Development and application of a simple machine learning algorithm for multiclass classifications

Yukako Toko (ytoko@nstac.go.jp)*, Toshiyuki Shimono (tshimono@nstac.go.jp)*,
Kazumi Wada (kwada@nstac.go.jp)*

Keywords: Multiclass classification, Autocoding, Naïve Bayes, Natural language processing

1. INTRODUCTION

1.1. Background and Proposal

Classification is often required in the process of survey statistics tabulation, and autocoding systems can contribute to the efficiency of classification. High-performance autocoding systems shorten the processing time and reduce costs. We have developed an autocoding system using a simple machine learning method for the Family Income and Expenditure Survey. This system works well as compared to a commercial classification system that uses a deep learning algorithm. This paper briefly introduces the technological methodology and the results including a comparison with the commercial system. The detail of the method and the results will be described in the full paper.

1.2. Machine learning

Machine learning is a 'field of study that gives computers the ability to learn without being explicitly programmed' and was first proposed by Arthur Samuel in 1959. Machine learning algorithms are often categorised by whether they are supervised or not. Both supervised and unsupervised algorithms search through data to look for patterns.

Currently, machine learning algorithms are being applied in many situations in business due to their following advantages.

- Rapid processing (faster than humans)
- Reduced costs
- Use of past data via learning.

The results of official statistics are expected to be published swiftly with high accuracy and involve enormous amounts of past data. Therefore, it would be worthwhile to consider applying machine learning algorithms to such a field.

1.3. Classification for the Family Income and Expenditure Survey

The Statistics Bureau in Japan conducts the Family Income and Expenditure Survey every month. The selected households are requested to manually fill out the family account books including both daily income and expenditures. In the process of tabulation, each item in the books is manually classified by trained staff into approximately 570 classes. Although it would be possible to reduce the processing time and costs, the automation of this survey's classification has not yet been performed due to the complexity of the classificatory criterion.

* National Statistics Center, Japan

2. METHODS

2.1. Basic concepts of our autocoding system

The basic concepts of our autocoding system are listed below:

- High rate of assignment with high accuracy
- Simple algorithm
- Swift processes

A high rate of assignment with high accuracy is essential to apply this system to official statistics. Official statistics are expected to provide results with high reliability because their results influence the decision-making process both in public and private sectors. The applicability of our system to other classification tasks depends on the simplicity of the algorithm. It is desirable to develop a versatile system to reduce development costs; therefore, each algorithm needs to be as simple as possible. Swiftness of the processes is another crucial factor for practical applications.

2.2. Learning method

The learning system is an application of a supervised algorithm. The system learns patterns of features from labelled training data. The main processes are comprised of morphological analysis, feature creation, and tabulation of the features using a label.

2.3. Classification method

The classification system borrows the idea of the Naïve Bayes classifier, although the algorithm does not simply apply the Naïve Bayes classifier method. The algorithm selects the prospective labels for each record together with the confidence scores for each label. The most promising label is output with its confidence score. The main processes are comprised of morphological analysis, feature creation, extraction of potential labels, calculation of confidence scores, and assignment of the most promising label.

3. RESULTS

3.1. Data

The volume of the labelled training data is approximately 4.56 million records and the data covers the entire year. An entire year's data are essential to achieve a high accuracy classification because the seasons have a strong influence on household consumption. Moreover, the prepared evaluation dataset for the classification contains approximately 0.65 million records, and it also covers an entire year.

3.2. Overview

Table 1 shows the performances of our system and a system using a deep learning algorithm, which was developed by a private firm for the same classification purpose. Our classifier succeeded in assigning labels to 99.3% of the records in the evaluation data, and 91.8% of the labelled records were given the correct assignments. In addition, after ordering all the records using the confidence score, the top 80.0% retained a correct prediction rate of 98%. Similarly, the top 92.8% retained a correct prediction rate of 95%. Currently, 98% is comparable to the accuracy of the labels assigned by trained staff. Therefore, only 20% of the data needs confirmation or assignment by trained staff because our system can classify approximately 80% of the data without manual confirmation.

Conversely, we find that the accuracy of our system does not differ significantly from the system using deep learning algorithm. The assignment rate of that system retaining a correct prediction rate of 98% was better than ours. However, with regards to the total performance, our autocoding system performed better than theirs.

Table 1. Results of the label assignment of our autocoding system and the deep learning system

		The number of records of data-set (1)	The number of assignments (2)	The number of correct prediction assignments (3)	The number of incorrect prediction assignments	The correct prediction rate (3) / (2)	The assignment rate (2) / (1)
Our autocoding system	Total	651,999	647,748	594,412	53,336	91.8%	99.3%
	98%-correctness area	651,999	521,730	511,295	10,435	98.0%	80.0%
	95%-correctness area	651,999	604,752	574,514	30,238	95.0%	92.8%
The deep learning system	Total	651,992	596,395	560,477	35,918	94.0%	91.5%
	98%-correctness area	651,999	532,766	522,049	10,717	98.0%	81.7%

3.3. Processing time

Table 2 shows the processing time of our system and the deep learning system. Our learning system processed 0.5 million records in approximately 2 min. 41 sec. and processed that classification system in 2 min. 35 sec. Although the processing time depends on the volume of the data, this is certainly faster than manual classification.

In addition, the processing time of our autocoding system was remarkably faster than that of the deep learning system, although their system comprises multiple layers to consider multiple factors in the input data; it already has the ability to be applied to more complicated classifications.

Table 2. Processing time of our autocoding system and the deep learning system

	condition	Processing time for learning for 0.5 million records	Processing time for classifying for 0.5 million records
Our autocoding system	Xeon, 3 GHz	2 min. 41 sec.	2 min. 35 sec.
The deep learning system	8 cores, 8 parallel processing	6 hour	2 hour

4. CONCLUSIONS

4.1. Summary of the findings

We presented an autocoding system using a simple machine learning algorithm. It succeeded in classifying the data into hundreds of categories with high accuracy. Its performance does not differ substantially from a commercial system using a deep learning algorithm. Furthermore, our system can be utilized to process data efficiently due to the swiftness of its processing speed.

4.2. Further discussions

To improve its accuracy, our system will require several additional techniques, such as those listed below.

- More relevant methods to choose the most appropriate label from several candidates
- Method to consider further information to assign labels such as family structure and the occupations of family members
- A method to consider additional orthographical variants (e.g., notation shaking)
- Knowledge of relevant editing techniques for big data

Although the algorithm requires further improvements, our system has the potential to be applied to broader fields in official statistics, such as occupational classification, industrial classification and other types of classification. In addition, it would contribute to improving official statistics.