ATLANTIS PRESS
Advances in Intelligent Systems Research (AISR), volume 152
Fifth Symposium of Risk Analysis and Risk Management in Western China (WRARM 2017)

K-means Cluster's Classification of Grade of Regional Strong Precipitation Events

Tiangui Xiao Tianyao Shen Fengrong Jing Chao Wang Ding Chen Yaze Wu Yao Huang

Chengdu University of Information Technology Atmospheric Sciences Academy, Chengdu 610225, China

Abstract

Based on the historical precipitation data of the stations in Sichuan Province from 1961 to 2013, 636 heavy precipitation processes were summarized by statistical analysis. First, the daily average of heavy rainfall was selected. Precipitation, rainfall intensity, coverage and duration were used as the index factors of grade division. Secondly, by analyzing the interaction between the index factors, the K-means method was used to analyze the similarity between the distances The heavy precipitation process is divided into five grades by 636 heavy precipitation. Finally, based on the clustering results, a hierarchical model of the operationalized heavy rainfall event in Sichuan Province is established for the risk assessment of the severe precipitation disaster and the forecasting and forecasting of the disaster.

Keywords

Regional; Strong Precipitation; K-means; Cluster

区域性强降水事件等级的 K-means 聚类划分

肖天贵 申天瑶 敬枫蓉 王超 陈丁 吴亚泽 黄瑶

成都信息工程大学大气科学学院, 成都 610225, 中国

摘要：基于已信息化的 1961—2013 年四川省各站点逐日的历史降水资料及 636 次强降水过程，研究了强降水事件等级划分。首先选取强降水过程的日平均降水量、降水强度、覆盖范围和持续时间四个指标作为等级划分的指标因子，其次通过对各指标因子之间相互作用的分析，提出以 K-means 方法根据各次过程指标间距离的相似性对强降水过程进行聚类分析划分等级，将 636 次强降水过程划分为五个等级。基于聚类结果建立四川省可业务化强降水事件等级模型，用于强降水灾害的风险评估及灾害的预报预警。

关键词： 区域性；强降水；k-means；聚类分析

1.引言

在全球气候变化的大背景下，极端气象灾害频发，如暴雨洪涝、干旱、高温热浪、低温冷害等日益频繁，对于地处青藏高原与我国东部平原的阶梯过渡区的四川盆地，海拔高度差大，地质构造复杂，受季风影响较大，更是我国自然灾害频繁发生且危害最严重的省份之一。就致灾气象事件而言，强降水导致的灾害尤为显著，如由强降水引发的洪涝、渍涝、崩塌、滑坡、泥石流和水土流失等灾害造成了国民经济和人民生命财产的巨大损失。对于四川省而言，每年因区域性强降水形成的突发性和持续性暴雨洪水灾害在致灾气象事件中所占比例高达 40%，已成为经济社会可持续发展的重要制约因素之一，如 2013 年出现的四川历史上罕见的“7.9”强暴雨（都江堰幸福站过程降水量超过 1000mm），不仅造成人民财产的损失，还使多人失去生命。因此加强对强降水致灾事件的等级划分及准确的风险评估的研究是迫切需要的，这是防灾减灾的重要依据。

四川盆地的降水呈现出显著的多尺度变化和极端区域性特征，强降水具有出现频率高、年际变化大、影响大、年内 5-10 月均可发生，以 7-8 月为盛。强降水引发的灾害的种类和程度随地形特点的不同而不同，强降水致灾过程的复杂性使得灾害的风险评估成为一项比较艰难的工作，但是由于灾害的预报预警的需要，强降水致灾的等级划分及风险评估又是必不可少的，科学准确的判断强降水事件的等级，由此做出灾害的风险评估，为当地政府开展防灾工作、确定减灾、救灾方案、制定灾后救援计划等提供有力的科技支撑和科学决策依据，避免造成重大的损失。评判强降水等级的指标不只有降水量，还有其他一些指标可反映强降水的强度、致灾的可能性，通过综合分析讨论，选取其中主要的指标进行等级划分，选取了日平均降水量、降水强度、覆盖范围以及持续时间作为描述强降水事件的指标。目前，分类方法多种多样，但由于各区域的地质构造不同，引起灾害的强降水的指标也不同，所以很难建立一个统一的分类方法。考虑到各次降水过程间相似性的大小各不相同，可将各次过程各指标间的相似性作为类属划分的准则。在各种聚类算法中，K-means 算法对大数据有较高的效率并且是具有可伸缩性的，能保证局部收敛，事先给定聚类数，根据某种准则进行快速聚类，特别是样本分布呈类内团聚状时，可以达到较好的聚类效果。因此本文以 K-means 聚类方法进行多元分类，依据各区域的指标数据库进行分类。分类过程属于无监督分类的范畴，希望运用到四川盆地的区域性强降水的等级划分中，以便进行四川省区域性强降水致灾事件的灾害评估的研究，为当地政府确定减灾、救灾方案、制定灾害救援计划等提供有力的支撑和依据。

2.理论方法

K-means 聚类算法是硬聚类算法的一种，也是典型的基于距离的聚类算法。以欧式距离作为相似度的测度，判断各个样本到聚类中心的欧式距离，以样本点到聚类中心的欧氏距离来迭代优化，根据欧式距离最小来将样本点重新分配给最近的聚类中心，这样将所有的样本划分完之后，新的聚类中心形成。算法采用误差平方和准则函数作为聚类准则函数，一次聚类结束之后判断误差平方和的大

小，若该值变化不大则算法收敛。具体算法如下：

(1) 设样本即被分类的对象为 $X=\{x_1, x_2, x_3\cdots, x_n\}$，其中 n 为样本数，X 中的每个对象又用 m 个描述属性 $A_1, A_2, \cdots, A_m$ (维度)表示，本文所对应的样本为 636 次强降水过程，即 n=636,m 个描述属性为日平均降水量、降水强度、覆盖范围以及持续时间，即 m=4；

(2)选取聚类个数 k，本文要将 636 次强降水过程划分为五个等级，因此令 k=5；

(3) 数据样本 $x_i=\{x_{i1}, x_{i2}, \cdots, x_{im}\}$，$x_j=\{x_{j1}, x_{j2}, \cdots, x_{jm}\}$，其中 $x_{i1}, x_{i2}, \cdots, x_{im}$ 和 $x_{j1}, x_{j2}, \cdots, x_{jm}$ 分别是样本 x_i 和 x_j 对应的 m 个描述指标 $A_1, A_2, \cdots, A_m$ 的具体取值。从数据样本中随机选取 k 个强降水过程作为聚类中心，选取的聚类中心样本的各次过程间的距离尽可能分开，以保证聚类的快速收敛；

(4)样本 x_i 和 x_j 之间的相似度通常用他们之间的距离 $d(x_i, x_j)$ 来表示，他们之间的距离越小，样本 x_i 和 x_j 越相似，差异度越小；距离越大，样本 x_i 和 x_j 越不相似，差异度越大。计算原样本数据的每次过程与聚类中心样本的欧式距离，判断距离大小，将每次过程划分到距离最小的质心类；欧式距离：

$$d(x_i, x_j)=\sqrt{\sum_{k=1}^{d}(x_{ik}-x_{jk})^2} \quad (1)$$

(5)计算新的聚类中心，计算各聚类子集的各指标的平均值作为新的聚类中心；

(6)使用误差平方和准则函数来评价聚类性能，将给定数据集 X,其中只包含描述属性，不包含类别属性。设 x 包含 k（k=5）个聚类子集，各个聚类子集的样本数量为 $n_1, n_2, \ldots\ldots, n_k$，各个聚类子集的均值代表点（聚类中心）分别为 $m_1, m_2, \cdots, m_k$。误差平方和准则函数公式：

$$E=\sum_{i=1}^{k}\sum_{p\in X}\left\|P-m_i\right\|^2 \quad (2)$$

判断准则函数收敛：①质心的值变化不大；②类内的样本变化不大。

(7)若一次迭代之后不收敛，则继续从第（4）步进行迭代，新的聚类中心取代旧的聚类中心，直到收敛为止。

由于 K-means 聚类算法属于硬聚类算法，所以存在以下缺点：一是对初始聚类中心的选取比较敏感；二是对离群点和孤立点较为敏感。

3. 资料及资料处理

3.1.资料选取

用已信息化的 1961～2013 年四川省逐日、逐小时历史降水等气象资料，根据 24 小时降水资料统计从 1961 年到 2013 年的强降水过程，将降水量大于等于 50mm 的降水过程，连续性的降水记为一次暴雨过程，记录一次暴雨过程的发生的时间、范围内的站点及其降水量，选取出现的 636 次暴雨过程。考虑到四川盆地的强降水特点，降水的时空分布极不均匀，夏季强降水中心比较集中，可以考虑分区，强降水致灾的因子和机理都比较复杂，且各种因子间存在相互作用，本文选取主要的因子忽略次要因子，选取日平均降水量、降水强度、覆盖范围和持续时间作为基本指标，以建立可业务化的分区域、分等级的四川

省强降水事件模型。各年强降水次数统计如下表：

表 1 强降水过程及其年份

强降水过程（次/年）	年份（年）
6	2002、2006
7	1976
8	1997、2010
9	1968、1975、1992、1993、2004
10	1972、1982、1986、1994、2008
11	1965、1970、1995、1996、2003、2005 2007、2011
12	1964、1987、1988、1991
13	1961、1967、1973、1974、1979、1990 1998、2001
14	1962、1966、1971、1977、1978、1980 1981、1989、2013
15	1969、1985、1999、2000、2009
16	1983、1984、2012
17	1963

3.2. 强降水评估指标数据处理

① 日平均降水量指标：

$$I_{js}=\frac{1}{n}\sum_{i=1}^{n}p_i \quad (i=1,2,\cdots\cdots，n) \tag{3}$$

其中，n 为区域内一次降水过程中降水量达到暴雨标准的观测站点个数，P_j 为其中第 i 个观测站点在被评估过程中的总降水量（单位：毫米），表征了区域的平均降水状况；

② 降水强度极值指标：

$$I_{qd}=\max(P_i) \quad (i=1,\cdots\cdots，n) \tag{4}$$

其中， max()为取最大值函数，P_j 为一次降水过程中第 i 个观测站点在暴雨过程中的 24 小时观测降水量（单位：毫米），这是一个强降水过程中的极值，用以表征过程降水最大值，值越大，致灾的可能性越大；

③ 覆盖范围指标：

$$I_{fw}=\frac{n}{N} \tag{5}$$

其中，N 为区域内观测站点总数（单位：个），n 为发生暴雨过程不重复（即一次过程多次发生暴雨重复站点记为一次）站点数总数，覆盖范围越广泛，说明此次强降水水汽充沛，可能是由较大的系统控制，致灾的可能性也越大；

④ 降水持续时间指标：

$$I_{tim}=T \tag{6}$$

其中，T 为过程持续时间（单位：天），持续时间较长的强降水过程致灾的可能性很大，可能引起暴雨洪涝、山洪泥石流等灾害现象。

4. 强降水事件等级评估模型建立

4.1. K-means 聚类划分等级

在聚类之前，先将所获取的数据资料（1961-2013 年 636 次强降水过程）进行处理，处理方法如下：通过 3.2 的方法，使用（3）、（4）、（5）和（6）的公式分别计算出各次强降水过程的评估指标按日平均降水量、降水强度、覆盖范围和持续时间的顺序排列，统计每次强降水过程的结束时间，然后与评估指标相对应。由于数据较多，因此使用 Fortran 编译器来处理。规范化的部分数据如下：

表 2 强降水评估指标数据

时间	日平均降水量(mm)	降水强度(mm)	覆盖范围（站点数/156）	持续时间（天）
19610616	75.4875	129.8	5.13E-02	1
19610620	66.41	88.5	6.41E-02	1
19610628	91.30081	306	0.3846154	6
19610706	85.88889	176.1	0.1730769	1
19610714	70.15285	218.8	0.3205128	3
19610724	70.61111	128.2	0.1153846	1
19610728	61.08	87.8	3.21E-02	1

续表 2 强降水评估指标数据

时间	日平均降水量(mm)	降水强度(mm)	覆盖范围（站点数/156）	持续时间（天）
19610802	87.90833	180.6	7.69E-02	1
19610808	51.244	140.3	0.1538462	2
19610814	91.26191	196	0.2564103	2
19610820	80.11875	167.9	0.2628205	2
19610824	71.275	102.1	5.13E-02	1
19610824	57.06667	63.1	3.85E-02	1
19620616	72.61667	97.7	7.69E-02	1

为解决上述 k-means 算法对离群点和孤立点的敏感的缺点，先对要进行分级的数据进行统计，先抽出其中的孤立点，将剩余的数据进行聚类，得到最终的聚类中心，对于之前排除的孤立点，计算孤立点与最终聚类中心的距离，决定孤立点属于哪一类。孤立点的统计如“图 1 强降水的指标因子的分布”所示（限于篇幅，图 1 略）。由以上的统计将如下孤立点抽出：

表 3 孤立点

结束时间	平均降水量	降水强度	覆盖范围	持续时间
19610629	91.30081	306	0.3846154	6
19930730	115.4192	524.7	0.1602564	2
19950824	151.7304	374.3	0.1474359	1
19960728	96.76	410.8	0.1858974	2
20130701	93.45574	415.9	0.3141026	3

再把处理后的数据进行聚类分级。K-means 聚类的关键在于聚类中心的选取，K-means 聚类算法是局部收敛的，聚类中心选得较好，收敛速度就较快，相反，聚类中心选得没那么理想聚类时间就会延长。本文初始聚类中心的选取如下：

表 4 初始聚类中心

	1 级	2 级	3 级	4 级	5 级
平均降雨量/mm	70.33	80.49	90.10	95.50	109.60
降雨强度/mm	88.00	150.30	206.33	241.28	311.76
覆盖范围	0.06	0.10	0.18	0.26	0.34
持续时间/d	1.0	2.0	2.0	3.0	4.0

根据四川省区域降水的特点将此次研究的强降水过程的样本划分为五类，从一级到五级表示从轻度到重度。聚类步骤如下：先计算 n（n=631）个样本与聚类中心 $Z_K(\mathrm{k}=1,2,3,4,5)$ 间的欧式距离，由公式（1）得到样本 X_{i} 与聚类中心 Z_k 间的欧式距离，然后比较距离大小，距离越小，该样本与该聚类中心越相似，则可把样本 X_{i} 划分到第 k 类，重复此过程就将 n 个样本划分为五类了，完成一次聚类过程。例如 1961 年 06 月 16 日的这一次强降水过程，根据强降水过程的数据即可算出其与初始聚类中心 Z_1,Z_2,Z_3,Z_4,Z_5 的欧式距离如下：

表 5 与初始聚类中心的欧式距离

$d(x_1,z_1)$	$d(x_1,z_2)$	$d(x_1,z_3)$	$d(x_1,z_4)$	$d(x_1,z_5)$
42.12	21.13	77.92	116.23	185.15

由此可以判断出 $d(x_1,z_2)$ 最小，所以把此次强降水过程划分到第 2 级，其他样本数据重复此过程，即可将所有的强降水过程划分到各个等级，最后再统计所以强降水过程，即可将样本数据划分为五个等级；再由误差平方和准则函数即公式(2)来计算并判断是否收敛，k-means 聚类方法是可以保证局部收敛的，设第 0 次聚类 E=0，其后的聚类依次计算出 E 值，E 的值是先递减的，递减到最小又开始

增加，此临界值的地方则可判断聚类收敛；若是收敛，划分结束；若是不收敛，继续进行迭代。迭代方法为将划分的各类求平均值作为新的聚类中心，然后继续从（1）开始，直到收敛为止。本文所使用的具体的样本数据在聚类过程当中，通过两次聚类就达到了收敛的目的，两次聚类完成后形成的新的聚类中心与初始聚类中心相差不大，E=205351.4，新的聚类中心如下表所示：

表 6 最终聚类中心

	1 级	2 级	3 级	4 级	5 级
平均降雨量/mm	68.62	79.11	87.96	94.40	100.14
降雨强度/mm	98.60	147.42	202.70	266.79	349.91
覆盖范围	0.078	0.13	0.19	0.22	0.25
持续时间/d	1.2	1.6	2.1	2.5	2.6

4.2. 划分结果

通过两次聚类过程，可将四川省 1963--2013 年间的 636 次强降水过程划分等级，然后再统计划分结果，统计每一等级的强降水过程及其频数，具体每类样本的数量如下表所示：

表7 强降水等级

等级	一级	二级	三级	四级	五级
频数	262	215	103	37	14

根据上述的划分则可列出重度暴雨过程即第五级的详细情况，如下表所示：

表8 等级划分

结束时间	平均降水量	降水强度	覆盖范围	持续时间
19610628	91.30081	306	0.3846154	6
19640722	92.06129	310.4	0.1538462	3
19740818	80.71923	309.2	0.2628205	3
19770910	93.90513	323.4	0.1858974	2
19830820	93.77143	314.7	0.3141026	3

续表8 等级划分

结束时间	平均降水量	降水强度	覆盖范围	持续时间
19840702	103.3762	362.3	0.2692308	1
19840730	113.8514	379	0.1730769	2
19880626	97.43889	315.3	0.1153846	1
19930730	115.4192	524.7	0.1602564	2
19950824	151.7304	374.3	0.1474359	1
19960728	96.76	410.8	0.1858974	2
19970816	83.8775	311.7	0.1858974	2
19980708	85.10271	356.6	0.3653846	4
20070708	81.55833	325.5	0.1217949	2
20130700	93.45574	415.9	0.3141026	3
20130720	81.75957	358.7	0.25	3

4.3. 强降水事件等级确定

文章已经把 1961-2013 年的历史强降水过程划分为了五个等级，建立了一个有效的聚类中心，对于单个的强降水过程可根据此聚类中心来划分。若有一个强降水过程，先由 3.2 的一系列公式计算出强降水评估指标，然后利用公式（2）计算出此次强降水过程与表 5 的最终聚类中心的欧式距离，然后判断此次强降水过程属于哪一级。具体步骤如下：

设一次强降水过程为 $X=\{x_1,x_2,x_3,x_4\}$，最终聚类中心为 Z，计算 x_i（i=1,2,3,4）与 Z_k（k=1,2,3,4,5）的欧式距离，公式为 $d(X,Z_k)=\sqrt{\sum_{i=1}^{d}(X_i-Z_{ki})^2}$。根据欧式距离的大小来划分 X 的等级。

5. 总结

本文选取了四个指标对强降水等级进行评判，意在建立一个具有普遍

性能应用到四川省的强降水等级评估，这与其他分级结果有一些出入，对于有几个强降水过程分级稍有不同，但大体一致，说明说明通过 K-means 聚类的方法对强降水过程分级有一定的意义。运用 K-means 聚类方法对强降水事件进行等级划分，能使聚类中心达到最优，方法简单可行，能快速的进行聚类，且聚类效果可行，综合考虑了强降水事件的四个指标因子对其进行聚类，等级划分合理，是强降水致灾风险评估的根据，因此能用到实际业务中去，为当地政府开展防灾工作、确定减灾、救灾方案、制定灾后救援计划等提供有力的科技支撑和科学决策依据，避免造成重大的损失。

基于可拓模型的四川暴雨灾害风险评估模型的建立是非常有意义的工作，对于气象业务有重要的促进作用。文中选取致灾因子、孕灾环境、承灾体、防灾减灾能力作为一级评估因子，并将单站暴雨过程总量、暴雨过程平均量、最大日暴雨量、地形、坡度、人口密度、地均 GDP、耕地比例、人均 GDP、公路里程作为二级评估因子，并通过区域暴雨特征分析提出的高原站和非高原站的暴雨灾害分级、通过对数正态分布函数获得重现期作为致灾因子等级评估标准等思路，对业务应用都是有实践意义的。但由于暴雨灾害影响的复杂性以及致灾因子资料收集的困难，风险评估模型建立的恰当还需进一步研究。

致谢

大学生创新创业训练计划项目（CX2015017）、中国气象局西南区域气象中心重大科研项目（西南区域014-5）、国家自然科学基金（91337215，41575066）、国家科技支撑计划（2015BAC03B05）、公益性行业（气象）科研专项（GYHY201406015）、国家重点基础研究发展计划（2013CB733206）。

参考文献

[1] 章国才.气象灾害风险评估与区划方法.气象出版社.2010.
[2] 王小玲,翟盘茂.1957-2004 年中国不同强度级别降水的变化趋势特征.热带气象学报,2008,24(5):459-466.
[3] 王千,王成,冯振远,叶金凤.K-means 聚类算法研究综述.电子设计工程.2012,20(7):21-24.
[4] 宋连春,肖风劲,叶殿秀.气象灾害影响及风险评估理论与实践.气象出版社.2012.
[5] 吴夙慧,成颖,郑彦宁,潘云涛.K-means 算法研究综述*.现代图书情报技术.2011(5):28-35.
[6] 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.
[7] 刘合香.模糊数学理论及其应用.科学出版社.2012,8.