ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 5,Sep. 2016,pp. 949－957
DOI:10. 16337/j. 1004-9037. 2016. 05. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于模糊聚类的神经元识别方法

张　晶　毕佳佳　张玉红　胡学钢

(合肥工业大学计算机与信息学院,合肥,230009)

摘　要:大脑是生物体内结构和功能最复杂的组织,其中包含上千亿个神经元。作为大脑构造的基本单位,神经元的结构和功能包含很多因素,其中神经元的几何形态特征就是一个重要方面。大脑中神经元的几何形态复杂多样,对其识别分类问题是一个难题。本文在模糊聚类的基础上根据神经元的几何形态建立了模糊集模型,并利用多数据库分类模型中的最优划分模型对模糊聚类分析法进行改进。将改进后的模糊聚类方法用于对神经元的识别分类,得到最优的分类结果。根据聚类的评价方法,与其他的聚类方法比较,证明了改进的模糊聚类方法能够得到更好的聚类效果。

关键词:神经元;模糊集;聚类;划分策略

中图分类号:TP311　　**文献标志码**:A

Recognition Method of Neuron Based on Fuzzy Clustering

Zhang Jing　Bi Jiajia　Zhang Yuhong　Hu Xuegang

(School of Computer and Information, Hefei University of Technology, Hefei, 230009, China)

Abstract: The brain is the most complex tissue in the structure and function of the organism, which contains hundreds of neurons. As a basic unit of the structure of the brain, the structure and function of neurons contain many factors, among which the geometric feature is an important aspect. The morphology of the neurons in brain is so complicated and diversiform that it is a problem to recognize the category of them. Here, we first establish the fuzzy set model based on fuzzy clustering according to the geometry of neurons. We use the optimal classification model of multi-database classification model to improve the fuzzy clustering method and classify the neurons. Then we can obtain the optimal classification result. According to the evaluation method of clustering, we can verify that the improved fuzzy clustering method can get better clustering effect compared with other methods.

Key words: neuron; fuzzy sets; clustering; partitioning strategy

引　　言

随着“人类脑计划”研究的开展,人们对大脑神经元的结构和功能的研究逐渐深入。基于神经元特性的认识,最基本的问题是对神经元的分类识别。如何识别区分不同类别的神经元,这个问题目前科学上仍没有解决。生物解剖区别神经元主要通过几何形态和电位发放两个因素。神经元的几何形态主要

基金项目:国家自然科学基金(61273292)资助项目;国家自然科学基金青年基金(61305063)资助项目。
收稿日期:2015-05-14;**修订日期**:2015-06-24

通过染色技术得到,电位发放通过微电极穿刺胞内记录得到。利用神经元的电位发放模式区分神经元的类别比较复杂,主要涉及神经元的 Hodgkin-Huxley 模型和 Rall 电缆模型的离散形式(神经元的房室模型)。本文只考虑神经元的几何形态,研究如何利用神经元的空间几何特征,通过数学建模给出神经元的一个空间形态分类方法,将神经元根据几何形态比较准确地分类识别,聚类分析是指从已经给定的数据集中寻找数据对象之间的内部结构,即所存在的有价值的数据分布模式。

目前模糊聚类算法由于适应性强,已经在神经元的分类上得到了广泛应用。聚类主要解决的问题是在没有先验知识的前提下如何实现满足上述要求的聚类。相对于普通的聚类分析,模糊聚类分析是一种软划分,它是将数学上的模糊理论应用到聚类分析上,根据事物之间相似性以及属于各个类别的不确定性建立起模糊聚类相似关系并进行分类。1965 年 L. A. Zadeh[1] 创立了模糊集合论,之后 Bellman 和 Kalabaff. Zadeh[2] 在模糊集合论的基础上提出了将模糊集应用到聚类问题上。随后 E. H. Ruspini[3] 又在 1969 年将模糊划分的概念引入了模糊聚类分析中,第一次系统地表述并研究了模糊聚类。现在模糊聚类的应用越来越广泛,模糊理论被应用于各个领域中,如图像分割[4]、天气预报、生物、医学诊断及模式识别等领域都取得了很好效果。模糊聚类方法发展至今种类繁多,但上述列举方法对大数据量的情况进行聚类效果不佳,并且对实时性要求较高的情况聚类效果更是难以令人满意,因此上述方法的实际应用有限。而基于目标函数的方法由于其众多优点在近年受到了人们的广泛欢迎,随着计算机在日常生活中的广泛应用和数据挖掘技术的发展,该类方法已经成为模糊聚类研究领域的热点问题。而在众多基于目标函数的聚类算法中,理论最为完善、应用也最广泛的是模糊 C-均值(Fuzzy C-means, FCM)类型[5] 的算法。

本文针对神经元的几何形态,通过模糊聚类技术,根据神经元的空间几何特征进行划分,提出了一种基于最优划分策略的神经元识别方法,该方法结合模糊聚类中的隶属度和多数据库分类模型中的最优分类模型,定义了一种最优划分策略,选择出最优的隶属度 λ 值。并且本文用提取的神经元的几何特征的数据对该方法进行了检验,并将其结果与传统的聚类方法 K-means 进行了比较。

1 背景模型

模糊聚类中的每个样本不再明确地属于某一类,它是根据隶属度来确定其属于哪一类别,即通过模糊聚类分析后,计算出样本属于各类别的不确定程度,根据其不确定度确定其类别,如此能够更准确地反映出现实世界。同时也可能会出现隶属度不确定的情况,这种情况同样可以用模糊聚类进行分类[6]。模糊聚类不但可以从原始数据中直接提取特征,还能对已经得到的特征进行优选和降维操作,以免造成“维数灾难”。常用的模糊聚类算法有:模糊 C-均值算法[6] 和基于划分的聚类方法[7] 等。

1.1 建立模糊相似矩阵

设样本集合 $X=\{x_1,x_2,\cdots,x_n\}$,n 为样本数目,每个样本 $x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$ 为一组特征数据。由于生活中许多实际问题都具有不同量纲,为了使这些问题中的不同数据可以进行比较,一般先将实测数据作平移变换,将其压缩到[0,1]区间。首先进行标准差变换,即

$$x'_{ik}=\frac{x_{ik}-\overline{x}_k}{s_k} \tag{1}$$

式中:$\overline{x}_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik}$;$s_k=\sqrt{\frac{1}{n}\sum(x_{ik}-\overline{x}_k)^2}$。在此步骤后获得初步转化后的矩阵。然后,采用极差变换,使得 x'_{ik} 均在[0,1]之间

$$x''_{ik}=\frac{x'_{ik}-\min\{x'_{ik}\mid 1\leqslant i\leqslant n\}}{\max\{x'_{ik}\mid 1\leqslant i\leqslant n\mid\}-\min\{x'_{ik}\mid 1\leqslant i\leqslant n\}}\quad i=1,2,\cdots,n;\ k=1,2,\cdots,m \tag{2}$$

最后建立模糊相似矩阵 $\boldsymbol{R}$,主要确定其相似系数,即 x_i 与 x_j 之间的相似程度。求相似系数的方法

很多种,其中最大最小法为

$$r_{ij}=\frac{\sum_{k=1}^{m}\min(x_{ik}-x_{jk})}{\sum_{k=1}^{m}\max(x_{ik}-x_{jk})}\quad i\geqslant 1;j\leqslant n;1\leqslant k\leqslant m \tag{3}$$

1.2　聚类

根据构造的模糊相似矩阵,利用最大树法[8]进行聚类。最大树法的思想是:首先根据实测数据计算出其相似矩阵,然后根据相似矩阵画出一颗最大树。这棵最大树将被分类元素作为其顶点,以模糊相似矩阵 $\boldsymbol{R}$ 的元素 r_{ij} 作其为权重。具体做法是:设 $\boldsymbol{R}$ 为论域 $X=\{x_1,x_2,\cdots,x_n\}$ 上的模糊相似矩阵。首先画出所有顶点 $X_i(i=1,2,\cdots,n)$,在模糊矩阵 $\boldsymbol{R}$ 中根据权重 r_{ij} 的值,从大到小依次画出树枝,并将权重标注上,直到所有的顶点连通为止。这个过程中要求不产生回路,最后就得到了一棵具有 $n-1$ 条边的最大树。在[0,1]区间内为 λ 取值,将权重小于 λ 的树枝减去,便得到一个不连通的图。那么各个连通的分支便构成了在 λ 水平上的一个类,所有的子图集合即形成一种聚类结果。

2　最优划分策略

在模糊聚类中,λ 为模糊相似矩阵 $\boldsymbol{R}$ 的隶属度,选择不同的 λ 值会将样本分为不同的类。模糊聚类就是在已经建立的模糊等价关系矩阵上,根据给定不同的 λ 水平进行截取,从而得到不同的分类。λ 的值越小,则划分类别就越少、越粗;λ 值越大,划分的类别就越多、越细。当选取到最优的 λ 值时,就会选择最合理的聚类结果。由于 λ 取值不同,划分的结果不同,所以 λ 的取值直接关系到聚类结果的准确性。一般的处理方法都是领域专家根据实际情况来进行指导分析,人工选取最优的划分方式,这势必受专家对此问题认识的程度的影响。为了得到较好的聚类效果,λ 的选择必然成为关键。基于此,本文利用多数据库分类模型中的最优分类模型[9],定义了一种最优划分策略。

定义 1　$\mathrm{Cluster}(X,\lambda)=\{xc_1^\lambda,xc_2^\lambda,xc_3^\lambda,\cdots,xc_k^\lambda\}$ 为样本集合 $X=\{x_1,x_2,\cdots,x_n\}$(n 为样本数目)的一个 λ-划分(λ 为一个划分中的样本集合数目),如果 $\mathrm{Cluster}(X,\lambda)$ 同时满足以下条件,则它是完备的 λ-聚类:

(1) $xc_1^\lambda\cup c_2^\lambda\cup c_3^\lambda\cup\cdots\cup c_k^\lambda=X$;

(2)对于 $\forall xc_i^\lambda,\forall xc_j^\lambda\in\mathrm{Cluster}(X,\lambda)$,有 $\forall xc_i^\lambda\cap\forall xc_j^\lambda=\varnothing(i\neq j,1\leqslant i,j\leqslant k)$;

(3)对于 $\forall x_i\in xc_l^\lambda,\forall x_j\in xc_m^\lambda(i\neq j,1\leqslant i,j\leqslant n;l\neq m,1\leqslant l,m\leqslant k)$ 则 $\mathrm{sim}(x_i,x_j)<\lambda$。

定义 2　假设 $xc_m^\lambda\in\mathrm{Cluster}(X,\lambda)$,$1\leqslant m\leqslant k$。它的 λ-划分局部距离 $\mathrm{InterValue}(xc_m^\lambda)$ 定义为

$$\mathrm{InterValue}(xc_m^\lambda)=\sum_{x_i,x_j\in c_m^\lambda}^{i\neq j}(1-\mathrm{sim}(x_i,x_j)) \tag{4}$$

式中:$1-\mathrm{sim}(x_i,x_j)$ 为 x_i 和 x_j 间的距离,代表了两个样本之间的差异程度。$\mathrm{InterValue}(xc_m^\lambda)$ 描述了一种划分之后某个集合内部的每对样本之间的距离,其值越低,说明类内每个样本的距离越短,即耦合度越高。

定义 3　假设 $\mathrm{Cluster}(X,\lambda)=\{xc_1^\lambda,xc_2^\lambda,xc_3^\lambda,\cdots,xc_k^\lambda\}$ 为 X 的一个完备 λ-聚类,$\mathrm{Cluster}(X,\lambda)$ 的 λ-划分总体距离定义为

$$\mathrm{OuterValue}(X,\lambda)=\sum_{m=1}^{k}\mathrm{InterValue}(xc_m^\lambda) \tag{5}$$

式中:$\mathrm{InterValue}(xc_m^\lambda)$ 为描述了一种划分之后,某个集合内部每对样本之间的距离。而 $\mathrm{OuterValue}(X,\lambda)$ 描述了一个 λ-划分所有样本内距离之和,同样其值越小越好。因此可以得出,需要的 k 值要尽可能的小。

引理 1 设 $T=\{V(T),E(T)\}$为最大最小树,$W(e_{ij})=\mathrm{sim}(x_i,x_j)$,$1\leqslant i,j\leqslant n$ 为相应边上的权值,且 $e_{ij}\in E(T)$。令 $\lambda=\{\lambda_1,\cdots,\lambda_{m-1}\}$,$\lambda_k\in W(E(T))$,$1\leqslant k\leqslant n$ 且 $\lambda_i\neq\lambda_j$,$1\geqslant i,j\geqslant m-1$,且 $1=\lambda_0>\lambda_1>\cdots>\lambda_{m-1}\geqslant 0$,则 $|\mathrm{Cluster}(X,\lambda_{k-1})|-|\mathrm{Cluster}(X,\lambda_k)|\geqslant 1$,$1\leqslant k\leqslant n$。

证明 假设 $|\mathrm{Cluster}(X,\lambda_{k-1})|-|\mathrm{Cluster}(X,\lambda_k)|<1$ 成立。令 $k=1$,则 $|\mathrm{Cluster}(X,\lambda_0)|-|\mathrm{Cluster}(X,\lambda_1)|<1$,即 $|\mathrm{Cluster}(X,1)|-|\mathrm{Cluster}(X,\lambda_1)|<1$。因为 $|\mathrm{Cluster}(X,\lambda_0)|=|\mathrm{Cluster}(X,1)|$,所以 $\lambda\geqslant 1$;又因 $\lambda=\{\lambda_1,\cdots,\lambda_{n-1}\}$,$1>\lambda_1\geqslant\cdots\geqslant\lambda_{n-1}\geqslant 0$,所以在 T 中不存在任何一个 $W(e_{ij})\geqslant 1$,所以,没有任何树的结点通过边相连,每个结点自成一类,故 $|\mathrm{Cluster}(X,1)|=n$。又因 $\mathrm{sim}(i,i)=1=\lambda_0>\lambda_1>\cdots>\lambda_{n-1}>0$,所以,在 $W(e_{ij})$中至少存在一个值等于 λ_1,使得 $W(e_{ij})=\mathrm{sim}(i,j)=\lambda_1$,故至少存在一条边 e_{ij},使得 $|\mathrm{Cluster}(X,\lambda_1)|\leqq n-1$。即 $|\mathrm{Cluster}(X,\lambda_1)|\leqq|\mathrm{Cluster}(X,1)|-1$。故 $|\mathrm{Cluster}(X,1)|-|\mathrm{Cluster}(X,\lambda_1)|\geqslant 1$ 与假设矛盾,结论成立。

定理 1 设 $T=\{V(T),E(T)\}$为最大最小树,$W(e_{ij})=\mathrm{sim}(i,j)$,$1\leqslant i,j\leqslant n$ 为相应边上的权值,且 $e_{ij}\in E(T)$,令 $\lambda=\{\lambda_1,\cdots,\lambda_{m-1}\}$,$\lambda_k\in W(E(T))$,$1\leqslant k\leqslant n$ 且 $\lambda_i\neq\lambda_j$,$1\geqslant i,j\geqslant m-1$,且 $1=\lambda_0>\lambda_1>\cdots>\lambda_{m-1}\geqslant 0$,则 $|\mathrm{Cluster}(X,\lambda)|$为单调递增序列。

证明 由引理 1 得,当 $1=\lambda_0>\lambda_1>\cdots>\lambda_{m-1}\geqslant 0$ 时,有

$$|\mathrm{Cluster}(X,1)|-|\mathrm{Cluster}(X,\lambda_1)|\geqslant 1 \tag{6-1}$$

$$\vdots$$

$$|\mathrm{Cluster}(X,\lambda_i)|-|\mathrm{Cluster}(X,\lambda_{i+1})|\geqslant 1 \tag{6-i}$$

$$\vdots$$

$$|\mathrm{Cluster}(X,\lambda_j)|-|\mathrm{Cluster}(X,\lambda_{j+1})|\geqslant 1 \tag{6-j}$$

$$\vdots$$

$$|\mathrm{Cluster}(X,\lambda_{m-1})|-|\mathrm{Cluster}(X,0)|\geqslant 1 \tag{6-(m-1)}$$

因此当 $\lambda_i>\lambda_j$,式(6-i)到式(6-(j-1))分别左右相加得:$|\mathrm{Cluster}(X,\lambda_i)|-|\mathrm{Cluster}(X,\lambda_j)|\geqslant j-i$。所以 $|\mathrm{Cluster}(X,\lambda_i)|-(j-i)\geqslant|\mathrm{Cluster}(X,\lambda_j)|$。又因为 $\lambda_i>\lambda_j$,且 $\lambda_0>\lambda_1>\cdots>\lambda_{m-1}$,所以 $j>i$,故 $j-i>0$。所以 $|\mathrm{Cluster}(X,\lambda_i)|>|\mathrm{Cluster}(X,\lambda_j)|$,$|\mathrm{Cluster}(X,\lambda)|$为单调递增序列。

引理 2 设 $T=\{V(T),E(T)\}$是最大最小树,$W(e_{ij})=\mathrm{sim}(i,j)$,$1\leqslant i,j\leqslant n$ 为相应边上的权值,且 $e_{ij}\in E(T)$,令 $\lambda=\{\lambda_1,\cdots,\lambda_{m-1}\}$,$\lambda_k\in W(E(T))$,$1\leqslant k\leqslant n$ 且 $\lambda_i\neq\lambda_j$,$1\geqslant i,j\geqslant m-1$,且 $1=\lambda_0>\lambda_1>\cdots>\lambda_{m-1}\geqslant 0$,则必存在 $\varepsilon>0$,使得 $\mathrm{OuterValue}(\lambda_k)-\mathrm{OuterValue}(\lambda_{k-1})\geqslant\varepsilon$,$1\leqslant k\leqslant n$。

当 λ 递减,相应的孤立结点将与另一些结点相连,更新类的划分。则每失去一个孤立点,加入到相应的类中,其余的点都要与这个点进行距离计算,更新 Value 值。由于原孤立点的 Value 为 0,OuterValue 在逐渐地增大。

证明 假设 $\mathrm{OuterValue}(\lambda_k)<\mathrm{OuterValue}(\lambda_{k-1})$,则令 $k=1$,$|\mathrm{Cluster}(X,1)|=n$,故 $\mathrm{sim}(i,j)=0$,所以使得 $\mathrm{InterValue}(1)=\sum_{di,dj\in\mathrm{class}_s^a}^{i=j}(1-\mathrm{sim}(i,j))=0$。而 $|\mathrm{Cluster}(X,\lambda_1)|\leqslant n-1$,由于至少存在一个 $W(e_{ij})=\lambda_1$,所以假设有 s 个 $W(e_{ij})$ 与 λ_1 相等,其他结点不变。故 $\mathrm{OuterValue}(\alpha_0)=\sum_{m=n_0}^{n_{-1}}\mathrm{InterValue}(xc_m^{\lambda_0})+\sum_{m=n_s}^{n_{s-1}}\mathrm{InterValue}(xc_m^{\lambda_s})$,对应的:$\mathrm{OuterValue}(1)=\sum_{m=n_0}^{n_{-1}}\mathrm{Value}(xc_m^1)+\sum_{x=n_s}^{n_{s-1}}\mathrm{Value}(xc_m^1)$,其中 $\sum_{m=n_s}^{n_{s-1}}\mathrm{InterValue}(xc_x^1)=\sum_{m=n_{s+1}}^{n_{s-1}}\mathrm{Value}(xc_m^{\lambda_0})=0$。用 k_i 表示 s 中相互连通的结点数,$\sum_{m=n_0}^{n_{-1}}\mathrm{InterValue}(xc_m^{\lambda_0})=\sum_{m=k_1}^{k_i}C_m^2(1-\mathrm{sim}(i,j))=\sum_{m=k_1}^{k_i}\frac{m}{2}((m-1)(1-\mathrm{sim}(i,j))$,$\sum_{i=1}^{t}k_i=s$,$m\geqslant 2$,令:$f(m)=(m-1)(1-\mathrm{sim}(i,$

$j))$,$m\geqslant 2$。由于$0<(1-\mathrm{sim}(i,j))\leqslant 1$,所以$0<f(m)\leqslant(m-1)$。对于$\sum_{m=n_s}^{n_{s-1}}\mathrm{InterValue}(xc_m^1)$,根据$\mathrm{sim}(i,j)>\lambda_1$同样可以分成$t$部分,每部分中连通的结点数也用$k_i$表示。则$\sum_{m=n_0}^{n_{s-1}}\mathrm{InterValue}(xc_m^1)=\sum_{m=k_1}^{k_i}0$。令$g(m)=0,m\geqslant 2$,故$g(m)=0<f(m)\leqslant(m-1)$。所以,$\sum_{m=n_0}^{n_{s-1}}\mathrm{InterValue}(xc_m^{\lambda_1})>\sum_{m=n_0}^{n_{s-1}}\mathrm{InterValue}(xc_m^1)>\sum_{m=n_0}^{n_{s-1}}\mathrm{InterValue}(xc_m^{\lambda_0})$。

综上所述,$\mathrm{OuterValue}(\lambda_1)>\mathrm{OuterValue}(\lambda_0)$与假设矛盾。所以存在$\varepsilon>0$,使得$\mathrm{OuterValue}(\lambda_1)-\mathrm{OuterValue}(\lambda_0)\geqslant\varepsilon$。

定理 2　设$T=\{V(T),E(T)\}$为最大最小树,$W(e_{ij})=\mathrm{sim}(i,j)$,$1\leqslant i,j\leqslant n$为相应边上的权值,且$e_{ij}\in E(T)$,令$\lambda=\{\lambda_1,\cdots,\lambda_{m-1}\}$,$\lambda_k\in W(E(T))$,$1\leqslant k\leqslant n$且$\lambda_i\neq\lambda_j$,$1\geqslant i,j\geqslant m-1$,且$1=\lambda_0>\lambda_1>\cdots>\lambda_{m-1}\geqslant 0$,则$\mathrm{OuterValue}(\lambda)$为单调递减序列。

证明　由引理 2 知,当$\lambda_0>\lambda_1>\cdots>\lambda_{m-1}$时,有

$$\mathrm{OuterValue}(\lambda_1)-\mathrm{OuterValue}(\lambda_0)\geqslant\varepsilon \tag{7-1}$$

$$\vdots$$

$$\mathrm{OuterValue}(\lambda_i)-\mathrm{OuterValue}(\lambda_{i-1})\geqslant\varepsilon \tag{7-i}$$

$$\vdots$$

$$\mathrm{OuterValue}(\lambda_j)-\mathrm{OuterValue}(\lambda_{j-1})\geqslant\varepsilon \tag{7-j}$$

$$\vdots$$

$$\mathrm{OuterValue}(\lambda_{m-1})-\mathrm{OuterValue}(\lambda_{m-2})\geqslant\varepsilon \tag{7-(m-1)}$$

因此当$\lambda_i>\lambda_j$时,式(7-(i+1))到式(7-j)分别左右相加得$\mathrm{OuterValue}(\lambda_j)-\mathrm{OuterValue}(\lambda_i)\geqslant(j-i)\varepsilon$。因为$j>i$,且$\varepsilon>0$;$\mathrm{OuterValue}(\lambda_j)>\mathrm{OuterValue}(\lambda_i)$,所以,$\mathrm{OuterValue}(\lambda)$为单调递减序列。

通过实验和理论分析可以得到,$k=|\mathrm{Cluster}(x,\lambda)|$和 OuterValue 在一个$\lambda$阈值的制约下,分别是增函数和减函数,无法同时达到最小。如何选择最优划分的λ值问题,就转化成如何在$k=|\mathrm{Cluster}(x,\lambda)|$和 OuterValue 寻求一个平衡点,故引入定义 4:

定义 4　假设$\mathrm{Cluster}(X,\lambda)=\{xc_1^\lambda,xc_2^\lambda,xc_3^\lambda,\cdots,xc_k^\lambda\}$为$X$的一个完备$\lambda$-聚类,定义$\mathrm{Distance}(X,\lambda)=|\mathrm{OuterValue}(X,\lambda)-k|$。这样寻求最优划分策略问题就转化为求最小的 Distance。根据如上定义求的λ,即为最优划分策略下的λ。

定理 3　$\mathrm{Distance}(\lambda)=|\mathrm{OuterVlaue}(\lambda)-|\mathrm{Cluster}(x,\lambda)||$在$\lambda\in[1,0]$内,存在最小值点。

证明　由定理 1 知,$|\mathrm{Cluster}(x,\lambda)|$在$\lambda\in[1,0]$区间内单调递增,$\lambda$越大,所划分的类就越多,$|\mathrm{Cluster}(x,\lambda)|$越大。所以,$|\mathrm{Cluster}(x,\lambda)|$的值域为$[1,n]$,$n$为树的结点数。

由定理 2 知,$\mathrm{OuterValue}(\lambda)$在$\lambda\in[1,0]$区间内单调递减,$\lambda$越大,类的内聚值越大,$\mathrm{OuterValue}(\lambda)$越大。所以 $\mathrm{OuterValue}(\lambda)$的值域为$[0,\beta]$,其中$\beta=\sum_{k=1}^{\mathrm{NoC}(0)}\frac{n_k}{2}((n_k-1)(1-\mathrm{sim}(i,j)))$,$n_k$表示在$\lambda=0$下不同类的结点的数量。

所以,当λ增大,$|\mathrm{Cluster}(x,\lambda)|$从$1\sim n$变化,$\mathrm{OuterValue}(\lambda)$从$\beta\sim 0$变化。当$\beta\leqslant 1$时,则在$\lambda_i(\mathrm{OuterValue}(\lambda_i)=\beta)$取得最小值。当$\beta>1$时,根据变化趋势,显然存在最小值。故存在最小值点,且在实验部分对其进行了验证。

3　实验与分析

大脑是生物体内结构和功能最复杂的组织,其中包含上千亿个神经细胞(神经元)。人类脑计划

(Human brain project, HBP)的目的是要对全世界的神经信息学数据库建立共同的标准,多学科整合分析大量数据,加速人类对脑的认识。

神经元复杂多样,对其识别仍是个难题。生物解剖方面主要透过神经元的两个因素来区分其类别,即几何形态和点位发放。其中,神经元的几何形态主要通过染色技术得到,电位发放通过微电极穿刺胞内记录得到,但是利用神经元的电位发放模式来识别神经元是比较复杂的。本实验针对神经元的几何形态,通过模糊聚类技术,根据神经元的空间几何特征进行划分,再以最优划分策略进行选择,目的在于能够根据神经元几何形态比较准确地分类识别。

3.1 实验数据准备

实验选取了 George Mason University 提供的 Neuro Morpho. Org 神经元源数据。神经元数据由<房室标号,房室类型,房室的 x 坐标,房室的 y 坐标,房室的 z 坐标,房室的半径,与该房室连接的母房室标号>这 7 元组组成。由于缺少了所属信息的描述,所以需要先通过 L-measure 软件对这个元组的数据进行处理,得到神经元的几何形态特征信息。首先将神经元数据通过 L-measure 软件得到 20 个几何特征属性数据,形成一个 20×20 的矩阵 $\boldsymbol{B}_1$,然后分别对 5 种神经元统计每个属性的均值,得到一个 5×20 的矩阵 $\boldsymbol{B}_2$。最后,先设论域 $X=\{x_1, x_2, \cdots, x_n\}$($n=25$ 为样本数目)为所需要研究对象,每个神经元 x_i 由一组特征数据 $\{x_{i1}, x_{i2}, \cdots, x_{im}\}$ 表示($m=20$, 为其几何属性),于是可以得到问题的原始数据矩阵 $\boldsymbol{A}_{25\times 20}$。对 $\boldsymbol{A}$ 矩阵进行平移变换、最大最小法得到的相似矩阵如图 1 所示。该图中的每一行的 20 个元素代表着神经元的几何特征值,如胞体表面积、干的数目等。对图 1 的相似矩阵利用最大树法进行聚类,得到的树如图 2 所示。

1 0.746 0.731 0.532 0.51 0.512 0.435 0.441 0.441 0.369 0.468 0.35 0.486 0.415 0.67 0.605 0.604 0.658 0.65 0.599 0.243 0.215 0.385 0.331 0.164
0.746 1 0.676 0.484 0.59 0.596 0.467 0.474 0.474 0.336 0.425 0.335 0.484 0.432 0.703 0.568 0.719 0.677 0.678 0.535 0.227 0.216 0.469 0.302 0.195
0.731 0.676 1 0.541 0.5 0.505 0.367 0.378 0.378 0.383 0.473 0.351 0.571 0.478 0.656 0.702 0.515 0.579 0.712 0.679 0.264 0.334 0.365 0.444 0.203
0.532 0.484 0.541 1 0.37 0.377 0.44 0.436 0.436 0.433 0.505 0.443 0.329 0.27 0.436 0.41 0.377 0.381 0.507 0.463 0.307 0.181 0.251 0.286 0.084
0.51 0.59 0.5 0.37 1 0.896 0.454 0.466 0.466 0.312 0.368 0.288 0.531 0.455 0.592 0.537 0.586 0.585 0.469 0.458 0.194 0.229 0.666 0.2 0.229
0.512 0.596 0.505 0.377 0.896 1 0.492 0.498 0.498 0.302 0.363 0.281 0.489 0.422 0.559 0.504 0.567 0.562 0.472 0.453 0.191 0.217 0.649 0.21 0.206
0.435 0.467 0.367 0.44 0.454 0.492 1 0.925 0.925 0.528 0.533 0.403 0.231 0.19 0.334 0.269 0.36 0.327 0.403 0.262 0.398 0.084 0.348 0.144 0.049
0.441 0.474 0.378 0.436 0.466 0.498 0.925 1 1 0.538 0.533 0.421 0.24 0.198 0.34 0.279 0.366 0.333 0.413 0.266 0.403 0.091 0.358 0.148 0.056
0.441 0.474 0.378 0.436 0.466 0.498 0.925 1 1 0.538 0.533 0.421 0.24 0.198 0.34 0.279 0.366 0.333 0.413 0.266 0.403 0.091 0.358 0.148 0.056
0.369 0.336 0.383 0.433 0.312 0.302 0.528 0.538 0.538 1 0.738 0.614 0.255 0.215 0.306 0.314 0.234 0.236 0.416 0.324 0.769 0.158 0.193 0.186 0.09
0.468 0.425 0.473 0.505 0.368 0.363 0.533 0.533 0.533 0.738 1 0.645 0.326 0.295 0.403 0.415 0.33 0.323 0.547 0.436 0.583 0.187 0.241 0.237 0.154
0.35 0.335 0.351 0.443 0.288 0.281 0.403 0.421 0.421 0.614 0.645 1 0.252 0.263 0.335 0.316 0.253 0.243 0.395 0.298 0.545 0.178 0.219 0.199 0.15
0.486 0.484 0.571 0.329 0.531 0.489 0.231 0.24 0.24 0.255 0.326 0.252 1 0.775 0.59 0.763 0.549 0.661 0.472 0.548 0.137 0.284 0.312 0.177 0.361
0.415 0.432 0.478 0.27 0.455 0.422 0.19 0.196 0.198 0.215 0.295 0.263 0.775 1 0.554 0.647 0.484 0.566 0.412 0.492 0.12 0.317 0.329 0.158 0.457
0.67 0.703 0.656 0.436 0.592 0.559 0.334 0.34 0.34 0.306 0.403 0.335 0.59 0.554 1 0.747 0.652 0.681 0.575 0.636 0.207 0.324 0.443 0.348 0.273
0.605 0.568 0.702 0.41 0.537 0.504 0.269 0.279 0.279 0.314 0.415 0.316 0.763 0.647 0.747 1 0.56 0.613 0.605 0.681 0.191 0.354 0.341 0.287 0.318
0.604 0.719 0.515 0.377 0.586 0.567 0.36 0.366 0.366 0.234 0.33 0.253 0.549 0.484 0.652 0.56 1 0.812 0.57 0.533 0.117 0.156 0.444 0.183 0.214
0.658 0.677 0.579 0.381 0.585 0.562 0.327 0.333 0.333 0.236 0.323 0.243 0.661 0.566 0.681 0.613 0.812 1 0.545 0.581 0.119 0.186 0.434 0.201 0.237
0.65 0.678 0.712 0.507 0.469 0.472 0.403 0.413 0.413 0.416 0.547 0.395 0.472 0.412 0.575 0.605 0.57 0.545 1 0.646 0.289 0.24 0.348 0.346 0.185
0.599 0.535 0.679 0.463 0.458 0.453 0.262 0.266 0.266 0.324 0.436 0.298 0.548 0.492 0.636 0.681 0.533 0.581 0.646 1 0.191 0.277 0.297 0.331 0.185
0.243 0.227 0.264 0.307 0.194 0.191 0.398 0.403 0.403 0.769 0.583 0.545 0.137 0.12 0.207 0.191 0.117 0.119 0.289 0.191 1 0.2 0.203 0.228 0.112
0.215 0.216 0.334 0.181 0.229 0.217 0.084 0.091 0.091 0.158 0.187 0.178 0.284 0.317 0.324 0.354 0.156 0.186 0.24 0.277 0.2 1 0.312 0.473 0.474
0.385 0.469 0.365 0.251 0.666 0.649 0.348 0.358 0.358 0.193 0.241 0.219 0.312 0.329 0.443 0.341 0.444 0.434 0.348 0.297 0.203 0.312 1 0.295 0.293
0.331 0.302 0.444 0.286 0.2 0.21 0.144 0.148 0.148 0.186 0.237 0.199 0.177 0.158 0.348 0.287 0.183 0.201 0.346 0.331 0.228 0.473 0.295 1 0.246
0.164 0.195 0.203 0.084 0.229 0.206 0.049 0.056 0.056 0.09 0.154 0.15 0.361 0.457 0.273 0.318 0.214 0.237 0.185 0.185 0.112 0.474 0.293 0.246 1

图 1 神经元数据的相似矩阵

Fig. 1 Similar matrix of neuron data

3.2 Distance 分布

在对该神经元数据进行聚类划分之前,先对定理 3 中 Distance 有最小值进行验证,本文选用了 3 个不同的数据集,分别对每个划分下的 Distance 进行实验,实验结果如图 3 所示。根据定理 3 以及实验验证所得,3 种数据集分别在阈值 λ 为 0.867, 0.91 和 0.99 时 Distance 取得最小值。通过该实验验证了 Distance 有最小值。

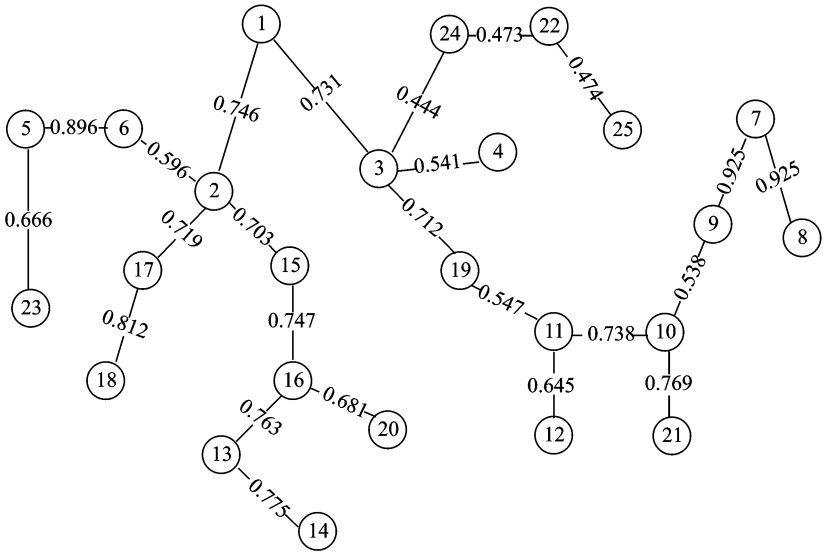

图 2　最大数聚类图

Fig. 2　Clustering graph of maximum tree

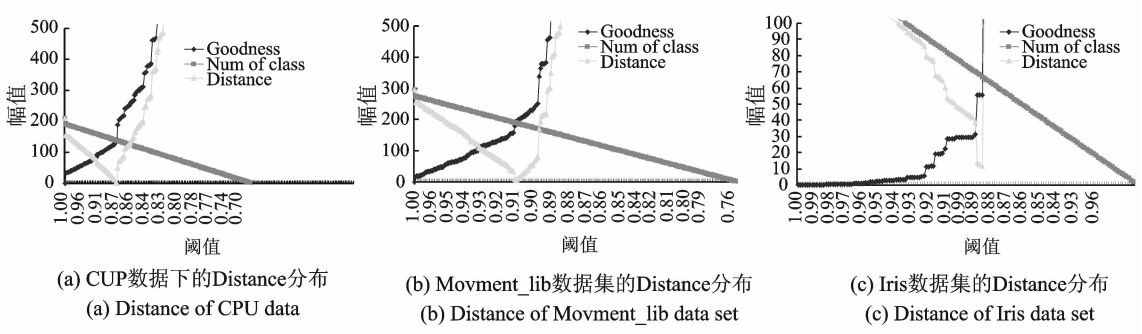

图 3　3 种数据集的 Distance 分布

Fig. 3　Distance distribution of the three data sets

3.3　模糊聚类

最大树法进行聚类后,对图 2 的树进行划分,令 λ 依次取得从大到小的值,将树按连接情况分成不同数量的类。因为图 2 中的树有 25 个结点,24 条边值对应于 24 个 λ 值,所以会产生 24 种聚类结果。根据最优划分策略,进行聚类结果的选择。通过实验所得,当 $\lambda>0.645$ 时,Distance($\boldsymbol{A}$,0.645)=1.616;当 $\lambda>0.596$ 时,Distance($\boldsymbol{A}$,0.596)=0.094;当 $\lambda>0.547$ 时,Distance($\boldsymbol{A}$,0.547)=1.902。所以当 $\lambda>0.596$时,分为如下的 8 类:

{{4},{23,5,6},{1,2,3,13,14,15,16,17,18,18,20},{11,12,10,21},{7,8,9},{22},{24},{25}},其中{{22},{23},{24},{25}}为加入的已知类,而{10,11,12}与{21}都属于 Motor Neuron 类,{5,6}与{23}都属于 InterNeuron 类,其余的为新类。所以有必要引入新的神经元类别。

3.4　基于已识类标签的比较

本文根据聚类算法的几种评价指标,将本文改进的模糊聚类方法与 K-means 方法相比较,验证该方法的有效性。评价方法分别为:(1)Purity。簇 C_k 的 Purity 定义为:根据已识类标识 $t\in T$,标识为该

类的数据占整个簇的比例，即 $\mathrm{Pur}(C_k)=\max(N_{tk}/N_k)$，其中 $t\in T$，N_k 为簇 C_k 的大小，N_{tk} 为该簇中标识为类 t 的数量。在整个划分中，纯净度 $P(C)$ 为所有簇的纯净度的均值，$P(C)\in[0,1]$，$P(C)$ 越大，纯净度越高，划分越好，但是可以预见，如果单一类的簇越多，纯净度也越高。极限情况，当每个实例自成一类，纯净度为 1，所以纯净度在一定程度上说明了聚类的质量。两种方法的比较结果如表 1 所示。

由表 1 中的实验结果可以看出，模糊聚类所认为的 3 个数据集的最优分类数分别为 138，189 和 67。与 Weka 平台下的 Simple K-means 进行比较，设置初始的聚类数分别为 138，189 和 67。由实验结果可知，相对于聚类算法 K-means 的质量，本文中改进的模糊聚类的质量略高。为了避免纯净度的片面性，需使用其他聚类质量评价方式对两种算法在相同分类数的情况下进行聚类质量的比较。

表 1 两种聚类方法的纯净度比较

Tab. 1 Purity comparison of two clustering methods

数据集	K-means	模糊聚类
CPU-class(140)	0.603	0.636
Movement_lib	0.135	0.200
Iris	0.080	0.253

(2) RI (Rand index) [10] 指标。RI 利用排列组合的方式对两类的划分进行评价，即

$$\mathrm{RI}=\frac{TP+FP}{TP+FP+TN+FN} \tag{8}$$

式中：$TP=|\{i,j\}|C_u(i)=C_u(j)\wedge C_v(i)=C_v(j)\}|$；$TN=|\{i,j\}|C_u(i)\neq C_u(j)\wedge C_v(i)\neq C_v(j)\}|$；$FP=|\{i,j\}|C_u(i)=C_u(j)\wedge C_v(i)\neq C_v(j)\}|$；$FN=|\{i,j\}|C_u(i)\neq C_u(j)\wedge C_v(i)=C_v(j)\}|$。$TP$ 和 TN 计算两个划分的一致性，而 FP 和 FN 计算其偏差。对于全局而言，需要计算任意两个类之间的一致性和偏差的和。由于 $\mathrm{RI}\in[0,1]$，所以只有在任意两个簇分配完全一致的情况下，才有 RI=1。RI 的值越小，表示这两个划分的差异就越大。因为 RI 值越小，那么聚类的质量就越好。实验结果如表 2 所示，从表 2 可知，模糊聚类的质量与 K-means 的质量差不多。

表 2 两种聚类方法的 RI 比较

Tab. 2 RI comparison of two clustering methods

数据集	K-means	模糊聚类
CPU-class(140)	0.984	0.983
Movement_lib	0.924	0.924
Iris	0.679	0.703

(3) 互信息。给定了 g 个类 V_h，其中 $h\in[1,g]$。经过模糊聚类算法，分成了 k 个簇，记为 C_i，$i\in[1,k]$，可以用类标签 V_h 评估聚类质量。为了评估一个单一的簇，可以用纯度 Purity 和 Entropy，然而整个的聚类应该用互信息去评价，互信息如下

$$Q(V,C)=\frac{1}{n}\sum_{l=1}^{k}\sum_{h=1}^{g}n_l^{(h)}\frac{\log\left[\dfrac{n_l^{(h)}\times n}{\sum_{i=1}^{k}n_i^{(h)}\sum_{i=1}^{g}n_i^{(i)}}\right]}{\log(k\times g)} \tag{9}$$

式中：$N_l{}^{(h)}$ 为一个簇 C_l 中被分到第 h 类的实例数；n 为总的实例数，实验结果如表 3 所示。

由表 3 可知，模糊聚类的互信息略大于 K-means，亦即在相同聚类数情况下，模糊聚类的质量略高于 K-means 算法。

表 3 两种聚类方法的互信息比较

Tab. 3 Mutual information comparison of two clustering methods

数据集	K-means	模糊聚类
CPU-class(140)	0.263	0.264
Movement_lib	0.187	0.185
Iris	0.138	0.143

4 结束语

本文在聚类的基础上提出了基于模糊聚类的神经元识别方法，该方法通过使用已知神经元类别做分类指导，利用模糊聚类法对未知的神经元进行分类，并利用多数据库定义的最优分类，合理地选取最优的隶属度 λ 的值，并根据此值进行分类，实验验证了该方法的有效性。与传统的聚类方法 K-means 相比，该模糊聚类方法提出了最优划分策略，因此聚类的质量更好一些。然而，由于分类过程中利用已知

类的中心点代表已知类,在极端情况下难免会出现类别误差。由于本文突出重点是用模糊方法识别新类和最优划分策略等多个方面,后期将进行更深入的比较工作。

参考文献:

[1] Zadeh L A. Fuzzy sets [J]. Information and Control, 1965,8(3):338-353.

[2] Bellman R E, Zadeh L A. Decision-making in a fuzzy environment[J]. Management Science, 1970,17(4):141-164.

[3] Enrique H R. A new approach to clustering[J]. Information and Control, 1969,15(1):22-32.

[4] 闫晓玲,王黎明,卜乐平. 基于多维彩色向量空间的火焰图像模糊聚类分割算法[J]. 数据采集与处理,2012,27(3):368-373.
Yan Xiaoling, Wang Liming, Bu Leping. Fuzzy clustering segmentation algorithm of flame image based on multi-dimensional color vector space[J]. Journal of Data Acquisition and Processing, 2012, 27(3):368-373.

[5] Ehrlich R, Bezdek J C, Full W. FCM: The fuzzy c-means clustering algorithm[J]. Computer & Geosciences, 1984,10(2/3):191-203.

[6] 陈健美,陆虎,宋余庆,等. 一种隶属关系不确定的可能性模糊聚类方法[J]. 计算机研究与发展, 2008, 45(9):1486-1492.
Chen Jianmei, Lu Hu, Song Yuqing, et al. A possibility fuzzy clustering algorithm based on the uncertainty membership[J]. Journal of Computer Research and Development, 2008,45(9):1486-1492.

[7] 张敏,于剑. 基于划分的模糊聚类算法[J]. 软件学报, 2004,15(6):858-868.
Zhang Min, Yu Jian. Fuzzy partitional clustering algorithms[J]. Journal of Software, 2004,15(6):858-868.

[8] Wu Z, Leathy R. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation [C]// IEEE Transactions on Pattern Analysis and Machine. [S. l.]: IEEE,1993: 1101-1113.

[9] Li Hong, Hu Xuegang, Zhang Yanming. An improved database classification algorithm for multi-database[C]// Proceedings of Third International Workshop. Hefei: Springer-Verlag Berlin Heidelberg, 2009:346-357.

[10] Rand W. Objective criteria for the evaluation of clustering methods [J]. Journal of the American Statistical Association, 1971, 66(336): 846-850.

作者简介:

张晶(1976-),女,副教授,研究方向:数据挖掘、领域知识,E-mail:jzhang_zj@163.com。

毕佳佳(1989-),女,硕士研究生,研究方向:智能计算理论与软件。

张玉红(1979-),女,副教授,研究方向:数据挖掘、机器学习。

胡学钢(1961-),男,教授,研究方向:数据挖掘、知识工程。