高技术通讯 2010年 第20卷 第7期:714~718

doi:10.3772/j.issn.1002-0470.2010.07.010

使用"分裂-合并"策略改进文本聚类集成算法的研究①

卢志茂②* 徐 森** 刘远超*** 顾国昌*

(*哈尔滨工程大学模式识别与自然计算研究室 哈尔滨 150001)
(**盐城工学院计算机工程系 盐城 224051)
(***哈尔滨工业大学智能技术与自然语言处理实验室 哈尔滨 150001)

摘 要 探讨了"分裂-合并"(DM)策略对文本聚类集成算法改进的效果。首先在聚类成员生成阶段运行使用DM策略的超球K均值(SKM)算法 r 次,每次生成较多的文本子簇,并根据子簇的相似性使用凝聚层次聚类方法合并这些子簇,得到 r 个聚类成员,随后在聚类集成阶段采用两个快速的谱聚类算法进行集成。在6组真实文本集上进行了实验,使用DM策略的两个聚类集成算法获得的平均标准化互信息(NMI)分别比改进前的算法提高了4.6和7.9个百分点,证明了DM策略可以有效提高文本聚类集成算法的聚类质量。

关键词 聚类集成,谱聚类,文本聚类,分裂-合并(DM),标准化互信息(NMI)

0 引 言

当前网络在迅猛发展,文本信息充斥整个网络,如何提高信息获取的效率已成为研究人员广泛关注的课题。聚类分析可以发现无结构文本集中的"潜在概念"(latent concept),并用这些概念来给出文本集的概要或者标签,因此,它可以有效地组织和搜索大规模文本集。此外,聚类分析在文本摘要、语义分析和导航搜索引擎的检索结果等方面也发挥了重要作用[1,2]。

由于文本数据的高维稀疏性,许多数据挖掘中的聚类算法并不能直接用于文本聚类。另外,在信息检索领域中,文本通常是数以亿计的,因此,对聚类算法的计算复杂度也有很高的要求[1]。

在所有的聚类算法中,K均值(K-means)算法由于简单高效而成为使用最为广泛的算法,它在向量空间模型中的扩展——超球K均值(spherical K-means,SKM)算法[2]已被证明是非常有效的文本聚类算法。然而SKM算法存在以下两个主要缺点:(1)它是基于梯度的方法,而其目标函数关于概念向量(concept vector)在 $\mathbf{R}^d$ 空间中并不是严格凹函数(d 是向量空间的维数),因此,不同的初值会收敛到不同的局部极值,即算法极不稳定;(2)它显式地假设最终的簇具有超球形状,而在实际应用中该假设往往并不成立。

近年来,研究表明聚类集成(cluster ensemble)可以有效克服K均值算法及其扩展算法的缺点,提高其精度及稳定性[3,4]。Strehl和Ghosh[3]提出了基于簇的相似度划分算法(cluster-based similarity partitioning algorithm,CSPA)、超图划分算法(hypergraph partitioning algorithm,HGPA)和元聚类算法(meta-clustering algorithm,MCLA)。本文作者在文献[4]中提出了基于相似度矩阵的谱算法(similarity matrix-based spectral algorithm,SMSA)和基于超边相似度矩阵的谱算法(hyperedges' similarity matrix-based spectral algorithm,HSMSA)。在多组真实文本集上进行的实验,获得了这两个谱算法都比CSPA、HGPA和MCLA更加优越的结果。

目前,对于SKM算法的第2个缺点还鲜有研究,这是因为文本数据存在于高维向量空间,会产生Richard Bellman所称的"维数灾难"(curse of dimensionality)现象。与此相关的是空空间(empty space)[5]现象,即高维空间本质上是稀疏空间,这使得传统的统计建模方法很难用于高维文本数据。另外,由于我们对高维空间缺乏可视化观察的能力,并且余弦相似度函数与欧氏距离函数不同,不满足度量(metric)的第三个性质(即三角不等式),因而我们就更难把握文本数据的真实分布情况[6]。

本文在文献[4]研究的基础上进一步探讨如何

① 863计划(2007AA01Z172),国家自然科学基金(60975042,60603092)和高等学校博士学科点专项科研基金(20070217043)资助项目。
② 男,1972年生,工学博士,教授,博士生导师;研究方向:模式识别,自然语言处理,数据挖掘;联系人,E-mail:lzm@hrbeu.edu.cn
(收稿日期:2009-04-01)

进一步提高文本聚类集成算法的聚类质量。为了提高 SKM 算法发现不规则形状文本簇的能力,产生精度更高的聚类成员,本文探讨了 CHEMALEON 算法[7]的"分裂-合并"(divide and merge,DM)策略对聚类集成算法的影响。

1 聚类集成相关研究

聚类集成一般可以分为两步:第一步把数据集作为输入,输出多个聚类结果,这一步称为聚类成员的生成(generation)阶段;第二步把不同的聚类结果作为输入,对它们进行组合,输出最终的结果,这一步称为组合(combination)/集成(ensemble)/融合(fusion)阶段。

在聚类成员生成阶段,可以通过选择不同的算法[3]、对一个算法选择不同的初值[3,4]和选择不同的对象子集等来产生聚类成员[3]。其中最常用的方法是采用 K 均值算法,随机选择不同的初始质心,运行 r 次,产生所需要的 r 个聚类成员。这种方法的最大优点是算法的计算复杂度低,实现简单、方便等,缺点是难以发现边界以及难以区分的簇和非球形簇。

Kuncheva 等[8]通过实验发现聚类成员间较大的差异度将能提高聚类集成的质量。聚类成员之间的差异度可以通过 Rand Index,Jaccard Index,Adjusted Rand Index,Mutual Information 等来衡量。Hadjitodorov 等[9]研究发现,聚类成员间的差异度与聚类集成质量之间的关系是非单调的,当差异度过大时,聚类集成的质量反而下降。并指出,适中的差异度将能得到较好的聚类集成结果。罗会兰等[10]通过多组实验发现,随着平均成员准确度的增加,集成性能相对于聚类成员的提高幅度也在增加。

与分类问题不同的是,聚类学习中的样本是无标签的,因此,由不同的聚类算法得到的划分结果存在一个类别/簇标签对应问题。另外,不同的聚类成员可能产生不同的簇个数,这使得簇标签对应问题更加困难。

文献中解决聚类集成问题的主要方法是由 Strehl 和 Ghosh[3]提出的 CSPA 算法,该算法调用了高效的图划分算法 METIS[11]。CSPA 对簇的形状不做强的假设,但是易受 METIS 算法参数的影响,例如不平衡因子(imbalance factor)ubfactor① 的选取将会对簇的大小产生很大的影响。

本文作者在文献[4]中将谱聚类算法引入到文本聚类集成问题中,使得算法具备谱聚类算法的以下几个主要优点:对簇的形状不做强的假设;实现简单,只需解决特征值分解问题;算法不存在局部最优解[12]。为了降低谱聚类算法的计算复杂度,文中分别通过"代数变换"和"间接求解"的方法避免了 n 阶方阵的特征值分解问题,设计并实现了两个快速的算法 SMSA 和 HSMSA,它们的计算复杂度都仅为 n 的一次多项式。在 TREC 和 Reuters 文本集上的实验结果证实了两个谱聚类算法的有效性,它们不但比 CSPA 更能发现聚类成员之间的一致性,而且聚类结果也与真实文本类别标签更加匹配。

2 使用 DM 策略的聚类集成算法

前文已经指出,聚类成员的精度对最终的集成性能有很大影响。SKM 算法显式地假设最终的簇具有超球形状,而这个假设在实际应用中可能并不成立。为了进一步提高聚类集成算法的性能,本文在聚类成员生成阶段引入 CHEMALEON 算法的关键思想——DM 策略提高基聚类算法的精度。

2.1 CHEMALEON 算法

CHEMALEON 算法[7]首先利用图划分算法 METIS 将 n 个对象的 k-近邻图划分为 m 个相对较小的子簇,然后再利用凝聚层次聚类方法,通过不断合并这些子簇来发现真正的簇。为确定哪两个子簇最相似,CHEMALEON 算法不但考虑子簇间的相对连接度(relative inter-connectivity),而且考虑子簇间的相对接近度(relative closeness),特别是簇本身的内部特征。研究表明,CHEMALEON 算法在发现具有高质量任意形状聚类方面能力比较强;算法的计算复杂度为 $O(nm + n\log n + m^2\log m)$[7]。

然而,CHEMALEON 算法的聚类结果易受 k-近邻图中的 k 值选取的影响。k 近邻图中每个点与其最近的 k 个近邻点相连接,该方法可以有效发现不同尺度的数据结构,但是 k 值的选择比较困难。当数据有不同密度时会出现问题,如果 k 值选得过大,会把低密度的簇与其它簇相连接,过小的 k 值,会把高密度的簇分成多个子簇。

2.2 使用 DM 策略的文本聚类集成算法

事实上,DM 策略与人类思维中"局部构成整体"的过程是相吻合的。例如,1999 年美国著名的

① 参考了 METIS 算法实现软件包中的算法使用手册,算法实现软件包根据以下网址下载:http://glaros.dtc.umn.edu/gkhome/views/metis。

科学杂志《Nature》刊登了 Lee 和 Seung[13]两位科学家的突出成果——非负矩阵分解(non-negative matrix factorization,NMF)。Lee 和 Seung 将原始的人脸图像分解为多个小图像的非负加权组合,而每个小图像恰好表示了诸如“鼻子”、“眼睛”、“嘴巴”等人脸局部概念特征。他们认为,与人类识别事物的过程相似,NMF 是一种优化的机制,近似于大脑分析和存储人脸数据的过程。

对于文本数据,使用 DM 策略,我们可以首先将大量无结构的文本数据聚为许多小的较纯的“子概念”,再根据这些子概念之间的相似度合并为所需的 k 个概念。这种方法首先产生细粒度的聚类结果,随后根据子簇间的相似性进行合并得到最终所需的 k 个文本簇,因此能够发现潜在的有意义的簇,而这些簇可能与真实的文本分布更相符。例如,对于 XOR 问题,若直接将其聚为两个簇显然得不到数据的真实分布,但如果首先将其聚为 4 个纯的子簇,随后根据子簇的相似性进行合并,那么我们就可以正确揭示其内部结构。

考虑到 k-近邻图的构造困难,而 SKM 算法简单快速,所以本文并不是用图划分算法 METIS 来获得子簇,而是首先用 SKM 算法来产生初始子簇,随后用凝聚层次聚类算法来合并产生最终的 k 个簇。然而,该方法会受 SKM 算法初值选取的影响,因此我们使用聚类集成技术提高结果的精度和稳定性。我们做的多组实验表明[4],与 CSPA、HGPA 和 MCLA 相比,SMSA 和 HSMSA 获得的聚类结果与真实类别标签更匹配。因此本文在聚类集成阶段使用 SMSA 和 HSMSA 算法来探究 DM 策略对聚类集成算法性能的影响。

本文首先采用 SKM 算法获得多个较纯的文本子簇,随后使用凝聚层次聚类算法获得 k 个文本簇。考虑到在以往的聚类实验中 Ward 算法获得的结果较其它层次聚类算法要好[1],所以本文在合并阶段采用 Ward 算法获得 k 个簇。在集成成员生成阶段使用 DM 策略得到的聚类集成算法主要步骤如下:

输入: $d \times n$ 的词-文本共现矩阵 $\boldsymbol{A}$,初始簇个数 $k_0 = \lfloor n^{1/2} \rfloor$,真实簇个数 k。

步骤 1 聚类成员生成阶段:

(a) 运行 SKM 算法 r 次,每次随机选取初值,产生 k_0 个子簇,每个文本根据词频-逆文档频率(term frequency-inverse document frequency,TF-IDF)加权,并对文本向量进行归一化处理,使其欧氏范数为 1;

(b) 使用 Ward 算法将每次聚类得到的 k_0 个子簇合并为 k 个文本簇,获得 r 个聚类成员。

步骤 2 聚类成员集成阶段:使用算法 SMSA 和 HSMSA 进行集成。

输出: k 个文本簇。

使用 DM 策略的聚类集成算法第 1(a)步需要运行 SKM 算法 r 次①,时间复杂度为 $O(rk_0I_1dn) = O(n^{3/2})$,其中 I_1 表示 SKM 算法所需的迭代次数,通常 r、$I_1 << n$。第 1(b)步通过使用堆来存放簇之间的相似度可以将计算复杂度降到 $O(k_0^2 \log k_0) = O(n \log n)$。算法第 2 步的计算复杂度为 n 的一次多项式[4]。根据以上分析,使用 DM 策略后的聚类集成算法总的时间复杂度为 $O(n^{3/2})$。

3 实验设计与结果分析

3.1 实验数据集

本文实验使用与文献[4]中相同的 6 个数据集,表 1 给出了这些数据集的具体描述。对于每个数据集,使用停用词表移去停用词,并且去掉出现在少于两个文本中的词。

表 1 实验数据集描述

数据集	文本个数	特征个数	类别个数
hitech	2301	13170	6
reviews	4069	23220	5
tr31	927	10128	7
tr41	878	7454	10
re0	1504	2886	13
re1	1657	3758	25

数据集 hitech 和 reviews 取自 San Jose Mercury 报纸,它们是 TREC[14]文本集的一部分(TIPSTER Vol. 3),hitech 包含了关于计算机、电子、健康、医疗、研究和科技方面的文章,而 reviews 包含了关于食物、电影、音乐、广播和饭店方面的文章,所有文本的类别标签唯一。数据集 tr31 和 tr41 取自 TREC-6[14]和 TREC-7[14]文本集,这些数据的类别对应于某个特殊类别的查询。数据集 re0 和 re1 取自 Reuters-21578 文本分类测试集[15],其标签被分为两个部分,对于每个数据集,仅选择有唯一类别标签的文本。

① 聚类集成算法的性能与 r 的选取有很大关系,文献[10]通过多组模拟实验研究了集体大小、差异性度量与 CSPA 算法集成准确度的相关性,指出最好生成大小为 15~20 的集体。本文研究的重点是 DM 策略而不是聚类集体的大小对聚类集成算法的影响,因此实验中将 r 设置为一个固定值 5。

3.2 评价指标

因为文本的类别标签已知,我们采用源自信息论的标准化互信息(normalized mutual information, NMI)[3]来量化聚类结果和已知类别标签的匹配程度。与纯度和熵等准则相比,NMI 对 k 值的选取没有偏好,因此成为近年来比较流行的评价指标。NMI 值越大,两个类别标签越匹配,当两个类标签一一对应时,NMI 值达到最大值 1。

设 X 和 Y 分别为聚类成员 $\mathbf{P}^{(a)}$ 和 $\mathbf{P}^{(b)}$ 表示的随机变量,其中 $\mathbf{P}^{(a)}$ 和 $\mathbf{P}^{(b)}$ 分别有 k^a 和 k^b 个簇。设 $I(X;Y)$ 表示 X 和 Y 之间的互信息,$H(X)$ 为 X 的熵,可以发现 $I(X;Y)$ 是一个测度,它没有上界。因为 $H(X) = I(X;X)$,所以 Strehl 和 Ghosh[3]使用几何均值来标准化互信息:

$$\mathrm{NMI} = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

设 n_h^a 为 $\mathbf{P}^{(a)}$ 中的簇 $\mathbf{C}_h$ 包含的对象个数,n_l^b 为 $\mathbf{P}^{(b)}$ 中的簇 $\mathbf{C}_l$ 包含的对象个数,$n_{h,l}$ 表示同时在 $\mathbf{C}_h$ 和 $\mathbf{C}_l$ 中的对象个数,则 $\mathbf{P}^{(a)}$ 和 $\mathbf{P}^{(b)}$ 之间的 NMI 值为:

$$\mathrm{NMI}(\mathbf{P}^{(a)},\mathbf{P}^{(b)}) = \frac{\sum_{h=1}^{k^a}\sum_{l=1}^{k^b} n_{h,l}\log\left(\frac{n\cdot n_{h,l}}{n_h^a n_l^b}\right)}{\sqrt{\left(\sum_{h=1}^{k^a} n_h^a\log\frac{n_h^a}{n}\right)\left(\sum_{l=1}^{k^b} n_l^b\log\frac{n_l^b}{n}\right)}}$$

3.3 实验结果与分析

我们把在集成成员生成阶段使用 DM 策略的聚类集成算法分别记为 SMSADM 和 HSMSADM。因为文本数据的真实结构未知,所以本节将通过实验比较使用和未使用 DM 策略的集成算法获得的 NMI 值。如果使用该策略得到的结果较未使用该策略得到的结果有较大的提升,说明 DM 策略可以有效提高文本聚类集成算法的聚类质量;因为在聚类成员生成阶段的合并阶段使用了 Ward 算法,它本身也给最终生成的簇强加了某种结构,所以使用 DM 策略后得到的 NMI 值也有可能会降低。

我们将 4 个聚类集成算法 SMSA、HSMSA、SMSADM 和 HSMSADM 在 6 个数据集上进行了实验,获得的 NMI 值如图 1 所示。由图 1 可以看出,SMSADM 除了在 re0 数据集上的 NMI 值低于 SMSA 外,在所有其它数据集上,都能获得比 SMSA 更高的 NMI 值。HSMSADM 算法在所有数据集上都获得了比 HSMSA 更高的 NMI 值。

为了更好地比较 DM 策略对 SMSA 算法和 HSMSA 算法聚类质量的影响,我们将这两个算法使用 DM 策略后获得的 NMI 值分别除以未使用 DM 策略获得的 NMI 值,其比值如表 2 所示。根据表 2,SMSADM 算法在 6 个数据集上获得的 NMI 值分别比 SMSA 算法提升了 5.0%、2.4%、15.4%、2.0%、-3.2% 和 4.6%;HSMSADM 算法获得的 NMI 值相对于 HSMSA 算法在 hitech、reviews 和 tr31 上提升很明显,分别提高 14.9%、8.1% 和 16.1%;在其它 3 个数据集上的性能也略有提高,分别提高 3.5%、3.8% 和 3.3%。

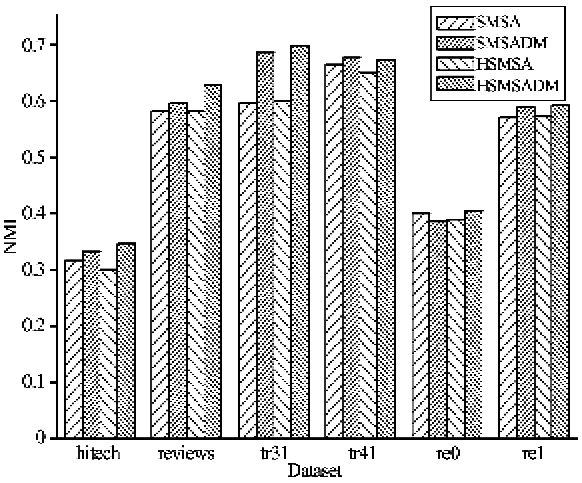

图 1 不同聚类集成算法获得的 NMI 值

表 2 SMSA 和 HSMSA 采用 DM 策略前后 NMI 值的比值

dataset	SMSA	HSMSA
hitech	1.050	1.149
reviews	1.024	1.081
tr31	1.154	1.161
tr41	1.020	1.035
re0	0.968	1.038
re1	1.031	1.033

通过计算,SMSADM 算法在 6 个数据集上获得的平均 NMI 值比 SMSA 算法提高了(0.546 - 0.522)/0.522≈4.6%;HSMSADM 算法在 6 个数据集上获得的平均 NMI 值比 HSMSA 算法提高了(0.558 - 0.517)/0.517≈7.9%。

上面的结果表明,DM 策略可以有效提高两个集成算法 SMSA 和 HSMSA 的聚类质量。当然,在实验中也出现了一些例外,例如,SMSA 算法在 re0 数据集上的 NMI 值出现了下降的情况。这是因为在聚类成员生成阶段使用了 Ward 算法来合并多个子簇,而 Ward 算法本身给最终的簇强加了某种结构,因此合并后得到的簇与真实的簇分布结构有可能出现不相符的情况。总的来看,两个算法使用 DM 策

略后在其它几组数据集上都一致获得了比较好的聚类效果。

4 结 论

本文在文献[4]的研究基础上探讨了CHEMALEON算法的关键思想——“分裂-合并”(DM)策略对聚类集成算法聚类效果的影响。在多组数据集上进行了实验,结果表明,DM策略可以有效提升聚类集成算法的聚类质量。

本文将初始子簇的个数设置为 $k_0 = \lfloor n^{1/2} \rfloor$,使得算法的计算复杂度从 $O(n)$ 提高到 $O(n^{3/2})$,当扩展到大规模信息检索领域时,使用DM策略显然会需要耗费大量的计算时间。因此,如何在适度增加算法复杂度的前提下,大幅提高算法的聚类质量值得进一步研究。

参考文献

[1] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining. MA, USA: Addison-Wesley, 2005. 487-647

[2] Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 2001, 42: 143-175

[3] Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining partitionings. *Journal of Machine Learning Research*, 2002, 3: 583-617

[4] 徐森, 卢志茂, 顾国昌. 解决文本聚类集成问题的两个谱算法. 自动化学报, 2009, 35(7): 997-1002

[5] Scott D W, Thompson J R. Probability density estimation in higher dimensions. In: Proceedings of the 15th Symposium on the Interface, Amsterdam, Holland, 1983. 173-179

[6] Duda R, Hart P, Stork D. Pattern Classification. Second Edition. New York, USA: John Wiley & Sons, 2001. 517-599

[7] Karypis G, Han E H, Kumar V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *IEEE Computer*, 1999, 2(8): 68-75

[8] Kuncheva L I, Hadjitodorov S T. Using diversity in cluster ensembles. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Hague, Netherlands, 2004. 1214-1219

[9] Hadjitodorov S T, Kuncheva L I, Todorova L P. Moderate diversity for better cluster ensembles. *Information Fusion*, 2006, 7: 264-275

[10] 罗会兰, 孔繁胜, 李一啸. 聚类集成中的差异性度量研究. 计算机学报, 2007, 30(8): 1315-1323

[11] Karypis G, Kumar V. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 1998, 20(1): 359-392

[12] Luxburg U V. A tutorial on spectral clustering. *Statistics and Computing*, 2007, 17(4): 395-416

[13] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, 401: 788-791

[14] TREC. Text REtrieval Conference. http://trec.nist.gov, 2007

[15] Lewis D D. Reuters-21578 text categorization test collection distribution 1.0. http://www.research.att.com/~lewis, 2007

Improvement of document cluster ensemble algorithms using divide and merge strategy

Lu Zhimao*, Xu Sen**, Liu Yuanchao***, Gu Guochang*

(* Pattern Recognition and Natural Compution Laboratory, Harbin Engineering University, Harbin 150001)

(** Department of Computer Engineering, Yancheng Institute of Technology, Yancheng 224051)

(*** Intelligent Technology & Natural Language Processing Laboratory, Harbin Institute of Technology, Harbin 150001)

Abstract

The influence of the divide and merge (DM) strategy on document cluster ensemble algorithms was explored. Firstly, the spherical K-means (SKM) algorithm utilizing the DM strategy was performed for r times in the ensemble member generation phase, and each time more document sub-clusters were obtained and the agglomerative hierarchical method was used to merge these sub-clusters according to their similarity to attain r ensemble members. Then, two fast spectral clustering algorithms were performed to ensemble the r clusterings. The experiments on six real-world document sets showed that the DM strategy increased the normalized mutual information (NMI) of the two cluster ensemble algorithms by 4.6 and 7.9 percentage in average, respectively. These results prove that DM strategy can effectively improve the performance of document cluster ensemble algorithms.

Key words: cluster ensemble, spectral clustering, document clustering, divide and merge (DM), normalized mutual information (NMI)