

Evaluation of the Classification of Higher Education Institutions by Multiclass Logistic Regression and Entropy

Masaaki Ida

National Institution for Academic Degrees and Quality Assurance of Higher Education, Tokyo, Japan

ida@niad.ac.jp

Abstract— This paper presents evaluation of the classification of higher education institutions by multiclass logistic regression and entropy. We examined the relationship between data item selection and entropy. By examining various item selection situations using real data, we gained insight into item selection and classification evaluation based on the entropy and f1-score of each class.

Keywords—Multiclass, Logistic regression, Eigenvalue distribution, Entropy, Higher education institution.

I. INTRODUCTION

In recent years, the scope of data science applications has rapidly expanded to various fields, including business, finance, education, and other social sciences. In addition, the amount of data and the number of data items being handled are becoming enormous. In this paper, financial data of higher education institutions is investigated as such large data set.

Multiclass logistic regression analysis has been used as a basic method for various multi-class classification problems [1], including image classification problems. In recent references, nonlinear randomized multiclass logistic regression analysis was proposed as a method to improve the classification accuracy of multiclass logistic regression by using randomization and nonlinearization [2].

Multiclass logistic regression analysis has been developed as a theoretically sophisticated method as an extension of regression analysis. Numerous applied studies on multiclass classification have been conducted. However, there are still issues to be resolved regarding the selection of data items and practical methods for evaluating classification quality.

In this paper, we focus on the relationship between f1-score which is related to accuracy, and entropy, and examine the followings:

- We use concrete examples to examine how item selection affects results.
- We also consider suggestions for improving classification.

II. MULTICLASS LOGISTIC REGRESSION

First, we will discuss the logistic regression [1]. Logistic regression is a well-known statistical method for predicting the probability of an event from multiple explanatory variables.

Along with support vector machine, it is widely used in multivariate analysis and supervised classification tasks in machine learning. The output of the method is a value between 0 and 1, that can be regarded as probability of class membership. Logistic regression produces this prediction based on a linear combination of features.

Multiclass logistic regression uses logistic regression as multiple classifiers simultaneously to perform multi-class classification tasks.

The formulation is as follows. The q -dimensional training data $\mathbf{x}$ is expressed as belonging to a certain class by 0 or 1. For such data set, the formulation of logistic regression is as follows. The linear combination of the inputs is $\theta^i \mathbf{x}$ in logistic regression for a class i , the output is a probability p^i ,

$$p^i = \sigma(\theta^i \mathbf{x}) \quad (1)$$

where $\sigma(t)$ is a Sigmoid function:

$$\sigma(t) = \frac{1}{1 + \exp(-t)} \quad (2)$$

The regression coefficient θ is determined by minimizing the cost function. By applying this method to the test data corresponding to each class, multiclass logistic regression coefficients can be obtained.

The numerical calculations in this paper partially utilize the Python libraries, which are popular tools for data analysis and machine learning [3]. We use these libraries with standard settings on solver, tolerance, etc. In this setup, multinomial type logistic regression is used.

III. EIGENVALUE DISTRIBUTION AND F1 SCORE

In general, when there are many data items in a classification problem, the selection of which data items to use is extremely important both theoretically and practically (data item selection problem). There are various methods for analysing the impact of the selection of data items on classification results. In this paper, the selection of data items is investigated by spectral analysis of the input data matrix to logistic regression.

For example, Figure 1 shows an example of the eigenvalue distribution of input data. This is for the MNIST dataset (a collection of images of handwritten characters from 0 to 9,

<https://www.tensorflow.org/datasets/catalog/mnist>), which is often used in data science experiments. In this paper, 1,000 training data were randomly taken, and the figure shows the eigenvalue distribution of the variance-covariance matrix of input data matrix.

As shown in this figure, there are a few large eigenvalues on the right side. These large eigenvalues are the dominant eigenvalues that represent the characteristics of the entire data matrix, and their eigenvectors indicate the direction of the dominant features of the data. The small eigenvalues on the left side are clustered, which is related to randomness. Also, fluctuations in these eigenvalue sets affect the overall calculation, so they need to be carefully considered.

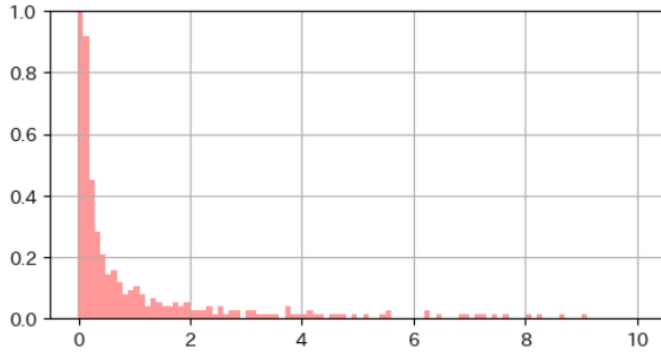


Figure 1. Eigenvalue distribution (eigenvalues and density)

Evaluating the distribution of eigenvalues, which have various important meanings, is essential for understanding the properties of a dataset. Possible evaluations of the input data matrix include the trace, determinant, and statistical indices such as the mean and variance of the matrix.

In this paper, we consider normalized spectral entropy (nsp) of the eigenvalue distribution defined as follows:

If each eigenvalue is λ_i , then nsp is as follows:

$$nsp = - \sum_{i=1}^n p_i \log_2 p_i / \log_2 n \quad (3)$$

$$p_i = \lambda_i / \sum_{i=1}^n \lambda_i \quad (4)$$

The following relationship can be seen between spectral entropy and clustering:

A large eigenvalue indicates the variance of the data in the corresponding direction. On the other hand, a large spectral entropy indicates that the variance is spread fairly evenly across multiple directions. Conversely, a small spectral entropy indicates that a few directions dominate.

In the impact on clustering, when entropy is high, data spreads in many directions, making clustering difficult. Conversely, when entropy is low, data tends to concentrate in a specific direction, making cluster boundaries clearer and making clustering easier.

In other words, the spectral entropy of the covariance matrix measures the diversity of the spread of the data and evaluates the clarity of the clusters.

In multi-class classification problems, metrics that measure the accuracy of classification are important. For the input data matrix, we consider accuracy. In addition, metrics that measure the accuracy of each class include precision, which indicates how many of the predictions were correct, and recall, which indicates how many truly correct predictions were made without missing any.

In this paper, we consider the *f1-score*, which is a metric that combines these two.

$$f1\text{-score} = 2 (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

Figure 2 shows a numerical example for f1-score and normalized spectral entropy. This figure is for the MNIST dataset mentioned above. The horizontal axis is the f1-score, and the vertical axis is the spectral entropy. The numbers in the circles correspond to single digits (e.g., 1, 2, 3, ...).

What can be seen from this figure is that for 1 and 0, which have high f1-scores, and the entropy value is low, that indicates stable classification. On the other hand, for 5 and 8, the entropy is high and the f1-score is low. This case indicates that many feature directions are required to represent these complex characters, which increases the entropy and makes recognition difficult.

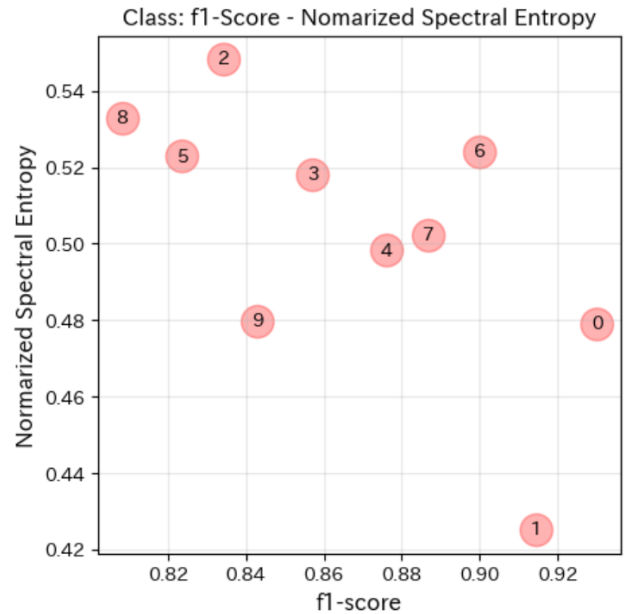


Figure 2. f1-score and spectral entropy for MNIST dataset

Generally, when entropy is low, the data is biased in a small number of directions, and if the data items related to that direction are selected, classification becomes easier, so it is thought that the f1-score will also increase. In other words, the figure shows a line sloping downward to the right. However, this does not necessarily the case. The two types of indicators

characterize datasets in different ways and require detailed consideration on a case-by-case basis.

IV. UNIVERSITY CLASSIFICATION ON FINANCIAL DATA

In this section, we consider the real-world application of classification using multiclass logistic regression. Note that this study is currently underway (work in progress), and the interpretation of the detailed discussion may change as it is refined in the future by the experts' knowledge in finance field.

As mentioned above, in classification problems, when there are many data items, the selection of data items from among them is the issue. It is important to analyse the impact of data item selection on the results. Here, we consider the relationship between the selection of data items and the spectrum of the data matrix mentioned in the previous section.

In this section, we apply this to the actual finance field. We attempt to classify higher education institutions (86 national universities in Japan) based on their financial indicators.

Various classifications of Japanese national universities have been considered. In this section, we consider a classification based on typical characteristics as follows [4]. This classification is based on the characteristics: history, size, affiliated university hospital, and field of study of the university.

TABLE 1. CLASSIFICATION BY CHARACTERISTICS

Class	Classification (university groups)
1	Former Imperial Universities
2	Large Universities (with university hospital)
3	Large Universities (no hospital)
4	Science and Engineering Universities
5	Universities of Humanities
6	Medical Universities
7	Education Universities
8	Graduate Universities

Let's explain the data in a little more detail. For example, Figure 3 is a scatter plot of two indicators for a particular fiscal year. The horizontal axis represents the debt composition ratio, which represents flow, and the vertical axis represents the subsidy dependence ratio, which represents stock. The color of each point corresponds to the eight university classes or groups.

As can be seen from this figure, universities from the same class tend to be located relatively close to each other, and similar classes tend to be located close. However, there are also cases where this does not necessarily follow this trend.

In this paper, we consider 33 types of financial indicators (data items). We used the data covering eight years from fiscal year 2012 to fiscal year 2019. Since this is a research paper, the names of the universities have been anonymized.

We selected several financial indicators from the 33 available. Below, we refer to these indicators as data items. We used six years of data from 2012 to 2017 as training data and two years of data from 2018 and 2019 as test data, and performed classification using multiclass logistic regression.

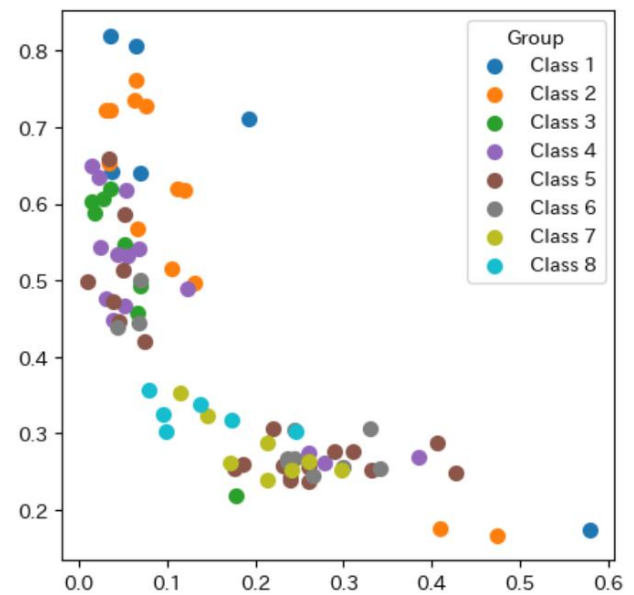


Figure 3. Scatter plot of universities for a certain year (Debt composition ratio, and Grant dependency, colors represent university classifications)

V. NUMERICAL EXPERIMENTS

In this section, we examine the relationship between the selected data items (financial indicators), accuracy, f1-score and entropy. The numbers in the following figures indicate the corresponding university classes in Table 1.

(1) Selection of two indicators (Fig. 4)

Indicators: Debt composition ratio, Grant dependence.

Normalized entropy of covariance matrix: 0.991.

Accuracy: 0.657.

In this simple setting, the accuracy is low due to the small number of data items (indicators). Especially, class 3 and 4 have high entropy value.

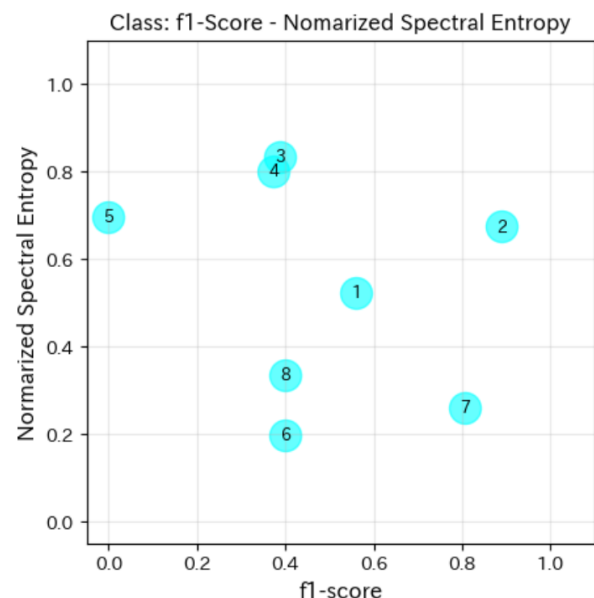


Figure 4. f1-score and spectral entropy for two data items

(2) Selection of four indicators (Fig. 5)

Indicators: Debt composition ratio, Grant dependency ratio, Education expenses ratio, Research expenses ratio.

Normalized entropy of the covariance matrix 0.859.

Accuracy: 0.889.

As the number of selected items (indicators) increased, the discrimination performance improved and accuracy increased. Regarding entropy, 1, 6 and 7 are low, and it is thought that there is a direction that clearly expresses each characteristic and so stable classified. Class 1, 6 and 7, shown in the bottom right of the figure, are considered to be easier to discriminate because high f1-score and low entropy.

The f1-score for class 5 tends to be significantly lower (outlier). This is likely due to the fact that class 5 is the class of humanities, and there is considerable variety in each university.

Class 3 and 4 have high entropy, and it is thought that there are multiple directions that express the data feature of the classes. As for entropy, low entropy does not necessarily mean high accuracy; it is possible for entropy to be relatively high for classification to be based on multiple criteria.

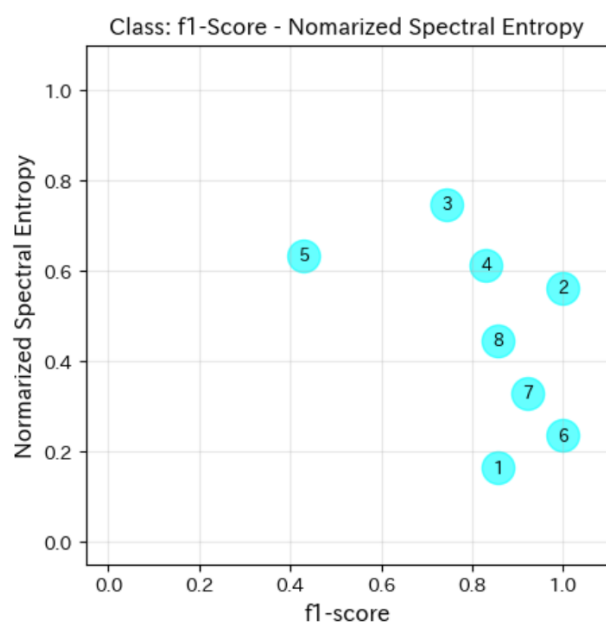


Figure 5. f1-score and spectral entropy for four data items

(3) Example of eight indicator selection (Fig. 6)

Indicators: Debt composition ratio, Grant dependency ratio, Education expenses ratio, Research expenses ratio, Research expenses per full-time faculty member, Growth rate of operating income, Personnel expenses ratio, Self-generated income ratio.

Normalized Entropy of Covariance matrix: 0.724.

Accuracy: 0.913.

Although the accuracy and whole entropy have improved, as can be seen from Fig. 5 and 6, there is little change in entropy of each class. Therefore, the data structure is thought to be similar to that in Figure 5.

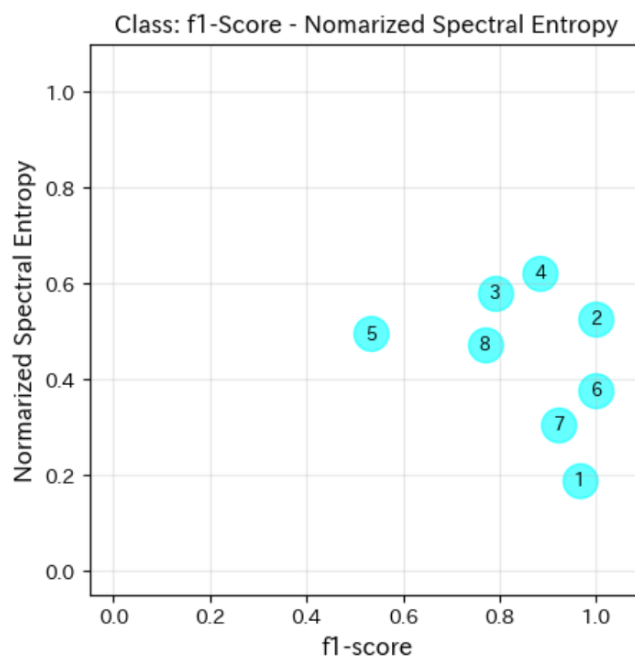


Figure 6. f1-score and spectral entropy for eight data items

(4) Selection of all (33) indicator selection (fig. 7)

Normalized Entropy of Covariance matrix: 0.725

Accuracy: 0.977

Basically, the f1-score increases as the number of data items increases. However, the entropy is not much different from that in Figure 6. It is thought that there is not much change in the data structure. Therefore, the appropriate number of items (indicators) to represent the data structure is thought to be smaller.

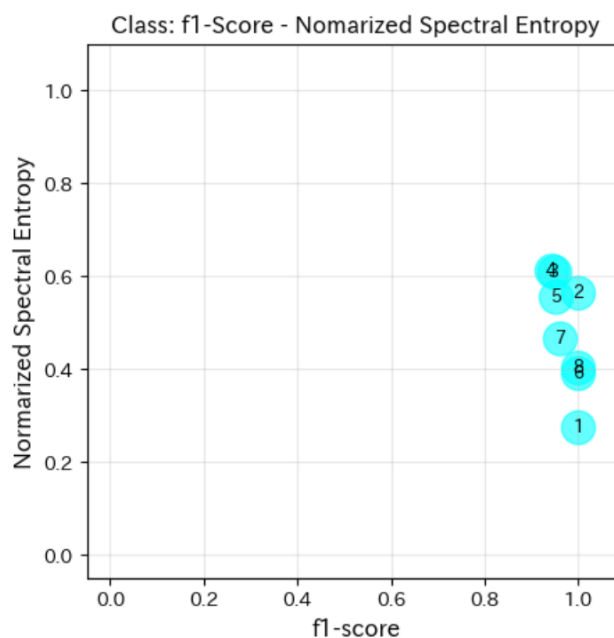


Figure 7. f1-score and spectral entropy for 33 data items

(5) Selections of (other) four indicator selection (Fig. 8 and 9)

Next, we investigate the differences in the selected items. In case of the number of indicators is 4, we selecte diffrent indicators from the case of (2):

Indicators: Debt composition ratio, Operating cash flow to future burden ratio, Current ratio, Quick ratio.

In this case, the selected indicators are from the set of similar financial indicators. Therefore, the f1-score is not so mutch high and the entropy is high.

Next, in case when we select the following four items:

Indicators; Maintenance and management cost ratio, Utilities cost ratio, Labor costs per full-time administrative and technical staff, and General administrative cost ratio.

These indicators are related to the management indicators common to the entire university set, and are not closely related to university classification. Therefore, as a result, the f1-score further decreases and entropy increases as seen from the Fig.9,.

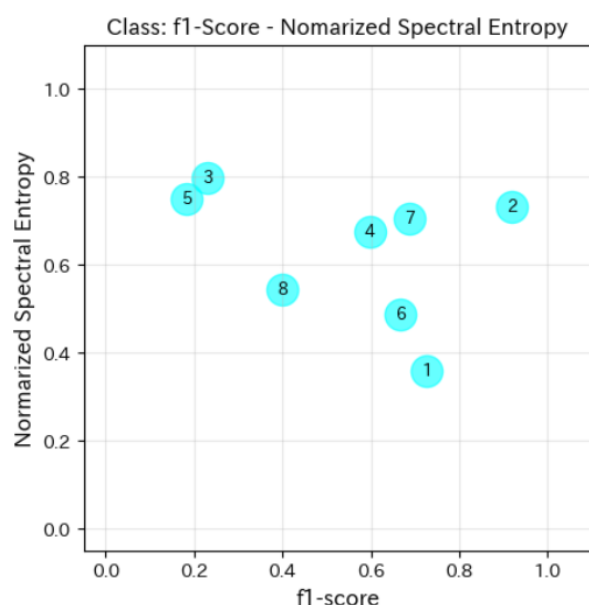


Figure 8. f1-score and spectral entropy for other four indicators

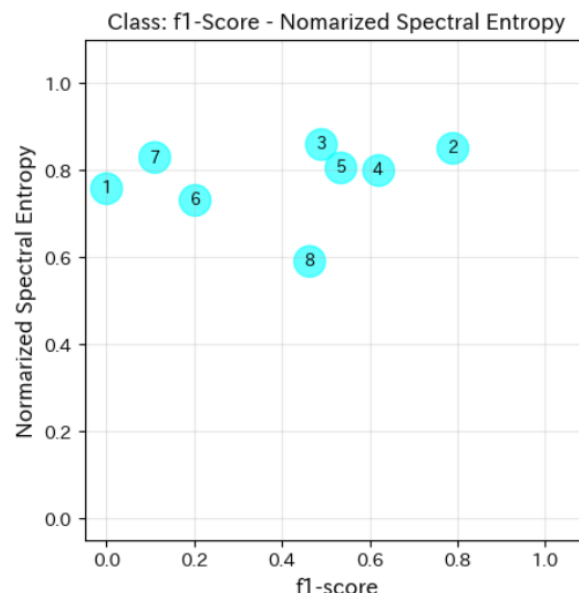


Figure 9. f1-score and spectral entropy for other four indicators

VI. CONCLUSIONS

In this paper, we investigated the effective evaluation method for classification by multiclass logistic regression. We also examine the relationship between data item selection and entropy.

By examining multiple item selection situations using real data, we gained insights into the evaluation method of data item selection and classification using the entire data and class data, entropy and f1-score. These insights were verified using specific numerical examples.

We hope that the knowledge and findings in this paper will lead to the refinement of various data classification methods, and further development of the application of big data analysis in the field of social science. We also hope that this analysis of organizational classification will provide practical recommendations for higher education administration.

REFERENCES

- [1] J. M. Hilbe, Practical Guide to Logistic Regression, Chapman and Hall/CRC , 2015.
- [2] M. Ida, Improvement of Multi-Class Logistic Regression by Dimensional Change, 2025, International Workshop on Intelligent Systems, doi: 10.1109/IWIS66215.2025.11142445, 2025.
- [3] Scikit-learn Python libraries for data analysis and machine learning, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed Dec 1, 2025).
- [4] NIAD-QE, National University Corporation Finances, URL: https://www.niad.ac.jp/support/university_finance/ (accessed Dec 1, 2025).