对简单向量距离文本分类算法的改进

王治和 杨延娇
(西北师范大学数学与信息科学学院 兰州 730070)

摘 要 分析了简单向量距离文本分类算法的不足,提出了相应的改进算法。把反馈思想引入简单向量距离分类模型,使文本分类系统具备了不断学习的能力。实验证明,改进后的文本分类模型适合于文本分类的需要,改善了原有分类器的性能。

关键词 文本分类,简单向量距离,反馈,分类模型

Improvement of the Vector Space Model Text Classifier

WANG Zhi-he YANG Yan-jiao
(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070, China)

Abstract The paper analyzed the shortcomings of vector space method and put forwards a better method to improve it. It introduced feedback learning into vector space method text classifier, let text categorization system have capability of self-learning. Experiment shows that the revised text categorization model is used to the need of text categorization, and improves the performance of former one.

Keywords Text categorization, Vector space method, Feedback, Categorization model

1 引言

一个好的文本分类方法能够和特征抽取算法相得益彰,取得满意的分类效果。基于向量空间模型的文本分类算法有类中心分类法、贝叶斯方法和神经网络方法等,其中简单向量距离分类法的应用比较广泛。

分类技术是模式识别技术的一个重要应用,模型的学习是其核心内容,各种不同分类方法的研究,在很大程度上是为了从不同的方面更好地提高模型的学习效果。在模型的改善中,反馈技术[1-5]是一种重要的研究方法,它将模型的输出通过一定方式返回到输入中,以改善模型的性能。大量研究成果表明了反馈对改善检索的作用。然而,作为信息挖掘重要应用的文本分类技术,反馈方法在此方面的研究却很少。为此,本文以简单向量距离分类方法为基础,着重探讨反馈技术在文本分类中的有效性,探讨了反馈技术在基于简单向量距离分类中的应用类型,分析了基于简单向量距离法的文本反馈学习技术的具体实现方法。

2 简单向量距离分类法

该方法的分类思路十分简单,根据算数平均为每类文本集生成一个代表该类的中心向量,然后在新文本到来时,确定新文本向量,计算该向量与每类中心向量的距离(相似度),最后判定文本属于与文本距离最近的类。具体步骤如下:

Step1 计算每类文本集的中心向量,计算方法为所有训练文本的算数平均。

Step2 新文本到来后分词,将文本表示为特征向量。

Step3 计算新文本特征向量与每类中心向量间的相似度,公式为:

$$\mathrm{Sim}_{ij}=\sin(d_i,c_j)=\frac{\sum_{k=1}^{m}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{m}W_{ik}^2)(\sum_{k=1}^{m}W_{jk}^2)}}$$

其中,d_i 为新文本特征向量,c_j 为第 j 类的中心向量,m 为特征向量的维数,W_k 为向量的第 k 维。

Step4 比较每类中心向量与新文本的相似度,将文本分到相似度最大的那个类别中。

通过分析发现,该文本分类方法可以划分为训练过程和分类过程,其训练过程就已经决定了系统所具备的分类能力,这个分类能力在将来的分类过程中是固定不变的,所以,简单向量距离法并不具备不断学习的能力。因此,本文提出了一种能够对分类结果进行反馈处理的新算法。该算法在传统的"训练→分类"算法的基础框架上加入反馈的过程,使得算法过程扩展为"训练→分类→反馈判断→反馈"。这种方式更贴近真正意义的机器学习,使得该算法具有一定程度的认知自主性。

3 基于反馈技术的简单向量距离分类法

简单向量距离分类算法简单直观,但对于类别界限不明显时,该方法性能不高。为此,本文以此方法的分类结果为基准来研究相关反馈对文本分类性能的提高。

对基于简单向量距离分类的反馈学习而言,经系统分类后证明是正确的文档,表明原分类模型已包含该文档的相关分类信息,因此它对反馈学习没有价值。而分类错误或不能分类的文档,则说明含有原模型中所不包含的分类信息,是进行反馈学习的重点。因此,基于简单向量距离分类反馈学习

到稿日期:2008-05-02

的重点是针对误分类文档或不能分类文档，而这些文档往往是整个分类结果集合的一小部分。因此，从理论上讲，该学习方法具有反馈样本少的优点。

对于分类处理，根据文档集合的不同可分为训练集、测试集和待分类集。不同的集合在反馈处理时有一定的不同。因此反馈学习也可划分为不同的类型。本文分别对训练文档、测试文档及待分类文档进行了反馈处理。其相关处理方法如下：

1)对训练文档的反馈

在训练文档训练完成后，通过获得的分类模型对训练文档进行封闭测试，产生分类结果集。对其中的错误分类文档集进行反馈处理。通常情况下，错误分类文档集远小于分类结果集。

设计其算法步骤如下：

输入：训练文档集 E，反馈阈值 α_i；

输出：经反馈处理后，各类别的模式矢量。

Step1　对训练文档集中的文档进行词条提取，去除停用词，然后统计词频。每篇文档生成一个向量 d；

Step2　计算向量 d 中每个词条的互信息量；

Step3　根据 TF-IDF 公式计算每个词条的权值 w_i；

Step4　生成特征向量表，每篇文档表示为向量$\langle t_1, w_1; t_2, w_2; \cdots; t_m, w_n\rangle$，$t_i$ 为特征项词条，w_i 为对应的权值；

Step5　对于每一类中的特征项词条 t_i，计算其在该类所有文档特征向量中权值的算术平均值 $\overline{w_i}$，作为该词条在类别特征向量中的权值；

Step6　构造类别特征向量 c_i：$\langle t_1, \overline{w_1}; t_2, \overline{w_2}; \cdots; t_m, \overline{w_m}\rangle$；

Step7　取出训练集中的一篇文档，设其特征向量为 d_v；

Step8　计算 d_v 与各类别特征向量 c_i 的 $\sin(d_v, c_i)$，最大值即为 d_v 所属类别，假设类别为 c_d；

Step9　若 c_d 与 d_v 的原始分类类别相符，则从训练集中新取一篇文档，转至 step7，否则的话进入反馈阶段，转至 Step10；

Step10　如果该文本与原始分类类别的类别相似度比较大的话(大于所设定的一个阈值 α_i)，查询并存储该文本对应的原始分类类别中心模式矢量 $c_i(p_{i1}, p_{i2}, \cdots, p_{ik})$及该主题类别所包含的所有特征项数目 k。

求出类别 c_i 几何平均前的类别中心模式矢量 $c_i(q_{i1}, q_{i2}, \cdots, q_{ik})$，其中：

$q_{ij}=n\times p_{ij}$，$(j=1,2,\cdots,k)$(该主题类别所包含的文本数为 n)　(1)

Step11　按照式(2)，计算类别 c_i 调整后的模式矢量 $(p_{i1}', p_{i2}', \cdots, p_{ik}')$，新的模式矢量代替 $c_i(p_{i1}, p_{i2}, \cdots, p_{ik})$，并用合适的数据结构存储起来；其中：

$$p_{ij}'=\frac{q_{ij}+w_{ij}}{n+1}(j=1,2,\cdots,k)\quad(2)$$

Step12　若训练文档集非空，转至 Step 7，否则，转至 step13；

Step13　反馈过程结束。

2)对测试文档的反馈

测试集的反馈处理流程与训练集基本一致，又有不同，测试集可以认为是具有人工分类信息的待分类文档。对于由训练集合产生的分类器，选取测试集合，由分类器进行分类处理，产生分类结果集，对其中的错误分类文档集进行反馈处理(训练过程和前面所介绍的相同，只给出对测试文档进行反馈的步骤)。

输入：测试文档，对应的主题类别 c_i，反馈阈值 α_i；

输出：对应主题类别 c_i 的模式矢量。

Step1　对测试文档进行预处理，构造特征向量 d_t；

Step2　计算 d_t 与各类别特征向量 c_i 的 $\sin(d_t, c_i)$，最大值即为 d_t 所属类别，假设类别为 c_d；

Step3　若 c_d 与 d_t 实际类别相符，则从测试集中新取一篇文档，转至 Step1，否则的话进入反馈阶段，转至 Step4；

Step4　如果该文本与对应的实际主题类别的类别相似度比较大的话(大于所设定的一个阈值 α_i)，查询并存储该文本对应的主题类别中心模式矢量 $c_i(p_{i1}, p_{i2}, \cdots, p_{ik})$及该主题类别所包含的所有特征项数目 k。

求出类别 c_i 几何平均前的类别中心模式矢量 $c_i(q_{i1}, q_{i2}, \cdots, q_{ik})$，其中：

$q_{ij}=n\times p_{ij}$，$(j=1,2,\cdots,k)$(题类别所包含的文本数为 n)

Step5　按照式(2)，计算类别 c_i 调整后的模式矢量$(p_{i1}', p_{i2}', \cdots, p_{ik}')$，新的模式矢量代替 $c_i(p_{i1}, p_{i2}, \cdots, p_{ik})$，并用合适的数据结构存储起来；其中：

$$p_{ij}'=\frac{q_{ij}+w_{ij}}{n+1}(j=1,2,\cdots,k)$$

Step6　反馈过程结束。

3) 对待分类分档的反馈

以上是对测试文档的反馈过程，我们知道待分类文档是实际分类应用中需要处理的大量无类别信息的文档，是分类模型在应用中的主要对象。为了使系统具备不断学习的能力，把反馈机制也引入待分类文档，其方法和测试文档的反馈过程类似。

输入：新文本的模式矢量 $D(w_1, w_2, \cdots, w_k)$，对应的主题类别 c_i，反馈阈值 α_i；

输出：对应主题类别 c_i 的模式矢量。

Step1　比较新文本 D 与对应主题类别 c_i 之间的相似系数值 β_{DC_i} 与反馈阈值 α_i 的大小，如果 $\beta_{DC_i}>\alpha_i$，则执行 Step2；如果 $\beta_{DC_i}<\alpha_i$，则转至 Step5，结束该算法。

Step2　查询并存储该文本对应的主题类别 c_i 的类别中心模式矢量 $c_i(p_{i1}, p_{i2}, \cdots, p_{ik})$和该主题类别所包含的所有特征项数目 k。

Step3　求出类别 c_i 几何平均前的类别中心模式矢量 $c_i(q_{i1}, q_{i2}, \cdots, q_{ik})$，其中：

$q_{ij}=n\times p_{ij}(j=1,2,\cdots,k)$

Step4　计算类别调整后的模式矢量$(p_{i1}', p_{i2}', \cdots, p_{ik}')$，用新的模式矢量代替 $c_i(p_{i1}, p_{i2}, \cdots, p_{ik})$，并用合适的数据结构存储起来；其中：

$$p_{ij}'=\frac{q_{ij}+w_{ij}}{n+1}(j=1,2,\cdots,k)$$

Step5　反馈过程结束。

4　实验及分析

为验证基于反馈技术的简单向量距离文本分类算法的有效性，文本进行了如下实验：我们选用中文自然语言处理开放

平台提供的语料库，文本类别为6个，依次是计算机、体育、军事、政治、教育、经济，其中计算机类200篇，体育类450篇，军事类250篇，政治类505篇，教育类220篇，经济类325篇，共1950篇。取1500篇作为训练文档集，余下的450篇作为测试文档集。根据反馈类型的不同，分别进行了训练文档和非训练文档的反馈学习实验。非训练分档应包括测试文档及待分类文档，在本实验中都用测试文档代替，采用待分类文档的反馈过程。

本试验在进行试验结果评估时采用的评估指标是查对率。

反馈实验结果如表1所示。

表1 分类反馈实验结果

查对率(%)		计算机	体育	军事	政治	教育	经济	平均
训练	反馈前	93.33	73.33	70.6	75.7	93.73	87.67	82.39
文档	反馈后	94.33	73.5	72.3	77.7	97.43	88.33	83.93
	初始训练集	82.35	72.2	66	73.2	93.4	77.5	77.44
	反馈1	83.8	72.2	69	75.3	94.8	76	78.51
非训练	反馈2	87.6	73.2	68.5	74.4	93.7	80	79.56
文档	反馈3	87.2	73	68	72.2	95.3	82	79.62
	反馈4	92.2	73	68.5	75.2	92.3	80	80.2
	反馈5	94.2	73.5	68.5	74.3	94.5	77	80.33

通过反馈实验可以看出：

1)反馈学习对分类性能有明显的提高作用，实验中训练文档反馈前后系统分类精度分别从82.39%提高到83.93%，分类精度提高了1个多百分点；而在非训练文档的反馈实验中，经5次反馈学习后，系统的分类精度从77.44%提高到80.33%，分别提高了2～3个百分点，反馈学习效果明显，而从非训练文档的实验结果可以看出，随着反馈的进行，分类性能逐步提高。

2)反馈训练具有反馈学习数据少的优点。

3)在反馈过程中，在文档本身存在较强兼类特性的情况下，可能会出现人工分类标准与实际训练样本标准不一致的情况，造成反馈学习后的某个类别分类性能暂时波动。但这不影响反馈学习对整体分类性能的提高，同时也说明了学习样本质量对分类性能影响的重要性。

总之，从对简单向量距离文本分类算法的研究可以看出，反馈学习是文本分类的一种有效的学习方法。可以通过较小的反馈文档数量，实现较大的分类性能提高，具有反馈样本少、效果提高明显的优点。因此，该算法是进行文本分类研究与应用的有效方法。

参考文献

[1] Rocchio J J. Relevance feedback in information retrieval[A]. The SMART Retrieval System Experiments in Automatic Document Processing [C]. Ne Jersey: Prentice Hall, Inc., 1971: 313-23

[2] Ide E. Relevance feedback in an automatic document retrieval systems [R]. Ithaca, NY: Cornell University, 1969

[3] Salton G. The SMART Retrieval System[M]. Englewood Cliffs N J: Prentice Hall, Inc., 1971

[4] Cox I J, Miller M L, Omohundro S M. Pichunter: Bayesian relevance feedback for image retrival system[A]// Int'l Conf. on Pattern Recognition [C]. Vienna, Austeia, 1996: 361-369

[5] Ho L J. Combining the evidence of differrent relevancefeedback methods for information retrieval[J]. Information Processing & Management, 1998, 34(6): 681-691

(上接第235页)

割图像，如图1所示。

图1 待分割图像

图2 加标记点和插值后的初始轮廓线

图3 最终分割后的效果图

图1中间部分为猪的第二胸椎，首先在其边缘交互添加标记点，利用样条插值，生成初始轮廓线，如图2所示。

以初始轮廓线作为零水平集，建立窄带，M-S模型参数取为：

$\mu=0.01\times255^2$，$\Delta t=0.5$，$\lambda_1=1$，$\lambda_2=1$，$v=0$，$\varepsilon=1$，图3所示为最终的分割结果，可以看出分割结果很好地逼近了目标的边缘。下一张切片可以直接利用此片的分割轮廓作为零水平集，因此可以实现只需一次标记就能实现某一器官的自动分割。

结束语 由于猪切片图像组织器官的复杂性，利用传统的分割方法很难保证分割的速度和精度，通过人工添加标记点，在目标周围建立样条插值曲线，以此曲线作为零水平集，建立窄带，再利用M-S模型完成目标的分割，不仅避免了M-S模型的全局迭代计算，缩短了计算时间，而且分割结果也更加准确。

参考文献

[1] Mumford D, Shah J. Boundary Detection By Minimizing Functions// Proceedings of Conference Computer Vision and Pattern Recognition. San Francisco, America, 1985: 41-44

[2] Chan F T, Vese L A. A Level Set Algorithm for Minizing the Munford-Shan Functional in Image Processing, 2000, 4: 161-168

[3] Chan F T, Vese L A. Active Contours without Edges. IEEE Transactions on Image Processing, 2001, 10(2): 266-277

[4] 李俊. 基于曲线演化的图像分割方法及应用研究. 博士学位论文. 上海交通大学, 2001: 1-10, 15-37, 57-64

[5] 李国有. 基于广义模糊集及主动轮廓线模型的图像分割方法研究. 博士论文. 燕山大学, 2006: 69-79

[6] 关治. 数值计算方法. 北京: 清华大学出版社, 1990: 126-153