Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Analysis of the Basic Concepts and Applications of Machine Learning in the Medical Field

Wenxing Wang *

Beijing Royal School, Beijing, China

* Corresponding Author Email: michaelwangwx@gmail.com

Abstract. The application of machine learning in cancer prediction has important academic value and practical significance, and can make important contributions to improving the accuracy and efficiency of early cancer detection. The article explores the key concepts of contemporary machine learning, the processes and steps involved in data processing, and its application in cancer prediction. It particularly focuses on the accuracy of different machine learning models in predicting specific cancers and their characteristics in medical applications. Due to its ability to handle large-scale and complex data, machine learning has become a crucial tool for improving the accuracy of early cancer detection. The article provides a detailed introduction to supervised learning models (such as SVM, decision trees, logistic regression) and unsupervised learning models (such as clustering analysis, PCA), discussing their concepts and conditions of application. The article also describes the performance of these models in breast cancer classification tasks, finding that SVM performed the best in terms of accuracy. Despite the significant potential of machine learning in medical prediction, challenges such as noise and missing values in clinical data, model interpretability, and personalized prediction remain. Future research should focus on improving data preprocessing methods, enhancing model interpretability, and promoting its application in clinical practice.

Keywords: Machine learning; Application; Medical field.

1. Introduction

Cancer has long posed a severe threat to human health and safety. According to statistics from the U.S. National Cancer Institute, approximately 40.5% of men and women will be diagnosed with cancer at some point in their lives. In 2024, out of 2,001,140 newly diagnosed cancer cases, 30.57% of patients are expected to die from the disease [1]. Early treatment of cancer has a higher success rate, typically because the cancer is smaller and less likely to spread. However, some cancers exhibit no obvious symptoms or cellular characteristics in the early stages, making them difficult to detect. Such cases are often untreatable and lead to a painful progression of the disease. Therefore, the ability to accurately and efficiently predict cancer symptoms is of paramount importance to human health.

To improve prediction accuracy and efficiency, machine learning is being applied to cancer prediction. Due to its ability to handle large-scale and complex data, machine learning algorithms can process vast amounts of data in a short time, providing real-time or near-real-time predictive results to support clinical decision-making. Additionally, machine learning models can continuously update and learn from new data, thereby improving prediction accuracy over time. Currently, mainstream machine learning models are categorized into two types: supervised learning models and unsupervised learning models. Supervised learning models include Logistic Regression, SVM, Decision Tree, and Random Forest (RF), while unsupervised learning models include Artificial Neural Networks (ANN), Reinforcement Learning (RL), Ensemble Learning Models, and Natural Language Processing (NLP), among others. For instance, Stanford University's Kohler and his colleagues are working on improving the identification of high-risk breast cancer cases by using the C-path algorithm to learn from pathological specimens.

The goal of this article is to explore the accuracy of different models in predicting specific cancers and the characteristics of their medical applications. It includes an analysis of the features of

supervised and unsupervised learning algorithms, the performance analysis of these models on medical data, and an examination of the methods used for validating and testing the results.

2. Machine Learning

2.1. Concept

Machine learning is a significant branch of artificial intelligence. As the name suggests, it enables machines to learn from vast and complex datasets and improve their performance over time without the need for human intervention to provide complex code. This field utilizes sophisticated algorithms and statistical models to identify predictive models and extract unique insights from large datasets. Machine learning is primarily divided into three types: supervised learning, which involves training models on labeled data to make predictions; unsupervised learning, where the machine learns the features of the provided data on its own, categorizes it, and uses it to predict and classify future data; and reinforcement learning, which refers to the process where the machine experiments and learns through trial and error in a given environment, using a reward mechanism to evaluate and judge data to achieve optimal results.

Today, machine learning has made significant contributions to productivity in the medical field. Its powerful computing capabilities assist medical professionals in processing thousands or even tens of thousands of medical images, experimental results, and various case samples, enabling analysis and diagnosis that are more accurate. It can detect minute changes that may be difficult for humans to recognize (within an acceptable margin of error). Additionally, machines can use existing data to predict future new data, facilitating preventive measures, reducing healthcare costs, and saving manpower resources, thereby significantly lowering disease mortality rates.

2.2. Common Machine Learning Methods

2.2.1. Supervised Learning.

Support Vector Machine, abbreviated as SVM, is mainly used for classification tasks and regression analysis. The most basic method is to find a hyperplane that separates data in such a way that the different types of data points are as distinct as possible. In a two-dimensional space, the dividing line is a point; in three-dimensional space, it is a plane; and in higher dimensions, the hyperplane is divided into geometric objects with multiple parts. SVM plots are based on selecting several support vectors, i.e., data points, and calculating the shortest distance from these points to the line segment, known as the margin, to find the most suitable separation.

For instance, when analyzing the relationship between the benignity of cancer and the size of the cancerous region, the size variable can be represented as a straight line, with the two ends representing benign or malignant states. These data points are placed at any point on the line, requiring a dividing point to achieve classification.

These are linear conditions, but another non-linear tool called the Kernel trick allows SVM to plot non-linear dividing points, lines, and planes, greatly enhancing its ability to classify special or complex data models.

However, SVM also has its drawbacks. For example, it struggles with classifying overlapping data points, and its accuracy significantly drops with outlier data. It finds it challenging to handle these overlapping or imbalanced data points [2].

A decision tree is a tree-structured model that classifies data into different categories based on a series of rules. Each training session starts from the Root Node, which represents the entire population or object under study. Each split in the tree applies a decision criterion based on features in the data. The decision to split a node is made based on information gain, which is calculated as the purity of the parent node minus the weighted purity of all child nodes. Purity, also known as Gini impurity, describes the degree of mixed categories within a node. When a node contains only one category of

data, the Gini impurity is 0; when two categories are equally present, the Gini impurity is 0.5. The decision tree always selects nodes with higher information gain to split, and stops splitting when it reaches a node where impurity is zero, converting that node into a Leaf node to prevent overfitting. The decision tree calculates the probability of classification based on the proportions observed in previous splits, making the entire process straightforward and direct [3].

However, due to the large and complex nature of modern datasets, training a decision tree can be resource-intensive, requiring substantial computational power, making it difficult to be widely adopted by all medical professionals. Moreover, when the tree is overly fitted, overfitting occurs. In such cases, analysts can manually adjust parameters to control the decision tree, such as setting its maximum depth so that it stops splitting at a certain level and uses the already recorded results for predictions. This may lead to a loss in accuracy, so a trade-off between the two needs to be made.

Random Forest is an improvement upon the decision tree; it trains multiple trees simultaneously and averages or aggregates their final results to enhance prediction accuracy. This ensemble model provides a better fit for potential outliers than a single tree result. It also allows analysts to adjust parameters for multiple trees according to specific domain requirements [3].

It is one of the simpler methods in supervised learning. It classifies and predicts by calculating the distance between the input sample and the already labeled samples, identifying the corresponding k-nearest neighbors (a controllable parameter), and using the proportion of labels to make predictions. However, this algorithm has some clear drawbacks. If the features of the input sample are too similar or average, it becomes challenging to accurately predict the outcome. Additionally, outliers in the data and the choice of the k value can significantly affect the prediction of new data [4].

The logistic regression model is typically used for predicting and determining categories, and it is generally represented by an S-shaped curve. It can use various continuous parameters, such as age, weight, height, and some non-continuous data, such as the presence of certain diseases, as the basis for training and judgment. The model sequentially calculates probabilities for all labeled data, fitting a probability curve to the dataset by computing the probability based on similar features. Finally, a threshold parameter is manually set, allowing the machine to classify the data by comparing the calculated feature probabilities with the manually set threshold. However, this algorithm's accuracy decreases when analyzing data with significant overlap, similar to SVM, as it also requires clearly defined boundaries in the data [5].

Deep learning is a branch of machine learning, usually associated with neural networks. It allows machines to learn data in a way that mimics the human brain through countless neurons and layered connections. When training a neural network, the features of the original labeled data are input into the input layer, where they are processed through weights in the connections and passed into the neurons in the next layer. Each neuron in the hidden layers stores bias parameters, which are added to the data by default. These data are then fed into an activation function to determine whether a specific neuron should be activated. This process is called propagation. This continues until the output layer, where the probabilities for each outcome are calculated to determine the final prediction [6].

The results of the training are compared with the labeled data for validation. If the predicted results do not match the actual results, the model will send the error magnitude and required adjustments back through the network, a process known as backpropagation. Based on this feedback, the model adjusts the parameters and weights of the nodes and connections accordingly. This entire process is repeated until the model reaches the desired prediction accuracy [6].

2.2.2. Unsupervised Learning.

Cluster analysis is an exercise in categorizing data into groups. The basic operational logic is that objects in the same group (called clusters) are more similar in a particular sense than objects in other groups (clusters).

Clustering algorithms refer to a set of algorithms and tasks rather than a specific algorithmic model. These algorithms have very different understandings of what clustering is and how to find clusters efficiently. Common algorithms include datasets with small distances between members of a group, dense regions in the data space, intervals, or specific statistical distributions. Thus, clustering can be formulated as a multi-objective optimization problem. Suitable clustering algorithms and parameter settings depend on the particular data set and the intended use of the results. Clustering analysis is not a task that can be accomplished at any time, but rather self-corrects and adjusts its parameters as it learns from the data [7].

Centroid-based clustering, the center of mass of a cluster is the arithmetic mean of all points in the cluster. Center of mass-based clustering organizes data into non-hierarchical clusters. Center of mass-based clustering algorithms are efficient but sensitive to initial conditions and outliers. Among these algorithms, k-means is the most widely used. It requires the user to define the number k of centers of mass and works better when the clusters are approximately equal in size.

Density-based clustering connects consecutive high-density regions into clusters. This method can discover any number of clusters and any shape. Outliers are not assigned to clusters. These algorithms have difficulty in dealing with clusters of different densities and high dimensional data.

Distribution-based clustering, this clustering method assumes that the data consists of a probability distribution, such as a Gaussian distribution. As the distance from the center of the distribution increases, the probability that a point belongs to that distribution decreases. Banded areas show the decrease in probability. Unless the type of data analyzed will be distributed according to an apparently fixed model, other algorithms should be used when the underlying distribution type of the data is uncertain.

Hierarchical clustering is a common method of grouping objects. It creates clusters so that objects in one cluster are similar to each other and different from objects in other clusters. Clustering is visualized in a hierarchical tree called a dendrogram. This type of clustering has 2 main advantages: There is no need to pre-specify the number of clusters. Instead, the dendrogram can be cut at the appropriate level; the use of dendrograms makes it easy to organize the data into a hierarchical structure that makes it easy to examine and interpret the clusters [7].

Principal component analysis, is a statistical technique for unsupervised learning. It projects high-dimensional data into a low-dimensional space through linear variation, in other words, it is a process of integrating data containing multiple variables, the steps of this process are standardizing the data, calculating the covariance matrix, eigenvalue decomposition to finally selecting the principal components and constructing a projection that analyzes each variable according to the weight of its influence on the classification. The use of this technique allows complex and large data to be visualized in a simple way without loss of accuracy. However, what cannot be avoided in this process is the decrease in the precision of the analyzed data, as the integrated data package does not accurately reflect the influence of each variable with respect to the different variables. The precise pairs of different combinations have been finally selected as a result of the analyst's own decisions [8].

3. Data Pre-processing

3.1. Concept

Data preprocessing is an important step in ML data analysis, any type of processing of raw data in order to prepare it for data processing programs. Data preprocessing techniques are used to train machine learning models and artificial intelligence models and perform inference based on these models.

Data prep organizes data into a concise or integrated pattern for easier and more efficient processing in machine learning and other data science tasks. Examples include analyzing variable selection and

changing the format of subject variables. These techniques are typically used in the initial stages of machine learning and AI development to ensure that the results analysts need is obtained.

Real-world data is messy and often created, processed, and stored by a variety of humans, business processes, and applications. As a result, datasets may be missing individual key pieces of information, contain manual input errors, duplicate data, or describe the same thing under different names. These problems can be curtailed at the recording stage but cannot be dealt with completely cleanly.

Machine learning and deep learning algorithms work best when the data is presented in a format that highlights the relevant aspects needed to solve the problem. This can significantly reduce the processing power and time required to train or reason about new machine learning or artificial intelligence algorithms [9].

3.2. Methods

Data analysis is the process of examining, analyzing and reviewing data to collect statistics about their quality. It begins with a survey of the available data and its characteristics. The analyst identifies datasets that are relevant to the problem at hand, inventories their important attributes, and makes assumptions about features that may be relevant to the proposed analysis or machine learning task. Here are a few key approaches. Involving data organization, data transformation, data reduction, feature selection and feature scaling.

The goal of data cleansing is to find the easiest way to correct quality issues, such as eliminating bad data, filling in missing data, or ensuring that the raw data is suitable for feature engineering. Raw datasets often contain a lot of redundant data, generated when phenomena are characterized in different ways, or data that is not relevant to a specific ML analysis task. Data reduction uses techniques such as principal component analysis to transform the raw data into a simpler form suitable for a particular use case, reducing extraneous data, reducing predictive accuracy and improving computational efficiency. Analysts consider how to organize different aspects of the data to achieve the most meaningful goals. This may include structuring unstructured data, combining salient variables where it makes sense, or identifying important areas to focus on. Analysts apply various libraries of feature engineering to the data to achieve the desired transformation. The dataset needs to be organized to strike a balance between the training time for the new model and the amount of computation required.

Another point, the data is divided into two groups. The first group is used to train machine learning or deep learning models. The second group is the test data which is used to measure the accuracy and robustness of the generated models. This second step helps to identify any problems in the assumptions used in data cleaning and feature engineering. For example, the ANN also adjusts its own parameter weights based on the test data [10].

4. Model Performance Evaluation

In machine learning, model performance evaluation uses model monitoring to assess the performance of a model in a particular task. In model monitoring, there are various approaches to model evaluation using metrics such as classification and regression. During model development and testing, continuous evaluation and testing can identify issues such as data drift and model bias so that the model can be retrained to improve performance.

It will be discussed mainly from the point of view of a binary classification problem, where, let's say, we have to find out whether a patient has cancer (positive) or is healthy (negative). Common terms that need to be clarified are: True positives (TP), predicted positive and are actually positive; False positives (FP), predicted positive and are actually negative; and True negatives (TN), predicted negative and are actually negative; False negatives (FN), predicted negative and are actually positive. These parameters are presented as a matrix, also called a confusion matrix. These parameters are presented as a matrix, also known as a confusion matrix [11].

The accuracy can be calculated by Eq:

$$(TP + TN)/(TP + FP + TN + FN) \quad (1)$$

The formula allows you to calculate the percentage of the overall true value.

Percentage of positive instances to total predicted positive instances:

$$TP/(TP + FP) \quad (2)$$

The denominator here is the model that predicts positively from the entire given dataset. That is, "how many times the model is correct when it says it is".

Percentage of positive instances to the total number of actual positive instances:

$$TP/(TP + FN) \quad (3)$$

The denominator here (TP + FN) is the number of correct instances that actually exist in the dataset. Think of this as "how many additional correct instances the model misses when it shows correct instances".

Negative instances as a percentage of the total number of actual negative instances.

$$TN/(TN + FP) \quad (4)$$

Thus, the denominator here (TN + FP) is the number of negative instances actually present in the dataset. It is similar to recall, but the change is in the negative instances. It is also a measure of the degree of category separation.

It is the harmonic mean of precision and recall. It takes into account the contribution of both, so a higher F1 score is better. The connected variables are multiplicative, and if one of the scores is low, the final F1 score drops significantly. Therefore, if the predicted positive outcome is indeed positive (precision) and no positive outcomes are missed and predicted as negative (recall), then the model will have a high F1 score.

One drawback is that precision and recall are equally important, so depending on our application we may need one to be higher than the other, and the F1 score may not be an accurate measure. Therefore, weighting the F1 score or looking at PR or ROC curves can be helpful.

PR images show the curves of precision and recall at different thresholds. Here we can get high precision and high recall. Depending on our application, we can choose the predictor and threshold.PR AUC is just the area under the curve. The larger its value the better.

ROC represents the receiver operating characteristic, which is plotted against TPR and FPR for different thresholds. As TPR increases, FPR also increases. We expect the threshold to be closer to the upper left corner of the image. Comparing different predictors on a given dataset also becomes easy and we can choose the thresholds according to the application at hand. the ROC AUC is simply the area under the curve. the larger its value the better.

Since there are no TN relationship variables in the precision-recall equation, they are very useful in unbalanced classes. In the case of class imbalance, when negative classes are in the majority. The method does not take into account too much the large number of TRUE NEGATIVES in the negative classes that are in the majority, thus providing better defense against imbalance. This is important when detecting positive classes is very important.

For example, detecting cancer patients has a high degree of class imbalance because only a very small number of all diagnosed cancer patients have cancer. We certainly do not want to miss a cancer patient and not detect it (recall), but then we want to make sure that the detected person has cancer (precision). Since the TN or negative category is considered in the ROC equation, it is useful when both categories are important to us [11].

5. Specific cases of analysis

In 2020, Javed Mehedi Shamrat and Abu Raihan did a classification experiment for benign and malignant breast cancers based on ML. The dataset used in this study is the Wisconsin Breast Cancer The dataset used in this study is the Wisconsin Breast Cancer Dataset (WBCD), provided by the University of Wisconsin Hospital. The dataset contains 699 records, each record represents the diagnosis of a breast cancer patient, and each record contains 11 attributes, such as Clump Thickness, Uniformity of Cell Size, and Marginal Adhesion. Target variables in the dataset were used to categorize each case as benign (2) or malignant (4).

In data preprocessing, the authors performed several key steps to prepare the dataset for use by machine learning algorithms. First, the "ID" column, which is irrelevant for predictive analysis, was removed to avoid introducing bias in model training. In addition, the dataset contains some missing values (denoted by “?”) in the dataset, and the authors used the 'dropna()' function to remove these incomplete records. After cleaning, 683 records and 10 attributes remained in the dataset. To ensure that there is no significant multicollinearity between features (which may adversely affect model performance), the authors generated a correlation matrix.

Shamrat uses six ML models; SVM, LR, KNN, RF, DT (Decision Tree) and NB (Fourier Transform). And uses evaluates the performance of these models using several statistical metrics, including accuracy, sensitivity (recall), specificity, precision, and F1 value. These metrics provide a comprehensive view of the models' ability to correctly classify breast cancer cases.

The performance of each model was tested and compared by the above metrics. The results show that the SVM classifier performs best in breast cancer prediction with a classification accuracy of 97.07%. Plain Bayes and Random Forest models followed with 97% accuracy. KNN, Decision Tree and Logistic Regression models also performed well with accuracy close to 96%.

Plain Bayes performed the best in terms of sensitivity at 100%, indicating that it is most effective in identifying malignant cases. However, it has a lower specificity of 92%, meaning that it is more prone to false positives. The F1 values of all models exceeded 95%, indicating that they showed a good balance between precision and recall [12].

6. Challenges and Gaps

A common issue in medication field is that clinical data often contain noise and missing values, which affect the training and prediction accuracy of the model. Although researchers addressed missing values by mean imputation in the final results, this is not the optimal solution. More advanced data cleaning and preprocessing methods are still needed.

Additionally, the interpretability of the model remains a significant challenge. Despite the excellent predictive accuracy of deep learning and ensemble methods, their complexity makes the models difficult to interpret. This is a major obstacle for medical professionals who need to understand and trust the model's predictions. Better model interpretation tools, such as LIME and SHAP, need to be developed to improve model transparency and interpretability.

Moreover, most existing models are based on group data, making it difficult to fully account for individual differences and achieve truly personalized disease prediction and treatment. Patients' health conditions and lifestyles change over time, posing a challenge for dynamically adjusting prediction models to maintain high accuracy. Adapting multiple parameter variables while addressing

various individual characteristics and multimodal data integration remains an important research direction. Many studies remain theoretical, lacking practical clinical application and translation. It is necessary to strengthen the research on the application of machine learning models in clinical practice.

7. Conclusion

This paper introduces the mainstream models and algorithms in machine learning in detail. Supervised learning models have higher computational efficiency and accuracy compared to the unsupervised models, but inevitably, the distortion and complexity of the original data will lead to some analytical obstacles. Such obstacles can seldom be handled by the model itself, requiring the analyst to adjust the parameter weights or perform some processing on the data itself, including classification, reduction and dimensionality reduction. Supervised learning does not require labeling the data set and can be used to classify and analyze the data directly according to the algorithm. However, this process will require the use of more arithmetic power, as well as analysis after the need for the analyst's autonomous judgment and choose the type of summary. These algorithms are widely used in the medical field for processing large amounts of small discrepancy data, such as medical image analysis. These models can process a large number of data features, organize and document them in a short period of time. The accuracy is also quite impressive, reaching up to 97% accuracy. This greatly improves the clinician's diagnostic efficiency, resource utilization and time costs, and accuracy, allowing for better medical solutions. Model evaluation and data preprocessing are indispensable steps in machine learning, using these methods allows analysts to obtain accurate judgments on the reference degree of model data, and also provides engineers with a reasonably comprehensive perception of the accuracy of the model, so that models and algorithms can be better improved.

References

[1] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J.W.W. Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. European Journal of Cancer, [online] 49 (6), pp.1374 – 1403. (2013).

[2] D. Mustafa Abdullah and A. Mohsin Abdulazeez. Machine Learning Applications based on SVM Classification A Review. Qubahan Academic Journal, 1 (2), pp.81 – 90. (2021).

[3] D.J. Wu, T. Feng, M. Naehrig, and K. Lauter. Privately Evaluating Decision Trees and Random Forests. Proceedings on Privacy Enhancing Technologies, 2016(4), pp.335–355. (2016).

[4] M.-L. Zhang and Z.-H. Zhou. A k-nearest neighbor-based algorithm for multi-label classification. [online] IEEE Xplore. (2005).

[5] A.S. Hess and J.R. Hess. Logistic regression. Transfusion. (2019).

[6] J. Schmidhuber. Deep learning in neural networks: An overview. Neural Networks, 61 (61), pp.85 – 117. (2015).

[7] Comparing EM Clustering Algorithm with Density Based Clustering Algorithm Using WEKA Tool. International Journal of Science and Research (IJSR), 5 (7), pp.1199 – 1201. (2016).

[8] A. Maćkiewicz and W. Ratajczak. Principal components analysis (PCA). Computers & Geosciences, 19 (3), pp.303 – 342. (1993).

[9] N. Kumari. Data Management, Data Analytics, and Business Intelligence Can Assist in Process Management and Process Improvement Efforts. SSRN Electronic Journal. (2018).

[10] K. Maharana, S. Mondal, and B. Nemade. A Review: Data Pre-Processing and Data Augmentation Techniques. Global Transitions Proceedings, [online] 3(1), pp.91 – 99. (2022).

[11] F.I. Syed, T. Muther, A.K. Dahaghi, and S. Negahban. AI/ML assisted shale gas production performance evaluation. Journal of Petroleum Exploration and Production Technology, 11 (9), pp.3509 – 3519. (2021).

[12] M. Javed, A.K.M. Md. Abu Raihan, Sazzadur Rahman, I. Mahmud, and R. Akter. An Analysis on Breast Disease Prediction Using Machine Learning Approaches. International Journal of Scientific and Technology Research, 9 (2), pp.2450 – 2455. (2020).