Summary of Text Categorization based on Maximum Entropy Model

Mingcai Li [a, *], Meiling Jin [b]

School of Management Science and Engineering, Shandong Normal University, Ji'nan 250014, China

[a]qilulimingcai@163.com, [b]307457267@qq.com

Abstract

Since 1990s, the maximum entropy model has been used in text categorization and achieves good results in Natural Language Processing since its framework and algorithm were established. On the basis of the Maximum Entropy Model, scholars improve it and make a more in-depth study. Using Maximum Entropy Model for text sentiment categorization has become a hot research topic in recent years. In this paper, the application of Maximum Entropy Model in text categorization is analyzed and classified into three categories: text categorization based on the original Maximum Entropy Model, text categorization based on improved Maximum Entropy Model and text emotion categorization based on Maximum Entropy Model. The authors consider that the existing text categorization using Maximum Entropy Model is to classify the text into a certain category directly, but not to give the probability that a text belongs to a category based on characteristic of Maximum Entropy Model. Therefore, the future research focus will be on the use of the maximum entropy model in t fuzzy classification information.

Keywords

Maximum entropy model, Text classification, Feature selection.

1. Introduction

With the advent of the era of big data, information categorization has become a hot issue in the field of information dissemination and processing. There are great challenges in classification of natural language text. The ME (Maximum Entropy) Model has been widely used in the field of text categorization, and scholars have improved the original ME Model to get better accuracy. At the same time, someone applies the ME Model for text emotion categorization, and compares categorization accuracy between ME Model and other common classifier in the same environment. Lots of experiments show that the ME Model has better performance than others.

In this paper, the application of ME Model in text categorization is analyzed and classified into three categories. The first category is text categorization based on the original ME Model. The second category is text categorization based on improved ME Model. With the development of emotion analysis research, there are more and more research about text sentiment categorization based on ME Model, and we consider this as the third category.

Scholars from around the world have similar research ideas in text categorization based on ME Model. With the deepening of research, some scholars improve the feature selection method, feature function, training set and so on to improve classification accuracy. In this paper, author cards research results, summarize the shortcomings of the existing research and put forward the future research directions combining with the characteristics and advantages of ME model.

2. Related work

Information Entropy [1] was first proposed by Shannon, and he defines information as "something to eliminate uncertainty". Information Entropy is used to represent a measure of uncertainty, when uncertainty is greater, the entropy is greater. After Jaynes proposing the principle of maximum entropy, scholars have established ME Model. Since ME Model's framework and algorithm were established, ME Model has been widely used in Natural Language Processing, especially in text categorization. It becomes most successful Machine learning method in Natural Language Processing in recent years.

2.1 Text classification based on the original maximum entropy model

Xuetian Chen and Ronglu Li [2] use Maximum Entropy Model for text and use Absolute-discounting to improve feature function. They collect more than Twenty thousand pieces of news to test the classification accuracy of the maximum entropy model and get good classification results. Ronglu Li etc.[3] compare and analyze its categorization performance using different approaches for text feature generation, different number of features and smoothing technique and find the best practices in their further study of Chinese text categorization. They compare it to Bayes, KNN and SVM, and show that its performance is higher than Bayes and comparable with KNN and SVM. Gu Bo and Liu Kaiying [4] compare the two models of decision tree and Maximum Entropy in text categorization. They do experiments with Boolean value condition and adding the frequency of words. The results show the Maximum Entropy performance is higher by 20% in precision than decision tree.

Li Junhui etc. [5] propose a approach to apply Maximum Entropy Model to Chinese text categorization and its main points include constituting feature functions and pre-disposing documents. They use removing html tags, word segmentation, filtering empty words and removing stop words to make the feature items more reasonable. They also propose feature-template which is combined with feature item’s weight while constituting feature functions. Their experiments show that the micro-average precision with Maximum Entropy Model based on the combination of feature-template and weigh is better than that with Maximum Entropy Model based on word- frequency.

Qu Zhiyi etc. [6] implement a classification system using Maximum Entropy which is based on keywords semantic repetition. They overcome the shortcomings of Chinese word segmentation.

Some scholars improve a high-accuracy maximum entropy classifier by combining an ensemble of classifiers with neural network voting [7]. This classifier has superior performance both over a single classifier as well as over the use of the traditional weighted-sum voting approach. Jun’ichi Kazama and Jun’ichi Tsujii[8] propose the use of box-type inequality constraints, where equality can be violated up to certain predefined levels that reflect this uncertainty. They evaluate the inequality Maximum Entropy models on text categorization datasets by a text categorization case. They demonstrate the advantages of Maximum Entropy Models with Inequality Constraints over standard Maximum Entropy estimation.

2.2 Text categorization Improved maximum entropy model

Fu Lin [9] extends the maximum entropy model to non-extensive entropy and proposes a non-extensive entropy model. She adds higher-order constraint to the model and constructs co-occurrence relation constraint between word group to improve text categorization accuracy. Zhang Xinhua [10] builds ME text classifier using semi-supervised learning. She considers that it is preferable to learn classifiers from a small amount of labeled, examples and a large, example of unlabeled data and such unlabeled data are easily and abundantly, available. It improves the efficiency of ME Model in text categorization.

Jihong Cai and Fei Song [11] explore the use of different feature selection methods for text categorization using maximum entropy modeling. They also propose a new feature selection method based on the difference between the relative document frequencies of a feature for both relevant and irrelevant classes. Their experiments show the improved feature selection performs better than the other feature selection methods. Some scholars use other feature selection method to improve ME Model. Qi Ruihua etc. [12] propose a mixed make feature selection method based on conditional maximum entropy gain and feature frequency. They also make feature compensation using Gauss prior smoothing to solve feature deletion problem, which can make text categorization more accurate. Masaki Murata etc. [13] also propose feature extraction method. They employ ME for natural language processing problems including machine translation and information extraction successfully.

Estimation of categories, extraction of important features, and correction of error data items is useful in text categorization.

X.-S. He and C.-C. Yang[14] find ME Model has different training results to different testing documents, that is, the stability of it is worse. Thus, they improve ME Model using boosting mechanism in order to advance its stability. Experimental results show that the improving method is valid in improving the stability and the classification accuracy of the text categorization algorithm based on maximum entropy model. Other scholars try to improve ME model in other ways. Li Xuexiang[15] propose an improved ME test categorization ,which fully combines c-means and ME algorithm advantages. The algorithm firstly takes Shannon entropy as ME Model of the objective function, simplifies classifier expression form, and then uses c-means algorithm to classify the optimal feature. This method can quickly get the optimal classification feature subsets, greatly improve text categorization accuracy, compared with the traditional text categorization. Zhang Aike[16] uses a similar method to improve ME Model. The improved maximum entropy c-means clustering text classification method combined the c-means clustering algorithm and the ME algorithm, and it can also fast obtain the optimal classification feature subset.

Alfons Juan, David Vilar and Hermann Ney[17] research the relationship between the naive Bayes and maximum entropy approaches to text classification. They show that both approaches are simply two different ways of doing parameter estimation for a common log-linear model of class posteriors and how to map the solution given by maximum entropy into an optimal solution for naive Bayes according to the conditional maximum likelihood criterion which can improve the classification accuracy.

There are some scholars add some other methods to ME to improve the classification accuracy. Chunyong Yin and Jinwen Xi[18] study the mobile text classification technology based on the maximum entropy model and implement the automatic classification system of texts in cloud computing, and through technical improvements, for a large number of documents in the network, given technical solutions in mobile environment. They propose the text classification methods and features of the maximum entropy model with improved information gain selection method and the pretreatment method and the MapReduce programming method. Xiao Xue[19] proposes hierarchical text categorization methods base on ME Model. She organizes categories into hierarchical structure according to the certain relations and the hierarchical classification performance of ME Model is better than that of plane methods.

2.3 Text emotion Categorization Based on Maximum Entropy Model

To improve the user experience and provide assistant decision for product improvement by mining useful information from comment data published by users on the net, text sentiment analysis is a hot topic in information analysis field. Automatically classifying user emotions can help us understand the preferences of the general public.

Bo Chen, Hui He and Jun Guo [20] research how to extract useful and meaningful language features and how to construct appropriate language models efficiently. They conduct a Global-Filtering and Local-Weighting strategy to select and evaluate language features in a series of n-grams with different orders and within various distance-windows. They adopt ME modeling methods to construct language model framework and construct two kinds of improved models with Gaussian and exponential priors respectively. This method is more suitable for the text subjectivity analysis task in experiments. Zhang Lei etc. [21] use ME Model to predict the opinion-relevant product feature relations.

Jun Li etc. [22] propose a multi-label Maximum Entropy (MME) Model for user emotion classification over short text. To improve the robustness of the method on varied-scale corpora, they further develop a co-training algorithm for MME and use the L-BFGS algorithm for the generalized MME model. Ma Changlin etc. [23] propose a novel topic and sentiment joint Maximum Entropy LDA model for fine-grained opinion mining. Based on the above research, Wang Meng[24] intensive studies how to add ME component to LAD model to distinguish viewpoint words, feature words and background words and do further clustering.

Huang Wenming etc. [25] established an emotion analysis model based on ME. Considering the characteristics that effect feature in each feature in each feature vector is few, they introduce a three-dimensional topic model as important complement of the model to assist sentiment. The emotion analysis model was verified to be solid through experiments based on comment data divided by the names of products.

3. Prospect

According to the research status of text categorization based on ME Model the characteristics of text categorization, and considering that there may be sparse sample and other issues in the process of classification, the author believes that text categorization based on ME Model research can proceed from the following aspects to find a breakthrough in the future.

First, the text categorization based on ME Model is a kind of machine learning essentially. Therefore, it is very important to select a appropriate training set. Whether the training set is reasonable or not will affect the final classification accuracy directly. A training set constituted with different category of information, which must have some information can belong to different categories. The author considers these information as fuzzy information. Such information should be eliminated from training sets. There may also be outliers in the training set, and outliers should also be removed. If we can solve the problem of the boundary blur and outlier in training sets, the categorization accuracy will be improved effectively.

Second, many scholars improve the feature selection method, but there are still problems. For the features that appear in different kinds of texts have a lower reference value of text categorization. Such features should have lower values in feature functions. Therefore, a new feature selection method can be constructed in the future, which can take into account the reference value of features in classification.

4. Conclusion

In this paper, the author introduces the recent research results of text categorization based on ME Model. Due to the advent of the big data, the number of information increase rapidly, thus, the research on text categorization has become a hot topic.

As one of the most successful text categorization methods, ME Model has been widely used. With the continuous improvement of feature selection methods, smoothing technology and the enlargement of experiments scale, the accuracy of text categorization based on ME Model is getting higher and higher. With the increasing of social media services and network services, the research of text sentiment categorization has become a hot topic. Using ME Model to the text sentiment categorization achieves good results.

At the same time, it is necessary to study the method of constructing training set and the feature selection method which is more suitable for the real environment.

To sum up, the author believes that there will be more research on the improvement of the training sets and the feature selection methods and text emotion categorization of text is still a hot research topic in the future.

References

[1] Claude Elwood Shannon. Bell System Technical Journal.1948, (27).

[2] Xuetian Chen, Ronglu Li. Text Categorization Based on Maximum Entropy Model. Computer Engineering and Applications. 2004,(35):78-79+195.

[3] Ronglu Li, Xiaopeng Tao, Lei Tang, Yunfa Hu. Using for Chinese Text Categorization. Advanced Web Technologies and Applications. 2004:578-587.

[4] Gu Bo, Liu Kaiying. Comparative Research of Decision Tree Model and Maximum Entropy Model in Text Classification. The eighth Joint Conference on Computational Linguistics Proceedings.2005:6.

[5] Li Junhui. A Text Classification Method Based on Maximum Entropy Model. The Second National Conference on Information Retrieval and Content Security Proceedings.2005:8.

[6] Qu Zhiyi, Li Yiwei, Zhang Yantang, Yang Shuguang, Zhang Feifei. A maximum Entropy Text Classification Based on Keywords semantic repetition. Journal of Guangxi Normal University (Natural Science Edition). 2007, (04):204-207.

[7] Philipp Koehn. Combining Multiclass Maximum Entropy Text Classifiers with Neural Network Voting. Advances in Natural Language Processing. 2002: 125-131

[8] Jun'ichi Kazama, Jun'ichi Tsujii. Maximum Entropy Models with Inequality Constraints: A Case Study on Text Categorization. Machine Learning. 2005,(60) : 159–194.

[9] Fu Lin. Text Categorization Based on the Non-Extended Maximum Entropy Model. Tianjin University. 2009.

[10] Zhang Xinhua. Building Maximum Entropy Text Classifier Using Semi-supervised Learning. National University of Singapore. 2011.

[11] Jihong Cai, Fei Song. Maximum Entropy Modeling with Feature Selection for Text Categorization. Information Retrieval Technology.2008: 549-554.

[12] Qi Ruihua, Yang Deli, Hu Runbo. Text Classification Algorithm Based on Maximum Entropy and Compensation Strategy for Unseen Features. Journal of Intelligence. 2010,(29)5:141-147.

[13] Masaki Murata, Kiyotaka Uchimoto, Masao Utiyama, Qing Ma, Ryo Nishimura, Yasuhiko Watanabe, Kouichi Doi, Kentaro Torisawa. Using the Maximum Entropy Method for Natural Language Processing: Category Estimation, Feature Extraction, and Error Correction. Cognitive Computation. 2010,(2)4:272-279.

[14] Shixing He, Chengcheng. Yang. Improvement of text categorization algorithm based on maximum entropy. Journal of Xi'an Petroleum University (Natural Science Edition). 2009, (24) 6:77-79.

[15] Li Xuexiang. Research of Text Categorization Based on Improved Maximum Entropy Algorithm. Computer Science. 2012,(39)6:210-212.

[16] Zhang Aike. Application of text categorization based on improved maximum entropy means clustering algorithm. Application Research of computers. 2012, (29)4: 1297-1299.

[17] Alfons Juan, David Vilar, Hermann Ney. Bridging the Gap between Naive Bayes and Maximum Entropy Text Classification. Proceedings of the 7th International Workshop on Pattern Recognition in Information Systems. 2007.

[18] Chunyong Yin, Jinwen Xi. Maximum Entropy Model for Mobile Text Classification in Cloud Computing Using Improved Information Gain Algorithm. Multimedia Tools and Applications. 2016:1-17.

[19] Xiao Xue. Hierarchical Text Categorization Methods Base on Maximum Entorpy Model. Computer and Network. 2015,(09):36-38.

[20] Bo Chen, Hui He, Jun Guo. Constructing Maximum Entropy Language Models for Movie Review Subjectivity Analysis. Journal of Computer Science and Technology. 2008,(23) 2: 231-239.

[21]Huang Wenming, Sun Yanqiu. Chinese Short Text Sentiment Analysis Based on Maximum Entropy. Computer Engineering and Design. 2017, (01):138-143.

[22] Jun Li , Yanghui Rao, Fengmei Jin, Huijun Chen, Xiyun Xiang. Multi-label Maximum Entropy Model for Social Emotion Classification over Short Text. Neurocomputing. 2016, (210):247-256.

[23] Ma Changlin, Xie Luodi, Si Qi, Wang Meng. Fine-grained Opinion Mining Based on Sentiment Dependency and Maximum Entropy Model. Computer Engineering and Science. 2015, (10): 1952-1958.

[24] Wang Meng. Opinion Mining Research Based on Topic and Sentiment Unification Maximum Entropy Model. Huazhong Normal University.2015.

[25]Zhang lei, Li Shan, Peng Jian, Chen Li, Li Hongyou. Feature-Opinion Pairs Classification Based on Dependency Relations and Maximum Entropy Model. Journal of University of Electronic Science and Technology of China. 2014, (03):420-425.