文章编号:1006-9348(2021)11-0290-05

基于机器学习的不确定数据增量式挖掘算法

文明瑶,廖伟国
(华南农业大学珠江学院,广东广州 510900)

摘要:以实现数据增量式精准挖掘为目的,提出基于机器学习的不确定数据增量式挖掘算法。以机器学习算法中的模糊c-均值聚类(FCM)算法为基础,通过主成分分析法筛选原始数据集中指标,利用Relief算法计算指标权重,实现FCM算法改进。改进FCM算法通过阈值定义目标函数,经样本数据分类、特征提取和聚类,使目标函数达到最小值,实现数据挖掘。实验结果表明,上述算法的数据样本分类符合率可达99.28%,分类准确率在98%左右,且分类耗时短、效率高;特征提取能力受数据量增加影响较小;在数据增量情况下,改进算法增量式挖掘准确率保持在95%~98%之间,且所需迭代次数少。

关键词:机器学习;增量式;数据挖掘;聚类
中图分类号:TP183 **文献标识码**:B

Incremental Mining Algorithm for Uncertain Data Based on Machine Learning

WEN Ming-yao, LIAO Wei-guo
(Zhujiang College, South China Agricultural University, Guangdong Guangzhou 510900, China)

ABSTRACT: In order to realize incremental accurate data mining, an incremental uncertain data mining algorithm based on machine learning is proposed. Based on the fuzzy c-means clustering (FCM) algorithm in machine learning algorithm, the indexes in the original data set were selected by principal component analysis, and the index weight was calculated by relief algorithm to improve the FCM algorithm. Then, the improved FCM algorithm defined the objective function through threshold, and made the objective function reach the minimum value through sample data classification, feature extraction and clustering to realize data mining. The experimental results show that the data sample coincidence rate of the algorithm can reach 99.28%, the classification accuracy is about 98%, and the classification time-consuming us short and the efficiency is high; The ability of feature extraction is less affected by the increase of data volume; In the case of data increment, the accuracy incremental mining of the improved algorithm remains between 95% ~ 98%, and the number of iterations is less.

KEYWORDS: Machine learning; Incremental; Data mining; Clustering

1 引言

广义上讲,使用“机器”模拟人类学习活动的学科,称为机器学习[1,2],在这里“机器”代表电子计算机,通过机器学习可获取新知识、新信息以及新技能,并可识别现有知识信息[3]。机器学习主要组成为:对人类学习过程进行解读并模拟;研究人类与计算机之间的语言接口;自动规划问题;以实现机器学习为目的,设计发现新信息的程序。机器学习涉及统计学、概率论、算法复杂度理论等多个学科[4],广泛应用于建筑、医学、数学、金融等多个领域。机器学习可从一类数据中自动分析其规律,并利用该规律对未知数据展开预先计算,因此,机器学习在处理庞大数据量时,效果显著[5]。

经过分析每个数据,从大量数据中找出其关联性,称为数据挖掘,现代电子技术迅猛发展,各种行业数据信息量成倍增长,并发展迅速[6,7],数据具有构成庞大、组成复杂、变化迅速等特征。为了实现数据的有效挖掘,杨阳[8]等人提出一种基于Spark的不确定数据集频繁模式挖掘算法,该算法在计算过程中需进行频繁式计算,计算精度高,但其计算耗时长,期望概率和权重需选择单一指标,计算过程复杂。许磊[9]等人提出基于模糊神经网络的异常网络数据挖掘算法,该算法依据异常网络数据进行相似度计算,其收敛性较差,

基金项目:2017年广东省高等教育教学研究和改革项目(706)
收稿日期:2020-12-03 修回日期:2021-02-08

造成计算结果不够精确。

从大量并且复杂多变的数据中，挖掘有效信息，是现代机器学习的重要方向。机器学习的主要技术之一是模糊 c－均值聚类算法，简称 FCM 算法，该算法应用范围广，计算精度高，可将目标函数进行优化，并从中获取聚类中心的隶属度，根据数据样本的隶属度达到将数据分类的目的。因此，为了解决现有方法存在的问题，本文提出基于机器学习的不确定数据增量式挖掘算法，使用改进 FCM 算法实现海量数据增量式挖掘，可提高计算结果的精确度与运算时间[10,11]。

2 基于改进 FCM 的不确定数据增量式挖掘

2.1 FCM 算法思想

由 $x_i(i=1,2,\cdots,n)$ 表示 n 个向量，FCM 算法将其分成 F 个模糊簇，通过计算得出每个簇的聚类中心，使目标函数达到最小，$l<w<+\infty$，其目标函数定义由式(1)表示

$$J_w(u,v)=\sum_{k=1}^{n}\sum_{i=1}^{f}(u_{ik})^{w}d(x_k,v_i) \tag{1}$$

其中，w 代表模糊权重指数，$\sum_{i=1}^{f}u_{ik}=1,\forall k,d(x_k,v_i)=\|x_k-v_i\|^2$，$u_{ik}\in(0,1)$。

更新聚类中心与隶属度可使目标函数最小，分别由式(2)、(3)表示

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^{w}x_k}{\sum_{k=1}^{n}u_{ik}^{w}} \tag{2}$$

其中，$i=1,2,\cdots,f$。

$$u_{ik}=\frac{1}{\sum_{j=1}^{f}\left(\frac{d_{ik}}{d_{jk}}\right)^{1/(w-1)}} \tag{3}$$

其中，$i=1,2,\cdots,f,j=1,2,\cdots n$。

FCM 计算流程由图 1 表示。

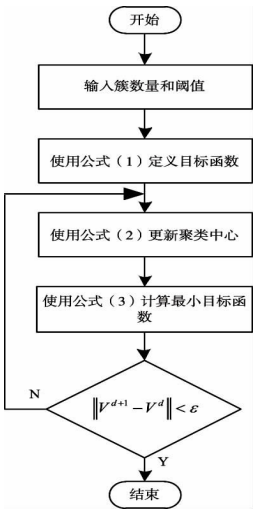

图 1 FCM 计算流程

1) 由 $X=(x_1,x_1,\cdots x_n)$ 表示 n 个数据集，从该数据集中选择初始聚类中心为 $V^0=(v_1,v_2,\cdots v_f,)$，阈值 $\varepsilon>0$。

2) 通过式(3)计算隶属度矩阵 U^d；基于当前的 U^d 通过式(2)计算聚类中心 V^{d+1}。按照以上步骤反复运算，当 $\|V^{d+1}-V^d\|<\varepsilon$ 时，运算终止，其中，d 表示当前迭代次数。

2.2 FCM 算法改进

为提升 FCM 算法的应用效果，在 FCM 算法中增加指标筛选和指标权重计算流程，对 FCM 算法进行改进。

假设每个样本数据集有 t 个特征指标，将指标筛选流程添加到 FCM 算法中，将 z 个指标通过主成分分析法从数据集中挑选出来，利用挑选出的指标实施样本数据集聚类。

假设 $W=(w_1,w_2,\cdots,w_z)$ 是 z 个指标的权重，符合 $\sum_{k=1}^{z}w_k=1$，利用 *Relief* 算法计算其权重 w，修正原型目标函数，可由式(4)表示

$$J'_m(U,V)=\sum_{i=1}^{f}\sum_{j=1}^{n}u_{ij}^2\times\sum_{k=1}^{s}[w_k\times(x_{jk}-v_{ik})]^2 \tag{4}$$

将 *FCM* 算法改进后，首先确定样本分类数 $f(2\leqslant f\leqslant n-1)$，设 q 和 ξ 分别代表参数和误差值，其初始矩阵由 $R^{(0)}$ 表示；然后依据初始矩阵 $R^{(0)}$，计算聚类中心 $V_i(i=1,2,\cdots,f)$，其由式(5)表示

$$V^{(l)}=\sum_{j=1}^{n}(u_{ij}^{(l)})^2\times x_j \tag{5}$$

依据式(5)计算新分类矩阵 R^*，更改划分矩阵，由式(6)表示

$$r_{ij}^{(l+1)}=\left[\sum_{h=1}^{f}\frac{\|w\times(X_j-V_i^{(l)})\|^2}{\|w\times(X_j-V_h^{(l)})\|^2}\right]^{-1} \tag{6}$$

式(6)中，$i=1,2,\cdots f;j=1,2,\cdots n$。

将 $r^{(l)}$ 与 $r^{(l+1)}$ 进行对比。将误差值设为 ξ，如果 $\xi>\|R^{(l+1)}-R^{(l)}\|$，则终止迭代，反之，返回式(5)重新进行计算。

2.3 基于改进 FCM 的增量式聚类算法

如果原始数据被聚类为 A、B、C、D、E、F 类，新增加的数据可能属于其中某一类，也可自行称为新类，还有可能与其中两类聚合形成新类[12]。其聚合示意图由图 2 表示。

由图 2 可知，新增数据使原始数据聚类增加了 2 组，其中，H 类属于自成一类，G 类将 A、B、C 类部分数据聚合形成新类。FCM 增量式聚类具体算法如下：

首先，由随机选取一个样本组成初始数据集，使用 FCM 对该数据集进行聚类，由函数 $Sub(U;f)=\max_{i=1}^{f}\max_{h=1,h\neq1}^{f}F(A_1,A_h)$ 最小的 f 确定聚类数目，可由式(7)计算：

$$F(A_1,A_h)=\frac{\sum_{i=1}^{n}(u_{i1}^m d_{i1}^2\wedge u_{ih}^m d_{ih}^2)}{\sum_{i=1}^{n}(u_{i1}^m d_{i1}^2)} \tag{7}$$

FCM 聚类最终结果是对数据集进行模糊划分：$A=\{A_1,A_2,\cdots,A_f\}$，其中，样本属于第 1 类的隶属函数由 A_1 代表，是

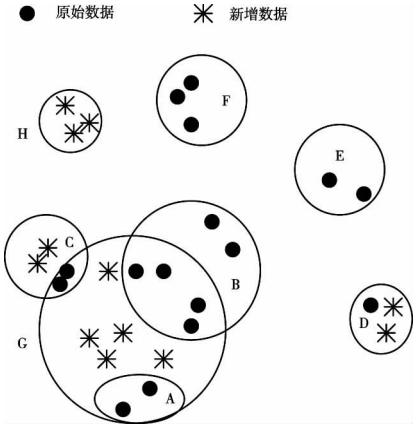

图 2　增量式聚类示意图

隶属矩阵 U 的第 1 行，分类效果最好的情况是 A_1 和 A_h 尽量分离，即 A_1 尽量不包含 $A_h(h\neq 1)$。通常而言，聚类的数量很小，若聚类数量很大，那么其精度不够准确，那么，可通过式(6)确定最优样本数据集分离数量 f。聚类数目可通过以上步骤自动获得，可规避人为规定分类数目的不准确性，完成初始聚类。

其次，基于初始聚类结果，将剩余样本数据实施分配和聚类。

1) 设 Q_i 与 Q_j 两个类之间的平均距离由式(8)表示

$$Dist(Q_i,Q_j)=\frac{1}{n_i n_j}\sum_{P\in Q_i}^{n}\sum_{P'\in Q_{ji}}^{n}|P-P'| \tag{8}$$

其中，Q_i 和 Q_j 分别代表计算出的第 i 类和第 j 类。

2) 设任意类到任意数据点的平均距离由式(9)表示:

$$d_{ik}=\frac{1}{n_i}\sum_{X_j}^{n}|X_j-X_k| \tag{9}$$

其中，n_i 代表第 i 类中包含的样本数量; n_j 代表第 j 类中包含的样本数量; X_k 代表剩余数据集中的数据。其阈值由式(10)表示

$$MaxDist=\underset{1\leqslant i,j\leqslant f,i\neq j}{Max}Dist(Q_i,Q_j) \tag{10}$$

分配规则如下:

若 $MaxDist\geqslant d_{ik}$，需将 X_k 划分到第 i 类中; 若 $MaxDist\geqslant d_{ik}$ 并且 $MaxDist\geqslant d_{jk}$，需合并第 i 类和第 j 类; 若 $MaxDist<d_{ik}$，X_k 将单独划分为一类。

3) 重复以上步骤，直至数据集无剩余数据为止。

通过上述流程，改进 FCM 算法可合并两类或多类数据，对非球形与椭球形分布的数据集进行较好聚类，从而形成新的类，可避免传统 FCM 算法的局限性。

3　实验分析

为验证所提算法的实际应用效果，本次实验采用内存为 4G、主频为 2.6G、windows 10 操作系统的台式电脑进行实验。以 5000 组数据作为实验样本，将该样本分为 A、B、C 三组，分别作为原始数据组，其中 A 组数据为 2500 组，B 组数据为 1000 组，C 组数据为 1500 组，阈值取 0.6。以基于 Spark 的不确定数据集频繁模式挖掘算法(文献[8]算法)和基于模糊神经网络的异常网络数据挖掘算法(文献[9]算法)作为对比方法，进行实验验证。

3.1　数据分类

良好的数据分类是实现数据增量式挖掘的基础，为此，分别使用所提算法与文献[8]算法、文献[9]算法对 A、B、C 三组数据样本实施增量式挖掘，对比三种算法分类性能，结果如表 1 所示。

表 1　三种算法分类对比

算法	数据组别	分类数(个)	符合率(%)
所提算法	A	96	98.76
	B	57	99.01
	C	63	99.28
文献[8]算法	A	81	85.43
	B	39	89.09
	C	44	89.27
文献[9]算法	A	76	85.06
	B	52	85.37
	C	43	84.18

分析表 1 可知，所提算法对 A、B、C 三组数据样本分类个数均高于文献[8]和文献[9]算法，说明所提算法分类详细。并且所提算法的分类符合率最高达到了 99.28%，远高于文献[8]算法、文献[9]算法，实验结果表明，所提算法数据分类能力更强。

统计上述实验过程中三种算法对 A 组数据的平均分类准确率，其结果如图 3 所示。

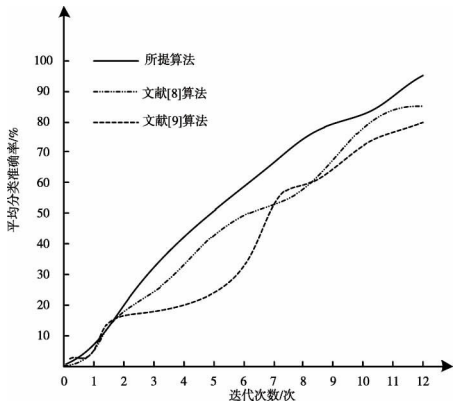

图 3　三种算法平均分类准确率

分析图 3 可知，所提算法的平均分类准确率要明显高于文献[8]、文献[9]算法，其中所提算法与文献[8]算法的准确率曲线较平缓，可见其分类较稳定，而文献[9]算法的曲线波动较大，分类稳定性较差，同时也影响其分类准确率。当迭代次数为 12 次时，所提算法的分类准确率达到了 98% 左

右，可见所提算法分类准确率更高。

三种算法挖掘 A 样本数据的耗时对比结果如图 4 所示。

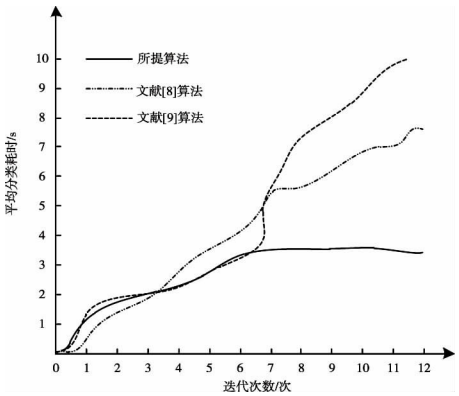

图 4　三种算法耗时对比

分析图 4 可知，文献［8］算法的最高耗时为 7.5s 左右；文献［9］算法分类耗时在迭代次数为 7 之前，趋势较平缓，迭代次数超过 7 时，呈迅速上升状态，并且耗时较长，最高已达到 10s；所提算法分类耗时在迭代次数为 7 之前，上升趋势比较明显，迭代次数超过 7 之后，其耗时曲线趋近直线，可见所提算法数据挖掘能力强、效率高。

3.2　特征提取

数据增量式挖掘的前提是数据特征提取，为此对比三种算法对不同数据样本的数据特征提取性能，其结果如图 5 所示。

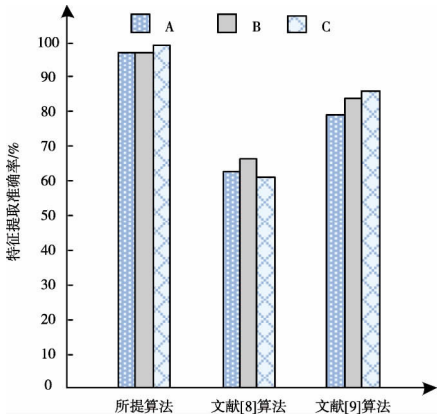

图 5　三种算法特征提取准确率

分析图 5 可知，文献［8］算法与文献［9］算法在对三组数据进行特征提取时，数据量不同提取准确度存在一定差异，可见传统算法在特征提取上有所欠缺，造成提取结果准确率不高。而所提算法在提取数据特征时，未受数据量影响，其特征提取准确率大致相同，因此，所提算法特征提取能力较强，不易受数据量影响。

3.3　增量式挖掘能力

为验证所提算法的数据增量式挖掘能力，以数据 A 为例，在数据样本 A 中，添加 9500 组新数据，分别测试三种算法数据增量式挖掘准确率，结果由图 6 表示。

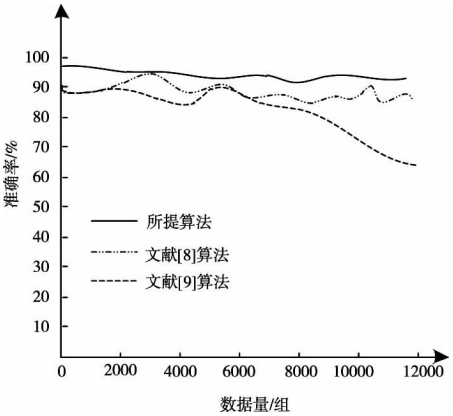

图 6　三种算法增量式挖掘准确率

分析图 6 可知，在数据增量情况下，三种算法的数据挖掘准确率整体上都是随着数据量的增加而降低，文献［9］算法从数据量增加至 6000 组后，准确率下降幅度增大；文献［8］算法与所提算法数据挖掘准确率较接近，但所提算法数据挖掘准确率高于文献［8］算法，并且所提算法随着数据量的增加，数据挖掘准确率曲线变化平缓，文献［8］算法挖掘准确率曲线呈波浪状，可见其数据挖掘准确率变化不稳定，且准确率低，所提算法准确率始终保持在 95% ~98% 之间，因此所提算法可在信息增量的情况下，对数据进行有效挖掘。

对比三种算法，在数据增量情况下的迭代次数，其结果由图 7 表示。

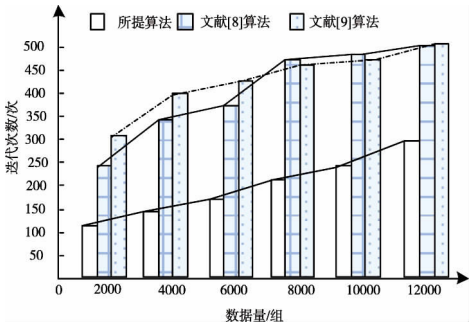

图 7　三种算法迭代次数对比

通过对比图 7 可知，迭代次数与数据量成正比，当数据量增加至 12000 组时，文献［8］算法与文献［9］算法迭代次数相差不大，在数据量为 6000 组时文献［8］算法迭代次数低于文献［9］算法，当数据量继续增加时，文献［8］算法迭代次数反而高于文献［9］算法，由此可知，文献［8］算法与文献［9］算法随着数据量的增加，计算结果不稳定；所提算法随着数

据量的增多,迭代次数呈平缓趋势增长,增长幅度较低,因此所提算法迭代次数少,数据增量式挖掘能力强。

4 结论

研究基于机器学习的不确定数据增量式挖掘算法,利用改进 FCM 算法通过设定数据目标函数,进行样本分类与特征提取,依据分类和特征提取结果实施数据聚类,实现数据增量式挖掘。实验结果表明:

1) 所提算法分类详细,符合率高,且符合率与样本数量成反比。

2) 迭代次数相同时,所提算法的平均分类准确率高。

3) 数据分类耗时短,分类能力强。

4) 所提算法受数据量增加影响较小,特征提取效果好,数据挖掘能力强。

5) 数据增量情况下,所提算法准确率降低幅度小,计算准确率高,其准确率在 95% ~98% 之间。

6) 数据增量情况下,所提算法迭代次数少,数据增量式挖掘能力强。

因此,所提算法具有极高实用性,可有效实现不确定数据增量式挖掘。目前世界中不确定性数据种类繁多,聚类关联性问题涉及面广泛,本文仅针对不确定数据增量式进行挖掘,无法满足多领域实际要求,因此需在分类效果评价、多类型数据、多领域重合等方面展开研究,从而提升聚类结果有效性和实用性。

参考文献:

[1] 崔治国,曹勇,武根峰,等. 基于机器学习算法的建筑能耗监测数据预处理技术研究[J]. 建筑科学,2018,34(2):94 -99.

[2] 任永功,高鹏,张志鹏. 一种利用相关性度量的不确定数据频繁模式挖掘[J]. 小型微型计算机系统, 2019,40(3):161 -165.

[3] 丁哲,秦臻,秦志光. 基于差分隐私的不确定数据频繁项集挖掘算法[J]. 计算机应用研究,2018,35(7):1942 -1946.

[4] 何保荣. 基于多目标决策的时间序列数据挖掘算法仿真[J]. 计算机仿真, 2019,36(11):243 -246.

[5] 叶福兰. 基于离群点检测的不确定数据流聚类算法研究[J]. 中国电子科学研究院学报, 2019,14(10):1094 -1099.

[6] 王菊,刘付显,靳春杰,等. 一种面向不确定数据流的模体发现算法[J]. 电子科技大学学报, 2017,46(1):81 -87.

[7] 李飞江,钱宇华,王婕婷,等. 基于样本稳定性的聚类方法[J]. 中国科学(信息科学) ,2020,50(8):1239 -1254.

[8] 杨阳,丁家满,李海滨,等. 一种基于 Spark 的不确定数据集频繁模式挖掘算法[J]. 信息与控制, 2019,48(3):257 -264.

[9] 许磊,王建新. 基于模糊神经网络的异常网络数据挖掘算法[J]. 计算机科学, 2019,46(4):73 -76.

[10] 陈力,费洪晓,丁海伦,等. 基于双决策树的数据采样方法[J]. 计算机工程与科学, 2019,41(1):134 -139.

[11] 叶炬锋. 基于维度根距离相似度量方法对单值和区间中性的聚类算法进行聚类算法[J]. 机床与液压, 2018,46(6):199 -208.

[12] 程舒通,徐从富,但红卫. 增量式隐私保护数据挖掘研究[J]. 计算机应用研究, 2018,35(7):2156 -2159,2171.

[作者简介]

文明瑶(1983 -),女(苗族),湖南湘西州人,硕士,讲师,研究方向:软件工程、数据挖掘。

廖伟国(1981 -),男(汉族),湖北天门市人,硕士,讲师,研究方向:计算机技术。

(上接第 224 页)

[12] I A J Radcliffe, M Taylor. Investigation into the effect of varus-valgus orientation on load transfer in the resurfaced femoral head: A multi - femur finite element analysis[J]. Clinical Biomechanics, 2007,22(7):780 -786.

[13] Hunt Kenneth J, Pereira Helder, Kelley Judas , et al. The Role of Calcaneofibular Ligament Injury in Ankle Instability: Implications for Surgical Management[J]. The American journal of sports medicine, 2019,47(2): 431 -437.

[14] Siegler S, Block J, Schneck C D. The mechanical characteristics of the collateral ligaments of the human ankle joint[J]. Foot & Ankle, 1988,8(5):234.

[15] Attarian D E, Mccrackin H J, Devito D P, et al. Biomechanical characteristics of human ankle ligaments[J]. Foot Ankle, 1985, 6(2): 54 -58.

[作者简介]

孟春玲(1966 -),女(汉族),山东人,教授,硕士研究生导师,主要研究领域为:有限元仿真分析;工程力学。

王 静(1993 -),女(汉族),山东省济宁市人,硕士研究生,主要研究领域为:有限元结构仿真。

高春雨(1985 -),男(汉族),北京人,博士研究生,主要研究领域为:脊柱方向。

叶宜颖(1982 -),女(汉族),中国台北人,博士研究生,主要研究领域为:脊柱及其相关疾病的研究。