Explain the Predictions with Multiple Rule Inference Methods

Xinyue Xu

Research School of Computer Science
Australian National University, Canberra, Australian
u6708515@anu.edu.au

Abstract. In this paper, **fully connected neural network, decision tree, ensemble learning** and **fuzzy classification system** are used to generate different models to classify SARS dataset, and the perfect accuracy of 100% was obtained. In order to explain the model prediction, interpretative rules are generated for different models in different ways. The neural network adopts weights to generate **causal index** and combines with **characteristic input** to generate rules and the decision tree observes the decision steps. The ensemble learning model is trained by using bagging estimator and eliminates semantic duplication rules by comparing similarities. The fuzzy classification system uses the **hyperbox recursive definition** to solve the overlap problem, using fuzzy clustering to classify the dataset and the generation of rules. The generated rules are analyzed to explain their possible application scenarios. The characteristics of the dataset are also analyzed. Finally, different methods are used to explain the predictions of SARS dataset from different points of view.

Keywords: Fuzzy classification system · Causal index · Decision tree · Ensemble learning · Fully connected neural network · Rule extraction.

1 Introduction

1.1 Background

In traditional Boolean logic, the truth value can only be the integer value 0 or 1. However, in real life, the truth value of many events may vary between completely true and completely false. In order to make up for the deficiency of classical logic, Prof. Lotfi A. Zadeh put forward the theory of fuzzy sets in 1965 [1]. Fuzzy logic is closer to human reasoning and it is convenient to generate approximate information and uncertain decisions. It is a tool for dealing with many inherently imprecise problems.
There are many applications of fuzzy logic, one of which is in medical decision making. Since most medical data are subjective or fuzzy, it can be well collected and represented by fuzzy logic-based methods. The most common area of application is computer aided diagnosis (CAD) [2] in medicine. Fuzzy logic can be used to describe the key features of various lesions. Doctors use a large number of statistical data to extract the key characteristics of the disease as an auxiliary basis.

1.2 Dataset

SARS-COV-1 dataset [3] is a fuzzy dataset that contains four diseases sets and 23 kinds of fuzzy values in each set. There are four diseases in total, which are hypertension, normal, pneumonia and SARS. There are a thousand observations for each disease. Raw data have no label values, and their file names represent the disease categories to which they belong. The 23 fuzzy values can be divided into four general symptoms, which are fever, blood pressure, nausea and whether there is abnormal pain. Fever data is monitored and collected four times a day at 8 a.m., 12 p.m., 4 p.m., and 8 p.m. Fever is divided into three fuzzy sets: slight, moderate, and high fever. Blood pressure is measured by systolic or diastolic blood pressure, which is divided into three grades: low, normal or high. Nausea, like fever, can be slight, moderate or high. Abnormal pain is divided into yes and no. Table 1 clearly lists the composition of the SARS dataset. Table 2 shows the statistical distribution of the dataset. This dataset will be trained using a variety of methods and the resulting model will be used to predict the disease a patient will have. It is essentially a disease classification problem.

Symptom	**Class**	**Type**
Fever	8 am	slight, moderate, high
	12 pm	slight, moderate, high
	4 pm	slight, moderate, high
	8 pm	slight, moderate, high
Blood Pressure	Systolic	low, normal, high
	Diastolic	low, normal, high
Nausea		slight, moderate, high
Abnormal Pain		no, yes

Table 1. SARS dataset symptom, classes and types.

column name	count	mean	std	min	max
fever_temp_8_s	4000.0	0.52442595	0.4275884749061481	0.0	1.0
fever_temp_8_m	4000.0	0.402175975	0.3629893003152231	0.0	1.0
fever_temp_8_h	4000.0	0.41300680000000006	0.41972740613715664	0.0	1.0
fever_temp_12_s	4000.0	0.5130529	0.44028848151902583	0.0	1.0
fever_temp_12_m	4000.0	0.412093575	0.3706425016606011	0.0	1.0
fever_temp_12_h	4000.0	0.412182575	0.4190504831231206	0.0	1.0
fever_temp_16_s	4000.0	0.49984685	0.45371025850177216	0.0	1.0
fever_temp_16_m	4000.0	0.41254427499999996	0.3705386567637221	0.0	1.0
fever_temp_16_h	4000.0	0.412087625	0.4188062570862606	0.0	1.0
fever_temp_20_s	4000.0	0.5000865999999999	0.4536593419627525	0.0	1.0
fever_temp_20_m	4000.0	0.414540325	0.3717170913213293	0.0	1.0
fever_temp_20_h	4000.0	0.41303165000000003	0.41970441777857576	0.0	1.0
blood_pre_sys_s	4000.0	0.475074075	0.47683223782180545	0.0	1.0
blood_pre_sys_m	4000.0	0.3005223	0.2964659283954648	0.0	1.0
blood_pre_sys_h	4000.0	0.412774625	0.41978111240167976	0.0	1.0
blood_pre_dia_s	4000.0	0.50046535	0.4536947229247657	0.0	1.0
blood_pre_dia_m	4000.0	0.299188875	0.2936420622014289	0.0	1.0
blood_pre_dia_h	4000.0	0.41106615	0.4178952404337381	0.0	1.0
nausea_s	4000.0	0.775246875	0.3898398421093697	0.0	1.0
nausea_m	4000.0	0.18718332499999998	0.3279421400510305	0.0	1.0
nausea_h	4000.0	0.189132575	0.3311923513077567	0.0	1.0
no_pain	4000.0	0.75	0.43306683863080586	0.0	1.0
have_pain	4000.0	0.18708897500000002	0.3280563541967014	0.0	1.0

Table 2. SARS dataset summary.

1.3 Problem Description

However, the problem to be solved in this paper is not so simple as the problem of disease classification, but to explain the generated model. That is, to come up with a set of rules that make the model have been used interpretive. The SARS dataset will be trained with multiple models: the traditional fully connected neural network, decision tree, ensemble learning model, fuzzy c-means clustering and fuzzy classification system. Fully connected neural network will use causal index and characteristic input to generate rules. The decision tree will automatically generate relevant rules during the process of judgment. Ensemble learning is a further improvement of decision tree model, which will balance the interpretability of decision tree and the modeling ability of random forest to generate a strong set of classification rules. Fuzzy c-means will give a overview of fuzzy clustering result. It is a simple representation of a fuzzy classification system. The fuzzy classification system classifies the whole data set in detail, and gives the clustering results and visualization results. The rules generated by each method will be visualized and analyzed on the test set.

1.4 Model Introduction

Fully connected neural network is a very traditional classification method. In this paper, the common three-layer fully connected neural network is used to train the training set. The structure is shown in Figure 1. Log softmax activation function is used in the output layer, because it outputs more stable values, performs well in digital representation, and avoids loss of precision. For the following calculation model of causal index to pave the way. Cross-Entropy loss is used to calculate the model loss, which is a very common loss function for classification problems. The weights of the trained neural network will be used to calculate the causal index, which will be explained in the method section.

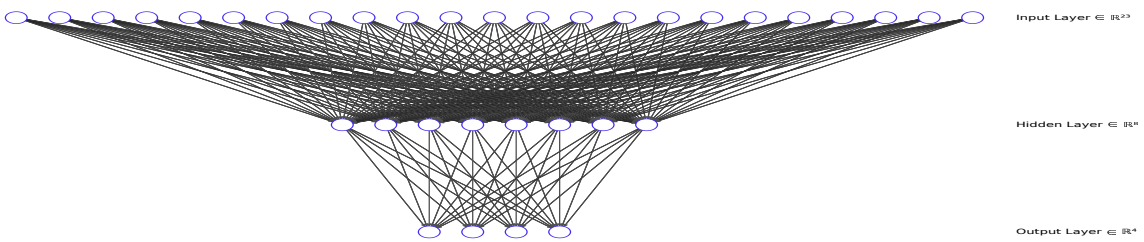


Fig. 1. Three layers fully connected neural network

Decision tree is a kind of tree based on strategic choice. In machine learning, decision tree is a prediction model. It represents a mapping relationship between object attributes and object values. Each node in the tree represents an object and each bifurcation path represents a possible attribute value. The path experienced from the root node to the leaf node corresponds to a decision test sequence. The decision tree can be a binary tree or a non-binary tree, or it can be regarded as a set of if-else rules, or it can be regarded as a conditional probability distribution in the feature space. What makes decision trees special in the field of machine learning models is the clarity of their information representation. The knowledge obtained by the training of decision tree directly forms the hierarchical structure. This structure preserves and presents knowledge in such a way that even non-experts can easily understand it. The branching path of a decision tree is the rule that interprets the decision tree model, which is intuitive.

Ensemble learning is the strategic generation and combination of multiple models (such as classifiers) to improve the performance of the model [4]. Although the decision tree is simple and intuitive, it also has the hidden trouble of over-fitting. Learning an optimal decision tree is considered an NP-complete problem [5]. In practice, the decision tree is built based on the heuristic greedy algorithm, which cannot guarantee the establishment of the global optimal decision tree. The introduction of Random Forest can alleviate this problem. Integrating the advantages of decision tree and random tree is the method adopted in this paper. Ensemble learning is generally divided into three categories: Bagging, Boosting and Stacking. The model will use bagging method to ensemble the decision tree and random forest to generate rules.

Fuzzy Clustering is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership level [6]. The membership level shows how strongly the data points are associated with a particular cluster. Fuzzy clustering is the process of calculating these member levels and deciding which one or which cluster data points belong to according to member levels. One of the most widely used fuzzy clustering algorithms is the Fuzzy C-means clustering algorithm.

Algorithm 1: Fuzzy C-Means [7]

Initialize centroids V_j, j=1...C;
Compute the degree of membership of all feature vectors in all the clusters;
$u_j.i = \frac{(1/d^2(X_i,V_j))^{1/(m-1)}}{\sum_{j=1}^{C}(1/d^2(X_j,V_j))^{1/(m-1)}}$;
while $\max_{ji} \mid u_{ij} - \hat{u}_j 1 \geq \varepsilon$ **do**
  $\hat{v}_j = \frac{\sum_{i=1}^{N}(u_{ji})^m X_i}{\sum_{i=1}^{N}(u_{ji})^m}$;
  and update the degree of membership u_{ji} to $\hat{u}_{ji}$;
end

Above is the pseudocode of the Fuzzy C-Means algorithm. It can be seen that compared with the traditional hard clustering, there are more steps to calculate the degree of membership.

2 Method

2.1 Data Preprocessing

Because all the data in the dataset is concentrated between 0 and 1. So we do not need to normalize the data. The four disease datasets are labeled: HighBP: 1, Normal: 2, pneumonia: 3, SARS: 4. This represents the classification label to which they belong. Numbers are used to facilitate training and calculation of the model. As for the 23 fuzzy attributes, to facilitate the generation of rules, they are given column names. This is consistent with the summary of the dataset in Table 2. Although the dataset does not require much processing, consider that this is a 23-dimensional data set. If we want to perform clustering visualization, the dimension must be reduced to a lower dimension. T-SNE [8] method is used to compress and visualize the dataset. T-Distribution Stochastic Neighbour Embedding is a machine learning method for dimensional reduction that helps us identify associated patterns. The main advantage of T-SNE is the ability to preserve local structures. This means that points that are close in distance in the higher-dimensional data space are projected to be close in the lower-dimensional data space. T-SNE also produces nice visualizations. Finally, the four separate datasets are assembled into one dataset, and the sequence is randomly shuffled. The ratio of the training set to the test set is listed as eight to two. The original 23-dimensional dataset and the 2-dimensional dataset after dimension reduction are stored respectively for use.

2.2 Causal Index and Characteristic Input

Input patterns are classified according to their impact on each particular output [9]. Defines the input which has influence on a certain class as the characteristic input of this class. In view of the large number of characteristic inputs, the arithmetic mean of characteristic inputs is calculated and their average is defined as characteristic ON pattern, predictions with respect to other classes are considered characteristic OFF pattern, and vice versa. There are four classes in total. Each class has a characteristic ON pattern and three characteristic OFF patterns. The characteristic inputs are calculated in the training set of the network and does not involve any data from the test set.
The causal index depends on the differentiability of the network activation function. Based on the rate of change of the influence of the input value x_i on the output value y_k, we can determine which inputs will have a greater influence on the classification. According to the weight of each neuron in the neural network, a weight matrix with the size of $23(features) * 4(classes)$ is obtained. We can clearly see the impact of each feature input on the output of each class. The calculation method [10] (as below) is to multiply and sum the weights of each layer to get the final causal index. Since the derivatives $f'(U_{k2}) \cdot f'(U_{j1})$ will get constants, we can ignore them.

$$\frac{dy_k}{dx_i} = \frac{dy_k}{dU_{k2}} \cdot \frac{dU_{k2}}{dh_j} \cdot \frac{dh_j}{dU_{j1}} \cdot \frac{dU_{j1}}{dx_i} \quad = f'(U_{k2}) \cdot f'(U_{j1}) \cdot \sum_j w_{jk} \cdot W_{ij}$$

Using the fully connected neural network completed by training, the weight of each layer is extracted and the causal index matrix can be obtained by calculating according to the above formula. The causal index of each class is analyzed. The three causal indexes with the largest absolute value are extracted to generate rules. If the causal index is positive, the feature is positively correlated with the class, while a negative number is negatively correlated with the class. The boundary value is determined by the characteristic ON input. If the correlation is positive, the feature must be greater than the boundary value, while the negative correlation is less than the boundary value. The rules generated for each feature are concatenated using AND. Each class generates a discriminant rule to interpret the predicted results.

2.3 Decision Tree Rules

Generally, the generation of decision tree starts from the root node, the corresponding features are selected. Then the segmentation points corresponding to the features of the nodes are selected. The nodes are split according to the segmentation points. The leaf node in the tree represents the result of a decision based on a series of judgments on the path from the root node of the tree to that leaf node. Impurity is used to select the features of nodes and feature segmentation points. Gini index is used to reflect the equilibrium degree of the distribution of samples of different categories in the sub-nodes of the decision tree, that is, the impurity. The Gini index is calculated by the following formula [11]:
Assumes the dataset has a total of C class, $[p(C \mid t)]$ is the relative frequency of class C samples in node t, then the Gini index of node t is:

$$Gini(t) = 1 - \sum_{c=1}^{c} [p(C \mid t)]^2$$

The rule generation is completed when the tree splits to a Gini index of 0. The conditional criteria of each node are extracted to obtain the final rule.

2.4 Ensemble Learning Rules

The approach of ensemble learning is a trade-off between the interpretability of decision trees and the modeling capability of random forests [12]. It introduces random forest to solve the over-fitting phenomenon caused by the over-complexity of decision tree in large datasets. Rules are extracted from a set of trees and a weighted combination of these rules is constructed by solving the L1 regularization optimization problem on the weights [13]. Extracting rules from a set of trees enables us to generate such a set of trees using existing fast algorithms, such as bagged decision trees or gradient enhancement. Rules that are too similar or duplicate are then removed based on the supported similarity thresholds. In this paper, Skope-Rules model is used to train the training set. The structure [14] of Skope-Rules is shown in Figure 2. The Bagging estimator trains multiple decision tree classifiers and each node in the Bagging estimator can be thought of as a rule. Performance filtering uses out-of-bag accuracy and recall thresholds to select the best rule. Semantic deduplication applies similarity filtering and comparison operators are used to combine variable names [14]. In the classification task, Skope-rules are useful for describing each cluster. Each cluster can be postprocessed and approximated with a set of interpretable rules.

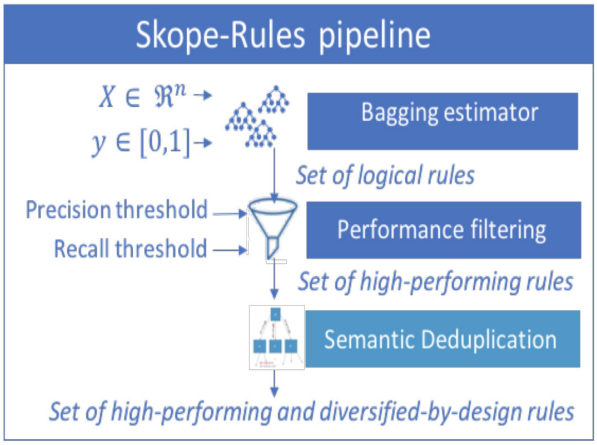


Fig. 2. Skope-Rules Structure

2.5 Fuzzy Classification System [15]

In addition to using neural network to automatically extract fuzzy rules, fuzzy rules can also be extracted directly from numerical data. For example, in [16], fuzzy rules are obtained by dividing the input space into fuzzy regions, the output space into regions, and determining in which fuzzy region each numerical input data is contained, and in which output region the corresponding output data is contained. While the above approach is very simple, it also requires pre-partitioning the input space. In [17], fuzzy rules with variable fuzzy regions are extracted for classification problems. This method solves the problem of the fuzzy system mentioned above. The input area of each class is represented by a set of hyperboxes that allow overlaps between hyperboxes of the same class, but not between different classes. If some data in the hyperbox belongs to a class, it is determined that the data is that class. When multiple classes overlap, the learning algorithm dynamically expands, divides, and shrinks the hyperbox. The system [15] introduces two types of hyperboxes: activation hyperboxes which define the existence regions for classes, and inhibition hyperboxes which inhibit the existence of data within the activation hyperboxes. The hyperboxes are defined recursively. Firstly, calculating the minimum and maximum values of data for each class. If there is a overlapping region between class i and j, it should be defined as an inhibition box. If in the inhibition hyperbox exists classes i and j, the system will define additional activation hyperboxes for these classes. Similarly, if there is overlap between these activated hyperboxes, we further define the overlap region as inhibition hyperboxes. In this way, the overlap of the activation hyperboxes is resolved recursively. The following figure is the schematic diagram of the fuzzy classification system. It is effective to use a set of hyperboxes to represent the data presence domain of a class to deal with classification problems with many input variables. The overlap between different classes is solved by the method of recursive definition. The boundary of the hyperboxes are the rules we want to generate.

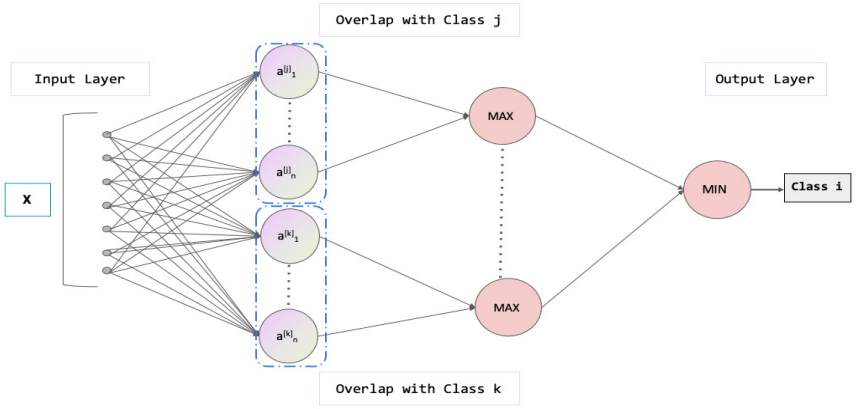


Fig. 3. Architecture of a fuzzy classification system (only the network for class i is shown)

3 Experiments and Results Discussion

After many experiments, since the accuracy of each model is always 100%, the following discussion is about the characteristics of using different models to generate the rules.

3.1 Causal Index Rules

The parameters of neural network training are very simple. Input size is 23 and output size is 4. The number of hidden neurons is set as 8, considering the dataset as 8 classes. The batch size is set as 10 and learning rate is 0.01. Epochs is set 500, but it can achieve 100% accuracy and 0 loss at around 51 epochs. According to the causal index calculation method, three features with the highest correlation were found for each class, and the positive and negative signs were used to judge whether they were greater than or less than the characteristic boundary value. The result is the following set of rules. The number before the colon represents the selected feature column.
For HighBP: 17: '>0.41692959375', 14: '>0.41889875000000004', 12: '<0.46736246875'
For normal: 14: '<0.41889875000000004', 17: '<0.41692959375', 12: '>0.46736246875'
For pneumonia: 14: '<0.41889875000000004', 12: '>0.46736246875', 17: '<0.41692959375'
For SARS: 14: '>0.41889875000000004', 17: '>0.41692959375', 16: '>0.3041014375'
Verified by the test set, the accuracy is 100%. It can be proved that this set of rules can explain the neural network model predictions for SARS dataset.

3.2 Decision Tree Rules

The precision of the decision tree is also 100%, and the depth of the tree is 4. The explanatory rules it generates are much more straightforward.
The rules for decision tree:
$IF \quad \mathrm{X}[22] \geq 0.25, \quad THEN \quad Class4.$
$IF \quad X[17] \geq 0.4, \quad THEN \quad Class1.$
$IF \quad X[6] \geq 0.499, \quad THEN \quad Class2.$
$IF \quad X[6] \leq 0.499, \quad THEN \quad Class3.$

The structure of the entire tree is shown in Figure 3. It can be seen that the decision tree only uses three features to perfectly classify the data, which verifies that the SARS dataset model trained by the decision tree is also very effective.

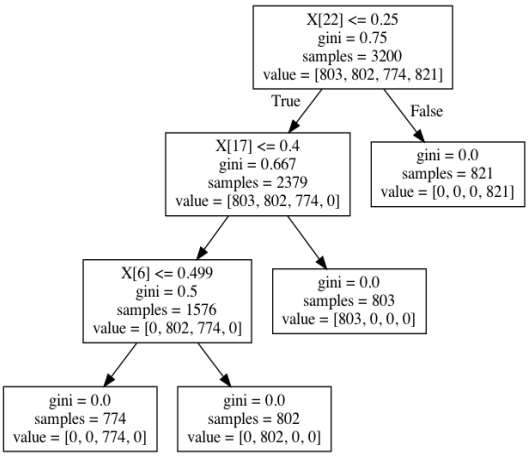


Fig. 4. Decision Tree Rules

3.3 Ensemble learning Rules

Compared with the above two methods of generating rules, the rules generated by ensemble learning are more detailed. It also provides a variety of possibilities for classification rules. For each category, four rules with an accuracy of 100% are generated in this paper and any of them can be used to accurately classify data. The reason for doing this is to test whether the model can generate a variety of rules. This tends to do well in the

Rules for SARS-COV-1 HBP
('fever_temp_12_s >0.9368999898433685 and fever_temp_16_s >0.9986999928951263', (1.0, 1.0, 2))
('blood_pre_dia_s <= 0.43650001287460327 and fever_temp_16_s >0.9986999928951263', (1.0, 1.0, 6))
('fever_temp_8_s >0.9999000132083893 and fever_temp_12_m <= 0.07154999673366547', (1.0, 1.0, 2))
('fever_temp_20_m <= 0.0010499999625608325 and blood_pre_dia_m >0.13900000229477882', (1.0, 1.0, 2))
Rules for SARS-COV-1 NOR
('blood_pre_sys_h <= 0.25 and fever_temp_12_h <= 0.4000000059604645', (1.0, 1.0, 2))
('blood_pre_sys_h <= 0.25 and fever_temp_20_s >0.5000000074505806', (1.0, 1.0, 2))
('blood_pre_dia_h <= 0.4000000059604645 and fever_temp_20_s >0.5000000074505806', (1.0, 1.0, 2))
('fever_temp_8_m <= 0.4000000134110451 and blood_pre_sys_s >0.4000000059604645', (1.0, 1.0, 2))
Rules for SARS-COV-1 PNE
('blood_pre_sys_m <= 0.0006500000017695129 and blood_pre_dia_s >0.9850499927997589', (1.0, 1.0, 2))
('blood_pre_sys_m <= 0.0006500000017695129 and fever_temp_16_h >0.4000000059604645', (1.0, 1.0, 8))
('blood_pre_dia_m <= 0.003650000086054206 and fever_temp_12_s <= 0.524349994957447', (1.0, 1.0, 2))
('fever_temp_20_m >0.366100013256073 and blood_pre_dia_s >0.9990499913692474', (1.0, 1.0, 2))
Rules for SARS-COV-1 SARS
('nausea_m >0.25', (1.0, 1.0, 12))
('nausea_h >0.25', (1.0, 1.0, 12))
('no_pain <= 0.5', (1.0, 1.0, 14))

Table 3. The rules for ensemble learning model.

face of less ideal datasets.

Because the symptoms of SARS are too obvious, only three rules are generated, and the repeated rules will be deleted automatically according to the model design. It can be observed that there are three evaluation parameters after the rule, the meaning of which are respectively the description degree, precision and performance item of the rule for the current cluster. The performance item is the number of extractions in the tree built during rule fitting. It can be seen that the integration model generates as many rules as possible while maintaining precision, which can be described in detail for the entire dataset.

3.4 Fuzzy System Rules

According to the hyperbox boundaries generated for each dimension we get the rules for each class, which is Table 4. These ranges of maximum and minimum values can be used to accurately locate each type of disease. Visualized classification results at low dimensions and high dimensions are shown in Figure 5. Compared with other methods, it has the following advantages and disadvantages:

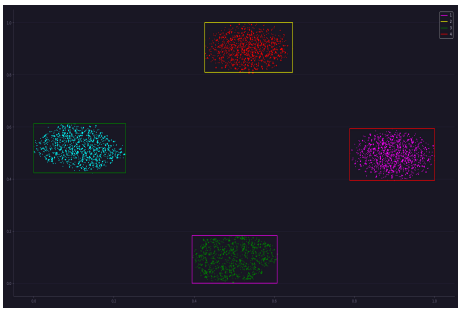
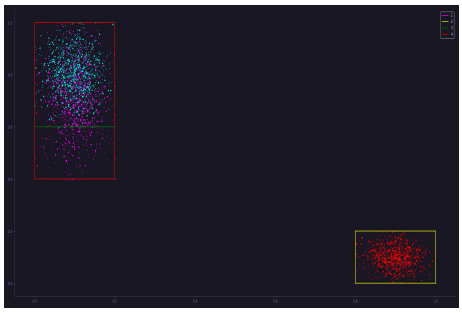
[2d]  [23d] 

Fig. 5. Clustering visualization results.

Advantages: The structure of the network is determined automatically by obtaining fuzzy rules and overlapping between classes. The fuzzy parameters can be modified to improve the generalization ability of the model. The model is relatively easy to implement and the principle is easy to understand. Recursive definition can be a good solution to the overlap problem.

Drawbacks: There is only one rule per feature per class, and test accuracy may be affected if the training set and test set are very different. The overlap problem only considers two hyperboxes at a time, and the network structure may become very complex when there are many categories.

column	1-min	1-max	2-min	2-max	3-min	3-max	4-min	4-max
fever_temp_8_s	1.0	1.0	0.8	1.0	0.0	0.2	0.0	0.2
fever_temp_8_m	0.0	0.0	0.0	0.2	0.6	1.0	0.4	1.0
fever_temp_8_h	0.0	0.0	0.0	0.0	0.8	1.0	0.5	1.0
fever_temp_12_s	1.0	1.0	0.8	1.0	0.0	0.2	0.0	0.1
fever_temp_12_m	0.0	0.0	0.0	0.2	0.6	1.0	0.5	1.0
fever_temp_12_h	0.0	0.0	0.0	0.0	0.8	1.0	0.5	1.0
fever_temp_16_s	1.0	1.0	0.8	1.0	0.0	0.2	0.0	0.0
fever_temp_16_m	0.0	0.0	0.0	0.196	0.6	1.0	0.5	1.0
fever_temp_16_h	0.0	0.0	0.0	0.0	0.8	1.0	0.5	1.0
fever_temp_20_s	1.0	1.0	0.8	1.0	0.0	0.2	0.0	0.0
fever_temp_20_m	0.0	0.0	0.0	0.2	0.6	1.0	0.5	1.0
fever_temp_20_h	0.0	0.0	0.0	0.0	0.8	1.0	0.5	1.0
blood_pre_sys_s	0.0	0.0	0.8	1.0	1.0	1.0	0.0	0.0
blood_pre_sys_m	0.2	0.5	0.0	0.2	0.0	0.0	0.5	1.0
blood_pre_sys_h	0.8	1.0	0.0	0.0	0.0	0.0	0.5	1.0
blood_pre_dia_s	0.0	0.0	0.8	1.0	1.0	1.0	0.0	0.2
blood_pre_dia_m	0.2	0.5	0.0	0.2	0.0	0.0	0.5	1.0
blood_pre_dia_h	0.8	1.0	0.0	0.0	0.0	0.0	0.5	1.0
nausea_s	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.2
nausea_m	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
nausea_h	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
no_pain	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
have_pain	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0

Table 4. The rules for fuzzy system.

4 Conclusion and Future Work

In this paper, multiple methods were used to generate models for the SARS dataset, all of which achieved an accuracy of 100%. And for each model, a different way is used to generate an explanatory set of rules to prove that the model's predictions make sense. For the neural network, the weight of each neuron is extracted and the causal index is generated. The method of combining characteristic input and causal index is used to generate the rules to explain the network. For the decision tree, the intuitionistic decision generating rules of each node are adopted, and the decision tree itself is a very good model of interpretation. Most interesting is the ensemble learning model, which combines the best of various tree-based approaches. It uses similarity to filter rules and generates as many rules as possible to maintain diversity. Fuzzy classification system takes fuzzy clustering as the basic method, uses hyperbox to define the area of each class and uses inhibition box to solve the overlap problem between classes. As a result, rules are generated for each feature of all classes. It is the most complete method of all.

In future work, we can use more complex datasets to compare the advantages and disadvantages of these methods. For the boundary conditions of neural network, we can try to use grid search method to obtain. The decision tree does not have much to improve, but it can be integrated and learned with more models. It turns out that ensemble learning is also very effective in generating rules. According to the characteristics of the dataset, different fusion methods can be adopted to select the model for integration. For the fuzzy clustering system, I tried the adaptive fuzzy clustering neural network. However, it does not get better results than the hyperbox generation clustering rule. The performance of using genetic algorithm to select parameters in adaptive neural network is not good (codes including in the support). Therefore, a better method of model selection should be found in the future. These are the work that can be tried in the future.

References

1. L.A. Zadeh, Fuzzy sets, Information and Control, Volume 8, Issue 3, 1965, Pages 338-353, ISSN 0019-9958. https://doi.org/10.1016/s0019-9958(65)90241-x

2. Juri Yanase, Evangelos Triantaphyllou, A systematic survey of computer-aided diagnosis in medicine: Past and present developments, Expert Systems with Applications, Volume 138, 2019, 112821, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2019.112821
3. Mendis, B. S., Gedeon, T. D., Koczy, L. T. (2005). Investigation of aggregation in fuzzy signatures, in Proceedings, 3rd International Conference on Computational Intelligence, Robotics and Autonomous Systems, Singapore.
4. Dietterich, T. G. (2002). Ensemble learning. The handbook of brain theory and neural networks, 2, 110-125. https://doi.org/10.4249/scholarpedia.2776
5. Laurent, Hyafil, and Ronald L. Rivest. "Constructing optimal binary decision trees is NP-complete." Information processing letters 5, no. 1 (1976): 15-17.
6. Bora, D. J., Gupta, D., Kumar, A. (2014). A comparative study between fuzzy clustering algorithm and hard clustering algorithm. arXiv preprint arXiv:1404.6059.
7. Zhong-dong Wu, Wei-xin Xie and Jian-ping Yu, "Fuzzy C-means clustering algorithm based on kernel method," Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003, 2003, pp. 49-54, https://doi.org/10.1109/ICCIMA.2003.1238099.
8. Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. 11 (2008).
9. Gedeon T. D. and Turner H. S.:Explaining student grades predicted by a neural network. In Proceedings of 1993 International Joint Conference on Neural Networks(1993)
10. Harry,S. T., Tamás,D.G.: Extracting Meaning from Neural Networks
11. Raileanu, L.E., Stoffel, K. Theoretical Comparison between the Gini Index and Information Gain Criteria. Annals of Mathematics and Artificial Intelligence 41, 77–93 (2004). https://doi.org/10.1023/B:AMAI.0000018580.96245.c6
12. https://github.com/scikit-learn-contrib/skope-rules
13. Jerome H. Friedman, Bogdan E. Popescu: Predictive learning via rule ensembles. The Annals of Applied Statistics, 2008, Vol. 2, No. 3, 916–954 https://doi.org/10.1214/07-AOAS148
14. http://2018.ds3-datascience-polytechnique.fr/wp-content/uploads/2018/06/DS3-309.pdf
15. S. Abe and Ming-Shong Lan, "A method for fuzzy rules extraction directly from numerical data and its application to pattern classification," in IEEE Transactions on Fuzzy Systems, vol. 3, no. 1, pp. 18-28, Feb. 1995 https://doi.org/10.1109/91.366565
16. P. K. Simpson, "Fuzzy min-max neural networks-Part 1: Classification," IEEE Trans. Neural Networks, vol. 3, no. 5, pp. 776–786, Sept.1992.
17. R. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, Part II, vol. 7, pp. 179-188, 1936.
18. Xu, X. (2021) "Explain the Predictions Using Rule Extraction via Causal index, Decision Tree and Ensemble Learning," 4rd ANU Bio-inspired Computing conference (ABCs 2021), paper 89, 7 pages, Canberra.