Journal of System Simulation

Volume 29 | Issue 12 Article 14

6-6-2020

Application of Distributed Clustering in Anomaly Detection of Farm Environment Data

Deng Li
1. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China;;2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai 200072, China;

Honglin Pang
1. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China;;2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai 200072, China;

Ling Wang
1. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China;;2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai 200072, China;

Minrui Fei
1. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China;;2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai 200072, China;

Application of Distributed Clustering in Anomaly Detection of Farm Environment Data

Abstract

Abstract: The massive farm environment data stored in the distributed system should be dealt with so as to provide abnormal environment reference and make preventive strategies for crop yield. *Considering the characteristics of the farm environment data, the Dirichlet Process Mixture Model (DPMM) clustering is implemented with the farm environment data on Hadoop and the anomaly detection method of the farm environment is proposed based on clustering analysis.* Under the framework of MapReduce, Map stage implements the distribution of the sample points to the models; Reduce stage completes the update of models and the number of clusters. The performance has been verified by experiments. The results of clustering and the index of suitable environment for tomato are compared to implement the anomaly detection. The analysis results show that the method can be applied to anomaly detection of large number of farm environment data.

Keywords

DPMM, distributed clustering, farm environment data, anomaly detection

第 29 卷第 12 期 系统仿真学报© Vol. 29 No. 12
2017 年 12 月 Journal of System Simulation Dec., 2017

分布式聚类在农场环境数据异常检测中的应用

邓丽 [1,2]，庞洪霖 [1,2]，王灵 [1,2]，费敏锐 [1,2]
(1. 上海大学机电工程与自动化学院，上海 200072；2. 上海市电站自动化技术重点实验室，上海 200072)

摘要：为了处理大量分布式存储的农场环境数据，为作物增产提供异常环境参考并制定预防策略，*本文结合农场环境数据的特点，在 Hadoop 平台中实现了对农场环境数据的 Dirichlet 过程混合模型聚类，并提出了基于聚类分析的农场环境异常检测方法。*在 MapReduce 框架下，Map 阶段完成样本点到模型的分配；Reduce 阶段对模型与类簇个数进行更新。通过实验验证了分布式 Dirichlet 聚类的性能，分析结果表明该方法可以应用于大量农场环境数据的异常检测。

关键词：Dirichlet 过程混合模型；分布式聚类；农场环境数据；异常检测
中图分类号：TP338.8 文献标识码：A 文章编号：1004-731X (2017) 12-3035-07
DOI: 10.16182/j.issn1004731x.joss.201712014

Application of Distributed Clustering in Anomaly Detection of Farm Environment Data

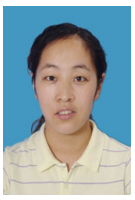
Deng Li[1,2], *Pang Honglin*[1,2], *Ling Wang*[1,2], *Minrui Fei*[1,2]
(1. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China;
2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai 200072, China)

Abstract: The massive farm environment data stored in the distributed system should be dealt with so as to provide abnormal environment reference and make preventive strategies for crop yield. *Considering the characteristics of the farm environment data, the Dirichlet Process Mixture Model (DPMM) clustering is implemented with the farm environment data on Hadoop and the anomaly detection method of the farm environment is proposed based on clustering analysis.* Under the framework of MapReduce, Map stage implements the distribution of the sample points to the models; Reduce stage completes the update of models and the number of clusters. The performance has been verified by experiments. The results of clustering and the index of suitable environment for tomato are compared to implement the anomaly detection. The analysis results show that the method can be applied to anomaly detection of large number of farm environment data.

Keywords: DPMM; distributed clustering; farm environment data; anomaly detection

引言

近年来随着农业生产的智能化，网络通信、自动控制及软件技术被大量的应用到农业生产技术中，包括温室大棚智能监控与农业专家知识库在内的解决方案积累了大量数据，分析挖掘数据中的规律能更好地指导生产，聚类作为一种数据分析方法，在建立作物产量预测模型，环境异常检测，植株品种特性划分等方面都有成功的应用实现，如 M. Gunasundari Ananthara 等[1]基于蜂巢聚类方法建立了作物产量的预测模型；时雷等[2]基于 STORM 和 COD 聚类算法设计了小麦生长环境数



收稿日期：2015-10-16 修回日期：2016-01-02;
基金项目：上海市科委重点项目(14DZ1206302);
作者简介：邓丽(1978-)，女，安徽，博士后，副教授，研究方向为机器学习、智能优化算法等；庞洪霖(1991-)，男，江苏，硕士生，研究方向为机器学习与分布式计算。

据异常检测的算法；潘渊洋等[3]针对传感器网络测量数据的异常检测问题，利用基于密度的 DBSCAN 聚类算法提取环境数据特征集进行检测；唐成龙等[4]针对网络攻击行为的特点，首先利用层次聚类改进了模糊 k 均值聚类的敏感性，再利用遗传算法克服局部最优的缺点，提出了基于特征选择的 AGFCM 聚类算法的异常入侵模型，来对网络攻击的类型进行检测。

然而，当前普遍使用的是一些传统聚类方法，如基于密度的 DBSCAN 聚类对用户设置参数敏感，不同密度的数据集需要设置不同的扫描半径和点数，运算复杂度过高[3]，而通过层次聚类改善的 FCM 聚类也只能发现球状的数据簇，聚类效果并不理想[4]；另一方面，随着农业生产规模的扩大，数据分布存储的情况下，单机处理大量分布式的数据面临的运算能力不足的缺陷已初步显现。

针对上述问题，本文提出在分布式平台上通过 Dirichlet 过程混合模型(Dirichlet Process Mixture Models，DPMM)聚类完成对环境数据的异常检测。

DPMM 是一种基于模型的聚类方法。该方法将混合模型引入 Dirichlet 过程，实现的对连续数值的聚类。DPMM 聚类可以应对复杂数据集，不用指定类簇初值，且兼具了模糊聚类与层次聚类的特性。DPMM 已被广泛应用于样本聚类、话题建模、标签提取等领域。如 Fox 等[5]采用 Dirichlet 过程混合模型的聚类性质，在机动目标跟踪中实现目标数量的确定；Orbanz 等[6]将 Dirichlet 过程引入马尔科夫模型中，对模型 Gibbs 采样实现图像分割区域个数的自动生成。

Hadoop 平台为分布式计算提供了解决方案，本文参考开源机器学习包，在 Hadoop 上完成对农场环境数据的分布式聚类，并结合番茄适宜生长指标，实现了异常环境检测。

1 农场环境数据的 DPMM 生成过程

1.1 农场环境数据

聚类数据大约四万条，来自种植番茄的大棚。实际采集到的数据受多种因素影响，数据集存在缺失值，需要进行数据清洗。参考常见的缺失值处理方法，对数据进行插入补全。例如温度 Temp 与相对湿度 RH(Relative Humidity)在 12 月 10 日-12 月 11 日之间出现九十条缺失，由于数据集采样周期为五分钟，属于定距型缺失，此处做线性插入补全，补全后部分数据如表 1 所示。

表 1 清洗后部分二维数据
Tab. 1 Part of the 2D Data After Cleaning

Temp (℃)	RH (g/m^3)	Date
…	…	…
15.30	90.81	10-12-2014 23:10:00
15.30	91	10-12-2014 23:15:00
15.23	91.09	10-12-2014 23:20:00
…	…	…
11.64	99.22	11-12-2014 06:45:00
11.57	99.31	11-12-2014 06:50:00
11.47	100	11-12-2014 06:55:00
…	…	…

聚类数据大约四万条，来自种植番茄的大棚。实际采集到的数据受多种因素影响，数据集存在缺失值，需要进行数据清洗。参考常见的缺失值处理方法，对数据进行插入补全。例如温度 Temp 与相对湿度 RH(Relative Humidity)在 12 月 10 日-12 月 11 日之间出现九十条缺失，由于数据集采样周期为五分钟，属于定距型缺失，此处做线性插入补全，补全后部分数据如表 1 所示。

考虑到在高维数据聚类中，不相关的或者冗余的特征会使类的结构不明显，本文选取温度和相对湿度，组成二维样本集进行聚类。

为排查异常环境数据，需要确定番茄生长的适宜环境范围。查阅番茄无土栽培的相关资料[7]，得知番茄的盛果期的环境温度白天一般控制在 25 ℃，上下波动范围在 5 ℃以内，夜晚保持在 15 ℃至 17 ℃，相对湿度保持在 80%至 90%。可以看出

DOI: 10.16182/j.issn1004731x.joss.201712014

温度和湿度的分布主要受均值和标准差条件制约，参考往年同一时间段内的数据，可认为近似服从正态分布，故选择高斯模型作为各个类簇的基础模型。

1.2 DPMM 的生成过程

本文采用 DPMM 对农场环境数据实现聚类，聚类前实现模型的初始化，然后通过模型对数据进行描述。

前述 1.1 节根据番茄生长的限制条件确定了使用正态分布模型描述环境数据簇，将正态分布作为混合模型的基础模型，引入 Dirichlet 过程。因为 Dirichlet 过程只能实现离散值聚类，因此引入混合模型，通过混合模型实现同分布值的聚类，通过类标签 si 建立模型比例 π_k 与模型参数 $\varphi_{s_i}|s_i=k$ 之间的联系[8-9]。

Dirichlet 过程可以看作趋向无穷维度的 Dirichlet 分布，其中参数 G_0 表示模型中各个组元的分布，α 决定各个模型的混合权重。

对照图 1 聚类图模型结构，以农场环境数据为例描述 DPMM 的产生过程，其中虚线框表示聚类迭代过程，虚线框以外表示模型的生成过程。

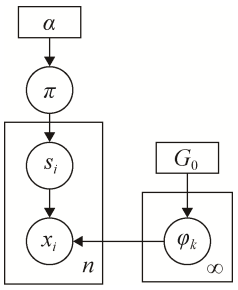


图 1 DPMM 聚类图模型结构
Fig. 1 Model Structure of DPMM Clustering

定义一组变量 $s=\{s_1,s_2,...s_n\}$表示 n 个样本与 k 个类簇的关系。$s_i=k$ 表示第 i 个样本属于第 k 类簇。当实际模型数远大于参数初值 k 时，可近似认为是变参贝叶斯过程。由 DP 过程对类标签 s_i 进行采样。然后，对各类簇个数进行归一化从而生成模型比例系数 π，作为混合模型各个成分比例的先验分布[10]。其中 $\Pi=\{\pi_1,...\pi_K\}$，$\pi_k\geq 0$，$k=\{1,...K\}$且 $\sum_{k=1}^{K}\pi_k=1$，π_k 表示第 k 个模型占总体的比例系数[10-11]。

用 $X=\{x_1,x_2,...x_n\}$ 表示农场环境数据集，样本个数为 n，此处 X 由 k 个正态模型混合生成，这里的 k 个模型都采用正态分布，参数集 $\Phi=\{\varphi_1,...\varphi_K\}$，其中 $\varphi_k=\{\mu_k,\sigma_k\}$，表示模型的均值和标准差。由于混合模型的各个成分都服从正态分布，且正态分布的先验共轭也为正态分布，参考番茄生长的适宜温湿度范围，选取 N(22.5,6.5)、N(7.5,1)与 N(85,5)、N(5,1)分别作为温度与相对湿度的 μ_k 、σ_k 的基准分布 G_0，用来产生模型参数 φ_K[12]。

混合模型的比例系数 Π 和初始参数 Φ 确定后，就可以用混合模型来拟合农场环境数据集实现聚类。

2 分布式聚类实现农场环境异常检测

本文的分布式聚类在 MapReduce 计算框架下实现 DPMM 的迭代求解过程。MapReduce 是一种运行在计算机集群上的用于离线数据分析的计算模型，能够利用数据并行性进行分布计算而后汇总结果的计算框架。样本数据由网络化文件系统 Hadoop Distributed File System(HDFS)统一管理，分散的存储在不同的计算机硬盘中[13]。

为了分析分布式存储的环境数据，本文在该计算机集群中利用 MapReduce 框架部署了分布式聚类程序实现计算。HDFS 中的样本数据格式参考表 1。前两列的温度(℃)和相对湿度(%)用于聚类计算，数据采集时间为主键，作为异常点排查的条件。

DPMM 聚类的主要思想是通过 Dirichlet 过程生成混合模型，通过 MapReduce 框架迭代更新模型，最终模型稳定在一定范围或达到最大迭代次数退出迭代，如果最终模型稳定，对各个模型进行异常判断，找出异常簇，并计算聚类正确率。DPMM 的异常检测实现流程图如图 2。

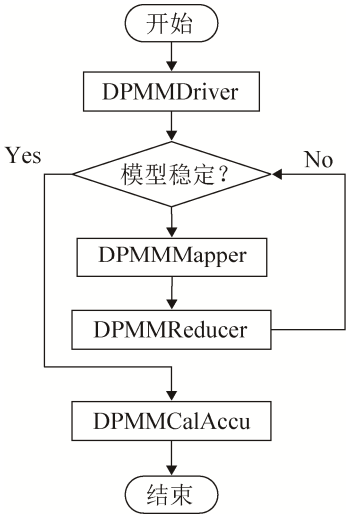


图 2 异常检测总体流图
Fig. 2 Flow Chart for anomaly detection

第一步，执行驱动类 DPMMDriver，完成任务的初始化。包括 MapReduce 框架的配置以及上文中描述的样本数据清洗和先验 DPMM 的生成。

第二步，判断是否满足退出条件。如图 2，完成先验模型生成后，会生成当前和上次的模型文件，通过做差与阈值比较模型的变化是否满足稳定的条件。

第三步，通过迭代过程实现模型的求解，分为 DPMMMapper 与 DPMMReducer 两部分。

DPMMMapper 部分继承框架的 map 类，用于将全部环境数据样本点分配给相应的模型。Map 函数首先从 HDFS 中读取农场环境数据文件和原始混合模型。对每条样本数据，计算模型产生样本的概率 *pdf*，作为考量样本点与模型相似度的指标；根据条件概率公式，求模型产生概率 *mixpi[k]*与样本由模型产生的概率 *pdf* 的乘积；其中模型产生的概率就是混合模型的比例系数 π_k；最后，对样本 *x[i]*的 *k* 种可能进行多项分布采样，求取 *x[i]*的所属类标签 *s[i]*。

DPMMReducer 继承框架的 reduce 类，用于更新模型参数，并删除空模型。DPMMReducer 部分读取带有类标签信息的样本数据，根据样本点所属的模型类编号更新模型参数 φ_k，并保留前一次模型参数用于判断退出迭代的条件，并逐条叠加统计每个模型包含的样本点数量，最后清除样本点为空的模型。最后更新迭代次数。

第四步，DPMMCalAccu 部分继承自 map 类，功能是根据 1.1 节查阅得到的番茄生长适宜指标先对类簇进行异常筛选，区分出正常类簇和异常类簇；并采用纯度计算来评价聚类效果，具体为逐条筛选异常数据，对正常类簇中的异常数据统计求取漏分率，对异常类簇中的正常数据统计求取错分率，最后剩下的部分即为正确率。

0	18.75328	3.53521	87.24564	6.54674	20124	true
1	19.34783	4.05113	87.66334	6.22345	28214	true
2	19.72352	4.51356	87.62146	8.17556	35236	true
3	20.52584	4.51462	80.88265	17.12334	42078	false
4	21.53567	5.75792	75.41623	21.32345	45682	false

(a) 异常检测结果

2	17.09	85.05	04-02-2015	02:45:00
2	17.19	84.05	04-02-2015	02:50:00
2	17.11	84.95	04-02-2015	02:55:00
2	17.01	85	04-02-2015	03:00:00
2	17.09	85	04-02-2015	03:05:00
2	17.19	84.05	04-02-2015	03:10:00
3	17.20	84	04-02-2015	03:15:00
2	17.01	84.95	04-02-2015	03:20:00
2	17.09	85	04-02-2015	03:25:00
2	17.19	84.05	04-02-2015	03:30:00
3	17.20	84	04-02-2015	03:35:00
2	17.01	84.95	04-02-2015	03:40:00
2	17.09	84.05	04-02-2015	03:45:00
3	17.19	84	04-02-2015	03:50:00
3	17.11	84	04-02-2015	03:55:00
2	17.01	84.95	04-02-2015	04:00:00
2	17.09	84.05	04-02-2015	04:05:00
3	17.19	84	04-02-2015	04:10:00

(b) 部分样本聚类结果

图 3 聚类结果
Fig. 3 Clustering Result

异常检测程序结束后，系统输出异常检测结果和聚类结果，并返回错分率、漏分率、正确率作为效果评价指标。样本聚类后，各个模型的最终参数、模型包含的样本个数和模型异常性都记录在异常检测结果中。图 3(a)中第一列为模型类标签，中间四列为二维正态模型的参数，依次为温度均值、温度标准差、湿度均值、湿度标准差，最后两列为各模型样本数和说明异常性的布尔量，true 表示类簇正常，反之为 false。

图 3(b)为样本点聚类结果，图中第一列为样本点最终所属的类标签，对应图 3(a)第一列，图中后四列分别为样本温度和湿度的属性值以及样本采集的时间。

3 实验测试结果与分析

3.1 运行环境的搭建与设置

本文的计算机集群硬件及软件版本配置情况如表 2 所示，该集群由三台计算机组成，名称如表 2 Node 一栏，MartinMaster 节点，同时作为 NameNode 与 DataNode，另外两台则只作为 DataNode 工作。

表 2 集群软硬件信息
Tab. 2 Cluster Software and Hardware Information

Node	MartinMaster	MartinPoint1	MartinPoint2
Processor	AMD FX(tm) 8300 Eight @3.30 GHz	Pentium(R) Dual E5300 @2.60 GHz	Pentium(R) Dual E2180 @2.00 GHz
Memory	16G DDR3	4G DDR3	6G DDR2
Storage	2TB	1TB	1TB
OS	Ubuntu14.05 (desktop edt)		
JDK	Jdk1.8.0_45		
Hadoop	Hadoop-2.6.0		
Mahout	Mahout-0.5		

Hadoop 底层由 java 实现，二代 Hadoop 要求在 linux 上运行且 jdk1.7 以上版本，所以为三台机器安装了 ubuntu 与 jdk1.8，在运行 Hadoop 时，三台机器要多次互相操作，本文对三台机器分别配置了 SSH(Secure Shell)远程控制协议，并相互交换了密钥，然后对 MartinMaster 安装配置了 Hadoop 并复制到另外两台上，最后安装了机器学习包 Mahout 用于调用正态分布模型。

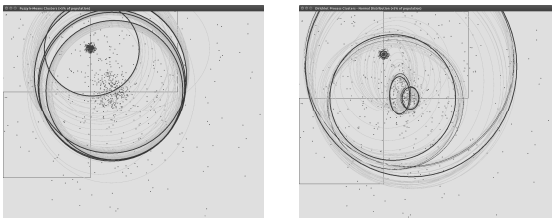
3.2 聚类对比与异常检测分析

实验测试分为两部分，首先通过小数据集聚类对比分析了分布式下 Fuzzy K-Means(FKM)与 DPMM 聚类的特点；然后通过大数据集异常检测分析实验对环境数据进行异常分析。

3.2.1 小数据集聚类对比实验

FKM 是一种结合模糊集与 K 均值的聚类，聚类的结果以隶属度的形式划归相应的数据簇，所以数据簇之间有交集是 FKM 与 DPMM 的相同点。演化自 K 均值的 FKM 是种基于划分的聚类，除了区分特征的方式不同外，DPMM 相较 FKM 的另一区别是 DPMM 类簇数不固定[14]。

如图 4(a)和(b)所示，为两种分布式聚类的结果。其中，线圈所示为最终的聚类簇。为突出 DPMM 的特点，该部分统一采用服从正态分布的随机样本集，样本点数 1 000。FKM 隶属度收敛阈值设为 0.001，模糊系数为 3，DPMM 集中度参数设为 1，采用正态分布作为基准分布。FKM 设定聚类簇为 5 个，DPMM 初始聚类簇为 10 个，最终收敛到 5 个。FKM 耗时 18 min，DPMM 耗时 24 min。



(a) FKM 实现 (b) DPMM 实现

图 4 两种分布式聚类结果
Fig. 4 Two Kinds of Distributed Clustering Results

对比图 4(a)(b)，(b)中 DPMM 选取的基准分布与样本点服从的分布同为正态分布，DPMM 可以在稀疏分布的样本点中区分出密集的样本簇，还能区分出呈椭圆状的二维正态分布点簇，针对该数据集，(b)的聚类效果明显较好。由此，针对不同的数据集为混合模型选择不同的基准分布，测试聚类效果，可以达到拟合甚至复现数据集的效果。所以，模型聚类也是重要的数据分析工具，然而，DPMM 复杂的计算也是其求解缓慢的主要原因。

3.2.2 大数据集异常检测分析

大数据集异常检测分析，该部分对农场数据集的聚类结果进行异常检测分析。图 5 为包含 45 000

余条样本的二维数据集聚类结果。最终数据集被分成 5 个类簇，模型参数依次为 μ_1，σ_1，μ_2，σ_2。

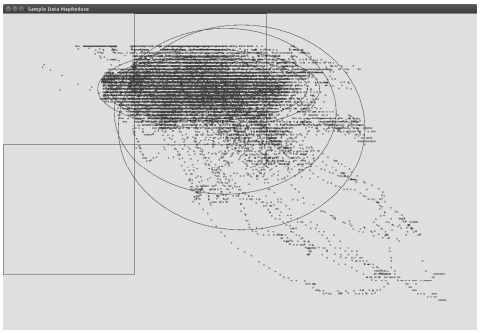


图 5　环境数据聚类结果
Fig. 5　Clustering Result of Environmental Data

图 5 纵坐标为相对湿度，横坐标为温度。该次聚类的漏分率、错分率、正确率依次为 5.21%、3.38%、91.41%，也就是约 41 700 条样本实现了正确分配，约 3900 条样本为漏分点和错分点。图 3(a)中的类簇 0、1、2 为正常类簇，类簇 3、4 为异常类簇，对照参考图 5，可以看出类簇之间是逐级包含的关系。所以异常点分布在类簇 2 以外的类簇 3、4 的边沿部分。

观察图 3(a)的类簇 3、4 的湿度均值和标准差，类簇 3 的湿度(%)为 80.88±17.12，类簇 4 湿度(%)为 75.41±21.32，两者的湿度跨度范围较大，故超出了番茄生长适宜指标的范围。

结合图 5 分析，首先观察图 5 上部，可以看到上部的样本紧贴湿度达到 100%的纵坐标分布，参考图 3(a)的异常检测结果，这些点划归类簇 4，相对湿度过高，空气中的水蒸气将出现液化现象，超出了适宜湿度范围，长期处于该状态会阻碍植株的呼吸作用，造成番茄缺氧。

然后，观察图 5 的左上角，样本点的密集主体位于该区域，对应番茄生长的适宜环境指标，说明果实生长的过程长时间控制在该环境中[15]。

观察图 5 右下侧，可以看到少部分样本点逐渐偏离最适番茄生长环境数据的主体区域，朝向高温低湿方向分布，并且在低湿度的离群点区域温度越高的部分，样本点更密集，说明较高温度的采样点时刻的蒸发量更大，相对湿度总体更低。由于采样间隔为五分钟，最右下侧呈带状分布的样本点所属的时间段，这些样本点必定是连续时间段内的采样点，应当注意，较高的温度可能滋生病虫害，并且湿度较低，长时间缺水干燥会影响果实饱满度。

4　结论

本文对目前农场环境数据，给出了一种分布式环境下基于分布式聚类的异常检测方法，通过搭建分布式平台，并在平台中部署基于模型的 DPMM 聚类完成对农场环境数据的聚类，通过将聚类结果与番茄最适环境指标对比，筛选异常环境数据，实现了异常检测。并通过实验验证实现的有效性。

参考文献：

[1] M. G Ananthara, Dr. T. Arunkumar, Ms. R. Hemavathy. CRY – An improved Crop Yield Prediction model using Bee Hive Clustering Approach for Agricultural data sets [C]//International Conference on Pattern Recognition, Informatics and Mobile Engineering, 2013. India: IEEE, 2013.

[2] 时雷, 马新明. 基于物联网的小麦生长环境数据采集与数据挖掘技术研究[D]. 河南: 河南农业大学, 2013: 61-64.
Shi Lei, Ma Xinming. Research on Data Collection and Data Mining of Wheat Growth Environment based on the Internet of Things[D]. Henan Agricultural University, 2013:61-64.

[3] 潘渊洋, 李光辉, 徐勇军. 基于 DBSCAN 的环境传感器网络异常数据检测方法[J]. 计算机应用与软件 2012, 29(11): 69-73.
Pan Yuanyang, Li Guanghui, Xu Yongjun. Abnormal Data Detection Method for Environment Wireless Sensor Networks Based on DBSCAN[J].Computer Applications and Software, 2012, 29(11): 69-73.

[4] 唐成华, 刘鹏程, 汤申生, 等. 基于特征选择的模糊聚类异常入侵行为检测[J]. 计算机研究与发展, 2015, 52(3): 718-728.
Tang Chenghua, Liu Pengcheng, Tang Shensheng, et al. Anomaly Intrusion Behavior Detection Based on Fuzzy Clustering and Features Selection[J].Journal of Computer Research and Development, 2015, 52(3): 718-728.

[5] Fox E B, Choi D S, Willsky A S. Nonparametric Bayesian methods for large scale multi-target

DOI: 10.16182/j.issn1004731x.joss.201712014

tracking[C]//Proceedings of the 40th Asilomar Conference on Signals Systems and Computers. Pacific Grove, USA: IEEE, 2006.

[6] Orbanz P, Buhmann J M. Nonparametric Bayesian image segmentation[J]. International Journal of Computer Vision (S0920-5691), 2008, 77(13): 25-45.

[7] 刘禅. 温室番茄生长期水分诊断研究[D]. 北京：中科院研究生院, 2011: 31-34.
Liu Chan. Research on Water Diagnosis of Greenhouse Tomato during Growth Period[D]. Beijing: Graduate University of Chinese Academy of Sciences, 2011: 31-34

[8] 何明胜. 基于 GPU 的 Dirichlet 算法并行计算设计与实现[D]. 北京: 北京邮电大学, 2012: 17-19.
He Mingsheng. Parallel Computing Design and Implementation for Dirichlet Algorithm Based on GPU[D]. Beijing: Beijing University of Posts and Telecommunications, 2012: 17-19.

[9] 周建英, 王飞跃, 曾大军. 分层 Dirichlet 过程及其应用综述[J]. 自动化学报, 2011, 37(4): 389-407.
Zhou Jianying, Wang Feiyue, Zeng Dajun. Hierarchical Dirichlet Processes and Their Applications: A Survey[J]. ACTA AUTOMATICA SINICA, 2011, 37(4): 389-407.

[10] Nicole L. B, Lishu L. Clustering digital forensic string search output[J]. Digital Investigation (S1742-2876). 2014, 11(4): 314-322.

[11] Lai Y, Zhongzhi S, Li D.Xu, et al. DH-TRIE Frequent Pattern Mining on Hadoop using JPA[C]. International Conference on Granular Computing, 2011. Taiwan: IEEE, 2011.

[12] Yonghong X, Aziguli W, Yantao W, et al.. Implementation of Time Series Data Clustering Based on SVD for Stock Data Analysis on Hadoop Platform[C]//9th Conference on Industrial Electronics and Applications, 2011. China: IEEE, 2014.

[13] Wentao F, Nizar B. Infinite Dirichlet Mixture Model and Its Application via Variational Bayes[C]//10th International Conference on Machine Learning and Applications, 2011. USA: IEEE, 2011.

[14] 杨来, 史忠植, 梁帆, 等. 基于 Hadoop 云平台的并行数据挖掘方法[J]. 系统仿真学报 2013, 25(5): 936-944.
Yang Lai, Shi Zhongzhi, Liang Fan, et al. Parallel Approach in Data Mining Based on Hadoop Cloud Platform[J]. Journal of System Simulation, 2013, 25(5): 936-944.

[15] 林兰芬, 于鹏华, 李泽洋. 基于聚类的农产品流通物联网感知数据时空可视化技术[J]. 农业工程学报, 2015, 31(3): 228-235.
Lin Lanfen, Yu Penghua, Li Zeyang. Using clustering to visualize spatial-temporal internet of things data in process of agricultural product circulation[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(3): 228-235.

(上接第 3034 页)

[9] Pioneer Hi-Bred International Inc. Method and System for Digital Image Analysis of Ear Traits[P].United States：13/192291, 2013-07-16.

[10] 王传宇, 郭新宇, 吴升, 等. 基于计算机视觉的玉米果穗三维重建方法[J]. 农业机械学报, 2014, 45(9): 274-279.
Wang Chuanyu, Guo Xinyu, Wu Sheng, et al. Three dimensional reconstruction of maize ear based on computer vision[J]. Transactions of the CSAM, 2014, 45(9): 274-279.

[11] WahabzadaMirwaes, Paulus Stefan, KerstingKristian, et al. Automated interpretation of 3D laserscanned point clouds for plant organ segmentation[J]. BMC Bioinformatics (S1471-2105), 2015, 16: 248.

[12] Lange C, Polthier K. Anisotropic smoothing of point sets[J]. Computer Aided Geometric Design (S0167-8396), 2005, 22(7): 680-692.

[13]李宝, 程志全, 党岗, 等. 三维点云法向量估计综述[J]. 计算机工程与应用, 2010, 46(23): 1-7.
Li Bao, Cheng Zhiquan, Dang Gang, et al. Survey on normal estimation for 3D point clouds[J]. Computer Engineering and Applications, 2010, 46(23): 1-7.

[14] 孟海东, 张玉英, 宋飞燕.一种基于加权欧氏距离聚类方法的研究[J]. 计算机应用, 2006, 26(增2): 152-153.
Meng Haidong, Zhang Yuying, Song Feiyan.A study of a weighted Euclidean distance clustering method[J]. Computer Applications, 2006, 26(S2): 152-153.

[15] 刘进, 武仲科, 周明全. 点云模型分割及应用技术综述[J]. 计算机科学, 2011, 38(4): 21-24.
Liu Jin, Wu Zhongke, Zhou Mingquan. Review of point cloud segmentation and application[J]. Computer Science. 2011, 38(4): 21-24.