面向分布式数据流大数据分类的多变量决策树

张 宇 [1, 2] 包研科 [1] 邵良杉 [2] 刘 威 [1]

摘 要 分布式数据流大数据中的类别边界不规则且易变, 因此基于单变量决策树的集成分类器需要较大数量的基分类器才能准确地近似表达类别边界, 这将降低集成分类器的学习与分类性能. 因而, 本文提出了基于几何轮廓相似度的多变量决策树. 在最优基准向量的引导下将 n 维空间样本点投影到一维空间以建立有序投影点集合, 然后通过类别投影边界将有序投影点集合划分为多个子集, 接着分别对不同类别集合的交集递归投影分裂, 最终生成决策树. 实验表明, 本文提出的多变量决策树 GODT 具有很高的分类精度和较低的训练时间, 有效结合了单变量决策树学习效率高与多变量决策树表示能力强的优点.

关键词 分布式数据流, 大数据, 分类, 几何轮廓相似度, 多变量决策树

引用格式 张宇, 包研科, 邵良杉, 刘威. 面向分布式数据流大数据分类的多变量决策树. 自动化学报, 2018, **44**(6): 1115−1127

DOI 10.16383/j.aas.2017.c160809

A Multivariate Decision Tree for Big Data Classification of Distributed Data Streams

ZHANG Yu[1, 2] BAO Yan-Ke[1] SHAO Liang-Shan[2] LIU Wei[1]

Abstract Considering the irregularity and variability of the class boundaries of distributed big data streams, when the univariate decision tree is used as the base classifier in an ensemble classifier, large amounts of base classifiers are needed to accurately approximate class boundaries. This will reduce the learning and classification performance of ensemble classifiers. This article proposes a multivariate decision tree based on geometric outline similarity (GODT). Firstly, by using the optimal reference vector, the n-dimensional data points are projected onto the one-dimensional space, thus a set of ordered projection points are established. Secondly, the set of projection points are divided into several subsets, and the intersections of different subsets are projected and divided by recursive projecting and splitting. Finally, a decision tree is built. Experimental results show that GODT has a better classification accuracy and requires less training time. It combines the high learning efficiency of univariate decision tree algorithm with the strong representation power of multivariate decision tree.

Key words Distributed data streams, big data, classification, outline similarity, multivariate decision tree

Citation Zhang Yu, Bao Yan-Ke, Shao Liang-Shan, Liu Wei. A multivariate decision tree for big data classification of distributed data streams. *Acta Automatica Sinica*, 2018, **44**(6): 1115−1127

分布式和流动性的大数据简称分布式数据流大数据, 广泛存在于大型电子商务网站的交易系统、网络监控系统、传感器网络、股票交易及银行业务等系统, 具有突发性、快速性、无限性、概念漂移、分散性、信息价值稀疏性等特点[1−3], 给此类数据的分类挖掘提出了极大的挑战.

针对数据流的分类问题, 基于集成学习[4−5] 的分类模型是较好的解决方法, 因其具有很高的抗概念漂移能力, 并且分类准确性也较高[6]. 很多集成分类模型采用决策树作为基分类器, 因为决策树学习效率高, 模型简单, 其内在的不稳定性可以提高集成分类模型的多样性[7]. 但是, 面对分布式数据流大数据分类, 现有基于决策树的集成分类模型面临一个急需解决的问题: 此类数据在线到达的数据量大且分散化, 类别的边界呈现易变性和不规则性, 而且这些有价值的边界信息相对于数据总体呈现稀疏性. 然而, 现有基于决策树的集成分类器多采用单变量决策树作为基分类器, 由于单变量决策树只能生成平行于坐标轴的决策边界, 因此需要较大数量的基分类器才能正确地近似表示类别边界, 这使得集成分类模型的学习性能和预测效率降低, 很难适应入侵检测等需要快速预测的应用.

鉴于上述问题, 本文提出了基于几何轮廓相似度的多变量决策树 (Decision tree based on geometric outline similarity, GODT). GODT 可以产生任

收稿日期 2016-12-14 录用日期 2017-04-18
Manuscript received December 14, 2016; accepted April 18, 2017
国家自然科学基金 (71371091) 资助
Supported by National Natural Science Foundation of China (71371091)
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 辽宁工程技术大学理学院 阜新 123000 2. 辽宁工程技术大学系统工程研究所 阜新 123000
1. School of Science, Liaoning Technical University, Fuxin 123000 2. Research Institute of System Engineering, Liaoning Technical University, Fuxin 123000

意角度的决策边界, 相比单变量决策树, 其表示能力更强. 另外, 最小交集分裂准则促使 GODT 可以快速发现类别边界, 而递归投影分裂策略可以有效降低中间节点的分裂次数, 因此 GODT 具有较低的学习时间. 在表示相同决策边界的条件下, 相比单变量决策树, GODT 作为基分类器所需的数量更少, 所以可有效解决因增加基分类器而由此产生的学习与预测性能下降问题. 本文研究的创新点主要包括两点: 1) 提出最小交集分裂准则. 基于类别相似度偏差最大化的方法求解最优基准向量, 使得在最优基准向量的引导下, 不同类别投影点集合的交集最小, 实现了类别归属不确定的样本集合最小化. 2) 提出递归投影分裂策略. 针对父节点中的投影重叠区域, 在其子节点重新计算最优基准向量, 这样可使得重叠区域的样本点经过重新投影之后, 被正确地分离开, 解决了投影重叠区域的分裂问题.

本文组织结构如下: 第 1 节介绍面向数据流分类的基于决策树的集成学习方法的相关研究; 第 2 节介绍经典的多变量决策树算法和几何轮廓相似度函数; 第 3 节详细阐述 GODT, 包括建立属性组合度量标准, 分析算法的原理, 设计和实现 GODT 算法; 第 4 节是实验, 从分类的准确性、训练时间和多样性等几个方面测试 GODT; 第 5 节是总结及后续的研究.

1 相关工作

近些年, 针对数据流的分类挖掘, 研究人员已经提出了很多基于决策树构建集成分类器的方法. Street 等[8] 于 2001 年提出 SEA 算法, 使用连续数据块训练 C4.5 生成彼此独立的基分类器, 并利用启发式的替换策略更新集成分类器来解决大规模数据流分类问题. 2009 年, Bifet 等[9] 提出了两种改进的 Bagging 学习方法: ADWIN Bagging 和 ASHT Bagging, 其中 ASHT Bagging 通过将多个可变尺寸的 HoeffdingTree 进行结合, 进而增加基分类器的多样性. 同年, Polat 等[10] 基于 C4.5 以及一对多 (One-against-all) 分类方法构建集成分类器, 此方法很好地解决了多类分类问题. 2011 年, Wozniak[11] 提出了 iDTt-NGE 学习算法, 利用广义嵌套范式 (Nested generalized exemplar) 算法协同训练多个基分类器 C4.5 来解决概念漂移问题. Abdulsalam 等[12] 通过调整分类模型的参数来优化随机森林的分类效果, 提出了基于随机森林的数据流分类方法, 该方法在多类分类问题中可以达到较高的准确性. 2012 年, Bifet 等[13] 提出了基于 "学习法" 的集成分类方法, 通过在指定维数的每个子空间上训练初级学习器 HoeffdingTree, 利用初级学习器的输出训练 sigmoid 感知器来生成次级学习器, 并验证了该方法在准确率、时间和空间消耗等方面优于 Bagging 方法. 2014 年, Ahmad 等[14] 提出了基于线性多变量决策树构建集成分类器的方法, 利用随机投影 (Random projection) 与随机离散化 (Random discretization) 相结合的方法训练基分类器以增强集成分类器的表示能力. 2016 年, Blaser 等[7] 提出了旋转特征空间的集成学习方法, 基于随机旋转的特征空间训练基分类器来提高集成分类器的多样性, 此方法特别适用于决策树的集成学习. 毛国君等[15] 提出了基于淘汰策略的集成分类方法, 使用微簇集合重构训练数据来训练基分类器 C4.5, 并利用淘汰策略生成集成分类器来解决分布式数据流分类问题.

上述相关工作主要从基分类器和集成学习方法两个角度开展的研究, 为本文的研究提供了可借鉴的理论和方法.

2 理论基础

2.1 多变量决策树

决策树采用自顶向下的归纳学习方法来创建, 分为单变量决策树 (Univariate decision trees)[16−17] 和多变量决策树 (Multivariate decision trees)[18−19]. 与单变量决策树不同, 多变量决策树的分裂判别标准建立在属性组合之上而非单一属性, 进而可以产生任意角度的超平面来分割数据集, 所以表示能力更强, 但同时计算复杂度也更高[14]. 根据属性组合的方式, 多变量决策树分为线性组合多变量决策树、布尔组合多变量决策树和非线性组合多变量决策树等, 其中最常用的是线性组合多变量决策树, 它利用属性的线性组合作为检验标准.

$$\sum_{i=1}^{k} a_i x_i \le C \tag{1}$$

其中, x_i 为第 i 个属性, a_i 为属性 x_i 的权重, C 为阈值.

多变量决策树同单变量决策树一样也会出现生长过盛的问题, 因此需要对其剪枝. 剪枝策略主要分为事前剪枝 (Pre-pruning) 和事后剪枝 (Post-pruning) 两种类型[20−26].

2.2 几何轮廓相似度函数

几何轮廓相似度函数[27] 是建立在伽罗华群 – 单纯型理论之上的多维对象相似性度量方法, 其主要思想是通过计算多维对象的几何轮廓相似度来反映多维对象的亲疏关系.

设多元随机向量 $X = (x_1, x_2, \cdots, x_m)$, 其中 x_j 是向量 X 的第 j 个特征变量 $(j = 1, 2, \cdots, m)$. 设 X 的观测样本集 $T = (X_1, X_2, \cdots, X_n)$, 称 X_i

为 T 的第 i 个样本 $(i=1,2,\cdots,n)$, 令 $X_i=(x_{i1}, x_{i2}, \cdots, x_{im})$.

定义 1. 点集 $Q_i=\{(j,x_{ij})|j=1,2,\cdots,m\}$ 称为样本 X_i 的几何轮廓.

定义 2. 样本 X_k 与 X_s 的几何轮廓相似度函数:

$$\rho_{ks}=\frac{1}{m}\sum_{j=1}^{m}\sqrt{\left(1-\frac{|d_{kj}-d_{sj}|}{R_j}\right)\left(1-\frac{|x_{kj}-x_{sj}|}{W_j}\right)} \tag{2}$$

其中,

$$d_{ij}=\begin{cases}x_{ij+1}-x_{ij}, & 1\le j\le m-1\\ x_{im}-x_{i1}, & j=m\end{cases} \tag{3}$$

$$R_j=\max_{i=1,\cdots,n}\{d_{ij}\}-\min_{i=1,\cdots,n}\{d_{ij}\} \tag{4}$$

$$W_j=\max_{i=1,\cdots,n}\{x_{ij}\}-\min_{i=1,\cdots,n}\{x_{ij}\} \tag{5}$$

给定 3 个样本 X_1, X_2 和 X_3, X_1 与 X_2 的相似度值记为 ρ_{12}, X_3 与 X_2 的相似度值记为 ρ_{32}. 若 $\rho_{12}>\rho_{32}$, 则样本 X_1 与 X_2 更相似; 若 $\rho_{12}<\rho_{32}$, 则样本 X_3 与 X_2 更相似. 因此, 几何轮廓相似度函数为构建多变量决策树提供了一种有效的非线性降维投影方法, 并且降维后的一元变量 ρ_{ij} 具有可解释性: 由于树中非叶子节点的每个分支弧都代表一个相似度区间, 且其对应的分支节点代表与父节点的基准向量相似度介于此区间的样本子空间, 因此非叶子节点的每一种划分都具有可解释性, 这使得基于几何轮廓相似度的多变量决策树具有可解释性.

3 基于几何轮廓相似度的多变量决策树 (GODT)

3.1 属性组合方法及其度量标准

定义 3. 对于给定的样本集 T (见第 2.2 节), 点集 $Q_i=\{(j,f(x_j))|j=1,2,\cdots,m\}$ ($f(x_j)$ 为随机变量 x_j 到实数域 $\mathbf{R}$ 的映射) 称为样本集 T 的基准几何轮廓; 多元随机向量 $S=(f(x_1),f(x_2),\cdots,f(x_m))$ 称为样本集 T 的基准向量.

利用式 (2) 计算样本 X_i 与基准向量 S 的相似度, 记为 Y_i, 如果 X_i 与 Y_i 一一映射, 之后只需建立基于 Y_i 的度量标准, 因为基于 Y_i 的度量标准与基于 X_i 的度量标准是等价的.

定理 1. 给定一个样本集 T (样本数为 n), $X_i=(x_{i1}, x_{i2},\cdots,x_{im})$ 为 T 的第 i 个样本, $S=(f(x_1), f(x_2),\cdots,f(x_m))$ 为 T 的基准向量, $Y_i=\rho(X_i)=\frac{1}{m}\sum_{j=1}^{m}\mathrm{sqrt}(1-\frac{|d_{ij}-d_{0j}|}{R_j})(1-\frac{|x_{ij}-f(x_j)|}{W_j})$, 其中 d_{ij} 见式 (3),

$$d_{0j}=\begin{cases}f(x_{j+1})-f(x_j), & 1\le j\le m-1\\ f(x_m)-f(x_1), & j=m\end{cases}$$

R_j 和 W_j 见式 (4) 和 (5), 则 X_i 与 Y_i 一一映射.

证明. 显然 ρ 是满射. 只需证 ρ 是单射, 即对任意 $X_i\ne X_k$, 均有 $Y_i\ne Y_k$, 包括三种情况:

1) 假定不等式 $f(x_j)>x_{ij}, x_{kj}$ 成立. 由 $X_i\ne X_k$, 不妨设 $f(x_1)>x_{i1}>x_{k1}>0$, $x_{ij}=x_{kj}$, $j=2, 3,\cdots,m$. 此时假设

$$Y_i-Y_k=\frac{1}{p}\left(1-\frac{|d_{i1}-d_{01}|}{R_1}\right)\left(1-\frac{|x_{i1}-f(x_1)|}{W_1}\right)-\left(1-\frac{|d_{k1}-d_{01}|}{R_1}\right)\left(1-\frac{|x_{k1}-f(x_1)|}{W_1}\right)=0 \tag{6}$$

即

$$\left(1-\frac{|d_{i1}-d_{01}|}{R_1}\right)\left(1-\frac{|x_{i1}-f(x_1)|}{W_1}\right)=\left(1-\frac{|d_{k1}-d_{01}|}{R_1}\right)\left(1-\frac{|x_{k1}-f(x_1)|}{W_1}\right) \tag{7}$$

$$\frac{1-\dfrac{|d_{i1}-d_1|}{R_1}}{1-\dfrac{|d_{k1}-d_1|}{R_1}}=\frac{1-\dfrac{|x_{k1}-f(x_1)|}{W_1}}{1-\dfrac{|x_{i1}-f(x_1)|}{W_1}} \tag{8}$$

由 $f(x_1)>x_{i1}>x_{k1}$, 可知 $|x_{k1}-f(x_1)|>|x_{i1}-f(x_1)|$, 进而

$$1-\frac{|x_{k1}-f(x_1)|}{W_1}<1-\frac{|x_{i1}-f(x_1)|}{W_1} \tag{9}$$

所以

$$1-\frac{|d_{i1}-d_{01}|}{R_1}<1-\frac{|d_{k1}-d_{01}|}{R_1} \tag{10}$$

即

$$|d_{i1}-d_{01}|>|d_{k1}-d_{01}|,\ |x_{i2}-x_{i1}|>|x_{k2}-x_{k1}|$$

由假设 $x_{i2}=x_{k2}$, 可得 $x_{i1}<x_{k1}$, 与已知 $x_{i1}>x_{k1}$ 矛盾, 所以 $Y_i\ne Y_k$, ρ 为单射.

2) 同理可证, 不等式 $f(x_j)<x_{ij}, x_{kj}$ 成立时 ρ 也是单射.

3) 若不等式 $x_{ij}<f(x_j)<x_{kj}$ 成立, 且 $f(x_1)=f(x_2)=\cdots=f(x_m)$, $R_1=R_2=\cdots=R_m$, $W_1=W_2=\cdots=W_m$, X_i 与 X_k 关于 S 对称时, 有 $X_i\ne X_k$, 使 $Y_i=Y_k$, 此情况为一种零概率事件. 除此之外, 同理可证 ρ 也是单射.

综上所述, 以概率 1 有 X_i 与 Y_i 一一映射. □

Y_i 的全体构成了一维数轴上的一个点集 $Y = (Y_1, Y_2, \cdots, Y_n)$. 现对 Y 中的元素从小到大排序, 得到有序集合 P ($P = \{Y_i | Y_j > Y_i\ (n \geq j > i \geq 1)\}$). P 中的每个元素 Y_i 表示样本 X_i 与基准向量 S 的相似程度, 并且从左至右相似程度越来越高. 假设样本集 T 的两个子集分别为 T_1 和 T_2, 且 T_1 中的样本同属类别 C_1; T_2 中的样本同属类别 C_2, T_1 中的样本 X_i 与 S 的相似度值 Y_i 构成的集合记为 P_1 ($P_1 \subset P$); T_2 中的样本 X_j 与 S 的相似度值 Y_j 构成的集合记为 P_2 ($P_2 \subset P$). 由于 T_1 中的样本同属类别 C_1, 因此 T_1 中的样本与基准样本 S 的相似程度是相当的, 也就是说 P_1 中的元素位置是相近的 (同处一个邻域), 同理 P_2 中的元素位置也是相近的, 而 P_1 中的元素与 P_2 中的元素是疏远的, 这样 P_1 与 P_2 在一维数轴上便形成了 2 个簇, 如图 1 所示.

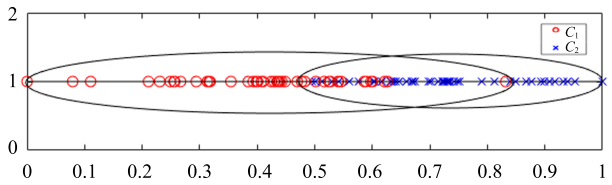


图 1 投影点集合 P_1 与 P_2 的位置关系

Fig. 1 The position relationship of projection point sets P_1 and P_2

P_1 和 P_2 的位置关系有不相交和相交两种. 对于第一种关系 (不相交), 很容易找到一个分裂点将 P_1 和 P_2 完全分割开, 但对于第二种关系 (相交), 这样的分裂点是不存在的, 然而实际上第二种关系是最多的. 针对此种情况, 采取逐步分裂的方法, 首先将 P_1 和 P_2 的差集部分从 P 中分离出去, 之后对交集部分进行逐步分裂, 其中每一步都执行相同的操作 (重新投影、排序、分组和分离非交集部分), 直到交集部分为空. 被分裂的交集部分越小, 则分裂的不确定性越小, 准确性越高, 整体的分裂次数越少, 为了实现被分裂交集部分最小化, 本文使用了变基投影策略. 通过基于类别相似度偏差最大化的方法求解最优基准向量, 使其引导下的不同类别的投影点集合的交集最小, 以实现样本集的最优分裂. 上述逐步分裂的整个过程构成了一棵决策树, 接下来本文介绍基于几何轮廓相似度的多变量决策树.

3.2 基于几何轮廓相似度的多变量决策树 (GODT)

基于几何轮廓相似度的多变量决策树, 在最优基准向量的引导下, 利用几何轮廓相似度函数把 m 维空间下的样本 X_i 投影到一维空间的数轴上, 并对投影点进行排序、分组, 得到一组有序集合 $P_1, P_2, \cdots, P_k$ (k 为类别数目), 计算 $P_1, P_2, \cdots, P_k$ 的交集与差集, 其中交集记为 $T_{p1}, T_{p2}, \cdots, T_{ps}$; 差集记为 $T_{q1}, T_{q2}, \cdots, T_{qt}$. 将 T_{qi} ($i = 1, 2, \cdots, t$) 标记为叶子节点, 而将交集 T_{pj} ($j = 1, 2, \cdots, s$) 标记为中间节点, 之后分别对每个中间节点递归地投影分裂, 直到满足一定条件分裂停止, 最终生成一棵决策树. 该算法在设计过程中需要解决关键问题.

1) 参数 R_j 和 W_j 的确定. 由式 (5) 可知 $W_j = \max_{i=1}^{n}\{x_{ij}\} - \min_{i=1}^{n}\{x_{ij}\}$, 因此需要分别确定 $\max_{i=1}^{n}\{x_{ij}\}$ 和 $\min_{i=1}^{n}\{x_{ij}\}$, 而 $\max_{i=1}^{n}\{x_{ij}\}$ 和 $\min_{i=1}^{n}\{x_{ij}\}$ 是总体 x_j 的最大值和最小值, 所以无法获得实际的取值. 但可以通过训练样本集中的 $\max_{i=1}^{T}\{x_{ij}\}$ 和 $\min_{i=1}^{T}\{x_{ij}\}$ 来估计总体 x_j 的 $\max_{i=1}^{n}\{x_{ij}\}$ 和 $\min_{i=1}^{n}\{x_{ij}\}$. 设 $s_1, s_2, \cdots, s_t$ 是来自随机变量 x_j 的样本, 对应样本值为 $x_{1j}, x_{2j}, \cdots, x_{tj}$, 由于 $s_1, s_2, \cdots, s_t$ 相互独立, 且 $t \geq 30$ 时, $x_j \sim \mathrm{N}(\mu, \sigma^2)$, 则 x_j 的似然函数为

$$L(\mu, \sigma^2) = \prod_{i=1}^{T} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

求解得 μ, σ^2 的极大似然估计量为 $\hat{\mu} = \bar{x}_j$, $\hat{\sigma}^2 = \frac{1}{t}\sum_{i=1}^{T}(x_{ij} - \bar{x}_j)^2$. 令 $M = \max(s_1, s_2, \cdots, s_t)$, $N = \min(s_1, s_2, \cdots, s_t)$, 则

$$\begin{aligned} P\{M \leq z\} &= P\{s_1 \leq z, s_2 \leq z, \cdots, s_t \leq z\} = \\ & P\{s_1 \leq z\}P\{s_2 \leq z\}\cdots P\{s_t \leq z\} = (F(z))^n \end{aligned} \tag{12}$$

$$\begin{aligned} P\{N \leq u\} &= 1 - P\{N > u\} = \\ & 1 - P\{s_1 > u, s_2 > u, \cdots, s_t > u\} = \\ & 1 - P\{s_1 > u\}P\{s_2 > u\}\cdots \\ & P\{s_t > u\} = 1 - (1 - F(u))^n \end{aligned} \tag{13}$$

若 $P\{M \leq z\} = P\{N \leq u\} = \alpha$, 则 $\Phi((z - \mu)/\sigma) = \alpha^n$, $\Phi((u - \mu)/\sigma) = 1 - (1 - \alpha)^n$, 进而求出置信水平为 α 时的总体最大值 z 和最小值 u, 于是可得 $W_j = z - u$.

由式 (4) 可知,

$$R_j = \max_{i=1,\cdots,n}\{d_{ij}\} - \min_{i=1,\cdots,n}\{d_{ij}\}$$

又由于

$$\max_{i=1,\cdots,n}\{d_{ij}\} \leq \max_{i=1,\cdots,n}\{x_{ij+1}\} - \min_{i=1,\cdots,n}\{x_{ij}\}$$

而

$$\min_{i=1,\cdots,n}\{d_{ij}\} \leq \min_{i=1,\cdots,n}\{x_{ij+1}\} - \max_{i=1,\cdots,n}\{x_{ij}\}$$

因此

$$R_j \le \max_{i=1,\cdots,n}\{d_{ij}\} \le \left(\max_{i=1,\cdots,n}\{x_{ij+1}\} - \min_{i=1,\cdots,n}\{x_{ij}\}\right) - \left(\min_{i=1,\cdots,n}\{x_{ij+1}\} - \max_{i=1,\cdots,n}\{x_{ij}\}\right)$$

进一步合并, 可得

$$R_j = \max_{i=1,\cdots,n}\{d_{ij}\} \le W_{j+1} + W_j$$

2) 最优基准向量的构建. $P_i \cap P_j$ $(i \neq j)$ 表示数轴上的某一区间 $[a, b]$, 区间内的投影点既有属于类别 C_i, 也有属于类别 C_j, 因此该区间内样本的类别归属是不确定的. 如果 $P_i \cap P_j$ $(i \neq j)$ 越大, 则 T 中样本类别归属的不确定性越高, 不可分的样本越多, 反之则越小. 通过改变基准向量 S, 可以减小 $P_i \cap P_j$ $(i \neq j)$, 若基准向量 S 与 C_i 中样本的相似度越高, 而同时与 C_j 中样本的相似度越低, 则 S 引导下的两类投影点集合 P_i 与 P_j 的交集越小, 当 P_i 与 P_j 的交集最小时, S 与 C_i, C_j 两类的相似度差异性最大, 此时 S 即为最优基准向量 S_b.

设 C_i 的均值向量为 $\overline{X} = (\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_m)^{\mathrm{T}}$, C_j 的均值向量为 $\overline{Y} = (\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_m)^{\mathrm{T}}$, $\overline{X}$ 与 $\overline{Y}$ 的偏差向量为 $d = (a_1, a_2, \cdots, a_i, \cdots, a_m)^{\mathrm{T}}$, $a_i = \overline{x}_i - \overline{y}_i$, 令 $v = (w_1a_1, w_2a_2, \cdots, w_ma_m)^{\mathrm{T}}$, 其中 w_1, w_2, $\cdots$, w_m 为权重系数, 基准向量 $S = \overline{X} + v$, 则 S 与 $\overline{X}$, $\overline{Y}$ 的相似度偏差为

$$f = \rho(\overline{X}) - \rho(\overline{Y}) = \sum_{j=1}^{m}\left[\left(1 - \frac{|w_{j+1}a_{j+1} - w_ja_j|}{R}\right)^{\frac{1}{2}} \times \left(1 - \frac{|w_ja_j|}{W}\right)^{\frac{1}{2}} - \left(1 - \frac{|(w_{j+1}+1)a_{j+1} - (w_j+1)a_j|}{R}\right)^{\frac{1}{2}} \times \left(1 - \frac{|(w_j+1)a_j|}{W}\right)^{\frac{1}{2}}\right] \tag{14}$$

若 $w_1, w_2, \cdots, w_m$ 取一组值使得 f 最大, 则 $S_b = S = \overline{X} + v$ 或 $S_b = S = \overline{Y} - v$. 接下来只需求解如下非线性方程组即可求得权重系数 $w_1, w_2, \cdots, w_m$, 进而得到最优基准向量 S_b.

$$\begin{cases} \dfrac{\partial f}{\partial w_1} = 0 \\ \dfrac{\partial f}{\partial w_2} = 0 \\ \quad\vdots \\ \dfrac{\partial f}{\partial w_m} = 0 \end{cases} \tag{15}$$

由于上述的相似度偏差 f 中存在开方和绝对值运算, 所以其偏导运算复杂, 且构成的方程组不易求解, 为了解决这一问题, 本文利用函数

$$g(w_1, w_2, \cdots, w_m) = \sum_{j=1}^{m}\left[\left(1 - \frac{(w_{j+1}a_{j+1} - w_ja_j)^2}{R_j^2}\right) \times \left(1 - \frac{w_j^2a_j^2}{W_j^2}\right)\right] - \left[\left(1 - \frac{[(w_{j+1}+1)a_{j+1} - (w_j+1)a_j]^2}{R_{j+1}^2}\right) \times \left(1 - \frac{(w_j+1)^2a_j^2}{W_{j+1}^2}\right)\right] \tag{16}$$

来度量基准向量 S 与 $\overline{X}$, $\overline{Y}$ 的相似度偏差, 并称其为相似差异度. 然后利用相似差异度 g 对 w_j $(j = 1, 2, \cdots, m)$ 求偏导并建立方程组, g 对 w_j 的偏导数如下:

$$\begin{aligned} \frac{\partial g}{\partial w_j} = {} & R_j^2W_j^2[2a_{j+1}W_{j-1}^2(a_{j+1} - a_j) + \\ & 6a_j^3a_{j+1}w_{j-1}^2 - 2a_j^2a_{j+1}^2w_{j-1}^2 + \\ & 6a_j^3a_{j+1}w_{j-1} - 4a_j^2a_{j+1}^2w_{j-1}w_j - \\ & 2a_j^2a_{j+1}^2w_j - 4a_j^2a_{j+1}^2w_{j-1} + 2a_j^3a_{j+1} - \\ & 2a_j^2a_{j+1}^2] + R_{j-1}^2W_{j-1}^2[(6a_j^3a_{j+1} + \\ & 2a_j^4)w_j^2 + 6a_j^3a_{j+1}w_{j+1} - 4a_j^2a_{j+1}^2w_{j+1} + \\ & 12a_j^3a_{j+1}w_j - 12a_j^4w_j - 2a_j^2a_{j+1}^2w_j + \\ & (12a_j^3a_{j+1} - 4a_j^2a_{j+1}^2)w_jw_{j+1} - \\ & 2a_j^2a_{j+1}^2w_{j+1}^2 + 2a_j^2R_j^2 - \\ & 2a_jW_j^2(a_{j+1} - a_j) + 2a_j^3(a_{j+1} - a_j) - \\ & 2a_j^2(a_{j+1} - a_j)^2] \end{aligned} \tag{17}$$

从式 (17) 可以看出, $\frac{\partial g}{\partial w_j} = 0$ 为一元二次方程, 因此, 其构成的方程组为非线性方程组.

$$\begin{cases} f_1(w_n, w_1, w_2) = \dfrac{\partial g}{\partial w_1} = 0 \\ f_2(w_1, w_2, w_3) = \dfrac{\partial g}{\partial w_2} = 0 \\ \quad\vdots \\ f_m(w_{m-1}, w_m, w_1) = \dfrac{\partial g}{\partial w_m} = 0 \end{cases} \tag{18}$$

利用多变量 Newton 方法求解上述方程组, 设 $F(x) = (f_1, f_2, \cdots, f_m)$, $x = (w_1, w_2, \cdots, w_m)$, 则 Jacobi 矩阵 $DF(x)$ 为

$$\begin{pmatrix} \dfrac{\partial f_1}{\partial w_1} & \dfrac{\partial f_1}{\partial w_2} & 0 & \cdots & 0 & \dfrac{\partial f_1}{\partial w_m} \\ \dfrac{\partial f_2}{\partial w_1} & \dfrac{\partial f_2}{\partial w_2} & \dfrac{\partial f_2}{\partial w_3} & 0 & \cdots & 0 \\ 0 & \dfrac{\partial f_3}{\partial w_2} & \dfrac{\partial f_3}{\partial w_3} & \dfrac{\partial f_3}{\partial w_4} & 0 & \cdots \\ \cdots & 0 & \dfrac{\partial f_4}{\partial w_3} & \dfrac{\partial f_4}{\partial w_4} & \dfrac{\partial f_4}{\partial w_5} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial f_m}{\partial w_1} & 0 & \cdots & 0 & \dfrac{\partial f_m}{\partial w_{m-1}} & \dfrac{\partial f_m}{\partial w_m} \end{pmatrix} \tag{19}$$

其中,

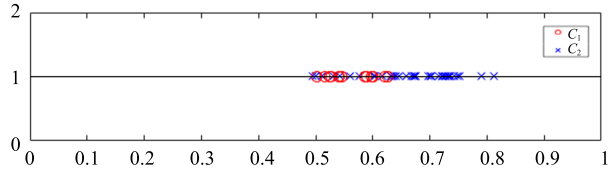
$$\begin{aligned} \frac{\partial f_i}{\partial w_{i-1}} &= R_i^2 W_i^2 [12a_j^3 a_{j+1} w_{i-1} - \\ &\quad 4a_j^2 a_{j+1}^2 (w_{i-1} + w_i) + \\ &\quad 6a_j^3 a_{j+1} - 4a_j^2 a_{j+1}^2] \\ \frac{\partial f_i}{\partial w_i} &= R_{i-1}^2 W_{i-1}^2 [12a_j^3 a_{j+1}(w_{i+1} + w_i) - \\ &\quad 4a_j^2 a_{j+1}^2 w_{i+1} - 24a_j^4 w_i + \\ &\quad 12a_j^3 a_{j+1} - 12a_j^4 - 2a_j^2 a_{j+1}^2] - \\ &\quad R_i^2 W_i^2 (4a_j^2 a_{j+1}^2 w_{i-1} + 2a_j^2 a_{j+1}^2) \\ \frac{\partial f_i}{\partial w_{i+1}} &= R_{i-1}^2 W_{i-1}^2 [12a_j^3 a_{j+1} w_i - \\ &\quad 4a_j^2 a_{j+1}^2 (w_i + w_{i+1}) + \\ &\quad 6a_j^3 a_{j+1} - 4a_j^2 a_{j+1}^2] \end{aligned} \tag{20}$$

接下来迭代求解如下矩阵方程即可求解权重系数 w_1, w_2, $\cdots$, w_m, 最后根据 C_i, C_j 所含样本的数量来确定最优基准向量 S_b.

$$\begin{cases} DF(x_k)t = -F(x_k) \\ x_{k+1} = x_k + t \end{cases} \tag{21}$$

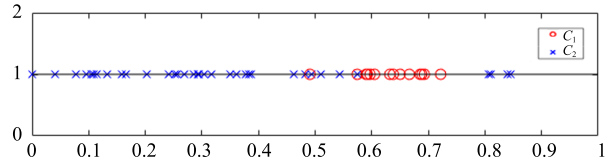
$$S_b = \begin{cases} \overline{X} + v, & n_i > n_j \\ \overline{Y} - v, & n_i \le n_j \end{cases} \tag{22}$$

虽然在最优基准向量 S_b 的引导下 $P_i \cap P_j$ $(i \neq j)$ 最小, 但是 $P_i \cap P_j \neq \phi$ $(i \neq j)$. 如果根据投影点的类别归属直接分裂 $P_i \cap P_j$ $(i \neq j)$, 将导致分裂不均衡和分裂效率降低, 分裂后的每一个子集含有极少相同类别样本 (极端情况, 甚至只含一个样本). 如此分裂直到终止, 最后生成的决策树非常复杂且过拟合严重. 为了实现 $P_i \cap P_j$ $(i \neq j)$ 最优分裂, 本文提出了递归投影分裂方法, 主要思想是重新计算投影点集 $P_i \cap P_j$ $(i \neq j)$ 对应的样本集合的最优基准向量 S_b^0, 使得在 S_b^0 的引导下, 父节点中交叉排列的不同类别样本点在子节点经过重新投影之后被成功的分离开, 如图 2 所示. 因此, 不同类别投影点集合的交集进一步缩小, 然后采用相同的方法对产生的交集递归地投影分裂, 直到满足一定终止条件, 这样便可实现对 $P_i \cap P_j$ $(i \neq j)$ 最大程度地正确分裂.



(a) 重新投影之前两类投影点集合的交集

(a) The intersection of two kinds of projection point set before re-projecting



(b) 重新投影之后两类投影点集合的交集

(b) The intersection of two kinds of projection point set after re-projecting

图 2 两类投影点集合的交集

Fig. 2 The intersection of two kinds of projection point sets

3) 终止策略. GODT 采用了两种终止策略: a) 若当前节点样本集所含样本同属一类, 则分裂停止, 并把该节点标记为叶子节点. b) 若当前节点样本集所含样本分属多类且所含样本数量超过阈值, 则分裂停止, 并把该节点标记为叶子节点, 该节点的类别标签与其父节点中包含样本最多的类别标签相同.

4) 剪枝. 虽然递归投影可以改善树的结构, 降低过拟合的程度, 但是过拟合不能完全消除, 为了进一步降低过拟合, 本文引入了事后剪枝策略. 现有的事后剪枝策略有很多, 其中经典的方法包括 Error-complexity pruning, Critical value pruning, Minimum-error pruning, Reduced-error pruning, Pessimistic pruning 和 Error-based pruning, 文献 [21−22] 对上述方法进行了详细的对比分析, 本文

选择了 Pessimistic pruning 作为 GODT 的剪枝方法, 因为此方法不会出现过度剪枝, 其剪枝后决策树的平均错误率与 C4.5 算法中使用的 Error-based pruning 方法相差甚微, 然而其运行时间更低.

基于几何轮廓相似度的多变量决策树算法流程如下:

算法 1. 基于几何轮廓相似度的多变量决策树算法

Input:
T : 训练样本集
$Minobj$: 节点最小样本数
R : 根节点
Output:
$Tree$: GODT 决策树
Function GODT (T, $Minobj$, R)
Begin
if (R 的所有样本同属类别 C)
标记 R 为叶子节点, 其类别为 C;
Return;
else if (R 的样本数小于 $Minobj$)
标记 R 为叶子节点, 其类别为 R 的父节点包含样本数最多的类别;
Return
else
计算最优基准向量 S_b;
利用式 (2) 计算有序投影点集合 $P_1, P_2, \cdots, P_j$;
计算 $P_1, P_2, \cdots, P_j$ 的差集与交集, 分别记为:
$T_{q1}, T_{q2}, \cdots, T_{qt}$;
$T_{p1}, T_{p2}, \cdots, T_{ps}$;
for each T_{qi} **in** $T_{q1}, T_{q2}, \cdots, T_{qt}$
创建 R 的子节点 L_i, 并将其标记为叶子节点;
存储分裂点 Y_i^b;
end for
for each T_{pj} **in** $T_{p1}, T_{p2}, \cdots, T_{ps}$
创建 R 的子节点 M_j;
存储分裂点 Y_j^b;
获取 T_{pj} 的样本集 T_j;
GODT (T_j, $Minobj$, M_j);
end for
end if
End
End Function

通过训练、剪枝, 最终生成一棵 GODT. 接下来, 应用 GODT 对未知类别样本进行分类, 分类过程如下:

步骤 1. 如果当前节点是叶子节点, 则待测样本 X_i 的类别输出为当前节点的类别标签.

步骤 2. 否则计算待测样本 X_i 与当前节点最优基准向量 S_b 的相似度 Y_i, 根据 Y_i 与 Y_j^b 的比较结果, 选择进入相应的子节点, 回到步骤 1.

在每个中间节点, GODT 主要完成以下计算任务:

1) 构建最优基准向量;

2) 计算样例与最优基准向量的几何轮廓相似度;

3) 排序和分裂训练集.

任务 1) 中最耗时的计算为求解权重系数 w_1, w_2, $\cdots$, w_m, 其训练时间复杂度为 $\mathrm{O}(m^3)$; 任务 2) 的训练时间复杂度为 $\mathrm{O}(nm)$ (n 为当前节点的训练样例数量); 任务 3) 采用快速排序方法, 其训练时间复杂度为 $\mathrm{O}(n\log n)$. 于是, GODT 在每个中间节点的总代价为 $\mathrm{O}(m^3)+\mathrm{O}(nm)+\mathrm{O}(n\log n)$. 若 $n \gg m$ 时, 则 GODT 在每个中间节点的训练时间复杂度为 $\mathrm{O}(nm)$; 若 $m \gg n$ 时, 则 GODT 在每个中间节点的训练时间复杂度为 $\mathrm{O}(m^3)$.

4 实验及结果分析

本节在分布式数据流大数据环境下共设计 3 个实验.

实验 1. 测试基于 GODT 的集成分类器 EGODT 的分类精度和训练时间, 并选择相关工作中的基分类器 C4.5[8, 10−11, 15], HoeffdingTree[9, 13] 和 Cart-LC[14] 构建对比集成分类器: EC45, EHoeffdingTree 和 ECart-LC;

实验 2. 在不同的基分类器数量下, 测试 4 种集成分类器的分类精度;

实验 3. 测试基分类器 GODT 的多样化程度, 并与 C4.5, HoeffdingTree 和 Cart-LC 作对比.

4.1 实验环境和数据

本文选择真实数据集 KDDCUP99[28], Record Linkage[28] 和 Heterogeneity Activity[29] 作为测试数据集. 数据的详细信息见表 1.

GODT 算法由 C# 语言编写实现, 并在 Visual Studio 2013 环境下编译执行, 所有实验全部在同一台工作站 (CPU: E5-2620, 内存: 40 GB) 上完成.

表 1 数据集

Table 1 Dataset

Dataset	Number of attributes	Type of attributes	Size	Number of class
KDDCUP99	42	Nominal, Numeric	5 209 460	23
Record Linkage	12	Numeric	4 587 620	2
Heterogeneity Activity	7	Numeric	13 062 475	7

4.2 分布式数据流环境下的测试

为了模拟分布式数据、流式大数据环境, 本文编写了一个生成数据流的软件工具 Stream Generator. Stream Generator 包含 1 个控制参数: 数据流速 dfs. 另外, 为了从局部节点获取训练数据, 本文编写了数据采集工具 Data Collector. Data Collector 包含 1 个控制参数: 滑动窗口大小 wt.

本节实验的分布式环境由 4 个局部节点和 1 个中心节点组成, 利用 VM Workstation 在 1 台工作站上构建 5 台虚拟机, 每台虚拟机分配 4 GB 内存, 利用其中 4 台作为局部节点 (Distributed node), 1 台作为中心节点 (Master node). 训练数据集和测试数据集分布存储在 4 个局部节点, 并在每个局部节点部署 Stream generator 来模拟数据流在线到达的过程以形成窗口数据. 为了更好地模拟数据随机到达的情况, 每个局部节点的数据流速 dfs 会随着时间不断变化, 其变化范围设置为 [1 000, 3 000]. 在中心节点部署 Data collector 和全局分类器, Data collector 为全局分类器收集训练数据, 而全局分类器负责训练分类器并将生成的分类器分发给局部节点, 局部节点利用最新的分类器替换现有分类器, 继续对在线到达的测试数据进行分类.

中心节点的全局分类器为集成分类器, 其更新学习策略为: 设集成分类器的基分类器数量为 n, 假设当前挖掘时间点为 t_i, 上一个挖掘点生成的集成分类器为 $ec_{i-1}=\{c_j^{i-1}|j=1,2,\cdots\}$ (c_j^{i-1} 为 ec_{i-1} 中第 j 个基分类器), 窗口 w_i 下的训练集为 T_i. 若 $m\le n$, 利用 T_i 训练第 $m+1$ 个基分类器 c_{m+1}^i; 若 $m>n$, 利用 T_i 训练 ec_{i-1} 中分类错误率最高的基分类器, 这样做既可以保证基分类器之间的差异性, 又实现了在相同的条件下训练基分类器.

4.2.1 集成分类器的精度测试

在 $n=10$ 的条件下, 随着 wt 的增加, 图 3 给出了 4 种集成分类器的分类精度变化情况.

从图 3 可以看出, 在 KDDCUP99 和 Heterogeneity Activity 数据集上, 随着 wt 的增加, 4 个分类器的精度都在逐步提升, 这是因为训练数据的数量与 wt 成正比, wt 越大, 训练数据越多, 则 4 个分类器学习质量越好. 然而, 这种学习特性在 Record Linkage 数据集上体现不明显, 这是因为此数据集为二分类数据集, 其类别边界相对简单, $wt=5$ 时的训练数据量便满足了绝大部分类别边界的泛化学习, 因此随着 wt 的增加, 分类精度提升幅度很小. 另外, 相比 EC45 和 EHoeffdingTree, EGODT 与 ECart-LC 的分类精度较好, 并且当 $wt=5$ 时, EGODT 的分类精度最高, 而此时的训练数据最少, 这体现了 EGODT 在小训练集环境下的学习优势, 这是因为 GODT 是基于类别边界构建的决策树, 其非叶子节点的每个分支弧都对应着一个分裂边界, 而这些分裂边界是由投影点的类别边界所决定的. 每个投影点代表随机样本与最优基准向量的几何轮廓相似度, 第 3.1 节已经讨论过同属一类的投影点是相近的, 它们所处的区域相对整个数轴是一定的, 相同类别的小样本集与大样本集的投影区域大致相同, 因此小样本集的投影点即可确定类别的近似边界, 这就是 GODT 能够在小训练集环境下获得较高分类精度的原因.

为了进一步说明 GODT 的这种学习特性, 图 4

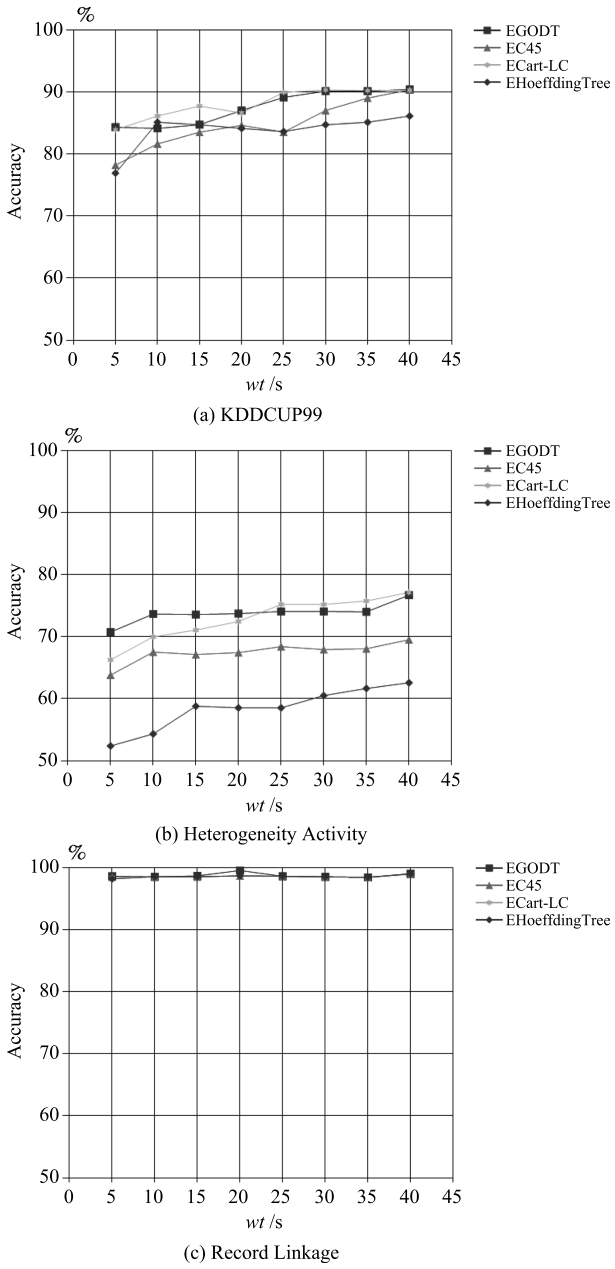
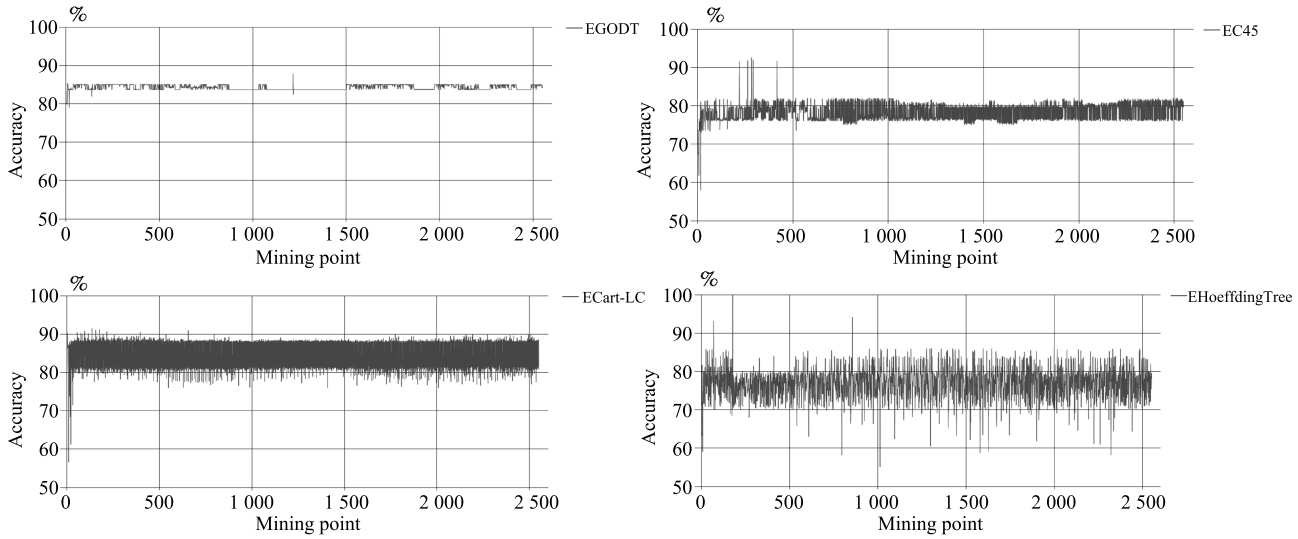
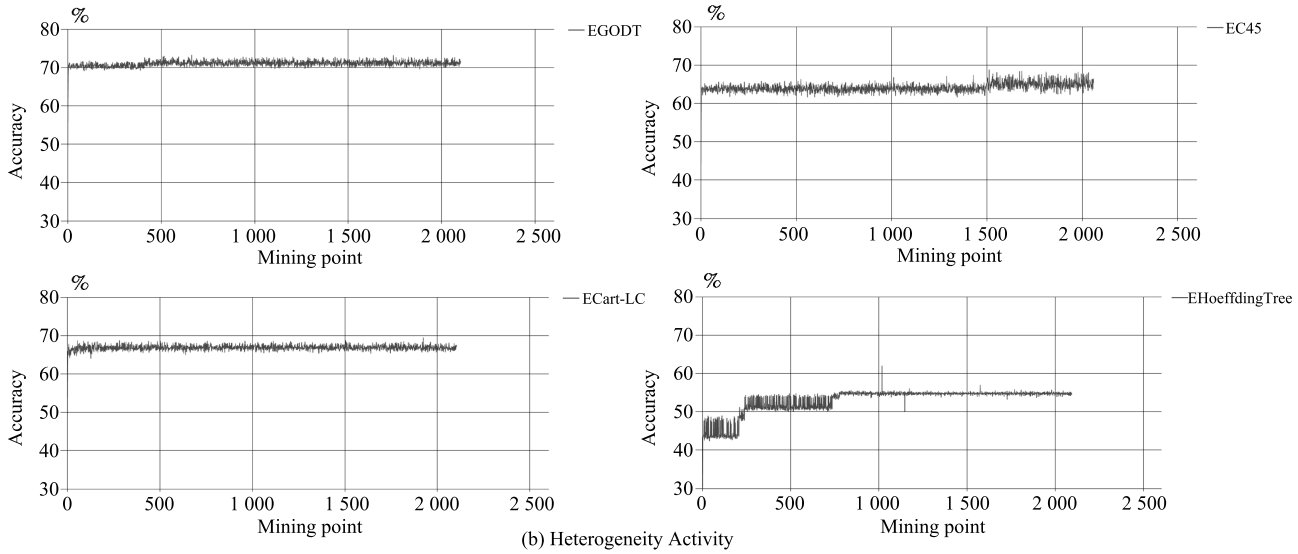


图 3 分类精度随滑动窗口大小的变化情况

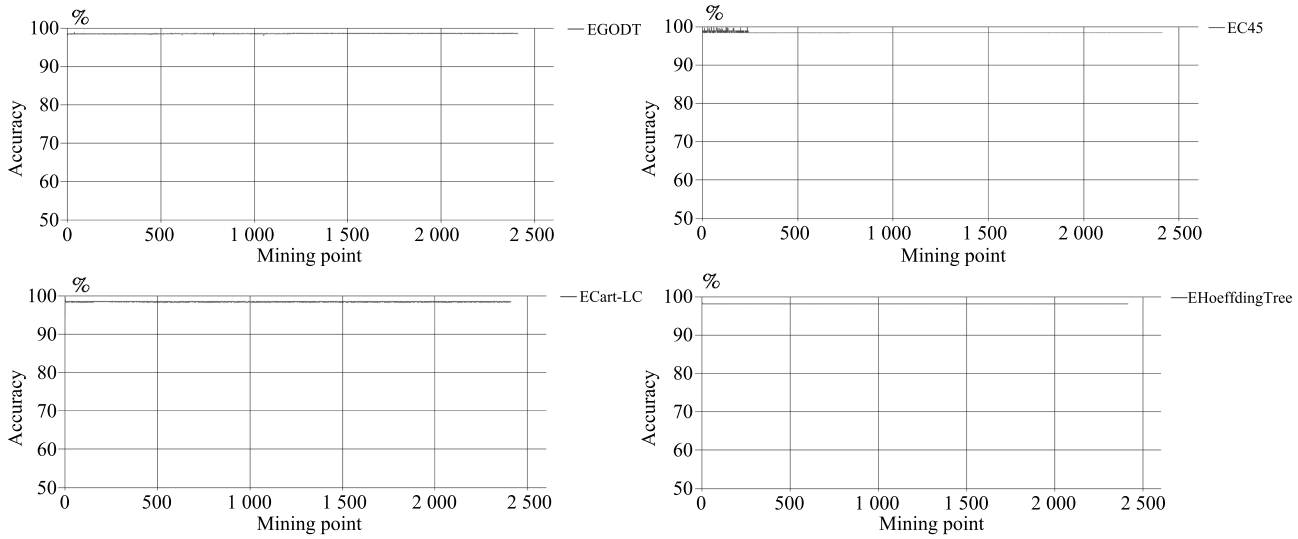
Fig. 3 The variation of classification accuracy with the sliding window size



(a) KDDCUP99



(b) Heterogeneity Activity



(c) Record Linkage

图 4　$wt = 5$ 时, 分类精度在整个挖掘序列的变化情况

Fig. 4　The variation of classification accuracy in the mining sequence when $wt = 5$

给出了 4 种集成分类器在 $wt = 5$ 时的整个挖掘序列的学习过程. 在 KDDCUP99 数据集的挖掘序列初始阶段, 4 种分类器的分类精度均出现了较大的波动, 这是由分类器的更新学习策略所导致的, 因为在初始阶段集成分类器中的基分类器数量少, 并且基分类器之间的差异大, 所以分类精度出现较大波动. 但是, EGODT 的波动范围比其他 3 种算法小很多, 并且随着基分类器数量的增多, EGODT 的波动很快收敛到一个较小的范围, 而且在后续的挖掘序列, 始终保持在这个较小的范围, 而 ECart-LC 的波动也很快收敛到一定的范围, 但波动范围较大; EC45 的波动范围在 700 点之后才趋于稳定; EHoeffdingTree 始终在较大范围波动. 对于 Heterogeneity Activity 数据集的整个挖掘序列, EGODT 和 ECart-LC 的波动范围较小, 而 EC45 在 1 500 点之后波动范围扩大, EHoeffdingTree 的波动范围在 750 点之后才趋于稳定. 由于 Record Linkage 数据集的类别边界简单且变化小, 因此 4 个分类器在整个挖掘序列的分类精度趋于稳定, 波动范围非常小. 图 4 验证了 EGODT 在较小的滑动窗口下可以获得较高的分类性能, 这种特性使得在分布式数据流大数据分类环境下, 能够进一步尝试缩小滑动窗口的大小来适应突发式或渐进式的概念漂移, 以及减少通信代价.

4.2.2 集成分类器的精度测试

由于每次更新学习只生成一个基分类器, 因此集成分类器的单次训练时间即为基分类器的训练时间. 图 5 给出了 4 种集成分类器的训练时间.

从图 5 可以看出, 当 wt 从 5 增加到 40, 4 种集成分类器一次更新学习的时间都会增加, 这是因为如果 wt 增大, 则 wt 时间段内收集的训练数据增多, 因此学习时间加长. 此外, 在 KDDCUP99 数据集上, EGODT 的训练时间小于其他 3 种分类器, 而在 Heterogeneity Activity 数据集上, EGODT 与 EC45 的训练时间相当, 并低于 ECart-LC. 在 Record Linkage 数据集上, EGODT 与 EC45、ECart 的训练时间接近, 并低于 EHoeffdingTree. 综上, GODT 的平均训练时间接近单变量决策树 C4.5, 而低于多变量决策树 Cart-LC, 这是因为 GODT 的递归投影策略使得每个非叶子节点的分裂至少产生一个叶子节点, 因此待分裂示例的数量从根节点便开始减少, 最终导致整体的分裂次数减少, 降低了决策树的生成时间. 同时, 随着学习示例的增多, EGODT 平均爬升幅度低于其他 3 种分类器, 主要是因为增加的绝大部分样例被投影到类别的非交集部分, 只有很少一部分投影到类别的交集部分 (类别边界区域), 因此分裂交集所增加的计算量很少, 进而训练时间增加的幅度很小.

4.2.3 不同基分类器数量下的测试

在 $wt = 30$ 的条件下, 随着 n 的增大, 图 6 给出了 4 种集成分类器的分类精度变化情况.

从图 6 可以看出, 对于 KDDCUP99 数据集, EGODT 与 ECart-LC 在 $n = 5$ 时的分类精度已经接近于 EC45 在 $n = 30$ 的分类精度, 而高于 EHoeffdingTree 在 $n = 30$ 的分类精度. 在 Heterogeneity Activity 数据集上, EGODT 在 $n = 5$ 的分类精度接近于 ECart-LC 在 $n = 10$ 的分类精度以及 EC45 在 $n = 20$ 的分类精度, 而高于 EHoeffdingTree 在 $n = 30$ 的分类精度. 由于 Record

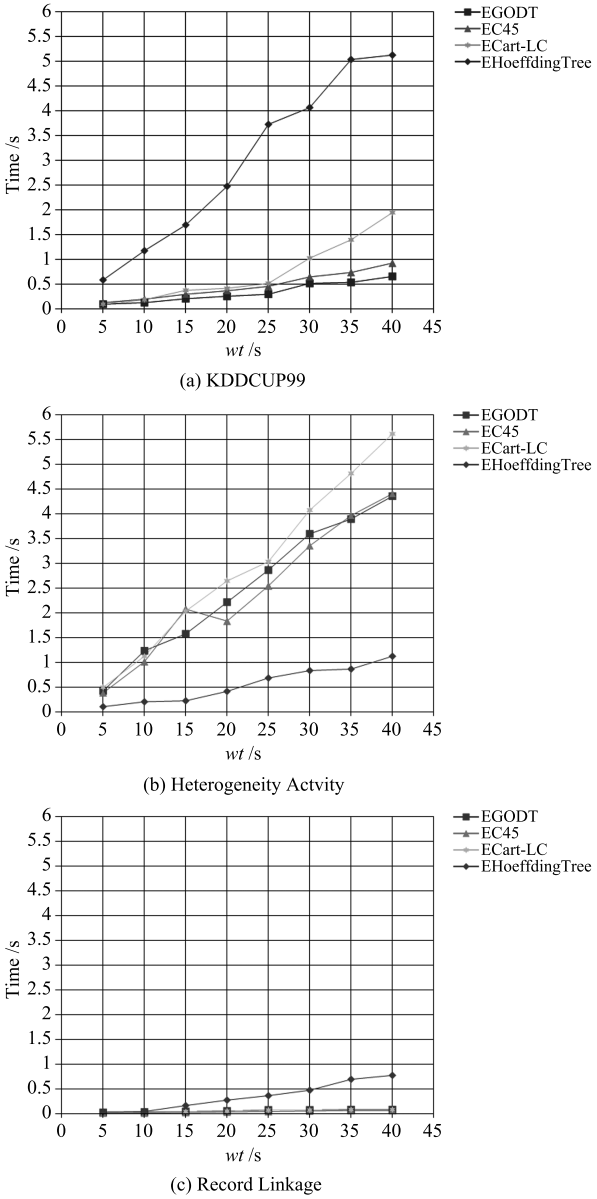


图 5 训练时间随滑动窗口大小的变化情况

Fig. 5 The variation of training time with the sliding window size

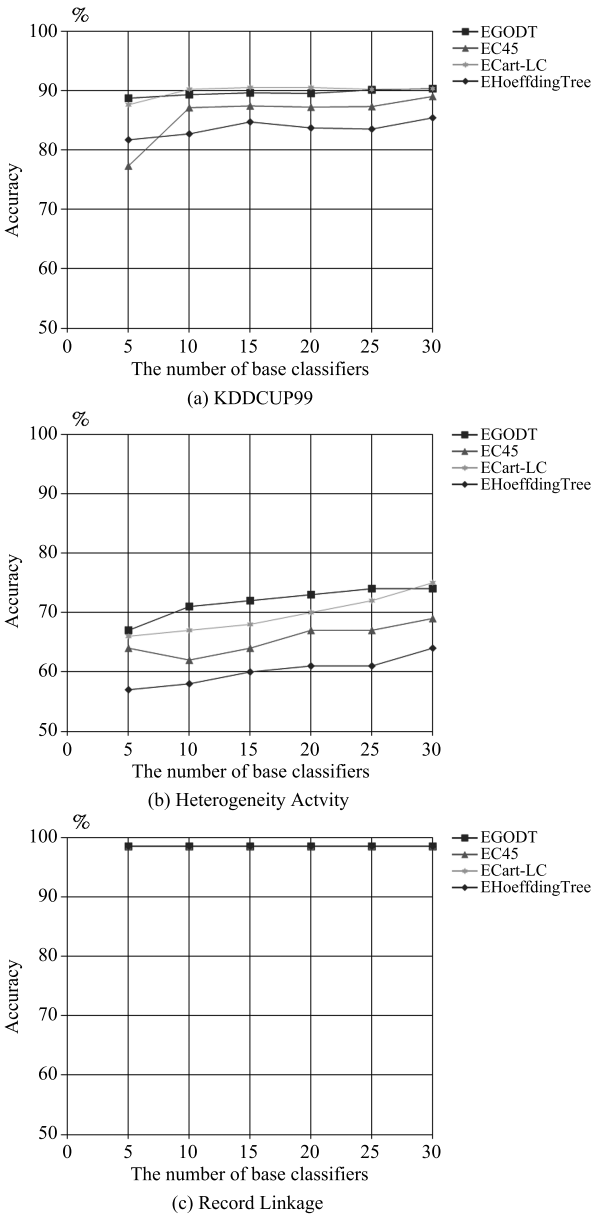


图 6　分类精度随基分类器数量的变化情况

Fig. 6　The variation of classification accuracy with the number of base classifiers

Linkage 数据集的类别边界简单, 因此 4 种分类器在 $n = 5$ 时即可完成绝大部分类别边界的表达, 随着 n 的增大, 4 种分类器的分类精度变化极小. 综上, 图 6 验证了 GODT 与 Cart-LC 的表示能力更强, 在表示相同决策边界时, GODT 与 Cart-LC 所需的数量更少.

4.2.4　基分类器的多样性测试

基分类器多样性的度量指标有很多, 本节选用不合度量 (Disagreement measure)[30] 作为评价指标, 因为它简单、易表述, 其数学描述如下:

$$dis_{ij} = \frac{b + c}{m} \tag{23}$$

其中, m 为样本总数, b 为分类器 i 预测为正例且分类器 j 预测为反例的样本数, c 为分类器 i 预测为反例且分类器 j 预测为正例的样本数.

在 $wt = 10$, $n = 10$ 的条件下, 表 2～5 给出了

表 2　EGODT 的基分类器间的不合度量

Table 2　The disagreement measure between base classifiers of EGODT

GODT	$c1$	$c2$	$c3$	$c4$	$c5$	$c6$	$c7$	$c8$	$c9$	$c10$
$c1$	0	0.43	0.52	0.55	0.51	0.41	0.43	0.43	0.41	0.48
$c2$		0	0.51	0.46	0.6	0.32	0.29	0.39	0.29	0.6
$c3$			0	0.24	0.24	0.56	0.6	0.56	0.6	0.52
$c4$				0	0.35	0.66	0.56	0.67	0.55	0.73
$c5$					0	0.53	0.64	0.52	0.57	0.58
$c6$						0	0.41	0.19	0.13	0.37
$c7$							0	0.12	0.19	0.68
$c8$								0	0.2	0.64
$c9$									0	0.6
$c10$										0

表 3　EC45 的基分类器间的不合度量

Table 3　The disagreement measure between base classifiers of EC45

C4.5	$c1$	$c2$	$c3$	$c4$	$c5$	$c6$	$c7$	$c8$	$c9$	$c10$
$c1$	0	0.31	0.44	0.44	0.35	0.3	0.31	0.31	0.29	0.35
$c2$		0	0.48	0.48	0.53	0.11	0.09	0.09	0.09	0.53
$c3$			0	0.02	0.53	0.55	0.53	0.52	0.51	0.55
$c4$				0	0.52	0.55	0.53	0.53	0.53	0.54
$c5$					0	0.48	0.51	0.53	0.49	0.04
$c6$						0	0.08	0.09	0.07	0.47
$c7$							0	0.02	0.07	0.53
$c8$								0	0.08	0.54
$c9$									0	0.49
$c10$										0

表 4　ECart-LC 的基分类器间的不合度量

Table 4　The disagreement measure between base classifiers of ECart-LC

Cart-LC	$c1$	$c2$	$c3$	$c4$	$c5$	$c6$	$c7$	$c8$	$c9$	$c10$
$c1$	0	0.39	0.31	0.31	0.3	0.37	0.46	0.46	0.32	0.32
$c2$		0	0.22	0.22	0.36	0.57	0.48	0.47	0.48	0.6
$c3$			0	0	0.15	0.61	0.5	0.5	0.5	0.5
$c4$				0	0.15	0.61	0.5	0.5	0.5	0.5
$c5$					0	0.46	0.52	0.52	0.39	0.39
$c6$						0	0.57	0.57	0.2	0.16
$c7$							0	0	0.38	0.5
$c8$								0	0.38	0.5
$c9$									0	0.12
$c10$										0

表 5 EHoeffdingTree 的基分类器间的不合度量

Table 5 The disagreement measure between base classifiers of EHoeffdingTree

HoeffdingTree	$c1$	$c2$	$c3$	$c4$	$c5$	$c6$	$c7$	$c8$	$c9$	$c10$
$c1$	0	0.26	0.53	0.54	0.38	0.3	0.36	0.31	0.24	0.47
$c2$		0	0.42	0.47	0.2	0.18	0.12	0.16	0.1	0.32
$c3$			0	0.45	0.45	0.43	0.58	0.58	0.46	0.45
$c4$				0	0.52	0.44	0.45	0.5	0.47	0.51
$c5$					0	0.54	0.57	0.56	0.58	0.07
$c6$						0	0.08	0.15	0.14	0.51
$c7$							0	0.19	0.12	0.58
$c8$								0	0.25	0.57
$c9$									0	0.56
$c10$										0

4 种基分类器在 3 个数据集上的平均不合度量结果.

从表 2～5 可以看出, GODT 的不合度量样本均值与方差为 $(\overline{dis}, D(dis)) = (0.46, 0.025)$, 而 C4.5, Cart-LC 和 HoeffdingTree 的不合度量样本均值与方差分别为 (0.36, 0.038), (0.40,0.025) 和 (0.38,0.028). GODT 的不合度量样本均值高于其他 3 种分类器, 而且其相邻基分类器的最小不合度量为 0.12, 而 C4.5, Cart-LC 和 HoeffdingTree 相邻基分类器的不合度量分别为 0.02, 0 和 0.08, 说明 GODT 的多样性更强. 另外, GODT 的方差与 Cart-LC 相同, 并低于另外两种分类器, 说明其多样性相对稳定. 因此, 基于 GODT 可以构建泛化能力更强的集成分类器.

5 结论与未来工作

本文提出了一种新的多变量决策树. 利用几何轮廓相似度函数将多维属性合成为一维属性, 建立了非线性属性组合方法, 并在此基础上提出了最小交集分裂准则, 这使得中间节点分裂的不确定性最小化. 此外, 通过对投影分裂过程的分析, 提出了递归投影策略, 并将该策略与最优分裂准则相结合, 形成了一种有效的分裂方法, 降低了中间节点分裂的不均衡性, 简化了决策树的结构. 同其他几种面向数据流的基分类器相比, 在分布式数据流环境下, 本文提出的多变量决策树具有较高的分类准确性和较低的训练时间, 为构建分布式数据流大数据环境下的集成分类模型提供了一种有效的基分类器.

分布式数据流大数据的分类挖掘需要整体的解决方法, 并非单一技术所能解决, 本文算法也仅仅是针对单变量决策树在分布式数据流大数据分类中的表示能力有限来研究对应的解决方法, 除此之外, 分布式数据流大数据的分类挖掘还面临数据类型多样化、大数据的形式化表达、多节点的概念漂移检测、复杂多分布的数据统计与样本重构方法等很多问题, 下一步工作的重点将围绕复杂多分布环境下的数据统计与样本重构方法, 开展相应的理论和算法研究.

References

1 Zhu Qun, Zhang Yu-Hong, Hu Xue-Gang, Li Pei-Pei. A double-window-based classification algorithm for concept drifting data streams. *Acta Automatica Sinica*, 2011, **37**(9): 1077−1084
(朱群, 张玉红, 胡学钢, 李培培. 一种基于双层窗口的概念漂移数据流分类算法. 自动化学报, 2011, **37**(9): 1077−1084)

2 Wu X D, Zhu X Q, Wu G Q, Ding W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(1): 97−107

3 Sun Da-Wei, Zhang Guang-Yan, Zheng Wei-Min. Big data stream computing: technologies and instances. *Journal of Software*, 2014, **25**(4): 839−862
(孙大为, 张广艳, 郑纬民. 大数据流式计算: 关键技术及系统实例. 软件学报, 2014, **25**(4): 839−862)

4 Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997, **55**(1): 119−139

5 Breiman L. Bagging predictors. *Machine Learning*, 1996, **24**(2): 123−140

6 Zhang P, Zhou C, Wang P, Gao B J, Zhu X Q, Guo L. E-tree: an efficient indexing structure for ensemble models on data streams. *IEEE Transactions on Knowledge and Data Engineering*, 2015, **27**(2): 461−474

7 Blaser R, Fryzlewicz P. Random rotation ensembles. *Journal of Machine Learning Research*, 2016, **17**(4): 1−26

8 Street W N, Kim Y. A streaming ensemble algorithm (SEA) for large-scale classification. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2001. 377−382

9 Bifet A, Holmes G, Pfahringer B, Kirkby R, Gavaldá R. New ensemble methods for evolving data streams. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France: ACM, 2009. 139−148

10 Polat K, Güneş. A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 2009, **36**(2): 1587−1592

11 Wozniak M. A hybrid decision tree training method using data streams. *Knowledge and Information Systems*, 2011, **29**(2): 335−347

12 Abdulsalam H, Skillicorn D B, Martin P. Classification using streaming random forests. *IEEE Transactions on Knowledge and Data Engineering*, 2011, **23**(1): 22−36

13 Bifet A, Frank E, Holmes G, Pfahringer B. Ensembles of restricted hoeffding trees. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2012, **3**(2): Article No. 30

14 Ahmad A, Brown G. Random projection random discretization ensembles-ensembles of linear multivariate decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(5): 1225−1239

15 Mao Guo-Jun, Hu Dian-Jun, Xie Song-Yan. Models and algorithms for classifying big data based on distributed data streams. *Chinese Journal of Computers*, 2017, **40**(1): 161−175
(毛国君, 胡殿军, 谢松燕. 基于分布式数据流的大数据分类模型和算法. 计算机学报, 2017, **40**(1): 161−175)

16 Quinlan J R. Induction of decision trees. *Machine Learning*, 1986, **1**(1): 81−106

17 Quinlan J R. *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

18 Breiman L, Friedman J H, Olshen R A, Stone C J. *Classification and Regression Trees*. Belmont, CA, USA: CRC Press, 1984.

19 Brodley C E, Utgoff P E. Multivariate decision trees. *Machine Learning*, 1995, **19**(1): 45−77

20 Ferri C, Flach P A, Hernández-Orallo J. Improving the AUC of probabilistic estimation trees. In: Proceedings of the 2003 European Conference on Machine Learning. Berlin, Heidelberg, Germany: Springer, 2003. 121−132

21 Mingers J. An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 1989, **4**(2): 227−243

22 Esposito F, Malerba D, Semeraro G, Kay J. A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, **19**(5): 476−491

23 Fournier D, Crémilleux B. A quality index for decision tree pruning. *Knowledge-Based Systems*, 2002, **15**(1−2): 37−43

24 Osei-Bryson K M. Post-pruning in decision tree induction using multiple performance measures. *Computers and Operations Research*, 2007, **34**(11): 3331−3345

25 Elomaa T, Kääriäinen M. An analysis of reduced error pruning. *Journal of Artificial Intelligence Research*, 2001, **15**(1): 163−187

26 Quinlan J R. Simplifying decision trees. *International Journal of Man-Machine Studies*, 1987, **27**(3): 221−234

27 Bao Yan-Ke, Zhao Feng-Hua. Measure axiom of outline similarity of multi-scale data and its calculation. *Journal of Liaoning Technical University (Natural Science)*, 2012, **31**(5): 797−800
(包研科, 赵凤华. 多标度数据轮廓相似性的度量公理与计算. 辽宁工程技术大学学报 (自然科学版), 2012, **31**(5): 797−800)

28 Bache K, Lichman M. UCI machine learning repository [Online], available: http://archive.ics.uci.edu/ml, January 1, 2016

29 Stisen A, Blunck H, Bhattacharya S, Prentow T S, Kjaergaard M B, Dey A, Sonne T, Jensen M M. Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition. In: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems. Seoul, South Korea: ACM, 2015. 127−140

30 Zhou Z H. *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2012.

张 宇 辽宁工程技术大学理学院讲师. 主要研究方向为数据流挖掘, 人体行为识别, 机器学习. 本文通信作者.
E-mail: vectorzhy@outlook.com
(**ZHANG Yu** Lecturer at the School of Science, Liaoning Technical University. His research interest covers data stream mining, human activity recognition, and machine learning. Corresponding author of this paper.)

包研科 辽宁工程技术大学理学院副教授. 主要研究方向为数据挖掘, 数据分析.
E-mail: baoyanke_9257@163.com
(**BAO Yan-Ke** Associate professor at the School of Science, Liaoning Technical University. His research interest covers data mining and data analysis.)

邵良杉 辽宁工程技术大学系统工程研究所教授. 主要研究方向为数据挖掘, 复杂管理信息系统.
E-mail: lntushao@163.com
(**SHAO Liang-Shan** Professor at the Research Institute of System Engineering, Liaoning Technical University. His research interest covers data mining and complex management information system.)

刘 威 辽宁工程技术大学理学院副教授. 主要研究方向为人工智能, 模式识别, 机器学习.
E-mail: lv8218218@126.com
(**LIU Wei** Associate professor at the School of Science, Liaoning Technical University. His research interest covers artificial intelligence, pattern recognition, and machine learning.)