



(12) 发明专利申请

(10) 申请公布号 CN 101944122 A

(43) 申请公布日 2011. 01. 12

(21) 申请号 201010289451. 3

(22) 申请日 2010. 09. 17

(71) 申请人 浙江工商大学

地址 310018 浙江省杭州市下沙高教园区学
正街 18 号

(72) 发明人 琚春华 郑丽丽 梅铮

(74) 专利代理机构 杭州九洲专利事务有限公
司 33101

代理人 韩小燕

(51) Int. Cl.

G06F 17/30(2006. 01)

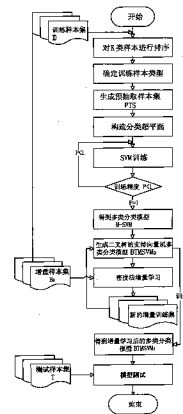
权利要求书 2 页 说明书 7 页 附图 2 页

(54) 发明名称

一种融合增量学习的支持向量机多类分类方法

(57) 摘要

本发明涉及一种融合增量学习的支持向量机多类分类方法。本发明的目的是提供一种融合增量学习的支持向量机多类分类方法,旨在减少样本的训练时间、提高分类器的分类精度及抗干扰性。本发明的技术方案是:1、在总的样本中随机抽取一部分作为训练样本集 D,另一部分作为测试样本集 T;2、对训练样本集 D 进行预抽取支持向量;3、用循环迭代法对预抽取训练样本集 PTS 进行支持向量机训练,得到多类分类模型 M-SVM;4、对多类分类模型 M-SVM 进行二叉树处理,得到支持向量机多类分类模型 BTMSVM₀;5、对多类分类模型 BTMSVM₀进行增量学习训练,得到模型 BTMSVM₁;6、将步骤 1 中的测试样本集 T 输入到多类分类模型 BTMSVM₁进行分类。本发明用于通过增量学习对海量信息进行高效多类分类。



1. 一种融合增量学习的支持向量机多类分类方法,其特征在于包括步骤:

步骤 1、在总的样本中随机抽取一部分作为训练样本集 D,另一部分作为测试样本集 T;

步骤 2、对训练样本集 D 进行预抽取支持向量,获得最终的预抽取训练样本集 PTS,剩下的训练样本作为增量样本集 B_0 ;

步骤 3、用循环迭代法对预抽取训练样本集 PTS 进行支持向量机训练,得到多类分类模型 M-SVM;

步骤 4、对多类分类模型 M-SVM 进行二叉树处理,得到基于二叉树的支持向量机多类分类模型 BTMSVM₀;

步骤 5、对多类分类模型 BTMSVM₀ 进行增量学习训练,得到新的多类分类模型 BTMSVM₁;

步骤 6、将步骤 1 中的测试样本集 T 输入到多类分类模型 BTMSVM₁ 进行分类。

2. 根据权利要求 1 所述的融合增量学习的支持向量机多类分类方法,其特征在于步骤 2 中所述的预抽取支持向量按以下步骤进行:

2. 1、对 K 类样本根据各类样本的数量进行由多到少排序,训练样本中出现最多的类为第 1 类,以此类推,直至第 K 类,形成分类序列集合 $\{s_1, s_2, \dots, s_k\}$;

2. 2、确定训练样本的类型,对于 K 类的训练样本,训练 K-1 个支持向量机,第 i 个支持向量机以第 i 类样本为正的训练样本,将第 i+1, i+2, \dots, K 类训练样本作为负的训练样本训练 SVM(i) (i = 1, \dots, K-1),第 K-1 个支持向量机将以第 K-1 类样本作为正样本,以第 K 类样本为负样本训练 SVM(K-1);

2. 3、从训练样本集 D 中的第 i 类样本集中选取一个样本,根据类均值距离法求其与第 i+1 类所有样本间的距离 $\delta_{i,j}$ (i = 1, 2, \dots, k, j = 1, 2, \dots, k):

$$\delta_{i,j} = \left| |m_i - m_j| \right|^2 - r_i - r_j$$

$$r_i = \frac{1}{n} \sum_{k=1}^{n_i} \|x_k^i - m_i\|^2$$

其中 $m_i = \frac{1}{n} \sum_{k=1}^{n_i} x_k^i$ 表示第 i 类样本集的均值向量, $\left| |m_i - m_j| \right|^2$ 为 i 类和 j 类均值向量间的距离, r_i 和 r_j 分别为 i 类及 j 类的类平均半径, n_i 为 i 类中的样本数目,称 $\delta_{i,j}$ 为 i 类和 j 类之间的类均值距离;

2. 4、找出这些距离中的最小值及与这个最小距离所对应的第 i+1 类中的样本,将其放在预抽取样本集合 PTS₀ 中;

2. 5、返回到步骤 2. 3, 遍历完训练样本集 D 中第 i 类中的所有样本;

2. 6、返回到步骤 2. 1, 遍历完训练样本集 D 中第 i+1 类中的所有样本;

2. 7、对预抽取样本集合 PTS₀ 中的样本做唯一化处理,获得最终的预抽取训练样本集合 PTS。

3. 根据权利要求 1 所述的融合增量学习的支持向量机多类分类方法,其特征在于步骤 3 中所述的循环迭代法按以下步骤进行:

3. 1、以步骤 2 最终获取的预抽取向量集 PTS, 构造数据集的分类超平面;

3. 2、利用所得到的分类超平面对训练样本集 D 进行测试, 计算训练样本集 D 中样本与分类超平面间距离 d, 以及训练正确率 P;

3. 3、当分类正确率 P 较小时, 减少加入预抽取训练样本集 PTS 中违反 KKT 条件 (请解

释一下 KKT 条件) 的样本的数量, 以求减少求解二次规划问题的规模; 当分类正确率 P 较大时, 加大加入预抽取训练样本集 PTS 中违反 KKT 条件的样本的数量, 以求减少循环迭代的次数, 并加入到预抽取训练样本集 PTS 中, 取代原预抽取训练样本集 PTS, 作为下次循环迭代的训练样本集;

3. 4、当训练正确率 $P < 1$ 时, 重复 3. 3; 当训练正确率 $P = 1$ 时跳出循环, 结束迭代, 训练结束, 得到多类分类模型 M-SVM。

4. 根据权利要求 1 所述的融合增量学习的支持向量机多类分类方法, 其特征在于步骤 4 包括:

4. 1、将步骤 3 训练得到的 SVM_1 作为二叉树的根节点, 将属于第 1 类的测试样本决策出来;

4. 2、将不属于第 1 类的样本通过 SVM_2 继续进行分类, 依次类推直到 SVM_{k-1} 将第 k 类样本训练完, 从而得到基于二叉树的 SVM 多类分类模型 $BTMSVM_0$ 。

5. 根据权利要求 1 所述的融合增量学习的支持向量机多类分类方法, 其特征在于步骤 5 中所述的增量学习按以下步骤进行:

5. 1、将增量样本集 B_0 输入 $BTMSVM_0$, 将增量样本集分为满足 $BTMSVM_0$ 的 KKT 条件和不足 KKT 条件的 2 个集合;

5. 2、对满足 KKT 条件的增量样本集和 $BTMSVM_0$ 的负例样本集分别进行密度法淘汰处理, 得到剩余集;

5. 3、将剩余集与 $BTMSVM_0$ 的正例样本集及增量样本集的不满足 KKT 条件的集合并得到新增量训练集 B_1 ;

5. 4、将新增量训练集 B_1 输入 $BTMSVM_0$ 进行训练, 得到新的多类分类模型 $BTMSVM_1$ 。

一种融合增量学习的支持向量机多类分类方法

技术领域

[0001] 本发明涉及智能信息处理和机器学习技术领域,尤其是一种融合增量学习的支持向量机多类分类方法。适用于在复杂属性条件下,通过增量学习的方式对海量信息进行高效多类分类。

背景技术

[0002] 支持向量机 (Support Vector Machine) 是近几年发展起来的一种机器学习方法,它是基于结构风险最小化原则构建的,具有很强的学习能力和泛化性能,能够较好地解决小样本、高维数、非线性、局部极小等问题,广泛的应用于模式分类和非线性回归。传统的支持向量机是针对二类分类问题而提出的,不能直接用于多类分类,但在实际的应用中更多的是多类分类问题,如何将传统支持向量分类机推广到多类分类仍是目前支持向量机领域研究的热点。

[0003] 增量学习技术相比于传统的学习方法具有明显的优越性,主要表现为:(1) 增量学习算法可以充分利用历史的训练结果,从而减少后继训练时间;(2) 无须保存历史数据且可舍弃无用样本减小训练集,从而减少了存储空间的占用;(3) 随着增量学习过程的不断进行,所获得分类器的分类精度将不断提高。

[0004] Syed. N(1999年)是最早开始基于支持向量机增量算法研究的学者,在该算法中给出了增量学习的增加策略,每次增量学习后只保留支持向量,丢弃其他的样本。G. Cauwenberghs(2003年)等人提出了在线增量训练的精确解释,即增减一个训练样本对 Lagrange 系数和支持向量的影响。该算法是有效的,但其缺点是忽略了新增样本分布对已有样本分布的影响,其训练结果并不令人满意。萧嵘等(2001年)提出了一种支持增量学习的方法 ISVM,其训练集的来源为支持向量、误分数据或有选择地淘汰一些样本而来,算法具有较高的学习精度。C. Domeniconi(2001年)提出了一种快速支持向量机增量学习算法。上述的这些算法的学习或存在振荡现象或因引入的淘汰机制缺乏新增样本对支持向量集影响的考虑而导致分类知识丢失。Liao DP 等人(2007年)提出一种基于密度法的增量学习淘汰算法,该方法能准确地提出边界向量,淘汰掉非边界向量,有效地淘汰掉无用样本,保留重要信息,既保证训练的精度又提高训练的速度,计算简单且易于实现。尽管关于样本的增量学习取得上述一些成果,但有关该领域的增量学习方法的研究还处于初级阶段,且都局限于在两类分类中样本的增加,然而实际中存在大量的多分类问题,如文本分类、信用评估、人脸识别等,因此针对多类别增量学习的研究很必要也很重要。

[0005] 目前对支持向量机多类分类算法的研究比较多,包括 1-a-r(one-against-rest) 算法、1-a-1(one-against-one) 算法、DDAGSVM(directed acyclic graph SVM) 算法等。Ying w 等人(2006年)在克服上述算法缺点的基础上提出基于二叉树的支持向量机多分类算法(简称 BTSVM),有较好的分类效果和分类效率,但是该算法的抗干扰能力较差,对训练数据的要求较高,在一定程度上阻碍了其应用。

发明内容

[0006] 本发明要解决的技术问题是：提供一种融合增量学习的支持向量机多类分类方法，旨在减少样本的训练时间、提高分类器的分类精度及抗干扰性。

[0007] 本发明所采用的技术方案是：一种融合增量学习的支持向量机多类分类方法，其特征在于包括步骤：

[0008] 步骤 1、在总的样本中随机抽取一部分作为训练样本集 D，另一部分作为测试样本集 T；

[0009] 步骤 2、对训练样本集 D 进行预抽取支持向量，获得最终的预抽取训练样本集 PTS，剩下的训练样本作为增量样本集 B₀；

[0010] 步骤 3、用循环迭代法对预抽取训练样本集 PTS 进行支持向量机训练，得到多类分类模型 M-SVM；

[0011] 步骤 4、对多类分类模型 M-SVM 进行二叉树处理，得到基于二叉树的支持向量机多类分类模型 BTMSVM₀；

[0012] 步骤 5、对多类分类模型 BTMSVM₀ 进行增量学习训练，得到新的多类分类模型 BTMSVM₁；

[0013] 步骤 6、将步骤 1 中的测试样本集 T 输入到多类分类模型 BTMSVM₁ 进行分类。

[0014] 步骤 2 中所述的预抽取支持向量按以下步骤进行：

[0015] 2. 1、对 K 类样本根据各类样本的数量进行由多到少排序，训练样本中出现最多的类为第 1 类，以此类推，直至第 K 类，形成分类序列集合 {s₁, s₂, ..., s_k}；

[0016] 2. 2、确定训练样本的类型，对于 K 类的训练样本，训练 K-1 个支持向量机，第 i 个支持向量机以第 i 类样本为正的训练样本，将第 i+1, i+2, ..., K 类训练样本作为负的训练样本训练 SVM(i) (i = 1, ..., K-1)，第 K-1 个支持向量机将以第 K-1 类样本作为正样本，以第 K 类样本为负样本训练 SVM(K-1)；

[0017] 2. 3、从训练样本集 D 中的第 i 类样本集中选取一个样本，根据类均值距离法求其与第 i+1 类所有样本间的距离 $\delta_{i,j}$ (i = 1, 2, ..., k, j = 1, 2, ..., k)：

$$[0018] \quad \delta_{i,j} = ||m_i - m_j|| |2r_i - r_j|$$

$$[0019] \quad \text{其中 } m_i = \frac{1}{n} \sum_{k=1}^{n_i} x_k^i \text{ 表示第 } i \text{ 类样本集的均值向量, } ||m_i - m_j||^2 \text{ 为 } i \text{ 类和 } j \text{ 类均值向量间的距离, } r_i \text{ 和 } r_j \text{ 分别为 } i \text{ 类及 } j \text{ 类的类平均半径, } n_i \text{ 为 } i \text{ 类中的样本数目, 称 } \delta_{i,j} \text{ 为 } i \text{ 类和 } i \text{ 类之间的类均值距离；}$$

[0020] 2. 4、找出这些距离中的最小值及与这个最小距离所对应的第 i+1 类中的样本，将其放在预抽取样本集合 PTS₀ 中；

[0021] 2. 5、返回到步骤 2. 3，遍历完训练样本集 D 中第 i 类中的所有样本；

[0022] 2. 6、返回到步骤 2. 1，遍历完训练样本集 D 中第 i+1 类中的所有样本；

[0023] 2. 7、对预抽取样本集合 PTS₀ 中的样本做唯一化处理，获得最终的预抽取训练样本集合 PTS。

[0024] 步骤 3 中所述的循环迭代法按以下步骤进行：

[0025] 3. 1、以步骤 2 最终获取的预抽取向量集 PTS，构造数据集的分类超平面；

[0026] 3.2、利用所得到的分类超平面对训练样本集 D 进行测试, 计算训练样本集 D 中样本与分类超平面间距离 d, 以及训练正确率 P;

[0027] 3.3、当分类正确率 P 较小时, 减少加入预抽取训练样本集 PTS 中违反 KKT 条件 (KKT 条件是使每个样本满足优化问题的条件, 在本发明中该条件定义为 $a_i = 0 \Rightarrow y_i f(x_i) \geq 1$, $0 < a_i < C \Rightarrow y_i f(x_i) = 1$, $a_i = C \Rightarrow y_i f(x_i) \leq 1$ 。其中非零的 a_i 为 SV, $f(x) = h$ 为 SVM 分类决策函数, $\{x_i, y_i\}$ 为训练样本, C 为指定常数。可知 $f(x) = 0$ 为分类面, $f(x) = \pm 1$ 为分类间隔面, 则 $a = 0$ 对应的样本分布在分类器分类间隔面之外, $0 < a < C$ 对应的样本位于分类间隔面之上, $a = C$ 对应的样本位于分类间隔面关于本类的异侧。违背 KKT 条件等价于 $y_i f(x_i) < 1$ 。) 的样本的数量, 以求减少求解二次规划问题的规模; 当分类正确率 P 较大时, 加大加入预抽取训练样本集 PTS 中违反 KKT 条件的样本的数量, 以求减少循环迭代的次数, 并加入到预抽取训练样本集 PTS 中, 取代原预抽取训练样本集 PTS, 作为下次循环迭代的训练样本集;

[0028] 3.4、当训练正确率 $P < 1$ 时, 重复 3.3; 当训练正确率 $P = 1$ 时跳出循环, 结束迭代, 训练结束, 得到多类分类模型 M-SVM。

[0029] 步骤 4 包括:

[0030] 4.1、将步骤 3 训练得到的 SVM_1 作为二叉树的根节点, 将属于第 1 类的测试样本决策出来;

[0031] 4.2、将不属于第 1 类的样本通过 SVM_2 继续进行分类, 依次类推直到 SVM_{k-1} 将第 K 类样本训练完, 从而得到基于二叉树的 SVM 多类分类模型 $BTMSVM_0$ 。

[0032] 步骤 5 中所述的增量学习按以下步骤进行:

[0033] 5.1、将增量样本集 B_0 输入 $BTMSVM_0$, 将增量样本集分为满足 $BTMSVM_0$ 的 KKT 条件和 不满足 KKT 条件的 2 个集合;

[0034] 5.2、对满足 KKT 条件的增量样本集和 $BTMSVM_0$ 的负例样本集分别进行密度法淘汰处理, 得到剩余集;

[0035] 5.3、将剩余集与 $BTMSVM_0$ 的正例样本集及增量样本集的不满足 KKT 条件的集合并得到新增量训练集 B_1 ;

[0036] 5.4、将新增量训练集 B_1 输入 $BTMSVM_0$ 进行训练, 得到新的多类分类模型 $BTMSVM_1$ 。

[0037] 本发明的有益效果是: 本方法通过将支持向量机和二叉树的基本思想相结合, 以类均值距离作为二叉树的生成算法, 让与其他类相隔最远的类最先分隔出来, 这种方法构造的最优超平面具有较好的推广性, 能有效的避免孤立点带来的问题, 使其具有一定的抗干扰性, 而且只需构造 $k-1$ 个 (k 为样本类别总数) 支持向量机分类器, 测试时并不一定需要计算所有的分类器判别函数, 从而可节省测试时间。同时将基于密度法的增量学习算法加入本发明的多类分类器, 在实现多分类能力的同时提升其增量学习的能力。

附图说明

[0038] 图 1 是本发明的工作流程图。

[0039] 图 2 是本发明中 $BTMSVM$ 增量学习过程。

具体实施方式

[0040] 本发明融合增量学习的支持向量机多类分类方法按以下步骤进行：

[0041] 步骤 1、在总的样本中随机抽取一部分作为训练样本集 D，另一部分作为测试样本集 T。

[0042] 步骤 2、对训练样本集 D 进行预抽取支持向量，进一步包括步骤：

[0043] 2.1、对 K 类样本根据各类样本的数量进行由多到少排序，训练样本中出现最多的类为第 1 类，以此类推，直至第 K 类，形成分类序列集合 $\{s_1, s_2, \dots, s_k\}$ ；

[0044] 2.2、确定训练样本的类型，对于 K 类的训练样本，训练 K-1 个支持向量机，第 i 个支持向量机以第 i 类样本为正的训练样本，将第 i+1, i+2, ..., K 类训练样本作为负的训练样本训练 SVM(i) (i = 1, ..., K-1)，第 K-1 个支持向量机将以第 K-1 类样本作为正样本，以第 K 类样本为负样本训练 SVM(K-1)；

[0045] 2.3、从训练样本集 D 中的第 i 类样本集中选取一个样本，根据类均值距离法求其与第 i+1 类所有样本间的距离 $\delta_{i,j}$ (i = 1, 2, ..., k, j = 1, 2, ..., k)：

$$[0046] \quad \delta_{i,j} = \left\| \|m_i - m_j\|^2 - r_i - r_j \right.$$

$$[0047] \quad \left. \text{其中 } m_i = \frac{1}{n} \sum_{k=1}^{n_i} x_k^i \text{ 表示第 } i \text{ 类样本集的均值向量, } \|m_i - m_j\|^2 \text{ 为 } i \text{ 类和 } j \text{ 类均值向量间的距离, } r_i \text{ 和 } r_j \text{ 分别为 } i \text{ 类及 } j \text{ 类的类平均半径, } n_i \text{ 为 } i \text{ 类中的样本数目, 称 } \delta_{i,j} \text{ 为 } i \text{ 类和 } i \text{ 类之间的类均值距离;} \right.$$

量间的距离， r_i 和 r_j 分别为 i 类及 j 类的类平均半径， n_i 为 i 类中的样本数目，称 $\delta_{i,j}$ 为 i 类和 i 类之间的类均值距离；

[0048] 2.4、找出这些距离中的最小值及与这个最小距离所对应的第 i+1 类中的样本，将其放在预抽取样本集合 PTS_0 中；

[0049] 2.5、返回到步骤 2.3，遍历完训练样本集 D 中第 i 类中的所有样本；

[0050] 2.6、返回到步骤 2.1，遍历完训练样本集 D 中第 i+1 类中的所有样本；

[0051] 2.7、对预抽取样本集合 PTS_0 中的样本做唯一化处理，获得最终的预抽取训练样本集合 PTS，剩下的训练样本作为增量样本集 B_0 。

[0052] 步骤 3、用循环迭代法对预抽取训练样本集 PTS 进行支持向量机训练，进一步包括步骤：

[0053] 3.1、以步骤 2 最终获取的预抽取向量集 PTS，构造数据集的分类超平面；

[0054] 3.2、利用所得到的分类超平面对训练样本集 D 进行测试，计算训练样本集 D 中样本与分类超平面间距离 d，以及训练正确率 P；

[0055] 3.3、当分类正确率 P 较小时，减少加入预抽取训练样本集 PTS 中违反 KKT 条件的样本的数量，以求减少求解二次规划问题的规模；当分类正确率 P 较大时，加大加入预抽取训练样本集 PTS 中违反 KKT 条件的样本的数量，以求减少循环迭代的次数，并加入到预抽取训练样本集 PTS 中，取代原预抽取训练样本集 PTS，作为下次循环迭代的训练样本集；

[0056] 3.4、当训练正确率 $P < 1$ 时，重复 3.3；当训练正确率 $P = 1$ 时跳出循环，结束迭代，训练结束，得到多类分类模型 M-SVM。

[0057] 步骤 4、对多类分类模型 M-SVM 进行二叉树处理，包括：

[0058] 4.1、将步骤 3 训练得到的 SVM_1 作为二叉树的根节点，将属于第 1 类的测试样本决策出来；

[0059] 4.2、将不属于第 1 类的样本通过 SVM_2 继续进行分类，依次类推直到 SVM_{k-1} 将第 K

类样本训练完,从而得到基于二叉树的 SVM 多类分类模型 BTMSVM₀。

[0060] 步骤 5、对多类分类模型 BTMSVM₀ 进行增量学习训练,进一步包括:

[0061] 5.1、将增量样本集 B₀ 输入 BTMSVM₀,将增量样本集分为满足 BTMSVM₀ 的 KKT 条件和
不满足 KKT 条件的 2 个集合;

[0062] 5.2、对满足 KKT 条件的增量样本集和 BTMSVM₀ 的负例样本集分别进行密度法淘汰
处理,得到剩余集;

[0063] 5.3、将剩余集与 BTMSVM₀ 的正例样本集及增量样本集的不满足 KKT 条件的集合并
得到新增量训练集 B₁;

[0064] 5.4、将新增量训练集 B₁ 输入 BTMSVM₀ 进行训练,得到新的多类分类模型 BTMSVM₁。

[0065] 步骤 6、将步骤 1 中的测试样本集 T 输入到多类分类模型 BTMSVM₁ 进行分类。

[0066] 下面将通过两个具体的案例来说明本发明的技术方案。

[0067] 以某商业银行过去 18 个月的一组信用卡用户数据为实验样本集。经过处理的实
验样本集共 920 条,划分为训练样本集 D 和测试样本集 T 两部分,其中训练样本 240 条,测
试样本包括 680 条记录。训练样本中首先抽取 80 条,此后是新增 20 条、30 条、40 条、70 条
的增量进行训练。

[0068] 实施例 1:未使用增量学习训练的 BTMSVM 实验结果

[0069] 第 1 步,对训练样本集进行预抽取得到 80 条记录组成的预抽取训练样本集 PTS;

[0070] 第 2 步,用 PTS 进行循环迭代法支持向量机训练,得到多类分类模型 M-SVM;

[0071] 第 3 步,对 M-SVM 进行二叉树处理,得到基于二叉树的支持向量机多类分类模型
BTMSVM₀,训练时间 10.92 秒,提取 680 条测试样本测试分类模型,分类正确率为 70.49%;

[0072] 第 4 步,新增 20 条训练样本,与原训练样本 80 条合并,共 100 条记录,输入 BTMSVM₀
进行训练,训练时间为 23.75 秒。因为不能利用已经训练好的分类器(训练样本为 80 时),
需要重新训练,所以实际花费的时间等于利用 80 条训练样本进行训练的时间与重新利用
100 条训练样本进行训练的时间的和,即 10.92 秒与 16.43 秒的和,共 27.35 秒;

[0073] 第 5 步,再依次新增 30、40、70 条训练样本,迭加后的训练样本分别为 130、170、240
条,输入 BTMSVM₀ 进行训练,训练时间分别为 46.98 秒、70.89 秒、108.87 秒。

[0074] 第 6 步,以上四组增量训练后,分别提取 680 条测试样本测试分类模型分类正确
率,分别为:72.34%,74.16%,75.47%,87.02%。

[0075] 实验结果如表 1 所示。

[0076] 表 1 未使用增量训练的实验结果

	训练样本数	测试样本数	分类正确率	训练时间(秒)	训练总时间(秒)
	80	680	70.49%	10.92	10.92
[0077]	100	680	72.34%	16.43	27.35
	130	680	74.16%	19.63	46.98
	170	680	75.47%	23.91	70.89
	240	680	87.02%	37.98	108.87

[0078] 实施例 2 :使用增量学习训练的 BTMSVM 实验结果

[0079] 第 1 步,对训练样本集 D 进行预抽取得到 80 条记录组成的预抽取训练样本集 PTS ;

[0080] 第 2 步,用 PTS 进行循环迭代法支持向量机训练,得到多类分类模型 M-SVM ;

[0081] 第 3 步,对 M-SVM 进行二叉树处理,得到基于二叉树的支持向量机多类分类模型 BTMSVM₀,训练时间 10.92 秒 ;

[0082] 第 4 步,提取 680 条测试样本集 T 经过多类分类模型 BTMSVM₀ 分类,分类正确率 70.49% ;

[0083] 第 5 步,将 20 条记录增量样本集 B₀ 与 80 条初始训练样本 PTS 输入 BTMSVM₀ 进行增量训练,增量训练时间 2.56 秒,经过密度法淘汰规则,共淘汰 5 条记录,剩余 95 条记录为新的增量训练集 B₁ ;

[0084] 第 6 步,将新的增量训练集输入进行训练,得到新的二叉树支持向量机多类分类模型 BTMSVM₁ ;

[0085] 第 7 步,提取 680 条测试样本经过 BTMSVM₁ 分类,通过第 1 次增量训练的分类模型分类正确率提高到 71.82% ;

[0086] 第 8 步,将 B₁ 作为初始训练集,新增 30 条记录为增量样本集 B₂,重复第 5 步到第 7 步,淘汰样本 8 条,分类正确率提高到 73.85% ;

[0087] 第 9 步,继续以 40 条和 70 条记录作为新增样本重复第 5 步到第 6 步,分类正确率分别为 74.09%和 86.97%。

[0088] 实验结果如表 2 所示。表 2 使用增量学习后的实验结果

	初始训练 样本数	增量 训练 样本 数	淘汰 样本 数	训练 样本 总数	测试 样本 数	分类正 确率	增量 训练 时间 (秒)	训练 总时 间 (秒)
[0089]	初始训练	/	/	80	680	70.49%	/	10.92
	80	20	5	95	680	71.82%	2.56	13.48
	95	30	8	117	680	73.85%	3.79	17.27
	117	40	12	145	680	74.09%	4.89	22.16
	145	70	10	215	680	86.97%	8.09	30.25

[0090] 实验结果表明,支持增量学习的 BTMISVM 模型可以充分利用以前的训练成果,大大减少训练时间,分类正确率也随着样本的增加而加速提高。本发明可以在一定程度上满足信用卡行为评估对于增量学习的需求。

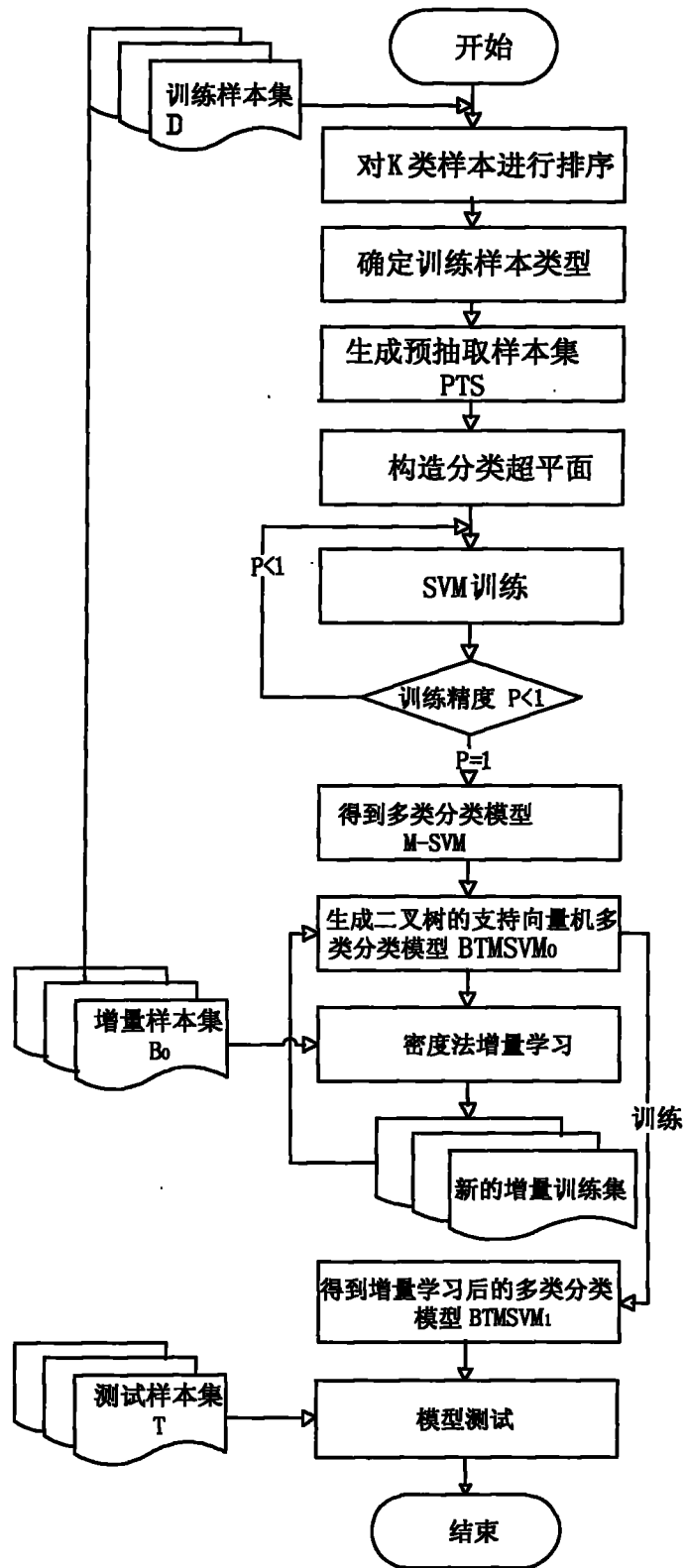


图 1

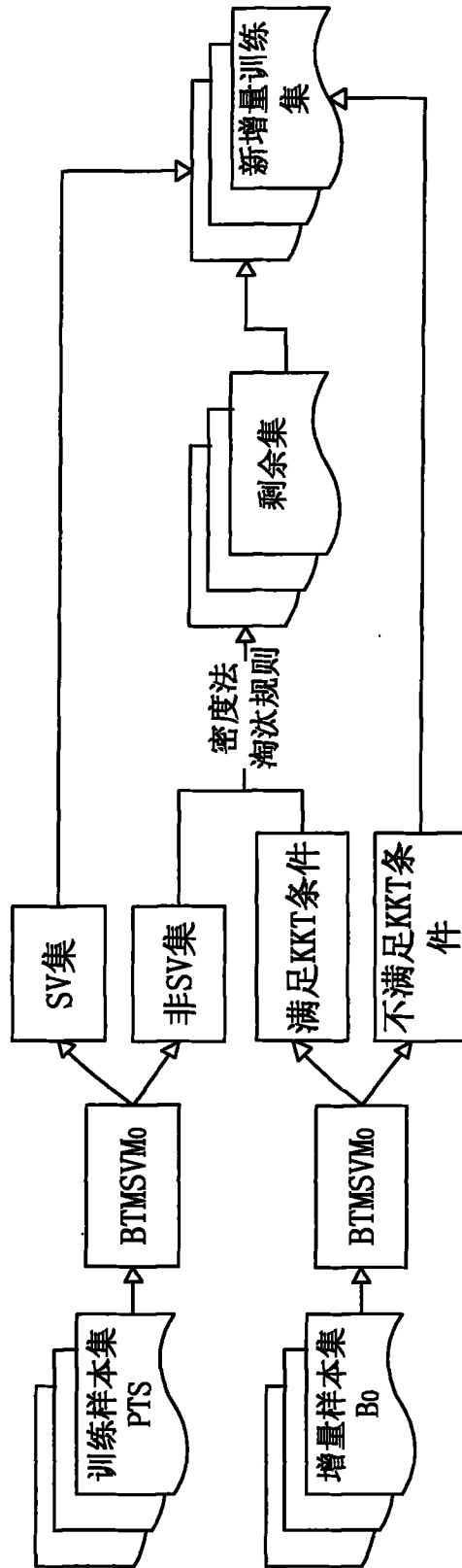


图 2