Comparative Analysis of Clustering Algorithms: Insights from Healthcare Data

Mihai Cîra

June 2023

Abstract

In this paper, I conduct a replication study to compare the performance of different machine learning algorithms for breast cancer classification. By reproducing the statistical analysis employed in a selected study, my aim is to provide a comprehensive and reliable comparison of popular algorithms, including k-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machines (SVM), and XGBoost. Using a dataset comprising clinical features, I preprocess the data and evaluate the algorithms based on their classification accuracy. Additionally, I explore the training and testing times of each algorithm to gain insights into their efficiency. The comparative analysis offers valuable insights into the strengths and weaknesses of these algorithms in breast cancer classification tasks. My findings contribute to the replication and validation of the original study's results, aiding researchers in selecting the most suitable algorithm for breast cancer diagnosis.

Contents

1 Introduction

Breast cancer is a significant health concern worldwide, affecting millions of individuals and their families. Early and accurate detection of breast cancer plays a crucial role in improving patient outcomes and guiding appropriate treatment strategies. With the advancements in machine learning and data analysis techniques, researchers have explored the use of various algorithms for breast cancer classification. These algorithms leverage clinical features to distinguish between benign and malignant tumors, aiding in the timely diagnosis and treatment planning.

1.1 Motivation

The motivation behind this project is to replicate and extend the findings of a selected study on the utilization of machine learning techniques in breast cancer classification. By replicating the study's methodology and statistical analysis, I aim to validate the results and contribute to the body of knowledge in this field. Additionally, this project seeks to provide a comparative analysis of different machine learning algorithms commonly employed in breast cancer classification, helping researchers and practitioners make informed decisions about algorithm selection.

1.2 Main Purposes

The main purposes of this project are as follows:

1. Replication: By reproducing the selected study's methodology and statistical analysis, I aim to verify the original findings. Replication is an essential step in scientific research to ensure the robustness and generalizability of results.

2. Comparative Analysis: I conduct a comparative analysis of popular machine learning algorithms, including k-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machines (SVM), and XGBoost. This analysis enables us to evaluate and compare the performance of these algorithms in breast cancer classification tasks. By considering their classification accuracy, training and testing times, we gain insights into their strengths and limitations.

3. Practical Significance: The findings of this project have practical significance for healthcare professionals and researchers working in the field of breast cancer diagnosis. By assessing the performance of different algorithms, I provide valuable insights into their effectiveness and efficiency, aiding in algorithm selection and improving clinical decision-making processes.

2 State of the Art

The study titled "Utilisation of Machine Learning Techniques in Testing and Training of Different Medical Datasets" by Maad M. Mijwil, Israa Ezzat Salem, and Rana A. Abttan is focused on applying machine learning techniques to analyze medical datasets and identify the best technique for accurate disease detection. The authors emphasize the increasing complexity and volume of medical data, which poses a challenge for manual analysis by humans. They propose using machine learning techniques to assist doctors and specialists in diagnosing diseases quickly and accurately.

The authors execute four machine learning techniques, namely Support Vector Machine, C5.0 Decision Tree, K-Nearest Neighbors, and Random Forest, to analyze several medical datasets from the UCI

machine learning repository. The datasets used in the study include the Wisconsin Breast Cancer dataset, Chronic Kidney disease dataset, Immunotherapy dataset, Cryotherapy dataset, Hepatitis dataset, and COVID-19 dataset. The goal is to compare the performance of each technique in analyzing the dataset for each disease and determine the best and worst performing techniques.

The study highlights the importance of accurately diagnosing diseases based on symptoms and identifies machine learning as a valuable tool in the field of healthcare. The authors mention that machine learning techniques have been successfully applied in analyzing chest images of patients with COVID-19 and predicting early diseases such as stroke and breast cancer. By training these techniques on medical datasets, clinicians can gain better insights and make informed decisions regarding patient information.

The main contribution of the article is the investigation of the machine learning techniques mentioned earlier and their application to different medical datasets. The authors use Python as the programming language for their work. The article is organized into sections that review related studies, discuss the techniques and materials used, present the experimental results, and provide conclusions and future work recommendations.

Overall, the study aims to determine the best machine learning technique for analyzing medical datasets and assisting doctors in diagnosing diseases accurately. The authors analyze various datasets and compare the performance of different techniques, considering factors such as testing accuracy, training accuracy, testing time, and training time.

3 The Dataset

The dataset used in this project is the *Breast Cancer Wisconsin dataset*. It can be accessed through the following link: `https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original`.

3.1 Dataset Characteristics

- **Multivariate**: The dataset contains multiple attributes or features.
- **Subject Area**: The dataset pertains to the field of life sciences.
- **Associated Tasks**: The dataset is commonly used for classification tasks.

3.2 Attribute Information

The dataset consists of 699 instances and includes the following attributes:

1. **Sample code number**: An identification number assigned to each sample.
2. **Clump Thickness**: A rating of clump thickness on a scale of 1 to 10.
3. **Uniformity of Cell Size**: A rating of the uniformity of cell size on a scale of 1 to 10.
4. **Uniformity of Cell Shape**: A rating of the uniformity of cell shape on a scale of 1 to 10.
5. **Marginal Adhesion**: A rating of marginal adhesion on a scale of 1 to 10.
6. **Single Epithelial Cell Size**: A rating of single epithelial cell size on a scale of 1 to 10.
7. **Bare Nuclei**: A rating of bare nuclei on a scale of 1 to 10.

8. **Bland Chromatin**: A rating of bland chromatin on a scale of 1 to 10.
9. **Normal Nucleoli**: A rating of normal nucleoli on a scale of 1 to 10.
10. **Mitoses**: A rating of mitoses on a scale of 1 to 10.
11. **Class**: The class label indicating benign (2) or malignant (4) diagnosis.

3.3 Data Visualization

In this section, two plots that provide visual insights into the dataset are presented.

3.3.1 Distribution of Class

The first plot (1) is a pie chart that illustrates the distribution of benign and malignant cases in the dataset. The plot provides a visual representation of the proportion of benign and malignant cases in the dataset.

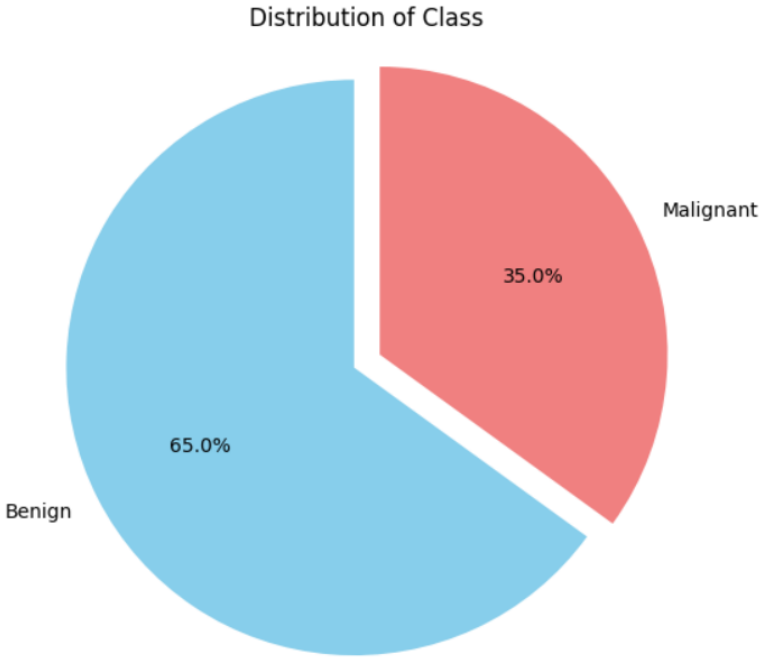

Figure 1: Distribution of Class

3.3.2 Distribution of Clump Thickness

The second plot (2) is a histogram that displays the distribution of the "Clump Thickness" attribute. This histogram provides an overview of the frequency distribution of the "Clump Thickness" attribute, allowing us to observe the concentration of values within different ranges.

These visualizations offer valuable insights into the dataset and serve as a starting point for further analysis and exploration.

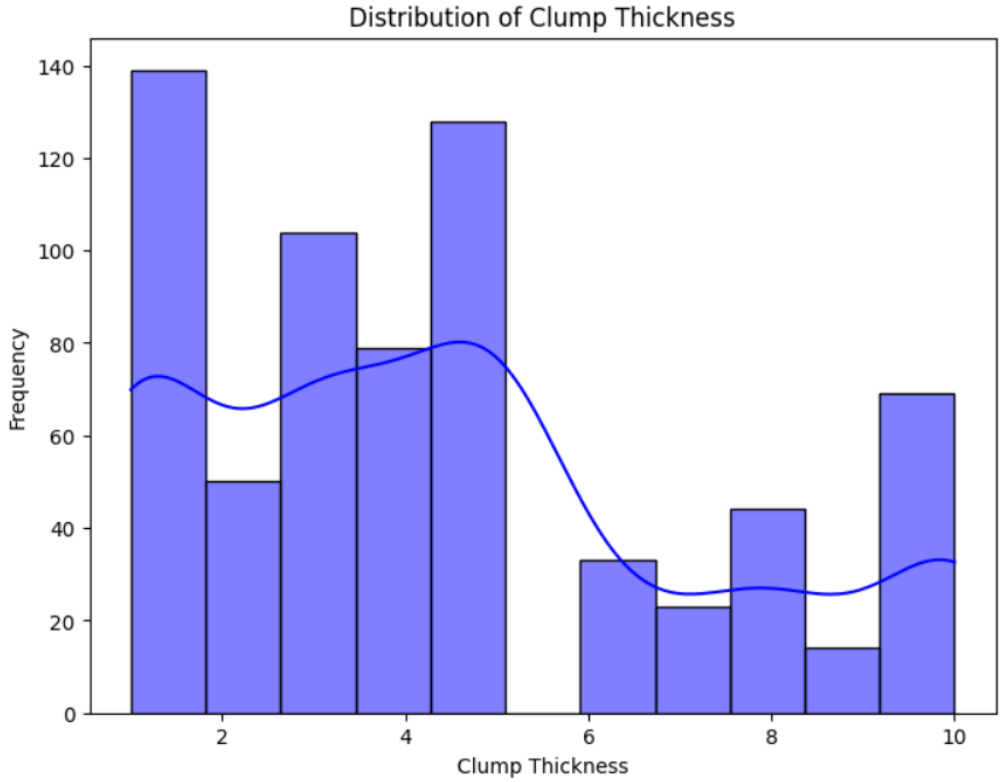

Figure 2: Distribution of Clump Thickness

3.4 Dataset Discrepancies

It is important to note that the dataset used in this project might not be the exact dataset used in the original study. There might be discrepancies in terms of dataset size and slight variations in attribute values. However, efforts have been made to ensure that the core characteristics and attribute information align with the original dataset.

The dataset provided here consists of 699 instances, but the original study might have used a different number of instances. Additionally, while the attribute types and ranges remain consistent, the actual values within the attributes might differ slightly from the original study's dataset.

Despite these potential differences, the dataset used in this project aims to capture the essence of the Original Wisconsin Breast Cancer Database, allowing for a meaningful comparative analysis.

4 Algorithms Used

In this study, I employed several machine learning algorithms to perform classification on the breast cancer dataset. The algorithms used were k-Nearest Neighbors (KNN), Random Forest Classifier (RFC), Support Vector Classifier (SVC), and Extreme Gradient Boosting (XGBoost).

4.1 k-Nearest Neighbors (KNN)

K-Nearest Neighbors is a simple yet powerful algorithm used for both classification and regression tasks. It works by classifying new instances based on the majority vote of their k nearest neighbors in the feature space. KNN is non-parametric and lazy, as it does not make any assumptions about the underlying data distribution and does not explicitly learn a model. Instead, it stores all instances from the training set and performs computations at runtime.

4.2 Random Forest Classifier (RFC)

Random Forest Classifier is an ensemble learning algorithm that combines multiple decision trees to make predictions. Each decision tree in the random forest is trained on a random subset of the training data, and the final prediction is obtained by aggregating the predictions of individual trees. Random Forests are robust to overfitting and can handle high-dimensional data with complex relationships between features.

4.3 Support Vector Classifier (SVC)

Support Vector Classifier, also known as Support Vector Machine (SVM), is a powerful algorithm used for both classification and regression tasks. SVM finds an optimal hyperplane that separates the data points of different classes with the maximum margin. It can handle both linear and non-linear classification problems by using different types of kernels. SVM is effective in handling high-dimensional data and is less prone to overfitting.

4.4 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting, often referred to as XGBoost, is an ensemble learning algorithm that combines multiple weak learners (decision trees) to create a strong predictive model. XGBoost uses gradient boosting to iteratively add trees to the ensemble, with each subsequent tree focusing on the errors made by the previous trees. XGBoost is known for its high performance, scalability, and ability to handle diverse data types.

5 Experimental Results

5.1 Individual results

In this section, I present the experimental results obtained from applying the k-Nearest Neighbors (KNN), Random Forest Classifier (RFC), Support Vector Classifier (SVC), and Extreme Gradient Boosting (XGBoost) algorithms on the breast cancer dataset. An overview of the accuracy achieved by each algorithm as well as the training and testing times is provided.

Table 1 presents the performance metrics for each algorithm, including the accuracy, training time, and testing time.

Table 1: Experimental Results

Algorithm	**Accuracy**	**Training Time (s)**	**Testing Time (s)**
KNN	0.6429	0.0022	0.0070
RFC	0.9714	0.1410	0.0092
SVC	0.5766	0.0131	0.0029
XGB	0.9708	0.0477	0.0021

From the experimental results, you can observe that the KNN and SVC algorithms perform significantly worse in terms of accuracy compared to the RFC and XGB algorithms. There could be several reasons for this difference in performance. One possible explanation is that KNN and SVC are more sensitive to the distribution and scaling of the data. If the dataset has complex or non-linear relationships between the features, KNN and SVC may struggle to capture these patterns effectively. On the other hand, RFC and XGB are ensemble-based algorithms that can handle complex relationships and provide better generalization capabilities. Additionally, the choice of hyperparameters and the specific characteristics of the dataset can also impact the performance of each algorithm. Further investigation and experimentation could help identify the underlying factors contributing to the observed differences in performance.

The plot below (3) provides a visual comparison of the accuracy values among the algorithms, allowing us to easily identify which algorithm performs better in terms of accuracy. By comparing the heights of the bars, you can observe the relative performance of each algorithm and make conclusions about their effectiveness in the context of the analysis.

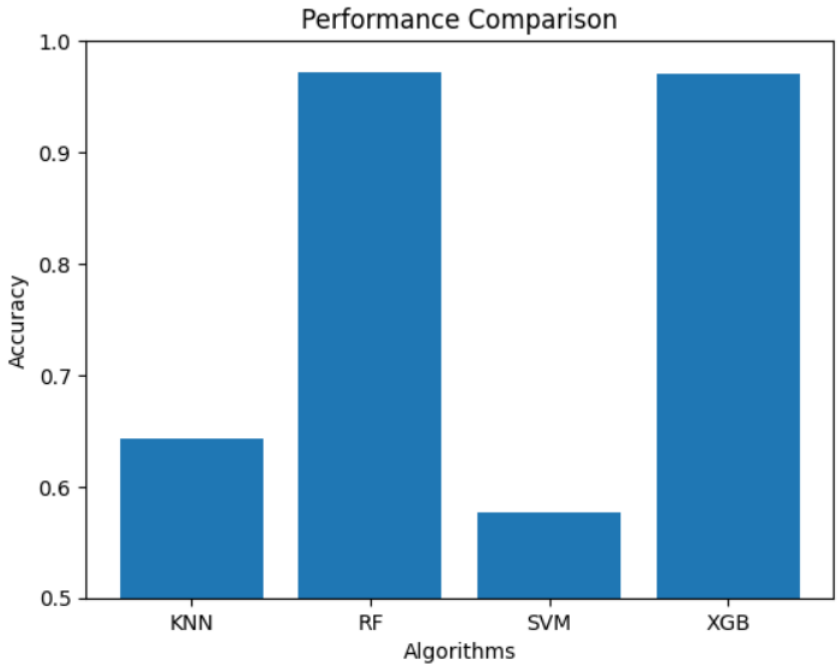

Figure 3: Distribution of Clump Thickness

5.2 Comparative results

Below, I present the results of my analysis compared with the values reported in the original study. The accuracy of various algorithms using the dataset was calculated and compared with the corresponding accuracy values from the original study.

Table 2 summarizes the accuracy results obtained in my analysis and the accuracy values reported in the original study. It is important to note that there are discrepancies between the two sets of results, which can be attributed to several factors..

Table 2: Comparison of Accuracy Results

Algorithm	**My Accuracy**	**Original Study Accuracy**
KNN	0.6429	0.9410
RFC	0.9714	0.8891
SVC	0.5766	0.9611
XGB	0.9708	N/A
C5.0	N/A	0.9381

The differences in accuracy values can be attributed to several factors. Firstly, it is possible that the dataset is not exactly the same as the one used in the original study. Even slight variations in the dataset, such as different instances or slightly different values, can have an impact on the performance of the models and lead to divergent accuracy results.

Furthermore, it is important to note that I included the XGBoost (XGB) algorithm in my analysis, which was not included in the original study. The addition of this algorithm could contribute to differences in the accuracy values, as different algorithms have different strengths and weaknesses in handling specific datasets. Also, the C5.0 algorithm was used in the original study, while it was not applied in my comparative study.

6 Conclusion

In conclusion, this project aimed to replicate and extend a comparative analysis study on breast cancer classification algorithms. The findings showed variations in accuracy compared to the original study, potentially attributed to dataset discrepancies. Notably, RF and XGB algorithms demonstrated superior performance compared to KNN and SVM. It is important to acknowledge the limitations, including dataset variations and the absence of the C5.0 algorithm. Further research is warranted to explore additional algorithms and datasets, with the goal of enhancing breast cancer classification outcomes.

Bibliography

1. K-Means Clustering Algorithm. [Online] Available: `https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning`.

2. AgglomerativeClustering. [Online] Available: `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html`.

3. The global k-means clustering algorithm. Paper: Likas, A., Vlassis, N., J. Verbeek, J. (2003). Pattern Recognition, 36(2), 451–461.

4. Random forest classifier for remote sensing classification. Paper: M. Pal (2005): Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26:1, 217-222.

5. Online training of support vector classifier. Paper: Lau, K. W., Wu, Q. H. (2003). Online training of support vector classifier. Pattern Recognition, 36(8), 1913–1920.

6. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. Paper: Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., Gifford, E. M. (2016). Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. Journal of Chemical Information and Modeling, 56(12), 2353–2360.

7. Machine learning for causal inference in Biostatistics. Paper: Rose, S., Rizopoulos, D. (2019). Machine learning for causal inference in Biostatistics. Biostatistics.