Tag Recommendation System for Data Catalog Site of Japanese Government

Yasuhiro Yamada
Institute of Science and Engineering, Academic Assembly, Shimane University, 1060 Nishikawatsu-cho, Matsue-shi, Shimane, 690-8504, Japan

Keywords: e-Government, Open Government Data, Tag Recommendation, Multi-Label Classification, Machine Learning.

Abstract: This paper proposes a tag recommendation system for a data catalog site of the Japanese government. The site publishes datasets that include files containing statistical data, government documents, and other files of the Japanese government. These datasets also each include the title, description, publication date, and tags, where a tag is a single-word or compound term which represents the content of a dataset. The system uses multi-label classification in machine learning to recommend tags for the datasets; multi-label classification is a method that outputs multiple tags for each input dataset. There are many tags already in datasets hosted on the site that appear infrequently. It is difficult to predict such infrequent tags from the datasets by multi-label classification. To deal with this problem, we use an existing oversampling approach which increases the data of infrequent tags in a training dataset for the learning process of the multi-label classification.

1 INTRODUCTION

Various government organizations have recently published datasets, each of which includes files containing, for example, statistics data and government documents, on their Web sites. Such datasets are referred to as open government data, and such sites are called data catalog sites. For example, the U.S. government publishes datasets on the site "Data.gov[1]." This site had 196,587 datasets in September 2017. The Japanese government publishes datasets on the Web site "DATA.GO.JP[2]." The site was hosting 27,169 datasets as of February 2021.

In addition to files, a dataset contains metadata such as the title of the dataset, description, publication date, and tags. This paper focuses on the tags of a dataset. A tag is a single-word or compound term which represents the content of a dataset. Examples of tags for DATA.GO.JP are "maps," "budgets" and "statistics survey result." A dataset often has multiple tags.

Tags are useful for getting a broad understanding of the content of a dataset before reading the dataset files. If the files of a dataset are large, or if a dataset has many files, then the tags are more important for understanding the dataset. The tags are also useful for users to search for the datasets they want. By inputting or selecting a tag on a search site, a list of datasets associated with the tag can be obtained.

In order for users to gain this benefit, the same tag must be assigned to datasets with the same or similar content. It is difficult to assign proper tags manually to a dataset because government employees or site managers need to first understand the content of the dataset in detail. Also, they need to remember the tags assigned to other datasets in the past.

Data catalog sites use many tags. Tags with general meanings appear in many datasets, whereas tags with detailed meanings are rare in the datasets of a site. The infrequent tags tend to express the concrete content of a dataset, and these tags are especially important for understanding the content of such a dataset. However, it is especially difficult to manually assign such tags to a dataset.

We propose a system which recommends tags for a dataset for DATA.GO.JP. To select appropriate tags for a dataset from among the tags which have been assigned to previously tagged datasets, we use multi-label classification in machine learning. Labels in

[1] https://www.data.gov

[2] This site was updated in March 2023. The new Japanese government data catalog site is called "e-Gov Data Portal" (https://data.e-gov.go.jp/info/ja/top). This paper describes our research for DATA.GO.JP.

Yamada, Y.
Tag Recommendation System for Data Catalog Site of Japanese Government.
DOI: 10.5220/0012260000003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 3: KMIS*, pages 325-331
ISBN: 978-989-758-671-2; ISSN: 2184-3228

multi-label classification correspond to tags. However, multi-label classification has difficulty predicting infrequent labels. We apply an existing oversampling technique which increases datasets with infrequent labels artificially in training data for learning in multi-label classification.

The Japanese government adopted four categories of tags: important data category of G8, category of basic tags of e-government information, category for business classification in e-government action plan, and prioritized fields in "Roadmap for Promotion of Open Data in Electronic Administration." The policy of the Japanese government in assigning tags is to select proper tags from those in these four categories. Additionally, government employees or site managers manually assign tags outside these four categories. In the present paper, we choose classifiers for predicting tags in each category by multi-label classification.

The remainder of the present paper is organized as follows. Section 2 describes related research. Section 3 shows the tags of datasets of open government data on DATA.GO.JP. Section 4 describes a tag recommendation system. Section 5 discusses the problem of the proposed system. Finally, our conclusions are presented in Section 6.

2 RELATED WORK

2.1 Multi-Label Classification for Imbalanced Data

We define multi-label classification. Let $L = \{l_1, l_2, \ldots, l_m\}$ be a set of labels which correspond to tags of datasets of open government data, and $D = \{(\boldsymbol{x}_1, Y_1), (\boldsymbol{x}_2, Y_2), \ldots (\boldsymbol{x}_n, Y_n)\}$ be a set of training examples, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $Y_i \subseteq L$. The learning task of multi-label classification is to make a classifier for L from D. Then, given an unlabeled example $\boldsymbol{x} \in \mathbb{R}^d$, the classifier predicts labels for $\boldsymbol{x}$.

Multi-label classification algorithms have been proposed for imbalanced data (Xu et al., 2016; Babbar and Schölkopf, 2017; Jain et al., 2016; Liu and Tsoumakas, 2019; Wu et al., 2020; Schultheis and Babbar, 2022; Yu et al., 2019). Imbalanced data include both frequent and infrequent labels in training data. These algorithms address the difficulty of predicting infrequent labels.

2.2 Resampling Training Data in Multi-Label Classification

Various methods for resampling training data for imbalanced data have been proposed. The oversampling method increases data of infrequent labels in training data by making feature vectors artificially (Chawla et al., 2002; Charte et al., 2015; Liu et al., 2022). This paper utilizes SMOTE (Chawla et al., 2002) for the oversampling. A brief explanation of SMOTE is given in Section 4. On the other hand, the undersampling method decreases data with frequent labels (Haixiang et al., 2017; Rao and Reddy, 2020).

2.3 Our Previous Work

We applied three multi-label classification methods for the tag recommendation (Yamada and Nakatoh, 2018) to 196,587 datasets on Data.gov, which is the data catalog site of the U.S. government. Specifically, we compared the following methods: support vector machine, random forest, and multinomial naive Bayes. The random forest method has obtained good results (Yamada and Nakatoh, 2018). However, our previous work did not deal with the prediction of infrequent tags.

3 TAGS OF JAPANESE OPEN GOVERNMENT DATA

This section describes tags used by the data catalog site "DATA.GO.JP" of the Japanese government. We collected 27,169 datasets from DATA.GO.JP in February 2021.

The tag assignment policy of this site is to select tags from those in each of the following four categories:

Category 1. important data category of G8 (16 tags),

Category 2. category of basic tags of e-government information (16 tags),

Category 3. category for business classification in e-government action plan (31 tags),

Category 4. prioritized fields in "Roadmap for Promotion of Open Data in Electronic Administration." (7 tags).

Tables 1 to 4 in appendix A show the tags in these four categories and their frequencies in the datasets. We see that the frequency of some tags is 0. In Categories 1 to 4, Japanese and English tags with the same meaning are assigned to the same dataset.

Also, the site managers assign appropriate tags for each dataset outside the four categories. **Category 5** comprises such tags. Table 5 in appendix A shows the top 20 most frequent tags with their frequencies. Figure 1 shows the frequencies of all the tags in Category 5. The vertical axis is in log scale. We can see that a small number of tags appear frequently in the datasets. We can also see that most of the tags appear once or twice in the datasets. Tags with general meanings tend to be assigned many times. On the other hand, tags with concrete meanings tend to appear rarely in the datasets.

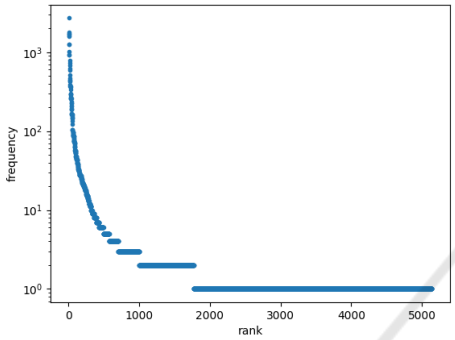

Figure 1: Frequencies of tags.

4 TAG RECOMMENDATION SYSTEM FOR DATA.GO.JP

This section describes a tag recommendation system for DATA.GO.JP (see Figure 2). This system utilizes multi-label classification which outputs multiple tags from a dataset.

The system learns a classifier for each of the four categories. Multi-label classification cannot predict the tags whose frequency in the training data is zero. Therefore, the system cannot predict some tags in Categories 1 to 4.

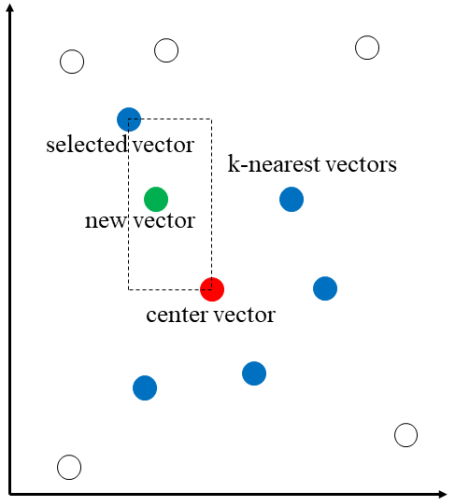

Figure 3: SMOTE.

Category 5 has many infrequent tags (see Figure 1). We use an oversampling approach to make a classifier for Category 5. The oversampling increases feature and label vectors of infrequent labels in the training data. We utilize SMOTE (Chawla et al., 2002) for the oversampling, whose process is illustrated in Figure 3. Circles in the figure express feature vectors. Given training data $D = \{(\boldsymbol{x}_1, Y_1), (\boldsymbol{x}_2, Y_2), \ldots (\boldsymbol{x}_n, Y_n)\}$, SMOTE executes oversampling as the following steps:

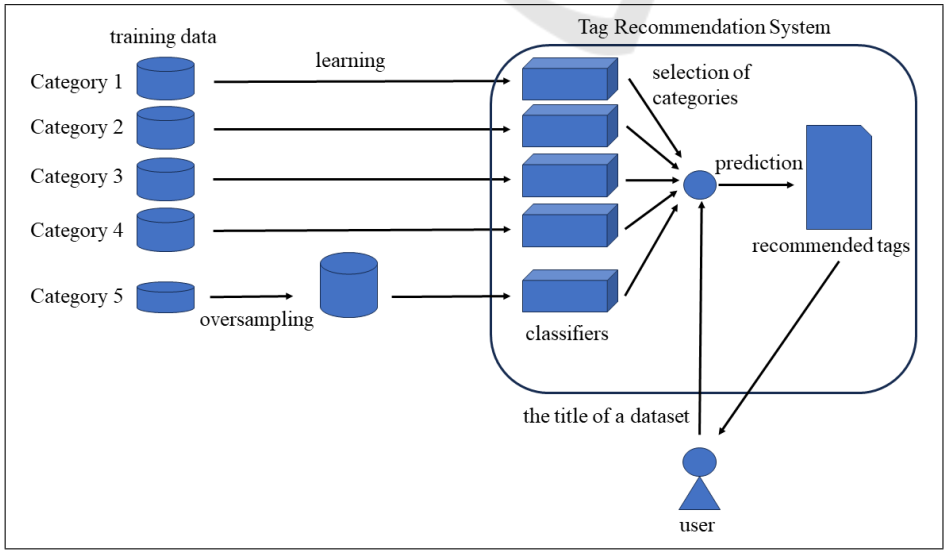

Figure 2: Tag recommendation system for DATA.GO.JP.

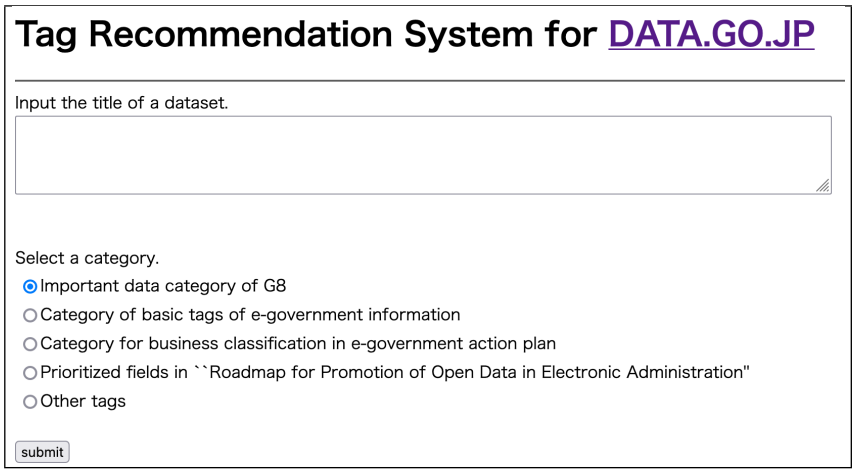

Figure 4: Input of tag recommendation system.

Recommended tags for "気象予報_天気予報・台風の資料"

Tag	Tag (English)	Value(0-1)
地球観測	earth_observation	1.0000
地図	maps	0.9948
社会的流動性と福祉	social_mobility and welfare	0.0527
犯罪と司法	crime and justice	0.0170
企業	companies	0.0124
統計	statistics	0.0124
政府の説明責任と民主主義	government_accountability and democracy	0.0122
財政と契約	finance and contracts	0.0069
選挙結果	elections	0.0039
交通とインフラ	transport and infrastructure	0.0015
科学と研究	science and research	0.0011
国際開発	global_development	0.0007

Figure 5: Output of tag recommendation system.

1. Select a feature vector $\boldsymbol{x}_i$ randomly from the feature vectors with an infrequent label l as a center vector $\boldsymbol{c}$ (red circle).
2. Calculate the k-nearest feature vectors (blue circles) in D for $\boldsymbol{c}$ by Euclidean distance.
3. Select a feature vector $\boldsymbol{x}_{near}$ (top left blue circle) from the k vectors randomly. Make a new feature vector (green circle) between $\boldsymbol{c}$ and $\boldsymbol{x}_{near}$.
4. Assigns the new feature vector to label l.

SMOTE repeats the process until the number of feature vectors of the infrequent label achieves the threshold.

The number of collected datasets is 27,169. The feature vector of each dataset was generated by nouns in the title and description of the dataset and their frequencies in the dataset. The feature and label vectors of infrequent tags in Category 5 were increased by SMOTE until the number of the vectors of each infrequent tag that appears fewer than 10 times in training data reached 10. The number of all datasets in the training data after the oversampling was 67,416.

We use the multinomial naive Bayes method (Manning et al., 2008) for multi-label classification. We implemented this method using scikit-learn[3] (Pedregosa et al., 2011).

Users input the title of a dataset and select a category of tags. Figure 4 shows a screen capture of the input of the system.

Figure 5 shows a screen capture of the output of the system for the input title “the documents of weather prediction, weather forecast and typhoon.” We use the function predict_proba() in scikit-learn which outputs probability estimates for tags. The col-

[3]http://scikit-learn.org/stable/

umn "Value" in Figure 5 is the estimates. Although automatically assigning tags to a dataset would be ideal, due to the accuracy problem of recommendations, it is assumed that the users select the proper tags from the recommended tags.

5 DISCUSSION

The frequency of some tags in Categories 1 to 4 is 0. Multi-label classification cannot predict tags which do not appear in the training data. Also, oversampling methods cannot increase feature and label vectors of such tags. One method that could be adapted to predict such tags is zero-shot learning.

It remains to check the accuracy of the tag recommendations. Various multi-label classification and oversampling methods have been proposed, so we need to conduct comparisons with such methods.

In the present study, we set the threshold of the frequency between infrequent tags and others to 10 in the datasets of DATA.GO.JP. It is important to automatically determine the threshold at which the oversampling is effective.

Since multi-label classification cannot output tags which do not appear in the training data, we need to extract new tags from a dataset. We previously proposed a method to extract particular noun phrases (Yamada et al., 2018; Yamada and Nakatoh, 2018) as characteristic phrases and words to use as new tags that represent the dataset.

6 CONCLUSION

This paper proposed a tag recommendation system for DATA.GO.JP which is the data catalog site of the Japanese government. The system utilizes multi-label classification in machine learning to recommend tags. SMOTE, which is an oversampling method, is used to increase the data of infrequent tags in the training data. The system is given the title of a dataset, and it recommends several tags using a classifier constructed in advance.

The work reported herein is part of ongoing research. Important remaining work includes increasing the accuracy of predicting infrequent tags and making new tags which are not included in previous datasets. Other data catalog sites of countries include Data.gov of the U.S. government. Our future work is also to develop a system for such sites.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP19K12715.

REFERENCES

Babbar, R. and Schölkopf, B. (2017). DiSMEC: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pages 721–729. ACM.

Charte, F., Rivera, A. J., del Jesus, M. J., and Herrera, F. (2015). MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 73:220–239.

Jain, H., Prabhu, Y., and Varma, M. (2016). Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 935–944. ACM.

Liu, B., Blekas, K., and Tsoumakas, G. (2022). Multi-label sampling based on local label imbalance. *Pattern Recognition*, 122:108294.

Liu, B. and Tsoumakas, G. (2019). Synthetic oversampling of multi-label data based on local label distribution. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*, page 180–193. Springer-Verlag.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rao, K. N. and Reddy, C. S. (2020). A novel under sampling strategy for efficient software defect analysis of skewed distributed data. *Evolving Systems*, 11:119–131.

Schultheis, E. and Babbar, R. (2022). Speeding-up one-versus-all training for extreme classification via mean-separating initialization. *Mach. Learn.*, 111(11):3953–3976.

Wu, T., Huang, Q., Liu, Z., Wang, Y., and Lin, D. (2020). Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision –*

ECCV 2020, pages 162–178. Springer International Publishing.

Xu, C., Tao, D., and Xu, C. (2016). Robust extreme multi-label learning. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284. ACM.

Yamada, Y., Himeno, Y., and Nakatoh, T. (2018). Weighting of noun phrases based on local frequency of nouns. In *Recent Advances on Soft Computing and Data Mining - Proceedings of the 3rd International Conference on Soft Computing and Data Mining*, pages 436–445. Springer.

Yamada, Y. and Nakatoh, T. (2018). Tag recommendation for open government data by multi-label classification and particular noun phrase extraction. In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018), Vol. 3*, pages 83–91.

Yu, H.-F., Zhong, K., Dhillon, I. S., Wang, W.-C., and Yang, Y. (2019). X-bert: extreme multi-label text classification using bidirectional encoder representations from transformers. In *NeurIPS 2019 Workshop on Science Meets Engineering of Deep Learning*.

APPENDIX A

This appendix shows the frequencies of tags in each of five categories. Note that tags in the following tables are originally written in Japanese.

Table 1: Category 1: Important data category of G8.

Tag	#
statistics	9,711
budgets	2,163
education	1,574
energy and environment	1,075
maps	947
health	909
finance and contracts	808
crime and justice	492
science and research	438
government accountability and democracy	434
companies	304
social mobility and welfare	176
global development	122
earth observation	47
elections	40
transport and infrastructure	10

Table 2: Category 2: Category of basic tags of e-government information.

Tag	#
statistics survey result	6,848
budgets and account settlement	2,023
white paper and annual report	1,423
information disclosure	994
evaluation and result	556
procurement	551
organization and institution	163
press release	98
council and working group	65
laws and regulations and notices	56
application and notification procedure	25
juridical competent authority	3
public comment	1
bill submitted to the diet	0
prior confirmation procedures and rules	0
press conference of minister	0

Table 3: Category 3: Category for business classification in e-government action plan.

Tag	#
environment	2,603
population and households	2,553
government except elsewhere classified	1,919
security	1,892
disaster	1,760
international	1,654
agriculture and forestry	1,621
land	1,033
business and household and economy	849
construction	614
climate	597
labor	523
justice	364
education and learning support	359
scientific research and professional and technical services	306
fisheries	252
medical and health care and welfare	167
sightseeing	167
manufacturing	161
mining and quarrying of stone	129
finance and insurance	126
accommodations and eating and drinking services	125
transport and postal services	117
information and communications	41
wholesale and retail trade	37
services not elsewhere classified	34
living-related and personal services and amusement services	33
industries unable to classify	28
compound services	3
electricity and gas and heat supply and water	3
real estate and goods rental and leasing	0

Table 4: Category 4: Prioritized fields in “Roadmap for Promotion of Open Data in Electronic Administration”.

Tag	#
statistics	9,711
budgets and final accounts and procurement	4,816
white paper and annual report	2,490
geospatial	1,499
disaster prevention and mitigation	1,324
transportation and tourism	396
code	103

Table 5: Top 20 most frequent tags in Category 5 (other tags).

Tag	#
finance	2,705
economy	1,772
2000012010019	1,711
5000012080001	1,701
energy	1,613
industry	1,572
expenditure	1,271
supply	1,244
contract	1,020
study	936
traffic	923
society	779
manufacture	684
manufacturing industry	616
law	588
observation	518
medical	471
strong economy that generates hope	443
ship	437
crime	426
gas	389