Question Classification Based on Incremental Modified Bayes

LI Ying-wei[1] YU Zheng-tao[1, 2] MENG Xiang-yan[1] CHE Wen-gang[2] MAO Cun-li[1, 2]
(1.The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051; 2.The Institute of Intelligent Information Processing, Computer Technology Application Key Laboratory of Yunnan Province, Kunming 650051)
E-mail:li1225yingwei@126.com

Abstract

How to use the incremental training corpus to improve the question classification accuracy rate in the process of question classification based on statistic learning. A question classification method based on the incremental modified Bayes was presented in this paper. The method used the modified Bayes and combined the incremental learning to correct the parameter by the incremental training set stage by stage, and established the question classification model based on the incremental modified Bayes. A question classification experiment was done in the domain of Yunnan tourism, the experimental results showed that the presented method evidently excelled than the modified Bayes method in the accuracy rate and the training time, the average accuracy rate was improved 3.3 percentage points than the accuracy rate of the modified Bayes method; the average training time was improved 39.1 percentage points than the training time efficiency of the modified Bayes method.

1. Introduction

Question classification is an important part in the question answering system, which is the basis of making the answer extraction strategy and determining the answer semantic type, the classification results can directly affect the accuracy rate of the answer extraction [1, 2]. At present, the question classification research mainly bases on the statistic learning, which through statistic learning to the real tagging question corpus extracts characters which can express each kind of question types, establishes the learning model, and then realizes the recognition of each question type, it has tremendous advantages. Literatures [3-9] research English question classification through statistic learning, the accuracy rate is above 80%, and the effect is better. Because there are many differences between Chinese questions and English questions, compared to English questions, Chinese questions have flexible inquiring ways, so there are many difficulty in Chinese question classification. Zhang Yu advanced a modified Bayes method to realize question classification [1], the accuracy rate was 72.4%, but this method needs a quantity of labeled training samples, and it is much difficult to get labeled training samples and deal with all types of training samples, many training samples probably were offered by the way of increment batch. In incremental text classification research, Gong Xiu-jun advanced an incremental Bayes classification model, which achieved good effect in incremental text classification.

This paper advanced a method of question classification based on the incremental modified Bayes; through the incremental learning, established the question classification model based on the incremental modified Bayes and a question classification experiment was done in the domain of Yunnan tourism.

2. Two ways of question classification model

2.1. Modified Bayes Question Classification Model [1]

The modified Bayes question classification method was presented based on the method Naïve Bayes text classification. Assume that the word distribution is mutually independent in a question, there are not any connections among the words, and the order among words is not important.

According to the independence of probability, the formula which is used to calculate the maximal value of the possibility is as follows:

$$P(c_i, \omega_j) = \frac{0.5 + N(c_i, \omega_j)}{V + \sum_{i=1}^{V} N(c_i, \omega_j)} \times \log \frac{V + 0.1}{M + 0.1} \quad (1)$$

In the formula (1), V expresses the total number of type; M expresses the number of character word ω_j which appears in M types; the constant 0.1 is an adjustment factor, it plays smoothness role.

2.2. Question classification model based on incremental modified Bayes

Based on the modified Bayes method, we provided a question classification method based on the incremental modified Bayes, its description was as follows:

The parameter is the prior probability when the events have happened, $P(\theta \mid I_0)$ is the probability density function, and I_0 is the prior information. According to the rules of Bayes, when the new sample S is offered, the formula which calculates the posterior probability density $P(\theta \mid S, I_0)$ based on the prior probability density $P(\theta \mid I_0)$ is as follows:

$$P(\theta \mid S, I_0) = \frac{P(S \mid \theta, I_0) \times P(\theta \mid I_0)}{P(S \mid I_0)} = \frac{P(S \mid \theta, I_0) \times P(\theta \mid I_0)}{\int [P(S \mid \theta, I_0) \times P(\theta \mid I_0)] d\theta} \quad (2)$$

In the formula (2), when the new sample was offered, the prior information $P(\theta \mid I_0)$ changed into $P(\theta \mid S, I_0)$, namely the posterior information changed into the prior information, which contains the sample information and the prior information:

The posterior information (I_1) = the prior information (I_0) + the sample information (S).

So, the posterior information can be used to be the next prior information. Its flow figure is as follows:

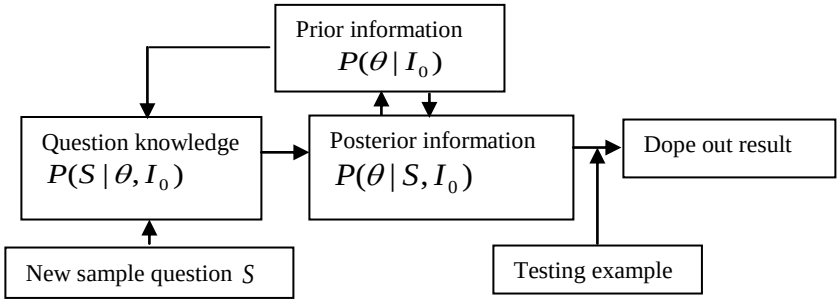

Figure 1. Incremental learning model

When the calculation of the incremental learning parameter was discussed, introduce the definition:

Definition 1: Assume the conditional distribution relative to the parameter θ of the sample S is $P(S|\theta,I_0)$, the prior distribution is $\pi(\theta)$, if $\pi(\theta)$ and the sample S determine the posterior distribution $P(\theta | S, I_0) = \frac{P(S|\theta) \times \pi(\theta)}{P(S)}$, which has the same distribution as $\pi(\theta)$,so $\pi(\theta)$ is named Dirichlet distribution.

Definition 2: Assume the event variable Y has states: $Y^1, Y^2, \cdots, Y^r$, the parameter vector θ = ($\theta_1, \theta_2, \cdots \theta_r$), in this formula $\theta_k = P(Y = Y^k | \theta, I_0)$ ($k = 1,2,...,r$) , if its distribution density is:

$$P(\theta | I_0) = Dir(\alpha_1, \alpha_2, ...\alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \times \prod_{k=1}^{r} \theta_k^{\alpha_k - 1} \quad (3)$$

In the formula (3), $\alpha = \alpha_1 + \alpha_2 + ... + \alpha_r$ .So the parameter vector θ has Dirichlet distribution about the prior information I_0. α_1 , α_2 ,…, α_r is named super parameter.

Through the above-mentioned learning process, the training sample $S = \{C_1, C_2, ..., C_n\}$, $C_n = \{q_1, q_2, ..., q_m\}$ in this formula, C_n expresses question type; q_m expresses some question. In the condition of no information Dirichlet prior assumption, get the estimation of the parameter:

Type probability:

$$\theta_r = P(C_r | \theta) = \frac{1 + count(C_r)}{|C| + |S|} \quad (4)$$

Type conditional probability:

$$\theta_{k|r} = P(q_i | C_r, \theta) = \frac{1 + count(\omega_i \cap C_r)}{|q_i| + count(C_r)} \quad (5)$$

In these formulas, $count(C_r)$ expresses the total number of the word which appears in the type C_r ; $|C|$ expresses the number of the type; $|S|$ expresses the total number of the word which appears in the training set; $count(\omega_i \cap C_r)$ expresses the number of the character word ω_i which appears in the type C_r ; ω_i expresses the character word; $|q_i|$ expresses the number of the character word in a question.

Finally, calculate the maximal value of formula (6), the corresponding type is the type of the question.

$$C_{\max} = \max \theta_r \times \theta_{k|r} \times \log \frac{|C|+0.1}{M+0.1} \quad (6)$$

In the formula (6), $\log \frac{|C|+0.1}{M+0.1}$ expresses using TF-IDF to debase the weight; M expresses the number of character word which appears in M types; the constant 0.1 is an adjustment factor, it plays smoothness role.

In the incremental learning process, mostly deal with the new incremental examples $T = \{C_1, C_2, \ldots, C_n\}$, $C_n = \{q_1^{'}, q_2^{'}, \ldots q_m^{'}\}$.Firstly, use the training set S to learn the classifier, use the formula(4)、(5)to estimate the parameter. To the new incremental examples T , use the current classifier to check up its label, if it is matching, reserve the classifier; otherwise, use the new sample to correct the parameter.

Use the formula(7)、(8) to correct the parameter and replace the formula(4)、(5),put them in the formula(6)to calculate the $C_{\max}$.

Type probability:

$$\theta_r = P(C_r \mid \theta^{'}) = \frac{1 + count(C_r) + count^{'}(C_r)}{|C| + |S| + |S^{'}|} \quad (7)$$

Type conditional probability:

$$\theta_{k|r} = P(q_i \mid C_r, \theta^{'}) = \frac{1 + count(\omega_i \cap C_r) + count^{'}(\omega_i \cap C_r)}{|q_i| + count(C_r) + count^{'}(C_r)} \quad (8)$$

In the formula(7)、(8), $count^{'}(C_r)$ expresses the total number of word which appears in the type C_r of the new incremental sample set; $|S^{'}|$ expresses the total number of the word which appears in the new incremental sample set; $count^{'}(\omega_i \cap C_r)$ expresses the number of the character word ω_i which appears in the type C_r of the new incremental sample set.

3. The experiment and the analysis

In order to test the effect of the classifier, a question classification experiment was put up in the domain of Yunnan tourism. In the experiment, we define 5 coarse types and 23 fine types.

In the domain of Yunnan tourism, we structure the scale of 5000 question instances as the training set S_1, and then add 3000 question instances based on S_1, form the scale of 8000 question instances as the training set S_2, likewise, add 4000 question instances based on S_2, form the scale of 12000 question instances as the training set S_3, add 3000 question instances based on S_3, form the scale of 15000 question instances as the training set S_4, add 7744 question instances based on S_4, form the scale of 22744 question instances as the training set D, and structure the scale of 500 question instances as the testing set T, and then we use the training sets S_1,S_2,S_3,S_4,D to train the Modified Bayes Classifier; and use the training set S_1 to train the Incremental Modified Bayes Classifier, and use the new 3000 question instances as incremental sample set and the training set S_1 to train the Incremental Modified Bayes Classifier. Likewise, use the different scale of incremental question instances as incremental sample set and the training set S_2, S_3, and S_4 to train the Incremental Modified Bayes Classifier respectively.

Table 1. Traveling domain question classification system

Coarse type	Fine type
景点(scenic spot)	景点简介(scenic spot synopsis), 景点位置(scenic spot position), 景点价格(scenic spot price), 景点交通(scenic spot transportation), 景点其他(scenic spot other)
地方(place)	地方简介(place synopsis)、地方位置(place position)、地方交通(place traffic)、地方气候(place climate)、地方其他(place other)
风土民情(local customs and practices)	土特产(native and specialty goods), 风味小吃(savor snacks), 风俗习惯(manners), 历史文化(history culture), 节日文化(festival culture), 民族歌舞(folk music and dance), 风土民情其他(local customs and practices other)
酒店(hotel)	酒店介绍(hotel introduction)、酒店位置(hotel position)、酒店价格(hotel price)、酒店星级(hotel star)、酒店其他(hotel other)
其他(other)	其他(other)

In the process of testing the Modified Bayes Classifier (MB) and the Incremental Modified Bayes Classifier (IMB), we choose the average accuracy rate of 5 coarse types (R5) and the average accuracy rate of 23 fine types (R23) to evaluate the results which were introduced in table 2.

Table 2. The testing results of 2 classifiers

Training set	Testing set	MB		IMB	
		R5 (%)	R23 (%)	R5 (%)	R23 (%)
S_1	T	78.4	65.6	79.2	64.1
S_2	T	79.6	72.8	82.4	75.3
S_3	T	82.5	76.5	86.5	79.1
S_4	T	84.6	78.4	88.1	85.1
D	T	86.2	81.8	90.2	83.2

Analyzing the results of the experiment, when we use the training set S_1 to train the incremental modified Bayes classifier, there is not any discrepancies between the incremental modified Bayes result and the modified Bayes result. The reason is that the classifier model is based on the modified Bayes method, we only use the training set S_1 to train the classifier, have no the process of incremental training. When we use the training set S_2,S_3,S_4,D to train the classifiers, the incremental modified Bayes classifier has preferable effect, compared to the modified Bayes classifier, its accuracy rate is respectively improved2.5%、2.6%、6.7%、1.4%,and its average accuracy rate is improved 3.3%. But in the classification process, part of the questions are not attributed to the correct type, the reasons are as follows:

1) Because the select training corpus is limited, so the corpus could not contain all domain terminology, in the incremental modified Bayes method, we could not calculate the frequency of the word, and then affect the accuracy rate. For example, if the domain terminology "玉龙雪山（YULONGXUESHAN）" is not contained in the type of "景点简介（scenic spot synopsis）", then the question "玉龙雪山的简介是什么？（what is the synopsis of YU LONG XUE SHAN?）" is not classified to the type of " 景点简介（scenic spot synopsis）" accurately.

2) The distribution of the training corpus is not even, such as there are 7986 question instances in the type of " 风土民情（local customs and practices）", but there are only 267 question instances in the type of " 其他（other）". The distribution of the instances is not even which can reduce the accuracy rate.

3) Because the type of the definition is overlapping, part of the questions could attribute to two or more types, for example, the question" 丽江有哪些景点？（what scenic spot does

LIJIANG have?）”, which can attribute to the type of“ 景点（scenic spot）” and the type of“ 地方（place）”, in the testing process, we do not consider the question of overlapping, so it will be wrong in the process of classification, and then affect the accuracy rate.

In the process of testing the training time of the Modified Bayes Classifier (TMB) and the Incremental Modified Bayes Classifier (TIMB), the training time was introduced in table 3.

Table 3. The training time

Training set	Testing set	TMB(min)	TIMB(min)
S_1	T	1.39	1.21
S_2	T	3.92	1.79
S_3	T	5.46	3.07
S_4	T	8.87	5.79
D	T	10.79	8.26

Analyzing the results of the experiment, when using the training set S_1, there are no contrast among the method, the reason is that we only use the training set S_1 to train the classifier, have no the process of incremental training, so the training time will have no contrast. When using the training set S_2, S_3, S_4, D, the incremental modified Bayes classifier has preferable performance, compared to the modified Bayes classifier, its training time is respectively reduced 2.13min、 2.39min、 3.08min、 2.53min, and the average training time is reduced39.1%. This shows that the incremental modified Bayes method can reduce the training time effectively.

4. Conclusions

This paper advanced a method of question classification through the incremental learning, established and realized the question classification model based on incremental modified Bayes（http://222.221.6.149:7001/LIIP/qc.jsp）, and then a question classification experiment was done in the domain of Yunnan tourism. The experimental results showed that the method was feasible and effective.

The future work should be how to use the unlabeled training samples and the incremental modified Bayes model to research the question classification, and how to improve the accuracy rate when we put the classification model in other domains.

5. References

[1] Zhang Yu, Liu Ting, Wen Xu, “Modified Bayesian Model Based Question Classification [J]”, Journal of Chinese information processing, 2005, 19(2):100-105.In Chinese.

[2] Wen Xu, Zhang Yu, Ma Jin-Shan,” Syntactic Structure Parsing Based Chinese Question Classification [J]”, Journal of Chinese information processing, 2006, 20(2):33-39. In Chinese.

[3] Li Xin, Roth Dan, “Learning question classifier [A]”, Proceedings of the 19th International Conference on Computational Linguistics [C], Taipei: Morgan Kaufmann Publishers, 2002, 556-562.

[4] Li Xin, Roth Dan, Small Kevin, “The role of semantic information in learning question classifiers [A]”, Proceedings of the 1st International Joint Conference on Natural Language Processing [C], Berlin: Spring-Verlag, 2004,451 - 458.

[5] Zhang Dell, Lee Wee Sun, “Question classification using support vector machines [A]”, Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval [C], New York:ACM Press, 2003,26 - 32.

[6] Hacioglu Kadri, Ward Wayne, “Question classification using support vector machines and error correcting code [A]”, Proceedings of HLT2NACCL2003[C], Edmonton, 2003, 28-30.

[7] Roth Dan, Cumby Chad, Li Xin, etal, "Question-answering via enhanced understanding of questions [A]", Proceedings of the 11th Text Retrieval Conference [C], Gait hersburg: N IST Special Publication, 2002,667 - 676.

[8] Hermjakob U, "Parsing and question classification for question answering [A]", ACL22001 Workshop on Open-Domain Question Answering [C], Toulouse, 2001, 255-262.

[9] Taira Jun Suzuki, Sasaki Yutaka,Maeda Eisaku, "Question classification using HDAG kernel [A]", ACL Workshop on Multilingual Summarization and Question Answering [C], Sapporo, 2003,61 - 68.

[10] Gong Xiu-jun, Liu Shao-hui, Shi Zhong-zhi, " An incremental Bayes classification model [J]" , Chinese Journal of Computer, 2002, 25(6)：645—650. In Chinese.

Authors

Li Yingwei, born in 1982, master. The research field is Intelligent Information Processing.