交通流量的局部区间模糊 C 均值聚类算法*

李泽军,曾利军

(湖南工学院 计算机科学系,湖南 衡阳 421002)

摘 要:提出了一种局部区间模糊 C 均值算法,该算法不仅能减少计算的复杂性、提高统计速度,同时能提高函数估算值的精确度。利用紧致性函数和分离性函数对局部模糊 C 均值聚类算法进行有效性验证,实验结果表明该算法具有较高的预测精度。

关键词:SVM;模糊 C 均值算法;预测;分离信息

中图分类号:TP391 **文献标识码**:A

Fuzzy C-means clustering algorithm for the local area of traffic flow

LI Ze Jun, ZENG Li Jun

(Hunan Institute of Technology, Hengyang, 421002, China)

Abstract: This paper proposes a fuzzy c-means (FCM) algorithm for the local area, which can reduce the computation complexity and increase the statistical speed. The accuracy of estimation is also improved. Verify the partial fuzzy c-means(FCM) algorithm on the basis of the compactness and separation of function. This method is proved to be more effective and precise.

Key words: SVM; fuzzy c-means algorithm; prediction; separation information

非线性函数估计的支持向量机算法 SVM[1](Support Vector Machine)是神经网络领域中的一种重要的方法,主要解决函数估值和时间序列预测的问题。近年来许多学者对此作了许多深入的研究,不仅形成了一套较为完备的理论与方法,而且其应用也是日趋广泛。但由于该方法复杂度高、精确度低,为此,本文提出了一种对交通流量的确定方法,即局部区间模糊 C 均值聚类算法(LAFCM),LAFCM 算法是一种可以被广泛使用的、基于目标函数优化的、无监督模糊聚类方法,它无需训练样本,通过迭代执行分类算法来提取各类的特征值。该算法对某高速公路段的交通流量进行聚类分析,利用 Matlab 进行仿真来判定该道路的使用效率及阻塞情况。实验结果表明模糊均真算法 FCM(Fuzzy C-Means)比 SVM 更具有较高的预测精度,该算法是可行的、有效的。

1 C 均值模糊聚类算法

聚类分析[2]是数据挖掘的重要工具之一,将数据对象划分为多个类或者簇,根据相似程度自动分类。FCM 的基本思想为[3-4]:设数据集 $X=\{x_i, i=1,2,\cdots,n\}$,其中每个元素包含多个属性的样本集合,$m_i(i=1,2,\cdots,c)$为每个聚类的中心,$c$ 为预定的分类数目,$u_i(x_i)$是第 i 个样本对于第 j 类样本的隶属度。则隶属度函数定义的聚类阻塞目标函数为:

$$J_f=\sum_{j=1}^{c}\sum_{i=1}^{n}[u_j(x_i)]^b \parallel x_i-m_i \parallel^2 \tag{1}$$

其中,$\parallel x_i-m_i \parallel$ 为元素与聚类中心的距离,b 是一个可以控制聚类结果的加权模糊程度参数$(b>1)$。

$$\sum_{j=1}^{c}u_j(x_i)=1 \quad i=1,2,\cdots,n \tag{2}$$

式(2)要求每一样本对于每一个聚类的隶属度和均为 1。在该式条件下求目标函数,即式(1)的极小值,令 J_f 对 m_i 和 $u_i(x_i)$的偏导数为 0,可得 m_i 和 $u_i(x_i)$必要条件:

$$m_j=\frac{\sum_{i=1}^{n}[u_j(x_i)]^b x_i}{\sum_{i=1}^{n}[u_j(x_i)]^b} \quad j=1,2,\cdots,c \tag{3}$$

$$u_j(x_i)=\frac{(1/\parallel x_i-m_j \parallel^2)^{1/(b-1)}}{\sum_{k=1}^{c}(1/\parallel x_i-m_k \parallel^2)^{1/(b-1)}} \quad i=1,2,\cdots,n, j=1,2,\cdots,c \tag{4}$$

* 基金项目:湖南教育厅科学研究项目(08C248);湖南工学院青年项目(HGQ0604)

利用迭代方法求解式(3)、(4),其算法步骤如下:

(1)取定初始聚类数目 c 和控制模糊程度的加权指数 $b(b\in[2,\infty])$,初始化各个聚类中心 m_i;

(2)重复执行以下两步运算,直到各个样本 i 的隶属度稳定:

①根据目前的聚类中心,用式(4)计算 $u_i(x_i)$;

②根据当前隶属度函数用式(3)更新各聚类中心。

当算法迭代收敛值很小时,即得到各类聚类中心和各个样本点对于各类的隶属度值,从而完成了模糊聚类划分。

2 局部区间模糊 C 均值聚类算法数学模型

假定数据集样本元素区间为 $I_k=[x_{k1},x_{k2},\cdots,x_{kn}]$,区间中值为 $\overline{x_k}$;区间宽度为 $\hat{x}_k$;第 i 个聚类区间 p_i 的中值为 m_i,区间宽度为 l_i。则样本集 I_k 到聚类原型 p_i 的距离可以推导为:

$$d_{ik}=(\|\overline{x_k}-m_i\|^2+\alpha\left|\hat{x}_k-l_i\right|^2)^{1/2} \tag{5}$$

在式(5)所定义的区间值距离中,常数 a 为 l_i 的影响因子。当常数 a 取大于 1 时,则区间的宽度对聚类的结果影响较大;当常数 a 取小于 1,则区间中值的位置对聚类结果的影响大;一般情况下,取 $a=1$。令:

$$j_m(u,p)=\sum_{k=1}^{m}\sum_{i=1}^{c}(u_{ik})^m(d_{ik})^2=\sum_{k=1}^{m}\sum_{i=1}^{c}(u_{ik})^m(\|\overline{x_k}-m_i\|^2+\alpha\left|\hat{x}_k-l_i\right|^2) \tag{6}$$

$$(1\leqslant m\leqslant\infty)$$

用拉格朗日乘数法将式(6)求极小值,可以得到:

$$u_{ik}=1/\sum_{l=1}^{c}(\frac{d_{ik}}{d_{lk}})^{1/(m-1)}=1/\sum_{l=1}^{c}\left[\frac{\sqrt{(\|\overline{x_k}-m_i\|^2+\alpha\left|\hat{x}_k-l_i\right|^2)}}{\sqrt{(\|\overline{x_k}-m_l\|^2+\alpha\left|\hat{x}_k-l_l\right|^2)}}\right]^{2/(m-1)} \tag{7}$$

同理可以分别得到:

$$m_i=\frac{1}{\sum_{k=1}^{n}(u_{ik})^m}\sum_{k=1}^{n}(u_{ik})^m\overline{x_k} \tag{8}$$

$$l_i=\frac{1}{\sum_{k=1}^{n}(u_{ik})^m}\sum_{k=1}^{n}(u_{ik})^m l_k \tag{9}$$

从式(7)、(8)、(9),设定数据集为 $x_k=[x_{k1},x_{k2},\cdots,x_{kn}]$,初始化类别数 C 和权重 M 的值,根据迭代终止条件就可以确定模糊分类矩阵 $[u_k]$,区间宽度 l_i 和聚类区间中心 m_i。假定初始化聚类类别数 C,其中 n 为样本数据的个数。设定迭代终止的条件为一极小 ε,用迭代计数器 B 来保存迭代的次数,则迭代的终止条件为:

$$\|m_i^b-m_i^{(b+1)}\|<\varepsilon \tag{10}$$

$$u_{ik}^b=1/\sum_{j=1}^{c}\left(\frac{d_{ik}^b}{d_{jk}^b}\right)^{2/(m-1)} \tag{11}$$

由多次迭代计算可求得:

$$m_i^{(b+1)}=\frac{\sum_{k=1}^{n}(u_{ik}^{(b+1)})^m\times\overline{x_k}}{\sum_{k=1}^{n}(u_{ik}^{(b+1)})^m}\quad i=1,2,\cdots,c \tag{12}$$

$$l_i^{(b+1)}=\frac{\sum_{k=1}^{n}(u_{ik}^{(b+1)})^m\times l_k}{\sum_{k=1}^{n}(u_{ik}^{(b+1)})^m}\quad i=1,2,\cdots,c \tag{13}$$

输出聚类区间宽度 l_i 和区间的中值及模糊分类矩阵[5-6] $[u_{ik}]$。

3 局部区间模糊 C 均值算法有效性分析

定义紧致性度量函数为:

$$set(c)=\sum_{i=1}^{c}\frac{\sum_{j=1}^{n}u_{ij}^m\|x_j-m_i\|}{\sum_{j=1}^{n}u_{ij}} \tag{14}$$

其中,$\sum_{j=1}^{n}u_{ij}^m\|x_j-m_i\|$ 为基于欧氏距离[7-8]的类内平方误差和,随着 c 的增加有单调减小的趋势;$\sum_{j=1}^{n}u_{ij}$ 随 c 的增加而减小,$1/\sum_{j=1}^{n}u_{ij}$ 作为每个聚类的一个权值,限制紧致性度量的单调递减。紧致性表示类内的内聚程度,$set(c)$ 值越小,表明模糊划分越好。

$S(A,B)$ 表示 2 个聚类模糊集的相似度,$sep(c)=1-s(A,B)$ 表示不同类间的隔离程度,即分离度。$sep(c)$ 值越大,表明模糊划分越好。

由于紧致性度量和分离性度量有不同标量,因此需要进行归一化处理,处理过程如下:

$$set^n(c)=\frac{set(c)}{\max\{set(2),set(3),\cdots set(n)\}} \tag{15}$$

$$sep^n(c)=\frac{sep(c)}{\max\{sep(2),sep(3),\cdots sep(n)\}} \tag{16}$$

其中 $c=2,3,\cdots n$。

因此定义有效性函数为:

$$V_n(c)=\frac{set^n(c)}{sep^n(c)} \tag{17}$$

分类要求最小的紧致性和最大的分离性,$V^n(c)$ 定义为紧致性和分离性的度量,越小的 $V^n(c)$ 值表明划分越好。因此,最优划分和最优聚类数可以通过求最小 $V^n(c)$ 值得到。其步骤描述如下:

(1)初始化参数 ε、m、$c(n)$;

(2)对 $c=2:c(n)$ 进行如下处理:

①执行 FCM 算法；

②用式(15)计算 $set(c)$，用式(16)计算 $sep(c)$。

(3)进行归一化处理，用式(17)计算有效性指标 $V^n(c)$；

(4)找到最小 $V^n(c)$，C 即为最优聚类数，对应的划分为最优划分。

4 实验结果与分析

本实验研究所采用的数据均来自于某高速公路监控中心数据库，利用 Matlab7 对数据进行仿真。图 1 所示为某高速公路 12 小时内的交通流量统计图。

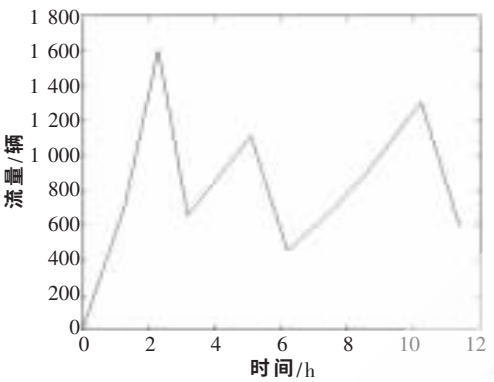

图 1 交通流量统计

针对图 1 中的数据进行采样，首先采用非线性函数估计 SVM 对交通流量进行模拟预测。然后利用 FCM 对交通流量的平均值和上下限区间进行聚类，将聚类结果与 SVM 的模拟预测结果进行比较分析，如表 1 所示。

表 1 SVM 预测与 FCM 聚类结果对比

时间/h	原始数据	SVM 预测分析车流量	FCM 聚类车流量
0.000 2	0	0	0
1.280 0	701	768	689
2.302 0	1 604	1 489	1 587
3.202 3	657	692	643
5.100 0	1 109	1 032	1 035
6.200 0	451	512	421
7.500 9	675	688	653
8.500 0	876	850	847
10.234 1	1 305	1 232	1 243
11.400 0	589	637	578

通过对表 1 的数据分析，利用相对误差公式可以得到如下结论：SVM 的平均相对误差为 0.52%，而采用 FCM 聚类的平均相对误差为 0.271%。采用 Matlab 进行仿真分析两种不同算法的失真情况，如图 2 所示。可以看出 FCM 的收敛程度比 SVM 预测要更快，且精度度更高。

本文采用模糊 C 均值聚类算法对统计数据的区间进行聚类分析，该算法根据 M 值的范围可以减少迭代次数，并利用紧致性函数和分离性函数对模糊 C 均值聚类算法进行有效性分析。实验结果表明，该算法是有效可行的，并且算法的相对平均误差比非线性函数估计的 SVM 算法要少、精度比 SVM 算法要高。

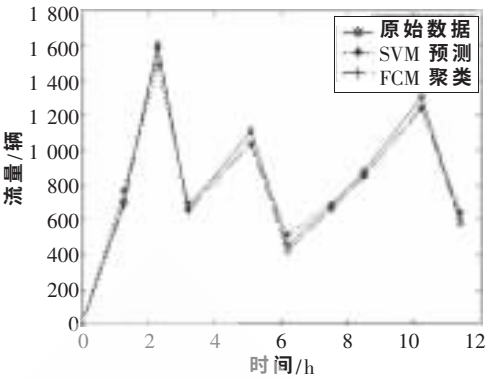

图 2 SVM 预测、FCM 聚类与原始数据比较曲线图

参考文献

[1] SRINIVASAN D, CHOY M C, CHEU R L. Neural networks for real-time traffic signal control [J]. IEEE Trans.on Intelligent Transportation System, 2006,7(3):261-272.

[2] PAKHIRA M K, BANDYOPADHYAY S, MAULIK U. Validity index for crisp and fuzzy clusters[J]. PaRem Recognition, 2004,37(3):487-501.

[3] BASU S, BANERJEE A, Mooney. Semi-supervised clustering by seeding [C]//Proceedings of the International Conference on Machine Learning ICML-2002, 2002:19-26.

[4] SETNES M. Supervised fuzzy clustering for rule Extraction [J]. IEEE Transactions on Fuzzy Systems, 2000,8 (4):416-424.

[5] PAKHIRA M K, BANDYOPADHYAY S, MAULIK U. A study of some fuzzy cluster validity indices,genetic clustering and application to pixel classification[J]. Fuzzy Sets and Systems, 2005,155(2):191-214.

[6] BRINK A D. Gray-level thresholding of images using a correlation criterion [J]. Pattern Recognition Letters. 1989,9 (5):335-341.

[7] LONG B, ZHANG F, WU X Y, et al. Special clustering for multi-type relational data. In: Cohen WW, Moore A, eds. Proc. of the23rd lntl Conf. on Machine Learning[C]. New York: ACM Press. 2006.

[8] 高新波，裴继红，谢维信.模糊 G 均值聚类算法中加权指数 M 的研究[J].电子学报，2000，28(4)：80-83.

(收稿日期：2009-10-10)

作者简介：

曾利军，男，1976 年生，硕士研究生，讲师，主要研究方向：数据挖掘，微分方程数值解，信息检索。

李泽军，男，1972 年生，硕士研究生，讲师，主要研究方向：数据挖掘，文本检索，模式识别。