

基于中心向量法与万有引力模型相结合的文本分类

王 宇, 刘 斌

(南京工业大学 计算机科学与技术学院, 江苏 南京 210000)

摘 要: 提出一种中心向量模型与万有引力模型相结合的文本分类的方法(CGMV), 引入了类内信息熵和类间边界置信度, 解决了传统模型下类间分布不均匀与类内个别样本分布稀疏从而导致分类准确度下降的问题. 经过实验证明 CGMV 算法提高了文本分类的准确率.

关键词: CMV; 文本分类; CGMV

中图分类号: TP393

文献标识码: A

文章编号: 1000—7180(2017)11—0119—05

Text Classification Based on Central Vector Method and Universal Gravitation

WANG Yu, LIU Bin

(School of Computer Science and Technology, Nanjing Tech University, Nanjing 210000, China)

Abstract: This paper presents a text categorization method (CGMV) that combines the central vector model with the gravitation mode. The intra - class information entropy and inter-class boundary confidence degree are introduced to solve the problem that the classification accuracy is degraded due to the heterogeneity of inter - class distribution and the sparse distribution of individual samples in the traditional model. Experiments show that CGMV algorithm improves the efficiency of text categorization.

Key words: CMV; text classification ;CGMV

1 引言

随着信息技术与互联网技术的普及和飞速发展, 各种非结构化、无序的文本信息成为网络信息的主导, 面对海量的文本信息, 如何将文本信息进行准确的分类, 便于有效的组织管理成为信息检索领域的一个严峻的挑战. 当前比较热门的方法就是基于统计学与机器学习理论将自然语言文本表示成机器可以识别的方式, 即自动文本分类 (Automatic Text Categorization) 技术, 或者简称为文本分类, 是指计算机将一篇文档归于预先指定的某一类的过程^[1]. 文本分类技术在信息检索、数字图书馆、自动文摘、文本过滤、单词语义辨析以及文档的组织和管理等多个领域得到了初步的应用^[2]. 文本分类技术得到了很多学者的关注, 他们基于统计学理论与模式识别理论提出了众多的文本分类的方法. 常见的有支

持向量机方法 (SVM)、最邻近的方法 (KNN)、贝叶斯算法 (BYS)、神经网络分类法 (NN) 以及中心向量法 (CV) 等等. 相比于其他的分类算法而言, KNN 方法与 SVM 方法分类准确度可能比较高. 但它们面对要处理海量的文本时, 分类速度却得到了更大的挑战. 中心向量法虽然分类效率极高, 但只适用于对于类别差别大的分类问题上, 对于较复杂的分类问题上表示出对于每个类别的类间分布较为敏感, 分类的准确率有所降低^[3].

本文提出一种中心向量模型与万有引力模型相结合的一种文本分类的方法. 首先为训练集中每一个类别构造一个质心向量, 作为区别其他类别的标签, 计算每篇待测文档与各个类别的对应的质心向量的距离, 然后引入了万有引力定律, 导入类内信息熵和类间边界置信度的概念来调整其训练模型. 经过实验证明文本提出的算法提高了文本分类的

效率.

2 相关工作

2.1 中心向量法

中心向量法^[4]是一种快速的建模方法. 适用于海量密集并且类别差别很大的文本分类中, 它的思想很简单: 利用 VSM 表示待分类的文档, 为每一个类别构建一个原型向量即质心, 即构造一个向量 c_j 来代表类别 j 中的文档, 原型构造为

$$c_j = \alpha \frac{1}{|c_j|} \sum_{d_j \in c_j} \frac{d}{\|d\|} - \beta \frac{1}{|D - c_j|} \sum_{d_j \in D - c_j} \frac{d}{\|d\|} \quad (1)$$

式中, $\|d\|$ 代表向量 d 的欧式距离, D 代表训练集, 两个参数识别相关文档和不相关文档的效果. 对于待分类文档, 计算其文档向量与各个类别对应的质心向量比较其余弦相似度. 归于最大的相似度所在的类别. 计算公式如下:

$$H(x) = \operatorname{argmax}_{c_j} (\langle c_j, x \rangle) = \operatorname{argmax}_{c_j} \frac{c_j}{\|c_j\|} * \frac{x}{\|x\|} \quad (2)$$

中心向量法虽然简单高效, 但它忽略了样本类间分布不均匀的, 该方法表现为对样本的分布过于敏感, 对此很多学者对应用于中心向量法在特征选择阶段做出了改进, 谭松波^[5]等人提出了 BUCC (batch update centroid classifier) 算法, 该算法的核心在于不断调整训练集中的类别比例来动态的改变质心的位置, 引入了学习率 Learnrate 来防止质心的调整过度, 调整后的质心的计算公式为

$$IC_j = C_j + ((\sum d^* - \sum d) * \text{Learnrate}) \quad (3)$$

式中, d^* 代表属于类别 c 的文本, d 代表不属于类别 c 的文本.

该算法提高了对训练集中大多数样本的正确分类. 陈震^[6]等人同样提出了不断动态的进行改变质心的位置的新算法, 运用两种并行框架 Mapreduce 和 BSP 来提出算法的并行策略. 该算法同样提高了分类器的性能.

2.2 万有引力模型

万有引力定律^[7]是牛顿在 1687 年提出的. 牛顿的普适的万有引力定律表示为

$$F = \frac{G * M_1 * M_2}{R^2} \quad (4)$$

式中, G 代表引力常数; M 表示两质点的质量; R 代表两个质点之间的欧式距离.

万有引力的思想是任意两个质点有通过连心线方向上的引力相互吸引. 因此我们可以将万有引力定律引入到我们数据分类的问题上. 质点间的引力与距离的关系正好可以类比为数据点的相似性和它们的距离. 质点的质量可以类比为数据集中选取的特征词对于数据类别的总贡献度. 于是我们自然想到可以将万有引力引入到我们研究的文本分类问题上. 测试文本对于某一个类别距离越近 (相似度越高), 属于该类别的可能性也就越大, 若相似度相差无几, 文本类别中特征词对于该类别的总的贡献度越大, 属于该类别的可能性也就越大.

3 中心向量法与万有引力相结合模型

3.1 模型的构建

设 $c_1, c_2, \dots, c_k$ 为训练集中的所有类别, 每一个类别有 $d_1, d_2, \dots, d_k$ 个文本, d_i 代表第 i 类别中的文本总数. 首先计算每个特征项的平均词频对于该类别的贡献, 计算公式如下:

$$DF(c_i, d_j, t_k) = \frac{d_f(i, j, k)}{|d_j|} \quad (5)$$

式中, $d_f(i, j, k)$ 代表在第 c_i 个类别中第 d_j 个文本的 k 维特征项, $|d_j|$ 代表文本 d_j 的特征项总和.

然后计算每一个特征项 t_k 对于类别 c_i 的贡献度记为 m , 计算公式如下:

$$m = \left(\frac{1}{S} \right) \sum_{j=1}^S DF(c_i, d_j, t_k) \quad (6)$$

式中, S 为类别 c_i 中的文本个数.

接着计算类别 c_i 中的特征词 t_k 的贡献度, 应等于类别 c_i 中所有文本的贡献度之和, 计算公式如下:

$$M_1(c_i, dx) = \sum_{k=1}^n d_f(x_k) * m \quad (7)$$

式中, $d_f(x_k)$ 表示特征项 t_k 在文本 d_x 中出现次数; n 为文本 d_x 的特征数总数.

设待测文本的特征项 t_k 对于文本 d_x 贡献度定义为单位贡献度计为 $M_2(t_k, d_x)$, 其数值上等于 1, 万有引力的常量为 G ; 相似度的参数记为 n .

最终待测文本对于类别 c_i 的引力可以表示为 F_i , 计算公式如下:

$$F_i = \frac{G * M_1(d_x, c_i) * M_2(t_k, d_x)}{[\text{sim}(d_x, c_i)]^n} \quad (8)$$

式中, $\text{sim}(d_x, c_i)$ 表示待测文本 d_x 与类别 c_i 的相似度, n 表示相似度的参数.

为了提高算法的精确度我们引入了两个概念,

类内信息熵与类间边界置信度。

类内信息熵^[8]:设某个类别 C_i 中有 N 篇文档,计算文 d_j 在该类的信息熵,用以反映特征项在预料中的分布情况。计算公式如下:

$$H(t_k, c_i) = - \sum_{j=1}^n \frac{t_f(t_k, d_j)}{t_f(t_k, c_i)} * \log_2 \frac{t_f(t_k, d_j)}{t_f(t_k, c_i)} \quad (9)$$

式中, $t_f(t_k, d_j)$ 表示特征词 t_k 在文档 d_j 出现的频率; $t_f(t_k, c_i)$ 表示特征词 t_k 在类别 c_i 中的频率。显然用 $H(t_k, c_i)$ 表示特征词 t_k 相对于类 c_i 的类内信息熵,类内信息熵越大,说明特征词在类内的分布越均匀,其分类效果也就越好。

类内边界置信度^[9]:我们考虑在使用中心向量法与万有引力结合时,在计算每个类别的质心时,忽略类别的边界对于分类效果的影响。例如待测文本 d_1 正好位于类别1的边界上,但是离类别2质心的距离 d_2 比待测文本离类别1的距离 d_1 要小,即 $d_1 > d_2$ 。我们会误把文档 d_1 归于类别2,造成分类精度下降,如图1所示。

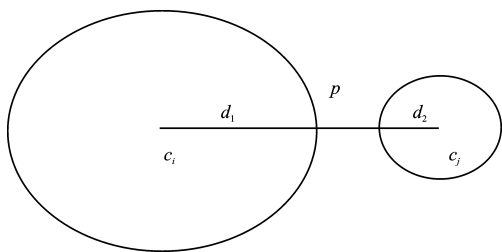


图1 分类精度下降

定义:类别 c_i, c_j 的边界距离为 $Border(c_i, c_j)$,计算公式如下:

$$Border(c_i, c_j) = Sim(b(c_i), x_i) \quad (10)$$

式中, x_i 满足下式:

$$\max_{i \in \{1, \dots, n\}} \{x_i \in c_i, x_j \in c_j, i \neq j\} \quad (11)$$

然后把类别边界代表此类别的真实分布的置信度设为 s 。显然由于误差的影响,训练样本不可能完全表示为所属类别的真实分布,故 $0 < s < 1$ 。文档 x_i 属于类别 c_i 的边界置信度就用以下公式表示:

$$Sim(x_i, c_i) \xrightarrow{change} Sim(x_i, c_i) * s \quad (12)$$

加入类内信息熵与边界置信度这两个概念,改进为以下这个表达式:

$$F_i = \frac{G * M_1(d_x, c_i) * M_2(t_k, d_x) * H(c_i, t_k)}{[Sim(d_x, c_i) * s]^n} \quad (13)$$

式中, $H(t_k, c_i)$ 代表特征词 t_k 对于类别 c_i 的类内信息熵, s 代表类间的置信度。

最后我们通过计算引力 $Class(d_i)$,将引力大小排序来最终确定该文本所属类别从而完成对待测文本的分类。计算公式如下:

$$Class(d_i) = \max(F_1, F_2, \dots, F_n) \quad (14)$$

3.2 算法步骤

名称:CGMV 文本分类算法

输入:训练集文本语料,测试集文本语料

输出:分类结果,文本所属的类别

(1)对文本进行预处理,去除停用词、虚词、标点符号等等。

(2)用于向量空间模型(SVM)来表示训练样本信息。

(3)对(2)表示的训练文本信息选择 CHI 方法进行特征选择。

(4)计算训练集中的每一个类别进行特征选择后,特征项对于文本总贡献度如式(7)与类内信息熵如式(9)。

(5)将测试集中未分类的文本 d_x 用向量空间模型表示。

(6)引入类间置信度 s ,计算待测文本 d_x 与每一个类别的边界距离如式(12)。

(7)依次计算每一个类别对于待测文本 d_x 的引力 F_i 如式(13),最后将待测文本分类到引力最大的那个类别如式(14)。

4 实验设计与分析

4.1 实验数据集

机器学习算法依赖于数据的分析与处理,数据的本质则决定了应用的机器学习算法是否合适,从而成为机器学习算法好坏的一个指标,数据的质量也会决定算法的可行性与准确性。我们研究的文本分类属于自然语言处理领域,在此领域中我们把所应用的数据集则成为语料库^[10]。

本文采用的语料库有3个,都是文本分类中最常用的语料库,它们分别为 TanCropV1.0, 20NewGroup 和复旦大学中文语料库。

TanCropV1.0 它是中科院研究所谭松波所编写的,该语料库共有12个大类,60个小类,一共14 150篇文档^[11],我们在每个类别中选取80%的文本作为训练文本,其余的20%的文本作为测试文本。

20NewGroup 由20 000多篇英文文章组成,分别是20个不同的类别^[12],本实验采用16 000篇文档作为其训练集,其余作为其测试集。

复旦大学中文语料库^[13]由复旦大学李荣陆提供的,它分为 20 个类别,其中 9 833 篇文档为测试语料;训练语料共计有 9 804 篇文档.训练语料和测试语料基本按照 1 : 1 的比例来划分.

4.2 评估指标

文本分类作为一个有监督的学习过程.通过训练过程得到文本与类别标签的一种映射.测试阶段将未知文本归到已知的类别标签.此时我们需要用指标来描述这种映射的好坏.一般我们采用准确率,召回率,F1 值.其分类结果可建立混合矩阵,如表 1 所示.

表 1 分类结果的混合矩阵

	确是该类的 文本数量	确是该类的 文本数量
判断是该类文档数	A	B
判断不是该类文档数	C	D

准确率 P (Precision)表示为

$$P = \frac{A}{A + B} * 100 \% \tag{15}$$

召回率 R (Recall)表示为

$$R = \frac{A}{A + C} * 100 \% \tag{16}$$

准确率和召回率从不同方面考虑分类器的性能,有时候我们需要把他们两个因素结合起来考虑,提出了新的评估标准 F1 值,计算公式如下:

$$F1 = \left(\frac{2 * P * R}{P + R} \right) * 100 \% \tag{17}$$

4.3 实验设计与结果分析

文本采用的 TanCropV1.0 语料库在预处理过程中已经采用了中科院的分词工具(ICTCLIS)进行了分词的处理,然后用于向量空间模型(SVM)来表示训练样本信息,采用卡放检测这种特征选择方法进行处理,并用 tfidf 进行加权,输出结果为文本所属类别.首先利用各个预料来判断本文提出的 CGMV 算法的可行性与准确性,接着比较朴素贝叶斯算法,中心向量法(CMV),与文本提出的 CGMV 算法进行对比分析,另外在实验结论的基础上,调整其参数分别赋予不同的值,从而找到最佳的参数取值.

4.3.1 分类算法参数的实验

CGMV 算法的参数 s 的最佳取值实验如下.

实验步骤:本实验分别在 TanCropV1.0, 20NewGroup,和复旦大学中文语料库这 3 个语料库做对比实验,采用 CHI 这种特征选择方法,并用布尔权值法进行加权,选用 CGMV 方法进行分类时,参数 s 从 0.75 设置到 0.95.比较实验结果找到

最佳的 s 的取值,如图 2 所示.

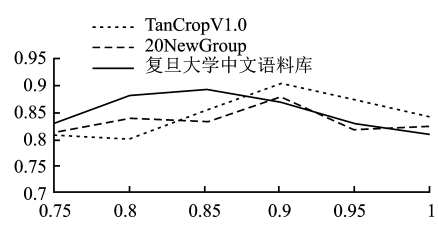


图 2 参数 s 的最佳取值实验比较

通过实验结果可以看出,结合着三种不同的语料库,当参数取值 0.87 时,分类效果比较理想.

CGMV 算法的参数 n 的最佳取值实验

实验步骤:分别在 TanCropV1.0, 20NewGroup 和复旦大学中文语料库这 3 个语料库做对比实验.参数 n 从 1 设置为 3,比较实验结果找到最佳的 n 的取值,如图 3 所示.

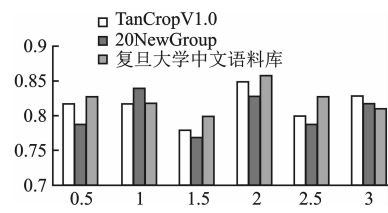


图 3 参数 n 的最佳取值实验比较

4.3.2 分类算法的对比试验

实验步骤:分别在 TanCropV1.0, 20NewGroup, 和复旦大学中文语料库这 3 个语料库做对比实验.并分别采用不同的分类方法.设置 s 为 0.85,并选用准确率,召回率,F1 值这 3 个指标来评价分类器的性能.如表 2 所示.

表 2 三种分类算法运行结果比较 %

		TanCrop V1.0	20New Group	复旦大学 中文语料库
BYS	P	68.25	69.35	71.34
	R	72.13	71.22	73.66
	F1	70.25	72.36	73.23
CMV	P	82.99	79.67	76.36
	R	78.23	78.39	77.13
	F1	80.19	81.12	78.39
CGMV	P	85.36	89.21	86.32
	R	83.27	88.25	88.11
	F1	86.11	89.56	87.07

通过实验结果可以看出,相比于 CMV 和 CGMV 方法,BYS 算法的准确度可能显得偏低,可能是因为运用朴素贝叶斯分类器时假定每一个类别的特征词之间相互独立,这只是一个理想的假设,而事

实情况的语料集很难满足这个限定条件. 而 CMV 算法虽然相比 BYS 分类精度有所提高,但很难保证训练文本集各个类别分布均匀,当各个类别相差不大时很难保证分类的准确. 本文提出的 CGMV 算法克服了传统的 CMV 算法对类别分布的差异较为敏感这一缺陷. 得到的实验结果在分类准确度上有了进一步的提高.

4. 3. 3 不同分类算法时间的对比实验

实验步骤:分类器的时间一般分为训练时间和测试时间,我们统计两者之和作为评价分类器效率的指标,需要指出的是训练时间和测试时间一般不包括文本预处理分词所经历的时间. 实验结果如图 4 所示.

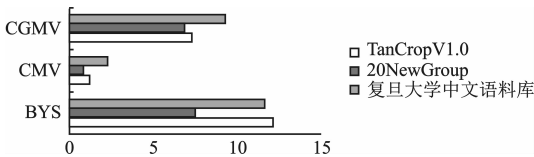


图 4 不同分类算法时间的对比实验

通过实验结果可以看出,相比于 BYS 和 CGMV,CMV 算法的分类速度时最快的,但以上实验表明其实此方法的分类准确率并不是很好. BYS 算法因为需要统计几个类别的后验概率,所以所花费的时间相对较长,CGMV 算法因为引入了类内信息熵和边界置信度,并多次比较样本向量之间的相似度,因此分类所需要的时间相对较多一点.

5 结束语

本文提出一种中心向量模型与万有引力模型相结合的一种文本分类的方法(CGMV). 引入了类内信息熵和类间边界置信度,解决了传统模型下类间分布不均匀与类内个别样本分布稀疏从而导致分类准确度下降的问题. 并且将算法运用到不同的数据集 TanCropV1. 0, 20NewGroup, 和复旦大学中文语料库中并与其他算法进行了多次比对实验. 实验证明文本提出的 CGMV 算法在分类准确度和分类时间上均达到了比较理想的结果.

参考文献:

[1] 朱坤红. 基于 VSM 模型和特征选择算法的中文文本自动分类研究[D]. 南昌:江西师范大学, 2011.

[2] 胡文静. 本分类技术进展[J]. 知识经济, 2011, 13 (10):99.

[3] 赵新苗, 冯向萍, 李永可, 等. 改进的中心向量算法在农业信息分类中的研究[J]. 计算机技术与发展, 2016, 26(8):146-151.

[4] 杨铭. 基于降维中心法的文本分类研究[D]. 长春:吉林大学, 2013.

[5] Straaten S V, Ford E C, Klis M V D, et al. Relation-between timing features and colors in the X-ray binary 4U 0614 + 09[J]. Astrophysical Journal, 2000, 540 (540):1049-1061.

[6] Mbaye I C. An improved parallel K-means clustering algorithm with MapReduce and Hadoop[D]. 长沙:湖南大学, 2015.

[7] 吴来初. 万有引力与向心力公式中 r 的不同含义[J]. 物理通报, 2010, 60(12):76-77.

[8] 陈国松, 黄大荣. 基于信息熵的 TFIDF 文本分类特征选择算法研究[J]. 湖北民族学院学报(自科版), 2008, 26(4):401-404.

[9] Guoyongfeng, Tanjianguo. Time evolution of information entropy for a stochastic system with double singularitiesdriven by quasimonochromatic noise[J]. Chinese Physics B, 2012, 12:99-103.

[10] Jake Miles. Chronic fatigue syndrome[J]. Journal of Clinical Virology37, 1994, 37:139-50.

[11] Hacker B R, Ratschbacher L, Webb L, et al. Exhumation of ultrahigh - pressure continental crust in east central China: Late Triassic - Early Jurassic tectonic unroofing[J]. Journal of Geophysical Research Solid Earth, 2000, 105364(10):339-340.

[12] Yanny B, Rockosi C, Newberg H J, et al. SEGUE: A spectroscopic survey of 240,000 stars with g = 14-20 [J]. Astron. j, 2009, 137(5):4377-4399.

[13] 林琛. 针对训练集分布偏斜问题的数字资源文本分类方法[J]. 现代图书情报技术, 2014, 30 (7) :24-33.

作者简介:

王 宇 男, (1992-), 硕士研究生. 研究方向为数据挖掘. E-mail:15062282096@163. com.

刘 斌 男, (1966-), 博士, 硕士生导师. 研究方向为模糊技术、人工智能、数据挖掘.