International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 1 (2018) pp. 539-544

Term Expansion and Powerlabel Set for Multi-Label Hierarchical on Short Document Classification

Zaid Farooq Salih and Sabrina Tiun

Faculty of Information Science and Technology,
National University of Malaysia, Bangi, Malaysia.

Abstract

The task of hierarchical classification is getting more challenging when handling short text. Short text documents contain a limited number of words which make it highly ambiguous regarding the difficulty of extracting contextual information. Several approaches have been proposed for the task of hierarchical text classification. However, such approaches have used the One-against-all mechanism which seems to be insufficient for the short text classification. Therefore, this paper aims to propose a combination of expansion method and Powerset-label mechanism for the short hierarchical classification using the Support Vector Machine (SVM) classifier. The expansion aims to handle the problem of 'something-like' that lies behind the short text by providing semantic correspondences using WordNet dictionary. On the other hand, the Powerset-label mechanism will be utilized with the SVM classifier in order to handle the problem of hierarchical text classification. To test the proposed method, a short text dataset of ACM has been used in the experiments which contain a vast amount of titles and keywords related to publication articles. Experimental results have showed that the expansion method has improved the hierarchical classification achieving an f-measure of 88.6%.

Keywords: Hierarchical Classification, Multi-label Classification, Text Expansion, Support Vector Machine

INTRODUCTION

The last decade has witnessed an expansion of the textual data over the internet. Such exponential growth has motivated many researchers to analyse such this data in order to identify meaningful patterns or extracting valuable information (Simoes et al. 2009). One of the main analysis tasks is the classification using the supervised machine learning techniques. These techniques aim at allocating an appropriate class label for each text document (Zhu & Goldberg 2009). The class label can be described as a set of predefined categories such as 'right' and 'wrong', or 'low', 'medium' and 'high'. The key issue behind such task lies on the historical or examples data that have been given the exact class label, this data is being utilized for training purposes (Kotsiantis et al. 2007). In this manner, the classification techniques will be adapted to training on such historical data and identify correlations among the attributes.

Recently, several issues have arisen in the field of text classification. One of these issues is dealing with short text documents. Short text documents contain a limited number of words which make it highly ambiguous regarding the difficulty of extracting contextual information (Lu et al. 2015). Basically, dealing with short text is considered to be a challenging task in terms of the performance of the classification (Song et al. 2014).

Another issue is the hierarchical or multi-label text classification. Unlike the traditional classification task, the multi-label text classification aims to identify one or more classes for a particular document (Sucar et al. 2014). In this case, the prediction will not be depending on certain classes, but rather it would provide a probability for each document (Santos & Rodrigues 2009). The single document may classify as one or more classes.

This paper aims to propose a combination of expansion method and Powerset-label mechanism for the short hierarchical classification using the Support Vector Machine (SVM) classifier.

Basically, the expansion of text has a significant impact on the hierarchical text classification in which the short text will be properly classified into their actual class labels. For example, assume a short text consists of a single word which is "traceability", this word would belong to several class labels such as financial tracking, historical tracking, program tracking. In order to determine which class label is suitable for such text, it is necessary to expand this word with more semantic correspondences. Let the expansion occur on such word as "traceability: verifying the functionality of application" in this manner, the text would accurately be classified as software or computer science class label.

RELATED WORK

Several research studies have addressed the problem of classifying short text documents. Some of these researchers have utilized semantic approaches, other have utilized statistical approaches, while the rest have combined both

techniques. For instance, Wang et al. (2014) have addressed the drawback behind the bag-of-word (BOW) approach in terms of classifying short text. BOW aims to handle the words contained in the short text separately in order to provide a clue for the category of such document. However, the problem of BOW when dealing with short text classification lies on the too limited number of words which make the classifier is unable to extract the contextual information. For this manner, the authors have proposed an approach called 'Bag-of-Concept' in which a knowledge source has been used to provide semantic correspondences such as synonyms and hyponyms. This can significantly enhance the process of extracting contextual information.

Moreover, Li & Qu (2013) have proposed an overcome solution for the problem of short text representation by using a feature extraction approach called Interested Term Count (ITC) in order to extract meaningful patterns from the short text. Such approach has significantly enhanced the process of enriching the short text which leads to better classification accuracy.

Kiritchenko et al. (2014) have proposed a user generated or so-called domain specific knowledge source for overcoming the problem of the sparsity of short text classification. The authors have concentrated on tweets from Twitter which likely contain a maximum of 140 characters. In this vein, the words inside the tweet have been analysed separately by processing it as a hashtag. Using the search engine of Twitter, the retrieval of these hashtags will significantly enhance the extraction of the contextual information of the tweets.

Yin et al. (2015a) and (2015b) have proposed a semi-supervised approach for short text classification. In fact, the supervised learning techniques are mainly depending on a predefined data where the class label for each instance is being provided. However, sometimes there are many cases where the data input is not labeled. The manual curation for each instance is a time and cost consuming. Therefore, semi-supervised learning comes to overcome this issue by giving labels for a small portion of the data then use such portion for the training. In particular, the authors have applied the concept of semi-supervised learning by providing labels for a few instances, then compute the similarity between each labeled instance and unlabelled instance. The process will continue until all the unlabelled instances are being labeled based on their similarity to the predefined ones. Finally, a Support Vector Machine classifier has been used to classify the instances. Experiment results showed a relatively good performance and a fair ability to handle large-scale data of short text.

Lu et al. (2015) have addressed the problem of short text classification by proposing a combination of semantic and statistical approaches. Since the short text is containing a limited number of words, the authors have utilized an external knowledge source in which the words of the short text is being employed to get related instances. In order to perform such search, the authors have utilized a Latent Semantic Analysis in order to identify the similarity between every word in the short text and the portions of the knowledge source. LSA has the ability to determine the semantic similarity between two sets of text using the statistics of co-occurrences. Finally, the keywords obtained from the knowledge source have been used to weight the terms of the short text.

Zhang & Zhong (2016) have enhanced the representation of short text in order to improve the classification process. The authors have utilized a statistical approach called Latent Dirichlet Allocation (LDA) which is similar to LSA. In this manner, both the words for the short text and the topics are being processed using the LDA. The results can be depicted as enriching the representation of short text. Experiment results showed an enhancement in the classification accuracy.

Santos & Rodrigues (2009) have addressed a new kind of short text classification which called ‘multi-label’ text classification where the single document is belonging to multiple class labels. Apparently, the traditional classifiers would face several challenges to handle such task. Therefore, the authors have utilized extensions of multiple classes such as NB, KNN, and SMO in order to handle the multi-label text classification task.

As shown in the related work, there are several approaches have been used for solving the problem of shortness and hierarchical taxonomy of the text. However, there is still a lack of generating contextual information from the text. The contextual information plays an essential role in terms of determining the class label of any text set. Specifically, when the text is underlying multiple class labels. Hence, this study aims to propose an expansion method in order to enrich the contextual information of text for the sake of hierarchical and multi-label text classification.

MATERIALS AND METHODS

This section aims to identify the phases of the research method in which the objectives are being accomplished. This study is composed of six main phases as shown in Fig 1. The first phase is the corpus where a set of short text documents that have multi-label classes are being processed as an input. Consequentially, the second phase which is normalization aims to normalize the documents by eliminating the noisy data such as digits, punctuations, and stopwords. The third phase is associated with the Part-Of-Speech (POS) tagging. In addition, the fourth phase is associated with the proposed expansion method which aims to expand the short text using a knowledge source. The fifth phase aims to apply a feature extraction using the modified version of TF-IDF which is Interesting Term Count (ITC). Finally, the sixth phase aims to apply the powerset-label mechanism using SVM to perform the hierarchical classification.

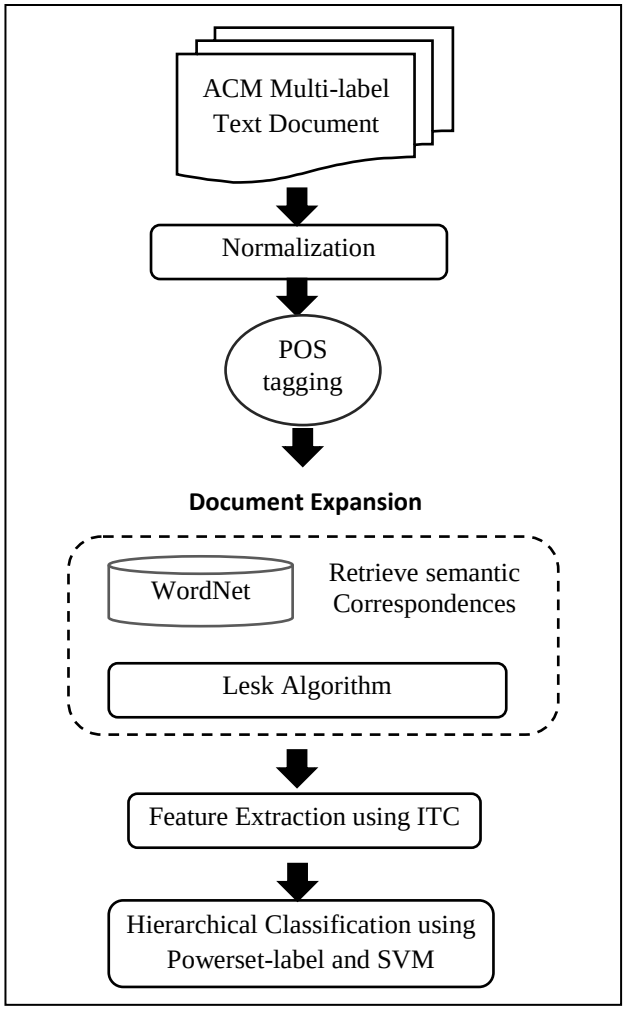

Figure 1. Proposed method's phases

Dataset

The dataset used in this study is called 'ACM Multi-label dataset' which consists of a variety of text document that has one or more class labels. This dataset contains documents with short text. It can be browsable from here (Santos & Rodrigues 2009). **Error! Reference source not found.** depicts the details of such dataset.

Table 1. Dataset details

Description	Quantity
Number of documents	86116
Number of documents with title (Average length)	86116
Number of documents with abstract	53963
Number of documents with one or more categories	54994
Number of documents with one or more keywords	23971
Number of documents with one or more general terms	51574

Table 1 shows a sample of the dataset in which each paper has four attributes including the Publication Id, Title, Keywords and the Class label. In fact, this study has excluded the abstract due to the long text that would be produced when using the abstract with titles and keywords. Therefore, this study will focus only on both title and keywords which are producing an average word count of 100 words thus, it would fit the case of short text classification.

Table 2. Sample of the dataset

Pub Id	Title	Keywords	Class label
1	Is a bot at the controls?: Detecting input data attacks	online games, cheat detection, cheating, security	C.3 SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS
2	User identification based on game-play activity patterns	-	H.1.2 Human factors, K.8.0 Games
3	Voronoi-based adaptive scalable transfer revisited: gain and loss of a Voronoi-based peer-to-peer approach for MMOG	peer-to-peer, Voronoi diagram, massively multiplayer online games, neighbor discovery, scalability, virtual worlds	C.2.1 Distributed networks
4	Skype4Games	Skype, distributed interactive applications, peer-to-peer	C.2.1 Distributed networks
5	Wildlife net-gamekeepers using sensor network	game modification, nature conservation, network monitoring, pattern recognition	K.8.0 Games, C.2.3 Network Operations
6	Adaptive & Delta;-causality control with adaptive dead-reckoning in networked games	causality control, consistency, networked racing game, simulation, subjective assessment	H.4 INFORMATION SYSTEMS APPLICATIONS, K.8.0 Games

As shown in Table 2, every paper has four columns including the Id, title, keywords and the class label. Note that, the keyword column is an optional column where some papers do not have related keywords (e.g. second row in the table). On the other hand, the class label can be represented as hierarchical and multi-label in some cases. The hierarchical class label consists of a parent node which can be formulated via letters from A to K. For example, if there is a general topic of 'Information Systems' then it would be represented as 'C0', C refers to the topic and the number 0 means that it is general. If the paper is related to the more detailed topic that belongs to the 'C' then it would be represented as C1, C2, and C3 (e.g. the first and last row in the table). Sometimes the paper may be classified as a child of the child such as in rows 2, 3, 4 and 5. In this manner, the class label will be represented in multiple indents such as 'C.2.1'. Finally, in terms of the multi-label aspect, some papers are belonging to multiple topics such as the paper in row 2 where the class labels are being formulated as H.1.2 and K.8.0.

Normalization

This phase aims to apply multiple pre-processing tasks such as stopwords elimination, numbers elimination and punctuations elimination. This kind of noisy data have an insignificant impact on the classification process, therefore, it is necessary to get rid of them. The steps of normalization can be explained as follows:

POS Tagging

In order to retrieve the correct sense, POS tagging has been used which aims to provide the syntactic tag for each word such as verb, noun, adjective and so on. POS tagging will play an essential role in terms of determining the required tag senses.

Expansion Method

The normalized documents will be processed using an external knowledge of WordNet in order to expand the short text by retrieving semantic correspondences. This process has the ability to overcome the problems of shortness by disambiguating the words.

Lesk Algorithm

Due to the variety of meanings that would be brought from the WordNet for each term, it is necessary to filter such meanings in order to acquire to most relevant ones. For this purpose, this study adopts the Lesk algorithm which aims to select the meanings that have mutual words.

Let A and B are two words that shared the same meaning. Assume that the word A is being occurred in three contexts C_{a1}, C_{a2} and C_{a3}, similarly, B is being occurred in three contexts C_{b1}, C_{b2} and C_{b3}, the Lesk between A and B can compute as:

$$Lesk\ (A,B) = Max\ C_{ai} \cap C_{bj} \quad (1)$$

where the Max is the maximum intersection between two contexts. This means that the two contexts that have the maximum similarity or matches in terms of the words, will be identified as similar contexts.

Feature Extraction Using ITC

The retrieved senses will be filtered using a statistical approach of Interesting Term Count (ITC) by focusing on the relevant senses and avoiding irrelevant ones. In addition, such statistical approach will be used for the text representation. The reason behind using ITC to represent the data lies in its capabilities in terms of identifying the most interesting terms. Such terms have a significant impact regarding the class labels in which the terms that have important weight would easily guide the classification process to acquire accurate results.

For the ITC, the mechanism of computing the frequency has been adopted to avoid the limitation behind the term frequency by making a new substitution. Such limitation can be expressed as the tedious task of counting the occurrences of insignificant terms by making a condition in which if the term has occurrence higher than 1, the log would be summed with 1, otherwise with zero. The ITC can be computed as follow (Li & Qu 2013):

$$Wid = \frac{\log(tf_{id}) * \log(\frac{N}{n_i} + 0.01)}{\sqrt{\sum_{i=0}^{n} log^2(tf_{id}) * log^2(\frac{N}{n_i} + 0.01)}} \quad (2)$$

Hence, ITC will be applied to the expanded and filtered results by Lesk algorithm in which the relevant synsets for each word will be brought. Each word in the synset will be represented in one column, while the values of ITC for each word will be represented in the instances. These values are indicating the significance of the words. Table 3 shows a sample of representation by ITC.

Table 3. Representation of ITC

No.	Term 1	Term 2	……..	Term n
Document 1	0.2984	0.3465		0.3255
Document 2	0.2765	0.4765		0.5765
.	.	.	……..	.
.	.	.	……..	.
Document m	0.6765	0.4765		0.8765

Hierarchical Classification

Unlike the conventional classification problem where the data instances belong to one of a predefined class label, multi-label classification aims to relate the data instances into multiple class label at the same time. Multi-label classification plays an essential role in different domain of interest such as text, music, images, and videos (Tsoumakas et al. 2010). Multi-label classification problem is considered to be one of the tasks related to multi-objective learning in which the

probability of associating the data instances is not certain but rather it took the context of the fuzzy concept. The uncertainty in the fuzzy concept implies that the data instances may relate to multiple class labels in a form of probability. For example, classifying text document into their corresponding topics would result in tackling some text documents that are belonging to different class labels. In contrast to the conventional classification where the document would be classified into a specific topic such as sport, politics, scientific and others, in the multi-label classification, the document could be classified as 20% sport, 30% politics and 50% scientific.

According to Zhang & Zhou (2014), there are two main methods for applying the multi-label classification namely; One-Against-All and Powerset-Label. The first method aims to break the multi-label class problem into multiple binary problems. Let C is a class label that belongs to C = {C1, C2, C3, C4} and the data X is the instances where X = {X1, X2, X3, X4}. Now in order to solve the problem of multi-label a matrix should be established as shown in Table 4.

Table 4. One-Against-All method

	C1	C2	C3	C4
X1	X1 -	X1 -	X1 +	X1 +
X2	X2 +	X2 -	X2 +	X2 -
X3	X3 -	X3 +	X3 +	X3 -
X4	X4 -	X4 +	X4 -	X4 +

Minus (-) indicate not belonging
Plus (+) indicate belonging

As shown in Table 4, each data instance has been distributed through each class label. Then, the classification will be turned into a binary classification whether the data instance is belonging to such class or not. For example, X1 has been classified in terms of C1, C2, C3, and C4. Results showed that X1 belongs to C3 and C4. However, this method has been criticized due to a specific limitation. Such limitation implies that the incorrect prediction of one class label will significantly affect the whole of multi-label classification.

Therefore, our study will utilize the second method which is Powerset-label where the multi-label classification problem will be applied on each data instance. Table 5 shows the matrix produced by powerset method.

Table 5. Powerset Matrix

	C1	C2	C3	C4
X1	X1 → 20%	X1 → 30%	X1 → 60%	X1 → 80%
X2	X2 → 10%	X2 → 70%	X2 → 30%	X2 → 90%
X3	X3 → 10%	X3 → 90%	X3 → 80%	X3 → 20%
X4	X4 → 20%	X4 → 60%	X4 → 30%	X4 → 70%

With a threshold of 50%

As shown in Table 5, the powerset matrix aims to identify a value that indicates the degree of belongingness of each data instance in accordance with each class label. With a threshold value, the classification can be easily determined whether the data instance is belonging to such class label or not. The threshold value is a margin that determines if the data instance belongs to the class or not. For example, X1 has been classified in terms of the four class labels, the values that above the 50% (the threshold value) have been provided in C3 and C4 therefore, X1 belongs to these class labels.

Note that, In this study, a discretization task has been applied in order to convert the hierarchy into a flat.

RESULTS

The proposed method has been evaluated using the common machine learning evaluation metrics which are precision recall and f-measure. SVM classifier has been adjusted into 80% for training and 20% for testing. Table 6 and Fig. 2 show the results of applying the powerset-label SVM and with the proposed expansion method.

Table 6. Experiment results

Method	Precision	Recall	F-measure
Powerset-label SVM	0.8269	0.7968	0.811571
Powerset-label SVM with the expansion	0.8842	0.8897	0.886941

As shown from Table 6 and Fig. 2, the proposed expansion method has improved the process of hierarchical classification by achieving 88.4% for precision, 88.9% for recall and 88.6% for f-measure. This was superior compared to the application of SVM without the expansion method where the precision was 82.6%, recall was 79.6% and f-measure was 81.1%.

This can demonstrate the efficiency of the proposed expansion method in which the short text has been expanded which leads to improve the contextual information. Such contextual information has led to determine the exact hierarchy of class label.

Comparing the obtained results of the proposed method with a related work of Santos & Rodrigues (2009) who used the same dataset with a different classifier (i.e. Naïve Bayes). Such related work has achieved an f-measure of 83%. Compared to the proposed method, it is obvious that the expansion reveals competitive performance for the task of hierarchical classification.

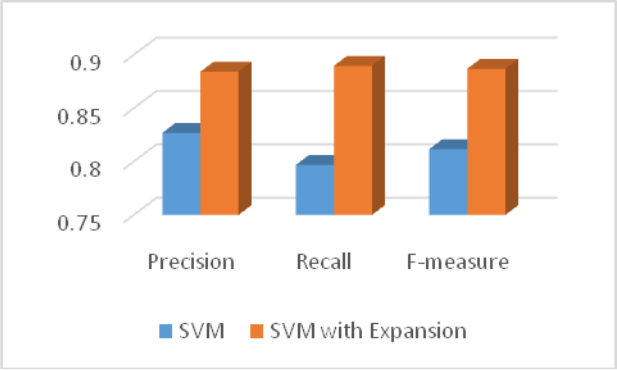

Figure 2. Impact of applying the proposed expansion method

CONCLUSION

This paper has introduced an expansion method for the process of hierarchical classification. Using a benchmark dataset, the proposed method has been tested using the powerset-label mechanism of SVM classification. Experimental results showed that the proposed expansion method has contributed toward improving the classification results. However, one of the limitations of this study lies on the restriction of levels for the classes hierarchy in which three levels have been adopted. Addressing further level would be a challenging task for future researches.

ACKNOLEDGMENT

This project is funded by MOHE under research code FRGS/1/2016/ICT02/UKM/01/14.

REFERENCES

[1] Kiritchenko, S., Zhu, X. & Mohammad, S. M. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50. 723-762.

[2] Kotsiantis, S. B., Zaharakis, I. & Pintelas, P. 2007. Supervised machine learning: A review of classification techniques.

[3] Li, L. & Qu, S. 2013. Short Text Classification Based on Improved ITC. *Journal of Computer and Communications* 2013.

[4] Lu, W., Huang, Y., Li, X., Zhang, Z. & Li, Y. 2015. Short text model based on Strong feature thesaurus.

[5] Santos, A. P. & Rodrigues, F. 2009. Multi-label hierarchical text classification using the acm taxonomy. *14th Portuguese Conference on Artificial Intelligence (EPIA)*, 553-564.

[6] Simoes, G., Galhardas, H. & Coheur, L. 2009. Information Extraction tasks: a survey. *Proc. of INForum*,

[7] Song, G., Ye, Y., Du, X., Huang, X. & Bie, S. 2014. Short Text Classification: A Survey. *Journal of Multimedia* 9(5). 635-643.

[8] Sucar, L. E., Bielza, C., Morales, E. F., Hernandez-Leal, P., Zaragoza, J. H. & Larrañaga, P. 2014. Multi-label classification with Bayesian network-based chain classifiers. *Pattern Recognition Letters* 41. 14-22.

[9] Wang, F., Wang, Z., Li, Z. & Wen, J.-R. 2014. Concept-based short text classification and ranking. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1069-1078.

[10] Yin, C., Xiang, J., Zhang, H., Wang, J., Yin, Z. & Kim, J.-U. 2015a. A new svm method for short text classification based on semi-supervised learning. *Advanced Information Technology and Sensor Application (AITS), 2015 4th International Conference on*, 100-103.

[11] Yin, C., Xiang, J., Zhang, H., Yin, Z. & Wang, J. 2015b. Short Text Classification Algorithm Based on Semi-Supervised Learning and SVM. *International Journal of Multimedia and Ubiquitous Engineering* 10(12). 195-206.

[12] Zhang, H. & Zhong, G. 2016. Improving short text classification by learning vector representations of both words and hidden topics. *Knowledge-Based Systems* 102. 76-86.

[13] Zhang, M.-L. & Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on* 26(8). 1819-1837.

[14] Zhu, X. & Goldberg, A. B. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* 3(1). 1-130.