ECMM409: Nature Inspired Computation

CA 2 Individual Report

Eta Team
Candidate number: 046781
January 8, 2021

1 Contribution

In this team project, after confirming the topic of the team project, we need to propose some natural heuristic algorithms that can be applied to this problem. I propose to try RNN (Recurrent Neural Network) and LSTM (Long Short-Term Memory) to solve this binary classification problem, and conduct experiments based on these two methods. Although these two algorithms are not the main algorithms we finally chose, preliminary research has allowed us to compare the advantages and disadvantages of various algorithms and help choose the most suitable algorithm.

In terms of code, we first complete the basic code, and then we all modify it together. Based on the results of the group discussion, I conducted multiple experiments in the AIS (Artificial Immune System) model, changing different thresholds and death rates, and working with the team members to determine the best parameters.

2 Algorithm survey

Based on the discussion at the meeting, we compared performances of different methods which has shown on Table 1. Finally, we chosen AIS algorithm as the main method and the LSTM as the control method to solve the water quality detection problem.

The loss of LSTM remains basically unchanged after the rapid decrease in the previous several periods. We agree that the rapid decline may be caused by highly skewed data, while the flat area is caused by the high imbalance of positive and negative samples in the data set. The macro F1-score of this algorithm is 0.5582.

Method	Dataset	TP	FP	TN	FN	F1-score	Macro F1
LSTM	Test	74	**0**	**241870**	2694	0.05207600	0.52327
AIS	Test	**1075**	572	241328	1693	**0.48697622**	**0.74042**
AIS	Training	**820**	123	120471	920	**0.61125606**	**0.80347**
FITORE	Training	472	636	**120670**	556	0.44194757	0.71852
RISHAB and et al.	Training	487	9707	111599	**541**	0.08679380	0.52145
LAM (Crazy)	Training	242	24392	96914	786	0.01886057	0.45195
ROMAN and et al.	Training	234	32662	88644	794	0.01379554	0.42752
LAM (Basic)	Training	297	45920	75386	731	0.01257276	0.38814

Table 1. Performance comparisons between different methods.

On the other hand, according to a written reference, LSTM is more suitable for image processing and text classification problems[1], so we decided to use it as a control experiment. Compared with that, AIS has better performance, with macro F1-sore at 0.74042 in test dataset.

3 Experimentation process

I tried the LSTM method in the first stage, but LSTM is very complicated, so I didn't make the final result. After the group meeting, we established the LSTM model and evaluated it. The evaluation results have been described before.

Our main method is artificial immune system combine with probability [2]. First, in the training stage, we simulate a group of cells, each of which has a different sensitivity to the environment. Then we set some different thresholds for the cells according to the joint distribution of the data set. If the preset number of feature values in a water quality sample exceeds the threshold of some cells, then these cells will die.

In addition, we simulated the immunity of random cells in an abnormal water sample, and it would affect its neighbors. Among the samples classified as abnormal, some dead cells are randomly selected to improve their immunity to this feature, while the remaining cells lose confidence. In a normal sample, the immunity of some cells is reduced, and the remaining cells gain confidence.

If the death rate in the cell population exceeds the predetermined extinction rate, then the water quality sample is classified as abnormal. Finally, we compare the results of the model with the true values of the data set to evaluate the performance of the method.

In the original AIS model, although we obtained very good results in the training set, the performance in the test set was very bad. Figure 1 shows the confusion matrix of the first AIS model in the training set and Figure 2 shows the performance confusion matrix in the testing set.

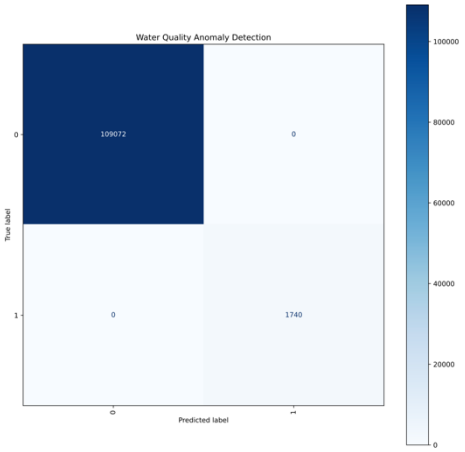

Figure 1. Training set

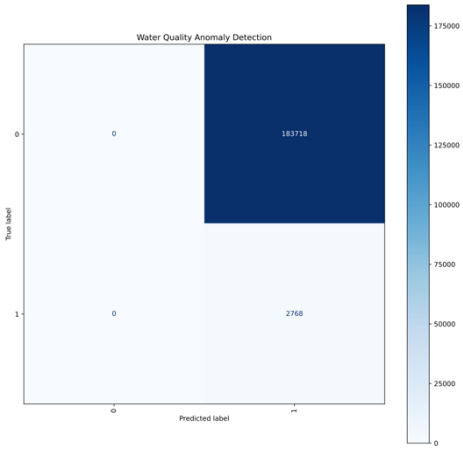

Figure 2. Testing set

Through the group meeting, according to the suggestions of the group members, I made the following adjustments in the original model:

i. In abnormal case, apart from gaining immunity on 25% of cells, remaining cells will lose confidence.
ii. In normal case, similarly, the cells will gain confidence, which will hence improve their probabilities.
iii. Given different death rate to each field instead of given a same. According to the correlation matrix, which is shown in Figure 3, it described how each field affecting the water quality. If it is a negative correlation, consider the cells whose probability is greater than the field death rate, if it is positive correlation, consider cells lower than the field death rate.

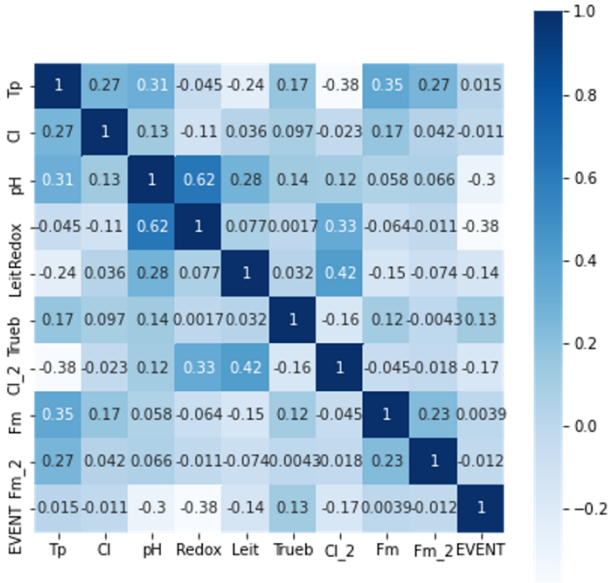

Figure 3. Correlation

Through the improved model, as shown in Table 1, very good results were obtained in the test set, macro f1-score of this method is 0.74042 and the standard f1-score is 0.48697622.

4 Teamworking process

In this team project, all members actively participated and made certain contributions. At the beginning of the project theme selection, everyone proposed items of interest, such as the traveling thief problem, DOTA2 AI Shadow Fiend Battle and 2048 Mini Games. Then we analyzed the difficulty of these problems through the voice conference and finally voted to select the water quality detection problem that is suitable for natural inspired algorithms to solve.

In the process of project realization, team members did not have a particularly obvious division of labor but made contributions in their respective areas of expertise. First, propose and evaluate the solutions that can be applied to solve the water quality testing

problems, and then write the basic code, and then everyone will continue to improve the methods on these foundations.

The biggest challenge is the determination of the main algorithm and the remote cooperation of the team members to complete the project, but at the same time this is also the biggest gain and fun of this project. Since there are many algorithms that can solve classification problems, it took us a lot of days to determine some solutions that belong to natural heuristic algorithms. In the process of implementation, we found that the effect of splitting the code into many blocks for different people is not good, which makes our code messy and difficult to adjust, so we choose to complete the basic code first and then modify it together.

5 Conclusion

In general, I think our natural heuristic method is successful because it performs well on both the training set and the test set and is better than GECCO's 2007 water quality testing challenge and achieved the expected results. In terms of teamwork, I think it can be improved. Due to the epidemic, team members can only communicate remotely. Most of the time, team members are busy with their own research and difficult to communicate in time. This makes our efficiency lower and some ideas cannot be displayed well. If we have more time, and face to face cooperate, I think our method will be further enhanced.

There are still some good algorithms that we have not been able to try, such as ant colony algorithm and particle swarm algorithm, both of which are representative natural heuristic algorithms, which are usually used to solve optimization problems. The ant colony algorithm simulates the intelligent behavior of the ant colony, combining time and pheromone to describe an optimal solution. In this problem, we can regard the different features in the water quality sample as the different destinations of the ants, and the value of the feature is the pheromone. If we can design a fitness function that can reflect the quality of water, and we can set one or more thresholds as a criterion for judging water quality by analyzing the training set data, the fitness function above the threshold is normal, and the fitness function below the threshold is unusual. However, because the ant colony algorithm tends to take the high pheromone as the optimal solution, but the different features in the water quality sample are not necessarily the higher the value, the better the water quality, so we also need to combine the correlation coefficient matrix analysis and calculate the pheromone based on the positive and negative correlation.

6 Reference

[1] WOO, S., TAY, L., & PROCTOR, R. (Eds.). (2020). Big Data in Psychological Research. Washington, DC: American Psychological Association. doi:10.2307/j.ctv1chs5jz

[2] Ying, K., & Lin, S. (2014). Efficient wafer sorting scheduling using a hybrid artificial immune system. The Journal of the Operational Research Society, 65(2), 169-179. Retrieved January 10, 2021, from http://www.jstor.org/stable/24502033