

基于信息熵的异常检测算法

张安勤, 吴蕊, 张挺

引用本文:

张安勤, 吴蕊, 张挺. 基于信息熵的异常检测算法[J]. 上海电力大学学报, 2020, 36(4): 386-390.

ZHANG Anqin, WU Rui, ZHANG Ting. Anomaly Detection Algorithm Based on Information Entropy[J]. *Journal of Shanghai University of Electric Power*, 2020, 36(4): 386-390.

您可能感兴趣的其他文章 (请使用火狐或IE浏览器查看文章)

Articles you may be interested in (Please use Firefox or IE to view the article)

基于NWPSOBP神经网络的异常用电行为检测算法

Detection Algorithm of Abnormal Electrical Behavior Based on NWPSOBP Neural Network

上海电力大学学报. 2020, 36(4): 357-363 <https://doi.org/10.3969/j.issn.2096-8299.2020.04.007>

基于模糊C均值聚类算法的区域用电特征分析

Fuzzy C-means Clustering-based Algorithm for the Analysis of Regional Electric Power Characteristics

上海电力大学学报. 2017, 33(2): 196-200,209 <https://doi.org/10.3969/j.issn.1006-4729.2017.02.017>

基于改进Canny检测与Hough变换的仪表图像识别算法

The Instrument Image Recognition Algorithm Based on Canny Detection and Hough Transform

上海电力大学学报. 2020, 36(2): 183-189 <https://doi.org/10.3969/j.issn.2096-8299.2020.02.015>

基于独特码符号距离的TDMA突发检测算法

TDMA Burst Detection Method Based on Unique Code Symbol Distance

上海电力大学学报. 2020, 36(4): 391-394 <https://doi.org/10.3969/j.issn.2096-8299.2020.04.013>

基于量测量突变检测与拓扑约束协同的不良数据检测

Bad Data Detection Based on Measurement Sudden Change Detection and Topology Constraint

上海电力大学学报. 2018, 34(1): 59-65 <https://doi.org/10.3969/j.issn.1006-4729.2018.01.011>

DOI: 10.3969/j.issn.2096-8299.2020.04.012

基于信息熵的异常检测算法

张安勤, 吴蕊, 张挺

(上海电力大学 计算机科学与技术学院, 上海 200090)

摘要:针对 K-means 异常检测算法检测性能低的问题,提出了一种结合信息熵与改进 K-means 算法的异常检测算法。该算法均匀地选出密度大于数据集平均密度的数据对象作为初始聚类中心,避免了初始中心的随机选择。在此基础上,引入了信息熵确定属性权重的方法来计算簇中数据点与该簇聚类中心的加权欧氏距离,通过对比簇中数据点的加权欧氏距离与该簇中所有数据点的平均加权欧氏距离来进行异常检测。实验表明,改进算法具有更高的检测率和更低的误检率,应用于电力负荷数据时检测率达到了 90.5%,能够有效地检测出异常的负荷数据。

关键词:信息熵; K 均值; 异常检测

中图分类号: TP312

文献标志码: A

文章编号: 2096-8299(2020)04-0386-05

Anomaly Detection Algorithm Based on Information Entropy

ZHANG Anqin, WU Rui, ZHANG Ting

(School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China)

Abstract: To solve the problem of low detection performance of the K-means anomaly detection method, an anomaly detection algorithm combining information entropy and improved K-means is proposed. The algorithm uniformly chooses the data object whose density is greater than the average density of the data set as the initial clustering center, avoiding the random selection of the initial center. Besides, the weighted Euclidean distance between the data point and the cluster center in the cluster is calculated according to the attribute weight based on the information entropy. Anomaly detection is performed by comparing the weighted Euclidean distance of the data point with the average weighted Euclidean distance of all data points in the cluster. Experiments show that the improved algorithm has higher detection rate and lower false detection rate. When the algorithm is applied to power load data, the detection rate reaches 90.5%. The abnormal power load data can be effectively detected.

Key words: information entropy; K-means; anomaly detection

随着电力系统的信息化程度不断提高,电网的数据规模飞速增长^[1]。如何对海量的电网数据

进行高效分析,快速准确地检测出异常的电力数据,是电网安全有效运行的重要保证。

收稿日期: 2020-03-18

通信作者简介:张安勤(1974—),女,博士,副教授。主要研究方向为数据挖掘和普适计算。E-mail: aqz612@sina.com。

基金项目:国家自然科学基金(41672114)。

异常检测可以发现数据集中与一般数据有差异的数据对象,即一些与众不同的数据^[2]。在电力领域中,异常检测可以作为数据研究的基础,如检测出异常的电力负荷数据能有效提高负荷数据的质量,对后续进行负荷预测及合理规划电网有着重要的作用^[3-4]。异常检测也可以直接用于数据分析,如异常用电检测和设备故障检测等。

异常检测技术主要可以分为基于监督、半监督和无监督3种类型。基于监督的异常检测方法的训练集由带标签的正常和异常数据构成,主要有概率统计方法^[5]、神经网络方法^[6]等;基于半监督的异常检测方法^[7]能够从大量无标签的数据及部分有标签的数据中挖掘出异常数据的信息;基于无监督的异常检测方法可以直接从无标签的数据集中检测出异常数据,主要有基于K-means算法的异常检测^[8]、基于AP聚类的异常检测^[9]和基于局部离群因子的异常检测^[10]等。基于监督的异常检测方法需要提前得到训练样本的标签信息,一般可通过人工标记来获得足够的训练样本,成本及代价较高;而电力负荷数据集中大多数为无标签的数据,采用无监督的异常检测方法具有代价小、简单、高效等优点。

近年来,研究者们提出了很多针对异常检测的相关方法。文献[11]通过划定时间窗口来提取相关特征,并运用核密度估计算法建模,实现了在无标签的海量数据中识别异常数据。文献[12]基于最小生成树提出了一种新的距离度量方法,能够比欧氏距离更好地进行样本间相似性度量,并通过捕获数据点或簇之间的相对连通性来进行异常数据的检测。文献[13]提出了一种异常检测模型,包含了特征提取、主成分分析、网格处理、局部离群因子计算等方法,可用于检测电力用户异常的用电模式。

K-means算法作为一种无监督聚类算法,方法简单且易于实现,可以广泛应用于异常检测且检测效果良好^[14]。但是存在以下两个问题:一是算法随机选择初始中心,易使结果陷入局部最优且初始中心可能包含异常数据点,导致聚类效果差且影响最终的异常检测率;二是在计算样本间的相似性时采用各属性无差别的欧氏距离来进行度量,没有考虑数据的属性权重,使得计算的样本间相似性不准确^[15]。

针对上述问题,本文结合信息熵与改进的K-means算法,提出了一种用于电力负荷数据的异常检测算法。首先,通过信息熵确定属性权重,在计算欧氏距离时引入属性权重来度量样本间的相似性;其次,计算每个样本的密度及数据集的平均密度,在样本密度大于平均密度的数据对象中根据密度及距离依次选出 K 个初始中心;最后,在K-means算法的迭代过程中,对比簇中数据对象到簇中心的加权欧氏距离与该簇平均加权欧氏距离,以检测数据是否异常。实验表明,改进的异常检测算法在检测率和误检率方面结果更优。对电力负荷数据进行异常检测时,相较于对比算法,改进算法的检测率提高了10.3%,误检率降低了1.7%,适合应用于电力领域。

1 K-means算法及属性权重的计算

1.1 K-means算法

K-means算法的主要思想是基于欧氏距离来计算样本间相似性,并根据样本间的相似程度将其划入所属的簇。传统K-means算法流程如下。

输入:样本集 $X = \{x_1, x_2, x_3, \dots, x_n\}$,聚类中心数 K 。

输出:簇 $P = \{P_1, P_2, P_3, \dots, P_K\}$ 。

步骤1 从 X 中随机选择 K 个样本作为初始中心 $\{c_1, c_2, c_3, \dots, c_K\}$ 。

步骤2 计算 X 中的样本与各簇中心 $c_i (1 \leq i \leq K)$ 的距离,并将样本划入与其最近的聚类中心所属的簇。

步骤3 根据簇中包含的样本,重新形成每个簇的中心点。

步骤4 重复步骤2和步骤3直到聚类中心不变,算法结束。

1.2 属性权重的计算

K-means算法在计算时并未考虑数据的属性权重对聚类的影响,将会导致算法的聚类结果不准确。信息熵是一种计算属性权重的经典算法,可以用来计算属性的离散程度。对于某项指标,其熵值越小,则该指标的离散程度越大且具有更大的信息量。因此,对属性而言,熵值越小时该属性对聚类的影响越大。将信息熵应用于K-means算法时,

可根据各属性的离散程度计算其权重,并通过计算加权欧氏距离来提高聚类效果和异常检测率。

信息熵计算属性权重的算法流程如下。

步骤1 假设数据集 X 由 n 个 m 维样本构成,则 X 可以表示为一个 $n \times m$ 的矩阵,即

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

步骤2 为了消除量纲对聚类结果的影响,对数据进行归一化处理,公式为

$$x_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (1)$$

式中: x_{ij} ——第 i 个样本的第 j 列属性值;

$\max(x_j), \min(x_j)$ ——数据集中第 j 列属性的最大值和最小值。

步骤3 计算第 i 个样本的第 j 列属性所占的比重,公式为

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^n x_{ij}} \quad (2)$$

步骤4 计算第 j 列属性的信息熵,公式为

$$E_j = -\frac{1}{\ln n} \sum_{i=1}^n P_{ij} \ln P_{ij} \quad (3)$$

步骤5 计算第 j 列属性的差异系数,公式为

$$q_j = 1 - E_j \quad (4)$$

步骤6 计算第 j 列属性的权重,公式为

$$w_j = \frac{q_j}{\sum_{j=1}^m q_j} \quad (5)$$

其中, $0 \leq w_j \leq 1$ 。

计算得到各属性的权重后,给出数据对象 a 和 b 的加权欧氏距离^[16],公式为

$$\text{dist}(x_a, x_b) = \sqrt{\sum_{j=1}^m w_j (x_{aj} - x_{bj})^2} \quad (6)$$

式中: x_{aj}, x_{bj} ——数据对象 a 和 b 的第 j 列属性取值。

2 异常检测算法

2.1 K-means 算法初始中心的选择

需要进行异常检测的数据集一般包含正常和异常两种数据。由于传统 K-means 算法的初始聚类中心是随机选出的,算法有可能将异常点选作初

始中心,从而影响聚类结果。异常数据一般具有以下特征:与正常数据相比,异常数据所占比例较小,数据集中的大多数数据都为正常数据;异常数据的密度较小,分布在稀疏的区域,比如在进行异常用电检测或计算机监控时,数据集中的大多数数据为正常数据,而异常数据大多处于正常数据所聚集的簇的边缘或者簇与簇之间的区域。根据异常数据的特点,可以选择密度大于数据集平均密度的数据点作为初始中心,使得初始中心不含异常点。

K-means 算法的聚类结果和聚类效率与初始中心有直接的关联。算法的初始中心越接近实际簇中心则聚类效率越高,而随机的选择初始中心则导致了聚类结果不够可靠。

初始聚类中心的选择算法如下。

输入:样本集 $X = \{x_1, x_2, x_3, \dots, x_n\}$, 最近邻数据点个数 t , 聚类中心数 K 。

输出:初始中心 $C = \{c_1, c_2, c_3, \dots, c_K\}$ 。

步骤1 计算 X 中数据点 $x_j (1 \leq j \leq n)$ 的密度为

$$D(x_j) = \frac{1}{\sum_{x_i \in G_t(x_j)} \text{dist}(x_i, x_j)} \quad (7)$$

式中: $G_t(x_j)$ —— x_j 的 t 个最近邻数据点集合。

步骤2 计算所有数据对象的平均密度 D , 由 $D(x_j) > D$ 的数据点形成新的数据集 X' 。其平均密度的公式为

$$D = \frac{1}{n} \sum_{j=1}^n D(x_j) \quad (8)$$

步骤3 在 X' 中,取 $D(x_j)$ 最大的数据点作为第 1 个初始中心 c_1 ,取距离 c_1 最远的数据点作为第 2 个初始中心 c_2 。对于 X' 中的样本 $x_j (j=1, 2, 3, \dots, n)$,计算其与前 $i-1 (3 \leq i \leq K)$ 个初始中心的加权欧氏距离 $\text{dist}(x_j, c_1), \text{dist}(x_j, c_2), \dots, \text{dist}(x_j, c_{i-1})$,选出其中的最小值 $\min(\text{dist}(x_j, c_1), \text{dist}(x_j, c_2), \dots, \text{dist}(x_j, c_{i-1}))$ 。按照上述方式依次计算 X' 中的每个样本与前 $i-1$ 个初始中心加权欧氏距离的最小值,并放入集合中,集合中最大值对应的数据对象为第 i 个初始中心 c_i ,直到选出 K 个初始中心,算法结束。即满足

$$\max \{ \min(\text{dist}(x_j, c_1), \text{dist}(x_j, c_2), \dots, \text{dist}(x_j, c_{i-1})) \} \quad (9)$$

由上述算法过程可知,实际的簇中心一般处于密度较大的区域且相互间具有一定的距离,改进算法在密度大于平均密度的数据点中均匀地选择初始聚类中心,满足了密度较大和保持距离的

要求。另外,改进算法有效避免了将异常点作为初始中心的情况,使算法不会从开始就陷入误差,提高了聚类效率。

2.2 基于信息熵与改进 K-means 算法的异常检测算法

本文根据异常数据的特征和分布模式提出了一种异常检测算法。该算法在聚类迭代时通过对比数据点到所属簇中心的加权欧氏距离与该簇中所有数据点到簇中心的平均加权欧氏距离来检测异常。在异常检测过程中采用加权欧氏距离,考虑了属性权重对异常检测计算的影响,使异常检测结果更加准确。

基于信息熵与改进 K-means 算法的异常检测算法如下。

输入:样本集 $X = \{x_1, x_2, x_3, \dots, x_n\}$, 最近邻数据点个数 t , 聚类中心数 K , 异常度阈值 η (X 不同时, η 的取值也不相同)。

输出:簇 $P = \{P_1, P_2, P_3, \dots, P_K\}$, 异常点集合 U 。

步骤 1 设数据集中每个数据对象的初始异常度为 $F_j = 0$ ($j = 1, 2, 3, \dots, n$), 并根据信息熵计算属性权重的算法来计算数据集 X 中各属性权重。

步骤 2 根据初始聚类中心的选择算法选择 K 个初始中心 $\{c_1, c_2, c_3, \dots, c_K\}$ 。

步骤 3 计算数据对象 x_j 与各簇中心 c_i ($i = 1, 2, 3, \dots, K$) 的加权欧氏距离, 并将 x_j 划入距其最近的聚类中心所在的簇。

步骤 4 在 K 个簇中, 若 x_j ($j = 1, 2, 3, \dots, n$) 与所属簇中心的加权欧氏距离大于该簇中所有数据对象到簇中心的平均加权欧氏距离, 则 F_j++ 。上述判断公式为

$$\text{dist}(x_j, c_i) > \frac{1}{|P_i|} \sum_{x \in P_i} \text{dist}(x, c_i) \quad (10)$$

式中: $|P_i|$ ——簇 P_i 中的数据对象个数。

步骤 5 重新生成每个簇的中心点

$$c'_i = \frac{1}{|P_i|} \sum_{x \in P_i} x \quad (11)$$

步骤 6 若 $c'_i \neq c_i$, 则将 c_i 更新为 c'_i , 并返回执行步骤 3。

步骤 7 若数据对象 x_j 的异常度 $F_j \geq \eta$, 则判断 x_j 为异常点, 将 x_j 并入 U 中。

步骤 8 聚类算法结束, 输出 P 和 U 。

由于数据对象的异常度是通过计算每次迭代时该数据对象与所属簇中心的加权欧氏距离来确定的, 因此算法选择的初始中心越接近实际的簇

类中心, 异常检测的性能越好。当初始中心含异常点时会严重影响异常检测的准确率, 而改进算法选择的初始中心不含异常点, 有效避免了上述问题。最后, 在异常计算中引入属性权重, 考虑了各个属性对异常检测的作用。

3 实验分析

实验数据采用 UCI 机器学习数据库的 Parkinson Disease Spiral Drawings Using Digitized Graphics Tablet 数据集^[17-18] (以下简称“帕金森数据集”), 以及某电厂一个月的电力负荷数据。将检测率和误检率作为异常检测性能的评价指标。

检测率 (Detection Rate, DR) 和误检率 (False Alarm, FA) 的定义如下^[19]

$$\text{DR} = \frac{\text{TP}}{\text{FN} + \text{TP}} \quad (12)$$

$$\text{FA} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (13)$$

其中: TP——被正确检测为异常的异常样本数;

FN——被错误检测为正常的异常样本数;

TN——被正确检测为正常的正常样本数;

FP——被错误检测为异常的正常样本数。

帕金森数据集包含静态螺旋、动态螺旋和定点稳定性 3 种测试。算法在每种测试中随机选择 1 000 条健康样本与 100 条患者样本作为数据集, 分别对 3 种测试类型进行计算, 并将患者样本当做异常数据来验证异常检测的效果。实验对比了改进算法与传统 K-means 算法、MinMax K-means 算法的异常检测性能^[20-21], 结果如表 1 所示。其中, 对比算法的数据结果为两种算法分别计算 100 次后的平均值^[22]。

表 1 3 种算法在帕金森数据集集中的实验结果对比 单位: %

测试类型	K-means 算法		MinMax K-means 算法		改进算法	
	检测率	误检率	检测率	误检率	检测率	误检率
静态螺旋	68.6	16.7	69.4	16.5	72.0	13.4
动态螺旋	62.5	19.2	64.8	18.0	78.0	17.4
定点稳定性	72.6	20.0	75.3	19.1	82.0	5.6

对于电力负荷数据集, 通过在其中随机添加一定比例的异常数据来对比 3 种算法的异常检测性能, 结果如表 2 所示。

表2 3种算法在电力负荷数据集中的
实验结果对比

算法	检测率	误检率
K-means 算法	77.5	18.1
MinMax K-means 算法	80.2	17.3
改进算法	90.5	15.6

实验结果表明,与 K-means 算法和 MinMax K-means 算法相比,改进算法的检测率更高且误检率更低。究其原因如下:首先,改进算法选择了更优的初始中心,不同的初始中心会导致在每次迭代时生成不同的簇类中心,而在迭代过程中数据对象的异常度计算与其所属的簇中心相关;其次,初始中心含有异常点,会使算法从一开始就陷入偏差,影响数据对象异常度的计算,进而影响最终的检测率和误检率,而改进算法有效避免了这种情况;最后,改进算法引入了属性权重的概念,根据各属性影响程度的不同,其在计算时的占比也不同,使得异常计算的结果更加准确,从而提高了异常检测的性能。

4 结 语

本文根据异常数据的特征和分布情况,结合信息熵与改进的 K-means 算法,提出了一种适用于电力负荷数据的异常检测算法。算法在选择初始聚类中心时有效避免了异常点并使得初始聚类中心均匀分布,聚类的效果更优。在此基础上,算法引入了属性权重和加权欧氏距离来计算数据点的异常度。实验证明,相较于其他对比算法,本文提出的改进算法有着更优的异常检测性能,能够有效检测出异常的电力负荷数据,为实现高精度的负荷预测提供了重要保障。

参考文献:

- [1] 韩博闻. 基于 Apriori 关联算法的配电网运行大数据关联分析模型[J]. 上海电力学院学报, 2018, 34(2): 163-168.
- [2] HE Y F, PENG Y, WANG S J, et al. A structured sparse subspace learning algorithm for anomaly detection in UAV flight data[J]. IEEE Transactions on Instrumentation and Measurement, 2017, 67(1): 90-100.
- [3] 李鹏辉, 崔承刚, 杨宁, 等. 基于 ARIMA-LSTM 组合模型的楼宇短期负荷预测方法研究[J]. 上海电力学院学报, 2019, 35(6): 573-579.
- [4] 李婧, 田龙威, 王艳青. 基于 GA-RBF 神经网络的电力系统短期负荷预测[J]. 上海电力学院学报, 2019, 35(3): 205-210.
- [5] 卫薇, 龙玉江, 钟掖. 基于概率统计模型的电力 IT 监控对象特征异常检测[J]. 山东农业大学学报(自然科学版), 2019, 50(4): 612-618.
- [6] 段茵, 陈恺焯, 刘昕, 等. 基于 BP 神经网络的纸张缺陷检测与识别研究[J]. 西安理工大学学报, 2018, 34(2): 235-239.
- [7] 刘亚丽, 孟令愚, 丁云峰. 电网工控系统流量异常检测的应用与算法改进[J]. 计算机系统应用, 2018, 27(3): 173-178.
- [8] 蒋华, 季丰, 王慧娇, 等. 改进 Kmeans 算法的海洋数据异常检测[J]. 计算机工程与设计, 2018, 39(10): 140-144.
- [9] 付迎丁, 兰巨龙. 基于核自适应的近邻传播聚类算法[J]. 计算机应用研究, 2012, 29(5): 1644-1650.
- [10] 张若雪. 自动识别异常波动: 机器学习在金融市场的一个应用[J]. 上海金融, 2018(11): 26-30.
- [11] 李海斌, 李琦, 汤汝鸣, 等. 一种无监督的数据库用户行为异常检测方法[J]. 小型微型计算机系统, 2018, 39(11): 114-122.
- [12] IMTIAZ A, ALDO D, YU D. Unsupervised anomaly detection based on minimum spanning tree approximated distance measures and its application to hydropower turbines[J]. IEEE Transactions on Automation Science and Engineering, 2019, 16(2): 654-667.
- [13] 庄池杰, 张斌, 胡军, 等. 基于无监督学习的电力用户异常用电模式检测[J]. 中国电机工程学报, 2016, 36(2): 379-387.
- [14] 贾凡, 严妍, 张家琪. 基于 K-means 聚类特征消减的网络异常检测[J]. 清华大学学报(自然科学版), 2018, 58(2): 137-142.
- [15] 樊蓉, 李娜. 基于特征选择的 K-means 聚类异常检测方法[J]. 网络安全技术与应用, 2018(4): 25-26.
- [16] 张丹丹, 游子毅, 郑建, 等. 基于改进的局部异常因子检测的优化聚类算法[J]. 微电子学与计算机, 2019, 36(11): 43-48.
- [17] ISENKUL M E, SAKAR B E, KURSUN O, et al. Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease[C]//The 2nd International Conference on e-Health and Telemedicine. Istanbul, Turkey, 2014: 171-175.
- [18] SAKAR B E, ISENKUL M E, SAKAR C O, et al. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings[J]. IEEE Journal of Biomedical and Health Informatics, 2013, 17(4): 828-834.
- [19] 邱雪松, 张珣, 宋彦斌, 等. 基于熵和线性关系的两级流量异常检测方法[J]. 北京邮电大学学报, 2018, 41(4): 56-62.
- [20] 左进, 陈泽茂. 基于改进 K 均值聚类的异常检测算法[J]. 计算机科学, 2016(8): 258-261.
- [21] 蒋华, 武尧, 王鑫, 等. 改进 K 均值聚类的海洋数据异常检测算法研究[J]. 计算机科学, 2019(7): 211-216.
- [22] 付卫红, 李丹. 基于滤波器阶数估计的卷积盲分离算法[J]. 华中科技大学学报(自然科学版), 2018, 46(6): 116-121.

(责任编辑 白林雪)