用 Naive Bayes 方法协调分类 Web 网页*

范 焱[1], 郑 诚[1,2], 王清毅[1], 蔡庆生[1], 刘 洁[1]

[1](中国科学技术大学 计算机科学与技术系,安徽 合肥 230027);

[2](安徽大学 计算机系,安徽 合肥 230027)

E-mail: tangfan@mail.hf.ah.cn

http://www.ustc.edu.cn

摘要: WWW 上的信息极大丰富,如何从巨量的信息中有效地发现有用的信息,是亟待解决的问题,而 Web 网页的正确分类正是其中的核心问题.针对超文本结构中的结构特征,提出了用 Naive Bayes 方法协调分别利用超文本页面中的文本信息和结构信息进行分类的方法.经实验验证,与只用单种方法对超文本进行分类的方法相比,综合分类法有效地提高了分类的正确率.

关键词: 超文本;Web;分类;机器学习;互联网;数据挖掘;信息检索;WWW

中图法分类号: TP181 **文献标识码**: A

当前,互联网上的信息极大丰富.然而,最终用户能消化吸收的信息量与时间之比呈常量.因此,如何自动、有效地从互联网的巨量信息中获取知识,是亟待解决的问题,而 Web 网页的正确分类是其中的核心问题.

针对这一问题,卡耐基-梅隆大学、IBM 公司等都进行了相应的研究,但当前对超文本分类基本上还是采用对平常文本分类的方法,未能有效地利用 Web 页面中的结构信息,如 title、head、超链接等.即使使用了相应信息,也未能有效地协调各算法进行分类,一般也只是采用投票法或最大值法[1,2],虽然有一定的效果,但稳定性差,对提高分类正确性的效果并不是很明显.而本文是针对超文本分类,用 Bayes 理论方法协调文本和超文本结构信息分类器,对 Web 页面进行分类.实验结果显示,与单独用基于 Bayes 方法的文本和结构信息方法的分类器相比,综合分类器的正确率提高了 5%以上;与单独使用基于文本相似性方法的分类器相比,综合分类器的正确率提高了 4%以上,而且结果相当稳定.

1 超文本分类理论基础

(1) 文本分类

方法 1. Naive Bayes 分类器

在文本分类研究中,Naive Bayes 分类与其他文本分类技术相比更有竞争性[3~7],因此获得了成功使用[8,9].基于 Bayes 定理的文本分类公式如下:

* **收稿日期**: 2000-02-24; **修改日期**: 2000-05-10

基金项目: 国家自然科学基金资助项目(69675016)

作者简介: 范焱(1968-),男,安徽合肥人,博士,工程师,主要研究领域为知识发现,机器学习;郑诚(1964-),男,安徽歙县人,主要研究领域为机器学习,知识发现;王清毅(1962-),男,安徽合肥人,博士,讲师,主要研究领域为知识发现;蔡庆生(1938-),男,重庆人,教授,博士生导师,主要研究领域为人工智能,机器学习,知识发现;刘洁(1972-),男,重庆人,博士,工程师,主要研究领域为机器学习.

$$c^* = \underset{c_j \in C}{\operatorname{argmax}} P(c_j | w_1, w_2, \ldots, w_n) = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i=1}^{n} P(w_i | c_j).$$

方法 2. 文档相似性分类器

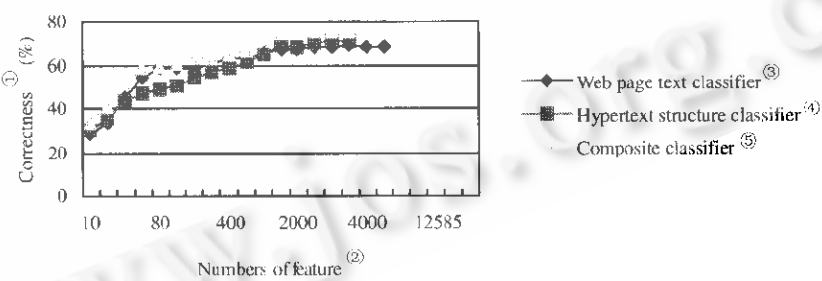
Naive Bayes 分类器假设特性是不互相依赖的,此假设在特性代表一个文档中所包含单词的情形下显然是不正确的. 因此,为了验证 Naive Bayes 方法在协调超文本分类中的通用性,我们还采用了信息检索中常用的文档间的相似形来分类文档. 此方法不需要 Naive Bayes 方法所需要的前提假设. 信息检索中常用两个文档的表示矩阵之间夹角的余弦值来表示它们之间的相似程度,

$$Similarity(x, y) = \cos\theta(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_j x_j^2 \sum_i y_i^2}}.$$

在基于以上两种方法的页面文本分类中,没有使用超文本页面中的任何结构信息,只是将此页面当作一个普通文本来看待.

(2) 超文本结构信息分类

超文本页面中含有大量有用的结构信息,而这些信息中可能会含有该页面标题、重要的子标题等含有该页面重要信息的内容. 如果将这些结构信息用于分类页面,一方面可以有效地提高分类精度,另一方面也可以减少计算的复杂度. 因此,我们将〈title〉和〈/title〉,〈hn〉和〈/hn〉之间的单词提取出来,分别用基于 Naive Bayes 理论方法和基于文档间相似性的分类器进行分类. 可以从图 1 中看出,超文本结构信息分类比页面文本分类的结果有所提高,说明可以用较少的文本信息,达到更好的分类效果,但是稳定性不是很好. 从图 2 中可以看出,在基于文档相似性的分类器中,超文本结构信息分类的结果与页面文本分类的结果相比没有提高,效果较差. 因此,对于不同的分类方法,超文本结构分类器的分类效果不同,稳定性不是很好,但是此方法提供了一个用较少的信息和计算复杂度对超文本进行分类的方法.



①正确率,②特性数,③页面文本分类器,④超文本结构分类器,⑤综合分类器.

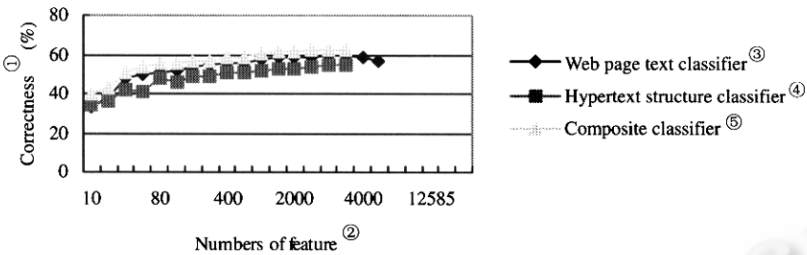
Fig. 1 The comparative figure of classification result of the classifier based on Bayes method

图 1 基于 Bayes 方法的分类器分类结果比较图

(3) 综合分类

页面文本分类器和超文本结构分类器是对 Web 页面的不同的信息内容进行分类的. 可否将两者结合,从而有效地促进分类结果? 基于这种考虑,我们探索了将这两种分类相结合的综合分类方法. 目前一般采用的是二者选最大值法等,但效果并不理想[1,2]. 而超文本结构分类和页面文本分类,二者之间不是互相依赖的,因此,我们尝试采用前面所提到的 Bayes 理论方法,将两个分类器看成是两个不互相依赖的条件因素,结合起来进行综合分类,其公式为

$$Classify = \underset{C_j \in C}{\operatorname{argmax}} \prod_{i=1}^{2} P(C_j | classify_i).$$



①正确率,②特性数,③页面文本分类器,④超文本结构分类器,⑤综合分类器.

Fig. 2 The comparative figure of classification result of the classifier based on similarity method

图 2 基于相似性方法的分类器分类结果比较图

利用上式即可结合文本分类器和超文本结构分类器进行综合分类.但是还有一点值得注意,由文本分类器和超文本结构分类器文档的平均长度相差太大可知,超文本结构分类器的单词数远远少于文本分类器中的单词数.因此,分类构成中就会出现偏差,为了解决这一问题,我们在计算中采取"单个分类器的平均概率"的方法,对不同分类器的结果进行预处理,见下式,其中 n 是 D 中出现的不同单词数.

$$P(C_j|classify_i)=n\sqrt{p(D|C_j)}.$$

从图 1 和图 2 中我们可以明显地看出,在基于 Bayes 方法和相似性方法的分类器的分类结果比较中,综合分类器的分类正确率都要明显好于文本分类器和超文本结构分类器单独分类时的效果,且结果稳定.

2 超文本分类实验

超文本分类实验分为训练过程和测试过程两种.在训练过程中,我们首先抽取训练文档中的 Web 页面的单字,将抽取的单字作为特性,再经过 stemming、单词选取等处理过程,从而学习各待分类文档的特征.在测试过程中,将测试集中的 Web 页面经过前面同样的处理过程后,分别用基于 Bayes 方法和文档相似性方法的文本分类器、超文本结构分类器和综合分类器进行测试,从而得出各分类器的分类结果.

(1) 特性选取

由于在文本训练中涉及大量的特性,因此特性规模的选取是超文本分类中的一个重要问题.例如,本实验从训练文档中抽取的特性,经过词根化和 stop list 两步处理后,实验所涉及的还有 16 789 个.一方面,特性规模过大,计算复杂度过大;另一方面,一些特性项在某些类中发生频繁,有更好的分类效率,而另一些特性项只会给分类带来噪音.所以应采取措施,选取对分类更有效果的特性项用于分类,而摒弃另一些噪音项.

单个特性的评价值能够通过机器学习过程中的特性选取方法获得.为了降低实验复杂度,在本实验中,我们将特性集限制在 2 000 个特性以内.对特性的选取我们采用了对文本分类实验非常有效的机器学习中的 Expected cross entropy 方法[10].此方法类似于决策树归纳中的 information gain 方法[11],但其效果要好于后者[12].

$$crossEntropyTxt(F)=P(F)\sum_i P(C_i|F)\log\frac{P(C_i|F)}{P(C_i)},$$

其中 F 表示特性,$P(F)$ 表示 F 发生的概率,$P(C_i)$ 表示第 i 类发生的概率值,$P(F|C_i)$ 是 F 发生在

第 i 类的条件概率,

(2) 训练和测试

我们用于训练和测试的数据集一部分来自卡耐基-梅隆大学搜集的几所美国著名大学的计算机系的 Web 页面,另一部分来自我们自己收集的互联网上其他美国大学的网页. 所用数据集涉及美国康奈尔大学(Cornell University)、德克萨斯大学奥斯汀分校(Texas University,Austin)、华盛顿大学(University of Washington)、威斯康辛大学(University of Wisconsin)等几所大学的计算机系的 Web 页面,并按照各网页的内容将网页分为 7 个类别:student,course,staff,faculty,other,project. 实验中我们共用了 2 700 个网页,其中 1 353 个网页用于训练,另外 1 347 个网页用于分类测试.

为了降低实验的复杂度,我们将特性数限制在 2 000. 表 1、表 2、图 3 和表 3 分别给出了基于 Bayes 方法的分类实验结果,表 4～表 6 和图 4 分别给出了基于文档相似性的分类实验结果.

Table 1 The classification result of the Web pages' text classifier based on Bayes method ($F=2000$)

表 1 基于 Bayes 方法的 Web 页面文本分类器的分类实验结果($F=2000$)

Actual classification① / Result classification②	Course③	Dept.④	Faculty⑤	Other⑥	Proj⑦	Staff⑧	Student⑨
	200	191	200	200	200	156	200
Course	185	1	3	23	3	0	6
Dept.	7	187	5	3	32	8	30
Faculty	5	1	132	9	26	6	34
Other	2	1	2	101	6	1	0
Proj	1	0	6	3	97	0	3
Staff	0	0	12	1	25	127	45
Student	0	1	10	60	11	14	85
Correctness⑩(%)	92.5	97.9	66	50.5	48.5	81.4	42.5

①实际分类,②结果分类,③课程,④系,⑤教员,⑥其他,⑦项目,⑧职员,⑨学生,⑩正确率.

Table 2 The classification result of the Web pages' hypertext structure classifier based on Bayes method ($F=2000$)

表 2 基于 Bayes 方法的 Web 页面超文本结构分类器的分类实验结果($F=2000$)

Actual classification① / Result classification②	Course③	Dept.④	Faculty⑤	Other⑥	Proj⑦	Staff⑧	Student⑨
	200	191	200	200	200	156	200
Course	183	7	18	35	26	7	18
Dept.	10	181	8	10	38	12	10
Faculty	0	0	119	9	11	10	35
Other	2	0	0	112	6	1	12
Proj	3	1	6	8	110	2	8
Staff	0	0	19	4	1	113	14
Student	2	2	30	22	8	11	103
Correctness⑩(%)	91.5	94.8	59.5	56	55	72.4	51.5

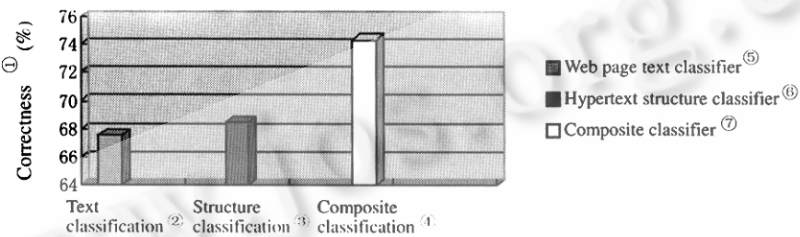
①实际分类,②结果分类,③课程,④系,⑤教员,⑥其他,⑦项目,⑧职员,⑨学生,⑩正确率.

我们以文本分类器的实验结果为基础,对几种分类器的分类结果进行比较. 从图 3 和图 4 可以看出,超文本结构分类器的分类结果同文本分类器相比,在基于 Bayes 的分类中,整体效果要稍好一些;在基于相似性的分类中,整体效果反而较差. 因此,超文本结构分类器的分类结果不够稳定. 但是,在整个训练和测试过程中,超文本结构分类所涉及的文本处理内容大幅度减少,有效地减少了计算复杂度.

从表 3 同表 1、表 2 以及表 6 同表 4、表 5 的分类结果比较中可以很明显地看出,超文本结构分类器与页面文本分类器相比,分类结果时好时坏,结果不够稳定;而综合分类器的分类效果明显好

于前两种分类器单独分类的效果,而且对于每一类的分类结果,效果都很稳定.

从图 3 和图 4 中可以很明显地看出,在基于 Bayes 方法的分类中,综合分类器的分类结果好于页面文本分类器和超文本结构分类器单独分类时 5%以上,综合分类器分类的正确率比前者提高了 6.75%,比后者提高了 5.79%;在基于相似性方法的分类中,综合分类器的分类结果也明显好于页面文本分类器和超文本结构分类器单独分类时 4%以上,综合分类器分类的正确率比前者提高了 4.09%,比后者提高了 9.28%.因此可以得出,用 Naive Bayes 方法协调超文本的文本分类器和超文本结构分类器,可以显著地提高分类正确率.



①正确率,②文本分类,③结构分类,④综合分类,⑤页面文本分类器,⑥超文本结构分类器,⑦综合分类器.

Fig. 3 The comparative figure of classification result based on Bayes method when features number=2000

图 3 特性值=2000 时基于 Bayes 方法的分类结果比较图

Table 3 The classification result of the Web pages' composite classifier based on Bayes method (F=2000)

表 3 基于 Bayes 方法的 Web 页面综合分类器的分类实验结果(F=2000)

Actual classification ① / Result classification ②	Course ③ 200	Dept. ④ 191	Faculty ⑤ 200	Other ⑥ 200	Proj ⑦ 200	Staff ⑧ 156	Student ⑨ 200
Course	189	2	7	33	10	1	8
Dept.	7	187	9	8	29	7	10
Faculty	0	0	133	6	18	8	35
Other	1	0	0	118	7	0	3
Proj	2	1	5	3	125	1	2
Staff	0	0	16	3	3	130	23
Student	1	1	30	29	8	9	119
Correctness ⑩ (%)	94.5	97.9	66.5	59	62.5	83.3	59.5

①实际分类,②结果分类,③课程,④系,⑤教员,⑥其他,⑦项目,⑧职员,⑨学生,⑩正确率.

Table 4 The classification result of the Web pages' text classifier based on similarity method (F=2000)

表 4 基于相似性方法的 Web 页面文本分类器的分类实验结果(F=2000)

Actual classification ① / Result classification ②	Course ③ 200	Dept. ④ 191	Faculty ⑤ 200	Other ⑥ 200	Proj ⑦ 200	Staff ⑧ 156	Student ⑨ 200
Course	157	3	4	10	2	3	4
Dept.	16	175	27	5	18	13	34
Faculty	6	0	103	12	29	14	29
Other	3	0	0	83	8	4	1
Proj	6	1	16	17	106	11	2
Staff	4	4	37	3	22	89	54
Student	8	8	13	70	15	22	76
Correctness ⑩ (%)	78.5	91.62	51.5	41.5	53	57.05	38

①实际分类,②结果分类,③课程,④系,⑤教员,⑥其他,⑦项目,⑧职员,⑨学生,⑩正确率.

Table 5 The classification result of the Web pages' hypertext structure classifier based on similarity method ($F=2000$)

表 5 基于相似性方法的 Web 页面超文本结构分类器的分类实验结果($F=2000$)

Result classification② \ Actual classification①	Course③	Dept.④	Faculty⑤	Other⑥	Proj⑦	Staff⑧	Student⑨
	200	191	200	200	200	156	200
Course	144	12	28	35	10	1	8
Dept.	24	161	19	12	36	16	8
Faculty	0	0	61	18	21	21	16
Other	3	1	3	97	13	13	21
Proj	6	2	3	12	85	7	5
Staff	2	2	35	2	10	66	37
Student	21	13	51	24	25	32	105
Correctness⑩(%)	72	84.29	30.5	48.5	42.5	42.31	52.5

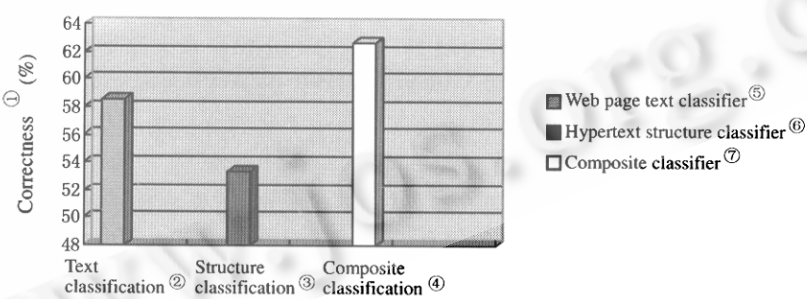
①实际分类,②结果分类,③课程,④系,⑤教员,⑥其他,⑦项目,⑧职员,⑨学生,⑩正确率.

Table 6 The classification result of the Web pages' composite classifier based on similarity method ($F=2000$)

表 6 基于相似性方法的 Web 页面综合分类器的分类实验结果($F=2000$)

Result classification② \ Actual classification①	Course③	Dept.④	Faculty⑤	Other⑥	Proj⑦	Staff⑧	Student⑨
	200	191	200	200	200	156	200
Course	159	4	7	26	4	2	6
Dept.	20	175	17	10	32	14	8
Faculty	0	1	103	12	16	20	19
Other	0	0	0	106	8	6	8
Proj	9	1	3	11	106	9	10
Staff	2	1	50	7	17	89	43
Student	10	9	20	28	17	16	106
Correctness⑩(%)	79.5	91.62	51.5	53	53	57.05	53

①实际分类,②结果分类,③课程,④系,⑤教员,⑥其他,⑦项目,⑧职员,⑨学生,⑩正确率.



①正确率,②文本分类,③结构分类,④综合分类,⑤页面文本分类器,⑥超文本结构分类器,⑦综合分类器.

Fig. 4 The comparative figure of classification result based on similarity method when features number = 2000

图 4 特性值=2000 时基于相似性方法的分类结果比较图

3 结束语

在本文中,我们提出了用 Naive Bayes 协调分类 Web 网页的方法,这是目前协调超文本分类中多种分类器的一种有效方法. 与现有的单个超文本分类器相比,以此方法为基础的综合分类器分类的正确率提高明显. 由于自动搜索引擎的需要,我们将进一步研究、探索,相信能提出更加有效的超文本分类方法. 目前,我们正将该方法用于我们设计的新的互联网搜索引擎和个人软件助理中.

References:

[1] Craven, M., DiPasquo, D., Freitag, D., *et al*. Learning to extract symbolic knowledge from the World Wide Web. Technical Report, MU-CS-98-122, School of Computer Science, CMU, 1998.

[2] Quek, C. Y. Classification of World Wide Web documents [MS. Thesis]. School of Computer Science, CMU, 1997.

[3] Paazzani, M., Billsus, D. Learning and revising user profiles: the identification of interesting Web sites. Machine Learning, 1997,27(3):313~331.

[4] Chakrabarti, S., Dom, B., Agrawal, R., *et al*. Using taxonomy, discriminants, and signatures for navigating in text databases. In: Jarke, M., Carey, M. J., eds. Proceedings of the 23rd International Conference on Very Large Databases (VLDB'97). San Fransisco, CA: Morgan Kaufmann Publishers, 1997. 446~455.

[5] Andrew, McCallum, Kamal, Nigam. A comparison of event models for naive bayes text classification. In: Sahami, M., ed. AAAI-98 Workshop on Learning for Text Categorization. Menlo Park: AAAI Press, 1998. 509~516.

[6] Apte, C., Damerau, F., Weiss, S. M. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 1994,12(3):233~251.

[7] Lang, K. News weeder: learning to filter net-news. In: Preditis, Russell, eds. Proceedings of the 12th International Conference on Machine Learning (ICML-95). San Fransisco, CA: Morgan Kaufmann Publishers, 1995. 331~339.

[8] Mitchell, T. M. Machine Learning. New York: McGraw-Hill, 1997.

[9] KontKanen, P., Myllymaki, P., Silander, T., *et al*. BAYDA: software for Bayesian classification and feature selection. In: Agrawal, R., Stolorz, P. E., Piatetsky-Shapiro, G., eds. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98). Menlo Park: AAAI Press, 1998. 254~258.

[10] Koller, D., Sahami, M. Hierarchically classifying documents using very few words. In: Fisher, D. H., ed. Proceedings of the 14th International Conference on Machine Learning (ICML-97). San Fransisco, CA: Morgan Kaufmann Publishers, 1997. 170~178.

[11] Quinlan, J. R. Constructing Decision Tree in C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993. 17~26.

[12] Dunja, Mladenic, Marko, Grobelnik. Feature selection for unbalanced class distribution and naive Bayes. In: Bratko, I., Dzeroski, S., eds. Proceedings of the 16th International Conference on Machine Learning (ICML-99). San Francisco, CA: Morgan Kaufmann Publishers, 1999. 258~267.

Using Naive Bayes to Coordinate the Classification of Web Pages*

FAN Yan[1], ZHENG Cheng[1,2], WANG Qing-yi[1], CAI Qing-sheng[1], LIU Jie[1]

[1](*Department of Computer Science and Technology, University of Science and Technology of China, Hefei* 230027, *China*);

[2](*Department of Computer Science, Anhui University, Hefei* 230039, *China*)

E-mail: tangfan@mail.hf.ah.cn

http://www.ustc.edu.cn

Abstract: There is a vast source of information in WWW. How to find the useful information from Internet is an exact issue to be solved. The correct classification of Web pages is the core. Based on the structure characteristics of hypertext, the method of Naive Bayes is adopted in this paper to coordinate the two classifiers that use the text document and hypertext structure. Compared with the two separate classifiers, the combining classifier promotes the correctness of Web pages' classification evidently and steadily.

Key words: hypertext; Web; classification; machine learning; Internet; data mining; information retrieval; WWW

* Received February 24, 2000; accepted May 10, 2000

Supported by the National Natural Science Foundation of China under Grant No. 69675016