(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107103332 B
(45)授权公告日 2018.06.26

(21)申请号 201710225520.6
(22)申请日 2017.04.07
(65)同一申请的已公布的文献号
申请公布号 CN 107103332 A
(43)申请公布日 2017.08.29
(73)专利权人 武汉理工大学
地址 430070 湖北省武汉市洪山区珞狮路122号
(72)发明人 刘芳 钟昊 李思瀚 童蜜 秦王晨 赵斐
(74)专利代理机构 湖北武汉永嘉专利代理有限公司 42102
代理人 李丹
(51)Int.Cl.
*G06K 9/62*(2006.01)

审查员 万洋

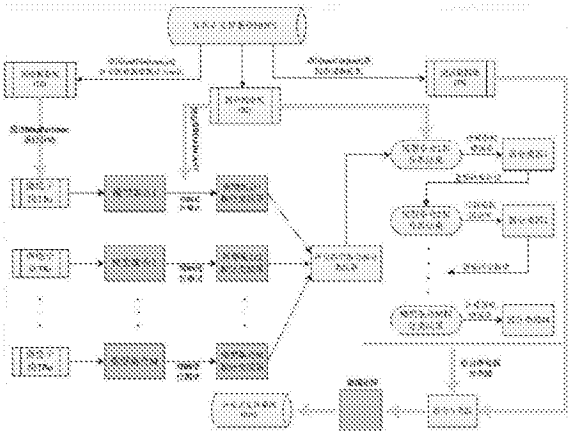
权利要求书3页 说明书10页 附图2页

(54)发明名称

一种面向大规模数据集的相关向量机分类方法

(57)摘要

本发明公开了一种面向大规模数据集的相关向量机分类方法，该方法针对传统相关向量机对较大数据集的性能降低问题，结合集成学习的思想，对方法的弱分类器获取方式进行改进，解决了方法运行过程中可能的数据局部不均衡问题，然后利用Spark分布式计算平台的特点将大规模数据集分类任务分布到集群的各个子节点上，从而有效提升模型训练的速度，实现相关向量机对大规模数据集的良好支持，通过集成学习中弱分类器的组合策略，将每个子节点得到的弱分类器结合，保证最终模型的准确率。对于大规模数据集的分类，本发明方法具有良好的效果。



CN 107103332 B

1.一种面向大规模数据集的相关向量机分类方法，其特征在于，该方法包括以下几个步骤：

1)数据采集：通过传感器对被测对象进行实时检测，采集被测对象的在不同工作状态下的数据；所述被测对象的工作状态包括正常工作状态和异常工作状态，工作状态根据分类需求划分；

2)获取样本集：对采集的数据进行样本提取，得到总体样本集并存储在数据存储模块上；

3)在Spark分布式平台下将总体样本集切分成若干等份，将切分后的样本集分配到集群子节点上，对样本的权值进行初始化，同时所有子节点共享一个提升样本数据集，用于提升训练；

4)在集群子节点上对样本集每一维进行训练得到若干个RVM弱分类器；

5)得到弱分类器后对共享的提升训练数据集进行测试，并将得到的结果返回至父节点，父节点中给提升训练集样本分配相同权值；

6)以相同数据集得到的若干个弱分类器为一组进行迭代运算，计算每一组弱分类器的带权误差；所述相同数据集是指步骤3)中切分后的样本集中的同一样本集；

7)选择一组训练子集所得的弱分类器对带权误差进行比较，选取每一组弱分类器中带权值误差最小的分类器，并计算该分类器对应的置信度；

8)对样本权值进行更新，选择另外一组训练子集所得的弱分类器为一组，再重复步骤6)和7)中操作选取带权值误差最小的分类器，循环迭代一次进行一次权值更新，直至所有训练子集选择完毕结束迭代；

9)根据步骤8)中所选取的弱分类器进行组合，得到被测对象工作状态的最终分类器，根据分类器结果对被测对象工作情况进行判断。

2.根据权利要求1所述的面向大规模数据集的相关向量机分类方法，其特征在于，所述步骤1)中，数据采集过程具体如下：

采用数据采集传感器对被测对象进行实时检测，被测对象需设置在工作正常和工作异常两种工作状态检测，数据采集传感器将检测所得的传感信号转为数值数据，同步传送至数据逻辑模块，相应获得被测对象的两组不同工作状态下的相关数据；两组数据的组成是被测对象在不同采样时刻以一定的采样频率所检测到的大量数据集。

3.根据权利要求1所述的面向大规模数据集的相关向量机分类方法，其特征在于，所述步骤2)中，样本集的获取过程具体如下：

2.1)特征分类：数据逻辑模块接收到来自数据采集模块的数据后，提取出能代表并区别该检测数据的一组特征参数，且该组特征参数包括W个特征量，并对特征向量进行编号；

2.2)样本集获取：经特征提取后的数据按检测时间组成样本，一条样本中包含相同检测时间下的W个特征量；样本数据集由训练样本、提升样本和测试样本以设定比例组成，对样本中工作正常状态下的样本和工作异常状态下的样本数量根据实际需要按设定比例调配。

4.根据权利要求1所述的面向大规模数据集的相关向量机分类方法，其特征在于，所述步骤3)中，样本$x_i$初始化权值采用以下公式

$D_1(x_i)=1/N, i=1,2,3...N$

其中N为样本的总数量，i为样本集的索引号。

5.根据权利要求1所述的面向大规模数据集的相关向量机分类方法，其特征在于，所述步骤3)中，若存在对数据不均衡的数据集，采用Smote算法生成若干个少数类别的新样本得到一个新的样本集。

6.根据权利要求5所述的面向大规模数据集的相关向量机分类方法，其特征在于，对于数据不均衡的数据集使用Smote算法生成新样本而得到新样本集方法如下：

步骤3.1)对于该类数据集中的每一个样本x，以欧氏距离为标准计算它到样本集中所有样本的距离，得到其k近邻；

步骤3.2)根据样本不平衡比例设置一个采样比例以确定采样倍率；

步骤3.3)对于每一个该类数据集中样本，从其k近邻中随机选择若干个样本，假设选择的近邻为$x_k$，分别与原样本x按照以下公式构建新的样本$x_{new}$；

$x_{new}=x+rand(0,1)\times(x_k-x)$。

7.根据权利要求1所述的面向大规模数据集的相关向量机分类方法，其特征在于，所述步骤4)中，在子节点上对样本集进行训练的方法采用如下具体步骤：

步骤4.1)对于初始样本集$(x_{i1},\ldots,x_{iM},y_i)$，$i=1,2,\ldots,N$，其中，$X=(x_1,\ldots,x_M)$是样本的特征向量，$y_i\in Y=\{-1,+1\}$表示每个样本所属工作状态类别；所得算法模型如公式：

$$y(x;w)=\sum_{n=1}^{N}w_nK(x,x_n)+w_0$$

其中，$w_n$代表对应样本的权值，$K(x,x_n)$为核函数，N为样本个数；

步骤4.2)在进行分类时，无法解析计算得到权值的后验概率，但是可利用拉普拉斯理论做近似计算：对于当前固定的α，使用二阶牛顿法求最大可能权值$w_{MP}$，计算以下公式，

$$\log\{p(t|w)p(w|\alpha)\}=\sum_{\alpha=1}^{N}[t_n\log yn+(1-t_n)\log(1-y_n)]-\frac{1}{2}w^tAw$$

其中$y_n=\sigma\{y(x_n;w)\}$，$A=diag(\alpha_0,\alpha_1,\ldots,\alpha_N)$；利用拉普拉斯方法，将对数后验概率进行二次逼近，对该公式进行二次求导可得式：

$$\nabla_w\nabla_M\log p(w|t,\alpha)|w_{MP}=-(\Phi^T/B\Phi+A)$$

Φ是N×(N+1)矩阵，$\Phi=[\varphi(x_1),\varphi(x_2),\ldots,\varphi(x_N)]^T$，$B=diag(\beta_1,\beta_2,\ldots,\beta_N)$，通过公式可得到协方差矩阵Σ；

步骤4.3再通过Σ和$w_{MP}$对超参数α进行更新，得到$\alpha^{new}=\gamma_i/w_{MP}^2$，其中$\gamma_i\equiv1-\alpha_i\Sigma_{ii}$，且$\Sigma_{ii}$是$\Sigma=(\Phi^TB\Phi+A)^{-1}$矩阵中的第i个对角元素；

模型在经过多次迭代计算后，许多$\alpha_i$会趋向于无穷大，从而与其对应的$w_i$就等于零，其中不为零的$w_i$所对应的训练样本即为算法的相关向量；由相关向量所确定的模型函数是一个高维的超平面，通过这个超平面可以近似的将测试样本划分在平面两侧，从而依据所属不同面得到分类结果。

8.根据权利要求1所述的面向大规模数据集的相关向量机分类方法，其特征在于，所述步骤6)中，弱分类器的带权误差计算公式如下，

$$\varepsilon_s=\sum_{i=1}^{N}w_i[H_s(x_i)-y_i]^2$$

其中，$\varepsilon_S$为弱分类器的带权误差，$H_S(x_i)$表示样本训练得到的弱分类器，s为迭代次数，$y_i$为真实值，$w_i$为当前样本权值。

9.根据权利要求1所述的面向大规模数据集的相关向量机分类方法，其特征在于，所述步骤8)中，权值更新的计算如式，

$$w_{s+1}(i)=w_s(i)\exp(-y_iH_s(x_i))$$

其中，$w_s(i)$为选取弱分类器前样本权值，$w_{S+1}(i)$为选取后样本对应权值。

10.根据权利要求1所述的面向大规模数据集的相关向量机分类方法，其特征在于，所述步骤9)中，在步骤8)中每次迭代时，基于最小二乘去做一个加权回归，最后根据所有回归函数的和得到最终的分类器，计算公式如式：

$$G_{\text{final}}(x)=sign(\sum_{s=0}^{T-1}H_s(x))$$ 。

# 一种面向大规模数据集的相关向量机分类方法

## 技术领域

[0001] 本发明涉及机器学习领域，尤其涉及一种面向大规模数据集的相关向量机分类方法。

## 背景技术

[0002] 随着互联网的迅猛发展，基于互联网统计和分析生成的数据大多都是大规模甚至海量数据。面对此类信息规模大、增长速度快的数据，如何高效、精确地对这些数据进行组织和分类是当前互联网信息时代的一大难题。使用传统平台进行数据分析时，通常的分类方法大多数是以串行的方式来完成相关计算任务，而且运行的平台在很大程度上受到了计算机性能的约束，当数据量增大时，方法性能会急剧下降，从而导致方法分类过程无法高效进行。

[0003] 相关向量机(relevance vector machine，RVM)是一种新的监督学习方法，与支持向量机(support vector machine，SVM)相比，它是一种基于贝叶斯的概率稀疏模型。通过在参数上定义受超参数控制的Gaussian先验概率，在贝叶斯框架下进行机器学习，利用自相关判定理论来移除不相关的点，从而获得稀疏化模型。由于在样本数据的迭代学习过程中，大部分参数的后验分布趋于零，而非零的参数所对应的学习样本和决策域的样本不相关，只代表数据中的原型样本，因此称这些样本为相关向量，体现了数据中最核心的特征。相关向量机最大的优点是极大的减少了核函数的计算量，并且在选取核函数时，不再受限于一定的条件，可选择的核函数大大增多。基于以上优点，相关向量机在诸多领域中性能表现都十分优秀。但是传统的相关向量机在面对大规模数据集进行分类和回归预测的时候，出现了效率极大的降低问题。

[0004] 近些年来，关于大数据分布式技术的发展十分迅速，诞生了很多分布式框架，通过用这些分布式框架来进行分布式计算可以有效的提升方法的运行效率，现在流行的大数据框架有Storm、Hadoop、Spark等。Storm适用于进行实时计算，Hadoop则更加适用于进行离线运算和对时延要求不高的任务，而Spark因为是基于内存进行运算，在进行迭代计算时，传统MapReduce(Hadoop框架核心编程计算模式)虽然具有自动容错、平衡负载和可扩展性的优点，但是因为它是采用非循环式的数据流模型，导致在进行迭代计算时要进行大量的磁盘IO操作，从而使得此类任务的性能受到极大限制，而在Spark中，通过RDD(Resilient Distributed Dataset，弹性分布数据集)将数据加载在内存中，便于之后的多次重用，使得它在处理迭代式计算时效果十分优秀。但是在Spark进行样本分区的过程中，样本是随机分布的，有可能导致某些分区中正负样本数目不均衡，从而导致得到的弱分类器整体性能受到影响，进而影响最终所得强分类器的性能。

[0005] Adaboost是基于PAC学习理论而建立的一套集成学习方法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)。Adaboost方法是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。

将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器融合起来，作为最后的决策分类器。使用Adaboost分类器可以排除一些不必要的训练数据特征，并将关键放在重要的训练数据上面。Adaboost方法是一种实现简单，应用也很简单的方法。它通过组合弱分类器而得到强分类器，同时具有分类错误率上界随着训练增加而稳定下降，不会过拟合等的性质，应该说是一种很适合于在各种分类场景下应用的方法。

## 发明内容

[0006] 本发明要解决的技术问题在于针对现有技术中传统相关向量机处理大数据集的性能降低问题的缺陷，提供一种面向大规模数据集的相关向量机分类方法，该方法基于集成学习思想，通过现有的AdaBoost的思想和RVM的相关特性，实现AdaBoost和RVM的结合，并对局部进行优化改进，再结合Spark平台在迭代运算方面的优秀表现，实现海量数据集下RVM方法的应用。

[0007] 本发明解决其技术问题所采用的技术方案是：一种面向大规模数据集的相关向量机分类方法，包括以下步骤：

[0008] 1)数据采集：通过传感器对被测对象进行实时检测，采集被测对象的在不同工作状态下的数据；所述被测对象的工作状态包括正常工作状态和异常工作状态，工作状态根据分类需求划分；

[0009] 2)获取样本集：对采集的数据进行样本提取，得到总体样本集并存储在数据存储模块上；

[0010] 3)在Spark分布式平台下将总体样本集切分成若干等分，将切分后的样本集分配到集群子节点上，对样本的权值进行初始化，同时所有子节点共享一个提升样本数据集，用于提升训练；

[0011] 4)在集群子节点上对样本集每一维进行训练得到若干个RVM弱分类器；

[0012] 5)得到弱分类器后对共享的提升训练数据集进行测试，并将得到的结果返回至父节点，父节点中给提升训练集样本分配相同权值；

[0013] 6)以相同数据集得到的若干个弱分类器为一组进行迭代运算，计算每一组弱分类器的带权误差；所述相同数据集是指步骤3)中切分后的样本集中的同一样本集；

[0014] 7)选择一组训练子集所得的弱分类器对带权误差进行比较，选取每一组弱分类器中带权值误差最小的分类器，并计算该分类器对应的置信度；

[0015] 8)对样本权值进行更新，选择另外一组训练子集所得的弱分类器为一组，再重复步骤6)和7)中操作选取带权值误差最小的分类器，循环迭代一次进行一次权值更新，直至所有训练子集选择完毕结束迭代；

[0016] 9)根据步骤8)中所选取的弱分类器进行组合，得到被测对象工作状态的最终分类器，根据最终分类器的计算结果对被测对象所处工作状态进行判断。

[0017] 按上述方案，所述步骤1)中，数据采集过程具体如下：

[0018] 步骤1.1数据采集：采用数据采集传感器对被测对象进行实时检测，被测对象需设置在工作正常和工作异常两种工作状态检测，数据采集传感器将检测所得的传感信号转为数值数据，同步传送至数据逻辑模块，相应获得被测对象的两组不同工作状态下的相关数据；两组数据的组成是被测对象在不同采样时刻以一定的采样频率所检测到的大量数据

集；

[0019] 按上述方案，所述步骤2)中，样本集的获取过程具体如下：

[0020] 2.1)特征分类：数据逻辑模块接收到来自数据采集模块的数据后，提取出能代表并区别该检测数据的一组特征参数，且该组特征参数包括W个特征量，并对特征向量进行编号；

[0021] 2.2)样本集获取：经特征提取后的数据按检测时间组成样本，一条样本中包含相同检测时间下的W个特征量。样本数据集由训练样本、提升样本和测试样本以6:1:3组成，对样本中工作正常状态下的样本和工作异常状态下的样本数量按实际需要设定比例调配。

[0022] 在实际工作中，被测对象大部分时间都是处于工作正常状态，只有极少数时间出现工作异常，为模拟此类条件需对样本中工作正常样本和工作异常样本按一定比例调配。训练样本包含60300条样本，其中60000条为工作正常状态下的数据样本，300条为工作异常时的数据样本，测试样本包含31000条样本，工作正常样本30000条，工作异常样本1000条，提升样本选用12000条样本，其中工作正常样本11000条，工作异常样本1000条。

[0023] 按上述方案，所述步骤3)中，样本$x_i$初始化权值采用以下公式

[0024] $D_1(x_i)=1/N, i=1,2,3...N$ (1)

[0025] 其中N为样本的总数量，i为样本集的索引号。

[0026] 按上述方案，所述步骤3)中，若存在对数据不均衡的数据集(样本分区过程中随机分布可能导致某些分区正负样本数目不均衡)，采用Smote方法生成若干个少数类别的新样本得到一个新的样本集；

[0027] 按上述方案，对于数据不均衡的数据集使用Smote方法生成新样本而得到新样本集方法如下：

[0028] 步骤3.1对于该类数据集中的每一个样本x，以欧氏距离为标准计算它到样本集中所有样本的距离，得到其k近邻；

[0029] 步骤3.2根据样本不平衡比例设置一个采样比例以确定采样倍率；

[0030] 步骤3.3对于每一个该类数据集中样本，从其k近邻中随机选择若干个样本，假设选择的近邻为$x_k$，分别与原样本x按照公式(2)构建新的样本$x_{new}$。

[0031] $x_{new}=x+rand(0,1)\times(x_k-x)$ (2)

[0032] 按上述方案，所述步骤4)中，在子节点上对样本集进行训练的方法采用如下具体步骤：

[0033] 步骤4.1对于初始样本集$(x_{i1},...,x_{iM},y_i)$，$i=1,2,...,N$，其中$X=(x_1,...,x_M)$是样本的特征向量，$y_i\in Y=\{-1,+1\}$表示每个样本所属工作状态类别，所得方法模型如公式(3)。

[0034] $$y(x;w)=\sum_{n=1}^{N} w_n K(x,x_n)+w_0 \quad (3)$$

[0035] 其中，$w_n$代表对应样本的权值，$K(x,x_n)$为核函数，N为样本个数。

[0036] 步骤4.2在进行分类时，无法解析计算得到权值的后验概率，但是可利用拉普拉斯理论做近似计算：对于当前固定的α，使用二阶牛顿法求最大可能权值$w_{MP}$，计算如公式(4)，其中$y_n=\sigma\{y(x_n;w)\}$，$A=diag(\alpha_0,\alpha_1,...,\alpha_N)$。利用拉普拉斯方法，将对数后验概率进行二次逼近。对式(4)进行二次求导可得式(5)：

[0037] $$\log\{p(t|w)p(w|\alpha)\}=\sum_{\alpha=1}^{N}[t_n \log yn+(1-t_n)\log(1-y_n)]-\frac{1}{2}w^{t}Aw \quad (4)$$

[0038] $$\nabla_w\nabla_M\log\ p(w|t,\alpha)\,|_{WMP}=-(\Phi^T/B\Phi+A) \quad (5)$$

[0039] Φ是N×(N+1)矩阵，$\Phi=[\varphi(x_1),\varphi(x_2),\ldots,\varphi(x_N)]^T$，B=diag(β1,β2,...,βN)，通过公式(5)可得到协方差矩阵Σ；

[0040] 步骤4.3再通过Σ和wMP对超参数α进行更新，得到$\alpha^{new}=\gamma_i/w_{MP}^2$，其中γi≡1-αiΣii，且Σii是Σ=(ΦTBΦ+A)-1矩阵中的第i个对角元素；

[0041] 模型在经过多次迭代计算后，许多αi会趋向于无穷大，从而与其对应的wi就等于零，其中不为零的wi所对应的训练样本即为方法的相关向量(Relevance Vector)。由相关向量所确定的模型函数是一个高维的超平面，通过这个超平面可以近似的将测试样本划分在平面两侧，从而依据所属不同面得到分类结果。

[0042] 本发明方法训练中产生的弱分类器不采用二值分类器，即分类结果为+1或-1。本发明方法对这一点进行了修改，弱分类器的结果是如式(6)中所示具有连续置信度的输出。

[0043] $$H(x)=P_w(y=1|x)-P_w(y=-1|x) \quad (6)$$

[0044] 即输出的值域是实数域范围，具体结果hj∈[-1,+1]。弱分类器结果的符号即是所属类别，而数值则是标识输出结果的置信度，这种方式更加贴近实际应用，使得结果更加平滑，不易出现“过拟合”现象。

[0045] 按上述方案，所述步骤6)中，弱分类器的带权误差计算公式如下，

[0046] $$\varepsilon_s=\sum_{i=1}^{N}w_i[H_s(x_i)-y_i]^2 \quad (7)$$

[0047] 其中，εS为弱分类器的带权误差，HS(xi)表示样本训练得到的弱分类器，s为迭代次数，yi为真实值，wi为当前样本权值。

[0048] 按上述方案，所述步骤7)中，根据步骤6)中所得弱分类器的带权误差εS计算该分类器对应的置信度，计算方法如下：

[0049] $$\alpha_s=\frac{1}{2}\log\left(\frac{1-\varepsilon_s}{\varepsilon_s}\right) \quad (8)$$

[0050] 按上述方案，所述步骤8)中，权值更新的计算如式(9)，

[0051] $$w_{S+1}(i)=w_s(i)\exp(-y_iH_s(x_i)) \quad (9)$$

[0052] 其中，ws(i)为选取弱分类器前样本权值，wS+1(i)为选取后样本对应权值。

[0053] 按上述方案，所述步骤9)中，在步骤8)中每次迭代时，基于最小二乘去做一个加权回归，最后根据所有回归函数的和得到最终的分类器，计算公式如式(10)。

[0054] $$G_{final}(x)=sign(\sum_{s=0}^{T-1}H_s(x)) \quad (10)$$

[0055] 本发明产生的有益效果是：本发明方法结合了集成学习方法Adaboost，常用的AdaBoost方法有Discrete AdaBoost、Real AdaBoost和Gentle AdaBoost等，通过每次循环过程中使用不同的弱分类器获取方式和样本权值更新的方式组合得到对应不同的AdaBoost，上述方法步骤1到步骤8是针对GentleAdaBoost与RVM的结合实现方法(GBR)。本发明方法针对不同规模大小的数据集的处理问题又进一步进行了改进，使得最终所有弱分类器所组成的分类器集合平均值比Gentle AdaBoost的更高。GBR方法在某些数据特征不是很多、数据规模不是很大的数据集下性能表现相对的比较稳定，所以本发明没有将其剔除，

详细数据和分析可以参考下面具体实施内容中的对比。

[0056] 本发明方法可记作All BoostRVM方法，简称ABR，具体改进如下。

[0057] 步骤4中，在每次循环内直接对子节点上的子样本集选取所有特征值进行整体训练，而不再进行逐一特征的切分然后再训练。这样改进整体训练后，对子节点训练后将只得到一个弱分类器，即一组样本子集得到一个弱分类器，步骤6和7中则改成所有组样本集的弱分类器进行迭代运算。

[0058] GBR方法在进行训练的过程中，首先是对数据进行了随机切分成相等数量的若干份分配到Spark集群的子节点上，然后再在子节点上对子节点中数据再次依据特征进行划分进行训练得到弱分类器，再计算弱分类器的错误率选择错误率最小弱分类器作为当前循环的结果。即训练过程中需要通过训练得到m*n个弱分类器(n代表数据切分的数量，和数据集大小和集群机器性能有关，m代表样本特征个数)，所以当训练数据较大时，因为过度切分数据和所需训练模型过多，会导致方法性能有所下降且方法运行时间较长，对此我们利用RVM的特性提出一种新的弱分类器获取策略。RVM在进行小样本分类时表现的性能较好，属于一种精度较高的一种分类器，通过上述方式改进对样本进行整体训练得到的最终弱分类器集平均准确率会比GBR更好。

## 附图说明

[0059] 下面将结合附图及实施例对本发明作进一步说明，附图中：

[0060] 图1是本发明方法的样本分类流程图；

[0061] 图2为本发明数据采集系统的结构示意图；

[0062] 图3是单机RVM、GAR和ABR在Image Segmentation上的F1值和分类正确率表格；

[0063] 图4是RVM、GAR和ABR在10000条数据下十次实验的F1值和分类正确率表格；

[0064] 图5是1000000数据样本下不同比例训练数据GAR和ABR的结果表格。

## 具体实施方式

[0065] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

[0066] 如图1所示，一种面向大规模数据集的相关向量机分类算法，包括以下几个步骤：

[0067] 步骤1通过数据采集模块和数据逻辑模块采集数据和样本提取，得到样本集并存储在数据存储模块上，数据采集系统的结构示意图如图2，具体实施步骤如下：

[0068] 步骤1.1数据采集：采用数据采集模块对被测对象进行实时检测，被测对象需设置在工作正常和工作异常两种工作状态检测，数据采集模块将检测所得的信号转为为数据，同步传送至数据逻辑模块，相应获得被测对象的两组不同工作状态下的相关数据；两组数据的组成是被测对象在不同采样时刻以一定的采样频率所检测到的大量数据集；

[0069] 步骤1.2特征分类：数据逻辑模块接收到来自数据采集模块的数据后，提取出能代表并区别该检测数据的一组特征参数，且该组特征参数包括W个特征量，并按顺序对特征向量进行编号；

[0070] 步骤1.3样本集获取：经特征提取后的数据按检测时间组成样本，一条样本中包含

相同检测时间下的W个特征量。样本集获取：经特征提取后的数据按检测时间组成样本，一条样本中包含相同检测时间下的W个特征量。样本数据集由训练样本、提升样本和测试样本组成。为了测试不同数量样本集情况下算法模型的性能，我们选择对不同数量的数据集分别进行测试，选定测试样本数目分别为：10000条、1000000条两种种。在这两种数据集上再分别用不同的数据划分方式来进行测试。10000条样本数据集中数据划分时选择60％做训练数据，10％做提升数据，30％做测试数据。

[0071] 步骤2在Spark分布式平台下将总体样本集切分成若干等分，将切分后的样本集分配到集群子节点上，对样本权值进行初始化，同时所有子节点共享一个数据集用于提升训练。样本$x_i$初始化权值计算公式如式(1)，其中N为样本的总数量，i为样本集的索引号。

[0072] $D_1(x_i)=1/N, i=1,2,3...N$ (1)

[0073] 步骤3对数据不均衡的数据集(样本分区过程中随机分布可能导致某些分区正负样本数目不均衡)用Smote算法生成若干个少数类别的新样本得到一个新的样本集，使用Smote算法生成新样本而得到新样本集具体实施方法如下：

[0074] 步骤3.1对于该类数据集中的每一个样本x，以欧氏距离为标准计算它到样本集中所有样本的距离，得到其k近邻；

[0075] 步骤3.2根据样本不平衡比例设置一个采样比例以确定采样倍率；

[0076] 步骤3.3对于每一个该类数据集中样本，从其k近邻中随机选择若干个样本，假设选择的近邻为$x_k$，分别与原样本x按照公式(2)构建新的样本$x_{new}$。

[0077] $x_{new}=x+rand(0,1)\times(x_k-x)$ (2)

[0078] 步骤4在子节点上对样本集每一维进行训练得到若干个RVM弱分类器，对样本集进行训练的方法可以划分如下具体步骤：

[0079] 步骤4.1对于初始样本集$(x_{i1},...,x_{iM},y_i)$，i=1,2,...,N，其中$X=(x_1,...,x_M)$是样本的特征向量，$y_i\in Y=\{-1,+1\}$表示每个样本所属类别，所得算法模型如公式(3)。

[0080] $$y(x;w)=\sum_{n=1}^{N} w_n K(x,x_n)+w_0 \quad (3)$$

[0081] 其中，$w_n$代表对应样本的权值，$K(x,x_n)$为核函数，N为样本个数。

[0082] 步骤4.2在进行分类时，无法解析计算得到权值的后验概率，但是可利用拉普拉斯理论做近似计算：对于当前固定的α，使用二阶牛顿法求最大可能权值$w_{MP}$，计算如公式(4)，其中$y_n=\sigma\{y(x_n;w)\}$，$A=diag(\alpha_0,\alpha_1,...,\alpha_N)$。利用拉普拉斯方法，将对数后验概率进行二次逼近。对式(4)进行二次求导可得式(5)：

[0083] $$\log\{p(t|w)p(w|\alpha)\}=\sum_{\alpha=1}^{N}[t_n\log ym+(1-t_n)\log(1-y_n)]-\frac{1}{2}w^tAw \quad (4)$$

[0084] $$\nabla_w\nabla_M\log p(w|t,\alpha)\big|_{w_{MP}}=-(\Phi^T/B\Phi+A) \quad (5)$$

[0085] 其中，Φ是N×(N+1)矩阵，$\Phi=[\varphi(x_1),\varphi(x_2),...,\varphi(x_N)]^T$，$B=diag(\beta_1,\beta_2,...,\beta_N)$，通过公式(5)可得到协方差矩阵Σ。

[0086] 步骤4.3再通过Σ和$w_{MP}$对超参数α进行更新，得到$\alpha^{new}=\gamma_i/w_{MP}^2$，其中$\gamma_i\equiv1-\alpha_i\Sigma_{ii}$，且$\Sigma_{ii}$是$\Sigma=(\Phi^TB\Phi+A)^{-1}$矩阵中的第i个对角元素。

[0087] 模型在经过多次迭代计算后，许多$\alpha_i$会趋向于无穷大，从而与其对应的$w_i$就等于零，其中不为零的$w_i$所对应的训练样本即为算法的相关向量(Relevance Vector)。由相关

向量所确定的模型函数是一个高维的超平面，通过这个超平面可以近似的将测试样本划分在平面两侧，从而依据所属不同面得到分类结果。

[0088] 训练中产生的弱分类器不采用二值分类器，即分类结果为+1或-1。弱分类器的结果是如式(6)中所示具有连续置信度的输出。

[0089] $H(x)=P_w(y=1|x)-P_w(y=-1|x)$ (6)

[0090] 即输出的值域是实数域范围，具体结果$h_j\in[-1,+1]$。弱分类器结果的符号即是所属类别，而数值则是标识输出结果的置信度，这种方式更加贴近实际应用，使得结果更加平滑，不易出现“过拟合”现象。

[0091] 步骤5得到弱分类器后对共享的提升训练数据集进行测试，并将得到的结果返回至父节点，父节点中给提升训练集样本分配相同权值；

[0092] 步骤6以相同数据得到的若干个弱分类器为一组进行迭代运算，计算每一组弱分类器的带权误差。弱分类器的带权误差计算公式如式(7)，$H_S(x_i)$表示样本训练得到的弱分类器，s为迭代次数，$y_i$为真实值，$w_i$为当前样本权值。

[0093] $$\varepsilon_s=\sum_{i=1}^{N}w_i[H_s(x_i)-y_i]^2 \quad (7)$$

[0094] 步骤7择一组训练子集所得的弱分类器对带权误差进行比较，选取每一组弱分类器中带权值误差最小的分类器，根据步骤6中所得弱分类器的带权误差$e_S$计算该分类器对应的置信度，计算方法如式(8)。

[0095] $$\alpha_s=\frac{1}{2}\log\left(\frac{1-\varepsilon_s}{\varepsilon_s}\right) \quad (8)$$

[0096] 步骤8对样本权值进行更新，选择另外一组训练子集所得的弱分类器为一组，再重复步骤6和7中操作选取带权值误差最小的分类器，循环迭代一次进行一次权值更新，直至所有训练子集选择完毕结束迭代，权值更新的计算如式(9)，$w_s(i)$为选取弱分类器前样本权值，$w_{S+1}(i)$为选取后样本对应权值。

[0097] $w_{s+1}(i)=w_s(i)\exp(-y_iH_s(x_i))$ (9)

[0098] 步骤9在步骤8中每次迭代时，基于最小二乘去做一个加权回归，最后根据所有回归函数的和得到最终的分类器，计算公式如式(10)。根据分类器结果对被测对象工作情况进行判断和预测。

[0099] $$G_{final}(x)=sign(\sum_{s=0}^{T-1}H_s(x)) \quad (10)$$

[0100] 实验结果中样本召回率体现了分类模型对正样本的识别能力，样本召回率越高，说明模型对正样本的识别能力越强。正确率体现了模型对负样本的区分能力，正确率越高，说明模型对负样本的区分能力越强。F1值是两者的综合，F1值越高，说明分类模型越稳健。

[0101] 图3所示为单机RVM、GBR和ABR算法对UCI数据集进行训练并测试的F1值和正确率。UCI数据集(Image Segmentation)是一个图像划分数据集，总样本数是2310，分为七类(GRASS、PATH、WINDOW、CEMENT、FOLIAGE、SKY、BRICKFACE)，每个样本有19个属性值和一个样本类别标签。实验过程数据划分比例6:1:3分别对应为训练数据、提升数据、测试数据，采用RangePartition策略对训练数据集RDD进行分区保证数据的均衡。

[0102] 对于GBR和ABR两个算法考虑到数据分区时的随机性，进行多次测试取平均值。另外，GAR和ABR算法在ImageSegmentation数据集上的训练时间和单机RVM训练时间的比率分

别为1.81和0.152，结合图3可以看出，虽然单机RVM的F1值和样本分类正确率都比GBR和ABR两个算法要高2%-4%，但是也相差不大，而ABR算法在样本训练时间效率上有十分显著的提升，提高了近6倍的效率。

[0103] 图4所示为在10000条样本数据集下单机RVM和GAR、ABR十次实验过程中F1值的最小值、最大值和平均值，以及对应分类正确率的最大值、最小值和平均值。通过图4可以初步得出，当样本个数达到10000时，GAR和ABR算法的分类结果与RVM分类结果的差距已经明显缩小，差距范围一般在1%～3%之间，其中性能表现最稳定的仍然是GAR，因为数据数量过小未比较分析其他差距，只能说明GAR和ABR算法的分类正确率达到要求。

[0104] 图5为1000000条样本数据集通过选择不同比例数据用于做训练数据后得到的GAR和ABR的分类结果的F1值、分类正确率的平均值。可以得出结论，当样本数量比较少时，RVM的运行性能最好。但是当样本数目逐渐变多时，通过提高训练样本的比例，GAR、ABR和RVM的训练差距逐步降低。且在训练样本比例达到60%时，GAR、ABR和RVM的性能基本持平。另外，在数据量很大时，选择GAR或者ABR时算法运行时间得到减少，尤其是ABR算法所需时间急剧减少，ABR算法效率大大提高。因此可知，GAR适合训练样本偏多但是特征维度不多的数据集，而ABR则更加适合与训练样本较多而且样本特征维度较高的数据集。在大规模的训练数据集的情况下，相比于单机RVM，本发明算法的模型训练时间大大减少，算法效率大大提高，而且在采样比例达到60%以上时，本发明算法的分类准确率十分接近于单机RVM。

[0105] 本发明方法一个具体应用实施例如下：

[0106] 为判断桥梁结构上是否存在裂缝损伤，需要对桥梁相关的各种物理量(温度、应变、位移、加速度等)进行测量。通过传感器收集到的数据由于实时监测和监测位置众多，信息量一般是十分的庞大。本发明算法可以应用于此类数据的分析和处理，对桥梁结构的各个部件的损伤情况进行判断和预测。

[0107] 步骤1通过数据采集模块和数据逻辑模块采集数据和样本提取，得到样本集并存储在数据存储模块上，具体步骤如下：

[0108] 步骤1.1数据采集：采用数据采集模块对被测对象进行实时检测，被测对象需设置在工作正常和工作异常两种工作状态检测，数据采集模块将检测所得的信号转为为数据，同步传送至数据逻辑模块，相应获得被测对象的两组不同工作状态下的相关数据；两组数据的组成是被测对象在不同采样时刻以一定的采样频率所检测到的大量数据集；本实例中被测对象为某大型桥梁建筑，数据采集传感器为光纤光栅加速度传感器，数据逻辑模块为对应光纤光栅监测软件系统。步骤1.1中关于被测对象的两种工作状态分别是工作正常状态对应桥梁某部位无裂缝，工作异常状态为桥梁某部位存在裂缝。

[0109] 步骤1.2特征分类：步骤1.1中光纤光栅加速度传感器数量为20个，选取大型桥梁的100米部分均匀分布，一个传感器在单位时间内采集到的数据即为一个特征向量，数据逻辑模块接收到数据后，按传感器编号对数据进行特征提取分类，并按传感器编号对特征向量进行编号。

[0110] 步骤1.3样本集获取：经特征提取后的数据按检测时间组成样本，一条样本中包含相同检测时间下的W个特征量。样本集获取：经特征提取后的数据按检测时间组成样本，一条样本中包含相同检测时间下的W个特征量。样本数据集由训练样本、提升样本和测试样本组成。训练样本集选用60300条传感器数据，其中无裂缝的样本60000条，有裂缝的样本300

条;测试集对应选用了31000条,其中无裂缝的样本30000条,有裂缝的样本1000条;提升集12000条,无裂缝的样本有11000条,有裂缝的样本则有1000条。每个样本有20个特征,分别对应着大型桥梁上部署的20个加速度传感器。

[0111] 步骤2在Spark分布式平台下将总体样本集切分成若干等分,将切分后的样本集分配到集群子节点上,对样本权值进行初始化,同时所有子节点共享一个数据集用于提升训练。样本$x_i$初始化权值计算公式如式(1),其中N为样本的总数量,i为样本集的索引号。

[0112] $D_1(x_i)=1/N, i=1,2,3...N$ (1)

[0113] 步骤3对数据不均衡的数据集(样本分区过程中随机分布可能导致某些分区正负样本数目不均衡)用Smote算法生成若干个少数类别的新样本得到一个新的样本集,使用Smote算法生成新样本而得到新样本集具体实施方法如下:

[0114] 步骤3.1对于该类数据集中的每一个样本x,以欧氏距离为标准计算它到样本集中所有样本的距离,得到其k近邻;

[0115] 步骤3.2根据样本不平衡比例设置一个采样比例以确定采样倍率;

[0116] 步骤3.3对于每一个该类数据集中样本,从其k近邻中随机选择若干个样本,假设选择的近邻为$x_k$,分别与原样本x按照公式(2)构建新的样本$x_{new}$。

[0117] $x_{new}=x+rand(0,1)\times(x_k-x)$ (2)

[0118] 步骤4、在子节点上对样本集每一维进行训练得到若干个RVM弱分类器,对样本集进行训练的方法可以划分如下具体步骤:

[0119] 步骤4.1对于初始样本集$(x_{i1},...,x_{iM},y_i)$,i=1,2,...,N,其中$X=(x_1,...,x_M)$是样本的特征向量,$y_i\in Y=\{-1,+1\}$表示每个样本所属类别,所得算法模型如公式(3)。

[0120] $$y(x;w)=\sum_{n=1}^{N} w_n K(x,x_n)+w_0 \quad (3)$$

[0121] 其中,$w_n$代表对应样本的权值,$K(x,x_n)$为核函数,N为样本个数。

[0122] 步骤4.2在进行分类时,无法解析计算得到权值的后验概率,但是可利用拉普拉斯理论做近似计算:对于当前固定的α,使用二阶牛顿法求最大可能权值$w_{MP}$,计算如公式(4),其中$y_n=\sigma\{y(x_n;w)\}$,$A=diag(\alpha_0,\alpha_1,...,\alpha_N)$。利用拉普拉斯方法,将对数后验概率进行二次逼近。对式(4)进行二次求导可得式(5):

[0123] $$\log\{p(t|w)p(w|\alpha)\}=\sum_{\alpha=1}^{N}[t_n \log yn+(1-t_n)\log(1-y_n)]-\frac{1}{2}w^t Aw \quad (4)$$

[0124] $$\nabla_w\nabla_M \log p(w|t,\alpha)\big|_{w_{MP}}=-(\Phi^T/B\Phi+A) \quad (5)$$

[0125] Φ是N×(N+1)矩阵,$\Phi=[\varphi(x_1),\varphi(x_2),.....,\varphi(x_N)]^T$,$B=diag(\beta_1,\beta_2,...,\beta_N)$,通过公式(5)可得到协方差矩阵Σ。

[0126] 步骤4.3再通过Σ和$w_{MP}$对超参数α进行更新,得到$\alpha^{new}=\gamma_i/w_{MP}^2$,其中$\gamma_i\equiv1-\alpha_i\Sigma_{ii}$,且$\Sigma_{ii}$是$\Sigma=(\Phi^T B\Phi+A)^{-1}$矩阵中的第i个对角元素。

[0127] 模型在经过多次迭代计算后,许多$\alpha_i$会趋向于无穷大,从而与其对应的$w_i$就等于零,其中不为零的$w_i$所对应的训练样本即为算法的相关向量(Relevance Vector)。由相关向量所确定的模型函数是一个高维的超平面,通过这个超平面可以近似的将测试样本划分在平面两侧,从而依据所属不同面得到分类结果。

[0128] 训练中产生的弱分类器不采用二值分类器,即分类结果为+1或-1。弱分类器的结

果是如式(6)中所示具有连续置信度的输出。

[0129] $H(x)=P_w(y=1|x)-P_w(y=-1|x)$ (6)

[0130] 即输出的值域是实数域范围，具体结果$h_j\in[-1,+1]$。弱分类器结果的符号即是所属类别，而数值则是标识输出结果的置信度，这种方式更加贴近实际应用，使得结果更加平滑，不易出现“过拟合”现象。

[0131] 步骤5、得到弱分类器后对共享的提升训练数据集进行测试，并将得到的结果返回至父节点，父节点中给提升训练集样本分配相同权值；

[0132] 步骤6、以相同数据得到的若干个弱分类器为一组进行迭代运算，计算每一组弱分类器的带权误差。弱分类器的带权误差计算公式如式(7)所示，$H_S(x_i)$表示样本训练得到的弱分类器，s为迭代次数，$y_i$为真实值，$w_i$为当前样本权值。

[0133] $$\varepsilon_s=\sum_{i=1}^{N}w_i[H_s(x_i)-y_i]^2 \quad (7)$$

[0134] 步骤7、择一组训练子集所得的弱分类器对带权误差进行比较，选取每一组弱分类器中带权值误差最小的分类器，根据步骤6中所得弱分类器的带权误差$\varepsilon_S$计算该分类器对应的置信度，计算方法如式(8)。

[0135] $$\alpha_s=\frac{1}{2}\log\left(\frac{1-\varepsilon_s}{\varepsilon_s}\right) \quad (8)$$

[0136] 步骤8、对样本权值进行更新，选择另外一组训练子集所得的弱分类器为一组，再重复步骤6和7中操作选取带权值误差最小的分类器，循环迭代一次进行一次权值更新，直至所有训练子集选择完毕结束迭代，权值更新的计算如式(9)，$w_s(i)$为选取弱分类器前样本权值，$w_{S+1}(i)$为选取后样本对应权值。

[0137] $w_{s+1}(i)=w_s(i)\exp(-y_iH_s(x_i))$ (9)

[0138] 步骤9、在步骤8中每次迭代时，基于最小二乘去做一个加权回归，最后根据所有回归函数的和得到最终的分类器，根据分类器结果对被测对象工作情况进行判断。计算公式如式(10)。

[0139] $$G_{final}(x)=sign(\sum_{s=0}^{T-1}H_s(x)) \quad (10)$$

[0140] 分类器结果中单机RVM模型的分类正确率和样本召回率为0.925和0.110，GBR算法为0.874和0.675，ABR算法为0.898和0.740，GBR和ABR相对于单机RVM的模型训练时间比分别为0.31和0.23。可以得出结论，虽然GBR和ABR的分类正确率比单机RVM稍低，且相差不大，在某些情况下可忽略不计，但是样本召回率要高出好几倍，说明本发明模型对正样本的识别能力比较强。在模型训练时间的比较上，GBR和ABR相对于单机RVM缩短了不少时间，效率大幅提高。上述结果说明了本发明算法在桥梁健康监测系统中对传感数据的分析和处理上相对于单机RVM的优势。

[0141] 应当理解的是，对本领域普通技术人员来说，可以根据上述说明加以改进或变换，而所有这些改进和变换都应属于本发明所附权利要求的保护范围。

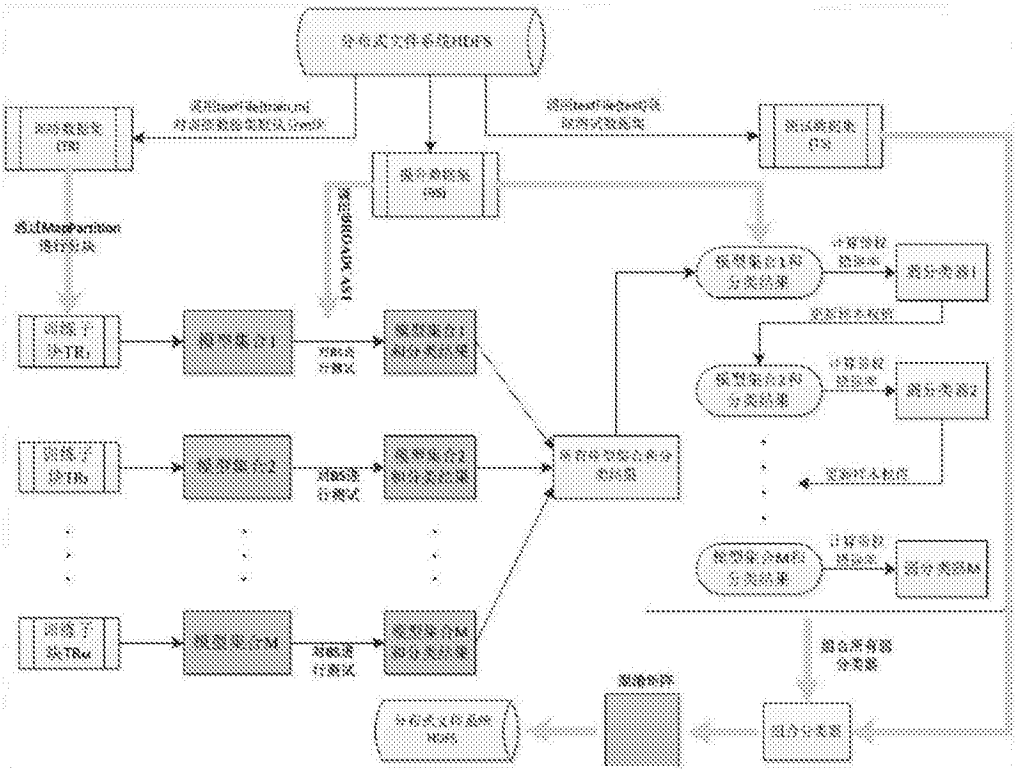


图1

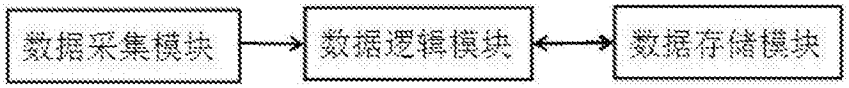



图2

| 类别 | RVM | | GAR | | ADR | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| GRASS | 95.0% | 96.4% | 88.6% | 93.2% | 92.2% | 95.1% |
| PATH | 92.8% | 94.1% | 89.8% | 95.1% | 91.4% | 94.0% |
| WINDOW | 82.3% | 85.9% | 82.0% | 86.1% | 79.8% | 84.0% |
| CEMENT | 78.8% | 86.9% | 73.0% | 84.5% | 77.1% | 84.1% |
| FOLIAGE | 79.9% | 90.2% | 70.0% | 86.3% | 69.9% | 87.1% |
| SKY | 96.8% | 96.5% | 97.9% | 99.0% | 95.8% | 89.2% |
| BRICKFACE | 97.5% | 98.7% | 90.1% | 95.4% | 96.3% | 94.1% |
| Average | 89.02% | 92.70% | 85.89% | 90.37% | 86.09% | 89.65% |

图3

| 类别 | GAR | | ABR | | RVM | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| min | 86.41% | 86.75% | 89.39% | 87.42% | 90.80% | 90.72% |
| max | 91.67% | 92.94% | 90.72% | 93.10% | 92.33% | 94.30% |
| average | 88.74% | 90.80% | 90.08% | 90.06% | 91.46% | 91.69% |

图4

| 训练样本比例 | GAR | | ABR | |
|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy |
| 20% | 83.33% | 85.12% | 87.13% | 88.01% |
| 40% | 87.54% | 88.23% | 88.34% | 89.22% |
| 60% | 88.81% | 90.27% | 90.13% | 90.18% |
| 80% | 89.02% | 91.02% | 91.72% | 91.63% |

图5