

一种基于混合概率模型的文本分类方法

吴新玲

(广东技术师范学院 计算机科学学院, 广东 广州 510262)

摘要: 提出了一种基于类的混合概率分类方法. 该方法为每一类文本独立选取能代表其本质特性的主要特征, 即不同类型的文本由不同的主要特征表示, 并基于各类的主要特征分别为每类文本建立相应的概率分布模型, 然后再根据朴素贝叶斯方法对未知类型的文本进行分类. 实验结果表明: 该方法简单有效且易于实现.

关键词: 文本数据挖掘; 文本分类; 属性选择; 概率模型; 多项分布

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2011)11-0133-04

A Text Classification Method Based on Hybrid Probability Model

WU Xin-ling

(Institute of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510262, China)

Abstract: A hybrid probabilistic classification method was proposed based on the class. The method selects the main features that can represent the essential characteristics of a class text independently for every class text. That is different class text will be expressed by different main features. Based on the main features of each class text, the probability distribution model will be established respectively for every class text. Then the unknown class label text can be classified by naive Bayesian method. The experiment results show that the proposed method is simple, effective and easy to implement.

Key words: text data mining; text classification; feature selection; probability model; multinomial distribution

1 引言

文本分类是文本数据挖掘中的一类重要任务, 尤其是随着 Web 技术的发展和广泛应用, Web 文本迅猛增加, 有效的自动文本分类技术更加需求.

文本分类的关键是文本的表示和文本分类模型的建立. 文本通常是非结构化数据对象, 要对其进行自动信息分类, 首先要解决的问题必然是结构化的表示问题, 因此文本表示就是如何将非结构化的字符串转换为结构化数据表示形式. 目前最常见的文本表示方法是向量空间表示模型^[1-4], 此外还有潜在语义索引模型、概率模型^[4]等. 文本经过结构化表示处理后, 就可以应用传统的分类算法进行分类模型训练, 如 K-近邻算法、贝叶斯算法、支持向量机方法^[5-7]等. 目前的文本分类系统多是如此.

传统分类算法通常是将每一个文本转换为相同特征组成的向量表示, 使得样本的向量空间非常稀疏, 这必然影响文本分类的效果和效率. 针对这一问题, 文中提出了一个基于不同特征概率分布的文本分类模型.

2 文本表示

向量空间模型(VSM)是目前最常见的文本表示模型, 在这种模型中, 文本中出现的词汇作为文本特征(或属性), 一篇文本由包含的特征值向量来表示. 例如, 假设文本集包含词汇 $t_1, t_2, \dots, t_m$, 则文本 d 表示为向量 $\vec{d} = (\nu_1, \nu_2, \dots, \nu_m)$, 其中 $\nu_i (i = 1, 2, \dots, m)$ 是特征 t_i 在 d 中的取值. 而根据 ν_i 的取值类型, 向量空间模型又有如下三种形式:

(1) 布尔形式(Boolean), 即 ν_i 是布尔值 0 或 1.

如果 d 包含词汇 t_i 则 $v_i = 1$, 否则 $v_i = 0$.

(2) 词频形式(Term frequency, TF). 假设 n_i 表示 t_i 在 d 中出现的次数, 则 v_i 是 n_i 的函数, 即 $v_i = \text{TF}(n_i)$. TF 可以简单的取值 n_i , 也可以定义为如下频率函数:

$$\text{TF}(n_i) = \frac{n_i}{\sum_{j=1}^m n_j}$$

(3) 词频 / 逆向文本频率形式 (Term frequency/Inverse document frequency, TFIDF). 这里 v_i 是词汇 t_i 在文本 d 中的 TF 值乘以一个逆向文本频率权重 (IDF) 所得. 词汇 t_i 的 IDF 权重是 t_i 在文本集中出现的文本比例. 例如, 在 Cornell SMART 系统中, t_i 的 IDF 权重定义如下:

$$\text{IDF}(t_i) = \log \frac{1 + |D|}{|D_{t_i}|}$$

其中 D 是文本样本集, D_{t_i} 是包含 t_i 的样本集, $|D|$ 表示集合的元素个数. 也可以简单地将 IDF 定义为 $|D| / |D_{t_i}|$.

对文本特征进行 IDF 加权处理, 是为了降低在许多文本中高频出现的词汇频率, 因为这样的词汇对文本分类作用会因为其在太多文本中出现而降低.

文中提出的文本分类方法是分类独立建模, 将采用简单的 TF 表示, 且 $\text{TF}(t_i) = n_i$.

3 文本特征选取

一篇简短的文本一般都会包含大量的不同词汇, 其向量空间的特征数相当庞大, 但由于语言文字的相关性, 其中许多词汇都是相互关联的, 且不同类型的文本会包含许多公共词汇, 对文本分类起不到分类的作用, 因此特征选取是文本分类不可缺少的关键步骤.

特征选取方法已有很多, 如信息增益、粗糙集方法、卡方检验等等. 这些选取方法主要是基于特征的独立性和分类能力来选取. 如信息增益方法优先选取能获得最多分类信息的特征; 粗糙集方法选取能完全区分不同类别对象的特征集, 去掉多余特征; 卡方检验则是通过检验特征属性与类别属性间的相关性, 选取与类别最相关的特征集. 文中基于不同类型文本分别建模的思想, 文本特征选取将分类进行, 即将样本数据集按类别划分, 然后为每一类样本数据集选取最能代表该类文本数据特性的主要特征, 最后基于这些主要特征建立分类模型.

最能反映每类文本数据特性的特征显然应该包

含在那些出现频度高的词汇中. 因此, 采用基于词汇频度的文本特征选取方法: 为每类文本样本数据集设置一个阈值 δ , 选取出现频度超过 δ 的词汇作为该类的分类特征. 此外, 还建一个“删除词汇表”(D-table) 和一个“同义词表”(S-table). 表 D-table 中包含了每类文本中都可能频繁出现, 起不到分类作用的公共词汇, 以及一些文本中经常会大量出现的, 与文本主题内容无关的词汇, 如英文中常用的介词、定冠词等, 在特征选取时, 将删除这些词. 表 S-table 中则包含了若干同义词集及其代表词, 如同义词集 {“people”, “person”, “human being”}, 并设置其代表词为“people”. 特征选取时, 凡属于同一个同义词集的词汇均转换为其代表词. D-table 与 S-table 可以基于作文知识和样本领域知识预先设置, 并可在学习中不断完善、健全. 归纳起来, 对每类文本的特征选取将经过如下主要步骤:

(1) 对每个样本文本, 参照表 D-table 删除无关键词;

(2) 对每个样本文本, 根据表 S-table, 转换全部同义词;

(3) 统计该类中每个剩余词汇的出现频率计数;

(4) 假设词汇 t_i 的频率计数为 n_i , 如果 $n_i > \delta$, 则作为该类文本的一个主要特征.

4 混合概率模型的建立及其分类

假设文本类别集合 $C = \{C_1, C_2, \dots, C_r\}$ 有 r 类; 文本样本数据集 $D = \{d_1, d_2, \dots, d_n\}$ 有 n 个样本实例; 将 D 按类别划分为 r 个不相交子集, $D_1, D_2, \dots, D_r$, 其中 D_k 对应类别为 C_k 的样本集, 并设 T_k 是 D_k 的主要特征集合 ($1 \leq k \leq r$). 令 $T = T_1 \cup T_2 \cup \dots \cup T_r = \{t_1, t_2, \dots, t_m\}$, n_{ij} 是特征 t_i 在文本 d_j 中出现的次数 ($1 \leq i \leq m, 1 \leq j \leq n$), 特征 t_i 在类 C_k 中发生的概率记为 $p(t_i | C_k)$. 则 $p(t_i | C_k)$ 可基于样本由下式估计.

$$p(t_i | C_k) = \sum_{d_j \in D_k} n_{ij} / \sum_{t_i \in T_k} \sum_{d_j \in D_k} n_{ij} \quad (1)$$

而类别 C_k 中文本 d_j 的概率分布则定义为如下的多态分布^[8]:

$$p(d_j | C_k) = \left(\sum_{i=1}^m n_{ij} \right)! \prod_{i=1}^m (p(t_i | C_k)^{n_{ij}} / n_{ij}!) \quad (2)$$

式(2)与朴素贝叶斯假设类似, 但考虑了文本中全部词汇及其各种可能的组合.

由式(1)可知: 如果类 C_k 不含词汇 t_i , 则 $p(t_i |$

$C_k) = 0$, 从而导致 $p(d_j | C_k) = 0$. 解决这一问题的常用方法就是采用拉普拉斯估计. 因此将式(1) 修改为:

$$p(t_i | C_k) = \begin{cases} (1 + \sum_{d_j \in D_k} n_{ij}) / (\sum_{t_i \in T_k} \sum_{d_j \in D_k} n_{ij} + r), & t_i \in T_k \\ 1 / (\sum_{t_i \in T_k} \sum_{d_j \in D_k} n_{ij} + r), & t_i \notin T_k \end{cases} \quad (3)$$

由贝叶斯公式可知文本 d_j 属于 C_k 类的概率为:

$$\begin{aligned} p(C_k | d_j) &= p(d_j | C_k) \cdot p(C_k) / p(d_j) \\ &\propto p(d_j | C_k) \cdot p(C_k) \\ &= p(C_k) \cdot (\sum_{i=1}^m n_{ij})! \prod_{i=1}^m (p(t_i | C_k)^{n_{ij}} / n_{ij}!) \\ &\propto p(C_k) \prod_{i=1}^m (p(t_i | C_k)^{n_{ij}}) \end{aligned} \quad (4)$$

因此,对任一未知类别的文本 d 可根据朴素贝叶斯模型进行分类,即文本 d 的类型是后验概率 $p(C_k | d)$ 值最大的类别. 而由式(4) 可知:对未知类别的文本 d ,只需计算每个类别 C_k 的 $p(C_k) \prod_{i=1}^m p(t_i | C_k)^{n_i}$ 的值即可,即文本 d 的类别是 $p(C_k) \prod_{i=1}^m p(t_i | C_k)^{n_i}$ 最大值对应的类别(其中 n_i 是特征 t_i 在 d 中出现的次数). 如果每类样本数相等,则就是 $\prod_{i=1}^m p(t_i | C_k)^{n_i}$ 值最大的类别.

5 实验

实验数据取自标准数据集 20 — newsgroup, 分别从 guns(talk. politics. guns)、mideast(talk. politics. mideast)、misc(talk. politics. misc) 三个组中随机抽取了 20 个文本作为建模样本数据. 经过删除无关词汇和同义词转换处理,并设置词汇频度阈值 $\delta = 6$,对 guns、mideast、misc 三类文本分别选取的主要特征如表 1 所示.

表 1 guns、mideast、misc 三类文本主要特征

类别	主要特征($\delta=6$)
guns	{case, claim, fbi, fire, firearm, get, government, gun, hear, know, make, post, see, start, state, weapon}
mideast	{armenians, get, go, Israel, jew, kill, know, land, nation, occupy, race, think, translate, Turkish, zionism}
misc	{behavior, child, cramer, death, different, drug, government, homosexual, law, make, president, response, right, sex, story, think, use, want}

根据公式(3)统计各类中主要特征发生的概率如表 2 所示,其中概率值的排列顺序与表 1 中特征顺序一致.

表 2 各类中主要特征的条件概率

类别	$t_i \in T_k$	$t_i \notin T_k$
guns	{0.0580, 0.0507, 0.0580, 0.0652, 0.0580, 0.0870, 0.0580, 0.1304, 0.0652, 0.0507, 0.0725, 0.0725, 0.0870, 0.0797, 0.0507, 0.0507}	0.0072
mideast	{0.0816, 0.0680, 0.0612, 0.0612, 0.1293, 0.0476, 0.0952, 0.0476, 0.0476, 0.0476, 0.0884, 0.0476, 0.1088, 0.0952, 0.0544}	0.0068
misc	{0.0461, 0.0526, 0.0658, 0.0526, 0.0461, 0.0724, 0.0658, 0.0855, 0.0658, 0.0724, 0.0526, 0.0592, 0.0461, 0.0461, 0.0461, 0.0921, 0.0526, 0.0789}	0.0066

从上述三类数据集剩余样本中随机抽取了四个文本作为测试数据,分别记为 dt_1, dt_2, dt_3, dt_4 . 这四个测试文本的实际类别、包含的主要特征和相应的 TF 值以及根据公式(4)计算的类别值见表 3 所示.

表 3 四个测试文本的实际类别、包含的主要特征和相应的 TF 值以及计算类别

测试文本	主要特征及 TF 值	实际类别	计算类别
dt_1	{firearm: 4, get: 2, gun: 17, know:1, see:1, get: 2, think:2, use:1}	guns	guns
dt_2	{ know: 1, make: 1, post:1, Israel:9, think: 1, want:1}	mideast	mideast
dt_3	{ claim: 8, fbi: 5, fire: 10, make: 2, see: 2, start:6, kill:1, death:1}	misc	guns
dt_4	{ get: 1, make: 2, think: 1, child: 6, drug: 2, president:2, want:2}	misc	misc

由表 3 可知文本包含某类的主要特征越多,属于该类的可能性就越大,这正符合建模的基本思想,即同一类型的文本包含更多共同的词汇. 表中 dt_3 的预测类别与实际类别不一致,因为它包含更多 guns 类的主要特征.

6 结束语

文中给出的建模方法是基于类进行的,即为每类文本独立选取主要特征建模,计算简单,每个文本

只需扫描一次,且无需存储样本向量空间.从上述实验数据可知,文中的混合概率模型对文本进行分类,样本是否典型很重要.样本数据的主要特征越集中,越能体现该类文本的特性,模型分类预测的准确率就越高.

实际应用时,可以通过动态调整文本类的主要特征来提高文中混合模型的准确率.即将模型分类错误的文本加入建模样本中,重新统计主要特征.因为新加入的样本数据不会影响原样本集的主要特征,因此新样本的加入只会促使分类准确率的提高.

参考文献:

- [1] Sparck K Jones, Willett P. Readings in information retrieval[M]. San Francisco, CA, USA: Morgan Kaufmann, 1997.
- [2] Sebastiani F. A tutorial on automated text categorization [C]// Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, AR: IEEE, 1999:7-35.
- [3] Fabrizio Sebastiani. Machine learning in automated text categorization[J]. ACM Comput Surv, 2002, 34(1):1-47.
- [4] 杨俊,陈贤富. 基于 KPCA 和 RBF 网络的文本分类研究[J]. 微电子学与计算机, 2010, 27(3):122-125.
- [5] Han J, Kamber M. Data mining: concepts and techniques[M]. [s. l.]:Morgan kaufmann publishers, 2006.
- [6] Hand D, Mannila H, Smyth P. Principles of Data Mining [M]. Beijing: Publishing House of Machinery Industry, 2003.
- [7] 黄双福,陈贤富. 基于改进 SVM 主动学习算法的入侵检测[J]. 微电子学与计算机, 2010, 27(3):75-77.
- [8] Zdravko Markov, Daniel T Larose. Data mining the web: uncovering patterns in web content, structure, and usage[M]. Hoboken, New Jersey: John Wiley & Sons Inc, 2007.
- [9] 王静,保文星. 一种基于差分算法的视频运动目标检测技术[J]. 计算机应用与软件, 2009, 26(12):68-70.
- [10] 孙凌,王硕,李清霞. 一种基于自适应域值混合差分的目标检测方法[J]. 计算机应用与软件, 2009, 26(10):94-97.
- [11] 刘国成,王永骥. 一种基于改进粒子滤波的多目标跟踪算法[J]. 控制与决策, 2009, 24(2):317-320.
- [12] Dyana A, Subramanian M P, Das S. Combining features for shape and motion trajectory of video objects for efficient content based video retrieval[C]//Seventh Inter-

- national Conference on Advances in Pattern Recognition, kolkata:IEEE, 2009:113-116.
- [13] Chaisorn L, Manders C, Rahardja S. Video retrieval—evolution of video segmentation, indexing and search [C]//proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology. Beijing, China:IEEE, 2009:16-20.
- [14] 张哲源,史颖欢,高阳. 基于色彩和形态特征的人阴影检测方法[J]. 微电子学与计算机, 2009, 26(10):86-89.
- [15] Tahayna B, Belkhatir M, alhashmi S. Motion information for video retrieval[C]//IEEE International Conference on Multimedia and Expo, New, York, NY:IEEE, 2009: 870-873.

作者简介:

吴新玲 女, (1963—), 教授. 研究方向为数据挖掘、软件开发方法.

(上接 132 页)

测[J]. 微电子学与计算机, 2005, 22(1): 118-121.

- [2] Meier T, Ngun K N. Video segmentation for content-based coding[J]. IEEE Trans on Circuits and Systems for Video Technology, 1999, 9(8):1190-1203.
- [3] Gupta S, Masound O, Martin R F K, et al. Detection and classification for vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2002, 3(1):37-47.
- [4] 王静,保文星. 一种基于差分算法的视频运动目标检测技术[J]. 计算机应用与软件, 2009, 26(12):68-70.
- [5] 孙凌,王硕,李清霞. 一种基于自适应域值混合差分的目标检测方法[J]. 计算机应用与软件, 2009, 26(10):94-97.
- [6] 刘国成,王永骥. 一种基于改进粒子滤波的多目标跟踪算法[J]. 控制与决策, 2009, 24(2):317-320.
- [7] Dyana A, Subramanian M P, Das S. Combining features for shape and motion trajectory of video objects for efficient content based video retrieval[C]//Seventh International Conference on Advances in Pattern Recognition, kolkata:IEEE, 2009:113-116.
- [8] Chaisorn L, Manders C, Rahardja S. Video retrieval—evolution of video segmentation, indexing and search [C]//proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology. Beijing, China:IEEE, 2009:16-20.
- [9] 张哲源,史颖欢,高阳. 基于色彩和形态特征的人阴影检测方法[J]. 微电子学与计算机, 2009, 26(10):86-89.
- [10] Tahayna B, Belkhatir M, alhashmi S. Motion information for video retrieval[C]//IEEE International Conference on Multimedia and Expo, New, York, NY:IEEE, 2009: 870-873.

作者简介:

姜 靛 男, (1985—), 硕士. 研究方向为基于内容的视频检索.

詹永照 男, (1962—), 教授, 博士生导师. 研究方向为多媒体技术、计算机图形学、人机交互技术.