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

Yukako Toko, Toshiyuki Shimono, Kazumi Wada
National Statistics Center, Japan

Introduction

Development an **auto-coding system**

Multiclass classification,
can classify into more than **570** classes.

For *"The Family Income and Expenditure survey"* – The Statistics Bureau of Japan

Concepts

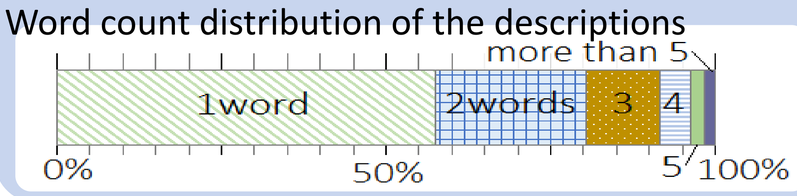
1. **High accuracy** with **high coverage**
2. **Quick** processes
3. **Simple** algorithm

Data

Approx. **4.56** million records for training
Approx. **0.65** million for evaluating

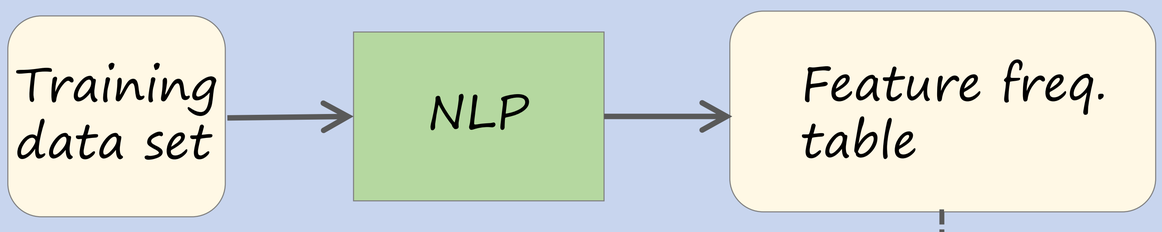
Free format, short descriptions on each item in the family account books

Word count distribution of the descriptions



Method - Learning part -

Supervised algorithm

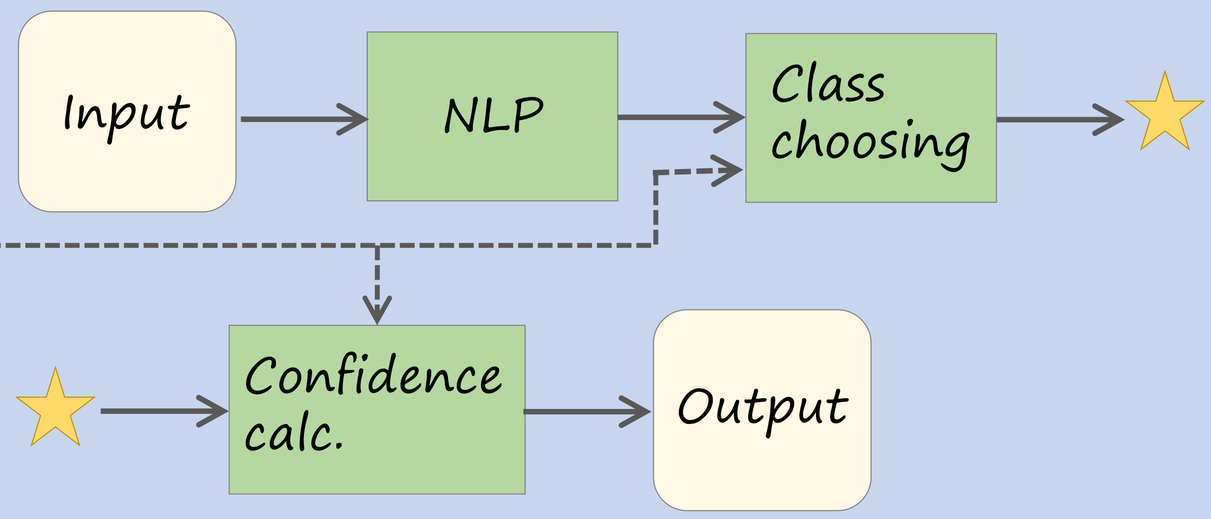


NLP (Natural Language Processing) ≈ Tokenizing:

1. Dividing the descriptions (training data set) into words
2. Feature extraction : unigrams, bigrams, and the entire sentences

Method - Classifying part -

Borrowed the idea of the Naïve Bayes classifier

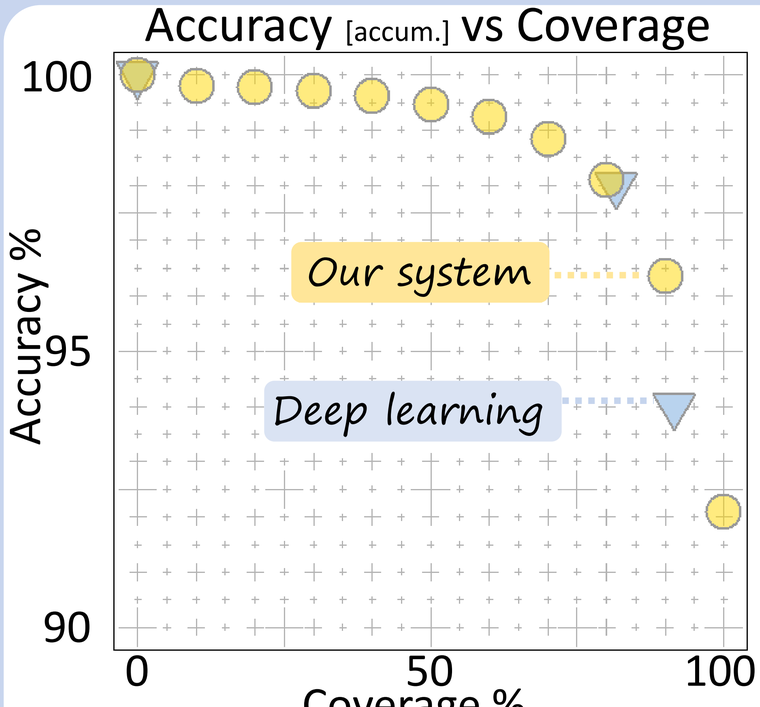


Class choosing: selecting the most promising one among the extracted prospective classes

Confidence calculation: calculating the confidence score of each output record

Experimental Result

Accuracy [accum.] vs Coverage



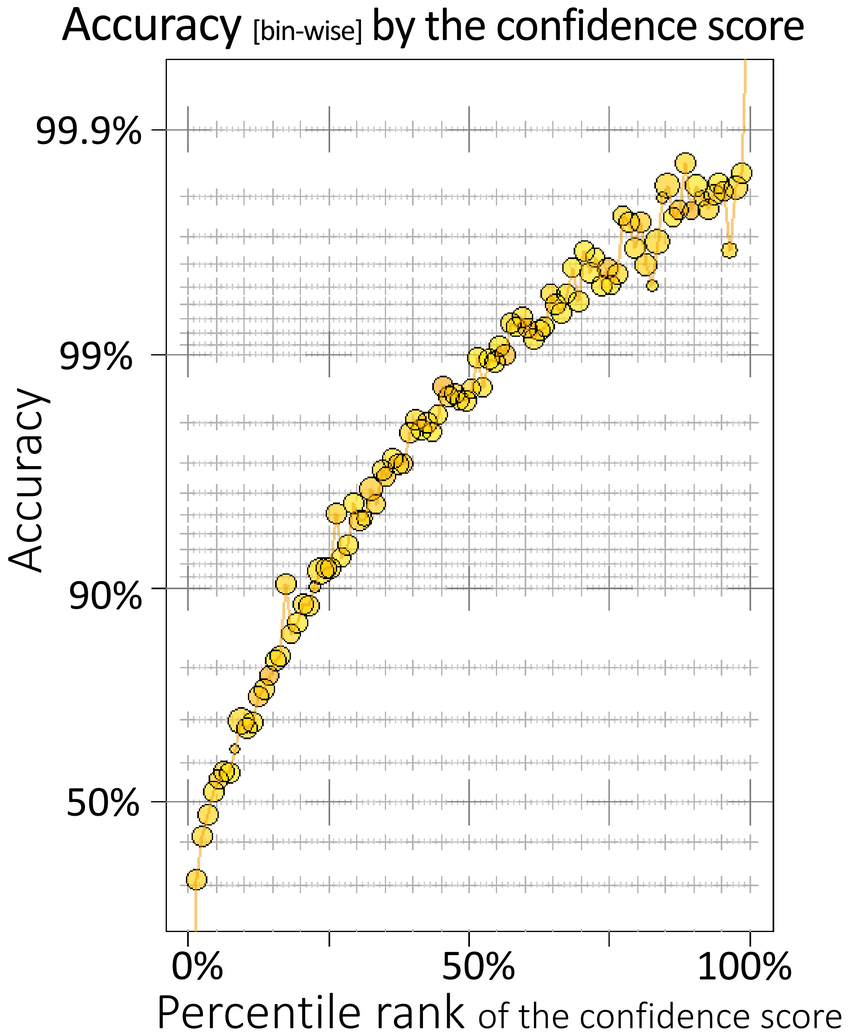
After sorting in the order of the confidence scores

The top 67% of dataset have retained accuracy of 99%

The top 80% of dataset have retained accuracy of 98%

		Total records (1)	Assigned records (2)	Correct prediction (3)	Coverage (2)/(1)	Accuracy (3)/(2)
Our system	98%-correctness area	651,999	521,730	511,295	80.0%	98.0%
	95%-correctness area	651,999	604,752	574,514	92.8%	95.0%
	Whole data	651,999	647,748	594,412	99.3%	91.8%
Deep learning system	98%-correctness area	651,999	532,766	522,049	81.7%	98.0%
	Whole data	651,992	596,395	560,477	91.5%	94.0%

Accuracy [bin-wise] by the confidence score



Running time

CPU : Xeon, 3 GHz

For learning :
< 6 min. / million records

For classifying :
< 6 min. / million records

Development cost

No cost for software and licenses

- Cygwin (GPL v.3)
- Perl (GPL,Artistic-1.0)
- MeCab (GPL, LGPL, BSD)
  MeCab is a morphological analyzer

Future works

Natural language problems e.g. synonyms, low freq. terms

Developing an algorithm considered further information e.g. family structures, regions, occupations

Improving our algorithm to select a more relevant one among multiple candidate classes