Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/44/20230156

A new approach based on machine learning to certain diseases

Ruyi Teng[1,4,7], Tianqi Zhu[2,5], and Shuyan Qiao[3,6]

[1]School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[2]School of Economics and Management, Wuhan University, Wuhan, China
[3]Faculty of Information Technology, Monash University, Melbourne, Australia

[4]2991718317@qq.com
[5]efhddjk25678@163.com
[6]sqia0012@student.monash.edu
[7]corresponding author

Abstract. Never in history is the importance of data science so emphasized in modern society. Focusing on obtaining conclusive results from the implicit features concealed in a huge amount of data, data science plays a remarkable role in various fields, including modern medical practice. Although the fascinating performance of cutting-edge technology is capable of coping with numerous diseases, certain diseases, such as breast cancer and Parkinson's disease, still compromise people's health since these diseases are difficult to predict and prone to exacerbate. In order to deal with that problem, we will introduce three different machine learning methods in our experiment to two different data sets to test the performance of classification. In the paper, we clarified the principle of each machine learning method (three different classifiers) at first. Then, we conducted our experiment, during which decisive parameters of classifiers were set by specific searching algorithms. Besides, we introduced metrics along with their principles for the evaluation of the numerical results, which were obtained by different classifiers. In the next step, we discussed the results by comparing the values of the metrics that represent the performance of a particular method. Therefore, we managed to obtain optimal classifiers for the two datasets. In the final stage of the paper, we discussed our experiment's limitations as well as prospects, which includes further application in other fields.

Keywords: Breast Cancer, Support Vector Machines, Random Forest, Naive Bayes.

1. Introduction

Certain diseases like breast cancer and Parkinson's disease have been demanding tasks for human beings since their detection, and the healing processes are painstaking. As for breast cancer, it is the second most common cause of death among American women. Breast cancer is dreadful if not accurately detected in the early stages since it usually metastasizes to the lungs, liver, brain, lymph nodes, and bones, making it irreversible for female patients to recover [1]. Therefore, accurate classification and prediction are needed to provide an alternative method to determine whether a tumor is benign or malignant based on data science.

Parkinson's disease (PD) is a common degenerative disease of the nervous system, commonly existing among elderly people. Parkinson's has severe clinical features, such as static tremors, rigidity, and a masked face. The diagnosis of Parkinson's disease nowadays mainly relies on medical history,

DOI: 10.54254/2755-2721/44/20230156

clinical symptoms, and signs. The accuracy of the Parkinson's disease diagnosis is a key factor determining the patient's future life quality. Many reports have shown that a certain amount of misdiagnosed patients end up with much more severe Parkinson's symptoms. To cope with this situation, introducing a classification method to determine whether a patient is at high risk of developing Parkinson's disease is indispensable [2].

Nowadays, with the growth and expansion of information, data analysis plays an increasingly important role in daily life. Data classification, as an important branch of data science, can extract features from data sets based on high-dimensional data and classify and predict data based on these implicit features. Basically, the classification methods mainly include Naive Bayes, SVMs and Kernel SVMs, Decision Trees and Random Forest, as well as Multi-layer Perceptron and so on [3-6]. The most commonly used scientific metric for quantitative evaluation of the classification performance includes OA and Kappa coefficient, and so on.

The basis of data classification is the data itself. In order to make the experimental results more convincing, we chose two authoritative data sets, namely the breast cancer data set and Parkinson's data set. These data sets are derived from scientific medical research as well as accurate statistical methods. The breast cancer data set comes from the Diagnostic Wisconsin Breast Cancer Database, whose features are calculated from a digital picture of a fine needle aspirate (FNA) of a breast tumour and characterise properties of the cell nuclei shown in the image. The breast cancer database also concerns the judgment of whether the breast mass is benign or malignant. The National Centre for Voice and Speech, Denver, Colorado, and Max Little of the University of Oxford collaborated to establish the Parkinson's disease data set. Biomedical voice measurements form the basis of this data set. By recording a patient's voice for less than 30 seconds, we are capable of telling whether a patient has Parkinson's disease or not.

2. Methods of Data Analysis

In the report, we used and compared multiple data classification methods to analyze two data sets and evaluate the performance of different ways on the same data set. At the same time, the performance of these methods on the two data sets was compared to obtain valuable conclusions. To approach our topic, an exhaustive introduction of machine learning methods will be presented first to lay a solid foundation for a better understanding of our topic.

2.1. SVM

The Support Vector Machine (SVM) is a supervised learning method for outlier detection, regression, and classification. A support vector machine creates a hyperplane or collection of hyperplanes in high infinite-dimensional space that can be used for data sorting, data regression, and other tasks. The classifier's ultimate goal is to produce the best classification, achieved by choosing the hyperplane with the largest distance from any class's nearest training data points, also referred to as the functional margin [7-9].

Let's focus on the basic principle of the SVMs. First of all, the desired hyperplane for classification can be represented in the following formula:

$$\omega^{\mathrm{T}}\mathbf{x} + b = 0 \tag{1}$$

where ω is the vector of parameters that possesses the same order of dimension as the data, and b is a common value. Assuming that P_i $(x_1,x_2,x_3...x_n)$ represents the ith of the samples, which possesses n features, indicating that the data are n-dimensional. Then, the distance between a sample point and a hyperplane is:

$$d = \frac{|\omega_1 * x_1 + \omega_2 * x_2 + \omega_3 * x_3 + \cdots + \omega_n * x_n + b|}{\sqrt{\omega_1^2 + \omega_2^2 + \omega_3^2 + \ldots + \omega_n^2}} \tag{2}$$

Since our task is a 2-class classification problem, the model's input is the n features of the data set, while the output is the condition of the tumors (benign or malignant) for the breast cancer data set, as

DOI: 10.54254/2755-2721/44/20230156

well as for the Parkinson's case. For the prediction or the classification of the SVMs, the output y_i is either 1 or -1, representing the 2-class classification. The n-dimensional vectors which satisfy both the equations: $\omega^{\mathrm{T}}\mathbf{x} + b = 1$ and $\omega^{\mathrm{T}}\mathbf{x} + b = -1$ are defined as X_s or the support vectors. By introducing the objective function, we are, therefore, capable of clarifying the ultimate goal of SVMs in mathematical form. The objective function is shown below:

$$arg\max_{\omega,b}\left(min\left(y*\left(\omega^{\mathrm{T}}\mathbf{x}+b\right)\frac{1}{|\omega|}\right)\right) \tag{3}$$

where ω is an n dimensional vector and b is a common value, the infinite combination of these 2 parameters represents myriad sets of hyper-planes.

Since the distance of the data to the hyperplane is shown above, the distance in the SVM problem can be represented: $d = \frac{1}{|\omega|}$

Thus, to guarantee the rationality of the classification, for the samples whose output y_i is 1, the correct hyper-plane should let $b + \omega^{\mathrm{T}}\mathbf{x} \geq 1$, while for the samples whose output y_i is -1, the hyperplane should satisfy the condition $b + \omega^{\mathrm{T}}\mathbf{x} \leq -1$, and vice versa.

To make the 'min-max' form objective function easier to understand, it can be stated as:

$$\min_{\omega}\tau(\omega) = \frac{1}{2}*|\omega|^2 \tag{4}$$

$$s.t.\, y_i*\left(b+\omega^{\mathrm{T}}\mathbf{X_i}\right) \geq 1i = 1,2,3\ldots,N \tag{5}$$

Then, we introduce the Lagrange Multiplier Approach to obtain the extreme value of the objective function under the constraint condition.

$$L(\omega,b,\mathbf{a}) = \tfrac{1}{2}*|\omega|^2 - \textstyle\sum_{i=1}^{N} a_i\left(y_i*\left(\omega^{\mathrm{T}}\mathbf{X_i}\right)-1\right) \tag{6}$$

where $a_i \geqslant 0$.

This is the primal problem for solving this kind of convex quadratic form programming problem, whose solution is stated below.

$$\textstyle\sum_{i=1}^{N} a_i * y_i * \mathbf{X_i} = \omega \tag{7}$$

$$\sum_{i=1}^{N} a_i * y_i = 0 \tag{8}$$

By selecting L ($L \leqslant$ the total number of the samples) support vectors, the coefficient vector of the optimal hyper-plane ω can be obtained by the formula:

$$\omega = \sum_{i=1}^{L} a_i * y_i * \mathbf{X_i} \tag{9}$$

By selecting one of the L support vectors $\mathbf{X_s}$, a corresponding common value b can be obtained by $b = y_i - \omega^{\mathrm{T}}\mathbf{X_i}$. In order to achieve a more stable b, we can select more X_s values randomly to get the average number of b. So far, the process of solving the parameters of the hyperplane has ended, and the hyperplane has been determined.

With the determined hyperplane, the SVMs are capable of predicting the classification by using the constructed model. By inputting the sample vectors into the support vector machine, certain classification results will be obtained. Therefore, it provides another way to determine whether the tumor is benign or malignant, as well as deciding whether a patient has the risk of developing Parkinson's.

DOI: 10.54254/2755-2721/44/20230156

2.2. Decision Tree and Random Forest

The decision tree is a supervised machine learning method to predict and determine used for classification and regression. Decision tree models are similar to human logic of identifying characteristics of things and classifying them on the basis of interactive rules. Decision tree structures consist of nodes and directed edges. These nodes have two types: internal nodes and leaf nodes. An internal node represents a characteristic or feature, while a leaf node symbolizes a class. So, decision trees can be thought of as groupings of if-then loops. Based on this, test an instance attribute starting at the root node, then assign the instance to its child nodes based on the test results. Each child node at this moment represents a value for a feature. Recursively test and assign the instance in this manner up until the leaf node, and then assign the instance to the leaf node's class [10]. Above is the procedure for classifying instances using a decision tree. So, we can conclude that in decision trees, the computational complexity is reduced by pruning and feature selection, and the outcome is easy for us to understand.

Now, let's focus on the algorithms for dividing data sets, which should include choices of characteristics, generation of a decision tree, and pruning process of a decision tree. The first algorithm is called Iterative Dichotomiser 3(ID3) [11]. The ID3 algorithm starts with the data set D. For each characteristic C of D, we calculate entropy $H(D)$ and information gain $g(C,D)$. In the end, we compare all the information gains and select the characteristic that produces the largest $g(C,D)$. To be more specific, entropy $H(D)$ is a way to measure the uncertainty in a given dataset D.

$$H(D) = -\sum_{x \in X} p(x) log_2 p(x) \tag{10}$$

In the formula (11), $X \in R^n$ represents the set of classes in dataset D, where n represents the number of elements in the set X and x represents an individual element within set X. If $p(x)=0$, we define that $0 * log_2 0 = 0$ in this situation [12]. Finally, we can define $Gain(C,D)$ as the information gain of dataset D and one of its characteristics C, which represents the difference between $H(D)$ and $Ent(C,D)$, where $Ent(C,D)$ means the entropy of characteristic C in the given dataset D and we assume that characteristic C divides dataset D into T parts, $D_1, D_2, \ldots, D_T$:

$$Gain(C,D) = H(D) - Ent(C,D) \tag{11}$$

$$Ent(C,D) = \sum_{t=1}^{T} \frac{|D_t|}{|D|} H(D_t) \tag{12}$$

Finally, we calculate every information gain of each characteristic and select the largest one [13]. So that's all for the ID3 algorithm. In the following, let's concentrate on the CART algorithm. CART, short for Classification and Regression Tree, is a versatile algorithm that can be employed for both classification and regression tasks [14]. In order to explain how CART plays a role, we will begin with the concept: Gini coefficient [15]. In classification, assume that there are T classes and the probability that a sample point belongs to class t is p_t. The Gini coefficient of the probability distribution can be defined as:

$$Gini(p) = \sum_{t=1}^{T} p_t (1 - p_t) \tag{13}$$

For a given dataset D, it's Gini coefficient can be defined as:

$$Gini(D) = 1 - \sum_{t=1}^{T} \left(\frac{|M_t|}{|D|}\right)^2 \tag{14}$$

where T is the number of classes in data set D and M_t is the subset of samples in set D belonging to class t. Next, characteristic C divides dataset D into T samples, $D_1, D_2, \ldots, D_T$ when characteristic C has T values. Hence, we can define Gini coefficient of dataset D under the condition of characteristic C:

DOI: 10.54254/2755-2721/44/20230156

$$Gini(D,C) = \sum_{t=1}^{T} \frac{|D_t|}{|D|} Gini(D_t) \tag{15}$$

So, in the CART algorithm, we calculate the Gini coefficients corresponding to all characteristics and their sharps and select the minimum one [13].

When we apply decision trees to classify a group of training sets, we can collect the results of multiple decision trees and select the category with the most votes as the final prediction. Such a choice process is known as ensemble learning. Random forest is an ensemble learning method when training with the same algorithm [15]. It will construct many different decision trees and select the majority choice, like voting. To be specific, the logic behind random forests is to calculate the average of multiple overfitting decision trees with large variances in order to reduce variance and build a strong algorithm with better generalization performance and less overfitting as well as smaller variances. The algorithm can be summarized in three simple steps. First of all, repeatedly (T times) select x samples with replacement from the original training set X. Next, for each sample, complete the machine learning of a single decision tree and repeat T times, so we can produce T decision trees. Eventually, calculate T decision trees separately, get T results and use a simple majority voting mechanism to determine the classification of a given dataset D waiting to be classified [10].

In conclusion, random forests are less interpretive than decision trees, but still have several advantages. For example, the optimal parameter can easily be chosen [13]. In general, the more trees that we use, the better performances that random forests show, while the computational cost will increase accordingly. Mostly, random forests, bagging methods, and so on will not overfit with the increasing of the number of trees. As different datasets have diverse attributes and complexity, results of random forest may differ from each other. Further more, selecting different parameters and samples may also play a role in classification. In the passage, we will apply the method random forest and its special attribute to datasets like Parkinson and breast cancer and pay attention to the impact of datasets on the model results.

2.3. Naive Bayes

A given example that is described by its feature vector is given the most likely class using a Bayesian classifier. Assuming that characteristics are independent of class when learning these classifiers can considerably simplify the learning process, that is,

$$P(\boldsymbol{X}|C) = \prod_{i=1}^{N} P(X_i|C) \tag{16}$$

where $\boldsymbol{X} = (X_1,...,X_n)$ is a feature vector and is a class. In spite of the simplicity of Naive Bayes, it has been proved to perform well in plenty of general cases [16].

Among many specific Naive Bayes models, the Gaussian Naive Bayes algorithm for classification is the most commonly used. In the Gaussian Naive Bayes model, the likelihood of the all-independent variables is assumed to conform the Gaussian Distribution. Based on that, the parameters of the model are calculated by Maximum Likelihood Estimation. Gaussian Naive Bayes model is a basic model when the features can be supposed to follow Gaussian Distribution.

Another popular Naive Bayes classifier is the Multinomial Naive Bayes. This model is generally used in text classification, especially in tf-idf vectors [17]. The distribution is parameterized by several vectors for each class, where we can take into account the size of the corpus vocabulary, and the probability of each word appearing in a sample belonging to a class. The parameters of the model are trained by a smoothed Maximum Likelihood Estimation. The smoothing priors account for features which are not present in the training set, and it can avoid 0 probabilities in further computations. In summary, Multinomial Naive Bayes is a good choice when dealing with text data.

DOI: 10.54254/2755-2721/44/20230156

Complement Naive Bayes is similar to the standard Multinomial Naive Bayes algorithm while it is ameliorated to suit imbalanced data sets scenario [18]. It uses statistics from the complement of each class to compute the model's weights. Complement Naive Bayes are more stable than those for Multinomial Naive Bayes empirically. Further, Complement Naive Bayes generally performs better than Multinomial Naive Bayes on text classification tasks thanks to its normalization stability.

The fourth Naive Bayes model is the Bernoulli Naive Bayes. Bernoulli Naive Bayes model is designed for data which is distributed according to multivariate Bernoulli distributions. In this algorithm, each feature is supposed to be a binary-valued random variable. It performs well in cases of text classification on shorter documents using occurrence vectors.

Categorical Naive Bayes is used when the features are numerable [19]. In this model, the probability distribution of each feature is calculated based on the appearing times on samples with smoothness factors. This model differs from the others in enumerating representation. Out-of-core Naive Bayes is designed to handle large data sets which may not be loaded at a time. It works by processing the data in mini-batches and updating the model parameters partially. It can be used with any kind of Naive Bayes models mentioned above with partial fit classifier interface in scikit-learn. Therefore, Out-of-core Naive Bayes are generally used when working with giant data sets.

3. Experiments

3.1. Experiment Settings

After the brief introduction of the data sets as well as the basic principles of the three methods of data analysis, we conducted experiments on the two data sets, namely, Breast Cancer and Parkinson's Disease, by utilizing the three aforementioned two data sets. The outline of our experiment can be briefly described in several subsections, including a brief data set introduction, which mainly focuses on the features of the data in the respective data sets. Then, a more comprehensive introduction is presented along with the four particular metrics utilized during the evaluation process.

The core part of the experiment is to classify two data sets, namely SVM, Decision Tree, and Random Forest, as well as Naive Bayes. We introduced data preprocessing to enhance the performance of the classifiers as well as the plausibility of the experiment.

After data preprocessing, what is also important is to call the function of data splitting in order to divide the data into a train set and a test set. There are several reasons why machine learning divides data sets into training and testing sets: To start with is the prevention of overfitting. During the training process of a model, if all available data is used to train the model, it will fit the training data well, but may not perform well on unseen data. Splitting the data set into training and testing sets can prevent that phenomenon. Apart from that, another way to prevent overfitting is K-fold cross validation, which is capable of coping with the problems in the process of adjusting the hyper-parameters, aiming to guarantee the optimal models for each machine learning method in distinct data sets.

K-fold validation is the process right after data-splitting which divides the data set into train set and test set. K-fold divides the training set into K parts where (K-1) parts are utilized as training while the left one is used for verifying the performance of the model by calculating the variance of the estimation error. The next step is to duplicate the previous step for K times, until every k-fold has been used for validation. Then calculate the mean and standard deviation of model performance by obtaining the model scores calculated for all K models previously. Then, each hyper parameter required for the model optimization followed the same previous two steps to obtain their variance respectively. Therefore, the optimal hyper parameters set is determined.

As for the selection of K, conventionally we choose 10 as the value of K. With regard to the larger data sets, K is set to be lower, approximately 5, which can lower the cost of calculation. In general, K fold validation is recommended when the data set's scale is relatively low in order to deal with the potential risk of overfitting.

Second is to evaluate the performance of the model: dividing the dataset into training and testing sets can be used to evaluate the performance of the model. Utilizing test datasets to evaluate models can

DOI: 10.54254/2755-2721/44/20230156

provide a better understanding of their performance on unprecedented data. Another reason is to verify the generalization ability of the model: In machine learning, we attempt to construct a model that can generalize to unprecedented data. Splitting the dataset into training and testing sets can help us validate the model's generalization ability. If a model performs poorly on the test set, it is likely that its generalization ability to new data is also poor.

After the data splitting, our group advanced to selecting classifiers. By introducing the SVMs, Decision Tree Random Forest, and Naive Bayes from the sci-kit-learn library, we select three classifiers to deal with our breast cancer dataset and Parkinson's dataset and conducts experiments which uses different algorithms to fit the classifiers and selects several crucial parameters for each classifier to ensure the optimal property of the three classifiers [20]. Then, the classifier with the best parameters screened in the previous steps was called, and a program that generated four evaluation parameters was written to obtain a total of 6 sets of evaluation parameter groups for a total of 2 data sets. Not only the process of parameter selection is visualized to describe the process of adjusting parameters more vividly, but also the effect of different parameters on the total performance of one particular classifier is shown in the visualizing process. The final step is to obtain the total twenty-four metrics in the experiment and draw scientific conclusions from the numerical results.

3.2. Data Set Introduction

In the report, our group has conducted several experiments on the two aforementioned data set by 3 different data analysis method mentioned earlier. The breast cancer data set is consisting of several features including useless ones such as sample codes, (which should be omitted in the data prepossessing process since it is no feature of breast cancer) useful features like Clump Thickness, Uniformity of Cell Size, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei and so on. The classification of whether the breast mass is benign or not is reflected in the Class column. When the class column shows the number '2', it means that the breast mass observed in the experiment is malignant, while the number '4' means benign state of the breast mass. As for the Parkinson's data set, several features like MDVP:Fo(Hz), MDVP:Jitter constructs the main feature of the dataset and the output of the dataset, also plays the role of classifying the patient's status, the number 1 in the table represents that the patient is healthy, while the number 0 represents that the patient has Parkinson's disease.

3.3. Data Preprocessing

Data preprocessing is an indispensable process since the data in the whole dataset is not flawless. There are many duplicate or missing data. For example, many data units in the breast cancer data set are blank spaces, by which the performance of all three classifiers will be significantly impacted. The classification results and possible prediction functions will be severely affected, because these missing data sets can not reflect the characteristics of the data set, thus undermining the plausibility of experiment. By using codes in python environment which are capable of dealing data frames, the processed data rule out duplicate or missing data and can be used in the following steps.

3.4. Data and Metrics

3.4.1. The Breast Cancer Dataset. The breast cancer dataset comprises a total of 569 instances, each of which represents a unique case [21]. For every instance, there are 30 different features extracted from the FNA images, providing a comprehensive set of characteristics that can be used for analysis and prediction. These features include various measurements such as radius, texture, smoothness, compactness, symmetry, and fractal dimension, among others. The primary objective of this dataset is to develop models and algorithms that can accurately classify breast masses as benign or malignant based on the provided features. This classification task is of significant clinical importance, as early detection and diagnosis of breast cancer greatly influence the treatment and survival rates of patients.

DOI: 10.54254/2755-2721/44/20230156

3.4.2. The Parkinson's Disease Dataset. The parkinson's disease dataset provides valuable information extracted from voice recordings, allowing researchers to explore the relationship between specific vocal characteristics and the presence of Parkinson's disease [22]. These features may include measures such as jitter, shimmer, fundamental frequency, noise-to-harmonics ratio, and other acoustic properties related to speech patterns. The dataset is particularly significant in the field of medical diagnostics as it contributes to the development of predictive models and algorithms for early detection and diagnosis of Parkinson's disease. Parkinson's disease is a progressive neurodegenerative disorder affecting motor function, and accurate identification can greatly aid in providing appropriate medical interventions and improving patient care.

3.5. Algorithms and Parameters Setting

GridSearchCV is a ubiquitous method for searching the optimal parameter by using cross-validation [20]. The process of grid search is to adjust the parameters in steps within a given parameter range, train the classifier using the adjusted parameters, and find out the most accurate parameter on the test set from all parameter sets. Basically, it is a process of training and comparison. This method can ensure that the parameter with the highest accuracy is found within the specified parameter range, and the search efficiency for a single parameter is relatively high. However, for classification methods with multiple parameters, the time and resources required for searching in this way are geometrically increased, which is also the main drawback of GridSearchCV.

Each of the three classifiers uses the grid search method to obtain the most accurate classifier. To make the grid search process more vivid, our chart will involve four function lines where the evaluation parameters mentioned earlier vary with the change of parameters of each classifier.

3.6. Numerical Results and Discussion

3.6.1. Outline of Numerical analysis. For these three different classifiers, our experiment fixed a data set at first and then adjusted the parameter set by using GridSearchCV to find out the optimal parameter set for each of the three classifiers. Then, we used the four metrics mentioned earlier to demonstrate the performance of the three classifiers after parameter adjustment, draw conclusions, and discuss them.

3.6.2. Parameter Setting of the Three Classifiers Used in Breast Cancer Data Set. The first data set processed in the experiment is the breast cancer data set. For the three different classifiers, the analysis is the first classification method selected for this data set is the SVMs. Due to the significant influence of three parameters of support vector machines, namely the penalty coefficient c and gamma as well as kernel [20]. Gamma is also an important parameter used to control the influence range of the kernel function. The gamma parameter is mainly used when the kernel function is RBF function or polynomial function. For the RBF kernel function, gamma defines the influence range of a single training sample on the model. A smaller gamma represents a larger influence range. Features with relatively long distances between samples can also be considered, and the Decision boundary is relatively smooth. On the contrary, if gamma is relatively large, it means that the training model pays more attention to the area near the sample, and the more details the decision boundary will contain, leading to a more complex decision boundary. Additionally, the parameter called cv is set to be the defaulted as 5 which is suitable for the large-scale data set mentioned earlier.

As for the polynomial kernel, gamma defines the similarity of features in the space. A smaller gamma value corresponds to a higher similarity between features, resulting in a smoother Decision boundary. A larger gamma value means that the similarity between features is lower, and the decision boundary is more complex.

Due to the fact that the two parameters gamma and kernel are discrete string type parameters, where gamma has two choices of 'scale' and 'auto', the kernel parameter has four choices of 'RBF', 'polynomial', 'sigmoid', and 'linear', and the penalty parameter C is continuous, we choose to fix gamma and kernel parameters to obtain the optimal parameter set by adjusting C. The visualized

DOI: 10.54254/2755-2721/44/20230156

graph of parameters setting for breast cancer data set by using SVM classifier is shown below (see Figure 1):

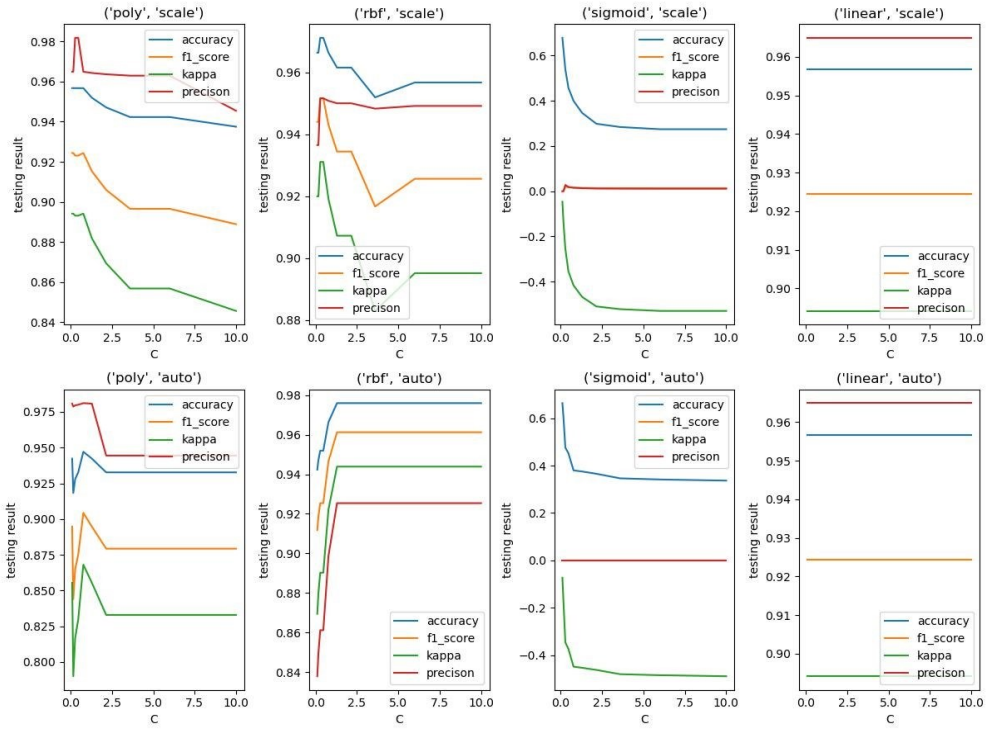

Figure 1. Changes of parameters in SVM.

The second classifier used in the analysis of breast cancer data set is random forest. Here we chose five parameters of random forest, known as n_estimators, criterion, max_depth, max_features and min_samples_leaf [20]. The n_estimators parameter represents the number of trees we choose in our forest and we test it in the range of 0 to 300. Parameter criterion is the way we deal with a decision tree. As we have discussed before, there are two methods to set a decision tree. One is to compare entropy, the other is to use Gini coefficient. So, the criterion has two choices of 'entropy' and 'gini'. Max_depth is the maximum depth of the tree and we test it in the range of 1 to 10. Similarly, min_samples_leaf represents the minimum number of samples we need to be at a leaf node. We choose the 1 to 10 numbers range. The last parameter max_features is the number of features we use when splitting. We select all odd numbers up to ten. In this experiment, we test the relationship between n_estimators and max_depth, the relationship between min_samples_leaf and criterion and the accuracy when parameter max_features changes alone.

The visualized graphs of parameters setting for breast cancer data set by using random forest classifier are shown in Figure 2, Figure 3 and Figure 4:

DOI: 10.54254/2755-2721/44/20230156

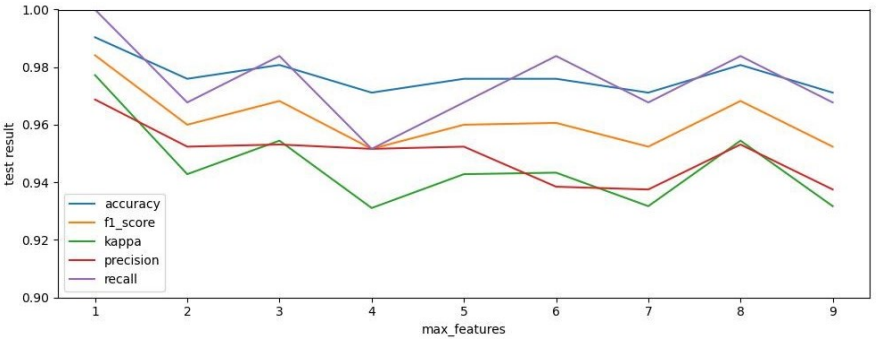

Figure 2. Changes of max_features in random forest.

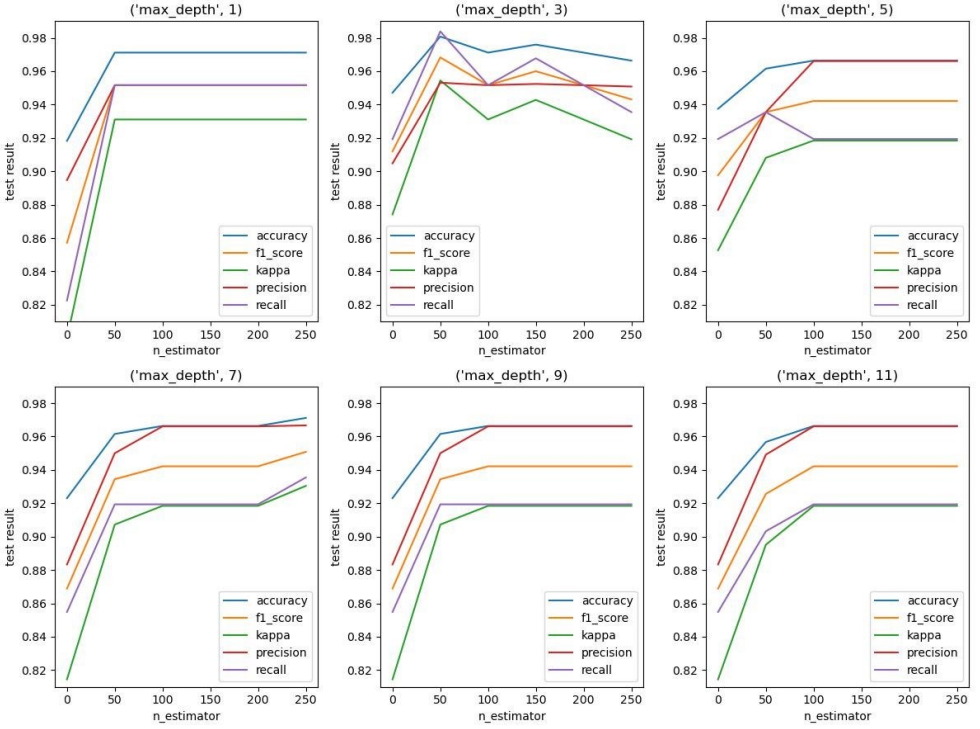

Figure 3. Changes of n_estimators and max_depth in random forest.

DOI: 10.54254/2755-2721/44/20230156

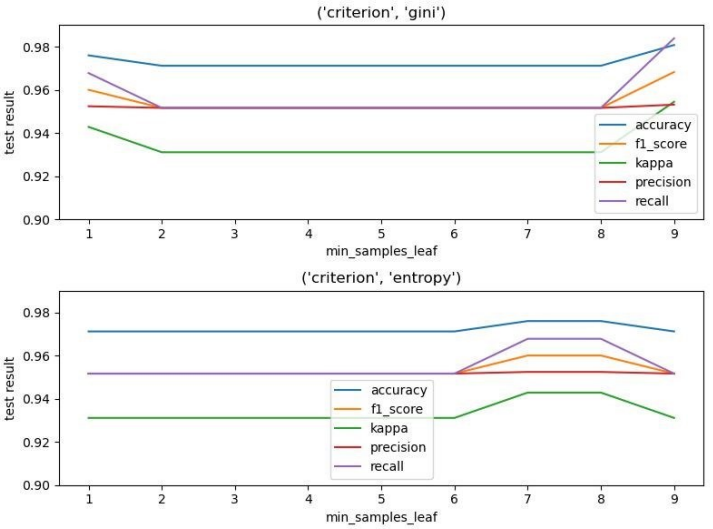

Figure 4. Changes of min_samples_leaf and criterion in random forest.

The last classifier for classification for the first data set is naive bayes. Since judging whether certain breast cancer is benign or malignant is a binary classification process, we choose Bernoulli Naive Bayes model. Due to the significant influence of two parameters of Bernoulli Naive Bayes, namely the coefficient alpha and binarize [20]. The alpha parameter is a Laplace or Leadstone smoothing. If it is set to 0, then it means no smoothing option at all. However, it should be noted that smoothing is different from artificially adding some noise to the probability, so the larger the setting, the lower the accuracy of the computational naive Bayesian (although the impact is not very large), and the Bool score gradually increases higher. The parameter binarize is the threshold for binarizing the feature. If it is set to None, it will be assumed that the feature has been binarized. We use every 0.1 interval point from 0 to 1 as the value of alpha, the same as binarize.

The visualized graphs of parameters setting for breast cancer data set by using Naive Bayes classifier are shown in Figure 5 and Figure 6:

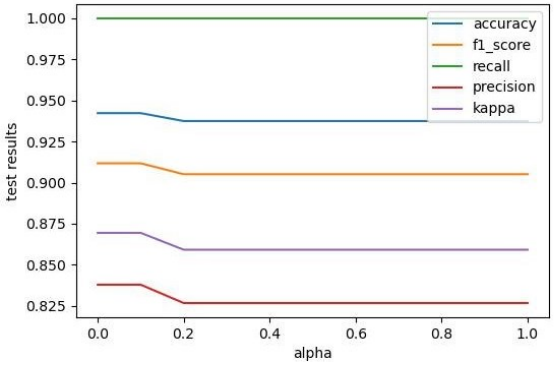

Figure 5. Changes of alpha in Naive Bayes.

DOI: 10.54254/2755-2721/44/20230156

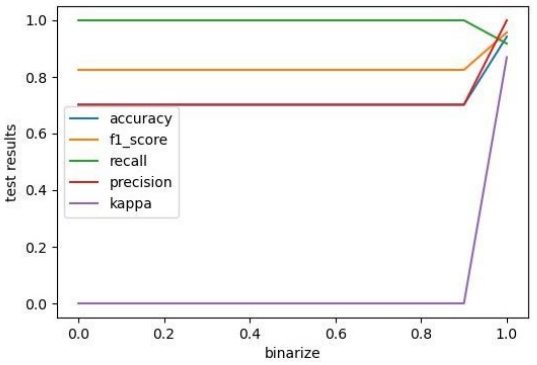

Figure 6. Changes of binarize in Naive Bayes.

It seems that when there is no smoothing option at all(alpha=0) and binarize equals 1.0, we have the best testing score.

3.6.3. Numerical Results of the Breast Cancer Data Set. After the crucial process of parameter setting for each of the three classifiers, what determines the performance on the breast cancer data set is the value of Accuracy, Recall, F1 Score, Kappa Coefficient. In our experiment around breast cancer data set, we obtained these metrics and the results are comprehensively reflected in Table 1 shown below:

Table 1. Numerical Results of Breast Cancer Data Set

	Accuracy	Precision	Recall	F1 Score	Kappa Coefficient
Kernel SVM	0.9663	0.9508	0.9355	0.9431	0.9192
Random Forest	0.9711	0.9516	0.9516	0.9516	0.9311
Naive Bayes	0.9423	0.8378	1.0000	0.9118	0.8694

3.6.4. Discussions of the Numerical Results of the Breast Cancer Data Set. Before the analysis of using the five metrics, we should sort out the most important metric in our experiment since all the metrics mentioned earlier serves for different purposes in disparate fields. Since our experiment aims to provide an alternative way of classifying and predicting on the basis of data science before medical diagnosis. Therefore, as for the two statistical errors, FN (false negative) is unbearable, because FN indicates that the model could have falsely predicted those patients who actually have the disease as healthy, therefore preventing them from receiving further medical examination. This circumstance may cause heavy pressure to the patient as well as heath care centre, indicating that false negative should be taken into consideration primarily.

Recall, the metric concerning FNs the most should be taken into consideration primarily. The second important metric should be F1 Score, which is the harmonic mean of Precision and Recall. The next metric should be Precision, since our experiment can bear the circumstance of FPs, which means that the patients who are healthy but diagnosed as potential breast cancer patients can do further medical examination to determine their health state [23].

From the list provided in the paper, Naive Bayes has an outstanding performance, an exciting Recall of 1 represents the classifier's ability of precluding false negative cases. Although the F1 score of the Random Forest is higher than the Naive Bayes along with Precision and Accuracy, it could prove that Naiye Bayes sacrifices these two metrics to preclude false negative, Naive Bayes should be the top choice among the three classifiers. The second-best performance is Random Forest, with same Recall and F1 score as well as Precision of the value of 0.9516, and its all metrics are superior to Kernel SVM, making Kernel SVM the last classifier when it comes to breast cancer data set.

DOI: 10.54254/2755-2721/44/20230156

3.6.5. Parameter Setting of the Three Classifiers Used in Parkinson's Data Set. For the Parkinson's disease data set, the primary step is parameter setting just like what we have done concerning the analysis of breast cancer data set. Similarly, in SVM, we select penalty coefficient c, gamma and kernel. In random forest, we choose five parameters: n_estimators, criterion, max_depth, max_features and min_samples_leaf. In Naive Bayes, we use Bernoulli method and select binarize and alpha [20].

3.6.6. Numerical Results of the Parkinson's Data Set. In our experiment around Parkinson's disease data set, we obtained these four metrics and the results are comprehensively reflected in Table 2 shown below:

Table 2. Numerical Results of Parkinson's Disease Data Set

	Accuracy	Precision	Recall	F1 Score	Kappa Coefficient
Kernel SVM	0.8814	0.8776	0.9773	0.9247	0.6485
Random Forest	0.9492	0.9362	1.0000	0.9670	0.8564
Naive Bayes	0.8136	0.8113	0.9773	0.8866	0.3872

3.6.7. Discussions of the Numerical Results of the Parkinson's Data Set. With regard to the Parkinson's disease data set, Random Forest stands out to be the best classifier. It has the best recall value as well as the rest four metrics. A very high rate of recall indicates that the classifier is capable of precluding the possibility of erroneously diagnosing patients who already develops PD as healthy. A high rate of f1 score represents its outstanding ability of reconciling precision and recall, therefore the model's capability of avoiding false positive and false negative which is more important in disease prediction. The Kernel SVM's performance is in the middle, since it has more modest Recall and Precision than Random Forest, along with Kappa coefficient and other metrics. Naive Bayes failed to continue its excellent performance in the breast cancer data set, with a recall of merely 0.9773. In addition, its kappa coefficient is quite low, making it the last choice for the classification and prediction of PD data set. Compared with the results of the Breast Cancer data set, Parkinson's data set has better recall value and higher rate of f1 score. There is an imbalance in the distribution of categories in the Parkinson's data set, due to the small proportion of classification category in Parkinson's data set and the model may tend to predict a large number of categories, thereby increasing the recall rate, but may reduce the accuracy rate and Kappa coefficient. In the same time, different optimal parameters setting intervals of the two data sets will also affect the results to a certain degree.

For the naive bayes classifier, it is based on the conditional independence between features, this assumption may not always hold true in practical situations. For the Parkinson's disease data set, the correlation between features are stronger than that of the breast cancer's. Certain features in the Parkinson's disease data set like the MDVP:Jitter(%) and MDVP:Jitter(Abs) are in strong correlation while most features in breast cancer data set are independent with each other. The numerical results indicates that the performance of naive bayes in breast cancer data set is much better than Parkinson's disease data set, which means Naive Bayes is suitable for classification problems with discrete features and satisfying the assumption of conditional independence between features, especially in fields such as text classification. However, in cases where there is strong correlation or continuous features between features, naive Bayes may not be the best choice.

Kernel SVM has different performances in the two data sets. Kernel SVM utilizes kernel trick to project low dimensional data into high-dimensional space, where non separable data in low dimensional space becomes more discrete and separable when projected into high-dimensional space. Low dimensional data itself contains less information, and it is difficult to ensure that the sample can be discretized after projection. Therefore, it can only be separated when projected to higher dimensions. Therefore, high-dimensional data contains more information, resulting in larger differences between samples, making it easier to achieve separability in relatively low dimensions after projection. Although kernel SVM is suitable for those complex small to medium-sized data sets, the complexity, dimension

DOI: 10.54254/2755-2721/44/20230156

as well as the linearity of breast cancer data set is higher than that of Parkinson's data set, which indicates that linear kernel SVM tends to be more suitable in coping with breast cancer data set, corroborating our numerical results.

4. Conclusion

In conclusion, the experiment around the three classifiers on the breast cancer data set and the Parkinson's disease data set has shown that although both of the three classifiers all have their applicable fields, there always exist an optimal solution. For the breast cancer data set, the best classifier is Naive Bayes classifier, who guarantees the least possibility of False negative. On the other hand, Random Forest is the optimal choice with regard to the Parkinson's disease due to its outstanding performance in all the five selected metrics.

In addition, the difference between the models also contribute to the different performance in the same data set.

So, it is clear that optimal classifier varies between different data sets. The difference between data sets, including the amount of the data, the dimension of the data, the feature of the data may cause completely different results. In order to cope with the problem, additional work considering data preprocessing like PCA (Principal Component Analysis) is expected. Other necessary works such as normalization process to reduce noise, introducing more parameters along with more efficient parameter setting methods are beneficial to the future performance of the classifiers.

Furthermore, our experiment merely took two data sets into consideration, which limits the practical use of the classifiers developed during the experiment. The application prospect of the classifiers should not be confined to classify and predict breast cancer and PD, instead, it could be extended to multiple branches in multiple fields. The classifiers can be applied to brain imaging, red blood cell carcinogenesis around medical image segmentation, or other branches such as credit evaluation, risk assessment in financial industry, quality control, fault detect in industrial sector. The future of data classification and prediction is undoubtedly promising.

Acknowledgement

Ruyi Teng and Tianqi Zhu contributed equally to this work and should be considered co-first authors.

References

[1] T. J. Key, P. K. Verkasalo, and E. Banks, "Epidemiology of breast cancer," *The lancet oncology*, vol. 2, no. 3, pp. 133–140, 2001.

[2] W. Poewe, K. Seppi, C. M. Tanner, *et al.*, "Parkinson disease," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–21, 2017.

[3] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565– 1567, 2006.

[4] A. Patle and D. S. Chouhan, "Svm kernel functions for classification," in *2013 International Conference on Advances in Technology and Engineering (ICATE)*, IEEE, 2013, pp. 1–9.

[5] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: From early developments to recent advancements," *Systems Science & Control Engineering: An Open Access Journal*, vol. 2, no. 1, pp. 602–609, 2014.

[6] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627– 2636, 1998.

[7] G. Zhang, "A modified svm classifier based on rs in medical disease prediction," in *2009 Second International Symposium on Computational Intelligence and Design*, vol. 1, 2009, pp. 144–147. doi: 10.1109/ISCID.2009.43.

[8] M.-P. Hosseini, M. R. Nazem-Zadeh, F. Mahmoudi, H. Ying, and H. Soltanian-Zadeh, "Support vector machine with nonlinear-kernel optimization for lateralization of epileptogenic hippocampus in mr images," in *2014 36th Annual International Conference of the IEEE*

DOI: 10.54254/2755-2721/44/20230156

Engineering in Medicine and Biology Society, 2014, pp. 1047–1050. doi: 10.1109/EMBC.2014.6943773.

[9] R. Li, K. Cui, R. H. Chan, and R. J. Plemmons, "Classification of hyperspectral images using svm with shape-adaptive reconstruction and smoothed total variation," in *IGARSS*, IEEE, 2022, pp. 1368–1371.

[10] H. Li and K. Yamanishi, "Text classification using esc-based stochastic decision lists," in *Proceedings of the eighth international conference on Information and knowledge management*, 1999, pp. 122– 130.

[11] J. Quinlan, "Induction of decision trees. mach. learn," 1986.

[12] R. Balian, "Entropy, a protean concept," *PROGRESS IN MATHEMATICAL PHYSICS*, vol. 38, p. 119, 2004.

[13] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.

[14] B. Ripley, "Pattern recognition and neural networks cambridge university press cambridge," *UK Google Scholar*, 1996.

[15] S. Raschka, *Python machine learning*. Packt publishing ltd, 2015.

[16] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, 2001, pp. 41–46.

[17] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, Citeseer, vol. 242, 2003, pp. 29–48.

[18] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, Springer, 2005, pp. 488–499.

[19] P. Fabian, "Scikit-learn: Machine learning in python," *Journal of machine learning research 12*, p. 2825, 2011.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in *2018 International Conference on Robots & Intelligent System (ICRIS)*, IEEE, 2018, pp. 157–160.

[22] S. Arora, V. Venkataraman, A. Zhan, *et al.*, "Detecting and monitoring the symptoms of parkinson's disease using smartphones: A pilot study," *Parkinsonism & related disorders*, vol. 21, no. 6, pp. 650– 653, 2015.

[23] H. Zhu, X. Liu, R. Lu, and H. Li, "Efficient and privacy-preserving online medical prediagnosis framework using nonlinear svm," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 838–850, 2017. doi: 10.1109/JBHI.2016.2548248.