doi: 10.12068/j.issn.1005-3026.2018.12.003

交互式数据探索框架的特征自适应技术

王蒙湘, 李芳芳, 于 戈
(东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

摘 要: 交互式数据探索是一组多样的发现式应用程序的关键技术,着重于交互、探索和发现;在许多场景和领域中广泛应用.以海量的学术文献数据探索为背景,对交互式数据探索的特征自适应技术进行研究.首先,提出一种适用于面向学术文献数据探索的特征自适应交互式数据探索框架FA-IDE(feature-adaptive interactive data exploration),在每次迭代过程中动态地调整特征子集,以满足用户兴趣多样性的需求.其次,针对该框架,提出特征子集的均匀度BFS(balance of feature subsets)评价准则,并给出了基于BFS的序列前向特征选择算法.再次,针对相关样本发现问题,提出划分等级建立方法,根据决策树模型对用户兴趣区域划分后,提出基于相似度的结果集排序策略.实验结果表明,所提出方法可有效提高用户探索效率和最终结果的准确性.

关 键 词: 交互式数据探索;主题提取;特征选择;样本发现;机器学习
中图分类号: TP 315 **文献标志码**: A **文章编号**: 1005-3026(2018)12-1685-06

Feature Adaptive Technology in Interactive Data Exploration Framework

WANG Meng-xiang, *LI Fang-fang*, *YU Ge*
(School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: YU Ge, E-mail: yuge@mail.neu.edu.cn)

Abstract: Interactive data exploration (IDE) is a key technique in a diverse set of discovery-based applications, which focuses on interaction, exploration and discovery and has a wide range of applications in many scenes and areas. The feature adaptive technology of interactive data exploration was studied in this paper with the background of massive academic literature data exploration. Firstly, a framework of interactive data exploration was presented, namely FA-IDE (feature-adaptive interactive data exploration) framework, which can dynamically adjust the subset of features during each iteration to meet the needs of the user's interest diversity. Secondly, according to this framework, the evaluation criteria of the balance of feature subsets (BFS) were proposed in the stage of exploration and a sequence forward feature selection algorithm based on BFS was also given. Besides, for the phases of related sample discovery, a division level establishment method was proposed. According to the decision tree model which can divide the user interest area, a strategy of result set sorting based on similarity was proposed. The results of experiments show that the accuracy and efficiency of the proposed method have been effectively improved.

Key words: interactive data exploration; topic extraction; feature selection; sample discovery; machine learning

交互式数据探索在挖掘大数据的数据价值方面具有重要作用.通常来说,交互式数据探索(interactive data exploration, IDE)是指用户在不十分明确自己查询输入的前提下,系统通过列举样例、协同过滤、机器学习等技术和方式与用户进行交互和反馈,从而逐渐接近用户的真实查询意图,最终提供给用户与其查询意图最匹配的查询结果或返回相应的查询语[1].交互式数据探索的

收稿日期: 2017-05-24
基金项目: 国家自然科学基金资助项目(61472071);中央高校基本科研业务费专项资金资助项目(N161604005);辽宁省自然科学基金资助项目(2015020018).
作者简介: 王蒙湘(1991-),女,内蒙古赤峰人,东北大学博士研究生;于 戈(1962-),男,辽宁大连人,东北大学教授,博士生导师.

关注点[2]是强调交互、探索和发现. 用户从海量的数据中用较小的精力,更准确地找到所需要的信息,其方式有别于用户通过搜索输入关键字找到所需信息的搜索过程[3].

对于数据探索这种大数据价值发现方式,其难点在于在不同场景和领域下,数据处理方式和交互式框架结构都有所不同. 因此,本文面向科学文献管理领域,具体讨论和研究海量文献数据的交互式数据探索关键技术与实现. 随着 Internet 的迅猛发展,网络上学术文献共享以及文献数量膨胀,产生了"信息迷向"和"信息过载"的问题. 其次,一些数据库和信息检索系统的交互方式为"提交查询 – 返回结果",这不能满足用户在查询过程中对信息需求的多样性与动态性[4].

针对以上问题,本文以面向学术文献的交互式数据探索关键技术作为研究内容,提出一种基于特征自适应的交互式数据探索的框架 FA – IDE(feature-adaptive interactive data exploration),在每次迭代过程中动态地调整特征子集,以满足用户兴趣多样性的需求. 其次,针对探索方法,提出了特征子集的均匀度 BFS(balance of feature subsets)评价准则,并给出了基于 BFS 的序列前向特征选择算法. 最后,针对相关样本发现问题,提出划分等级建立方法,根据决策树模型对用户兴趣区域划分后,提出基于相似度的结果集排序策略.

1 交互式数据探索概述

交互式数据探索与传统的 Web 数据探索查询不同,它的特点可概括为三个方面[1]. 一是查询动态性,即输入数据动态性[2]和数据信息的动态性;二是交互反馈性[1],即根据用户的探索行为,对查询进行动态调整以精确地预测结果;三是学习主动性,即引入学习主动性,将机器学习方法应用于数据探索的各个阶段以提高数据探索的效率和准确性[5].

近年来,已经存在很多原型系统针对交互式数据探索中的数据探索进行处理. DICE 系统[6]和 Blink 系统[7]都支持海量数据的交互式查询,但两者为提高交互式响应时间及快速返回查询结果而牺牲了结果准确性. AIDE 系统[5]是一种支持 IDE 的自动化用户导航系统,但存在属性分布偏斜等问题. SnapToQuery 系统[8]是一种基于 Snapping 技术的反馈机制,通过"快照"用户可能感兴趣的查询,指导用户在查询规范过程中提出互动反馈. 但不同的连接和聚合可视化效果不同,泛化能力差,基于运动捕捉的 Snap 界面存在抖动、敏感等问题.

2 基于特征自适应的交互式数据探索框架

本文提出的基于特征自适应的交互式数据探索框架 FA – IDE,如图 1 所示,改变传统的根据关键字等的输入方式,采用基于手动标记示例的数据探索方式.

在数据预处理阶段,给定一个数据库 D,在数据库中提取包含 m 个文献样本的初始文献数据集 $S(x_1, x_2, \cdots, x_m)$,假设用户已经决定探索 n 个属性,即文献的主题,每个文献样本 $\boldsymbol{x}_i$ 由 d 个属性(主题)描述,$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ 是 d 维样本空间中的一个向量,其中,x_{ij} 表示 $\boldsymbol{x}_i$ 在第 j 个属性上的取值,d 为样本 $\boldsymbol{x}_i$ 的维数. 文献样本经过 LDA(latent Dirichlet allocation)模型[9]的处理后,构建出文献的主题特征向量 $\boldsymbol{T} = (t_1, t_2, \cdots, t_k)$,其中,$t_i$ 表示第 i 个主题,进而生成文献 – 主题模型.

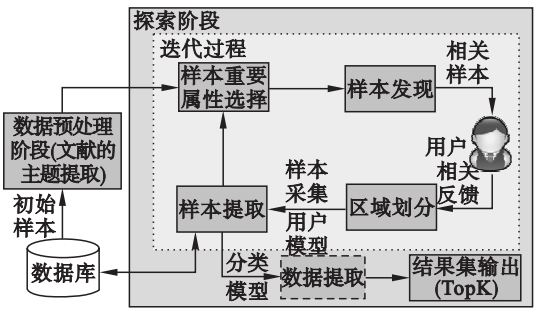

图 1 FA – IDE 系统框架

Fig. 1 System framework of FA-IDE

在探索阶段,系统进行样本重要属性选择. 在每次迭代过程中,从所有特征集合 $F = \{f_i \mid i = 1, \cdots, k\}$ 中选择出特征子集 $X = \{f_1, f_2, \cdots, f_k\}$,每次迭代过程动态地调整特征子集,随后,对探索空间根据重要属性进行网格划分,进入相关样本发现阶段. 将采样样本提供给用户进行反馈,并要求用户按特征将探索任务标记感兴趣与不感兴趣的样本,且这个分类在下一个迭代中不能修改.

当用户提供反馈时,迭代指导过程开始. 用户的相关性反馈要求按属性将样本分类,用户标记的第 i 个样本记为 $(\boldsymbol{x}_i, \boldsymbol{y}_i)$,其中 $\boldsymbol{y}_i \in Y$ 是样本 $\boldsymbol{x}_i$ 的标记,Y 是所有标记的集合或输出空间. 已标记的样本用于训练分类模型 M 以划分用户的兴趣

区域. 此时,将每次迭代选取的相关样本集与基于网格划分提取出的样本集合并,在下次迭代中,将提取的样本一起反馈给用户进行标记,达到提高系统精度的目的.

当用户显式地终止该过程时,即用户认为相关样本的集合已经达到了令其满意的大小或用户不想再标记更多样本时,迭代过程结束. 最后,系统将分类模型进行数据提取并把结果集排序后输出.

3　FA – IDE 中的探索方法

3.1　基于特征选择的样本重要属性选择

本文提出特征子集的均匀度 BFS 的评价准则,将类别上所有样本数据的评估特征均匀度因子引入 DFS(discernibility of feature subsets)准则[10],使其具备评价数据偏斜度的能力,并结合序列前向选择方法 SFS(sequential forward selection)实现特征选择算法,以达到在迭代过程中动态调整特征子集的目的.

定义 1　对于 $C(C\geqslant 2)$ 类分类问题,设训练样本集为 $\{(\boldsymbol{x}_m, y_m) \mid \boldsymbol{x}_m \in \mathbf{R}^k, k>0, y_m \in \{1,\cdots,C\}, C\geqslant 2, m=1,\cdots,n\}$,其中,$n$ 是训练样本集规模,k 是样本空间维数,$\| y_m \mid y_m = j, m=1,\cdots,n \| = n_j, j=1,\cdots,C$,$n_j$ 为第 j 类的样本个数,则含有 i 个特征的特征子集的均匀度 BFS_i 定义为

$$\mathrm{BFS}_i = \lambda \frac{\sum_{j=1}^{C} \| \bar{\boldsymbol{x}}^{(j)} - \bar{\boldsymbol{x}} \|^2}{\sum_{j=1}^{C} \frac{1}{n_j - 1} \sum_{m=1}^{n_j} \| \bar{\boldsymbol{x}}_m^{(j)} - \bar{\boldsymbol{x}}^{(i)} \|^2} + \mu \frac{1}{\sum_{j=1}^{i} \chi^2}. \quad (1)$$

式中:λ 是区分度因子的权重系数;μ 是均匀度因子的权重系数;χ^2 为卡方分布,检验样本在各个特征上是否分布均匀;$\bar{\boldsymbol{x}}$ 表示包含前 i 个特征的子集在全部数据集上的均值向量;$\bar{\boldsymbol{x}}^{(j)}$ 表示在第 j 类数据集上的均值向量;$\boldsymbol{x}_m^{(j)}$ 表示第 j 类中第 m 个样本对应前 i 个特征的特征向量.

定义 2　设文献样本 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$,$x_{ik}$ 为样本 $\boldsymbol{x}_i$ 第 k 个主题的概率值,文献样本 $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \cdots, x_{jk})$,$x_{jk}$ 为样本 $\boldsymbol{x}_j$ 第 k 个主题的概率值,本文使用余弦距离定义文献 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的相似度:

$$\mathrm{Dis}_{ij} = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\| \boldsymbol{x}_i \| \cdot \| \boldsymbol{x}_j \|} = \frac{\sum_{i=1,k=1}^{k} x_{ik} x_{jk}}{\sqrt{\sum_{i=1,k=1}^{k} (x_{ik})^2} \cdot \sqrt{\sum_{i=1,k=1}^{k} (x_{jk})^2}}. \quad (2)$$

基于 BFS 的序列前向特征选择算法设计如下:

设 $F = \{f_i \mid i = 1, \cdots, k\}$ 为全部特征组成的集合,X 为被选择特征组成的子集;L 为训练集且初始为空集,即 $L = \varnothing$;基于 BFS 评价准则,计算特征均匀度最大的 k 个特征,即 $X = \{f_1, f_2, \cdots, f_k\}$;在 $X = \{f_1, f_2, \cdots, f_k\}$ 上,基于网格划分算法从全部数据集 D 提取初始样本集,与标准对照集 A 比对结果,得到标记数据集 U;令 $L = L \cup U$;根据式(1),使用数据集 D 和训练集 L 计算 F 中每个特征的 BFS;选择最重要的特征 $f_{\max} = \max\{f_i, i = 1, \cdots, k\}$,令 $F = F - \{f_{\max}\}$;$X = X \cup \{f_{\max}\}$;max BFS = min.

输入:全部数据集 D,标准对照集 A;

输出:特征子集 X;

```
1. for 1 to k do//选择的特征个数 k 个
2. begin
3.     for each f_i ∈ X do
4.     begin
5.       temp X = X ∪ {f_i};
6.       根据式(1)计算 BFS(temp X);
7.       if BFS(temp X) > max BFS then
8.         begin
9.         max BFS = BFS(temp X);
10.        selected_feature = f_i;
11.      end //end of if
12.    end //end of for
13.     X = temp X;
14.   F = F − selected_feature;
15.  end //end of for
```

3.2　相关样本发现

相关样本发现过程是数据探索的关键步骤,它从还未探索的区域收集样本并识别单一相关样本,通过展示给用户不同数据区域的样本发现相关样本. 本文的方法是使用网格划分技术[7],从每个网格单元发现样本以保证最大程度覆盖全部探索空间.

定义 3　定义参数 η 为具体需要划分的区域数量,即划分等级,代表了划分的粒度级别. 如果将归一化后的属性划分为 η 个等宽区域范围,那么在整个探索空间上将创建 η^d 个网格单元.

定义 4　定义参数 σ 为每个单元格在每个属性上覆盖的范围:

$$\sigma = 100/\eta. \quad (3)$$

设一个网格的虚拟中心为 O,以这个中心点为基准,在每个维度的一定范围内检索一个随机样本.

定义 5　定义参数 τ 为每个单元的采样距离,取决于每个属性域的划分等级. 即

$$\tau = \sigma/2\ . \tag{4}$$

定义 6 每个探索任务在 R 个样本的 d 维空间上执行,即 $R=\{r_1,r_2,\cdots,r_i\}$,其中,r_i 表示在第 i 个网格中获得的样本. 对于 $R\subseteq D$ 一个给定的用户,相关反馈样本集被划分成两类,即感兴趣样本集 $R_r\subseteq R$ 和不感兴趣样本集 $R_{ir}\subseteq R$.

基于网格划分方法的样本发现算法设计如下:

输入:探索空间全部样本集合 $D=\{d_1,d_2,d_3,\cdots,d_n\}$,$d$ 个探索属性集 $X=\{f_1,f_2,\cdots,f_k\}$;

输出:相关反馈样本集 R;

设划分等级为 η;函数 U 为用户模型;

1) 对任意 f_i 划分成等宽的 η 个区域,其中 $i=1,2,\cdots,d$,得到 η^d 个网格单元;

2) 对第 i 个网格单元,从 D 中使用式(2),选择出距离虚拟中心 O 最近的样本 d_i;

3) 对 R 中的每一个样本,使用 U 进行标记,得到 R_r 和 R_{ir};

4) 对感兴趣样本集 R_r 中的每一个样本所在的网格单元重复 2) 和 3) 两个步骤;

5) 在步骤 2) 至 4) 中,相关反馈样本集 R 为样本发现阶段反馈给用户标记的样本集,R_r 中每个样本所在的网格构成用户的兴趣区域.

交互式数据探索依赖于决策树分类器来划分用户兴趣区域,通过用户已标记的样本数据对探索空间进行划分. 本文使用 CART 决策树生成分类模型,得到最终用户兴趣区域.

3.3 基于相似度的结果集排序

在用户兴趣区域中,对结果集的排序问题,分为单兴趣区域和多兴趣区域讨论. 兴趣区域中的已标记样本表示用户反馈时标记的感兴趣样本,未标记样本表示用户可能感兴趣的样本.

图 2 所示是在二维空间中用户的单兴趣区域的表示以及样本的分布情况. 单兴趣区域内,排序过程设计如下:

1) 对兴趣区域内所有已标记样本计算几何中心 G;

2) 得到中心后,对所有样本根据式(2)计算与 G 的距离;

3) 根据计算值进行快速排序,得到排序结果.

在多兴趣区域中,当位于不同区域的多个样本拥有相同的用户兴趣相似度时,需要对这些样本进行排序. 图 3 所示是二维空间中用户的多兴趣区域及样本的分布情况.

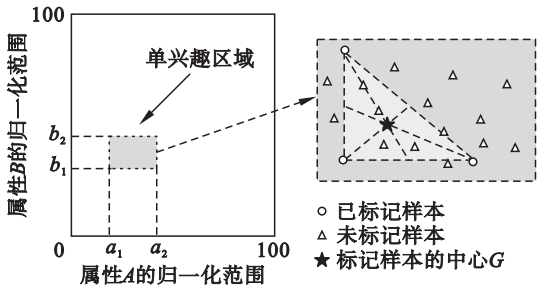

图 2 二维空间中用户的单兴趣区域及样本分布

Fig. 2 User's single interest area and sample distribution in two-dimensional space

定义 7 在多兴趣区域的情况下,对于任意兴趣区域,兴趣区域权重为已标记样本与该区域内所有样本数的比值,区域内样本总数记为 N_{total},已标记样本数记为 N_{marked},则区域权重公式为

$$\omega = N_{\text{marked}}/N_{\text{total}}. \tag{5}$$

因此,对于任意两个文献 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$,加权相似度公式为

$$\text{Dis}_{\omega} = \omega\cdot\text{Dis}_{ij} = \omega\cdot\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\cdot\|\boldsymbol{x}_j\|}\ . \tag{6}$$

多兴趣区域内,排序过程设计如下:

1) 对任意兴趣区域,计算兴趣区域内所有已标记样本的几何中心 G;

2) 对任意兴趣区域,使用式(5)计算区域的兴趣区域权重;

3) 对该区域内每个样本,使用式(6)计算与 G 的加权相似度;

4) 根据计算值进行快速排序,得到多兴趣区域的排序结果.

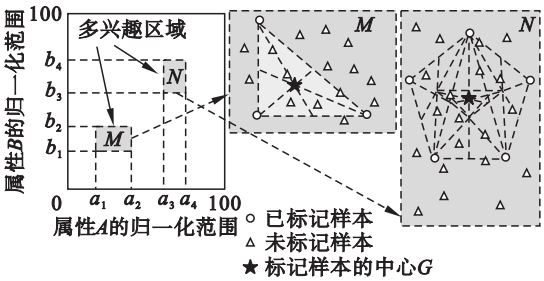

图 3 二维空间中用户的多兴趣区域及样本分布

Fig. 3 User's multi area of interest and sample distribution in two-dimensional space

4 实验验证及结果分析

4.1 数据说明及实验方法

本文实验数据集为 12 种计算机科学类期刊,包括:《软件学报》、《计算机科学与探索》、《计算机研究与发展》等 2012 ~ 2016 年 14 428 篇文献

的题目、作者、发表时间及摘要信息. 本文的数据预处理阶段，通过 LDA 主题模型[9]对文献的主题进行提取，通过交叉验证，根据不同的主题数 k 进行实验，画出 Topic_number – log$p(w|k)$ 曲线[11]，得到后续实验选择主题个数为 30.

实验方法：通过给定一个目标样本集作为标准对照集，每次迭代过程中，使用标准对照集对样本进行标记，达到模拟用户的目的. 根据这个标准对照集，标记每次从迭代过程中提取的新样本集，用户标记的样本是否为感兴趣的结果，取决于这些样本是否包含在标准对照集中；同时使用该标准对照集对最终样本集的准确率、召回率以及 $F1$ 值进行评价.

标准对照集选择方法：通过不同的标准对照集选择方法来模拟不同情况下的迭代目标. 实验通过改变标准对照集的复杂性来模拟以下迭代情况[2]，分别将大、中、小三种区域作为最终兴趣区域，即将该区域中的所有样本作为标准对照集；实验中在各个维度上归一化后，大区域为宽度大于 6 小于 9 的区域，中区域为宽度大于 3 小于 6 的区域，小区域为宽度大于 0 小于 3 的区域.

4.2　BFS 中的参数设置

根据实验数据的特点以及特征选择方法中关于类似参数的设置经验，将参数值设置为：$\lambda = 0.8, \mu = 0.2$；$\lambda = 0.6, \mu = 0.4$；$\lambda = 0.4, \mu = 0.6$；$\lambda = 0.2, \mu = 0.8$. 该实验的方法是在不同维数下，基于 BFS 的特征选择算法 $F1$ 值达到 70% 时，通过标记样本的数量来确定 λ 和 μ 的值.

由图 4 可知，当维度数为 2，3，4 时，μ 取较大值时迭代效率更高. 维度较低时，每轮给用户反馈的样本较少，即在 BFS 中 λ 作为区分度因子的训练数据较少. 当探索属性维数为 5 时，由于每轮迭代用户需要标记更多的样本，BFS 中区分度因子 λ 依赖于训练数据集的大小，故 λ 值较大时效果更好. 因此，本文选择 $\lambda = 0.4, \mu = 0.6$ 为 BFS 的权重系数.

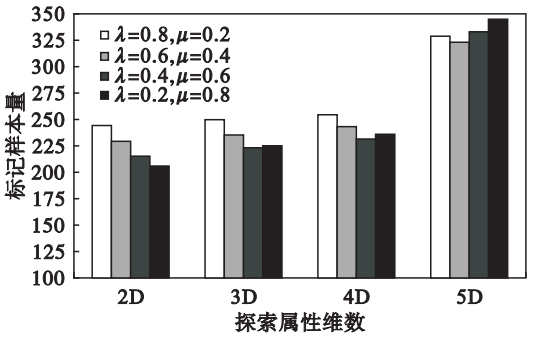

图 4　不同维度下 BFS 中参数 λ 和 μ 对标记样本量的影响

Fig. 4　Influence of parameters λ and μ on BFS in different dimensions

4.3　单目标区域下，标记样本总量与 $F1$ 值的关系实验

图 5 和图 6 分别描述了样本总量为 7 200 和 14 400 时，三种特征选择算法对应的迭代效率对比.

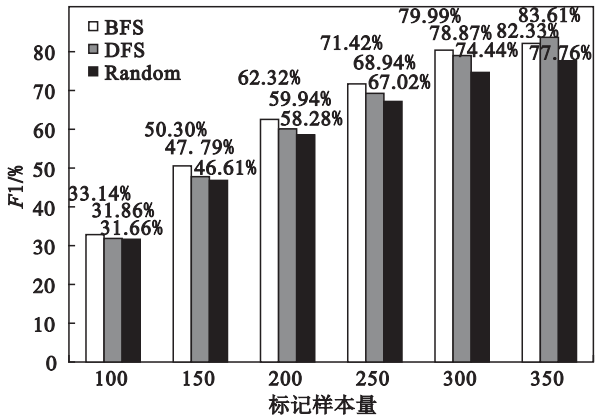

图 5　标记样本总量为 7 200 时三种算法 $F1$ 值的比较

Fig. 5　Comparison of $F1$ of three algorithms in the total number of samples at 7 200

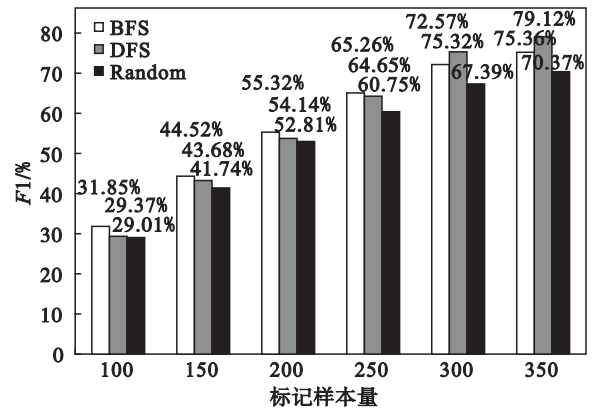

图 6　样本总量为 14 400 时三种算法 $F1$ 值的比较

Fig. 6　Comparison of $F1$ of three algorithms in the total number of samples at 14 400

由图可知，BFS 算法和 DFS 算法在不同数据量下的迭代效率均优于 Random 算法，且 BFS 算法迭代性能略优于 DFS 算法，$F1$ 值可提高 1% ~ 3%. 另外，由于本文提出的 FA – IDE 框架及基于 BFS 的序列前向特征选择算法，提供给用户进行反馈的样本对总体样本具有较好的覆盖度，样本总量的增加只会增加数据分布的密度，但不会对网格空间的收敛速度产生影响. 因此，数据量的成倍增长对迭代性能没有产生显著的影响.

4.4　不同区域下三种算法的迭代效率对比实验

图 7 所示为用户兴趣目标区域为大、中和小

三种规模时,三种特征选择算法达到相同 $F1$ 值的迭代效率. 由图 7 可知,BFS 算法和 DFS 算法总是优于 Random 算法. 在目标区域为大区域时,BFS 算法达到相同 $F1$ 值时,需要标记的样本数量更少且效率更高. DFS 算法依靠训练集进行特征选择,更关注于用户的兴趣,当目标区域较小时,往往需要标记更多样本才能准确寻找到目标区域. 因此,在这种情况下,DFS 算法略好于 BFS 算法.

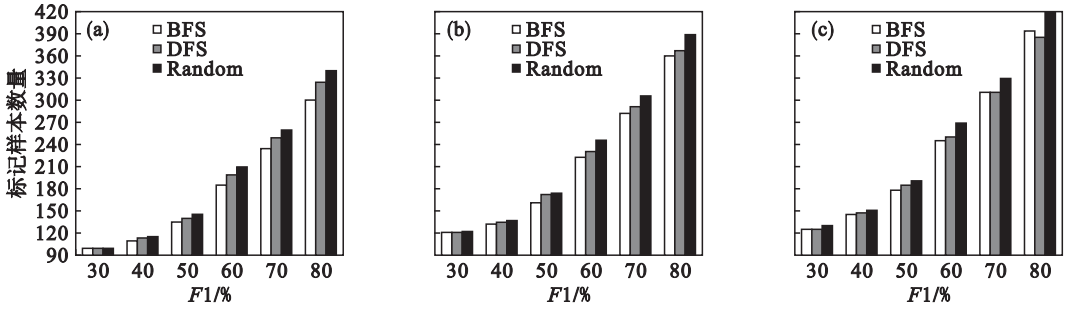

图 7 不同区域下三种算法的迭代效率对比

Fig. 7 Iteration efficiency comparison of three algorithms in different regions

(a)—大区域;(b)—中区域;(c)—小区域.

5 结 语

本文阐述了交互式数据探索的基本概念、存在的问题,提出了一种基于特征自适应的交互式数据探索框架 FA-IDE,可根据用户兴趣多样性需求,每次迭代过程中动态地调整特征子集. 该框架将整个交互式数据探索划分为数据预处理和探索两个阶段. 在探索阶段,基于数据集在特征子集上的分布情况,提出了特征子集的均匀度 BFS 评价准则,根据评价准则,给出了基于 BFS 的序列前向特征选择算法. 其次,针对相关样本发现问题,提出划分等级建立方法,利用决策树模型对用户兴趣区域划分后,提出基于相似度的结果集排序策略. 实验结果表明,与现有的方法相比,本文提出的基于特征自适应的交互式数据探索框架及相应的探索方法可以有效提高用户迭代效率和最终结果的准确性.

参考文献:

[1] 王蒙湘,李芳芳,谷峪,等. 交互式数据探索综述[J]. 计算机科学与探索,2017,11(2):171-184.
(Wang Meng-xiang, Li Fang-fang, Gu Yu, et al. Survey on interactive data exploration [J]. *Computer Science and Exploration*, 2017, 11(2): 171-184.)

[2] Ellermann J, Dorn K. Explore-by-example: an automatic query steering framework for interactive data exploration[C]// ACM SIGMOD International Conference on Management of Data. Snowbird, 2014: 517-528.

[3] Dimitriadou K, Papaemmanouil O, Diao Y. Interactive data exploration based on user relevance feedback [C]// IEEE International Conference on Data Engineering. Atlanta, 2014: 292-295.

[4] Du X Y, Chen J, Chen Y. Research on big data exploration [J]. *Journal of Communication*, 2015, 36(12): 77-88.

[5] Dimitriadou K, Pappaemmanouil O, Diao Y. AIDE: an active learning-based approach for interactive data exploration [J]. *IEEE Transactions on Knowledge & Data Engineering*, 2016, 28(11): 2842-2856.

[6] Kamat N, Jayachandran P, Tunga K, et al. Distributed and interactive cube exploration [C]// IEEE International Conference on Data Engineering. Atlanta, 2014: 472-483.

[7] Agarwal S, Iyer A P, Panda A, et al. Blink and it's done: Interactive queries on very large data [J]. *Proceedings of the VLDB Endowment*, 2012, 5(12): 1902-1905.

[8] Jiang L, Nandi A. SnapToQuery: providing interactive feedback during exploratory query specification [C]// Proceedings of the VLDB Endowment. Kohala Coast, 2015: 1250-1261.

[9] Newman D, Asuncion A U, Smyth P, et al. Distributed inference for latent Dirichlet allocation[C]// Conference on Neural Information Processing Systems. Vancouver, 2007: 1-6.

[10] 谢娟英,谢维信. 基于特征子集区分度与支持向量的特征选择算法[J]. 计算机学报,2014,37(8):1704-1718.
(Xie Juan-ying, Xie Wei-xin. Feature selection algorithm based on feature subset identity and support vector machine [J]. *Journal of Computers*, 2014, 37(8): 1704-1718.)

[11] Griffiths T L, Steyvers M. Finding scientific topics [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101(sup1): 5228-5239.