基于内容的医学图像检索中的相关反馈技术

沈 晔[1,2] 夏顺仁[1*] 李敏丹[2]

[1](浙江大学生物医学工程与仪器科学学院 教育部生物医学工程重点实验室,杭州 310027)

[2](中国计量学院信号与信息处理系,杭州 310018)

摘 要: 建立一个高效、准确的医学图像检索系统是目前具有挑战性的任务。由于相关反馈(RF)技术有效地解决了"语义鸿沟",成为基于内容的医学图像检索系统中提高检索性能的关键技术。文中根据 RF 算法采用的检索模型,从基于距离度量的模型、基于概率统计分类模型和基于机器学习模型三个方面,对有代表性的算法进行了分析与评价,并重点分析了基于机器学习的 RF 算法。最后对医学图像检索中 RF 技术的发展进行了展望。

关键词: 基于内容的医学图像检索; 语义鸿沟; 相关反馈; 机器学习; 小样本

A Survey on Relevance Feedback Techniques in Content-based Medical Image Retrieval

SHEN Ye[1,2] XIA Shun-Ren[1*] LI Min-Dan[2]

[1] (*College of Biomedical Engineering and Instrument Science, The Key Laboratory of Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou* 310027)

[2] (*Department of Signal and Information Processing, China Jiliang University, Hangzhou* 310018)

Abstract: Efficiently and accurately searching for similar medical images in databases is a challenging task. Relevance feedback (RF), a effective approach to bridge "semantic gap" and boost image retrieval, has become a key part of content-based medical image retrieval (CBMIR). RF algorithms can be categorized into three classes according to the retrieval models adopted in them. They are distance-based model, probabilistic model and machine learning based model. Following this categorization, representative algorithms were analyzed and evaluated in this paper with an emphasis on the machine learning algorithms. At last, some promising research directions about RF in CBMIR were discussed.

Key words: content-based medical image retrieval; semantic gap; relevance feedback; machine learning; small sample size

中图分类号 R318 **文献标识码** A **文章编号** 0258-8021(2009)01-0128-09

引言

随着医学成像技术的发展和医院信息网络系统(如 PACS、HIS、RIS)的普及,医院每天产生出大量包含病人生理、病理和解剖信息的医学图像,成为医生进行临床诊断、病情跟踪、手术计划、预后研究、鉴别诊断的重要依据。传统的医学图像检索方法是人工对图像进行文字标记,然后基于一些关键字来检索图像。但人工标注具有很强的主观性,并且文本难以完全表达图像所包含的丰富的语义内容。从医学图像本身提取灰度、形状、纹理、拓扑等底层视觉特征和高层语义特征,构成描述图像内容的特征向量,并以此作为建立索引和匹配准则的客观依据来检索所需图像,称为基于内容的医学图像检索(content-based medical image retrieval, CBMIR)[1]。近年来,CBMIR 技术已成为生物医学工程领域十分活跃的研究方向。

在一个典型的 CBMIR 系统中,通过提取待检索图像和图像库中目标图像的底层视觉特征(如颜色、纹理或形状等)形成特征向量,然后选择合适的相似

收稿日期: 2008-05-13, 修回日期: 2008-08-13

基金项目:国家自然科学基金资助项目(60772092)

*通讯作者。 E-mail: srxia@zju.edu.cn

度匹配技术来实现待检索图像与图像库中目标图像的匹配,返回相似度最高的前 N 幅图像作为系统初步的检索结果。但是,两幅底层特征相似度匹配程度高的图像在语义内容上可能相差很大,存在着所谓的"语义鸿沟"(semantic gap)[2],成为制约基于底层特征的图像检索性能的主要原因。如何解决"语义鸿沟"问题是 CBMIR 的核心技术,也是 CBMIR 研究中最活跃的领域。目前,缩减"语义鸿沟"的技术有:基于区域的图像检索、基于机器学习技术的图像语义分类、相关反馈(relevance feedback,RF)[3]、自适应相似度匹配函数、结合文本的 CBMIR 等[2],其中 RF 技术效果最为显著。

本研究对 CBMIR 系统中的 RF 技术进行了综述,对新近的 RF 算法进行了具体的分析与评价,并就 CBMIR 中 RF 存在的问题及解决的方法进行了探讨,最后对 RF 技术的发展进行了展望。

1　RF 技术

在 20 世纪 90 年代中期,文本检索领域提出的 RF 技术被引入到基于内容图像检索领域,以减少"语义鸿沟"。RF 是 CBMIR 的技术核心,特征选择、降维和相似度匹配等技术都需要与 RF 结合来提高系统性能。在过去十年左右的时间里,各种 RF 算法也不断涌现,目前 CBMIR 中所用的 RF 技术大都是从通用基于内容图像检索系统(content-based image retrieval, CBIR)中移植而来。

在 RF 的交互过程中,用户对系统返回的检索结果给出相关程度的判断。系统根据用户的反馈信息进行学习,通过改变特征空间、优化相似度的匹配公式或更新学习机等方法,使得检索结果更符合用户的主观感受,提高系统的语义检索能力。更多学者愿意将 RF 看成是一个学习问题,它是目前性能表现最好、应用最广泛的技术,但也表现出与其他学习问题一些不同的特点:①相对于图像特征向量的维数(几十到几百维)来说,每轮反馈中标记的训练样本太少,面临小样本学习问题;②反馈样本中的正例样本数要远小于负例样本数,而且相关的类别数也要远小于不相关类别数,面临训练样本的不对称性问题;③CBMIR系统有实时性的要求;④CBMIR 中存在大量的未标记样本。如何利用这些未标记的样本来提高学习算法的性能是一个值得研究的问题。

2　RF 算法

RF 算法的分类可以从不同角度来考虑。如按反馈中的用户模式,可分为"贪婪的"(greedy)和"合作的"(cooperative)两种;按检索目的来分类,可分为目标搜索(target search)和类别搜索(category search);按用户相关判断的度量方式,可分为一类判断、二值判断、多级判断和回归问题;根据 RF 算法所采用的检索模型,可把算法分为基于距离度量的方法、基于概率统计分类模型的方法和基于机器学习的方法,文中根据基于模型的分类分别论述。

2.1　基于距离度量的模型

在该检索模型下,RF 算法的主要策略是优化查询矢量、优化相似度度量公式和特征空间变换。

优化查询矢量也被称为查询点移动(query point movement,QPM),在反馈中把计算正例样本的中心点作为新的查询是启发性的 QPM 算法,但该方法只有正例样本为单一分布时才有较好的效果,而且没有利用负例样本。查询扩展(query expansion)则克服了 QPM 的缺点,它假设正例样本集是由多个单一分布的样本集组成。对每个样本集采用 QPM 技术生成新的查询,形成多点查询来进一步检索图像,但遗憾的是,负例样本同样没有被利用。

相似度度量公式优化也就是距离度量公式中的权值向量的优化问题,被称为权值调整(reweighting)。权值调整的原则是加强能使反馈中的正例样本聚拢并能将正例样本和负例样本有效分开的特征分量,否则减弱该特征分量。在文献[4]中,将所有正例样本在各特征分量上的标准方差的倒数作为该分量的权值。如果正例在某个特征分量上有相似的值(标准方差值小),那么这个特征分量就能较好地反映用户的需求,分配较大的权值。相反,如果正例分散(标准方差值大),那么该特征分量对当前查询就不重要了。但该方法同样是基于正例样本服从单一类型分布的假设之上,然而实际上该假设往往不能满足,另外该方法没有利用负例样本。根据实际样本的分布规律,利用负例样本信息来设计 QPM 算法和优化度量公式是以后进一步研究的方向。另外,文献[5]将权值调整看作是基于某个优化准则的权值优化学习。

特征空间变换法是通过空间变换改变特征空间中图像的分布,把一个特征空间中非聚类的点集映射到另一个特征空间中聚类的点集,与监督的降维技术在某种意义上是一致的。文献[6]的工作可以看成对特征空间进行一个白化变换,但它都只利用了正例样本的信息。判别分析(discriminant analysis, DA)是被广泛应用的空间变换方法[7],通过这种变

换,使得样本在变换得到的空间上的类内离散度与类间离散度的比值最小。由于 CBMIR 中负例样本往往来自数目未知的多个类别,正例样本则常常属于同一类别,因而将负例样本视作一类的 FDA (fisher discriminant analysis) 或单独一类的 MDA (multiple discriminant analysis),在实际应用中都得不到理想的效果。文献[7]注意到正例和负例在数量和分布上的不对称性,提出了一种有偏判别分析 (biased discriminant analysis,BDA) 算法来寻找一个变换。在变换后的空间中,正例更聚拢,同时负例被分散。BDA 也被看作是一个 $1 + x$ 分类器,它假设负例样本有 x 个未知类别。但用户只对其中一个类感兴趣,它仅需要将正例类别从其他 x 个类别中区分出来。BDA 合理地处理了负例样本,在实际中得到了比其他 DA 算法更好的性能表现。

DA 技术是基于正例样本服从高斯分布的假设,而对于实际应用的非高斯分布数据及非线性分布数据却得不到理想的效果。Zhou 等在 BDA 的基础上,提出基于核的 BDA 算法 (biased discriminant analysis using kernel, KBDA)。通过核技术将非线性可分的样本映射到线性可分的核空间上,在该空间进行 BDA 变换[7],该算法比 BDA 在线性空间和非线性核空间上有更好的性能。

尽管 DA 技术是流行的小样本学习方法,但小样本仍然是 DA 算法的一个巨大挑战。通过正则化来解决解决小样本问题是以往的常用技术,然而正则化并不是解决问题的理想选择。文献[8]将直接的思想应用到 KBDA 中,提出了 DKBDA 算法来解决小样本问题。DKBDA 去除了关于正例中心矩阵的负例离散度的零空间,然后将正例样本类内离散度矩阵最小特征值对应的特征向量作为核空间最具判别度的基本向量来构成空间变换。基于 Corel 图像库的大量实验表明,该算法相比 KBDA 与 SVM 算法,有更好的性能表现。

2.2 基于概率统计分类模型

基于距离度量的模型是在图像类别几何可分的前提下展开的,实际应用中的图像库内容丰富,往往不能满足该要求,此时可用概率统计的方法进行分类。该模型的典型算法是基于生成模型 (generative model) 的概率密度估计的分类方法,而高斯混合模型、Bayes 网络等是常用的生成模型。每类模型的参数在该类训练样本上分别估计,它表示了类别的概率分布或内部结构。该方法在少量样本学习上可得到较高的泛化性能、对噪声/异常模式具有抗拒性。基于贝叶斯分类器的 RF 算法被 Cox 等人首次用于图像检索系统中,但当时没有考虑负例样本。概率统计模型在性能上比欧式距离等几何划分图像类的方法更优越,但如果贝叶斯估计分类方法是建立在假设特征向量分布为混合高斯分布的基础上,则这种假设对小概率事件会造成错误判断,影响结果的精确度。文献[6]在假设图像类的特征服从高斯分布的情况下,采用极大似然法来估计高斯分布的参数。文献[9]中采用 DCT 系数上的混合高斯模型来表示特征,在图像的局部特征上采用贝叶斯推断来进行 RF 学习,达到不需图像分割就能对图像区域查询的目的。文献[10]针对训练样本少的困难,提出了一种基于贝叶斯规则的 RF 概率框架,利用了全体样本的分布特点,以提高检索性能。文献[11]提出 RF 框架,同时利用了正例样本和负例样本的概率模型,以实现特征选择。它利用了正例样本来学习生成模型,通过该模型得到每个特征的似然度。另外,利用正例样本与负例样本学习判别模型,通过判别模型得到每个特征对正、负样本的区分能力。然后,根据特征的似然度与区分能力,对其分配不同的权值,达到特征选择的目的。该方法利用了负例样本,并结合生成模型和判别学习的优点,提高了系统的性能。

近年来,结合生成模型和判别模型的混合生成-判别学习[12]方法受到了广泛关注,该方法首先对每一类模式建立一个生成模型,然后用判别学习准则对生成模型的参数进行优化。文献[13]中结合判别学习的 Bayes 网络也可以看作是混合生成-判别学习模型,该模型结合了判别学习与生成模型的优点,为基于概率模型的 RF 技术提出了一个新的思路,值得进一步研究。

2.3 基于机器学习的模型

近年来,许多学者将 RF 技术看作是一个学习问题,比如一类学习问题、两类分类问题等。基于机器学习的模型相比以上两种模型有更好的性能表现和推广性,是目前研究及应用最多的 RF 技术。

小样本学习是基于机器学习的 RF 技术的第一个难题,这也是大部分的学习方法不能直接应用于 CBIR 中的一个主要原因。解决的办法有两种:一是设计适合小样本学习的学习机,如 BiasMap[7],而 DA 和 SVM 是目前表现最好的两个小样本监督学习机;二是利用未标记样本来解决训练样本少的问题。由于 CBMIR 中存在大量未标记样本,利用未标记样本的半监督学习技术来解决小样本问题的方法近年来

被广泛采用,如何有效利用未标记样本也是以后工作中的主要研究方向。

典型的机器学习算法将正例和负例样本的分布视为近似一致。然而,CBMIR 中正例样本往往是属于同一类别,负例样本可能来自多个类别,并且 RF 中负例样本数要远大于正例样本。正例与负例样本的不对称性是 RF 中另一个需要解决的学习问题,因此在 RF 中需要对正例和负例样本有不同的处理方法,如 BDA、一类 SVM 等。近来集成学习技术在解决样本不对称问题上表现突出。另外,RF 是在线的人机交互过程,用户对反馈过程是没有耐性的。为了满足系统的实时性要求,RF 中应避免使用耗时的学习过程。采用适合的降维和特征选择算法得到图像最佳的特征表示,从而提高系统的效率。

由于 RF 中训练样本少及样本不对称的特点,使得基于监督学习的 RF 方法在实际中不能得到理想的结果。近年来,许多学者将半监督学习、主动学习、集成学习、长期学习和多示例学习等学习技术应用于 RF 中,用来解决上述学习问题。下面按所采用的学习技术,对基于机器学习的 RF 技术进行分类论述。

2.3.1　监督学习

在监督学习方法中,SVM 由于在有限样本下有良好的推广能力,在 RF 中被广泛使用。SVM 的本质是找到一个分界面,使得从该分界面到最近的正例样本和负例样本的距离最大,并且利用核技术将特征空间映射到更高维的特征空间,在该空间样本线性可分,从而完成对非线性可分样本的分类。核函数是 SVM 的核心,同一个训练集,不同的核及参数得到不同的分类结果。常用的核函数有 Gaussian RBF 核、Polynomial 核、Sigmiod 核等,其中 Gaussian RBF 核在各种应用中性能良好,是研究者的首选。核函数的参数选择方法在以往的论文中很少进行讨论,留一交叉校验法(leave-one-out)和交叉校验(cross-validation)是常用的参数选择方法。文献[14]提出了一种利用标记样本进行核参数估计的新方法。算法提出一个最大化准则,目的是使核空间上正负例样本最好地分开,采用了拟牛顿法对这种非线性优化的方法求解,得到最优的参数。相比以往的方法,该方法增加了小部分的相应时间,取得了更好的性能,为核参数的优化估计提供了一条新的思路。但是,以往的核参数估计算法都十分耗时,不能满足检索任务的实时要求,因此设计一个实时的在线核参数估计方法是今后进一步的研究方向。

尽管 SVM 是表现较好的小样本学习方法,但 RF 中样本过小时其性能将变差,这主要由下列原因引起:①SVM 分类器在小样本训练时不稳定,其最佳分界面在训练样本很少时对样本比较敏感。在 RF 中,通常用户仅对少量样本进行标记,而且标记不一定正确,所以在样本少和标记不准确的情况下,系统性能将下降。②当负例样本数远大于正例样本数时,SVM 的最佳分界面将有偏离负例样本,这样会将部分不相关样本误判为相关类别。③当样本数明显小于特征维数时,可能引起过学习(overfitting)问题,得不到良好的泛化性能。

因此,对于基于 SVM 的 RF 技术来说,小样本和样本的不对称性是两个必须解决的问题。利用未标记样本和正确处理负例样本,是目前两个主要研究方向。文献[15]针对标记样本少的学习问题,提出一种伪标记(pseudo-label)的概念来扩大训练样本集。算法基于标记样本图像往往呈现出相似的局部特性的假设,通过两级聚类得到正例和负例样本聚类结构。然后,通过 K-NN 算法找到与聚类中心最近的图像作伪标记样本,其类别与最近的聚类中心的类别相同。接着,利用聚类信息估计伪标记样本的软相关隶属度,离同类聚类中心越近且离不同类别的聚类中心越远的伪标记样本的置信度高,反之则低,标记样本为最高置信度。伪标记样本加入原标记样本集,得到了最后的训练集,解决了小样本问题。最后,采用 FSVM 对训练样本集进行训练,它考虑了训练样本的置信度,置信度决定了该伪标记样本对学习机贡献的高低,很好地处理了未标记样本被错误分类而引入的噪声问题。该算法很好地解决了小样学习的问题,并获得了很好的性能表现,但并没有处理样本的不对称问题。

如何有效利用负例样本来解决样本不对称问题,是提高 RF 中学习机性能的另一个有效手段,如 BSVM、BDA 等。但无论是将负例样本看成一类,还是将每个负例视为单独一类,其处理方法都不理想。Tao 提出了更合理的处理方法,将所有正例样本视为同一类,而负例样本划分在一些小的子类中,每个子类有一个简单的分布[16]。算法具体分为以下几个主要步骤:①通过 K-NN 聚类对负例样本聚类;②在每个负例聚类与正例聚类之间建立一个子分类器,可以在一定程度上解决样本不对称的问题;③集成多个子分类器进行 RF 分类,不仅保存了每个子分类器的优点,而且大大提高了系统的泛化能力。该算法将负例样本看作自然聚类的多类别分布,如

何找到负例样本的语义类别结构值得进一步研究。

在 RF 中,SVM 通常使用样本到分界面的距离作为度量,对图像进行相似度排序,然而并没有明确的证据表明这个距离是对图像最好的相似度度量方法,特别在学习得到的分界面大大偏离了真实分布的情况下,这样的度量方法将表现很差。以前也有一些学者对 SVM 的排序问题进行了研究和改进,但并没有引起广泛的关注。文献[17]提出了一个 CSM 距离度量方法。基于用户反馈样本学习 SVM 分类器,将 DB 中的图像分为相关、不相关两类。在分类界面以内的图像也就是相关类别的图像,按与查询图像的欧式距离进行相似度排序。该算法不仅考虑了图像的语义相关性,还应用欧式距离大大提高了检索性能。以后可尝试使用 OPL (optimization learning)和 SVM 构建 RF,使用 OPL 来得到最优的相似度方法,对相关类别图像进行排序,不相关图像按分界面的距离排序。

2.3.2 半监督学习

在监督学习中,要达到好的推广性需要大量的标记样本,而这些标记样本在搜索时很费力,而未标记样本相对比较容易得到。例如,在计算机辅助医学图像分析中,可以从医院获得大量的医学图像作为训练样本,但要求医学专家把这些图像中的病灶都标识出来则不现实。近年来,人们开始利用未标记样本来解决小样本学习问题,提高学习机器的性能,这称为半监督的学习方法。半监督学习中有两个常用的基本假设,即聚类假设和流形假设。聚类假设是指处在相同聚类中的样本有较大的可能拥有相同的标记;流形假设是指处于一个很小局部邻域内的样本具有相似的性质,其标记也应该相似。和聚类假设着眼整体特性不同,流形假设主要考虑模型的局部特性。根据半监督学习算法的工作方式,可以大致将现有的半监督学习算法分为三大类。第一类算法是基于生成模型的分类方法,可以看成是在少量有标记样本周围进行聚类,是早期直接采用聚类假设的做法[18]。第二类算法是基于图正则化框架的半监督学习算法[19-21],可直接或间接地利用了流形假设,它们通常先根据训练例及某种相似度度量建立一个图,图中节点对应样本,边为样本间的相似度。然后,定义所需优化的目标函数,并使用决策函数在图上的光滑性作为正则化项来求取最优模型参数。文献[19]使用高斯随机场以及谐波函数来进行半监督学习,首先基于训练例建立一个图,图中每个节点就是一个(有标记或未标记)样本,然后求解根据流形假设定义的能量函数的最优值,从而获得对未标记样本的最优标记。文献[21]在根据样本间的相似性构造图之后,利用图的局部平滑性,让样本的标记信息不断向邻近的未标记样本传播,直到达到全局稳定状态。基于图的半监督学习是近年来半监督学习的一个显著成就。基于图的半监督学习在 CBIR 中的应用还没有引起广泛关注,在构造图的方法及处理桥点(bridge-points)等问题上需要深入研究。第三类算法是协同学习(co-training)算法。Blum 等提出最早的协同训练算法[22],之后很多研究者对其进行了研究并取得了很多进展,使得协同学习成为半监督学习中最为重要的范例。Zhou 等人将协同学习引入 CBIR 中,提出了基于协同学习的半监督 RF 方法[23-24]。

协同学习是在具有充分、冗余的属性集的样本集上,分别训练得到两个分类器。每个分类器对未标记样本分类,将置信度高的样本进行标记,并加入到另一个分类器的训练样本集中,更新另一个分类器,直到分类器性能满足要求。这是很有吸引力的半监督学习方法,但两个对属性集的约束对许多实际应用来说太严格。Goldman 和 Zhou 扩展了协同学习算法,通过使用两个不同的决策树算法,从同一个属性集上训练出两个不同的分类器,以代替以上两个约束条件。采用十折交叉验证,找出置信度高的样本进行标记,但过于耗时,不能满足实时性要求。为了进一步放松协同训练的约束条件,Zhou 等提出了一种无约束条件的 Tri-training 算法[25]。该算法使用三个分类器来处理标记置信度估计及对未见样本的预测问题,还可以利用集成学习来提高分类器的泛化能力。首先对有标记样本集进行可重复取样(bootstrap sampling),获得三个有标记训练集,然后从每个训练集产生一个分类器。在协同训练过程中,各分类器所获得的新标记样本都由其余两个分类器协作提供,具体来说,如果两个分类器对同一个未标记样本的预测相同,则该示例就被认为具有较高的标记置信度,并在标记后被加入第三个分类器的有标记训练集。在对未见样本进行预测时,Tri-training 算法不再像以往算法那样挑选一个分类器来使用,而是采用集成学习中经常用到的少数服从多数的投票(majority voting)方法,实现对未见样本的预测。但是,初始分类器往往比较弱,未标记样本可能被错误标记,从而给第三个分类器的训练引入噪声,使得半监督学习的性能常常不稳定。为了降低噪声影响,有必要使用一些更可靠的误差估计技术,但这会

在一定程度上增大算法的开销。CBMIR 中的一个特定查询,通常在 DB 中只有小部分图像相关,所以未标记样本被标记为不相关更合理。针对协同学习中的噪声问题,文献[24]采用了一种保守的样本选择策略。对每一轮 RF,每个学习机仅利用负例标记置信度最高的两个未标记样本,有效降低了对样本的错误标记而引入的噪声。文献[26]提出了 Co-Forest 算法,有效地降低了噪声的影响。该算法是在 Tri-training 上的有效扩展,在协同学习的基础上引进了集成学习算法(Ramdom Forest),从而提高系统性能。与 tri-training 中采用三个分类器不同,该算法采用 N 个分类器,解决了未标记样本的误标记问题,大大减少了噪声的引入,并且通过集成学习大大提高了学习机的泛化能力及系统的性能。该算法成功应用于乳腺癌的 CAD 中,实验表明,未诊断样本的使用大大提高了 CAD 系统的性能。另外,如果未标记样本的数据分布不同于标记样本,则使用未标记样本反而会降低学习机的性能。如何利用与标记样本一致的未标记样本,避开不一致的未标记样本,需进一步研究。

2.3.3　主动学习

主动学习是通过主动选择样本让用户标记,减少需要标记的样本数量和预期的反馈次数。主动学习有两种方法:不确定采样(uncertainty sampling)和基于委员会的采样(committee-based sampling)。前者训练一个学习器,用它最没有把握确定类别的样本来询问用户,让用户进行标记。后者训练多个学习器,把学习器之间不能达成共识的样本询问用户。主动学习选择最有信息量的样本给用户标记,加快学习的过程,减少了预期的反馈回合,目前已成为提高 RF 性能的有效方法。在文献[27]中,采用了 SVM 和主动学习相结合的算法 SVMactive,将最靠近 SVM 边界的样本点视为最能提供信息量的样本,这也是应用最广泛的主动学习方法。但这类主动学习技术对学习机的分界面估计很敏感,特别在 RF 初期,训练样本很少,此时训练得到的分界面不准确,主动学习的效果不理想。在文献[28]中,采用了集成学习技术来提高 SVM 在 RF 的初期的稳定性,取得了很好的效果。文献[29]针对 SVM 反馈初期分界面不准确的问题,提出了边界修正技术,以此来修正首轮反馈后的分类界面,它将边界移向最不确定的区域,使相关样本都为正值而其他样本为负值。文献[29]还认为,当有一组数据出现在分界面附近时,选择最靠近分界面样本的传统方法不是最优的策略;提出一个基于平均误差最小化的准则来选择反馈图像,也就是选择能使新分类器的泛化误差最小化图像。相比传统的 SVM active 方法,该算法有更好的性能表现。主动学习结合半监督学习的方法,是另一个近年来被研究较多的主动学习技术。Zhou 将半监督学习中两个协同学习器给出相反判断的样本作为最能提供信息的样本图像询问用户[23]。在此基础上,Zhou 提出了 SSAIRA 算法,将半监督学习和主动学习的优点集成到了 RF 中[24]。其中,半监督学习用来解决小训练样本的问题,主动学习用来提高 RF 返回样本的信息量。提出了新的算法,对误判引入的噪声和样本的不对称问题进行了合理的处理。在算法中,主动学习是选择两个协同学习机而同时具有很低置信度的样本,也就是分类结果接近于零的样本供用户标记,提高了返回样本的信息量。在协同学习中,采用了只利用负例未标记样本的保守策略,降低噪声的引入,提高了学习机的性能。由于 CBMIR 中的负例样本属于多个不同的语义类别,靠近负例样本的图像很可能属于与该负例样本同一个语义类别。本算法试图通过推广负例样本,找到一个未知的、较好的语义类别表示。具体地说,将每个负例样本与 K 个近邻样本平均产生的虚拟样本来代替负例样本供学习所用,部分地解决了样本的不对称问题。然而,不同的负例样本,其近邻属于同一类的样本数目是变化的,所以使用一个自适应的 K 而非固定的近邻区域可能更符合要求。

2.3.4　集成学习

集成学习(emsemble learning)是一种新兴的机器学习范式,使用多个学习机解决同一个问题,能显著地提高学习系统的泛化能力。集成学习有两个算法族:Boosting 和 Bagging。一个集成学习通常分为两步:第一步产生多个子分类器,分类器之间需要差异化;第二步组合不同的子分类器,产生最后性能更优的分类器。子分类器的产生方法分为并行产生和串行产生两类。Bagging 是一个典型的并行方法,此类方法还有 random subspce 和 random forest 等。第二类中比较有代表性的算法是 AdaBoost。在 Bagging 算法中,在子分类器训练阶段,训练集由原始训练集 Bootstrap 获得。最后,通过简单投票或贝叶斯投票等方法集成子分类器,形成最终的分类器,可显著提高不稳定学习器的泛化能力。

目前,集成学习与半监督学习及主动学习等方法相结合,应用于 RF 中,并得到了很有希望的成果,集成学习技术使得学习机更加稳定和强大[24,26]。

文献[28]将集成学习应用到 RF 中来处理 RF 所面临的学习问题,是目前集成学习在 RF 中应用的最高水平。具体地,算法针对 SVM 分类器在 RF 初期样本少而且不对称的情况下出现的不稳定、分界面偏离及过学习等问题,采用了集成学习技术来提高 SVM 的性能。文中首先提出了 AB-SVM(asymmetric bagging SVM)算法,与通常 bagging 方法不同,非对称 bagging 仅对负例样本进行可重复取样。基于正例样本集与可重复取样得到的负例训练集来训练 SVM 子分类器,解决了样本不对称的问题。另外,针对特征维数过高引起的过学习问题,提出以 RS-SVM(random subspace SVM)来解决该问题。RS 与 bagging 类似,不同的是 RS 是对特征空间进行可重复取样。因此,基于 RS 的 SVM 子分类器的训练样本的特征维数会大大减少,从而解决了样本特征维数过高的问题。最后,通过采用投票的方式集成多个 SVM 子分类器,提高了 SVM 子分类器的稳定。基于 Corel 图像数据库的大量实验,表明集成学习技术的引入大大提高了 RF 的性能。以往的研究结果表明,集成学习是提高 RF 中学习机性能的有效手段。对于如何将集成学习更好地应用到 CBMIR 中,该算法起到了很好的提示作用。

2.3.5 长期学习

从一个长期学习(long-term learning)的观点来看,用户在以往 RF 中积累起来的日志(log)数据包含重要的图像语义信息,可以加速学习过程,还可以克服训练样本少的学习问题。近来两个长期学习技术[30-31]被成功应用于 RF 中,其良好的性能表现显示了可通过长期学习利用历史反馈信息提高 RF 性能的重要性。长期学习已经成为 CBIR 的一个新的研究方向。

文献[30]提出一个基于 log 反馈的 CBIR 系统。该算法将基于 log 的 RF 问题看作是一个相关函数 f_q 的学习问题,f_q 用于度量 DB 中图像与查询图像之间的相似度。与传统的基于在线反馈样本生成相关函数不同,文中提出了一个基于 log 反馈和在线反馈样本的相关函数算法,即

$$f_q(z_i) = \frac{1}{2}(f_g(z_i) + f_x(z_j))$$

式中,z_i 为 DB 中的一幅图像,f_g 为基于 log 训练而得到的相关函数,f_x 为传统的、基于在线反馈样本得到的相关函数(如支持向量机(SVM))。

f_g 的生成思想是:如果两幅图像的相关性越高,那么它们在以前的反馈 log 中,在同一反馈回合被同时判为相关的次数就越多,这个次数间接地描述了两幅图像的相关性。基于以上假设,本算法将 z_i 与当前反馈的正例样本的最大相关性减去与负例样本的最大相关性大小生成了 $f_g(z_i)$。当前反馈的正例样本表征了当前查询的语义概念,因此 z_i 与查询图像的相关性越高,那么它与当前反馈的正例样本间就有更高的相关性,相反,与负例样本间的相关性就越小,$f_g(z_i)$ 就越大。由上可知,相关函数 f_g 是基于反馈 log 生成的 DB 中图像与查询图像相关性的合理度量方法。对于一个特定的查询,通过 $f_g(z_i)$ 可以得到 DB 中每幅图像与查询图像的相关性大小。同时,本算法选择了相关性高的图像作为标记样本,加入在线反馈的标记样本,集中训练 f_x,解决了 RF 学习的小样本问题。实验结果表明,作为一种新的 RF 模式,基于 log 数据的长期学习技术大大提高了传统 RF 技术的性能。但 log 中存在的错误 RF 数据给相关函数的学习引入了噪声,如何有效地抑制噪声的影响,提高算法的运算效率,是今后值得研究的方面。

用户在 RF 过程中,对样本的标记隐含地表达了该样本所包含的语义概念,在当前反馈回合中被标记的正例样本就包含了当前的查询概念。在不同的检索任务中,检索的图像概念是不同的,一幅图像可以包含多个概念。文献[31]利用历史反馈数据,将图像以往的反馈信息表示为图像的一个虚拟特征(virtual feature,VF),也就是将图像在历史反馈中被标记的语义概念都保存在 VF 中。在 VF 中,包含了以往用户对图像的语义判断,因此相比底层特征,能更好地表示图像的语义特征。通过图像间虚拟特征的语义相似度的估计,可以得到图像间的语义相关性。随着反馈的进行,图像之间的相似度随着 VF 的更新而被动态地更新。在一个动态的 DB 中,用户的主观概念可能发生转移。算法中通过对检索性能的监控,自动将概念调整到一个新的概念上去。实验表明,将 VF 与传统 RF 技术结合,大大提高了 RF 技术的性能。以 VF 来表示历史反馈数据中隐含的图像语义信息,是一种崭新的长期学习技术,为今后对长期学习技术的研究开拓了思路。

2.3.6 多示例学习

多示例学习(multiple-instance learning)是一种新型的机器学习框架,它是与监督学习、非监督学习和强化学习并列的第四种示例学习框架。在此类学习中,训练集由若干个具有概念标记的包组成,每个包

包含若干没有概念标记的示例;若一个包中至少有一个正例,则该包被标记为正,若一个包中所有示例都是负例,则该包被标记为负。通过对训练包的学习,希望学习系统尽可能正确地对训练集之外的包的概念标记进行预测。由于多示例学习能很好地刻画图像内容描述的歧义性问题,为 CBMIR 提供一个新的、很有希望的技术方法。

经过过去几年的研究,常用的机器学习方法基本上都有了对应的多示例学习版本,但遗憾的是,不同的学习方法在向多示例学习转化时并没有一个一般性的方法或法则。文献[32]对 DD 算法[33]进行了扩展,并将其用于 CBIR 中。在该应用中,从图像中选择出一些方差较大的区域来生成图包,但效果并不理想。由于包生成方法对图像的检索性能有显著影响,文献[34]对此进行了研究。研究者在一种基于 SOM 神经网络的图像分割方法的基础上,提出了一种包生成方法。该方法将整幅图像作为包,而将从图像中分割出的不同区域作为包中示例,并以该区域的颜色特征作为描述示例的属性。实验显示,该方法显著优于前面的包生成方法。多示例学习技术可用来减少 CBIR 中存在的查询语义上的歧义性,而利用多示例学习的 CBIR 检索性能在很大程度上取决于图像包生成算法的优劣。设计一个高效的图像包生成方法,是以后多示例学习在 CBIR 应用中值得研究的关键技术。文献[35]将计算机辅助诊断(CAD)看作是一个有少量正例包和大量负示例的多示例学习问题,对多示例学习来解决 CAD 问题进行了探索性的研究,该领域还处于起步阶段。如何将现有的多示例学习技术及如多示例回归、广义多示例学习等新技术付诸应用,是 CBMIR 和 CAD 领域的一个很有希望的研究方向。多示例学习有效解决乳腺癌 CAD,也是今后的一个研究方向。

3　展望

医学成像技术的发展和 PACS 系统的日益普及,对基于内容的医学图像检索系统提出了严峻的挑战。RF 作为提高检索性能的一种有效手段,已成为 CBMIR 中必不可少的核心技术,成为图像检索中的研究热点,并取得了许多成果。基于机器学习的 RF 技术由于有更为优异的性能表现,而且近年来学者们在将 RF 视为一个学习过程的观点上达成了共识,所以该技术成为目前应用和研究最广泛的 RF 技术。由于 CBMIR 系统的特点,基于 CBMIR 的 RF 技术存在训练样本少、训练样本不对称、实时性强、未标记样本多的特点。目前主要围绕 RF 存在的几个学习问题,改进现有的机器学习算法,以提高 RF 的性能,并尝试将多示例学习和长期学习技术等引入 RF 中,得到了很有希望的成果。

该领域未来的研究方向和发展趋势包括:

(1) 设计适合 RF 的小样本学习机。DA 和 SVM 是目前表现最好的两个小样本监督学习机,但 RF 中过少的样本使它们表现得并不理想。设计真正的小样本学习机是一个研究方向,但难度相对比较大。有效利用 CBMIR 中存在的大量未标记样本来解决训练样本少的问题,是一个可行的方案。

(2) 有效利用负例样本来解决样本不对称问题。如何利用负例样本的语义类别结构来合理处理负样本值得研究。利用新的集成学习方法来解决 RF 中的学习问题,提高检索性能,是 RF 的一个重要研究方向。

(3) 多示例学习与长期学习学习技术在 CBMIR 领域中的应用还处于起步阶段,但其取得的进展很有希望,可能成为 CBMIR 中的下一个研究热点。

(4) 为了满足系统的实时性要求,在 RF 中应避免过于耗时的学习过程。设计有效的特征提取、选择和降维技术,得到图像最佳的特征表示,是获得高性能与高效率兼顾的学习方法的关键。

参考文献

[1] Muller H, Michoux N, Bandon D, *et al*. A review of content-based image retrieval applications-Clinical benefits and future directions [J]. Med Informat, 2004, **73**:1 - 23.

[2] Rahman M, Bhattacharya P, Desai B. A framework for medical image retrieval using nachine learning and statistical similarity matching techniques with relevance feedback[J]. IEEE Transactions on Information Technology in Biomedicine, 2007, **11**(1):58 - 69.

[3] Zhou X, Huang T. Relevance feedback for image retrieval: A comprehensive review[J]. Multimedia Syst, 2003, **8**(6):536 - 544.

[4] Rui Yong, Huang TS, Ortega M, *et al*. Relevance feedback: A power tool in interactive content-based image retrieva1[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1998, **8**(5):644 - 655.

[5] Rui Yong, Huang TS. Optimizing learning in image retrieval[A]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)[C]. Hilton Head Island: IEEE Computer Society, 2000. 236 - 243.

[6] Rui Yong, Huang TS. Optimizing learning in image retrieva1[A]. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition[C]. South Carolina: IEEE Proceedings, 2000, **1**:236 - 243.

[7] Zhou X, Huang T. Small sample learning during multimedia retrieval using BiasMap[A]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. Hawaii: IEEE CVPR, 2001, **1**:11 - 17.

[8] Tao Dacheng, Tang Xiaoou, Li Xuelong, *et al*. Direct kernel biased discriminant analysis: a new content-based image retrieval relvevance feedback algorithm [J]. IEEE Transactions on Multimedia, 2006, **8**(4): 716 - 727.

[9] Vasconcelos N, Lippman A. Bayesian relevance feedback for content-based image retrieval [A]. In: Proceedings of IEEE Workshop on Content-Based Access to Image and Video Libraries [C]. South Carolina: IEEE Proceedings, 2000. 63 - 67.

[10] Wu Hong, Lu Hanqing, Ma Songde. The role of sample distribution in relevance feedback for content-based image retrieval [A]. In: Proceedings of IEEE International Conference on Multimedia and Expo [C]. Lausanne, Switzerland: IEEE ICME, 2002, **1**: 225 - 228.

[11] Chevaleyre Y, Zucker JD. Solving multiple-instance and multiple-part learning problems with decision trees and decision rules. Application to the mutagenesis problem. [A]. In: Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (LNAI 2056) [C]. Ottawa, Canada: Berlin: Springer, 2001. 204 - 214.

[12] Holub A, Perona PA. Discriminative framework for modeling object classes[J]. CVPR, 2005, **1**: 664 - 671.

[13] Greiner R, Su Xiaoyuan, Shen Bin, *et al*. Structural extension to logistic regression: discriminative parameter learning of belief net classifiers[J]. Mchine Learning, 2005, **59**(3): 297 - 322.

[14] Wang Lei, Gao Yan, Chan Kap Luk, *et al*. Retrieval with knowledge-driven kernel design: an approach to improving svm-based CBIR with relevance feedback [A]. In: Proceedings of the 10th IEEE International Conference on Computer Vision [C]. Beijing: IEEE Computer Society, 2005. 1355 - 1362.

[15] Kui Wu, Kim-Hui Yap. Fuzzy SVM for content-based image retrieval[J]. IEEE Ccmputational Intelligence Magazine, 2006, **1**(2): 10 - 16.

[16] Tao Dacheng, Li Xuelong, Stephen JM. Negative samples analysis in relevance feedback[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, **19**(4): 568 - 580.

[17] Guo Dong, Ma Weiying, Zhang Hongjiang, *et al*. Learning similarity measure for natural image retrieval with relevance feedback[J]. IEEE Transactions on Neural Networks, 2002, **13**(4): 811 - 820.

[18] Nigam K, McCallum AK, Thrun S, *et al*. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000, **39**(2/3): 103 - 134.

[19] Zhu Xiaojin, Ghahramani GB, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions[A]. In: Proceedings of the 20th International Conference on Machine Learning (ICML'03) [C]. Washington, DC: AAAI Press, 2003. 912 - 919.

[20] Zhou Dengyong, Bousquet O, Weston J, *et al*. Learning with local and global consistency[A]. In: Proceedings of NIPS(16) [C]. Cambridge, MA: MIT Press, 2004. 321 - 328.

[21] Wang Fei, Zhan Changshui. Label propagation through linear neighborhoods[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, **20**(1): 55 - 66.

[22] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[A]. In: Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98) [C]. New York, USA: ACM, 1998. 92 - 100.

[23] Zhou Zhihua, Chen Kejia, Yuan Jiang. Exploiting unlabeled data in content-based image retrieval [A]. In: Proceedings of the 15th European Conference on Machine Learning (ECML'04) [C]. Pisa, Italy: LNAI 3201, 2004. 525 - 536.

[24] Zhou Zhihua, Chen Kejia, Dai Hongbin. nhancing relevance feedback in image retrieval using unlabeled data [J]. ACM Transactions on Information Systems, 2006, **24**(2): 219 - 244.

[25] Zhou Zhihua, Li Ming. Tri-training: Exploiting unlabeled data using three classifier[J]. IEEE Trans Knowl. Data Eng, 2005, **117**(11): 1529 - 1541.

[26] Li Ming, Zhou Zhihua. Improve computer-aided diagnosis with machine learning techiniques using undiagnosed samples[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2007, **37**(6): 1088 - 1098.

[27] Tong S, Chang E. Support vector machine active learning for image retrieval [A]. In: Proceedings of the ninth ACM international conference on Multimedia Multimedia [C]. Ottawa, Canada: ACM, 2001. 107 - 118.

[28] Tao Dacheng, Tang Xiaoou, Li Xuelong, *et al*. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intellegence, 2006, **28**(7): 1088 - 1099.

[29] Gosselin PH, Cord M. Active learing methods for interactive image retrieval [J]. IEEE Transactions on Image Processing, 2008, **17**(7): 1200 - 1211.

[30] Hoi S, Lyu M, Jin R. A unified log-based relevance feedback scheme for image retrieval [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, **18**(4): 509 - 524.

[31] Yin Pengyeng, Bhanu B, Chang Kuang, *et al*. Long-term cross-session relevance feedback using virtual features [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, **20**(3): 352 - 368.

[32] Yang C, Lozano-Pérez T. Image database retrieval with multiple-instance learning techniques [A]. In: Proceedings of the 16th International Conference on Data Engineering [C]. Washington, DC: IEEE Computer Society, 2000. 233 - 243.

[33] Maron O, Lozano-Pérez T. A framework for multiple-instance learning[A]. In: Proceedings of NIPS(10) [C]. Cambridge, MA: MIT Press, 1998. 570 - 576.

[34] Zhou Zhihua, Chen Kejia, Zhang Minling. A novel bag generator for image database retrieval with multi-instance learning techniques [A]. In: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence [C]. Sacramento, : IEEE, 2003. 565 - 569.

[35] Dundar M, Fung G, Krishnapuram B, *et al*. Multiple-instance learning algorithms for computer-aided detection [J]. IEEE Transactions on Biomedical Engineering, 2008, **55**(3): 1 015 - 1021.