Ensemble Method for Classification in Imbalanced Patent Data

Eleni Kamateri[1] and Michail Salampasis[1]

[1]*Department of Information and Electronic Engineering, International Hellenic University (IHU), Alexander Campus, Sindos 57400, Thessaloniki, Greece*

Abstract

This study presents an ensemble method for patent classification addressing the imbalance patent data problem. To achieve this, the dataset is divided into two data partitions based on the codes' representation magnitude. These partitions are trained separately by two identical classifiers and their results are combined using a stacking meta-classifier. Experiments are conducted using two benchmark patent datasets. The first results showed that the proposed combination of classifiers improves the imbalance patent data problem and outperforms the baseline classifiers, other combinations of classifiers and recent state-of-the-art techniques for patent classification.

Keywords

Patent, Classification, Ensemble, Imbalance data, Single-label, Sub-classes, Ensemble method, Deep learning, Word embeddings1

1. Introduction

Patent classification is an important task of the patent examination process dealing with the assignment of one or more classification codes from a classification scheme. The most widely used classification scheme is the International Patent Classification (IPC) which contains approximately 70,000 different IPC codes. The correct assignment of classification codes is quite important as it ensures that patents with similar technical characteristics will be clustered together under the same classification codes, something which is crucially important for many subsequent tasks, such as patent management and search, technology characterization and landscape [1, 2]. However, the high numbers of classification codes, along with their complex and heterogeneous definitions make the patent classification a challenging task.

The manual patent classification, which is performed by patent officers when a patent application arrives, includes the finding of relevant classification codes through the hierarchical descriptions of classification codes in the classification scheme. However, it can be very time consuming, tedious and strongly dependent on patent officer's ability and experience [3]. This is the reason why automatic tools for selecting the relevant classification codes are needed.

Research efforts in automated patent classification [4-7] utilize Natural Language Processing (NLP) techniques and Machine Learning (ML)/Deep Learning (DL) models for effective patent modelling and representation, and automatic classification. Most of these patent classification efforts used various simplifications when applied, e.g., working mostly with well-represented codes having many training samples or targeting the higher levels of the classification hierarchy, still they do not attain acceptable performance, i.e., one close to human performance.

The accuracy of the classification model mainly depends on the quality of the dataset and the classification algorithm. The data-related factors which could reduce the accuracy of a patent classification model are many, such as the complex/broad concepts expressed by classification codes, the ambiguous vocabulary or new terminology used, the overlapping concepts among classification codes (which increases as we go down in the level hierarchy), and, last but not least, the imbalanced patent dataset problem. This means that some classification codes have a large number of patent samples and thus high representation magnitude in the dataset. These codes are called major codes/classes. On the other side, there are some other classification codes which have very few patent samples and thus low representation magnitude These codes are called minor codes/classes.

Classification models trained by imbalanced datasets usually have a very poor prediction ability on minor codes. In order to solve the imbalanced dataset issue, lots of research efforts have been carried out. Improvements are mainly based on two directions, the

PatentSemTech'23: 4th Workshop on Patent Text Mining and Semantic Technologies, *July 27th, 2023, Taipei, Taiwan.*

ekamater@hotmail.com (E. Kamateri); msa@ihu.gr (M. Salampasis); 0000-0003-0490-2194 (E. Kamateri); 0000-0003-4087-125X (M. Salampasis);

dataset level and the algorithm level [8]. On the dataset level, the main strategy is to use resampling methods. Over-sampling and under-sampling methods have been introduced to resample the data to get a balanced dataset [9-11]. On the algorithm level, the main idea is to adjust the algorithms to improve the accuracy of models, such as introducing an ensemble method [12, 13].

In this study, we adjust the ensemble architecture for patent classification presented in [14] to address the imbalance patent data problem. More specifically, we divide the dataset into two partitions using the codes' representation magnitude, i.e., a partition with the major codes and a partition with the minor codes, and train two classifiers of the same type with patents from each partition separately. Then, we combine the outcomes of the two classifiers using a meta-classifier. The experiments showed that the proposed combination of classifiers improves the imbalance patent data problem and outperforms the baseline classifiers, the previous combinations of classifiers (presented in [14]) and recent state-of-the-art techniques for patent classification.

2. Motivation

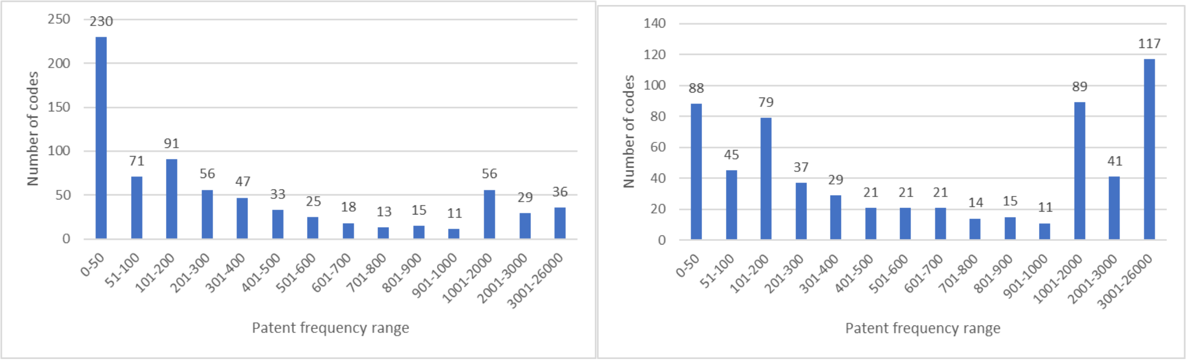
The classification scheme contains numerous codes, of which a varying number is assigned to each patent [15, 16]. The distribution of patents across classification codes is quite unbalanced following a Pareto-like distribution [16]. About 80% of all patent documents are classified in about 20% of the classification codes, meaning that some classification codes present quite low and other quite high patent frequency.

Similar to the real-life distribution of patents across codes, the distribution of patents across codes in test collections is quite unbalanced. For example, in the CLEFIP-0.54M dataset[2] which originates from the CLEF-IP 2011 (see Section 4 for more information), each code has a mean frequency of 740 patents with a standard deviation of 1,930 patents and a median frequency of 169, which is a more informative statistic compared to mean for imbalance datasets where there exist many frequency outliers. Similarly, in the USPTO dataset, each code has a mean frequency of 3,177 patents with a standard deviation of 12,710 patents and a median frequency of 578. Moreover, 392 codes (53.63% of all 731 codes) in the CLEFIP-0.54M dataset and 212 codes (33.76% of all 628 codes) in the USPTO dataset have a low patent frequency between 1 and 200 patents (Figure 1a and 1b).

Trying to explore whether the code's patent frequency affects the performance of the patent classification models, Figure 2a displays the accuracy of a state-of-the-art DL model, the Bi-LSTM [17], when applied to a range of patent frequencies in the subclass category of the IPC 5+ level hierarchy using the 60 first words of the abstract section from the CLEFIP-0.54M dataset.

As it is observed, high accuracies can be attained as the patent frequency of codes increases, meaning that the number of patent samples representing a specific classification code plays a significant role in the code's distinguishability and finally in the code's performance. Considering that the accuracy of the classification model across all codes is 63.76%, we assume that an adequate accuracy (see the "threshold" line in red – Figure 2a) is achieved for codes represented by more than 500 patent samples.

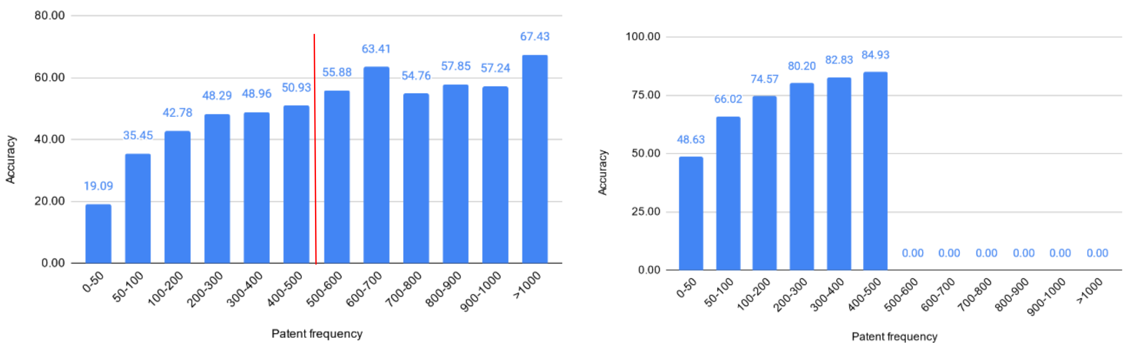
Especially, for classification codes with low representation magnitude the accuracy achieved is quite low affecting significantly the total accuracy of the classifier. e.g., the accuracy for codes with patent frequency between 0 and 50 patents is only 19.09%. Therefore, the idea behind this study is that if we had a classifier focusing only on these low-represented codes, better performance would be achieved. This is also validated in Figure 2b where the accuracy achieved by a similar classifier trained only with low-represented codes is presented.



1a **1b**

Figure 1a, b: The unbalanced distribution of patent frequency across the 731 and 628 main classification codes of the CLEFIP-0.54M and USPTO dataset, respectively.

[2] https://drive.google.com/drive/folders/1tfBsUkQwIpwwgDyw28EOZctaiiJqZr1Q



2a

2b

Figure 2a: The accuracy of a state-the-art patent classification model, the Bi-LSTM, as a function of codes' patent frequency organized into groups of subsequent codes. **Figure 2b:** The accuracy of the same model trained only on patents with low-represented codes.

3. Ensemble method for classification of imbalance patent data

An ensemble architecture for automated patent classification has been introduced by Kamateri and Salampasis in [14]. The architecture consists of individual classifiers that can be of any number and any type, while they can be trained with the same or different parts of the patent document. Each classifier produces a list of probabilities for all labels based on its whole or partial knowledge about the patent. Then, the probabilities for a specific label derived from all individual classifiers are combined and a total probability is calculated for this label. The label with the maximum probability consists the predicted label for the patent. The combination of probabilities of the individual classifiers can be aggregated using simple/weighted averaging, voting, stacking or other combination techniques.

In this study, we apply this ensemble architecture for automated patent classification to address the imbalance patent data issue equipped it with two baseline classifiers and a meta-classifier (Figure 3). The first classifier is trained with high represented classification codes, while the second classifier is trained with low represented codes. Thus, each classifier specializes in a portion of codes having high and low patent frequency, respectively. This means that if a patent application characterized with a classification code of low frequency is submitted to the first classifier specializing to high-represented codes, the classifier will not be able to classify this patent application correctly since the specific classifier is not (probably) trained with similar patents. Conversely, if this patent application belonging to a classification code of low frequency is submitted to the second classifier, which is more delicate to detect codes with low patent frequency, there are better chances to be properly classified under the correct classification code corresponding to the described invention. In such cases, an appropriate combination of two baseline classifiers can better approximate such a boundary by dividing the data space into smaller and easier-to-learn partitions. Then, a meta-classifier is trained on the features that are outputs of the baseline classifiers to learn how to best combine their predictions (stacking). More specifically, the meta-classifier will distinguish if the described invention of a patent application belongs to a high or a low represented classification code and, respectively, coordinate the operation (sigmoid stacking classifier) or selecting the more appropriate (softmax stacking classifier) of the two baseline classifiers to classify a receiving patent application.

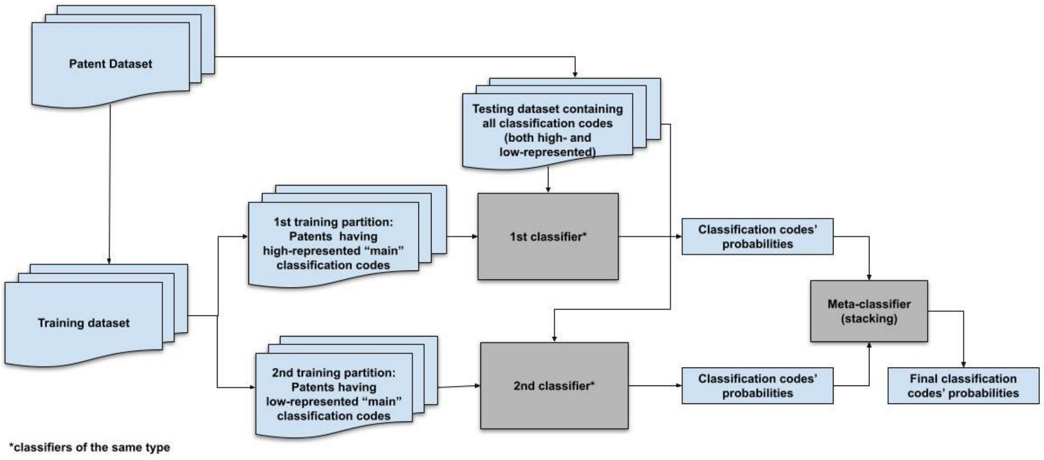

Patent Dataset
Testing dataset containing all classification codes (both high- and low-represented)
1st training partition: Patents having high-represented "main" classification codes
1st classifier*
Classification codes' probabilities
Training dataset
Meta-classifier (stacking)
2nd training partition: Patents having low-represented "main" classification codes
2nd classifier*
Classification codes' probabilities
Final classification codes' probabilities
*classifiers of the same type

Figure 3: Ensemble architecture for automated patent classification focusing on the imbalance patent data.

4. Data collection

To evaluate the real-world performance of the proposed ensemble method for imbalance patent data, two patent benchmark datasets have been used: the USPTO-2M and the CLEFIP-0.54M.

4.1. USPTO-2M 1.1

The USPTO-2M is a large-scale dataset prepared for patent classification [6]. The raw patent data have been obtained from the online website of the United States Patent and Trademark Office (USPTO) from 2006 to 2015. The dataset contains 2,000,147 patents with the title and abstract sections. in 637 categories at the subclass level.

4.2. CLEFIP-0.54M

The CLEFIP-0.54M contains English patents of CLEF-IP 2011 with the main classification code and all the following six patent sections: Title, Abstract, Description, Claims, Applicants and Inventors. In total, the dataset contains 541,131 patents classified in 731 subclass codes of which 276,794 come from the European Patent Office (EPO) and 264,337 from the World Intellectual Property Organization (WIPO)[3] .

5. Experimental setup

The ensemble architecture presented in [14] is instantiated in this study as a single-label classification task at the subclass (3rd) level category of the IPC 5+ level hierarchy. More specifically, the aim is to identify the main classification code. In the CLEFIP-0.54M dataset, this information is available by the dataset. In the USPTO dataset, we assume that the first code is the main classification code in cases where many codes are given to a patent.

An ensemble of bidirectional LSTM classifiers was employed, since this ML method has been proved in [14] to attain better results than other DL methods. Each classifier was trained on codes of different patent frequency: low-represented codes with patent frequency between 0-500 patents and high-represented codes with patent frequency over 500 patents, respectively. The outcome probabilities of individual classifiers were used as input for a meta-classifier using the stacking technique. The meta-classifier is a neural network having two dense layers. The second dense layer is activated with a softmax or a sigmoid activation in order to obtain a probability distribution over all targeted labels/codes.

With respect to the patent representation, the first 60 words from the patent part of interest (e.g., title, abstract, etc.) were used after undertaking a sequence of preprocessing steps (cleaning punctuation, symbols and numbers, and stop word removal). The feature words were then mapped to embeddings using a domain-specific pre-trained language model which has been created on a patent dataset, proposed by Risch and Krestel [4].

The dataset was split into training, validation and testing sets (80:10:10). Batch size was set to 128, epochs for baseline classifiers to 15 and epochs for meta-classifier to 20.

6. Results

In each experiment, two baseline classifiers have been trained on two different data partitions. The first classifier was trained on patents belonging to high-represented codes, having patent frequency over 500 patents, while the second classifier was trained on patents of low-represented codes, with patent frequency between 1 and 500 patents. Table 1 presents the Accuracy attained by each classifier i) when it is tested on the same data partition where it was previously trained, named as "Testing on the same data partition", and ii) when it is tested in the entire dataset, containing both data partitions with known and unknown data, named as "Testing on the entire dataset". It also presents the Accuracy of the meta-classifier combining the outcomes of the two baseline classifiers using a stacking technique. Last, it presents the Accuracy of the ensemble of classifiers combining sigmoid predictions from different patent sections.

In both datasets, the accuracy is much improved when a stacking technique is applied combining the predicted probabilities acquired by individual classifiers specialized in high- and low-represented codes, respectively. Moreover, the stacking technique

Table 1
Accuracy at subclass level

	Section	Classifier 1 - Training on high-represented codes		Classifier 2 - Training on low-represented codes		Meta-classifier combining classifier 1 & 2		Ensemble of sigmoid predictions for all patent sections		Baseline classifier trained on the entire dataset [14]	Ensemble of predictions for all patent sections (Weighted average) [14]
		Testing on the same data partition	Testing on the entire dataset	Testing on the same data partition	Testing on the entire dataset	Softmax	Sigmoid	Average	Weighted average		
USPTO	Title	55.34%/54.28%		65.43%/ 1.50%		54.65%	55.39%	61.98%	62.11%	53.44%	59.92
	Abstract	59.85%/59.86%		71.79%/1.65%		59.86%	60.64%			58.61%	
CLEFIP-0.54M	Abstract	68.02%/63.91%		65.72%/9.37%		67.69%	68.14%	75.36%	75.40%	63.76%	70.39%
	Description	70.59%/66.43%		71.23%/10.16%		69.47%	71.10%			66.46%	
	Claims	68.64%/64.59%		64.42%/9.52%		68.23%	68.88%			64.56%	

[3] CLEFIP-0.54M 2022 (accessed 18/12/2022), https://github.com/ekamater/CLEFIP2011_XML2MySQL

using the sigmoid activation seems to slightly outperformed the stacking classifier using the softmax activation. It is also clear that the proposed method provides better results than those obtained from recent state-of-the-art techniques [14, 18, 19].

7. Conclusions

In this study, a novel ensemble method for patent classification is presented addressing the imbalance patent data problem which is one of the most significant factors that reduces the accuracy in automated patent classification. The results showed that the proper combination of classifiers can attain significantly improved accuracy compared to baseline classifiers and existing classification techniques. Moreover, the combination of the knowledge gained from multiple classifiers could address the problem of low patent sample representation for codes, a phenomenon that is relatively common in the patent domain as the IPC/CPC taxonomy evolves with new codes introduced, codes partitioned into sub-categories, etc.

Acknowledgements

H.F.R.I. Hellenic Foundation for Research & Innovation

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 10695).

References

[1] M. Salampasis, G. Paltoglou, A. Giahanou, Report on the CLEF-IP 2012 Experiments: Search of Topically Organized Patents. Conference and Labs of the Evaluation Forum, 2012.

[2] E. Perez-Molina, F. Loizides, Novel data structure and visualization tool for studying technology evolution based on patent information: The DTFootprint and the TechSpectrogram. World Patent Information 64 (2021) 102009. doi: https://doi.org/10.1016/j.wpi.2020.102009.

[3] T. Montecchi, D. Russo, Y. Liu, Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search. Advanced Engineering Informatics 27(3) (2013) 335-345. doi: https://doi.org/10.1016/j.aei.2013.02.002.

[4] M. F. Grawe, C. A. Martins, A. G. Bonfante, Automated patent classification using word embedding. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 408-411. doi: https://doi.org/10.1109/ICMLA.2017.0-127.

[5] L., Xiao, G., Wang, & Y. Zuo, Research on patent text classification based on word2vec and LSTM. In 2018 11th International Symposium on Computational Intelligence and Design (ISCID), 2018, pp. 71-74. doi: https://doi.org/10.1109/ISCID.2018.00023.

[6] S. Li, J. Hu, Y. Cui, J. Hu, DeepPatent: patent classification with convolutional neural networks and word embedding. Scientometrics, 117(2) (2018) 721-744. doi: https://doi.org/10.1007/s11192-018-2905-5.

[7] J. Risch, R. Krestel, Domain-specific word embeddings for patent classification. Data Technologies and Applications 53 (2019) 108-122. doi: https://doi.org/10.1108/DTA-01-2019-0002.

[8] H. Feng, W. Qin, H. Wang, Y. Li, G. Hu, A combination of resampling and ensemble method for text classification on imbalanced data. In: Wei, J., Zhang, LJ. (eds), Big Data – BigData 2021. BigData 2021. Lecture Notes in Computer Science, volume 12988. Springer, Cham. doi: https://doi.org/10.1007/978-3-030-96282-1_1.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (2002) 321-357, doi: https://doi.org/10.1613/jair.953.

[10] G. E. Batista, A. L. Bazzan, M. C. Monard, Balancing training data for automated annotation of keywords: a case study. In WOB, 2003, pp. 10-18.

[11] B. Krawczyk, M. Koziarski, M. Woźniak, Radial-based oversampling for multiclass imbalanced data classification. IEEE transactions on neural networks and learning systems, 31(8) (2019) 2818-2831. doi: 10.1109/TNNLS.2019.2913673.

[12] Y. Zhao, A.K. Shrivastava, K.L. Tsui, Imbalanced classification by learning hidden data structure. IIE Transactions. 48 (7) (2016) 614–628. doi: https://doi.org/10.1080/0740817X.2015.1110269.

[13] C. Cao, Z. Wang, IMCStacking: cost-sensitive stacking learning with feature inverse mapping for imbalanced problems. Knowledge-Based Systems. 150 (2018) 27–37. doi: https://doi.org/10.1016/j.knosys.2018.02.031.

[14] E., Kamateri, M. Salampasis, 2022. An Ensemble Architecture of Classifiers for Patent Classification. In proceedings of the 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech), 2022, pp. 6-7. doi: https://doi.org/10.34726/3550.

[15] M. R. Gouvea Meireles, G. Ferraro, S. Geva, Classification and information management for patent collections: a literature review and some research questions. Information Research, 21, 1, (2016) 7051-29.

[16] K. Benzineb, J. Guyot, Automated patent classification. In M. Lupu, K. Mayer, J.Tait & A. J. Trippe (Eds.), Current challenges in patent information retrieval, Springer, London, 2011, pp. 239-262. doi: https://doi.org/10.1007/978-3-642-19231-9_12.

[17] E., Kamateri, V., Stamatis, K., Diamantaras, & M. Salampasis,. Automated Single-Label Patent Classification using Ensemble Classifiers. In 2022 14th International Conference on Machine Learning and Computing (ICMLC), 2022, pp. 324–330. doi: https://doi.org/10.1145/3529836.3529849.

[18] M. Sofean, Deep learning based pipeline with multichannel inputs for patent classification. World Patent Information 66 (2021) 102060. doi: https://doi.org/10.1016/J.WPI.2021.102060.

[19] D. Tikk, G. Biró, A. Törcsvári, A hierarchical online classifier for patent categorization. In Emerging technologies of text mining: Techniques and applications, 2008, pp. 244-267. doi: https://doi.org/10.4018/978-1-59904-373-9.CH012.