# Knowledge Discovery Solution for CDMC 2015

Yuki Maruno and Eri Nakahara

Abstract
CDMC 2015 is a data mining competition to solve advanced, real-world problems. Participants were asked to solve three tasks: e-News2015 categorization, Trade Me Comments Sentiment Classification, and Complex Disease Diagnosis. We took part in the competition and our team is the first place winner of CDMC 2015. This paper describes our solution for the competition.
Key words：CDMC2015, Data Mining Competition.

## 1. Introduction

The 6th International Cybersecurity Data Mining Competition (CDMC 2015) is a challenging, multi-month research and practice competition, focusing on application of knowledge discovery techniques to solve advanced, real-world problems. The competition is associated with the 8th International Workshop on Data Mining and Cybersecurity (DMC2015), which is an associated event to the 22nd International Conference on Neural Information Processing (ICONIP2015), Istanbul, Turkey.

In this competition, participants are required to solve all of the following tasks, Task 1: e-News2015 categorization, Task 2: Trade Me Comments Sentiment Classification, and Task 3: Complex Disease Diagnosis.

The four of the third year students whose major is programming participated in the competition and successfully got the first place. It was essential for us to perform data visualization and statistical analysis to solve these challenging tasks because the students did not have any experience of machine learning. We took enough time to perform these analyses in order to investigate which method/technique is suitable for each task by using Ruby and R, which are open-source programming languages and suitable for text mining and data analysis. The following sections describe our solution in detail.

# 2. Task 1: e-News2015 categorization

## 2－1. Task Description

The e-News2015 dataset was collected from nine newspapers: The New Zealand Herald [1], The Australian [2], The Press [3], Yahoo News [4], BBC [5], The New York Times [6], The Independent [7], The Times [8], and Herald Sun [9], on five topics of business, entertainment, sport, technology, and travel, respectively.

Each document of the dataset was labelled manually by skimming over the text in advance. The objective of this task is to classify these documents into five categories for each testing instance. In the provided data files, each document was formatted as one line pure text. The punctuation and stop word were removed in advance. Fig. 1 shows the example of the data.

GeI kUXU RxjIek Xxe mAeOWrjJOeAX DUXe mAeTWezXekrJ
uxe ajMeDAOeAX ZAXeAkR Xj
QmaYJ VDmRUkeDR zUA RXZrr XjW Gc

Fig. 1. Example of the data

The statistical information of the training dataset is summarized in Table 1.

Table 1. The statistical information of the training dataset

| Topic | N. of Documents |
|---|---|
| Business | 256 |
| Entertainment | 266 |
| Sport | 263 |
| Technology | 281 |
| Travel | 273 |

## 2－2. Our Method

As shown in Fig. 1, each text for this task has been encrypted. We first restored the encrypted text to its original form. In order to make a decoding table, we focused on the word frequencies in English, that is, how many times each word appears in a particular text.

We made a list of letters and words in frequency order with e-News2015 training data, which includes 581,957 words. For comparison, we also made a list of words with Alice 's Adventures in Wonderland [10], which includes 29,461 words. Table 2 (a) shows the top 20 words of e-News2015 training data, and Table 2 (b) shows the top 20 words of Alice's Adventures in Wonderland.

Compared with Table 2 (a) and (b), we conducted a step by step interpretation to make a decoding table for e-News2015 data. The encrypted data in Table 2 (a) was decoded to its original form as shown in Table 3. For validation of our decoding table, we decoded the whole text of training data. Finally, we got a decoding table as shown in Table 4. After validation, we decoded the test data with our decoding table. The encrypted data example (Fig. 1) in the previous section is decoded to its original form as shown in Fig. 2.

Once we get a decoding table, we can obtain the class label by searching the source of the document with the first few sentences of the article since the newspaper articles are openly available on the web.

Table 2. (a) Top 20 words of e-News2015 training data, (b) Top 20 words of Alice 's Adventures in Wonderland

(a)

| Index | Word | Count | Index | Word | Count |
|---|---|---|---|---|---|
| 1 | Xxe | 31831 | 11 | IZXx | 4477 |
| 2 | Xj | 16596 | 12 | IUR | 3974 |
| 3 | UAk | 14998 | 13 | ZX | 3543 |
| 4 | U | 14866 | 14 | UX | 3330 |
| 5 | jw | 13943 | 15 | UR | 3264 |
| 6 | ZA | 11035 | 16 | Ye | 2993 |
| 7 | wjD | 6009 | 17 | wDjO | 2877 |
| 8 | ZR | 5625 | 18 | RUZk | 2766 |
| 9 | jA | 5435 | 19 | YJ | 2642 |
| 10 | XxUX | 5327 | 20 | UDe | 2616 |

(b)

| Index | Word | Count | Index | Word | Count |
|---|---|---|---|---|---|
| 1 | the | 1672 | 11 | Alice | 386 |
| 2 | and | 839 | 12 | you | 352 |
| 3 | to | 789 | 13 | was | 350 |
| 4 | a | 666 | 14 | that | 261 |
| 5 | of | 600 | 15 | as | 255 |
| 6 | she | 499 | 16 | her | 242 |
| 7 | it | 476 | 17 | with | 220 |
| 8 | said | 452 | 18 | at | 214 |
| 9 | in | 411 | 19 | on | 200 |
| 10 | I | 396 | 20 | all | 186 |

Table 3. Encrypted words to original form

| Encrypted | Original | Encrypted | Original |
|---|---|---|---|
| Xxe | the | IZXx | with |
| Xj | to | IUR | was |
| UAk | and | ZX | it |
| U | a | UX | at |
| jw | of | UR | as |
| ZA | in | Ye | be |
| wjD | for | wDjO | from |
| ZR | is | RUZk | said |
| jA | on | YJ | by |
| XxUX | that | UDe | are |

Table 4. Decoding table for e-News 2015 data

| Encrypted | Original | Encrypted | Original | Encrypted | Original | Encrypted | Original |
|---|---|---|---|---|---|---|---|
| A | n | N | U | a | g | n | b |
| B | q | O | m | b | Q | o | s |
| C | G | P | H | c | Z | p | p |
| D | r | Q | R | d | r | q | j |
| E | X | R | s | e | e | r | l |
| F | F | S | k | f | A | s | K |
| G | N | T | x | g | m | t | V |
| H | W | U | a | h | D | u | T |
| I | w | V | C | i | E | v | Y |
| J | y | W | p | j | o | w | f |
| K | I | X | t | k | d | x | h |
| L | z | Y | b | l | O | y | L |
| M | v | Z | i | m | u | z | c |

New data showed the unemployment rate unexpectedly
The goverment intends to
Rugby Crusaders can still top NZ

Fig. 2. Example of the decoded data

## 3. Task 2: Trade Me Comments Sentiment Classification

### 3－1. Task Description

The data was collected from Trade Me [11], a New Zealand famous online shopping website. Comments from seller and buyer were collected in pure text, and features were extracted by counting the number of occurrence for 25 key words. The goal of this task is to classify these comments into three sentiment categories: Positive, Negative and Neutral. The given training dataset is a matrix in 26 columns. The last column is the class label identified as an integer in the range of 1-6: Buyer Positive (1), Buyer Neutral (2), Buyer Negative (3), Seller Positive (4), Seller Neutral (5), and Seller Negative (6).

The statistical information of the training data is summarized in Table 5.

Table 5. The statistical information of the dataset

| Comments from | N. of Positive | N. of Neutral | N. of Negative |
|---|---|---|---|
| Buyer | 27739 | 2671 | 4291 |
| Seller | 22191 | 3745 | 3442 |

### 3－2．Our Method

Based on our investigation of the data, we added two more features, which were the number of non-zero elements for original 25 features, and the sum of the number of all features. In total, 27 features were used.

For the prediction, we used classification tree [12], which is a machine-learning method for constructing prediction models from data. Fig. 3 shows the part of the classification tree trained with the training data. We applied the models on test data to obtain the class label.

## 4．Task 3: Complex Disease Diagnosis

### 4－1．Task Description

The dataset is a collection of clinical data on bowel, eye, heart, liver and lung diseases. The given features represent symptoms, background characteristics and disease situations of patients. The objective of this task is to predict the type of disease for each testing instance. Each training data gives a numerical matrix with the last column as the type of disease (i.e., the class label). In the example of heart diseases, label 1 represents atherosclerotic, and label 0 means native coronary artery.

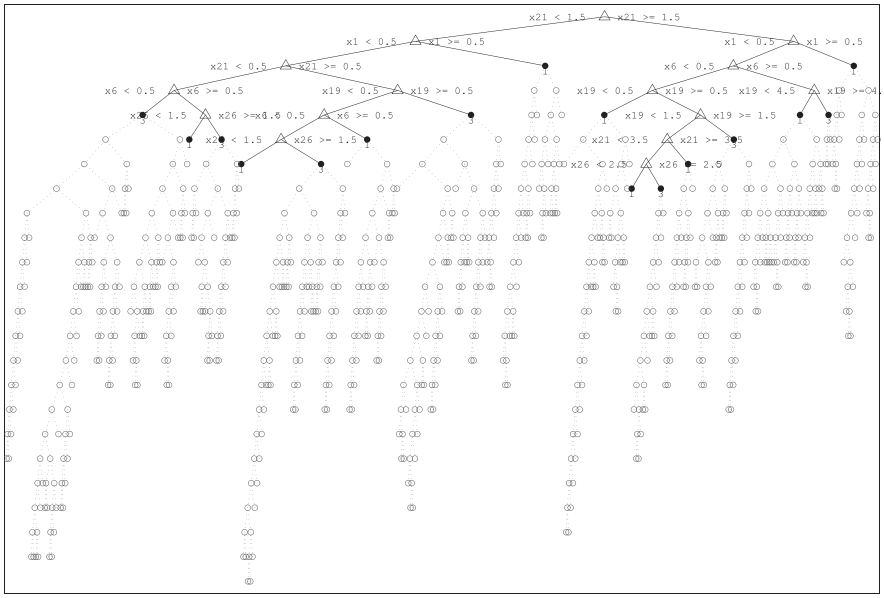


Fig. 3. Classification tree

The statistical information of the dataset is summarized in Table 6.

Table 6. The statistical information of the dataset

| Subset | N. of Class | N. of Feature | N. of Train | N. of Test |
|---|---|---|---|---|
| Bowel | 4 | 13 | 35 | 15 |
| Eye | 3 | 9 | 34 | 16 |
| Heart | 2 | 12 | 34 | 16 |
| Liver | 6 | 12 | 34 | 16 |
| Lung | 4 | 11 | 35 | 15 |

## 4－2. Our Method

The given features represent symptoms, background characteristics and disease situations of patients. The training data was used to learn the rules for the disease diagnosis. We concatenated the features and extracted the rules. In the example of bowel diseases, there are only four patterns of features, which are "1101100000000", "0010001110000", "0110010100000" and "0010000011110". We found out these patterns are corresponded to the class label 1, 2, 3 and 4, respectively as shown in Table 7.

Table 7-11 are the rules obtained from training data. The disease diagnoses, which means determination of the class label for test data, were conducted based on the rules.

Table 7. Rule for bowel disease

| Concatenated Features | Class label |
|---|---|
| 1101100000000 | 1 |
| 0010001110000 | 2 |
| 0110010100000 | 3 |
| 0010000011110 | 4 |

Table 8. Rule for eye disease

| Concatenated Features | Class label |
|---|---|
| 101000001 | 1 |
| 000001002 | 0 |
| 010010001 | 0 |
| 000001004 | 0 |

Table 9. Rule for heart disease

| Concatenated Features | Class label |
|---|---|
| 000000000001 | 0 |
| 000001111111 | 0 |
| 111011111101 | 0 |
| 010101000001 | 1 |

Table 10. Rule for liver disease

| Concatenated Features | Class label |
|---|---|
| 001100000010 | 0 |
| 100100011000 | 0 |
| 010011100000 | 5 |
| 100000000111 | 7 |

Table 11. Rule for lung disease

| Concatenated Features | Class label |
|---|---|
| 01000110000 | 0 |
| 00100000000 | 0 |
| 10010111000 | 0 |
| 11000000111 | 0 |
| 10001100000 | 1 |

## 5．Conclusion

The tasks in this competition were challenging for us. The most challenging one was Task 2. We solved the three tasks by attentively performing the data visualization and statistical analysis, which leads to a good result for the categorization and classification tasks. As a result, our team is the first place winner of CDMC 2015.

### Acknowledgements

The members of our team are Eri Nakahara, Misako Matsumoto, Yuki Matsuura, Akiho Muto, who were the third year students of the Department for the Study of Contemporary Society at Kyoto Women's University, and Prof. Hideo Konami. We thank Ako Okada and Aoi Hisayama for the useful comments and support.

### References

1. New Zealand Herald http://www.nzherald.co.nz
2. The Australian http://www.theaustralian.com.au
3. The Press http://www.stuff.co.nz
4. Yahoo News http://news.yahoo.com
5. BBC http://www.bbc.co.uk
6. The New York Times http://www.nytimes.com
7. The Independent http://www.independent.co.uk
8. The Times http://www.timesonline.co.uk
9. Herald Sun http://www.heraldsun.com.au/
10. Carroll, L. (2015). Alice 's Adventures in Wonderland. 11th ed. [ebook] Gutenberg, Available at: http://www.gutenberg.org/ebooks/11.
11. Trade Me http://www.trademe.co.nz
12. Ripley, B. D.: Pattern Recognition and Neural Networks. Cambridge University Press (1996)