基于组合分类器的相关反馈算法研究

陆晓艳 周 良 丁秋林
(南京航空航天大学信息科学与技术学院 南京 210016)

摘 要 基于内容的矢量图形检索系统可以通过使用相关反馈算法获得较好的检索性能。提出了一种新的基于组合分类器的相关反馈算法,该算法以每一个正负反馈样本作为唯一的训练样本,形成各个独立的最近邻分类器,融合各个分类器的预估结果,计算库中每个图形的相关分数,并引入贝叶斯查询点移动技术来优化相关分数。实验结果表明,该算法在进一步提高矢量图形检索系统查准率的同时,还能保证系统的查全率。

关键词 组合分类器,贝叶斯查询点移动,相关反馈,矢量图形检索
中图法分类号 TP391 **文献标识码** A

Research on Relevance Feedback Algorithm Based on Combining Classifiers

LU Xiao-yan ZHOU Liang DING Qiu-lin
(College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract High retrieval performances in content-based vector graphics retrieval system can be attained by adopting relevance feedback algorithms. A new relevance feedback approach based on combining classifiers was proposed, which combines the expected results from the independent nearest neighbor classifiers with only one training sample formed by each positive or negative feedback sample, computes the relevance score of every vector graphics and optimizes the relevance score by introducing the technique called "Bayesian Query Shifting". The results of the experiment show that the algorithm not only can further improve the precision of the vector graphics retrieval system but also can ensure the recall of the system.

Keywords Combining classifiers, Bayesian query shifting, Relevance feedback, Vector graphics retrieval

目前,在基于内容的图形检索技术中所抽取的图形特征基本上是图形的底层几何特征,它们与图形的实际语义是脱离的,但底层几何特征目前尚无能力辨别出图形中所包含的所有图元信息。因此,无论采用何种特征,使用何种距离测度,最终决定两幅图形是否相似,还是取决于实际用户。另外,由于侧重点的不同,不同的用户判断图形的相似性也存在不同的标准。为此需要研究如何使系统自动适应这种特定的需求,从而实现更好的检索性能。相关反馈技术则从机器学习的角度出发,把检索过程看作一个人机协同工作的过程,利用人对图形语义的认知和理解来弥补计算机在这方面的不足,是提高系统检索性能的一种有效方法。因此,相关反馈技术的研究逐渐成为一个比较活跃的研究方向,适用于不同领域的相关反馈技术也不断涌现。文献[1]已经对这些相关反馈技术做了较全面的综述。

早期的相关反馈工作是采用基于距离的度量方法来提高系统的检索性能。在这种检索模型下,相关反馈的主要策略有查询向量转移以及调整特征权重。后来系统中引入概率框架来描述检索问题。文献[2]中提出了一个基于贝叶斯决策论的相关反馈技术,在库中先预估所有图像的相关概率,并根据相关性将图形返回给用户。近年来,相关反馈已经被归结为不同类型的监督学习问题[3,4]、二类(相关或者不相关)或者是(1+X)类的分类问题。(1+X)分类问题也可以称之为有偏学习问题,认为矢量图形数据库中的类别的个数是不知道的,而用户只对其中一类感兴趣。根据这些学习问题的特点,支持向量机(Support Vector Machines, SVM)和判别分析(Discriminant Analysis, DA)这两种机器学习技术已经被广泛地使用。实验结果也表明这两种方法相对于前面提到的相关反馈技术可以获得更好的检索结果。

本文提出了一种新的基于组合分类器(Combining Classifiers, CC)的相关反馈算法,以每一个正负反馈样本作为唯一的训练样本形成各个独立的最近邻分类器,融合各个分类器的预估结果,计算库中每个图形的相关分数,并通过引入查询点移动技术来优化相关分数。实验结果证明了本文提出的算法的有效性和系统检索性能的提高。

1 组合分类器

组合分类器是一个众所周知的根据不同的线索进行信息融合的方法[5]。通过某种组合技术,将多个分类器的预测结果进行融合,从而产生一个新的分类器,并用新分类器对样本进行分类。如果融合得当,组合分类器的性能比任何单个分

到稿日期:2009-06-26 返修日期:2009-09-04 本文受国防基础科研重大专项基金项目资助。
陆晓艳(1985—),女,硕士生,主要研究方向为人机交互技术,E-mail:lucylu85721@yahoo.com.cn;周 良(1966—),男,副教授,主要研究方向为人机交互技术、信息系统与信息安全、知识工程;丁秋林(1935—),男,教授,主要研究方向为CIM, DSS, MIS。

类器都优越。

如果把组合分类器看作一个完整系统，则它由系统输入、单分类器设计、组合结构和融合规则4部分组成[6]。系统输入是指输入的表示方式及单个分类器输入的确定，在一般化的系统中，主要是针对单分类器的输入。因为一般情况下，在接受一个输入时，每个单分类器都要得到它的独立结果。单分类器设计是指各个分类器学习算法的构造和相关参数的定义。组合结构是各单分类器的组合方式，它有并联和串联两种类型。对一个新样本，并联总是把所有单分类器的结果都并行融合起来，这样几乎总是能得到此样本属于某类别的相对概率，然后输出可能性最高的那个类别。而串联方式是把一系列分类器前后相接，后面的分类器注意力集中到它前面分类器所发生的预测错误上，通过训练使之成为一个有效的整体。融合规则是各单分类器输出信息的组合方式，它是整个系统的核心。一旦上面4个部分被确定，那么一个完整的组合分类器系统也就确定了。

Kittler在文献[5]中系统地研究了组合分类器融合，给出了组合分类器融合的一个理论框架，并在该框架基础上得到了融合的5个规则：积规则、和规则、最小规则、最大规则以及投票规则。其中约束条件最为苛刻的和规则表现出了最好的分类性能，因此本文采用求和规则来融合各个分类器。测试样本 I 属于 j 类别的求和规则如下：

$$j=\mathrm{argmax}_m[(1-R)P(\omega_m)+\sum_{i=1}^{R}P_I(\omega_m\mid x_i)] \tag{1}$$

式中，R 是单分类器的个数，$P(\omega_m)$ 为类别 m 的先验概率，$P_I(\omega_m\mid x_i)$ 是在 i 分类器中测试样本 I 属于类别 m 的概率。

2 基于组合分类器的相关反馈算法

将组合分类器引入相关反馈，可以充分利用每一个正负反馈样本所提供的信息[7]。如果把组合分类器看作是一个完整的系统，它的系统输入则是库中的每一个矢量图形。在基于内容的矢量图形检索的应用环境中，很难提供一个具有高度推广能力的分类器，以便于判别不同用户对矢量图形相似概念的理解。而最近邻算法在这种应用环境中是非常有效的，因此本文采用最近邻算法构造单分类器。若使用和规则来融合各个单分类器，系统的组合结构则是并联的组合方式。

当正反馈样本较少时，单纯依靠组合分类器计算得到的相关分数并不是很可靠。为此，本文引入查询点移动技术实现相关分数优化，以提高矢量图形检索的准确率。本文提出的基于组合分类器的相关反馈检索流程如图1所示。

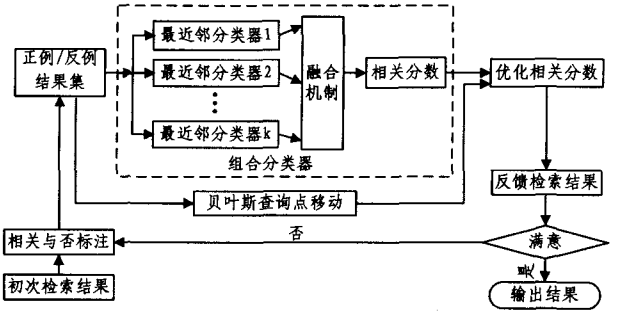

图1 相关反馈检索流程

2.1 相关分数计算

将用户反馈的被标识为相关或者是不相关的每一个图形作为各个最近邻分类器的唯一训练样本，矢量图形数据库中的每一个图形作为测试样本。在以训练样本为正反馈样本的分类器中，离训练样本越近的测试样本的相关概率密度越大。反之，在训练样本是负反馈样本的分类器中，离训练样本越近的测试样本的相关概率密度越小。

因此，假定对于一个给定的被标识为相关的图形 q_+ 所形成的最近邻分类器中，每一个测试样本 I 是相关的概率密度为

$$P_I(r\mid q_+)=\exp(-\parallel I-q_+\parallel) \tag{2}$$

式中，r 表示测试样本 I 是相关的，并通过使用指数函数进行归一化。从式(2)中可以看出，若测试样本 I 与正反馈样本 q_+ 的距离越短，则 I 的相关的概率密度就越大。

同样，对于一个被标识为不相关的图形 q_- 所形成的最近邻分类器中，每一个测试样本 I 是不相关的概率密度为

$$P_I(n\mid q_-)=\exp(-\parallel I-q_-\parallel) \tag{3}$$

$$P_I(r\mid q_-)=1-P_I(n\mid q_-) \tag{4}$$

式中，n 表示测试样本 I 是不相关的。从该式中可以看出，若测试样本 I 与负反馈 q_- 距离越短，不相关的概率密度就越大，相应的相关的概率密度就越小。可以通过式(4)计算以 q_- 为训练样本的最近邻分类器中 I 相关的概率密度。

根据式(2)－式(4)可以计算库中的每个图形在各个最近邻分类器中的相关程度，通过使用求和规则来融合各个独立的最近邻分类器，计算库中每一个图形的相关分数。由式(1)所提供的求和规则是用来判断测试样本的所属类别，它不但要计算每个测试样本在各个分类器中的分类结果，还要计算各个类别的先验概率 $P(\omega_m)$。在相关反馈中求和规则并不是用来判断测试样本所属的类别，而是用来计算各个测试样本的相关分数，将相关分数排在前 k 的图形返回给用户。而先验概率 $P(r)$ 对每个测试样都是一样的，对相关分数的排名并不产生影响，因此相关分数的计算公式中不需要考虑先验概率。

$$r\_score_{CC}(I)=P_I(r\mid(Q_+,Q_-))\doteq\frac{\sum_{q_+\in Q_+}P_I(r\mid q_+)+\sum_{q_-\in Q_-}P_I(r\mid q_-)}{k} \tag{5}$$

根据式(5)可以计算库中每个图形的相关分数，其中 k 表示每次反馈图形的数量，Q_+ 是正反馈集，Q_- 是负反馈集。

2.2 相关分数优化

单纯依靠组合分类器计算的相关分数还是存在一定问题的。每一次交互中，用于用户反馈的样本太少，而且很可能只有为数不多的图形是相关的，此时同时远离正反馈样本和负反馈样本的图形就会被赋予很高的相关分数。在这种情况下，最后返回给用户的图形很可能多数是不相关的，这样式(5)并不是一个很好的计算每个图形相关分数的方法。因此本文提供了一个与测试样本到相关图形区域的距离相关的项来调整相关分数。

选择原始的查询点作为后续的每一次反馈的查询点是不合理的，因为原始的查询点的邻居也许只包含为数不多的相关图形。因此，为了使查询点向一个发现相关图形可能性更高的区域移动，本文引入了贝叶斯查询点移动技术(Bayesian Query Shifting，BQS)。实验结果也表明它的检索性能比Rocchio向量转移算法更好[8]。新的查询向量的计算公式如下：

$$Q_{BQS}=m_R+\frac{\sigma}{\parallel m_R-m_N\parallel}\left(1-\frac{k_R-k_N}{\max(k_R,k_N)}\right)(m_R-m_N) \tag{6}$$

式中，m_R，m_N 分别是正反馈样本和负反馈样本的均值向量，σ 是对应的标准方差，k_R，k_N 分别代表正反馈样本和负反馈样本的数目。从式(6)可以看出，当正反馈样本数量很少时，查询点就会大幅度地向相关性可能性更高的区域移动。

$$r\_score_{BQS}(I)=\frac{1-\exp(1-\| I-Q_{BQS}\| /\max_I \| I-Q_{BQS}\|)}{1-e} \tag{7}$$

式(7)提供了一个基于 BQS 的相关反馈分数，通过使用高斯模型使基于 BQS 的相关分数归一化，也因此可以与 2.1 节提供的相关分数相结合。结合后的相关分数的计算公式如下：

$$r\_score(I)_{opt}=(\frac{n/k}{1+n/k})r\_score_{BQS}(I)+(\frac{1}{1+n/k})r\_score_{CC}(I) \tag{8}$$

式中，k，n 分别代表每次正反馈样本的数量和负反馈样本的数量。当正反馈的个数为 0 时，各自的权重为 1/2；随着正反馈数量的增加，BQS 的权重也会随着减少；当没有负反馈时，BQS 的权重将为 0。系统根据式(8)计算的相关分数将排在前 k 的图形返回给用户。

3 实验结果及性能评价

为了测试本文所提出的一种新的基于组合分类器的相关反馈算法的检索性能，本文以 Visual C#. net 为平台构建了一个原型系统，分别用 SVM、文献[9]中提供的基于最近邻的相关反馈算法(NN&BQS)和本文提出的基于组合分类器的相关反馈算法(CC&BQS)进行测试。实验中采用的图形数据库是某 CAD 软件自带标准件图库，库中有 1900 多张标准件工程图，包括螺母、螺栓、柳钉以及轴承等 40 几个类别，每一小类的图形个数均超过 40。每次反馈系统将相关分数排在前 20 的图形反馈给用户，用户只标记他认为相关的图形，其它的图形则默认都是不相关的。如果用户对反馈结果满意，则输出反馈结果，结束反馈，否则进入下一轮反馈。

本文使用查准率(precision)和查全率(recall)来评价各个算法的检索性能，实验结果如图 2、图 3 所示。

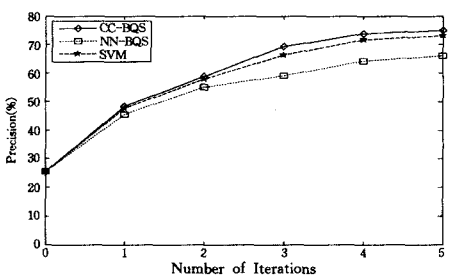

图 2 查准率比较

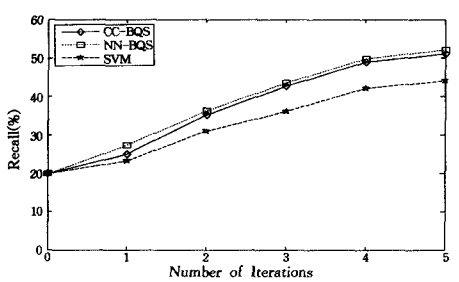

图 3 查全率比较

通过实验结果可以看出，基于组合分类器的相关反馈算法与基于最近邻的相关反馈算法相比，在查全率方面没有明显的优势，但是查准率有了一定程度的提高。与传统的 SVM 算法相比，不管是基于组合分类器的相关反馈算法还是基于最近邻的相关反馈算法，查全率都比 SVM 高很多，基于最近邻的相关反馈算法的查准率与 SVM 相比还是有所欠缺，不过本文提出的基于组合分类器的相关反馈算法的查准率则稍微高于 SVM。这说明本文提出的算法能够充分利用用户的反馈信息有效提高系统的检索性能。

结束语 相关反馈对于基于内容的矢量图形检索具有重要意义。如何减少反馈的次数，提高系统的检索性能，是相关反馈亟待解决的问题。本文提出了一种新的基于组合分类器的相关反馈算法，以每一个正负反馈样本作为唯一的训练样本形成各个独立的最近邻分类器，融合各个分类器的预估结果，并引入贝叶斯查询点移动技术来优化相关分数。实验结果表明，本文提出的基于组合分类器的相关反馈算法能够有效地提高系统的查准率，同时还能保证查全率，并在有限次的反馈后，使检索结果符合用户的主观要求。

参 考 文 献

[1] Zhou Xiang Sean, Huang T S. Relevance Feedback for Image Retrieval: A Comprehensive Review[J]. ACM Multimedia Systems, 2003, 8(6): 536-544

[2] Cox I J, Miller M L, Minka T P, et al. The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments[J]. IEEE Transactions on Image Processing 9, 2000, 9(1): 20-37

[3] Zhang Lei, Lin Fuzong, Zhang Bo. Support Vector Machine Learning for Image Retrieval[C]// Proc. IEEE Int'l Conf. Image Processing. 2001: 721-724

[4] Tao Dacheng, Tang Xiaoou, Li Xuelong, et al. Direct Kernel Biased Discriminant Analysis: A New Content-based Image Retrieval Relevance Feedback Algorithm[J]. IEEE Transactions on Multimedia, 2006, 8(4): 716-727

[5] Kittler J. On Combining Classifiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 226-239

[6] 蒋林波，蔡立军，易叶青. 一个新的多分类器组合模型[J]. 计算机工程与应用，2008，44(17)：131-135

[7] Deselaers T, Paredes R, Vidal E. Learning Weighted Distances for Relevance Feedback in Image Retrieval[C]// Pattern Recognition, 2008 19th International Conference on. Tampa, FL, USA, 2008

[8] GGiacinto E, Roli F. Query Shifting Based on Bayesian Decision Theory for Content-based Image Retrieval[C]// SSPR&SPR. Ontario, Canada, 2002

[9] Giacinto G. A Nearest-neighbor Approach to Relevance Feedback in Content Based Image Retrieval[C]// Proceedings of the 6th ACM International Conference on Image and Video Retrieval. 2007: 456-463