基于内容图像检索中的顺序回归问题*

吴 洪+, 卢汉清, 马颂德

(中国科学院 自动化所 模式识别实验室,北京 100080)

Ordinal Regression in Content-Based Image Retrieval

WU Hong+, LU Han-Qing, MA Song-De

(National Laboratory of Pattern Recognition, Institute of Automation, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: +86-10-62542971, Fax: +86-10-62551993, E-mail: hwu@nlpr.ia.ac.cn, http://www.nlpr.ia.ac.cn

Received 2003-08-26; Accepted 2004-01-07

Wu H, Lu HQ, Ma SD. Ordinal regression in content-based image retrieval. *Journal of Software*, 2004,15(9): 1336~1344.

http://www.jos.org.cn/1000-9825/15/1336.htm

Abstract: Relevance feedback, as a key component of content-based image retrieval, has attracted much research attention in the past few years, and a lot of algorithms have been proposed. Most current relevance feedback algorithms use dichotomy relevance measurement—relevance or non-relevance. To better identify the user's needs and preferences, a refined relevance scale should be used to represent the degree of relevance. In this paper, relevance feedback with multilevel relevance measurement is studied. Relevance feedback is considered as an ordinal regression problem, and its properties and loss function are discussed. A new relevance feedback scheme is proposed based on a support vector learning algorithm for ordinal regression. Since the traditional retrieval performance measures, such as precision and recall, are not appropriate for retrieval with multilevel relevance measurement, a new performance measure is introduced, which is based on the preference relation between images. The proposed relevance feedback approach is tested on a real-world image database, and promising results are achieved.

Key words: content-based image retrieval; relevance feedback; ordinal regression; preference relation; SVM (support vector machine)

摘 要: 相关反馈技术是基于内容图像检索研究的一个重要组成部分.近年来,人们对相关反馈算法开展了许多研究工作,并提出了多种算法.目前,多数的相关反馈算法都是基于二值的相关度量——相关或不相关.为了更好地辨别用户的需要和偏好,就需要考虑相关性在程度上的差异而采用更精细的度量尺度.探讨了支持多级相关度量的相关反馈问题,指出相关反馈问题可以看成是一个顺序回归问题,并讨论了它的特点和损失函数.基于

* Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1998030502 (国家重点基础研究发展规划(973)); the CAS Innovation Project of China (中国科学院创新基金)

作者简介: 吴洪(1971－),男,四川成都人,博士,副研究员,主要研究领域为多媒体信息检索,机器学习;卢汉清(1961－),男,博士,研究员,博士生导师,主要研究领域为图像与视频处理,多媒体信息检索;马颂德(1946－),男,博士,研究员,博士生导师,主要研究领域为计算机视觉,模式识别,图像与视频处理.

一个顺序回归的支持向量学习算法,提出了一种新的相关反馈方案.由于传统的检索性能度量(比如查准率和查全率)不适合多级相关度量的情况,采用了一种建立在图像间偏好关系上的检索性能度量.在现实世界图像数据库上的实验结果验证了所提出相关反馈方法的有效性.

关键词: 基于内容的图像检索;相关反馈;顺序回归;偏好关系;支持向量机

中图法分类号: TP391 **文献标识码**: A

在过去十多年里,随着数字技术的迅速发展和广泛运用,产生了大量的数字图像、视频等多媒体数据.如何有效地从这些数据中找到需要的信息成为一个迫切的问题.在这样的背景下,基于内容的图像检索技术吸引了广泛的注意,成为一个热门的研究领域.该领域早期的研究集中在低层特征的研究上,如颜色、纹理、形状特征.但这样的检索系统性能很有限,主要是因为图像低层特征和高层语义间的间隔以及人类感知的主观性.因此,把人的因素结合到检索过程中就很有必要.基于这样的考虑,相关反馈技术被引入到图像检索领域.相关反馈是一种交互式检索过程,在这个过程中,用户被要求对系统当前的检出结果给出相关性的判断,系统再动态地学习用户反馈以更好地把握用户的信息需求,给出更好的检索结果.在过去的几年里,相关反馈技术逐渐成为基于内容图像检索的一个重要组成部分,人们对相关反馈开展了许多研究工作,并提出了不少算法.

目前,多数的相关反馈算法都是基于二值的相关度量——相关或不相关.一些相关反馈算法[1,2]只把标记为相关的事例用于学习,在这种情况下,相关反馈问题本质上是一个密度估计问题.另一些算法[3~6]同时考虑标记为相关和不相关的事例,从而把相关反馈问题对应于一个两类分类问题.但是,为了更好地把握用户的需要和偏好,就应该考虑相关性在程度上的差异而采用更精细的相关尺度.Yao[7]从用户偏好的概念出发证明了用多值的顺序尺度(ordinal scale)来度量用户相关判断的合理性.在图像检索领域,也有少数相关反馈算法[1,2,8]考虑了多级相关度量,但它们把相关性度量的数值直接用到学习算法的运算中.我们指出,通常采用的多值相关尺度属于顺序尺度.由于顺序尺度不同取值间的差别没有定义,并且顺序尺度的取值在严格单调增变换后仍能反映所度量对象间的关系.因此,在算法中直接使用相关性度量的数值是有问题的.

在本文中,我们主要研究支持多级相关度量的相关反馈算法,并且指出正如在信息检索领域的一些工作[9,10]那样,图像检索中的相关反馈可以看成是一个顺序回归问题.在顺序回归中,不直接使用相关性度量的绝对数值,而是利用由相关性度量值所反映的事例间序的关系来构造损失函数.在我们的相关反馈方案中,采用了在文献[9,10]中提出的顺序回归的支持向量学习算法.在二值的相关度量情况下,查准率和查全率[11]作为对检索性能的度量被广泛地采用.在多级相关度量的情况下,直接使用多级相关尺度数值的性能度量是有问题的[7].这里,我们建议采用建立在图像间偏好关系概念上的归一化基于距离的性能度量[7],简称为 NDPM(normalized distance-based performance measure).最后,通过在现实世界图像库上的实验,比较了我们的方法和其他两种支持多级相关度量的相关反馈方法,实验验证了我们的方法的有效性.在本文中,虽然一些结论是借用自信息检索领域,但我们认为,图像检索和信息检索在我们所关心的方面有着共同特点.在下面的叙述中,我们将互换地使用图像和文档这两个概念.

本文第 1 节阐述为什么相关反馈可以被看成一个顺序回归问题,并且介绍其损失函数.第 2 节介绍用于顺序回归的支持向量学习算法及其核扩展.第 3 节介绍归一化基于距离的性能度量.第 4 节是实验的设置和结果.最后是结论和以后的工作.

1 相关反馈作为顺序回归问题

1.1 相关尺度与顺序回归

面向用户的相关性研究[12](user-oriented relevance research)表明,用户对文档相关性的判断存在于从非常相关经过部分相关到不相关的一个连续的区域.二值的相关性度量虽然简单,但却是一个粗糙的度量.在文献[12]中,作者的研究还进一步指出部分相关的事例在用户的检索过程中也起着重要的作用.另外,在检索系统的研究方面,一些研究人员[1]也指出,一个理想的系统应该能够允许用户指明一个图像在多大程度上符合他的需

求.因此,有必要在检索中引入更细致的相关度量尺度.

然而,在信息科学领域,对于如何度量相关性至今还没有形成共识.在二值相关尺度被广泛采用的同时,人们也尝试了许多其他的相关尺度.其中类别评定尺度(category-rating scale)使用得最为普遍.人们尝试过从 3 级~11 级不同类别数的类别评定尺度来度量相关性.在文献[8]中,采用了 5 级的相关尺度,其取值为:highly relevant,relevant,no-opinion,non-relevant,highly non-relevant,对应的数值为 3,1,0,–1,–3,这些数值是以经验性的方法确定的(ad-hoc).在文献[1]中,采用大于 0 的整数(用户点击鼠标的次数)来度量不同程度的相关性.

在文献[7]中,作者引用度量学理论[13]说明了,如果用户的偏好(preference)满足两个基本公理——非对称性(asymmetric)和负传递性(negative transitive),则用户的判断就可以用一个多值的顺序尺度(ordinal scale)来度量.用户偏好可以定义为文档集上的一个二元关系.具体地,给定一个文档集 D 和它上面的一个关系 $\succ$,对于 D 中的两个文档 d 和 d',$d \succ d'$ 表示用户在这两个文档中更偏好于 d,或说 d 比 d' 更相关,或 d 优于 d'.这个关系被称为严格偏好关系.同时,如果两个文档间不存在严格的偏好,就认为它们无差别(indifferent),可以用"~"来表示无差别关系.非对称性是说,一个用户不能既认为 d 比 d' 好,同时也认为 d' 比 d 好.负传递性是说,一个用户不认为 d 比 d' 好,也不认为 d' 比 d'' 好,那么他就不认为 d 比 d'' 好.满足这两个公理的严格偏好关系是一个严格弱序(strict weak order).我们把这两个公理看成是用户在给出相关判断时应遵守的原则.这时,无差别关系~就是一种等价关系,它对 D 构成一个划分.这样在 D 的商集 D/~上就可以定义一个严格线序(strict linear order).严格线序是一个弱序,且任何两个不同的元素都可以比较.于是,文档就可以被安排在几个(对应商集的元素个数)级别上,在高级别中的文档优于低级别中的文档,而在同一级种的文档间无差别.这样就可以用一个预先定义的顺序尺度来度量用户的相关判断.在度量学理论[13]中,顺序尺度是一种把事物依次排列起来的尺度,它的不同取值间存在着顺序关系,但没有定义取值间的差别.人们也常常用数值来表示这些取值,但这些数值间的减法是没有意义的,并且它们在严格单调增变换后仍能反映所度量的关系.可以看出,前面提到的相关度量尺度都属于顺序尺度.另一方面,用预先定义的相关尺度度量的用户判断,例如一个 3 值的尺度{相关,部分相关,不相关},也可以很容易地用偏好关系来表示.例如,所有相关的文档优于部分相关的文档,而部分相关的文档又优于不相关的文档.

在相关反馈中,用户根据系统所采用的相关尺度为检出图像给出相关判断.然后,相关反馈算法根据用户反馈的图像及对应的相关值来学习从图像特征到相关值的映射,以预测未标记图像的相关值.最后,系统按相关值的顺序把图像返回给用户.因为相关尺度是顺序尺度,所以相关反馈中对顺序尺度变量的预测就是一个顺序回归问题.在信息检索领域,已有一些工作[9,10]对顺序回归算法进行了研究,但在图像检索领域还未见到相关研究的报道.本文尝试把这一问题的研究引入到图像检索中.

1.2 顺序回归的风险公式

顺序回归问题与分类问题以及传统的回归问题一样,都是监督学习问题.给定一个由对象和它们对应的目标构成的训练集,监督学习的任务就是找到从对象到目标值间的映射.更形式化地,给定一个训练集 $S=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n \sim P_{XY}^n$ 和一个从输入空间 X 到输出空间 Y 的映射 h 的集合 H,一个学习过程的目标就是选择一个 $h^* \in H$,使得定义在损失函数 $l: Y \times Y \mapsto R$ 上的风险函数 $R(h^*)$ 最小.风险函数 $R(h)$ 是在概率测度 P_{XY} 下损失 $l(y,h(\boldsymbol{x}))$ 的期望 $R(h)=E_{P_{XY}}[l(y,h(\boldsymbol{x}))]$.在未知分布 P_{XY} 的情况下,多数算法采用经验风险最小化准则,选择使损失在训练集上的平均值 $R_{emp}(h)$ 最小的 h^*.经验风险最小化存在着过学习的问题.结构风险最小化准则在最小化经验风险和控制函数的复杂性之间加以折衷,从而使学习机器有更好的推广性.因此,定义合适的损失函数对于监督学习是非常重要的.在分类问题中,Y 是一个有限且元素间无序的集合.这种情况下,通常采用 0-1 损失函数.当 Y 是一个度量空间(metric space)时,这就是一个回归问题.这时,损失函数可以考虑度量空间的结构.而在顺序回归问题中,Y 是一个有限元素的集合,且元素间存在序的关系,它对应于顺序尺度.由于顺序尺度的特点——不同取值间存在着顺序关系但没有定义取值间的差别,因此在定义顺序回归的损失函数时就存在一些问题.与传统回归问题不同,顺序回归的 Y 不是一个度量空间;与分类不同,简单的 0-1 损失函数不能反映 Y 中元素间的顺序.由于不能找到合适的建立在真实的 y 和预测的 $\hat{y}=h(\boldsymbol{x})$ 上的损失函数 $l(y,\hat{y})$,Herbrich 建议的损失函数 $l_{pref}(\hat{y}_1,\hat{y}_2,y_1,y_2)$ 作用在真实的 y 对 (y_1,y_2) 和预测的 y 对 $(\hat{y}_1,\hat{y}_2)$ 上[9,10],具体介绍如下:

考虑一个输入空间 $X \subset R^d$,其中每个元素表示为一个 d 维特征向量 $\boldsymbol{x} = (x_1,...,x_d)^T \in R^d$.进一步地,我们考虑这样一个输出空间 $Y = \{r_1,...,r_m\}$,它对应于顺序尺度,其元素为顺序尺度的取值,也称为阶.各阶之间存在顺序关系 $r_m \succ_Y r_{m-1} \succ_Y ... \succ_Y r_1$.$\succ_Y$ 表示不同阶之间的序,在相关反馈的应用中,可以解释为"比…更相关".给定一个训练样本集 $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset X \times Y$,我们考虑一个由特征到阶的映射构成的模型空间 $H = \{h(\cdot): X \mapsto Y\}$.通过以下规则,每个映射 h 可以确定一种输入空间中元素间的序.

$$\boldsymbol{x}_i \succ_X \boldsymbol{x}_j \quad \Leftrightarrow \quad h(\boldsymbol{x}_i) \succ_Y h(\boldsymbol{x}_j) \tag{1}$$

一个顺序回归模型的任务就是要找出这样的映射 h^*_{pref},由它得到 X 空间的序有最少颠倒的对 $(\boldsymbol{x}_1, \boldsymbol{x}_2)$.给定两个训练样本 $(\boldsymbol{x}_1, y_1)$ 和 $(\boldsymbol{x}_2, y_2)$,我们要区分两种不同的情况: $y_1 \succ_Y y_2$ 和 $y_2 \succ_Y y_1$.这样,下面的风险函数给出了产生颠倒的概率:

$$R_{pref}(h) = E[l_{pref}(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), y_1, y_2)] \tag{2}$$

其中

$$l_{pref}(\hat{y}_1, \hat{y}_2, y_1, y_2) = \begin{cases} 1 & \text{if} \quad y_1 \succ_Y y_2 \\ & \text{and} \quad \text{not}(\hat{y}_1 \succ_Y \hat{y}_2) \\ 1 & \text{if} \quad y_2 \succ_Y y_1 \\ & \text{and} \quad \text{not}(\hat{y}_2 \succ_Y \hat{y}_1) \\ 0 & \text{else} \end{cases} \tag{3}$$

经验风险最小化准则建议采用能最小化以下经验风险的映射,

$$R_{emp}(h; S) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n l_{pref}(h(\boldsymbol{x}_i), h(\boldsymbol{x}_j), y_i, y_j) \tag{4}$$

由于式(3)的损失函数中不考虑有相同阶的样本对,可以考虑用对应不同阶的样本对来构造一个新训练集.用符号 $\boldsymbol{x}^{(1)}$ 和 $\boldsymbol{x}^{(2)}$ 来表示一个样本对中的第 1 个和第 2 个特征,这个新训练集 $S': X \times X \times \{-1,+1\}$ 可以这样构造:用 S 中所有满足 $y_i^{(1)} \succ_Y y_i^{(2)}$ 或 $y_i^{(2)} \succ_Y y_i^{(1)}$ 的元素来构成集合 $S' = \{((\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}), \Omega(y_i^{(1)}, y_i^{(2)}))\}_{i=1}^t$,其中 $\Omega(y_1,y_2)= sign(y_1 \# y_2)$,# 是阶差, $r_i \# r_j = i - j$,t 是 S' 的大小.在文献[9,10]中,作者给出了一个重要的理论结论.假设一个大小为 n 的训练集 S 是以概率测度 P_{XY} 取自 $X \times Y$ 的.那么,对于每一个映射 $h: X \mapsto Y$,下面的等式成立:

$$\frac{n^2}{t} R_{emp}(h; S) = R^{0-1}_{emp}(h; S') = \frac{1}{t} \sum_{i=1}^t l_{0-1}(\Omega(h(\boldsymbol{x}_i^{(1)}), h(\boldsymbol{x}_i^{(2)})), \Omega(y_i^{(1)}, y_i^{(2)})) \tag{5}$$

考虑到每个映射 $h \in H$ 都可以根据下面的方式定义一个函数 $p: X \times X \mapsto \{-1,0,+1\}$,

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \Omega(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2)) \tag{6}$$

这个结论说明,一个映射 h 在样本集 S 上的经验风险与相应的映射 p 按 0-1 损失函数在样本集 S' 上的经验风险相差一个既不依赖于 h 也不依赖于 p 的常数比例因子 t/n^2.这样,顺序回归问题就可以简化为在特征对 $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ 上的一个分类问题.我们知道,根据训练集 $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ 可以定义 X 上的偏好关系 $\succ_X$,$\prec_X$ 和无差别关系 $\sim_X$.因此,函数 p 很好地表示了偏好关系 $\succ_X$,$\prec_X$ 和无差别关系 $\sim_X$.值得注意的是,原来的问题是根据 S 找到从特征到阶的映射 h,通过考虑阶之间的顺序,现在转化为求从特征对到 $\succ_X$,$\prec_X$ 和 $\sim_X$ 3 类的映射 p.但是,反过来的推导就不一定成立,只有对于那些能在 X 上定义具有非对称性和负传递性的关系(严格弱序)的 p,才存在满足等式(6)的 h.这样的问题也被称为偏好学习问题.同时,我们也注意到对非对称性和负传递性的要求缩小了可行的函数 p 的空间.

2 线性效用模型与顺序回归的支持向量学习算法

保证非对称性和负传递性的一个简单的方法就是把每个特征映射到一个实数:$u: X \to R$.这个值可以看成是

一个文档对于用户的效用(utility)[13].顺序尺度 $\{r_1,...,r_m\}$ 可以看成是对连续变量 $u(\boldsymbol{x})$的一个粗略的测量.这样,我们可以把阶看成实数轴上的线段:

$$y=r_i \quad \Leftrightarrow \quad u(\boldsymbol{x})\in[\theta(r_{i-1}),\theta(r_i)] \tag{7}$$

这里的函数 u(效用)和 $\boldsymbol{\theta}=(\theta(r_0),...,\theta(r_m))^T$ 由训练数据确定.让我们考虑效用函数的线性模型[14]:

$$u(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+b \tag{8}$$

它通过公式(7)把对象映射到阶,同时可以假设 $\theta(r_0)=-\infty$ 和 $\theta(r_m)=+\infty$.我们知道,在下面的条件下,$u(\boldsymbol{x})$对于训练集 S' 中的第 i 个样本不产生错误:

$$z_i\boldsymbol{w}^T\boldsymbol{x}_i^{(1)}>z_i\boldsymbol{w}^T\boldsymbol{x}_i^{(2)} \quad \Leftrightarrow \quad z_i\boldsymbol{w}^T(\boldsymbol{x}_i^{(1)}-\boldsymbol{x}_i^{(2)})>0 \tag{9}$$

其中 $z_i=\Omega(y_i^{(1)},y_i^{(2)})$.这里,偏好关系通过特征的差 $\boldsymbol{x}_i^{(1)}-\boldsymbol{x}_i^{(2)}$ 来表示,特征的差可以看成是由一对特征组合而成的特征.假设属于类 $z_i=+1$ 和 $z_i=-1$ 的 d 维特征 $\boldsymbol{x}_i^{(1)}-\boldsymbol{x}_i^{(2)}$ 间的间隔有限,可以构造更强的约束,

$$z_i[\boldsymbol{w}^T(\boldsymbol{x}_i^{(1)}-\boldsymbol{x}_i^{(2)})]\geq 1-\xi_i,i=1,\ldots,t \tag{10}$$

其中,非负的 ξ_i 表示第 i 条约束违反的程度.最大化间隔的权向量 $\boldsymbol{w}^*$ 可以通过在约束(10)下最小化平方范数 $\|\boldsymbol{w}\|^2+C\sum_{i=1}^{t}\xi_i$ 来求得.这种方法很类似于在支持向量分类器中采用的规范化超平面的思想[15].在文献[9]中,作者还给出了运用结构风险最小化准则的理论依据.进一步地,通过引入拉格朗日乘子和关于 $\boldsymbol{w}$ 作优化得到一个对偶问题,

$$\max_{\substack{0\leq\boldsymbol{\alpha}\leq C1\\ \boldsymbol{\alpha}^Tz=0}}\left[1^T\alpha-\frac{1}{2}\alpha^T\boldsymbol{Z}^TQ\boldsymbol{Z}\alpha\right] \tag{11}$$

其中 $\boldsymbol{z}=(z_1,...,z_t)^T$, $\boldsymbol{Z}=diag(\boldsymbol{z})$, $Q_{ij}=(\boldsymbol{x}_i^{(1)}-\boldsymbol{x}_i^{(2)})^T(\boldsymbol{x}_j^{(1)}-\boldsymbol{x}_j^{(2)})$.然后,可以通过二次优化方法求解.给定式(11)解的一个最优向量 $\boldsymbol{\alpha}^*$,最优的权向量 $\boldsymbol{w}^*$ 可以写为训练集中特征差的线性组合(Kuhn-Tucker 条件).

$$\boldsymbol{w}^*=\sum_{i=1}^{t}\alpha_i^*z_i(\boldsymbol{x}_i^{(1)}-\boldsymbol{x}_i^{(2)}) \tag{12}$$

进一步,可以估计各阶之间的边界.根据式(10),对于所有 $\xi_i=0$ (等价的 $\alpha_i<C$)的训练样本,效用值的差大于或等于 1.那么,如果 $\Theta(k)\subset S'$ 是由训练集中这样的样本构成的集合,其对应的 $\xi_i=0$,阶差为 1,较小的阶为 r_k ,这样,对 $\theta(r_k)$ 的估计为

$$\theta(r_k)=\frac{u(\boldsymbol{x}_1)+u(\boldsymbol{x}_2)}{2} \tag{13}$$

其中 $(\boldsymbol{x}_1,\boldsymbol{x}_2)=\underset{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\Theta(k)}{\arg\min}\ [u(\boldsymbol{x}_i,\boldsymbol{w}^*)-u(\boldsymbol{x}_j,\boldsymbol{w}^*)]$.也就是说,阶 r_k 的最优边界 $\theta(r_k)$ 在阶 r_k 和阶 r_{k+1} 的最相邻(在效用的意义上)对象的效用值的中点.在估计了所有的边界之后,一个新的对象就可以根据等式(7)映射到某一阶.我们在这里要强调的是,作为相关反馈的学习算法,不必计算每幅图像对应的相关值,而是直接根据它们效用值的大小来排列它们.

对非线性效用函数情况的推广类似于非线性支持向量机的推导过程[15].我们引入一个映射 $\Phi(\boldsymbol{x}):X\mapsto Z$,把 X 映射到一个所谓的特征空间 Z,我们就得到一个非线性的效用函数:

$$u(\boldsymbol{x})=\boldsymbol{w}^T\Phi(\boldsymbol{x}) \tag{14}$$

假设这个特征空间 Z 是一个再生核 Hilbert 空间,那么,它可以被一个核函数 $K:\boldsymbol{X}\times\boldsymbol{X}\to R$ 唯一确定,并且核函数有这样的性质: $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\Phi(\boldsymbol{x}_i)^T\Phi(\boldsymbol{x}_j)$.这样, Q_{ij} 可以简化为

$$Q_{ij}=K(\boldsymbol{x}_i^{(1)},\boldsymbol{x}_j^{(1)})-K(\boldsymbol{x}_i^{(1)},\boldsymbol{x}_j^{(2)})-K(\boldsymbol{x}_i^{(2)},\boldsymbol{x}_j^{(1)})+K(\boldsymbol{x}_i^{(2)},\boldsymbol{x}_j^{(2)}) \tag{15}$$

从而避免对映射 $\boldsymbol{x}\mapsto\Phi(\boldsymbol{x})$ 的计算.最后,效用函数可以按如下公式计算:

$$u(\boldsymbol{x})=\boldsymbol{w}^T\Phi(\boldsymbol{x})=\sum_{i=1}^{n}\alpha_iz_i(K(\boldsymbol{x}_i^{(1)},\boldsymbol{x})-K(\boldsymbol{x}_i^{(2)},\boldsymbol{x})) \tag{16}$$

3 图像检索中的性能度量

3.1 归一化基于距离的性能度量

在传统二值相关度量的情况下,查准率和查全率[11]是两个广泛运用的检索性能度量.对于多级相关度量,一些改进的度量也被提了出来,如 Keen 的修正查准率和查全率[7].但是在这些性能度量中,直接使用相关度量数值的做法不一定有意义,因为这样的性能度量对于相关度量数值的严格单调增变换不具有不变性.另一种定义性能度量的方法是,利用由这些相关性度量的数值导出的元素间序的信息.在这种性能度量中,我们采用了归一化基于距离的性能度量 ndpm[7]作为图像检索中的性能度量.下面给出它的具体说明.

用户的偏好关系可以根据集合形式表示为

$$\succ=\{(d,d')\mid \text{对于文档 } d,d' \text{,用户更偏好于 } d \text{,或者说 } d \text{ 比 } d' \text{ 更相关}\} \tag{17}$$

这里,我们同样假设用户的偏好符合非对称性和负传递性.这样的偏好关系也被称为用户排列.检索系统的输出结果通常是文档的顺序排列,可以称为系统排列.对检索系统有效性的度量可以通过检验用户排列和系统排列的一致性来定义.在一个文档集合 D 上的两个排列 $\succ_1$ 和 $\succ_2$ 间的距离可以通过下列公式来计算:

$$\beta(\succ_1,\succ_2)=\sum_{d,d'}\delta_{\succ_1,\succ_2}(d,d') \tag{18}$$

这里,求和是在所有元素不分先后的文档对上进行的.如果 $\succ_1$ 和 $\succ_2$ 给出 d 和 d' 间的偏好判断相同,则 $\delta_{\succ_1,\succ_2}(d,d')=0$;如果 $\succ_1$ 和 $\succ_2$ 中的一个认为 d 和 d' 等价,则 $\delta_{\succ_1,\succ_2}(d,d')=1$;如果 $\succ_1$ 和 $\succ_2$ 给出相反的偏好判断,则 $\delta_{\succ_1,\succ_2}(d,d')=2$.

为了让系统的性能度量与系统如何排列用户认为等价的文档无关,应该采用一种称为可接受的排列准则.这种可接受的排列可以通过任意重新排列在用户偏好中等价的文档来得到.用 $\Gamma_u(D)$ 来表示一个用户排列 $\succ_u$ 对应的所有可接受的排列构成的集合.下面是一个基于距离的性能度量(dpm):

$$dpm(\succ_u,\succ_s)=\min_{\succ\in\Gamma_u(D)}\beta(\succ,\succ_s) \tag{19}$$

这个度量是基于单个查询的,但度量系统的检索性能时,通常采用在一组查询上性能度量的平均值.这就需要同等地评价在不同查询上的检索性能.这时,归一化基于距离的性能度量可以通过除以最大距离来定义.

$$ndpm(\succ_u,\succ_s)=\frac{dpm(\succ_u,\succ_s)}{\max_{\succ\in\Gamma(D)}dpm(\succ_u,\succ)} \tag{20}$$

其中, $\max_{\succ\in\Gamma(D)}dpm(\succ_u,\succ)$ 是 $\succ_u$ 与所有排列的最大距离.归一化基于距离的性能度量可以按如下公式计算:

$$ndpm(\succ_u,\succ_s)=\frac{2C^-+C^u}{2C} \tag{21}$$

其中 C^- 是在系统排列 $\succ_s$ 和用户排列 $\succ_u$ 中给出不同偏好判断的文档对的数目,C 是用户排列 $\succ_u$ 中所有文档对的数目,C^u 是这样的文档对的数目,对中的文档在系统排列中被认为是无差别(等价)的,在用户排列中则不然.在现实的检索系统中,系统排列可以看成是一个严格线序,所以 C^u 等于 0.ndpm 是基于一个查询的,对于一个查询集合,可以用 ndpm 的均值作为性能度量.

3.2 ndpm与R的比较

由于信息检索的目的与顺序回归的目的之间存在着差异.作为顺序回归问题,只是把文档映射到不同的相关级别.而在信息检索中,系统往往要把文档进一步按线序排列输出,就是说,在同一相关级别上的文档也需要排序.正是由于这种目的上的差异,导致在两者性能度量上的差异.现在,我们来分析、比较一下所引入的检索性能度量 ndpm 和顺序回归的经验损失函数 $R_{emp}(h;S)$ (见式(4)).由于顺序回归的经验损失函数和对应的偏好学习的经验损失函数 $R^{0-1}_{emp}(h;S')$ 只相差一个常数比例因子(见式(5)),因此,在这里用 $R^{0-1}_{emp}(h;S')$ 来度量顺序回归算法的性能,并把它简记为 R^{0-1}_{pref}.在计算 ndpm 时(见式(21)),分母中 C 是用户排列的文档对数;分子 C^- 是系统排列和用户排列相矛盾的文档对数.在计算 R^{0-1}_{pref} 时,分母是 S'的大小,等于 $2C$;分子是函数 p 错分的 S'中的样本对数. R^{0-1}_{pref} 和 ndpm 的差别源自于两个问题对样本对是否错分的不同判断上,下面我们分几种情况来考虑.

任取 S'中的一个样本 $((\boldsymbol{x}_i^{(1)},\boldsymbol{x}_i^{(2)}),\Omega(y_i^{(1)},y_i^{(2)}))$,不失一般性地,假设 $\Omega(y_i^{(1)},y_i^{(2)})=+1$,即 $y_i^{(1)}\succ_Y y_i^{(2)}$.考虑采用效用函数 $u(\boldsymbol{x})$把特征映射到实数,并假设检索系统的系统排列是按效用函数大小排序得到.同时,阶按式(7)对应于实数轴上的区间.有如下分析:当 $u(\boldsymbol{x}_i^{(1)})>u(\boldsymbol{x}_i^{(2)})$ 时,有 $h(\boldsymbol{x}_i^{(1)})\succ_Y h(\boldsymbol{x}_i^{(2)})$ 或 $h(\boldsymbol{x}_i^{(1)})=h(\boldsymbol{x}_i^{(2)})$.对应地,在计算检索性能度量ndpm时算0个错分,在计算 R_{pref}^{0-1} 时,算0个或1个错分;当 $u(\boldsymbol{x}_i^{(1)})=u(\boldsymbol{x}_i^{(2)})$ 时,有 $h(\boldsymbol{x}_i^{(1)})=h(\boldsymbol{x}_i^{(2)})$.对应地,在计算 ndpm 时,由于受排序(sorting)算法的影响,算 0 个或 1 个错分,在计算 R_{pref}^{0-1} 时,算 1 个错分;当 $u(\boldsymbol{x}_i^{(1)})<u(\boldsymbol{x}_i^{(2)})$ 时,有 $h(\boldsymbol{x}_i^{(2)})\succ_Y h(\boldsymbol{x}_i^{(1)})$ 或 $h(\boldsymbol{x}_i^{(1)})=h(\boldsymbol{x}_i^{(2)})$.对应地,在计算 ndpm 时算 1 个错分,在计算 R_{pref}^{0-1} 时也算 1 个错分.综合以上 3 种情况,我们有 $R_{pref}^{0-1}\geq ndpm$.

4 实验结果

在这个实验中,我们比较采用支持向量顺序回归算法(SVOR)的相关反馈方法和在文献[1,2]中采用的相关反馈方法(WT)以及采用基于 one-against-rest 策略的多类支持向量机的方法(SVM-MC).实验在现实世界的图像库上进行.我们构造了一个杂类的图像数据库,它由 1 200 幅从 Corel 图像集选出的图像组成.这些图像分别属于 12 个类,每个类 100 幅图像.这些类是大象、老虎、花、鹰、人物、云、海滩、建筑、气球、飞机、秋天和纹理.实验中采用的视觉特征是颜色直方图、颜色距、基于小波的纹理和边缘方向直方图.颜色直方图是在 HSV 颜色空间把 HS 量化成 8×4=32 的颜色上统计得到;在 3 个颜色通道上的 1~3 阶距被用来构造颜色距,并和 24 维的 PWT 小波特征和 8 维的边缘方向直方图一起来构造 73 维的特征向量.其中的每个特征分量通过高斯归一化使其取值在[−1,1]区间.

本实验采用基于事例的查询方式(query by examle)进行检索.40 个查询图像被随机地选出,10 幅来自老虎类,10 幅来自云类,10 幅来自人物类,10 幅来自海滩类.在文献[16]中,Choi 和 Rasmussen 研究了用户在美国历史图像的检索中进行相关判断所采用的标准.其中与图像内容有关的主要有:主题性(topicality)、准确性(accuracy)和完整性(completeness).主题性是指图像和用户的任务有关;准确性是指图片能够准确地表示用户的信息需求;完整性是指图像中包含用户需要的细节.在确定查询的标准答案(被认为与该查询相关的图像)时,我们采用了这些标准.简单起见,在实验中我们采用了 3 级的相关尺度:相关、部分相关、不相关.实验的标准答案是这样构造的:对于来自老虎类的查询图像,所有其他类的图像被认为是不相关的,来自老虎类的图像进一步分为相关的或部分相关的.我们把能看清完整形态的老虎图像标为相关,而特写的或有较多遮挡的老虎图像标为部分相关.如图 1 所示.来自云、人物、海滩类的查询图像对应的标准答案也以类似的方式构造.同时,我们还设计了一个自动标记机制来代替反馈中的人工标记.对于每个查询的第 1 轮检索,系统按每幅图像的特征到查询图像特征的欧氏距离由小到大的顺序来排列图像;此后,系统自动完成 3 轮相关反馈.在每一轮反馈中,自动标记程序根据标准答案把系统返回的前 20 个未标记图像标记为相关、部分相关或不相关.

在 SVOR(support vector learning for ordinal regression)和 SVM-MC 中,我们采用高斯核 $K(\boldsymbol{x},y)=e^{-\gamma\|\boldsymbol{x}-y\|^2}$,取 $\gamma=0.1$,参数 C=1000.所有的算法都用 C 语言编程,二次优化采用 Platt 的 SMO 方法[17]求解.在 WT(whitening transform)方法中,我们引入了正则化因子来避免协方差矩阵奇异化.

在评价系统的性能时,我们求取了 40 个查询的 ndpm 平均值.越小的 ndpm 意味着系统排列越接近可接受的用户排列,对应于更好的检索性能.我们还列出了在前 100 个检出图像中相关和部分相关图像的命中率(hit-rate),以更直观地比较它们的性能,其中命中率越高,性能越好.实验结果显示在表 1~表 3 和图 2 中.“iter0”表示第 1 轮检索,就是还没有进行相关反馈;“iter1”表示第 1 轮相关反馈,依次还有第 2 轮、第 3 轮.从实验结果我们可以看出,SVOR 在检索性能上优于 WT 和采用多类支持向量机的方法(SVM-MC).这说明,SVOR 可以更有效地利用相关度量取值间序的信息.

(a) Relevant
(a) 相关的

(b) Partial relevant
(b) 部分相关的

Fig.1 The ground-truth of queries from tiger class

图 1 老虎类查询的几个标准答案例子

Table 1 Hit-Rate after each round of relevance feedback

表 1 每轮相关反馈后的命中率

Hit-Rate (relevant/partial relevant)	Iter0	Iter1	Iter2	Iter3
WT	9.6/21.8	11.8/25.6	13.1/28.1	14.6/31.6
SVOR	9.6/21.8	15.6/27.3	18.9/35.5	21.3/42.3
SVM-MC	9.6/21.8	14.1/24	17.1/30.6	19.4/41.1

Table 2 R_{pref}^{0-1} after each round of relevance feedback

表 2 每轮相关反馈后的 R_{pref}^{0-1}

R_{pref}^{0-1}	Iter0	Iter1	Iter2	Iter3
SVOR	--	0.453	0.359	0.287
SVM-MC	--	0.492	0.351	0.296

Table 3 ndpm after each round of relevance feedback

表 3 每轮相关反馈后的 ndpm

Ndpm	iter0	iter1	iter2	iter3
WT	0.2	0.219	0.196	0.172
SVOR	0.2	0.16	0.098	0.065
SVM-MC	0.2	0.184	0.128	0.073

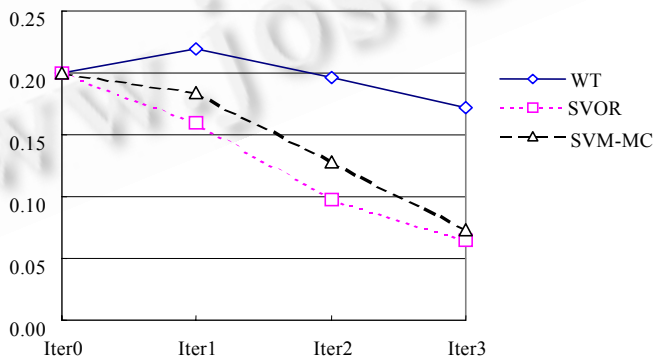

Fig.2 ndpm after each round of relevance feedback

图 2 每轮相关反馈后的 ndpm

5 结 论

在本文中,我们主要研究了支持多级相关度量的相关反馈算法,多级相关度量可以看成是对二值相关度量的扩展和补充.我们指出,在这种情况下的相关反馈可以看成是一个顺序回归问题,并介绍了它的特点和损失函数.图像检索中传统的性能度量主要是基于二值相关度量.这里,我们引入了一种新的性能度量,它建立在图像间的偏好关系上.进一步地,我们开发了一种基于顺序回归支持向量学习算法的相关反馈方法.实验结果证实了我们的方法的有效性.

在用偏好学习的方法解顺序回归问题时,以成对的样本来构成新样本用于学习.所以随着反馈的进行,用户标记的增多,训练样本的数目和相应的训练时间都增加得非常快.我们在将来的工作中,将探索其他解顺序回归问题的思路以及更快的算法,以使该相关反馈方案更具实用性.

References:

[1] Ishikawa Y, Subramanya R, Faloutsos C. MindReader: Query databases through multiple examples. In: Ashish G, Oded S, Jennifer W, eds. Proc. of the 24th VLDB Conf. New York: Morgan Kaufmann Publishers, 1998. 218~227.

[2] Rui Y, Huang TS. Optimizing learning in image retrieval. In: Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR). Hilton Head Island: IEEE Computer Society, 2000. 236~243.

[3] Meilhac C, Nastar C. Relevance feedback and category search in image databases. In: Proc. of IEEE Int'l Conf. on Multimedia Computing and Systems. Florence, 1999. 512~517.

[4] Wu Y, Tian Q, Huang TS. Disciminant EM algorithm with application to image retrieval. In: Proc. of the IEEE Conf. Computer Vision and Pattern Recognition. Hilton Head Island: IEEE Computer Society, 2000. 222~227.

[5] Zhang L, Lin F, Zhang B. Support vector machine for image retrieval. In: Proc. of the IEEE Int'l Conf. on Image Processing. Thessaloniki, 2001. 721~724. http:// research.microsoft.com/users/leizhang/Paper/ICIP01.pdf

[6] Su Z, Zhang HJ, Ma SP. An image retrieval relevance feedback algorithm based on the Bayesian classifier. Journal of Software, 2002,13(10):2001~2006 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/2001.pdf

[7] Yao YY. Measuring retrieval effectiveness based on user preference of documents. Journal of the American Society for Information Science, 1995,46(2):133~145.

[8] Rui Y, Huang TS, Ortega M, Mehrotra S. Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Trans. on Circuits and Systems for Video Technology, 1998,8(5):644~655.

[9] Herbrich R, Graepel T, Ober-Mayer K. Regression models for ordinal data: A machine learning approach. Technical Report, TR-99/03, Berlin, 1999. http://stat.cs.tu-berlin.de/~guru/abstract-HerGraeOber99a.html

[10] Herbrich R, Graepel T, Obermayer K. Support vector learning for ordinal regression. In: Proc. of the 9th Int'l Conf. on Articial Neural Networks. 1999. 97~102. http:// http://ni.cs.tu-berlin.de/publications/abstract-herb99c.html

[11] Van Rijsbergen DJ. Information Retrieval. London: Butter-Worths, 1979. 112~140.

[12] Spink A, Greisdorf H, Bateman J. From highly relevant to nonrelevant: Examining different regions of relevance. Information Processing and Management, 1998,34(5):599~622.

[13] Roberts F. Measurement Theory. Addison Wesley, 1979. 101~111.

[14] Wong SKM, Yao YY, Bollmann P. Linear structure in information retrieval. In: Yves C, LGI-IMAG, eds. Proc. of the 11th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Grenoble: ACM Press, 1988. 219~232

[15] Vapnik V. Statistical Learning Theory. Beijing: Tsinghua University Press, 2000. 85~125 (in Chinese).

[16] Choi Y, Rasmussen EM. Users' relevance criteria in image retrieval in American history. Information Processing and Management, 2002,38(5):695~726.

[17] Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Scholkopf B, Burges C, Smola A, eds. Advances in Kernel Methods: Support Vector Machines. Cambridge: MIT Press, 1998. 185~208.

附中文参考文献:

[6] 苏中,张宏江,马少平.基于贝叶斯分类器的图像检索相关反馈算法.软件学报,2002,13(10):2001~2006. http://www.jos.org.cn/1000-9825/13/2001.pdf

[15] Vapnik V;张学工,译.统计学习理论的本质.北京:清华大学出版社,2000.85~125.