Hidden Topic Models for Multi-Label Review Classification: An Experimental Study

Thi-Ngan Pham, Thi-Thom Phan, Phuoc-Thao Nguyen and Quang-Thuy Ha
Vietnam National University, Hanoi (VNU), College of Technology (UET),
144, Xuan Thuy, Cau Giay, Hanoi, Vietnam
{nganpt.di12, thompt_53, thaonp_53, thuyhq}@vnu.edu.vn

Introduction

In recent years, Multi-Label Classification (MLC) becomes an important task in the field of Supervised Learning. The MLC tasks are omnipresent in real-world problems, in which an instance could belong simultaneously to different classes. In this paper, an MLC model experimental study on user reviews on Vietnamese hotels is showed. We enriched the data features by using a hidden topic method for short documents of user reviews. We also used mutual information for feature selection. Experiments on user reviews on about one thousand Vietnamese hotels are showed.

Figure 1 describes a proposed model for Multi-Label Classification based on using the Hidden Topic Probability Model, which had been determined by the hidden topic model on the web pages crawled from Vietnamese web sites on tourism and hotels. Sample dataset to train and to evaluate the multi-label classifier is crawled from one website consisted hotel reviews in Vietnamese.

The features determined by the preprocessing step will be adding abstract features by using the hidden topic probability model. After that, a feature selection method based on MI is applied to improve the features for the classifier.

In the last step, a multi-label classifier is build based on a binary transformation method. The classifier will be applied for new reviews.

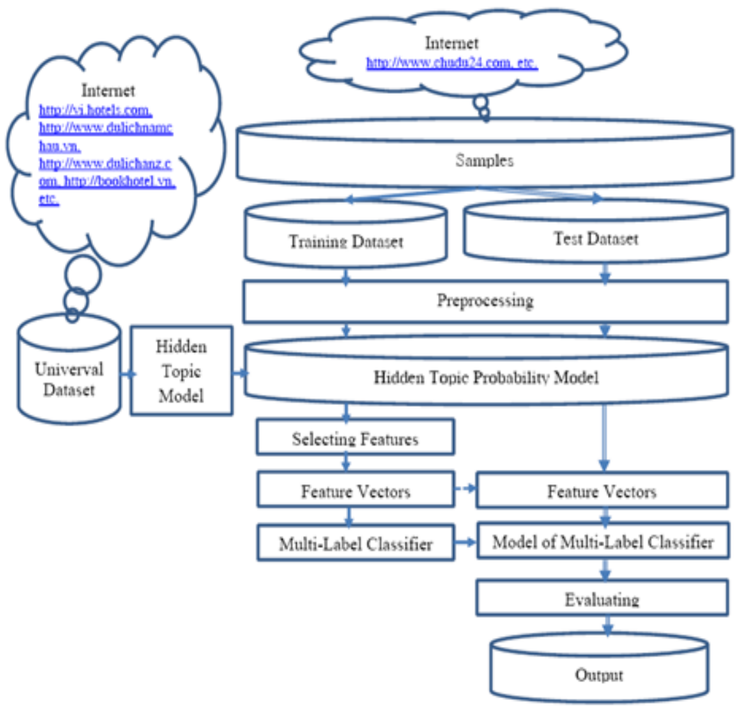

Figure 1. A Hidden Topic Model for Multi-Label Review Classification

Experiments and Results

The process of experiment is described as follows.

- Processing data: Preprocessing data, creating learning data for the classification model, creating data for the LDA model and converting data into vectors,
- Creating function to select features: Using method of MI to select feature set,
- Building classification function: Using method of transformation binary classifier,

Evaluating reputation of 1000 hotels using the most optimal model. In order to evaluate the effect of the solution using the hidden topic model and feature selection, three experiments had been done:

- Experiment 1 (denoted by the TF.IDF case): Classifying with TF.IDF features only (the baseline case);
- Experiment 2: Classifying with the combination of TF.IDF features and the hidden topic features. We considered three cases of LDA with 15 hidden topics (denoted by TF.IDF + LDA_15 topics case), LDA with 20 hidden topics (denoted by TF.IDF + LDA_20 topics case), and LDA with 25 hidden topics (denoted by TF.IDF + LDA_25 topics case);
- Experiment 3 (denoted by the TF.IDF + LDA_20 topics+Feature Selection case): Classifying with the combination of TF.IDF features and the hidden topic features (TF.IDF + LDA_20 topics case) and using feature selection based on MI.

A 5-fold cross-validation based on the Precision, the Recall and the F1 has been applied. The Precision indicates the percentage of system positives that are true instances, the Recall indicates the percentage of true instances that the system has retrieved, and the F1 is a combination of the two measures as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The results of the experiments are described in the Table 2. The experiments show that solutions to enrich features based on hidden topic probability model and to select features based on MI give an 1% improvement to the performance of the MLC.

Table 2. The results of the experiments

Average of 5-folds valuation	Precision	Recall	F1
TF.IDF	0.6764	0.7025	0.6804
TF.IDF + LDA_15 topics	0.6798	0.7056	0.6842
TF.IDF + LDA_20 topics	0.6827	0.7125	0.6883
TF.IDF + LDA_25 topics	0.6793	0.7075	0.6844
TF.IDF + LDA_20 topics+Feature Selection	0.6835	0.7108	0.6890

This paper shows an experimental study on the MLC models on Vietnamese reviews. There are some solutions to improve the features for MLC. Firstly, the feature set has been enriched by adding the hidden topic features. Secondly, The combination feature set has been improved by using a feature selection method based on mutual information. Some experiments have been implemented and gave an 1% improvement to the performance of the MLC. The "universal dataset" in domain of user reviews on Vietnamese hotels with small size may be not enough for a effective hidden topic probability model.

The work should be upgraded by some skilful solutions. Firstly, the universal dataset for hidden topic model should be extended. Secondly, the method to select features for MLC should be modified. Lastly, advanced solutions for building multi-label classifiers should be considered.

1. David M. Blei (2012). Probabilistic topic models. *Commun. ACM (CACM)* 55(4):77-84.
2. David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research (JMLR)* 3:993-1022.
3. Manoranjan Dash, Huan Liu (1997). Feature Selection for Classification. *Intell. Data Anal. (IDA)* 1(1-4):131-156.
4. Gauthier Doquire, Michel Verleysen (2011). Feature Selection for Multi-label Classification Problems. *IWANN 2011*:9-16.
5. Gauthier Doquire, Michel Verleysen (2012). A Comparison of Multivariate Mutual Information Estimators for Feature Selection. *ICPRAM 2012*:176-185.
6. André Elisseeff, Jason Weston (2001). A kernel method for multi-labelled classification. *NIPS 2001*:681-687
7. Jana Novovicová, Antonín Malík, Pavel Pudil (2004). Feature Selection Using Improved Mutual Information for Text Classification. *SSPR/SPR 2004*: 1010-1017.
8. Daniel Ramage, David Hall, Ramesh Nallapati, Christopher D. Manning (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. EMNLP 2009:248-256
9. Juho Rousu, Craig Saunders, Sándor Szedmák, John Shawe-Taylor (2006). Kernel-Based Learning of Hierarchical Multilabel Classification Models. *Journal of Machine Learning Research* 7: 1601-1626.
10. Carlos Nascimento Silla Jr., Alex Alves Freitas (2011). A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov. (DATAMINE)* 22(1-2):31-72.
11. Grigorios Tsoumakas, Ioannis Katakis (2007). Multi-Label Classification: An Overview. *IJDWM (JDWM)* 3(3):1-13.
12. Grigorios Tsoumakas, Ioannis Katakis, Ioannis P. Vlahavas (2010). Mining Multi-label Data. *Data Mining and Knowledge Discovery Handbook 2010*:667-685.
13. Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, Ioannis P. Vlahavas (2008). Multi-Label Classification of Music into Emotions. *ISMIR 2008*:325-330.
14. Grigorios Tsoumakas, Min-Ling Zhang, Zhi-Hua Zhou (2012). Introduction to the special issue on learning from multi-label data. *Machine Learning* 88(1-2): 1-4.
15. Vanessa Gómez-Verdejo, Michel Verleysen, Jérôme Fleury (2007). Information-Theoretic Feature Selection for the Classification of Hysteresis Curves. *IWANN 2007*: 522-529.
16. Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, and Quang-Thuy Ha (2011). Classification and Contextual Match on the Web with Hidden Topics from Large Data Collections, *IEEE Transactions on Knowledge and Data Engineering* 23(7): 961-976, July 2011.
17. Min-Ling Zhang, Jose M. Pena, Victor Robles (2009). Feature selection for multi-label naive Bayes classification, *Information Sciences* 179 (2009): 3218–3229.
18. Min-Ling Zhang, Zhi-Hua Zhou (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition (PR)* 40(7):2038-2048.
19. Yin Zhang, Zhi-Hua Zhou (2010). Multilabel dimensionality reduction via dependence maximization. *TKDD* 4(3).
20. Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, Yu-Feng Li (2012). Multi-instance multi-label learning. Artif. Intell. 176(1): 2291-2320 (2012).