基于 K-Means 聚类的农产品价格异常数据检测①

韩 琳, 吴华瑞, 顾静秋

(北京农业信息技术研究中心, 北京 100097)
(国家农业信息化工程技术研究中心, 北京 100097)
(农业部农业信息技术重点实验室, 北京 100097)
(北京市农业物联网工程技术研究中心, 北京 100097)

摘 要: 全国各地各个年份的农产品市场价格数据量庞大, 而海量的农产品的市场价格数据中无可避免存在超出市场正常价格范围的异常价格元素, 这对搜索引擎农产品市场价格的统计分析与预测造成了影响. 从市场价格大数据中发现离群点并计算出价格边界成为有待解决的问题, 为此, 本研究在数据挖掘聚类技术 K-means 算法的基础上, 提出了基于 K-means 聚类的农产品市场价格异常数据检测并计算出农产品市场价格边界, 测试及实践结果表明该方法提高了聚类的精确率和稳定性, 实现了价格异常点检测与价格边界的计算.

关键词: 海量农业数据; 聚类; K-means 算法; 离群点; 市场价格; 异常检测

Abnormal Agricultural Price Data Detection Based on K-Means Clustering

HAN Lin, WU Hua-Rui, GU Jing-Qiu

(Beijing Agricultural Information Technology Research Center, Beijing 100097, China)
(National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China)
(Key Laboratory of Agricultural Information Technology of Ministry of Agriculture, Beijing 100097, China)
(Research Center of Beijing Agricultural IOT Engineering Technology, Beijing 100097, China)

Abstract: Vertical search engine of the ministry of agriculture needs to collect the market price data of agricultural products in various years from all over the country. It can not be avoided that the massive agricultural market price data has abnormal price point, which has an impact on the analysis and forecast of the agricultural market price. It needs to be solved to find market price data outliers and calculates the price boundary. Therefore, on the basis of the traditional data mining clustering K-means algorithm, this study achieves the outlier data detection and calculation of the boundary of the price of agricultural products, test and practice results show that the method improves the clustering accuracy and stability and achieves the calculation of the price of outlier detection and border price.

Key words: massive agricultural data; clustering; K-means algorithm; outlier; market price; abnormal detection

1 引言

随着农业垂直搜索引擎的发展, 农业大数据的分析和处理变得越来越重要. 搜索引擎提供全国各个市场的农作物价格信息, 方便农民对农作物市场价格信息和趋势的把握. 但是全国的农产品市场价格数据量庞大, 品种繁多, 海量的价格信息中可能存在超出市场范围内的异常价格, 这些少量的异常价格数据会影响搜索引擎对农产品市场价格的分析与判断, 如何使用数学分析方法从这些数据中检测出异常价格并计算出每个品种的市场价格边界具有了特殊意义.

目前异常数据挖掘方法有基于统计的方法、基于距离的方法, 基于偏离的方法、基于密度的方法和基于聚类的异常点检测算法等. 聚类[1-4]是对于静态数据分析的一门技术, 是通过静态方法将物理或抽象对象的集合分成由类似的对象组成多个类的过程, 使得同一簇中的对象之间具有较高的相似度, 而不同簇中的

① 基金项目:国家科技支撑计划(2013BAJ10B15)
收稿时间:2016-06-16;收到修改稿时间:2016-07-25 [doi:10.15888/j.cnki.csa.005641]

对象差别很大. 它是分析数据并从中发现有用信息的一种有效手段. 目前常用的聚类算法包括划分法[5]、层次法[6]、基于密度的方法[7]、基于网格的方法[8]和基于模型的方法[9]. 其中基于划分方法的 K-means 算法的聚类技术具有以下特点: 产生的离群点集和它们的得分可能非常依赖所用的簇的个数和数据中离群点的存在性; 聚类算法产生的簇的质量对该算法产生的离群点的质量影响非常大, 但该算法发现离群点是高度有效的, 且因为其简单、快速处理大规模数据的有效性而得到广泛的应用.

农产品市场价格数据信息达到 1300 万条以上, 针对大规模的市场价格数据, 本文提出一种基于 K-means 的市场价格计算方法, 该算法满足了 K-means 应用数据信息庞大的基础, 有效提高了聚类精度, 解决了实际问题.

2　传统K-means 算法

传统 K-means 算法[10-12]的基本思想是: 随机地选择 k 个对象, 每个对象初始代表了一个聚类中心, 对剩余的每个对象根据其与各个聚类中心的距离, 将它赋给最近的聚类, 然后重新计算每个聚类的平均值, 作为新的聚类中心. 不断重复这个过程, 直到准则函数 E 收敛.

$$E=\sum_{i=1}^{k}\sum_{x=c_i}\left|x-\overline{x_i}\right|^2$$

其中, E 为数据集中所有对象的平方误差和, K-means 算法步骤如下:

① 随机选择 k 个数据元素作为初始聚类中心;

② 计算每个数据元素与聚类中心的距离, 根据距离将它们分到距离最近的簇;

③ 重复计算每个聚类中元素的平均值, 更新聚类中心;

④ 重复②和③, 直到收敛.

3　基于K-means的市场价格异常数据检测算法

在传统K-means算法的基础上, 做了以下改进: 首先, 选择初始种子结点. 由于市场价格需要计算最小和最大两个边界值, 且依据数据样本自身分布特点, 经过反复试验, 当K=3时, 聚类结果簇数量比例明显, 因此确定K值为3[13]. 在选择初始种子元素时, 将每个农作物品种的价格数据进行排序, 选取3个不同的从小到大排列的价格元素来作为种子元素, 以此来获取三个聚类中心构成的堆, 分别为最小堆、中间堆和最大堆. 基于离群点检测的K-means 算法的基本思想是: 每次聚类完成之后, 然后再利用改进后的K-means算法对最小堆进行聚类计算, 一直向左遍历, 直到检测到最小堆中异常数据为止; 当检测到最小堆中异常价格数据后回到*ROOT*结点, 对最大堆进行聚类计算, 直到检测到最大堆中异常数据为止. 如果遍历4层后依然没有找到则停止遍历. 最后得到一个类似树的结构, 算法的思想如图1所示.

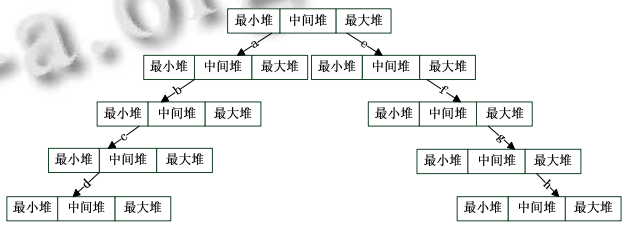

图 1　基于 K-means 市场价格边界算法树结构

其中, 针对市场价格数据, 应用一次 K-means 算法构成一个结点, 由于 K-means 算法先将价格数据从小到大排列, 然后围绕 K=3 进行聚类计算, 最后聚合生成 3 个堆. 因此按照堆中价格大小, 把这三个堆分为最小堆、中间堆和最大堆. 堆与堆之间离群点数据的偏离程度则通过计算价格个数的比值 a, b, c, d, e, f, g 和 h 来体现. 分别计算 $Sum(a+b+c+d)$ 和 $Sum(e+f+g+h)$的值, 如果 $Sum(a+b+c+d)$大于一定的阈值, 检测出最小堆的异常数据, 获得最小价格边界, 同理, 如果 $Sum(e+f+g+h)$大于一定的阈值, 则检测出最大异常数据, 获得最大价格边界. 其中, a, b, c 和 d 的计算公式相同, 均为父结点最大堆和中间堆数量之和与最小堆的数量之比; e, f, g 和 h 的计算公式相同, 均为父结点最小堆和中间堆数量之和与最大堆的数量之比. 经过反复试验, 阈值界定为 78.3. 即当 $Sum(a+b+c+d)$或者 $Sum(e+f+g+h)$大于阈值 78.3, 即表示价格超过了异常数据范围.

$$a|b|c|d=\frac{Count(\text{最大堆}+\text{中间堆})}{Count(\text{最小堆})}$$

$$e|f|g|h=\frac{Count(\text{最小堆}+\text{中间堆})}{Count(\text{最大堆})}$$

算法具体步骤如下:

① 对某个农作物品种价格进行排序, 选取 3 个不

同数据对象作为初始聚类中心;

② 计算其他数据对象与选取的数据对象间的距离;

③ 计算每个非离群点数据对象与聚类中心的距离, 根据距离将对象划分到距离最近的聚类;

④ 重复计算每个聚类中对象的平均值, 更新聚类中心;

⑤ 重复③和④, 直到准则函数 E 收敛;

⑥ 统计最大堆、中间堆和最小堆的价格数据个数;

⑦ 程序从根结点向左遍历, 计算最大堆与中间堆价格个数与最小堆得个数之比, 重复以上过程, 将以上比值相加, 超过一定阈值则找到最小价格异常点;

⑧ 再从根结点向右遍历, 计算最小堆和中间堆价格数据个数与最大堆个数之比. 重复以上过程, 将这些比值相加, 超过一定阈值则找到最大价格异常点.

算法描述如下. 以下代码使用 Java 程序编写, 其中 KMeans 类封装 K-means 算法, Cluster 方法实现对数据集 *s* 排序并以 *k*=3 来聚合计算.

Step 1. 首先使用K-means算法将*s*数据处理变成*g*数组.

```
g = new KMeans().cluster(s, k);
root = g;
t = getBorderPrice(g,k);
//统计每个簇的个数、最小和最大值到t数组
```

Step 2. 循环迭代左边界

```
int num=1;
double sum=0.0;
while(num!=5){
  if(t[0][0]>0){
//如果个数比值之和大于阈值78.3退出循环
    if(sum>=100) {
        min = t[1][1];
        break;
    }
    if(t[0][1] = = t[0][2]){
//最大最小值相等则无法继续聚类跳出循环
        min = t[0][1];
        break;
     }
    if(num = = 4){
//迭代4次还未找到异常数据跳出循环
        min = t[0][1];
        break;
    }
  g = new KMeans().cluster(g[0], k);
//对左边界继续聚类计算
  t = getBorderPrice(g,k);
  num++;
//计算中间堆和最大堆个数与最小堆数量比值之和
  sum = sum + (t[1][0] + t[2][0])/t[0][0];
 }
```

Step 3. 与 Step 2 同理循环迭代计算右边界.

4 农业市场价格异常数据分析

在部署好的环境中进行相关的实验, 验证本文提出的基于K-means农产品市场价格异常数据检测方法, 本文实验环境为: CPU 为 Intel i5-4590(3.3GHZ)、内存为 8GB、操作系统为 Window7, 编程语言为 Java. 农产品市场价格数据包括副食品、瓜果、粮棉油糖、蔬菜和水产品等类别, 全国各个省市地区历年的农产品市场价格数据, 例如茄子、生姜、芹菜、山药、大葱、鸡蛋、牛肉、白鲢活鱼、粳米、菠菜、苹果、和菠萝等 11914 个品种的 1300 万条市场价格数据, 测试与验证本文改进 K-means 算法实现统计市场价格边界计算方法针对某个品种算法准确性以及随着不同品种数据量的增加时算法查找异常价格点的准确性.

4.1 白鲢活鱼价格异常数据检测与分析

以上述农产品市场价格数据信息水产品类别中白鲢活鱼为例检测算法. 全国各地各个批发市场历年的白鲢活鱼市场价格数据组成数据对象集 *S*, 该集合包含 50447 条记录.

输入: 全国各地各个批发市场历年的数据白鲢活鱼 50447 个数据对象集 *S* 和聚类数 *k*=3, 将数据对象集合 *S* 中的价格数据提取到数组中, 以便下一步算法的处理和应用. 其中价格单位均为元/公斤.

表 1 白鲢活鱼数据集

地区	批发市场	种类	名称	价格	日期
安徽	宿州砀山	水产品	白鲢活鱼	7.0	2014.1
湖北	孝感	水产品	白鲢活鱼	5.8	2012.2
湖南	常德	水产品	白鲢活鱼	6.0	2010.4
湖南	邵阳	水产品	白鲢活鱼	600.0	2011.5
江苏	常州	水产品	白鲢活鱼	6.5	2012.6

输出: 遍历树结构中每个结点最小堆、中间堆和最大堆的价格数据数目、最小值和最大值, 由于a+b+c+d=21.2<阈值 78.3, 则未找到最小异常数据, 最小边界是 2.2, 由于 e+f=841>阈值 78.3, 则找到最大异常数据 600-620, 从而得到最大边界是 72. 由以上过程可以得到白鲢活鱼的最小边界值: 2.2, 最大边界值: 72.0. 实验结果、最小堆和最大堆的层次迭代如下表 2 和表 3 所示. 最终得到的最小和最大边界以及异常数据如表 4 所示. 其中价格单位为元/公斤.

表 2　最小堆层次迭代过程

堆	数量	最小价格	最大价格	比率
小	40303	202	9.11	a=7673+2471)/40303=0.2
中	7673	9.13	15.2	
大	2471	15.4	620.0	
小	7645	2.2	5.3	b=(16993+15665)/7645=4
中	16993	5.37	6.85	
大	15665	6.9	9.11	
小	1205	2.2	4.07	c=(2178+4262)/1205=5
中	2178	4.1	4.7	
大	4262	4.73	5.3	
小	91	2.2	3.1	d=(534+580)/91=12
中	534	3.2	3.8	
大	580	3.83	4.07	

表 3　最大堆层次迭代过程

堆	数量	最小价格	最大价格	比率
小	40303	2.2	9.11	e=40303+7673)/2471=19
中	7673	9.13	15.2	
大	2471	15.4	620.0	
小	2344	15.4	21	f=(2344+124)/3=822
中	124	10.0	72.0	
大	3	600.0	620.0	

表 4　白鲢活鱼价格边界以及异常数据

名称	左边界	右边界	异常数据
白鲢活鱼	2.2	72.0	600.0-620.0

4.2 农产品价格异常数据批量检测与分析

采用基于 K-means 聚类的农产品市场价格异常数据检测方法对农产品市场价格数据信息进行批量计算与分析, 全国各地各个省市批发市场农产品市场价格数据如表 5 所示, 计算出的市场价格边界表以及异常价格数据如表 6 所示. 从海量品种中计算异常数据, 经过反复测试, 阈值 78.3 能很好地统计海量品种的异常数据, 针对白萝卜, 有一条价格数据 32 缺漏没有记为异常数据, 对于这条数据, 右边界计算时 e+f+g+h<阈值 78.3 时, 该数据是正常数据, 但其实它是异常的, 我们认为它是缺漏异常数据, 误判数据则是本应为正常数据, 而被误判为异常数据. 本文实现的算法数据量统计如表 7 所示. 其中价格单位为元/公斤.

表 5　农产品市场价格表

地区	批发市场	种类	名称	价格	日期
北京	八里桥	蔬菜	长茄	6.50	2013.1
安徽	宿州砀山	水产品	鲢鱼	2.00	2014.1
山西	长治紫坊	副食品	牛肉	44.00	2015.4
河北	永年中原	瓜果	苹果	5.00	2013.1
湖北	武汉	蔬菜	洋葱	1.30	2013.4
湖南	吉首	粮棉油糖	黑豆	12.00	2010.4
天津	武清大沙	蔬菜	萝卜	0.80	2011.5
云南	曲靖师宗	副食品	鸡蛋	8.00	2012.6
河南	郑州刘庄	蔬菜	蒜苔	2.80	2015.5
浙江	杭州	粮棉油糖	粳米	3920.0	2010.3

表 6　农产品异常价格数据表

名称	左边界	右边界	异常数据
鲢鱼	2.20	72.00	600.00
茄子	0.04	39.00	400.00
萝卜	0.16	7.00	104.00
鸡蛋	2.90	34.0	60.00
牛肉	20.00	117.00	240.00
黑豆	3.85	21.00	51.00
芹菜	0.04	70.00	380.00
生姜	0.45	44.00	950.00
粳米	1.98	775.00	3920.0
山药	0.05	95.00	701.00
菠菜	0.04	63.00	340.00

表 7　算法准确度统计表

品种	测试数据量	异常数据量	缺漏异常数据	误判数据
生姜	179857	165	0	0
白萝卜	208803	119	1	0
菠菜	219187	141	0	0
白鲢鱼	51771	3	0	0
茄子	222938	36	0	0
芹菜	236639	205	0	0
牛肉	80216	143	0	0
山药	107758	1	0	0

4.3 本文实现的算法与传统 K-means 算法对比实验

本文实现的算法相比传统 K-means 算法做了以下改进: 针对初始数据进行排序, 选择三个不同的从小到大排列的数据作为种子结点, 避免选择离群点作为初始聚类中心, 构造出三个堆以从小到大顺序排列的

堆, 方便了边界价格的检索与计算. 分别统计在不同数量级别本文实现的算法与传统 K-means 算法在对市场价格数据进行聚合计算时, 两个算法计算的准确率, 如下图 2 所示, 横坐标表示数据集农产品市场价格数据规模, 纵坐标表示异常价格检测的准确度, 本文实现的算法随着农产品市场价格数据集中数据规模的增大, 逐渐呈现出稳定增长的趋势. 而传统 K-means 算法则因为聚类中心的不稳定影响了结果的稳定性.

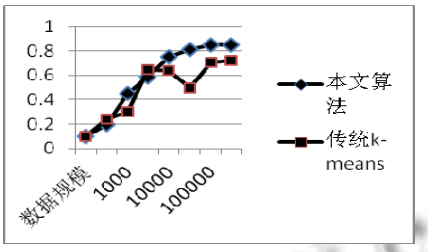

图 2 本文算法与传统 K-means 对比图

5 结语

针对农业搜索引擎的数据库中农产品数据信息庞大的特点, K-means 算法快速、简单, 对大数据集有较高的效率并且是可伸缩性的, 时间复杂度近于线性, 非常适合挖掘大规模数据集. 因此, 为解决根据已有海量的农产品市场价格数据来检索异常价格和计算价格边界的问题, 本文采用数据挖掘聚类技术 K-means 技术进行分析, 改进传统 K-means 算法并提出了基于 K-means 的农产品市场价格异常数据检测方法. 实验结果表明, 改进的算法可以很好地筛选出离群点数据, 找到异常价格元素从而获得农产品市场价格边界. 目前该算法已经应用于农业智能搜索引擎市场价格分析模块, 效果显著.

参考文献

1 数据挖掘概念与技术.范明,孟小峰译.北京:北京机械工业出版社,2010.
2 傅德胜,周辰.基于密度的改进K均值算法及实现.计算机应用,2011,31(2):432–434.
3 刘伟,刘露,陈荦等.基于离群点检测的 K-means 算法.渤海大学学报,2014,(1):34–38,48.
4 蒋盛益,李庆华.一种增强的K-means聚类算法.计算机工程与科学,2006,(11):56–59.
5 黄韬,刘胜辉,谭艳娜.基于 K-means 聚类算法的研究.计算机技术与发展,2011,21(7):54–57.
6 汪中,刘贵全,陈恩红.一种优化初始中心点的K-means算法.模式识别与人工智能,2009,22(2):299–304.
7 Khan SS, Ahmad A. Cluster center initialization algorithm for K-means clustering. Pattern Recogintion Letters(S0167-8655), 2004, 25(11): 1293–1320.
8 MacQueen. Some methods for classification and analysis of multivariate observationsC. Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley. University of California Press. 1967.
9 应磊,王儒敬.异常农产品价格数据检测.计算机系统应用,2010,19(4):178–179.
10 韩凌波,王强,蒋正锋,等.一种基于改进的 K-means 初始聚类中心选取算法.计算机工程与应用,2010,46(17):150–153.
11 张玉芳,毛嘉莉,熊忠阳.一种改进的 K-means 算法.计算机应用,2003,8(23):31–34.
12 Elio L, Edgar A. Parallel algorithms for distance-based and densit-based outliers C. Proc of International Conference on IEEE. 2005. 767–776.
13 张健沛,杨悦,杨静,等.基于最优划分的 K-Means 初始聚类中心选取算法.系统仿真学报,2009,21(9):2586–2590.