CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 41 卷第 12 期 2019 年 12 月
Vol. 41, No. 12, Dec. 2019

文章编号:1007-130X(2019)12-2211-06

面向排序学习的层次聚类特征选择算法*

孟昱煜,陈绍立,刘兴长

(兰州交通大学电子与信息学院,甘肃 兰州 730070)

摘　要:大型搜索系统对用户查询的快速响应尤为必要,同时在计算候选文档的特征相关性时,必须遵守严格的后端延迟约束。通过特征选择,提高了机器学习的效率。针对排序学习中快速特征选择的起点多为单一排序效果最好的特征的特点,首先提出了一种用层次聚类法生成特征选择起点的算法,并将该算法应用于已有的 2 种快速特征选择中。除此之外,还提出了一种充分利用聚类特征的新方法来处理特征选择。在 2 个标准数据集上的实验表明,该算法既可以在不影响精度的情况下获得较小的特征子集,也可以在中等子集上获得最佳的排序精度。

关键词:特征选择;排序学习;层次化聚类;贪婪搜索

中图分类号:TP391　　文献标志码:A

doi:10.3969/j.issn.1007-130X.2019.12.016

A hierarchical clustering based feature selection algorithm for ranking learning

MENG Yu-yu, CHEN Shao-li, LIU Xing-chang

(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: Large search systems are especially necessary for quick response to user queries. At the same time, strict backend delay constraints must be observed when calculating the feature relevance of candidate documents. Feature selection can improve the machine learning efficiency. Considering the characteristics that most of the initial points of fast feature selection in ranking learning start from the single feature, which has the best ranking effect, this paper first proposes an algorithm of generating initial points of fast feature selection by hierarchical clustering, and applies the algorithm to two existing fast feature selection algorithms. In addition, a new method that makes full use of clustering features is proposed to deal with feature selection. Experiments on two standard datasets show that the proposed algorithm can obtain a smaller feature subset without affecting the accuracy and obtain the best ranking accuracy on a medium subset.

Key words: feature selection; ranking learning; hierarchical clustering; greedy search algorithm

1　引言

识别最有效的特征子集是一种在训练集上进行特征选择的过程,而识别的特征子集可以被用来学习原有任务的模型,是模式识别的关键问题之一[1]。在机器学习过程中,存在着各种各样的挑战,特别是高维数据的处理。随着数据维数的不断增加,必然会产生大量的无关冗余信息。学习模型趋于过拟合将导致模型表示能力减弱,数据特征的有效性难以表达。已有实验表明,特征选择,即不包括数据中的某些特征,只选择特征的一个子集,并不会削弱数据的表现能力。特征选择作为降维、提高数据质量的方法,引起了研究者的广泛关注。

* 收稿日期:2018-06-11;修回日期:2018-11-15
基金项目:甘肃省自然科学基金(1606RJZA003);甘肃省住房和城乡建设厅项目(JK2015-15)
通信地址:730070 甘肃省兰州市兰州交通大学电子与信息学院
Address: School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, Gansu, P. R. China

当前,排序学习[2]广泛应用于信息检索、自然语言处理以及数据挖掘[3]。迄今为止,排序函数的构建所使用的特征已过百余种,例如:基于内容的特征,如 TF-IDF(Term Frequency-Inverse Document Frequency)、BM25(Best Match 25);基于链接的特征,如 PageRank、HITS(Hyperlink-Induced Topic Search);基于点击数据的用户行为的特征等[1]。然而,这些特征之间并不是完全无关的,比如 TF(Term Frequency)和 IDF(Inverse Document Frequency)这2个特征本身就是 BM25 特征的组成部分,在排序学习中如果把它们同时应用,就会造成一定的信息冗余,模型计算量也会变大,更甚者会影响排序结果。

在排序学习应用特征选择的手段,国内外已经做了如下工作:Geng 等[4]明确强调特征选择是排序学习必不可少的,并且提出了贪婪搜索算法 GSA(Greedy Search Algorithm),会在 GSA 的每一步中选择具有最大相关性的特征,并以此特征与其它剩余特征的相似性来惩罚未选择特征的相关性权重,特征选择也就迭代地进行下去;Hua 等[5]提出一种基于聚类的特征选择方法,首先用 K-menas 聚合相似的特征,然后选择最相关的特征;Pan 等[6]使用集成回归树去研究贪婪和随机 Wrapper 方法,该方法以相似性判断来降低相关性;Lai 等[7]通过处理凸优化问题,将内嵌式方法应用在特征选择与排序模型构建的同一步骤上;Gigli 等[8]根据最小冗余最大相关性准则,在排序学习上提出了3种快速特征选择算法;Naini 等[9]采用一种贪婪多样化的方法来处理特征选择问题;Dang 等[10]提出一种封装式方法,使用最佳优先搜索和坐标上升贪婪地将一组特征分割成子集。

在目前的排序学习领域中,对特征分析的研究相对较少,针对于如何提高特征重组和选择的有效性,本文首先以初始点选择为依据,提出一种新的方法应用在2种快速特征选择算法 GSA[4]和 NG-SA(Naïve Greedy Search Algorithm)[8]上;然后组合层次聚类和特征选择2种算法,提出一种新的 HC-GSA(Hierarchical Clustering for Greedy Search Algorithm)算法应用在特征选择上。出于模型训练效率和训练的耦合度考虑,本文特征选择算法主要是基于过滤式的算法,就过滤式算法本身特性而言,它也是一种极为有效的特征选择算法。相对而言,包装式和内嵌式算法在计算代价上的要求会更高,致使这2种算法在涉及连续或分类特征上的排序学习情景并不适合。为了评估所提的算法,本文使用排序学习中的 LambdaMART 模型在标准数据集上进行实验,并与基准算法(GSA、NG-SA)进行对比。实验结果证明了该算法的有效性,即在不影响甚至提高排序质量的情况下,能够选择到尽可能小的特征子集。

2 特征选择算法

特征选择算法一般分为3类[8,11]:第1类是过滤式算法。该算法的预处理步骤是特征选择,与学习过程无关。过滤算法根据每个特征计算的分数进行特征选择。第2类称为包装算法。该算法将模型学习视为一个黑盒,对每个特征子集进行评分。第3种算法称为嵌入式算法。该算法将特征选择作为模型训练的一部分。

2.1 概述

特征选择的目的是从整个特征集合中选取部分特征,而这些部分特征的代表性足够强,冗余性或者相似性足够小。与 GSA、NAGS 类似,首先,需要定义每个特征的相关分数和特征之间的相似性分数。然后,在最大相关和最小冗余准则的基础上,利用改进的算法最大化特征与排序结果之间的相关性,最小化特征与排序结果之间的相似性。

2.2 特征相关性评测

在训练数据中找到数据特征和文档相关标签之间的一种健壮的相关性的度量,是选择关联函数的主要目的。特征相关性的一种重要标准就是排序质量,Gigli 等[8]也已证明 LM-1(LambdaMART-1)是一种较好的相关性评测方式。本文还利用 LM-1 对特征之间的相关性进行了评价,即在 LambdaMART 排序模型中训练了单个特征,用 NDCG@10 值来衡量特征与类别之间的相关性。

2.3 特征相似性评测

在特征相似性的计算上,Geng 等[4]研究了 Kendall's τ 方法,用其去计算2个不同特征 f_i 和 f_j 分别在同一种排序学习模型上所得结果的差异性来表示相似性。与此类似,Gigli 等[8]使用了 Spearman's rank 相关系数作为特征间的相似性。这2种方法用在这里在原理上并无区别。在本文实验中尤为关注特征间的关系,因此用皮尔逊相关系数来直接衡量2个特征 (f_i,f_j) 间的相似性:

$$PCC(f_i,f_j)=\frac{cov(f_i,f_j)}{\sqrt{var(f_i)\cdot var(f_j)}} \quad (1)$$

其中,2 个特征的协方差表示为 $cov(\boldsymbol{f}_i,\boldsymbol{f}_j)=\sum_{k=1}^{n}(f_i^{(k)}-\overline{f_i})(f_j^{(k)}-\overline{f_j})$,单个特征的方差表示为 $var(\boldsymbol{f}_i)=\sum_{k=1}^{n}(f_i^{(k)}-\overline{f_i})^2$,$f_i^{(k)}$ 为 $\boldsymbol{f}_i$ 中第 k 个元素,$\overline{f_i}$ 为 $\boldsymbol{f}_i$ 中各个元素的平均值。

2.4 优化方式

目前提出的特征子集选择算法要么基于特征的重要性,要么基于特征的重要性和相似性,还有一些两者都不涉及。如上所述,本文采用的非线性特征选择方法是基于 mRMR(minimal Redundancy Maximum Relevance)[12] 的,它平衡考虑了特征的重要性和相似性。在排序学习中,mRMR 提供了一种新的评价方法,即利用其重要性和相似度来获得最大相关和最小相似度的特征子集,可用如下数学问题表示:

$$\max[imp(\boldsymbol{f}_i)-c\sum sim(\boldsymbol{f}_i,\boldsymbol{f}_j)] \tag{2}$$

其中,$imp(\boldsymbol{f}_i)$ 表示特征 $\boldsymbol{f}_i$ 的重要性,$sim(\boldsymbol{f}_i,\boldsymbol{f}_j)$ 表示特征 $\boldsymbol{f}_i$ 与 $\boldsymbol{f}_j$ 的相似性,显然 $sim(\boldsymbol{f}_i,\boldsymbol{f}_j)=sim(\boldsymbol{f}_j,\boldsymbol{f}_i)$,$c$ 是一个超参数,可调节相似性与相关性的权重关系。

3 算法流程

式(2)中的优化问题是一个典型的 0-1 整数规划问题。截至目前,对此类问题没有有效的解决办法,一种可能的办法只能是进行穷举搜索,但是其时间复杂度太高,在实际中很难应用,故而需要更多的实际解决方案。为了解决这些问题,本文对已知的 2 种算法进行了改进,提出了 2 种新的算法,详细描述如 3.1 节和 3.2 节所示。

$F=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n\}$ 表示 n 个特征,$r(\boldsymbol{f}_i)$ 表示特征 $\boldsymbol{f}_i$ 与排序结果的相关性,$s(\boldsymbol{f}_i,\boldsymbol{f}_j)$ 表示特征 $\boldsymbol{f}_i$ 和 $\boldsymbol{f}_j$ 的相似性,它是一种对称评估,即 $s(\boldsymbol{f}_i,\boldsymbol{f}_j)=s(\boldsymbol{f}_j,\boldsymbol{f}_i)$。

3.1 通过层次聚类产生初始点的贪婪选择算法

通过层次聚类产生初始点的贪婪特征选择算法 HCIP-GSA(Hierarchical Clustering generate Initial Points for GSA)/HCIP-NGSA(Hierarchical Clustering generate Initial Points for NGSA)根据第 2 节定义的相关性和相似性计算方法,为避免特征选择的子集陷入局部最优以及增加特征多样性,本节算法结合层次聚类产生初始点。首先,对特征进行层次聚类,产生 x 个类,然后选择每个类中相关性最大的特征添加进初始点集合中,最后从该集合中每个特征开始,进行快速特征选择(GSA、NGSA),具体流程描述如下:

算法 1 通过层次聚类产生初始点的贪婪特征选择算法

输入:特征数据集 G。

输出:特征子集 T_i。

特征全集记为 $G=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n\}$,聚类后的特征全集以 S 表示,通过层次化聚类方法产生 x 个特征子集,即 $S=\{s_1,s_2,\cdots,s_x\}$;

初始点集合置为空,即 $S_0=\varnothing$;

For s_i in S

  初始点选择:$S_0=S_0\cup\arg\max_{\boldsymbol{f}_i\in s_i}(r(\boldsymbol{f}_i))$

End for

For $\boldsymbol{f}_i$ in S_0

  以 $\boldsymbol{f}_i$ 为特征选择初始点,分别用 GSA、NGSA 算法进行选择,将得到的特征点放入 T_i;

End for

分别输出特征子集 $T_1,T_2,\cdots,T_x$。

3.2 结合层次聚类的特征选择

结合层次聚类的特征选择算法 HC-GSA 与 3.1 节中算法类似,但本节算法是通过层次聚类完成快速特征选择。首先考虑将特征分为 x 个类,其中应用了层次聚类算法,也可使用自适应算法;然后选择每个类中相关性最大的特征,并按相关性排序后依次添加进初始特征子集;再将每个不为空的子类中与已选特征相似性最小的特征选择出来,按相关性排序后依次加入有序特征子集;最后直到所有子类中子集为空,则结束算法。

算法 2 结合层次聚类的特征选择算法 HC-GSA

输入:特征数据集 G。

输出:有序的特征子集 T。

特征全集的特征共 n 个,表示为 $G=\{\boldsymbol{f}_1,\cdots,\boldsymbol{f}_n\}$,有序的特征子集表示为 $T=\varnothing$;

通过层次化聚类,将聚类后的特征全集 S 划分为 x 个类,即 $S=\{s_1,\cdots,s_x\}$;

初始特征子集 $S_0=\varnothing$;

For s_i in S

  $\boldsymbol{f}=\arg\max_{\boldsymbol{f}_k\in s_i}r(\boldsymbol{f}_k)$;

  $S_0=S_0\cup\{\boldsymbol{f}\}$;

  $s_i=s_i\backslash\{\boldsymbol{f}\}$;

End for

$T=T\cup\{sort(r(S_0))\}$;

当类集合 S 不为空:

$t=\varnothing$;

For s_i in S

$f = \arg\min_{f_k \in s_i} s(T, f_k)$;

$t = t \cup \{f\}$;

$s_i = s_i \backslash \{f\}$;

End for

$T = T \cup \{sort(r(t))\}$;

输出 T。

4 实验设置

4.1 数据集

本文在2个标准数据集上进行实验,第1个是OHSUMED,来自于Letor 3.0,语料来源于医学检索任务,应用在许多信息检索实验中。它包含106个查询,从在线医药信息库MEDLINE中提取了45维特征,包含了16 140个文档-查询对和3种相关性标注2,1,0。另1个是MQ2008,出自于Letor 4.0[13],语料来自于2008版的TREC Million Query语料库,包含了784个查询和15 211个文档-查询对,拥有46维特征,其结构类似于第1个数据集。这2个标准数据集被分为5个文件夹,每个文件夹包含1个训练集、1个验证集和1个测试集,其数据量大小比例约为3∶1∶1,每个文件夹下的数据集均可用来进行交叉验证。先对数据集做预处理,去除掉全是0的无用特征,以提高训练排序学习模型的效率。清理掉这些特征后,实际特征在OHSUMED中为36个,在MQ2008中为40个。

4.2 评价指标

实验中采用的排序模型评估方式是在信息检索中广泛使用的NDCG(Normalized Discount Cumulative Gain)[3]。NDCG常被用来计算有多个相关性评判等级时的排序精度。对每一个查询,可根据该查询返回的文档相关性大小顺序和实际文档相关性大小顺序计算NDCG值。n表示所处的位置,排在第n位的NDCG值可计算如下:

$$N(n) = Z_n \sum_{j=1}^{n} \frac{2^{R(j)} - 1}{\log(1+j)} \quad (3)$$

其中,$R(j)$表示排在第j位的得分,Z_n是归一化因子,用来保证一个完美排序的NDCG值在n处等于1。对于查询来说,所检索的文档数应小于n,超过n个的文档视为无效文档,也就是说NDCG只计算检索到的文档。3种相关类型,在本文实验的数据集中可概括为完全相关、部分相关和不相关,此处考虑把完全相关定义为正,剩下的为负。

4.3 排序模型

LambdaMART[14]是Learning To Rank中的一个Listwise算法,应用在许多排序学习场景。比如Yahoo!Learning to Rank Challenge比赛中的参赛者就是用这个模型夺得冠军的,它是基于LambdaRank算法和MART(Multiple Additive Regression Tree)算法实现搜索结果排序问题向回归决策树问题的转化。在本文实验中,参数设置如下:叶子节点最少抽样数($min\_samples\_leaf$)为50,最大叶子节点个数(max_leaf_nodes)为7,其余参数取默认值。

5 实验结果与分析

5.1 MQ2008数据集上的实验结果

2种改进算法HCIP-GSA和HCIP-NGSA在数据集MQ2008上的NDCG@10折线图如图1和图2所示,即先通过层次聚类产生初始点,其初始子集特征编号为{28,34,1,14,7},其中LM-1值最高的是34,它也是基础算法的NDCG@10的变化曲线。

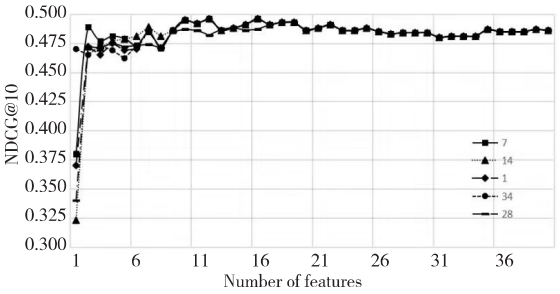

Figure 1 NDCG@10 of HCIP-GSA on MQ2008

图1 HCIP-GSA在MQ2008上的NDCG@10的值

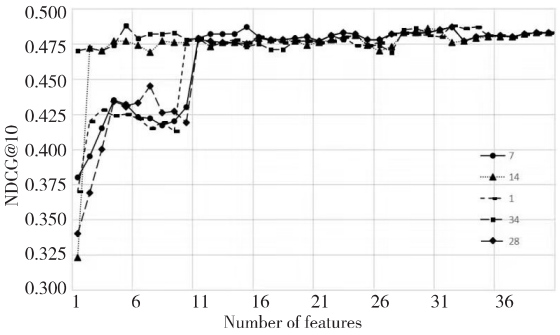

Figure 2 NDCG@10 of HCIP-NGSA on MQ2008

图2 HCIP-NGSA在MQ2008上的NDCG@10的值

如图1所示,在数据集MQ2008上,HCIP-GSA在特征全集的前10%~20%上已经取得了相对较好的结果,当特征个数增加时,排序的精度几乎与最终排序的精度相同。也就是说,改进后的算

法在较小的子集上可以获得更好的性能,并且比在全特征集上具有更高的排序精度。但是,图 2 中的差异并不明显,基本结果与原始算法相似或略高。在图 3 中,所提出的算法在前 40%的特征子集上取得了最好的结果,明显优于基本算法,排序精度明显高于全局值。

本文算法与已有的 2 种算法相比,NDCG@10 的值均高出 1%左右,且高于在特征全集上的结果 1%左右,因此实验证明这种算法是有效的,即能够获得有效的特征子集。表 1 列出了图 1～图 3 所展示的实验结果的具体数值比较,较好结果用加粗的黑体标出。

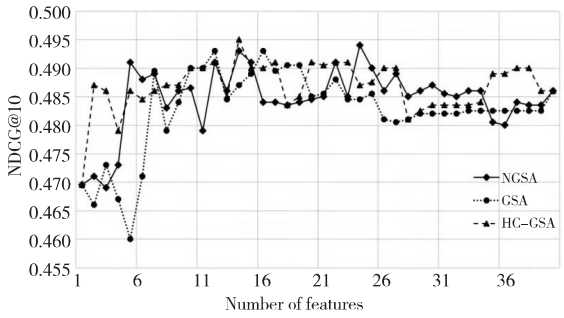

Figure 3 NDCG@10 of HC-GSA on MQ2008

图 3 HC-GSA 在 MQ2008 上的 NDCG@10 的值

Table 1 Comparison of NDCG on MQ2008 dataset

表 1 MQ2008 数据集上的 NDCG 值比较

Subset	GSA $c=0.01$	HCIP-GSA $c=0.01$	NGSA	HCIP-NGSA	HC-GSA $x=5$
10%	0.472 6	**0.487 3**	0.472 2	**0.476 5**	**0.487 3**
20%	0.482 8	**0.488 2**	0.490 0	0.490 0	0.487 3
30%	0.492 4	0.492 4	0.490 0	0.490 0	0.490 5
40%	0.492 9	0.492 9	0.490 0	0.490 0	**0.494 5**
50%	0.492 9	0.492 9	0.490 0	0.490 0	**0.494 5**
Full	0.485 9	0.485 9	0.485 9	0.485 9	0.485 9

5.2 OHSUMED 数据集上的实验结果

图 4 和图 5 是 2 种改进算法在数据集 OHSUMED 上的 NDCG@10 折线图,特征在层次聚类后的初始子集编号为{22,35,5,9,34},其中特征 9 的 LM-1 值最高,基础算法也是以编号 9 的特征开始的。与在数据集 MQ2008 上得出的结论类似,算法 HCIP-GSA 获得的较高排序准确率是在前 10%～20%的特征全集上,而 HCIP-NGSA 在前 40%左右的特征全集上获得了较好的排序准确率。从图 6 中可得出类似结论,即所提算法在 20%左右的特征子集上获得了较好的结果,优于 2 种基本算法,且明显优于在特征全集上获得的结果。

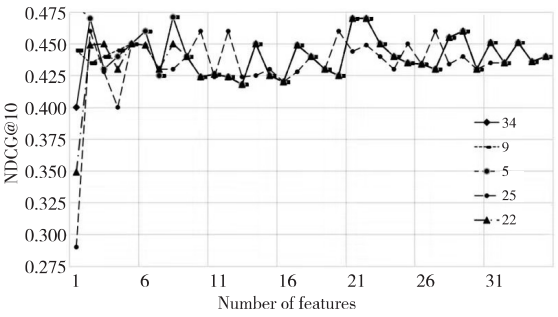

Figure 4 NDCG@10 of HCIP-GSA on OHSUMED

图 4 HCIP-GSA 在 OHSUMED 上的 NDCG@10 的值

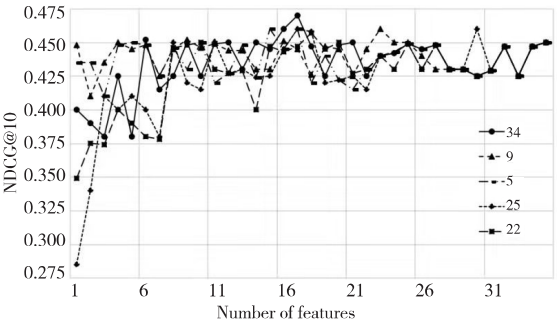

Figure 5 NDCG@10 of HCIP-NGSA on OHSUMED

图 5 HCIP-NGSA 在 OHSUMED 上的 NDCG@10 的值

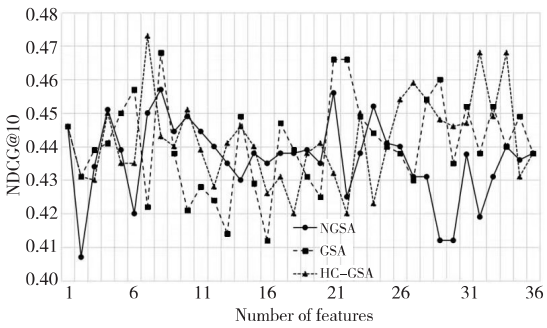

Figure 6 NDCG@10 of HC-GSA on OHSUMED

图 6 HC-GSA 在 OHSUMED 上的 NDCG@10 的值

实验结果和 5.1 节中的实验结果分析类似,在特征子集上获得的 NDCG@10 值比在特征全集上的高 5%左右,在该数据集上取得的改进效果更加明显。图 4～图 6 所示实验结果的具体数值比较如表 2 所示,较好结果均以黑体加粗表示。

Table 2 Comparison of NDCG@10 on OHSUMED dataset

表 2 OHSUMED 数据集上的 NDCG@10 比较

Subset	GSA $c=0.01$	HCIP-GSA $c=0.01$	NGSA	HCIP-NGSA	HC-GSA $x=5$
10%	0.445 6	**0.467 2**	0.451 5	0.451 5	0.449 6
20%	0.456 8	**0.468 3**	0.451 5	0.451 8	**0.471 9**
30%	0.468 3	0.468 3	0.455 8	0.455 8	0.471 9
40%	0.468 3	0.468 3	0.455 8	**0.459 9**	0.471 9
50%	0.468 3	0.468 3	0.455 8	**0.466 5**	0.471 9
Full	0.437 9	0.437 9	0.437 9	0.437 9	0.437 9

在实验中,当选择的特征数量增加时,排序精度并不总是提高甚至反而降低,这说明了特征选择的必要性。原因是当有更多的特征时,由于测试集的过度拟合导致效果不好,这在分类或回归问题中是很常见的。

通过实验,首先验证了特征选择的必要性,即通过特征选择算法,以 NDCG@10 的值作为评价指标,所选的特征子集能够获得较好的排序结果,且排序质量高于在特征全集上的值;其次,本文算法的排序质量高于已有的 2 种快速特征选择算法,有效性得以验证。另一方面,该算法的处理过程并不复杂,略优于现有的经典算法,具有较高的实际应用意义。

6 结束语

本文对 2 种快速特征选择算法进行了改进,并将其(HCIP-GSA、HCIP-NGSA)应用在排序学习上;提出一种新的算法 HC-GSA 进行特征选择。HCIP-GSA、HCIP-NGSA 均是以层次聚类产生初始点,新特征的选择以不同的初始点开始,而 HC-GSA 则是利用层次聚类特性,对所选特征与已选子集关系做了衡量,从而保证每次所选特征为最佳。在 2 个标准数据集 MQ2008、OHSUMED 上的实验也表明,本文算法在较小的特征子集上获得了较高的排序准确率。在对数据进行预处理时花费了一点时间,但这点时间对于模型训练来说无关紧要。

未来还需要做的是:首先,数据集的选择,类型更多、维度更大的数据集应当考虑加入,测试所提算法的有效性和通用性;其次,特征选择算法值得继续探索,引入另外的算法框架,改进本文算法,相似性和相关性评价标准也可引入多样化作为一种特征;此外可考虑是否能将所提算法应用于其他方向上,比如分类或是回归问题。

参考文献:

[1] Hua Gui-chun, Zhang Min, Kuang Da, et al. Feature analysis methods for learning to rank[J]. Computer Engineering and Applications, 2011, 47(17): 122-127. (in Chinese)

[2] Qin Tao, Liu Tie-yan, Xu Jun, et al. LETOR: A benchmark collection for research on learning to rank for information retrieval[J]. Information Retrieval Journal, 2010, 13(4): 346-374.

[3] Li H. Learning to rank for information retrieval and natural language processing[M]. Williston: Morgan & Claypool Publishers, 2011.

[4] Geng X, Liu T-Y, Qin T, et al. Feature selection for ranking [C]// Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007: 407-414.

[5] Hua G, Zhang M, Liu Y, et al. Hierarchical feature selection for ranking[C]// Proc of International Conference on World Wide Web, 2010: 1113-1114.

[6] Pan F, Converse T, Ahn D, et al. Greedy and randomized feature selection for web search ranking[C]// Proc of the 2011 IEEE 11th International Conference on Computer and Information Technology, 2011: 436-442.

[7] Lai H, Tang Y, Luo H X, et al. Greedy feature selection for ranking[C]// Proc of International Conference on Computer Supported Cooperative Work in Design, 2011: 42-46.

[8] Gigli A, Lucchese C, Nardini F M, et al. Fast feature selection for learning to rank[C]// Proc of the 2016 ACM International Conference on the Theory of Information Retrieval, 2016: 167-170.

[9] Naini K D, Altingovde I S. Exploiting result diversification methods for feature selection in learning to rank[C]// Proc of European Conference on Information Retrieval, 2014: 455-461.

[10] Dang V, Croft B. Feature selection for document ranking using best first search and coordinate ascent[C]// Proc of ACM SIGIR Workshop on Feature Generation and Selection for Information Retrieval, 2010: 1-5.

[11] Yao Ming-hai, Zhao Lian-peng, Liu Wei-xue. Research on bagging classification algorithm based on feature selection [J]. Computer Technology and Development, 2014, 24(4): 103-106. (in Chinese)

[12] Qiu Ke-li, Guo Zhong-wen, Liu Qing, et al. Feature selection algorithm based on redundancy analysis[J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40(1): 36-41. (in Chinese)

[13] Qin Tao, Liu Tie-yan. Introducing LETOR 4.0 datasets[J]. arXiv preprint arXiv: 1306.2597, 2013.

[14] Burges C J C, Svore K M, Bennett P N, et al. Learning to rank using an ensemble of lambda-gradient models[J]. Journal of Machine Learning Research, 2011, 14: 25-35.

附中文参考文献:

[1] 花贵春,张敏,邝达,等.面向排序学习的特征分析的研究[J].计算机工程与应用,2011,47(17):122-127.

[11] 姚明海,赵连朋,刘维学.基于特征选择的 Bagging 分类算法研究[J].计算机技术与发展,2014,24(4):103-106.

[12] 仇利克,郭忠文,刘青,等.基于冗余分析的特征选择算法[J].北京邮电大学学报,2017,40(1):36-41.

作者简介:

孟昱煜(1975-),女,河北张家口人,硕士,副教授,研究方向为数据挖掘。E-mail:529267338@qq.com

MENG Yu-yu, born in 1975, MS, associate professor, her research interest includes data mining.