

用于skyline排序的主动学习技术

程蔚蔚¹ Eyke Hüllermeier
马尔堡大学数学与计算机系
{cheng,eyke}@mathematik.uni-marburg.de

摘要: Skyline 查询是当前数据库研究非常热门的课题。它在多目标决策等领域扮演着重要的角色。Skyline 查询的一个重要问题是其输出结果往往非常庞大并且无序。从应用的角度来讲, 将查询结果基于用户的偏好进行排序是一种有效的解决方案。本文提出一种运用主动学习技术的排序方法。其基本思路是根据用户对中间结果的反馈, 通过引入效用函数来更加有效的对用户的偏好进行建模。

关键词: skyline 排序 主动学习 效用方程

Ranking skylines using active learning techniques

Weiwei Cheng¹ and Eyke Hüllermeier
Mathematics and Computer Science
University of Marburg
{cheng,eyke}@mathematik.uni-marburg.de

Abstract: Skyline queries have received considerable attention in the field of databases in recent years. It is important for several applications involving multi-criteria decision making. An important problem of skyline queries is the answer sets can become extremely large. From application point of view, a system response in terms of a ranking of objects, ordered according to the user's preferences, would hence be more desirable than an unordered set. In this paper, we propose a method for constructing such a ranking in an interactive way. The key idea of our approach is to ask for user feedback on intermediate results, and to use this feedback to improve, via the induction of a latent utility function, the current ranking so as to represent the user's preferences in a more faithful way.

Key Words: skyline, ranking, active learning, utility function

1. 引言

Skyline 最先由 Borzsonyi 等人于 2001 年提出[1]。自那之后 skyline 的概念在数据库领域被广泛接受并受到很大的关注, 很多相关的方法, 尤其是计算 skyline 的方法被提了出来。Skyline 查询属于典型的基于用户偏好的查询(preference query): 一个 d 维的 skyline 包含了给定对象集中所有“不逊于”(not dominated) 其他对象的对象。规范的说, 给定对

¹ 本文通讯作者。Corresponding author.

象 $\mathbf{a}=(a_1, a_2, \dots, a_d)$, $\mathbf{b}=(b_1, b_2, \dots, b_d)$, 如果 \mathbf{a} 中任意一个属性 a_i 都不比 \mathbf{b} 对应的属性 b_i 差, 并且 \mathbf{a} 中至少有一个属性优于 \mathbf{b} , 那么我们说 \mathbf{a} 优于 \mathbf{b} (\mathbf{a} dominates \mathbf{b})。我们可以把 skyline 理解为给定对象集中一组经过最优化的子集; 在经济学领域, 这样的最优被称为帕类托最优 (Pareto optimality)。

在这篇文章里, 我们介绍一种特别的信息检索 (IR) 技术, 它综合了 skyline 计算与排序; 我们使用机器学习技术来提取用户的偏好, 并最终对 skyline 里的对象进行排序。在 IR 领域, 使用用户建模与人机交互技术来增强信息系统的互动性、自主性是当前非常活跃的研究方向, 本文所述的技术与这个趋势是完全相应的。

本文安排如下: 我们将在第二节重点介绍我们的 skyline 排序方法。第三节为实验章节。本文的综述以及未来工作的重点将在第四节介绍。

2. Skyline 排序

如前文所述, 我们的目标是对 skyline 在帕类托最优的基础上根据用户的偏好进行再优化, 使 skyline 查询的结果更加贴近用户的需要。为了实现这一目的, 一个重要的问题就是如何对用户的偏好进行建模。解决这个问题的常用方法是使用效用函数 (utility function)。效用函数 $U: \mathbf{0} \rightarrow \mathbf{R}$ 为每一个对象 $\mathbf{a} \in \mathbf{0}$ 分配一个属于实数的效用值。当一个效用方程被确定之后, 排序就非常直接: 如果 $U(\mathbf{a}) > U(\mathbf{b})$, 则 \mathbf{a} 的排序高于 \mathbf{b} ($\mathbf{a} \succ \mathbf{b}$)。显然, 每个不同用户的效用函数不尽一致, 需要通过算法学习获得。

一个合理的效用函数应该是单调的。如何在学习时确保效用函数的单调性是一个具有挑战性的课题。规范的说来, 假定所有的属性值越大越好, 对于所有的 $\mathbf{a}, \mathbf{b} \in \mathbf{0}$ 我们应当有

$$(\mathbf{a} \geq \mathbf{b}) \Rightarrow (U(\mathbf{a}) \geq U(\mathbf{b})) \quad (1)$$

其中 $\mathbf{a} \geq \mathbf{b}$ 表示 $a_i \geq b_i$, $i=1 \dots d$ 。一个非常有意思的现象是, 很多标准的机器学习算法都不能保证这个基本的单调概念。换句话说, 一个由学习算法, 比如说决策树 (decision tree), 所实现的效用方程 U 往往不能满足单调性, 即便用于建立这个模型的所有训练数据都满足这个属性。

这篇文章从一个非常简单的模型开始着手。我们考虑如下的线性效用函数

$$U(\mathbf{a}) = \langle \mathbf{w}, \mathbf{a} \rangle = w_1 a_1 + \dots + w_d a_d. \quad (2)$$

对于这样的线性模型, 保证单调性相对容易。对于模型 (2), 单调性事实上等价于权向量 \mathbf{w} 的非负, 即对于所有 $i=1 \dots d$, $w_i \geq 0$ 。

虽然结构简单, 但线性模型 (2) 在实际应用中却有着一系列的优点。首先, 它非常易于解释: 权值 w_i 直接对应着相应属性值的重要性。因此, 我们可以很容易的在模型 (2) 中整合背景知识, 比如属性 A_i 至少两倍重要于属性 A_j 等价于 $w_i \geq 2w_j$; 另外, 从机器学习的角度来讲, 线性模型也非常有吸引力: 大量现有的学习技术是基于线性模型的。并且, 非线性模型可以通过“核化” (kernelization) 线性模型得到。[2]

2.1 学习算法

提供给学习机的训练数据 \mathbf{T} 包含以 $\mathbf{a} \succ \mathbf{b}$ 形式出现的成对偏好, 其中 $\mathbf{a}, \mathbf{b} \in \mathbf{S}$ (skyline)。由此, 当前的学习目的就是找出一个尽可能满足这些用户偏好的效用方程, 并且, 该效用方程应该满足单调性约束 (1)。此外, 所学习出的效用方程要具备良好的概括性 (generalization ability)。

在线性效用模型 (2) 的前提下, 我们的学习任务本质上可以转化为一个二元分类问题 (binary classification): 偏好 $\mathbf{a} \succ \mathbf{b}$ 所导出的约束 $U(\mathbf{a}) > U(\mathbf{b})$ 等价于 $\langle \mathbf{w}, \mathbf{a} - \mathbf{b} \rangle > 0$ 和 $\langle \mathbf{w}, \mathbf{b} - \mathbf{a} \rangle < 0$ 。这样一来, 我们可把 $\mathbf{a} - \mathbf{b}$ 和 $\mathbf{b} - \mathbf{a}$ 分别看作分类问题中的正例 (positive example) 和反

例(negative example)。

在机器学习中，二元分类问题是一个被深入研究的问题，相应的算法有很多。本文使用贝叶斯点(Bayes point machine)学习机，一类模拟最优贝叶斯决策的算法。以往的研究表明，贝叶斯点可以通过版本空间的质心(center of mass of the version space)来近似求得[3]。具体来说，我们基于训练数据的不同排列产生一个感知器的集合(ensemble of perceptrons)，这个感知器的集合形成了对版本空间的采样；通过这个采样，我们可以近似求得贝叶斯点。采用这个方法的主要优点是：首先，它容许我们相对容易的保证单调属性；另外，在接下来的章节中我们将看到，这个感知器的集合在向用户索取反馈信息的主动学习过程中也起着重要的作用。

感知器学习(perceptron learning)是经典的机器学习技术，它是一种根据错误(error-driven)渐进调整假设空间(hypothesis space，即权向量 \mathbf{w})的算法。为了保证单调性，我们对原算法作了以下改变：每当对 \mathbf{w} 的升级产生一个负值 $w_i < 0$ 的时候，该值被设置为0。换句话说，原来的升级算法被修改为一个受限的升级算法。在训练数据无噪音的情况下，原本的感知器学习算法已经被证明能够于有限次的升级之内实现聚合[4]；使用类似的证明技术，我们可以证明修改后的学习算法同样可以实现聚合[5]。篇幅所限，这里将跳过完整的证明。

版本空间的质心(或者说贝叶斯点)是通过集合中感知器权值的均值来近似得到的。显然，每个感知器的单调性保证了最后输出模型的单调性。

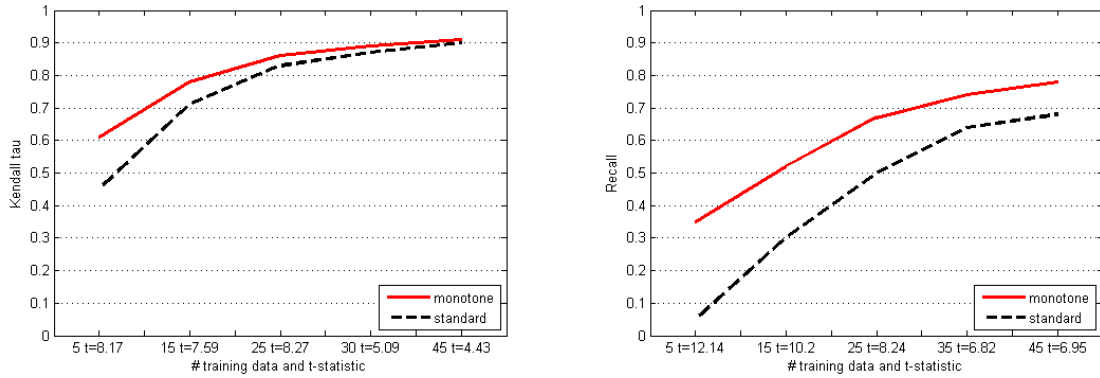
2.2 获取用户反馈

在用户对当前排序结果不满意的情况下，我们提供了一个基于用户反馈的学习过程。用户将被问及其对于对象 \mathbf{a} 、 \mathbf{b} 的偏好，她/他的反馈 $\mathbf{a} \succ \mathbf{b}$ 或 $\mathbf{b} \succ \mathbf{a}$ 将用来扩展训练数据 \mathbf{T} 。选择问题对象 (\mathbf{a}, \mathbf{b}) 最简单的方式是随机的从 $\mathbf{S} \times \mathbf{S}$ 挑出。当然，对于学习机来说不同的对象对所包含的信息量显然不一样，因此在选择问题对象的时候，我们希望尽可能选择信息量最大的对象对，或者说最有助于改善当前学习结果的对象对。如何有目标的选择最有意义的学习数据，正是主动学习(active learning)的核心问题。

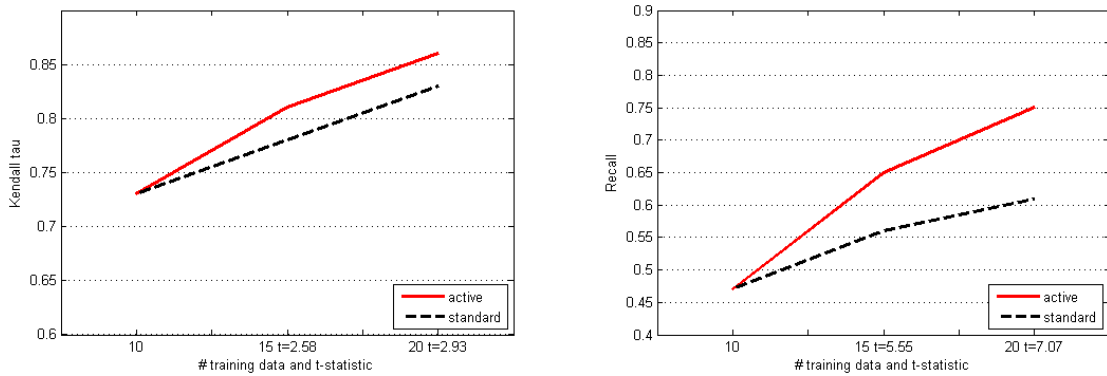
近些年在机器学习领域关于主动学习的研究非常活跃，所提出的方法以启发式的近似算法为主：通常，理论直接导出的算法往往由于复杂度的问题无法直接实现。本文所采用的方法大致可以被视为委员会选择算法(query by committee, QBC)的一个延伸。其主要思想是选择可以使“委员会”中的不一致达到最大的训练数据。这里的委员会指的是一组事先训练出来的模型的集合。直观地来看，这样的数据学习难度较大，最有可能改变当前的模型，也就因此可被认为是信息量最大的训练数据。[6]

结合我们的方法，委员会即是在2.1节讲到的感知器的集合。对于对象对 $(\mathbf{a}, \mathbf{b}) \in \mathbf{S} \times \mathbf{S}$ ，当委员会中的两个感知器其中一个推出 \mathbf{a} 优于 \mathbf{b} ，另一个推出 \mathbf{b} 优于 \mathbf{a} ，我们则说，这两个成员不一致。当然，如何具体挑选信息量最大的训练数据，换句话说如何挑选能够在委员会中产生最大不一致性的训练数据，可以有很多不同的实现方法。本文采用如下的实现步骤：设 $\mathbf{W} = \{\mathbf{w}_1 \dots \mathbf{w}_m\}$ 为组成委员会的感知器的集合，我们首先基于余弦相似度确定集合中差异最大的两个模型；考虑由这两个模型分别产生的两个序列 π_1 、 π_2 ，我们采用从上往下的方式选择第一对不同的对象对 (\mathbf{a}, \mathbf{b}) 作为提出给用户反馈的问题。换句话说，这个对象对出现在 π_1 和 π_2 最先出现不同对象的位置上。举例来讲，假设 π_1 为 $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}$ ， π_2 为 $\mathbf{a}, \mathbf{b}, \mathbf{d}, \mathbf{e}, \mathbf{c}$ ，对象对 (\mathbf{c}, \mathbf{d}) 即为所需。²

² 原则上，当 π_1 、 π_2 相同的时候，我们需要制订额外的策略来选择对象对；鉴于这种情况发生的概率较低，在此不做额外的叙述。



图一：单调学习机 vs. 非单调学习机



图二：主动学习 vs. 非主动学习

3. 实验结果

在这一节里，我们展示与本文所提方法相关的实验结果。我们希望通过实验验证本文中提出的基于机器学习方法的 skyline 排序策略高效并且可靠；我们将测试并探讨保证单调的学习机和主动学习策略所带来的改进。

本节所述的实验是建立在人工合成数据之上的。所采用的人工数据重复采样 50000 个在 9 维单位空间内均匀分布的数据记录。首先，我们使用 BNL 算法得出给定数据的 skyline [1]，然后对于每次实验，我们随机产生一个权向量 w （其元素同样是在 $[0, 1]$ 均匀分布的）用以计算 skyline 中对象的效用值并以此导出一个序列。此序列将被作为产生训练数据以及提供反馈信息的依据（即 ground truth）。

我们将综合考虑两类不同的评估方法，相关系数 (Kendall tau) 和查全率 (recall) [7]。这里，查全率定义为预测的 top-K 对象中属于真实 top-K 对象的比例。在本节里，我们考虑 $K=10$ 的情形。本节中展示的实验结果，是 100 次独立实验的平均值。为了展示实验结果的统计显著性，我们一并给出成对 t 检验 (paired t-test) 的输出结果。

图一展示了保证单调性的学习机在学习质量上的改进。左图和右图分别是基于两种不同评估方式的实验结果。从机器学习的角度解释，附加单调性条件相当于缩小了假设空间的范围，从而降低了查找难度。成对 t 检验证实了保证单调性所带来的显著差异。

图二所述的实验检验了我们所提出的主动学习算法的有效性。主动学习相对于传统的被动学习方法（随机从对象空间内挑选训练数据），能够为学习机提供更富含信息量的训练数据；这对于改进排序质量非常有效。

4. 总结与展望

本文探讨了一系列基于 skyline 排序的机器学习技术。据我们所知，在相关领域内，这是一次全新的尝试。本文的基本目的是通过学习用户的偏好使得 skyline 查询的结果更加贴近用户：一个有序的结果比无序的集合更加符合用户的实际使用需要。

实验结果验证了本文的基本假设，通过与用户的交互我们可以在很短的过程内构造一个准确的序列。从技术角度上讲，保证单调性以及主动学习策略极大的加快了信息检索的进程。

在今后的研究中，我们希望逐步扩大已有的成果，将该方法应用到成熟的信息检索系统中去。目前，更为复杂的，基于实际数据的测试正在进行之中。我们同样也在不断改良所采用的方法，以期获得更好的学习效果。

参考文献

- [1]. S. Borzsonyi, D. Kossmann, and K. Stocker. The skyline operator. In IEEE Conf. On Data Engineering, 421–430, Heidelberg, Germany, 2001.
- [2]. B. Schölkopf and J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.
- [3]. R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. In Journal of Machine Learning Research, 1:245–279, 2001.
- [4]. M. Minsky, and S. Papert. Perceptrons: Expanded Edition. MIT press, 1988.
- [5]. W. Cheng. Interactive ranking of skylines using machine learning techniques. Master thesis, University of Magdeburg, Germany, 2007.
- [6]. H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In Computational Learning Theory, 287–294, 1992.
- [7]. M. Kendall. Rank correlation methods. Charles Griffin, London, 1955.

作者简介

程蔚蔚

德国马尔堡大学 (University of Marburg) 数学与计算机系博士研究生，Hüllermeier 教授主持下的 KEBI 实验室成员。主要研究方向为机器学习、数据挖掘，并在相关领域的国际重要期刊及会议上发表过论文多篇。曾担任多个国际会议的审稿人。现为 ECML/PKDD-08 Preference Learning 委员会成员。德国马格德堡大学 (University of Magdeburg) 计算机硕士学位，郑州大学计算机与工商管理双学士学位。

Eyke Hüllermeier

德国马尔堡大学计算机与数学系教授。主要从事人工智能、机器学习、数据挖掘、模糊系统以及生物信息学等领域的研究工作。IEEE 及 IEEE Computational Intelligence Society 会员，European Society for Fuzzy Logic and Technology (EUSFLAT) 组委会成员，Fuzzy Sets and Systems、Soft Computing、Advances in Fuzzy Systems 等杂志编委会成员，EUSFLAT working group on Learning and Data Mining 联合协调人，IEEE CIS Task Force on Machine Learning 负责人，Case-Based Approximate Reasoning (Springer) 一书的作者。在国际重要期刊会议上发表论文逾百篇。