Use Multifarious Classifiers to classify the Anesthetic Questionnaire Results

May 9, 2023

Abstract

Artificial intelligence is currently a galloping and hot topic today and the scientific community. It is becoming apparent that artificial intelligence can be used in various fields, including mapping, analysis, data processing, image processing, and more. And as one of the most prevalent of AI, machine learning undoubtedly provides the theoretical approach and practical basis for implementing AI. In this course, INT104, the main content is a few of the more popular methods in machine learning to accomplish data processing. In this presentation, the author will "clean" and process the given data based on what is taught in the class, more on the author's learning and thinking inside and outside the classroom. The steps involved are broadly dimensionality reduction, classification, and clustering. Due to the diversity of the different parts of the method, the author will use Python, a standard programming language in machine learning, to analyse the mathematical logic of a few pre-selected techniques and select the most appropriate one based on the results of several tests and the purpose of the data processing, to clarify its rationality, describe the process and draw a picture of the results. The results will be illustrated, the method described, and an image will be produced.

1. Introduction

Currently, medical care uses much machine learning knowledge, such as image processing of lesions, genetic testing, disease prediction, etc. According to machine learning techniques, based on the high speed, unbiased and empirical nature of the machines, they can be trained based on their large amount of data and reflect the results more objectively. The comfort care and targeted treatment it provides can significantly benefit patients. In this experiment, the anaesthetic approach to today's drugs still requires individualised feedback. Different patients have different individual situations and treatment goals, so a questionnaire with 15 questions containing yes, no, and other answers was set up. After collecting more than 5,000 questionnaires and scoring the options for each questionnaire, the results allowed for the patient's most appropriate type of anaesthesia. Specifically, the mathematical process and some detailed principles will be shown in the appendixes.

2. Methodology

The authors worked broadly on data dimensionality reduction, classification, and clustering in this assignment. After several accuracy-oriented experiments, the methods used for each part were finalised. In particular, in data dimensionality reduction, the Principle Component Analysis (PCA) method was used for cleaning after comparing the variance ratios; in classification, various models of classifiers (logistic regression, naive Bayesian, SVM, etc.) were tested and measured according to different validation methods, and the final decision was made to use the logistic regression (LR) algorithm as the classifier to process the dimensionality reduced data and use cross-validation as the validation techniques, after observing the results obtained using different clusters, the final decision was made to use K-Means for clustering the data.

3. Data Analysis

3.1. Procedure Reasons

Here, the data needs to be analysed first. Because it serves as the infrastructure for this assignment, any changes or errors can significantly impact the following work. The authors have therefore decided to prioritise the analysis of the data. Then decide on the purpose of the next step.

3.2 Data Observation

First, observe the data distribution, which has 5345 rows, of which the first row is the introduction, 17 columns, the first column is the index, and the last column is the label. Then observe the type of data containing numerous blocks of 0, 1, and 2 in total, with 2 being the exception data as required. Returning to the provided task sheet, it states that labels containing "2" can be deleted directly but does not clarify how blocks containing 2 in the specific data are to be manipulated. This brings us to the difference between using a binary classifier and a multi-classifier. That is, whether the 2's in the block are treated as a separate class. In general, multiclassification will be a little more complex than a classifier because multi-classifiers require more performance from the system, and decision boundaries can be a little more rigid to carve out than classifiers [1]. In this regard, it needs to be tested using different classifiers after dimensionality reduction, with the highest accuracy as a guide to decide which class of classifier to use.

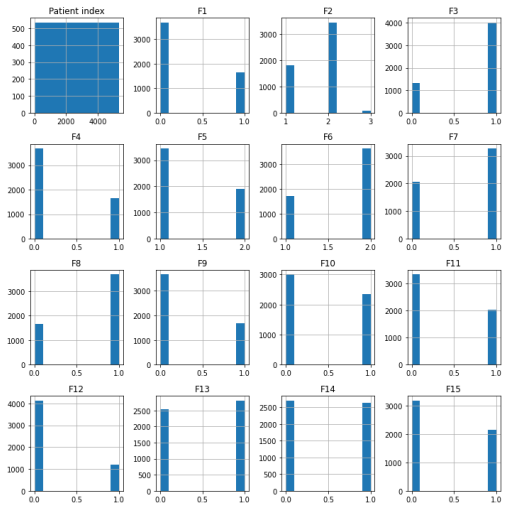

Fig.1 Concrete data distribution

3.3 Specific Operations

After observation, 14 rows of data are marked with 2 in the label column, with a ratio of (14/5344) *100% = 0.26198%. As they are anomalies and the proportion is low, they can be eliminated according to the requirements of the task sheet above. After selecting the dimensionality reduction method, given a dataset, multiple candidate classifiers were selected, followed by testing their accuracy (focusing on dichotomous and multi-classifiers) [1] and finally deciding on the final choice.

4. dimensionality reduction

Multiple dimensionality reduction algorithms are based on scikit-learn, each adapted to a different situation. Therefore, the authors will determine the specific scientific and justified dimensionality reduction method based on experimental results. Based on the more popular dimensionality reduction algorithms, and in conjunction with what has been learned in class, I will then compare the advantages and disadvantages of these dimensionality reduction methods in the order of Principal Component Analysis (PCA), Singular Value Matrix Decomposition (SVD), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA). First, I will explain the reasons for choosing these dimensionality reduction methods as a pre-selection option and discarding other classifiers. Then, I will compare the advantages and disadvantages of these methods. Since this is currently a task-oriented approach, the authors will infer the optimal choice of the downscaling method after combining the accuracy of the classifications.

4.1 Reasons for Dimension Reduction

Dimensionality reduction refers to reducing the feature quantity or dimensionality of features in a high-dimensional data set. The impact of noise and redundancy in the data is more severe when the number of dimensions is high. Reducing the dimensionality of data can, on the one hand, provide a quality database for subsequent classification and improve accuracy; on the other hand, a suitable data reduction process makes computer processing less cumbersome and faster. Its purpose can vary depending on the application and needs [2].

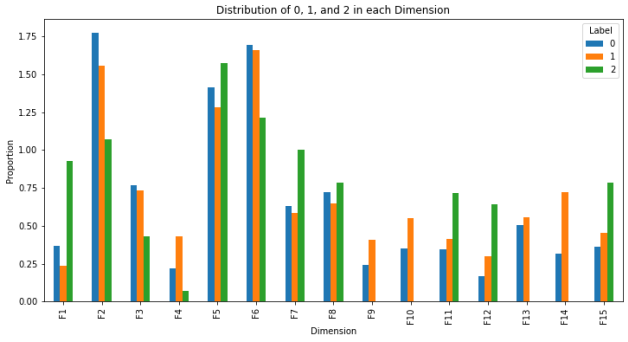

Fig.2 Data distribution in the fifteen dimensions

4.2 Pre-selection requirements

Based on the previous data analysis and the requirements described in the task sheet, the data is binary distributed (mostly 0 and 1). Based on this data characteristic, the performance of the classifier using a Gaussian database may be reduced due to the dimensionality of 15 dimensions and the fact that the data within the features are discrete and do not conform to a strict Gaussian distribution; at the same time, it should be more effective based on the use of non-linear dimensionality reduction methods.

4.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most extensively used data dimensionality reduction algorithms. It transforms the original data into a set of linearly uncorrelated dimensions through a linear transformation that reduces data dimensionality. It extracts the main features of data and is commonly used in the dimensionality reduction of high-dimensional data. It can reduce the number of features while retaining the primary information in data [3]. The basic principle is to project the data into a new coordinate system by finding the directions in the data with tremendous variance. These directions with the highest variance are principal components and are linearly independent combinations between the original features.

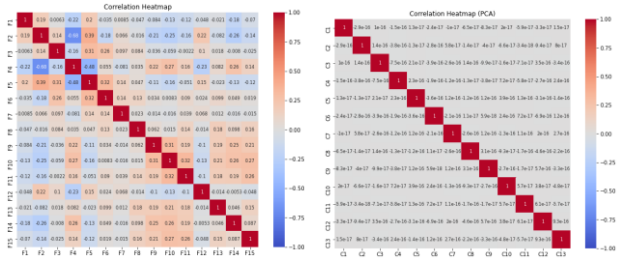

Fig.3, 4 Correlation heatmap before(left) and after(right) PCA

4.3.1. Advantages and drawbacks of PCA

Advantages of the PCA algorithm:

1. Simpler to operate.
2. Reduces the computational overhead of the algorithm.
3. Easy to remove noise.
4. Underlying mathematical logic is strong.

Disadvantages:

1. The eigenvalue decomposition has some limitations; the transformed matrix needs to be square.
2. The result of PCA may not be optimal in the case of non-Gaussian distribution.

4.4 Singular Value Decomposition (SVD)

The Singular Value Decomposition algorithm (SVD) is a standard matrix decomposition method used to reduce the dimensionality of data, extract key features, and solve problems such as systems of linear equations. Similar to PCA, both PCA and SVD are statistically based dimensionality reduction methods for extracting the essential features from high-dimensional data. Both can project the original data into a low-dimensional space by a linear transformation.

However, SVD has an advantage: the dimensionality reduction results of PCA are orthogonal to the principal components and have a better interpretation [4]. In contrast, the dimensionality reduction result of SVD contains singular vectors, which may not be entirely orthogonal but has more excellent numerical stability. Also, PCA is usually used to deal with covariance matrices and is suitable for data with linear relationships. SVD can deal with arbitrary matrices and is suitable for data with non-linear relationships.

4.5 Comparison of PCA and SVD

As both dimensionality reduction algorithms are based on changes to the matrix, they can both be evaluated using the 'cumulative variance ratio', which is a measure of the variance of the data captured by PCA by mapping the original data to the principal components. A high proportion of cumulative variance means that the dimensionality reduction still retains more of the variance of the original data, implying that the dimensionality reduction results are better able to explain the variation in the data. In data processing, the closer the cumulative variance ratio is to 1, the better the dimensionality reduction effect is. In general, we want the cumulative variance ratio to be more excellent than 80% or even 90%.

Through Python code and drawing, then know that PCA is slightly better than SVD below 14 dimensions, and at 14 dimensions, the two reach almost the same. However, since the SVD algorithm can only downscale to less than one dimension of the data, the cumulative variance ratio of PCA is greater than that of SVD at dimension 15. At dimension 15, the cumulative variance ratio of PCA reaches 1.00, but to prevent overfitting, a dimension with a cumulative variance ratio of around 0.9 is eventually chosen. Therefore, for these two similar algorithms, to achieve optimal dimensionality reduction results, the authors chose PCA and discarded SVD and chose to drop to dimension 13.

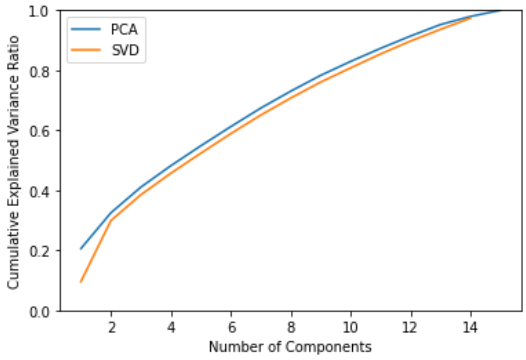

Fig.5 Comparison of PCA and SVD in cumulative Explained Variance Ration

4.6. Linear Discriminant Analysis (LDA)

The basic idea of linear discriminant analysis is to minimise the inter-class scatter of samples in the resulting subspace while maximising the intra-class scatter [4]. In terms of substantial completion, a set of linear transformations is performed for the samples, and then samples from the same class are clustered together, and those from different classes are taken apart [5],[6]. Its high separability within the data is well suited to such 0,1 binary distribution data. Therefore, LDA is used as a dimensionality reduction alternative. The high-dimensional pattern samples are projected into the optimal discriminative vector space. Unlike PCA, an unsupervised learning method, LDA is a supervised dimensionality

reduction method.

4.6.1 Strengths and weaknesses of LDA

Strengths:

1. LDA performs better for the sample classification information relies on the mean rather than the variance.
2. Prior knowledge experience of the category can be used in the dimensionality reduction process.

Weaknesses:

1. LDA descends to a maximum of k-1 dimensions of the category number and cannot be used if the dimensionality after the **dimensionality reduction is greater than k-1**.
2. LDA may over-fit the data.
3. When the data depends on variance, dimensionality reduction could be more effective.

4.7 Reasons for Excluding LDA

It is worth acknowledging that LDA is more compatible with the need for data dimensionality reduction and the form of the data. When the data is more discrete, using LDA, mapped in a particular dimension, can better guarantee its higher separation, for example, a higher classification effect. Also, after the authors' verification, during which they controlled for its accuracy and performance using the same classifier and the same number of folds, it was slightly better than PCA by about 2.5 percentage points on average.

SVM main accuracy for LDA by cross-validation: **72.420%**.

However, the setting of its parameters needs to be observed. One noticeable difference between LDA and PCA in terms of how it treats data dimensionality reduction is that LDA reduces to a maximum of k-1 dimensions of the number of categories, whereas PCA does not have this limitation. Specifically, in this case, LDA divides the data into two categories, that is, 0 and 1. Therefore, the LDA dimensionality reduction has and can only be reduced from the original 15 dimensions to 2-1 = 1 dimension. The authors believe that if the data is reduced from 15 dimensions to more than half (around eight dimensions), it can be very good at reducing resource consumption and improving accuracy [7]; However, using LAD to minimise the data directly to 1 dimension would be very informative for the data to lose. The accuracy may be inflated (since the training resources are only for 1-dimensional data, the training accuracy will be significantly improved), and there may be overfitting, which is detrimental to the model's training.

Therefore, for a method like LDA, with such a large amount of data loss, even with a high accuracy rate, the LDA algorithm will be prioritised as far back as possible on balance.

4.8 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is also a scikit-learn-based method for dimensionality reduction. Its strategy for finding implicit factors or components from multidimensional statistics extends principal component analysis (PCA) and factor analysis (FA). It assumes that the subcomponents are non-Gaussian signals and are done statistically independently of each other. Similar to the PCA algorithm, a linear transformation and ICA pre-process the data in much the same way as PCA before dimensionality reduction. For example, centroiding (subtracting the mean) is very similar. Therefore, the advantages of ICA will not be presented here, but only some of its features.

One of the most important features is how different data are treated: Because ICA assumes that components are statistically independent, it is well adapted to data containing non-Gaussian distributed details. Furthermore, ICA can separate independent components from mixed signals, which is better for solving mixed problems. Based on these features, it was decided to try ICA as a kind of extension of PCA for dimensionality reduction work.

Regarding parameter selection, after several choices of target dimensions, and controlled use of the same classifier and cross-validation stack, the highest accuracy was eventually found to be 70.919% after dimensionality reduction to 12 dimensions.

SVM main accuracy for PCA by cross-validation: **70.919%**.

After comparison as above, based on accurate measurements and applicability. Even ICA has a little better accuracy than PCA, but for clustering, PCA was far better than ICA. Principal Component Analysis (PCA) was finally chosen as the dimensionality reduction method for this data.

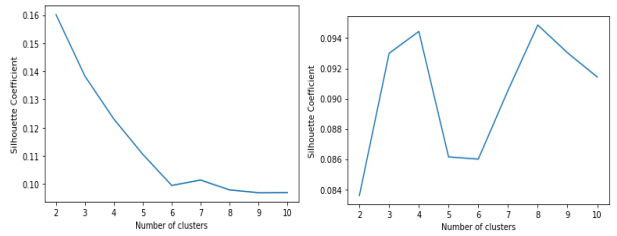

Fig.6,7 Local best Silhouette Coefficient for PCA (left) and ICA (right)

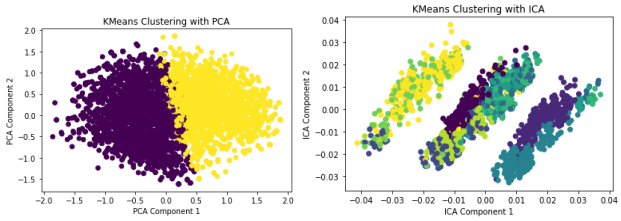

Fig.8,9 Local best Clustering effect diagram for PCA (left) and ICA (right)

5.1 Data processing

Most data are distributed with 0 and 1, but there are still some data with 2, which is a large amount of data. Totally 11.18% of the original data are 2's, which is anomalous data. But after using PCA to reduce the dimension, the number of 2 decreased, and the portion is 10.5%.

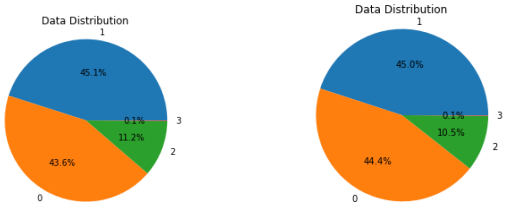

Fig.10,11 Data distribution before PCA (left) and after PCA (right)

5. Classification

For classification problems, the classification algorithm finds the mapping that maps the 'x' input to the 'y' discrete output. This "classification algorithm" is a classifier. In other words, the data will be classified into different categories by the classifier.

For validation. Firstly the purpose of the task needed to be clarified. The answers to the questionnaire required to be classified, so naturally, the most important thing to consider was accuracy, and it was therefore decided to use accuracy as the criterion.

Then, for the verification method, it is natural to pursue accuracy, so the cross-verification method is used to obtain the most accurate results. Given that the data consists of more than 5,000 lines at a moderate level, ten folds are an excellent way to process.

Similar to dimensionality reduction, the choice of classifier also needs to be based on the characteristics of the data. It should use a supervised algorithm classifier to fit the data. As mentioned in 3.1 and 4.1, the current data are discrete and significant in number, so avoiding models suitable for Gaussian distributed data should result in better classification performance. Of the supervised learning algorithms commonly used: regression analysis, logistic regression, K-nearest neighbours, decision trees, Naive Bayesian, and support vector machines, the following were pre-selected as the models of choice: support vector machines, Naive Bayesian, and logistic regression classifiers.

In general, multi-classifiers do not perform as well as binary classifiers and are generally less accurate than binary classifiers. First, consider the source of these data. According to the task sheet, these data are outliers and can be removed directly. In other words, these data may be due to data collection errors, anomalies or other circumstances that do not match expectations. Therefore, in these terms, even though the proportion of 2 is large, at 11.18%, it can still be treated as noise, and this approach is more appropriate.

5.2 Support vector machines (SVM)

Support vector machine is a linear classifier that bifurcates data according to supervised learning, and its decision boundary is the maximum distance hyperplane solved for the learned samples [8]. SVM performs well with small and medium-sized datasets and has high computational efficiency. At the same time, SVM can effectively classify high-dimensional space, which is suitable for data sets with multiple features, and this algorithm is based on binary classification [9]. However, the merits of SVM are that it is sensitive to noise and missing data, and some of the 2s in the dataset are noisy, which can impact SVM. One of the lesser and more critical parameters is that different Margin Classification applies to different situations.

Curved margin: Data cannot be partitioned by straight lines or hyperplanes in some cases and needs to be divided using non-linear boundaries such as curves.

Elastic margin: Allow for some outliers and noise, allowing for some degree of edge or internal error and tolerating anomalies in the data.

Soft margin: the model allows for some misclassification points on the classification boundary to accommodate noise or the presence of some degree of data overlap.
Hard margin: it requires that all data be accurately classified and that the data be strictly linearly separable.
Large margin: classification is performed by finding a hyperplane with the maximum interval such that the distance between positive and negative samples is as considerable as possible.

There is a defined value "C" in the penalty degree. The C-value corresponding to a hard margin is larger. The penalty for each failure to classify correctly is more substantial.

In this data, it is linearly separable. However, as there is more noise and some outliers, a soft boundary, a smaller C-value, should, in theory, be used as a parameter. However, during the actual testing using the parametric grid, instead of the highest accuracy being presented when C is used as the generally smallest value (0.1), the highest accuracy is shown when it is set to the median value of 1. Analysis of the possible reasons for this may be due to its more pronounced decision boundaries and more discrete data, which are effectively reconciled with the problem of more noise.

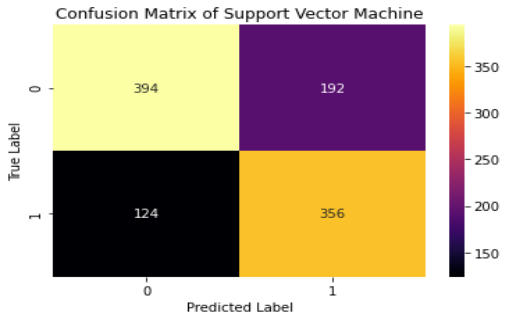

Fig.12 Confusion Matrix of Support Vector Machine

The best gamma and kernel for the highest accuracies by grid search are in the following:

C	Best gamma	Best Kernel	Accuracy
10	0.1	rbf	0.71529
1	0.1	rbf	0.71529
0.1	0.01	rbf	0.70497

Ultimately, the highest average accuracy obtained using the SVM classifier was **71.529%**.

5.3 Naive Bayes

Bayes' theorem describes the likelihood of an event based on prior knowledge of some situation related to the matter. Naive Bayes is a probability-based classification algorithm for solving classification problems based on Bayes' theorem and feature independence assumption [9],[10].
The central assumption of the Naive Bayesian algorithm is that features are conditionally independent; for instance, each component is independent of the others for a given class. This assumption may not hold in practical problems, but in many cases, Naive Bayes can still produce better classification results.

There are three submodules of Naive Bayes in Scikit-learn:
1. **Gaussian**: applied to continuous variable features, which are assumed to follow a Gaussian distribution.
2. **Polynomial**: commonly used for text classification, where the features are words, and the value is the number of occurrences of the word.
3. **Bernoulli**: each feature is Boolean, true or false, and can be replaced by 0/1.

In this problem, the data fits perfectly into the Bernoulli model; that is, the 0's and 1's in the questionnaire can be considered as Boolean data. And several parameters can be adjusted, one of which is the smoothing parameter [11]. It is used when specific feature values do not appear under a particular category in the training data. The smoothing parameter controls the model's estimation of unseen feature values, and commonly used smoothing parameters include Laplace smoothing and Lidstone smoothing.

However, even though Bernoulli's method of Naive Bayes seems to fit the data best, whichever model is in Naive Bayes, the results are less accurate than SVM, regardless of which smoothing is attempted.

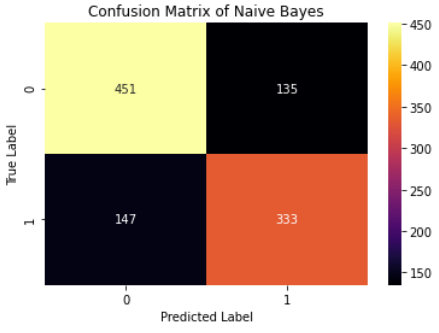

Fig.13 Confusion Matrix of Naïve Bayes

Laplace Smoothing average accuracy by cross-validation: **68.611%**
Lidstone Smoothing average accuracy by cross-validation: **68.630%**

Whichever way is the highest, the Naive Bayesian accuracy was **68.63%**.

5.4 Logistic regression

Logistic regression is a generalised linear model with many similarities to multiple regression analysis but is typically used to predict the probability of a binary output variable. It is based on a linear regression model, transformed by a logistic function (also known as a Sigmoid function) that converts continuous predictive values into probability values and makes classification decisions based on a threshold (usually 0.5, with greater than 0.5 being positive and less than 0.5 being negative) and thus classification decisions.

In 5.2, logistic regression has the same penalty parameter, C, and the reasoning is similar to the margin. In addition, to prevent overfitting, logistic regression also introduces a regularisation in the loss function (cost function), which can be either L1-regularised or L2-regularised [12]. To simplify, L1 regularisation is suitable for problems with features and sparsity, while L2 regularisation is suitable for dealing with covariance problems and maintaining a balance of overall features.

There are some common types of solvers: this is the solver for small data sets based on the axis descent method.

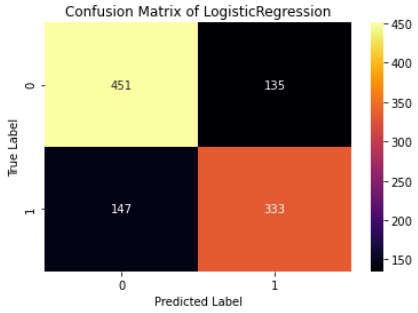

Fig.14 Confusion Matrix of Logistic Regression

Based on experiments with grid parameters, the following approach was finalised:
Best parameters: **'C':0.1,'penalty':12, 'solver': sag**.
Best accuracy:**72.186%**.

The logistic regression achieved an accuracy of **72.186%**.

Overall, the highest accuracy rate of 72.186% is achieved using the logistic regression classifier.

6. Clustering

Clustering, is a statistical method for studying classification problems. Clustering is an unsupervised learning method that divides and organises objects in a dataset according to similarity, forming clusters with internal closeness and external distinction. In the field of data science, grouping data allows for clearer access to data information and visualisation. At the same time, clustering can help us to identify outliers and outliers in a data set [13]. By organising data points into clusters, outliers are often grouped into separate clusters or are more different from other clusters, thus facilitating the detection and analysis of outliers.

Commonly used clustering methods include K-Means clustering, hierarchical clustering, DBSCAN clustering and Gaussian Mixture Models (GMM). The principles differ, and so do the approaches applied.

6.1 The K-Means algorithm

The K-Means algorithm is the most used clustering method and is widely used and simple in many industrial-grade data science and machine learning applications. In simple terms, the K-Means algorithm calculates the distance of each data to a centroid, divides the data points into whichever class they are close to, and calculates the average of the positions of the issues in each class as the new cluster centres [11]. Then, according to the required clusters, until the centre of each class stays mostly the same after each iteration to complete the clustering process. It has the advantage of being very fast, as the knowledge needed to calculate the distances between points and group centres is less computationally intensive.

For the evaluation of the clustering results, the authors decided to choose data visualisation and standard parameters to observe the clustering effect.

The Elbow function was applied as a criterion, a method used to determine the optimal number of clusters in K-Means clustering. It is based on the analysis of the clustering error about the number of clusters K. Visually, as K increases, the clustering error usually decreases, as more clusters can fit the data better. However, as K

continues to grow, the improvement in clustering error gradually diminishes. Therefore, the value of K at the inflexion point is considered to provide a reasonable number of clusters while avoiding over-fitting, a point generated by a compromised algorithm.

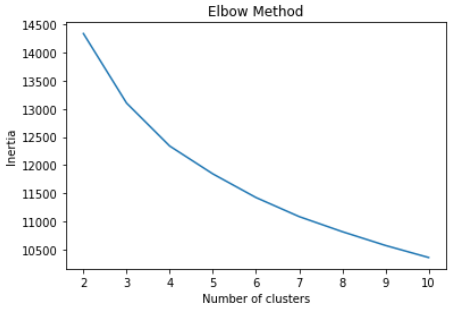

Fig.15 Elbow method curve for K-Means

On the other hand, there is another criterion, silhouette coefficient, which is a metric used to assess the quality of clustering and measure the tightness and separation of the clustering results. In simple, the contour coefficient calculates the difference between the average distance between the sample and the nearest samples from different clusters and the distance from the sample to other samples in the same cluster. The closer to 1, the closer the sample is to other samples in the same cluster, the further it is from the nearest other clusters and the better the clustering result; the closer to -1, it is on the contrary. Also, the authors will use data visualisation to view the effects based on what is shown after clustering.

Ultimately, as the elbow function did not reveal obvious inflexion points, it was decided to use the number of clusters corresponding to the topic with the most significant contour coefficient: 2 to obtain the final clustering effect plotted as follows:

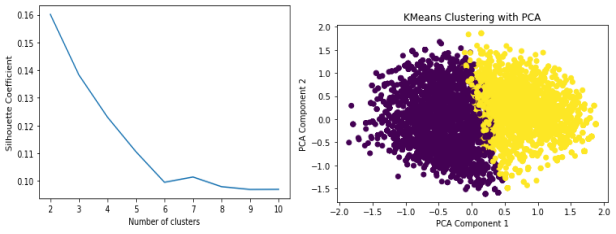

Fig.16 Silhouette Coefficient for K-Means Fig.17 Clustering effect diagram for K-Means

Silhouette coefficient is **0.16036283490454256**.

6.3 Hierarchical clustering

Hierarchical clustering is based on the idea of gradually merging or splitting samples. The similarity or distance relationship between data samples is represented by treating each data point as a separate cluster at the beginning and then gradually merging them until all clusters are merged into the same cluster species, which contains all the points. In this way, the hierarchy of clusters is represented as a tree diagram.

Consistent with the approach in 6.2, a combination of the two rubrics is taken (Elbow serves for K-Means):

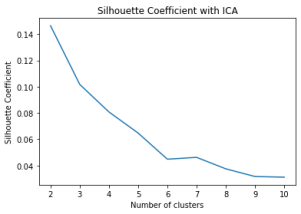

Fig.18 Silhouette Coefficient for hierarchical clustering

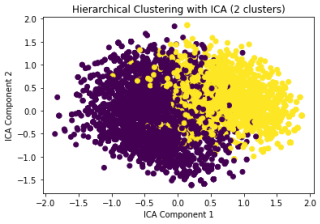

Fig.19 Clustering effect diagram for hierarchical clustering

Silhouette coefficient is: 0. 14645721409970333**.**

Obviously, the silhouette coefficient is smaller than K-Means, and the clustering effect is also inferior to K-Means (There are more mixed parts in the middle and has no clear separation).

Finally, the K-Means method is chosen.

7. Deficiency and Innovation in this Coursework

Until now, the author has finished the specific statement of the report, while there are still some areas to improve in this report.

Firstly, there are still plenty of kinds of algorithms that the author hasn't tested. Each of them may have better performance than the current ones. What can do is continuously update optimal classifiers to get better accuracy. Secondly, due to the limitation of the pages, the author hasn't demonstrated the detailed processes in the main body. Yet, the author put some detailed or general principles in the appendix. If the reader wants to know more, can move to the appendix to find more detailed info.

Also, there are still some possible ways to get more accurate results. For example, it can combine the classifiers and use respective strengths to fix or train the data so that the classifiers fit to present items.

8. Conclusion

In this report, based on extensive experiments and evaluations, Principal Component Analysis (PCA) was chosen as the final dimensionality reduction method, reducing the data to 13 dimensions. This choice was aimed at solving the overfitting problem and achieving optimal performance in classification and clustering tasks. Also, it avoids the overfitting problem. By using multiple classifiers and tuning their internal parameters, the highest accuracy was achieved (by the logistic regression method). In addition, data visualisation techniques enhance the understanding of the data. The results and methods presented in this report demonstrate the authors' competence in data analysis and highlight the iterative nature of machine learning.

The results and methods presented throughout the report result from intensive testing and refinement. Machine learning is characterised by constant trial and error and constant updating, and the authors have established a reliable analytical approach by conducting numerous experiments and learning from their mistakes. This report demonstrates the authors' ability to process and analyse given data, combining theoretical understanding, practice, and critical thinking.

8. Reference:

[1] S. Kang, S. Cho, and P. Kang, "Constructing a multi-class classifier using the one-against-one approach with different binary classifiers," Neurocomputing, vol. 149, no. Part B, pp. 677–682, 2015. doi: 10.1016/j.neucom.2014.08.006.

[2] B. Zhao and B. Zhao, "Research on the Application of Machine Learning Classification Based on Data Reduction," Modern Information Technology, vol. 2, no. 2, pp. 144-145, 2018. doi: 10.3969/j.issn.2096-4706.2018.02.052.

[3] V. Gray, "Principal component analysis: methods, applications, and technology," Mathematics Research Developments, Novinka, 2017. [Online]. Available: https://search-ebscohost-com.ez.xjtlu.edu.cn/login.aspx?direct=true&db=cat01010a&AN=xjtlu.0001130781&site=eds-live&scope=site. [Accessed: May 15, 2023].

[5] J. Zhang et al., "The Iterative Solution of Linear Discriminant Analysis and Its Application," Periodical of Ocean University of China, vol. 45, no. 11, pp. 119-124, 2015. doi: 10.16441/j.cnki.hdxb.20130328.

[6] Z. Liu, J. Wang, Y. Zhang, and H. Li, "Modified Linear Discriminant Analysis Method MLDA," COMPUTER SCIENCE, vol. 37, no. 11, pp. 239-242, Nov. 2010. doi: 10.3969/j.issn.1002-137X.2010.11.057.

[7] D. Tao, X. Li, X. Wu and S. J. Maybank, "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 10, pp. 1700-1715, Oct. 2007, doi: 10.1109/TPAMI.2007.1096.

[8] S. Zhang, J. Li, X. Wang, and Y. Chen, "Twin proximal least squares support vector regression machine based on heteroscedastic Gaussian noise," Journal of Intelligent & Fuzzy Systems, vol. 44, no. 2, pp. 1727-1741, Feb. 2023. doi: 10.3233/JIFS-211631.

[9] Z. Su, et al., "Review on Support Vector Machine Based on Bayes' Theorem," COMPUTER APPLICATIONS AND SOFTWARE, vol. 27, no. 5, pp. 179-193, May 2010. doi: 10.3969/j.issn.1000-386X.2010.05.053.

[11] S. Rana, R. Kanji, and S. Jain, "Comparison of SVM and Naïve Bayes for Sentiment Classification using BERT data," in 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), Aligarh, India, pp. 1-5, 2022. doi: 10.1109/IMPACT55510.2022.10029067.

[12] A. Bar-Hillel, I. Bilik and R. Hecht, "Naive Bayes nearest neighbour classification of ground moving targets," 2013 IEEE Radar Conference (RadarCon13), Ottawa, ON, Canada, 2013, pp. 1-5, doi: 10.1109/RADAR.2013.6586125.

[13] K. She, C. Dai, and Y. Ding, "Application of Logistic Regression and Principal Component Analysis in TCM diagnosis and treatment," in 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2021, pp. 726-729, doi: 10.1109/AEMCSE51986.2021.00150.

[14] Gauss Naïve Bayes (2022) Available at: https://awesomeopensource.com/project/odubno/gauss-naive-bayes (Accessed: 16 May 2023).

9. Appendix

9.1 Brief description of the basic mathematical process of PCA

Firstly, in mathematical statistics, variance is used to calculate the difference between a variable and the overall mean.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n}$$

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

The covariance (cov) is used to identify correlations in high-dimensional data.

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Let $X=(X_1,X_2,X_3,\ldots\ldots,X_N)^T$ The matrix:

$$C = (c_{ij})_{n\times n} = \begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{pmatrix}$$

is the covariance matrix of n-dimensional random variables.

And $c_{ij} = \mathrm{cov}(X_i,X_j)$, i, j=1,2,3,……,n

9.2 Evaluation criteria

Available metrics for evaluating classification models: accuracy (most used), precision and recall, F1 score, AUC & ROC.

Accuracy measures the ratio between the number of samples correctly classified by the model and the total number of samples. Accuracy can be used as a basic indicator of overall classifier performance but can be biased when dealing with unbalanced datasets.

Precision: A measure of the proportion of the sample whose model predicts a positive class that is a positive class. Precision is concerned with the accuracy with which a model predicts a positive class and is used in situations where false positives (incorrectly predicting a negative class as a positive class) are more costly.

Recall: A measure of the ratio between the number of samples that the model correctly predicts as positive classes and the number of samples that are positive classes. The recall concerns the rate at which the model checks out positive classes and is applicable when false negatives (incorrectly predicting positive classes as negative classes) are costly.

F1 Score: Combines precision and recall and is the summed average of precision and recall. F1 Score combines the predictive accuracy and the completeness of the model and is suitable for scenarios where precision and recall are balanced.

AUC and ROC: used to evaluate the performance of the classification model under different thresholds.

According to this assignment, the most crucial thing is correctly classifying each patient, so accuracy is the most prominent and will be used as the criterion in the report.

9.3 Parameter regularisation methods

Where the penalty term used for L1 regularisation is the sum of the absolute values of the model coefficients (weights), it reduces the complexity of the model by making some coefficients zero, enabling feature selection, for example, sparsification of the model.

The penalty term used for L2 regularisation is the square root of the sum of squares of the model coefficients (weights). It pushes the model coefficients towards smaller values by penalising the sum of squares of the coefficients but does not force them to be 0.

9.4 Small dataset solver

"sag": this is the optimisation algorithm for stochastic mean gradient descent, which uses the average of the gradients of each sample to update the model parameters.

"newton-cg": this is a Newtonian optimisation algorithm which uses an approximation of Newton's method to update the model parameters. It supports L2 regularisation.

"saga": This is an improved version of the "sag" algorithm. The "saga" algorithm performs better with large data sets and high-dimensional features.

9.5 Brief description of the basic mathematical process of Naïve Bayes

As the Bayes' theorem:

Bayes' theorem

P(A, B)=P(A|B)P(B);

P(A, B)=P(B|A)P(A):

P(A| B)=P(B|A)P(A)/P(B);

Plain Bayes ignores the interactivity of the samples and treats all elements as unrelated. Then, according to Equals above:

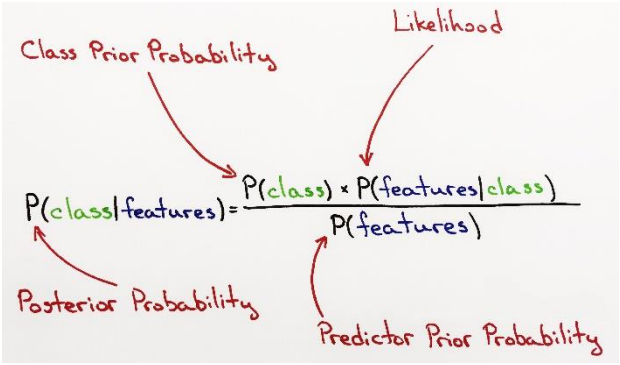

Fig.20 Naïve Bayes classification formula [14]

Finally, get the different possibilities from different classes, and choose the highest possible as the decided class.