Breast cancer is characterized by the uncontrolled growth of cells in breast tissue. Failure to control these cells can have adverse effects on the entire body. Men can also be affected by breast cancer, which carries a higher mortality rate[15]. Several risk factors, including increasing age, obesity, excessive alcohol consumption, family history of breast cancer, history of radiation exposure, reproductive history, tobacco use, and postmenopausal hormone therapy, can elevate the risk of developing breast cancer. Approximately half of breast cancer cases occur in women without any identifiable risk factors other than being female and over 40 years old. The treatment of breast cancer is a lengthy and

challenging process, involving diagnosis, surgery (if necessary), and therapy[24](radiation, chemotherapy, and medication).

The inadequacy of doctors and general practitioners persists in developing countries such as Indonesia. Data illustrates a ratio of 0.4 doctors per 1000 individuals in the population[25], marking it as the second lowest in South East Asia. This critical situation drives healthcare researchers to explore novel strategies for improving healthcare services for the populace, particularly through the utilization of technology.

A multitude of studies have been conducted in the realm of healthcare technology[6,10], including the development of a heart disease prediction system. One study utilized the k-nearest neighbor (kNN) machine learning algorithm to identify the type of heart disease in patients, while another research project proposed a heart disease prediction system using a Hybrid Random Forest with a Linear Model(HRFLM)[18]. Furthermore, data mining algorithms like kNN and Bayesian were employed to predict diabetes [5] in patients. In the context of breast cancer, a study focused on predicting benign or malignant breast cancer [22] using data mining techniques such as naïve bayes, RBF Network, and J48 Decision Tree, while [9]another study investigated the same topic using five different machine learning algorithms: C4.5, support vector machine (SVM), naïve bayes, and kNN.

Breast cancer diagnosis involves classifying the tumor to determine if it is benign or malignant. Malignant tumors are the ones that indicate cancer. It would be beneficial to have a system that can predict whether breast cancer will recur or not after the patient has undergone treatment for a certain period of time[2]. The aim of this research is to develop a more accurate method for predicting breast cancer recurrence using the k-NN algorithm. The effectiveness of k-NN classifiers depends on the choice of distance or similarity measure. Therefore, it is important to investigate the impact of using different distance measures when comparing biomedical data. In this study on predicting breast cancer recurrence, we have utilized both conventional and innovative distance measures such as Average (L_1, L_∞) , Bray-Curtis, Chebyshev, Jaccard, Kulczynski, Matusita and Squared Euclidean. We have assessed the performance of k-NN with these distances and found that the Chebyshev distance yielded the best results.

2.Preliminaries

The k-Nearest Neighbor (k-NN) algorithm was initially proposed by Evelyn Fix and Joseph Hodges in 1951[19]. This model has found applications in various including classification,regression,Pattern Recognition[26],ranking models[7],categorization of text[14],recognition ofobjects[3],and even in the field of medicines[12,13,17]. It is often referred to as Lazy Learning because it does not derive a distinct function from the training data; instead, it simply memorizes the training dataset. The k-NN algorithm operates by identifying the k-Nearest data points in the input sample for classification or predicting the value for regression based on the values or labels of the neighboring points. In classification, it determines the appropriate classification for the data, while in regression, it produces a numerical value for the object. This research examines the methods for detecting the recurrence of breast cancer, yielding significant outcomes. The algorithm accomplishes this by utilizing different distance functions to determine the closest neighbors of a specific object. These neighbors are determined by considering overlapping characteristics such as the age of the patient, mass shape, margin, or density.

3.Dataset

The breast cancer dataset utilized in this Paper has been obtained from the UCI-Machine Learning Repository.

https://archive.ics.uci.edu/ml/datasets/breast+cancer.

“Breast Cancer Data Set” (https://archive.ics.uci.edu/ml/datasets/breast+cancer)

4.Distance metrics

This article provides the mathematical formulas that measure the distance between two vectors x and$y$, each with numerical attributes. The distance metric $d_m(x, y)$ evaluates the distance between x and y depending on the chosen metric m. The definitions and terms are outlined by Abu Alfeilat [1].

4.1 Average (L_1, L_∞) distance

Average (L_1, L_∞) is computed as the arithmetic mean of the Manhattan and Chebyshev distances.

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

4.2 Bray-Curtis distance

In the fields of ecology and environmental science [20], the Bray-Curtis measurement is frequently utilized to characterize relationships between variables. This measurement can be viewed as a modified version of the Manhattan distance, with the total sum of values being used to normalize the difference between vectors x and$y$. The resulting distance metric will range from 0 to the maximum value when vector values are positive.

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

4.3 Chebyshev distance

The Chebyshev distance, also known as the maximum value distance[8], Lagrange distance[21], and chessboard distance[16], is utilized to distinguish between two objects based on variations in a single dimension[23]. This metric is defined on a vector space, where the distance between two vectors is calculated as the largest difference along any coordinate dimension.

$$d_{chebyshev} = max_i|x_i - y_i| \qquad (3)$$

4.4 Jaccard distance

The Jaccard distance serves as a metric for quantifying the dissimilarity between sets of samples. It acts as a counterpart to the Jaccard similarity coefficient [11]and is derived by subtracting the Jaccard coefficient from one. This distance[4]metric provides a measure of dissimilarity between sets.

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

4.5 Kulczynski Distance

In contrast to the Soergel distance, this technique makes use of the minimum function rather than the maximum.

$$d_{kulczynski} = \frac{\sum_{i=1}^{n}|x_i - y_i|}{\sum_{i=1}^{n} min(x_i, y_i)} \qquad (5)$$

4.6 Matusita distance

The Matusita distance is determined by finding the square root of the squared chord distance.

$$d_{\text{Matusita}} = \sqrt{\sum_{i=1}^{n}\left(\sqrt{x_i} - \sqrt{y_i}\right)^2} \qquad (6)$$

4.7 Squared Euclidean distance

The squared Euclidean distance is derived from the sum of squared differences, without the square root being applied.

$$d_{\text{Squared Euclidean}} = \sum_{i=1}^{n} (x_i - y_i)^2 \quad (7)$$

5.Results and Discussions

The accuracy of the appropriately classified data for the prognosis of breast cancer has been determined and is displayed below.

Table 1: The scores among all evaluated k values for the data

Distance	Training Accuracy	Testing Accuracy
Average(L_1, L_∞)	0.848	0.72
Bray-Curtis	0.810	0.75
Chebyshev	**0.860**	**0.75**
Jaccard	0.848	0.75
Kulczynski	0.810	0.75
Matusita	0.803	0.7
Squared Euclidean	0.848	0.72

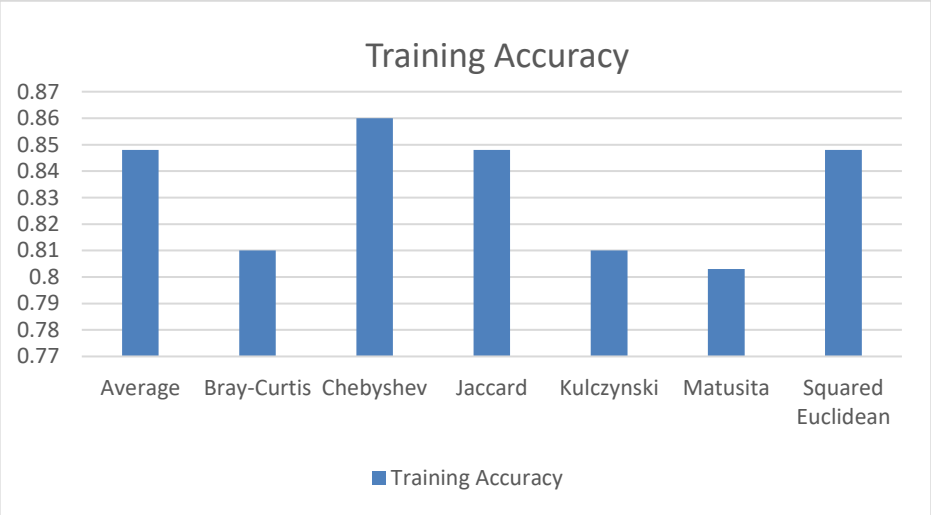

Figure 1: The Chebyshev distance outperformed the other distances for training accuracy.

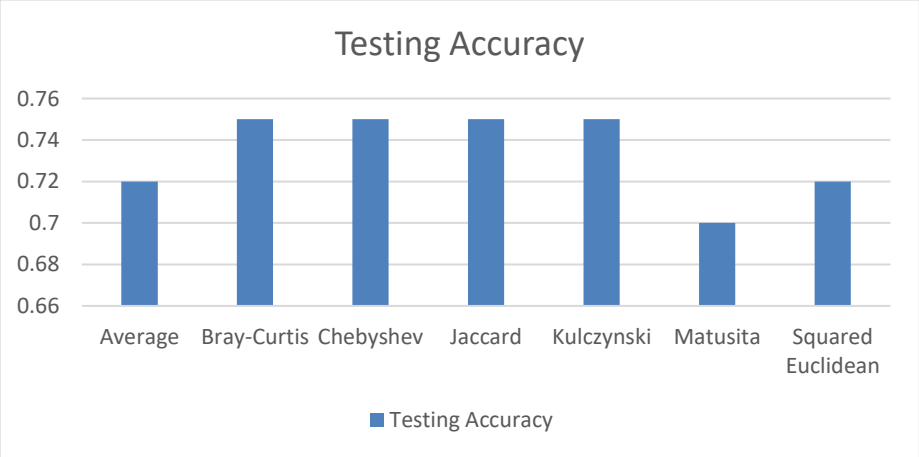

Figure 2: The Performance of all distances gives the similar results for Testing Accuracy.

6.Software implemented

The Python Programming language version 3.7.1 was used for scripting purposes within Anaconda3. The k-NN algorithm was implemented to compute various distances, with libraries from the scikit-learn package version (0.20.1) being utilized.

7.Conclusions

Machine learning plays a vital role in medical applications, bringing about significant advancements. Through the analysis of kNN classification on cancer data using various distance measures, important distinctions between the measures and data sets are revealed. The study highlights the exceptional accuracy achieved with the Chebyshev distance measure. However, it also emphasizes that no single measure can be universally optimal for all data sets. Therefore, it is advisable to evaluate multiple measures on reference data resembling the actual data when choosing a distance measure for a particular study.

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]Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR - Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, J Health Med Inform 2013,4:2,http://dx.doi.org/10.4172/2157-7420.1000124.

[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] Cesare, S., and Xiang, Y. (2012). Software Similarity and Classification. Springer.

[5]D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining," in Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS 2017, 2018.

[6]Enriko, I. K. A., Suryanegara, M., & Gunawan, D, "Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters," J. Telecommun. Electron. Comput. Eng., vol. 8, no. 12, pp. 59–65, 2016.

[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]Grabusts, P. (2011). The choice of metrics for clustering algorithms. Environment. Technology. Resources , 70–76.

[9]H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," in Procedia Computer Science, 2016.

[10]I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease diagnosis system with k-nearest neighbors method using real clinical medical records," in ACM International Conference Proceeding Series, 2018

[11]Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et du Jura.

[12]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.

[13]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.

[14]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

[15]National Breast Cancer Foundation Inc., http://www.nationalbreastcancer.org/about-breast-cancer

[16] Premaratne, P. (2014). Human computer interaction using hand gestures. Springer.

[17]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.

[18]S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, 2019.

[19]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.

[20]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.

[21]Todeschini, R., Ballabio, D., & Consonni, V. (2015). Distances and other dissimilarity measures in chemometrics. Encyclopedia of Analytical Chemistry.

[22]V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," J. Algorithms Comput. Technol., 2018.

[23]Verma, J. P. (2012). Data Analysis in Management with SPSS Software. Springer.

[24]A.G Waks and E.P . Winer, "Breast Cancer Treatment: A Review ," JAMA-Journal of the American Medical Association.2019.

[25]W. Bank, "Physicians (Per 1,000 People)," World Bank Report, 2020. [Online]. Available: https://data.worldbank.org/indicator/SH.MED.PHYS. ZS?most_recent_value_desc=true. [Accessed: 15-Feb2021].

[26]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