

新的模糊聚类有效性指标

赵娜娜, 钱雪忠, 冯振华

(江南大学 物联网工程学院, 江苏 无锡 214122)

摘 要: 为了更有效地确定模糊聚类算法的最佳聚类数, 提出了一个新的有效性指标. 该指标考虑了聚类的紧致度、重叠度、分离度. 紧致度是用来衡量类内相似程度; 重叠度是用来衡量类间的重叠程度; 分离度用来衡量类间的分离程度. 利用该指标可以找到最符合数据自然分布的聚类数. 实验结果证明, 新的指标均能发现最优聚类数, 从而克服了模糊 c 均值 (FCM) 算法聚类数需要预先设定的缺点, 并能够准确地判断含有交叠子类的最佳聚类数.

关键词: FCM 算法; 聚类指标; 紧致度; 重叠度; 分离度

中图分类号: TP391.9

文献标识码: A

文章编号: 1000—7180(2016)08—0121—05

Novel Validity Index for Fuzzy Clustering

ZHAO Na-na, QIAN Xue-zhong, FENG Zhen-hua

(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: A novel validity indice is proposed to determine the optimal number of clusters for fuzzy clustering. The novel validity indice considers the degree of compactness, the degree of overlapping, and the degree of separation. The compactness measures the similarity within a cluster. The degree of overlapping measures the overlap between clusters. Meanwhile, the degree of separation is used to measure the degree of the clear between clusters. The optimal cluster number can be effectively found by the proposed index. The experimental results show that the optimal cluster number are obtained which are general used in FCM algorithm. the new index overcomes the shortcomings of FCM that the cluster number must be pre-assigned and works well in the situations when there are overlapping subcluster in the clusters.

Key words: Fuzzy C-means Clustering; Cluster Validity Indice; Compactness; Overlapping; separation

1 引言

聚类有效性问题在数据挖掘的领域主要是用于处理数据集, 是一种重要的无监督学习技术, 没有类的先验知识可用. 因此如何确定最佳类数, 是研究聚类问题的一个基本问题^[1]. 如何建立合适的聚类指标来验证 FCM 算法最终聚类结果的优劣是聚类分析的核心问题之一. 已有许多学者对比进行研究.

针对经典的 FCM, 人们提出了许多有效性指标^[2-4], 这些指标按照计算过程中是否使用样本结构信息通常分为两类: 一类仅使用隶属度矩阵, 如划分系数 V_{pc} , 划分熵 V_{pe} 等; 另一类除使用隶属度矩阵外, 还使用数据集本身, 如 Xie-Beni 的 V_{XB} , Fukuya-

ma-Sugeno 的 V_{FS} , Kwon 的 V_K , Rezaee 的 V_{CWB} 等; 大部分的指标仅仅考虑聚类中心, 很少有考虑到度量多种几何结构的数据集. 特别在有重叠结构的数据集和含有交叠子类的数据集的研究中并不显著.

本文基于类内紧致度、类间分离度以及类间重叠度, 提出了一种新的聚类有效性指标. 它既考虑了数据的几何结构, 又考虑了不同类的形状和大小. 实验证明该指标能够有效地确定由模糊 C-均值 (FCM) 所得模糊划分的最优划分和最优聚类数.

2 相关工作

2.1 FCM 聚类算法的背景

FCM 算法通过最小化目标函数来实现样本的

最优划分. 假设将 n 个样本 $X = \{x_1, x_2, x_3, \cdots, x_n\}$ 聚成 c 类. 在满足约束条件:

$$\begin{cases} \sum_{i=1}^c u_{ij} = 1, 1 \leq j \leq n \\ 0 \leq u_{ij} \leq 1, 1 \leq j \leq n, 1 \leq i \leq c \\ 0 \leq \sum_{j=1}^n u_{ij} \leq n, 1 \leq i \leq c \end{cases} \quad (1)$$

求得目标函数:

$$J = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \|x_i - v_j\|^2 \quad (2)$$

最小的模糊划分.

式中, x_i 为样本 i 的特征向量, v_j 是第 j 类的聚类中心, u_{ij} 为样本 i 对于 j 类的隶属度, m 为模糊加权指数, 文献[5] 从聚类有效性的实验中得到 m 的取值为 $[1.5, 2.5]$, 本文研究取 $m = 2$. FCM 算法通过不断更新 v_j 和 u_{ij} 来优化目标函数. 根据拉格朗日条件极值理论可得到更新公式:

$$u_{ij} = 1 / \sum_{k=1}^c \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}} \quad (3)$$

$$v_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m} \quad (4)$$

FCM 算法的步骤如下:

- (1) 设定最大的迭代次数 T , 迭代的终止条件误差为 $\epsilon > 0$, 给定聚类数目 c , 模糊加权指数 $m = 2$, 且满足 $\sum_{i=1}^c u_{ij} = 1$, 令初始迭代次数 $k = 0$;
- (2) 利用式(3) 更新隶属度矩阵 U ;
- (3) 利用式(4) 更新聚类中心 V ;
- (4) 如果 $\|V^{(k)} - V^{(k+1)}\| < \epsilon$, 则停止, 否则令 $k = k + 1$, 返回(2), 直到达到最大迭代次数 T .

2.2 传统聚类有效性指标

聚类算法是一种无监督学习, 预先不知道给定数据集的结构, 因此无论采用哪一种聚类算法, 最终的聚类结果都需要进行有效性验证和质量评价, 这就是聚类有效性问题, 也是聚类过程中的一个重要的环节. 目前有许多聚类有效性指标, 一些比较经典的指标也存在一定的缺陷, 下面将阐述这些经典指标存在的优点和缺点. 如表 1 所示.

表 1 传统有效性指标的优缺点对比表

有效性指标	优 点	缺 点
划分系数 V_{PC} 划分熵 V_{PE}	函数简单, 运算量小.	与数据的集合结构缺乏直接的联系, 并且存在单调的趋势.
改进的划分系数 V_{MPC} 改进的划分熵 V_{MPE}	函数简单, 运算量小, 克服了单调的趋势.	仍然与数据的几何结构缺乏直接的联系.
Xie 和 Beni 的有效性指标	第一个真正意义结合了数据几何特征的模糊聚类有效性指标	该指标存在两个缺点: 当 $c \rightarrow n$ 时, V_{XB} 将单调递减趋于 0. 当 $m \rightarrow \infty$ 时, V_{XB} 将很不稳定. 在上述两种情况下, 指标将失去判断能力.
S. H. Kwon 的有效性指标	与数据集的几何结构联系紧密, 并克服了单调的趋势.	当时仍然不能够保证当 $c \rightarrow n$ 时, 指标值接近于 0 的趋势.
鲍正义的有效性指标 V_B	与数据集的几何结构联系紧密, 并克服了单调的趋势.	但是当 $m \rightarrow \infty$ 时, 该指标不是很稳定

3 新的有效性指标的提出

充分考虑数据集的几何结构和不同大小的集群, 利用类内紧致性和类间分离度的比值再结合类间重叠度定义了一种新的聚类有效性指标. 本文定义的指标充分考虑了数据集本身的信息和它的几何结构, 故能够较准确地确定由模糊 C -均值算法(FCM) 所得模糊划分的最优划分和最优聚类数, 具有较好的性能和较好的稳定性.

3.1 新的紧致度

用来衡量类内紧凑度, 可以采用类内加权平方

误差和表示, 公式如下:

$$\text{com}(c; U) = \left(\frac{c+1}{c-1} \right)^{1/2} \sum_{i=1}^c \frac{\sum_{j=1}^{N(i)} u_{ij}^m \|x_j - V_i\|^2}{\omega_i} \quad (5)$$

式中, $\omega_i = \sum_{j=1}^{N(i)} u_{ij}$ 是作为每个类的权值, 即针对每一个不同的类都有对应着一个不同的权重值. 其权值的意义代表是每一个不同的类对数据集的支持度. $N(i)$ 表示第 i 类中所包含数据样本的个数.

随着 c 增加时, $\left(\frac{c+1}{c-1} \right)^{1/2}$ 和权重 ω_i 均减小, 则

$1/w_i$ 通过给每个类加权从而抑制度量值的减少.

当有噪声点被作为单独一类,则有 $1/w_i = 1$, 此时该类的权重将会比其他的类更大,因此为使紧致度指标更具有鲁棒性,增加了 $\left(\frac{c+1}{c-1}\right)^{\frac{1}{2}}$ 进行调整.

3.2 新的分离度

类间数据的分离程度是衡量聚类结果有效性的主要标准和基本条件之一. 基于 FCM 算法,本文提出了模糊聚类有效性指标中的分离度的定义,其公式如下:

$$\text{sep}(c;U) = E_1 \times \max_{i \neq k} \|V_i - V_k\| \quad (6)$$

式中, E_1 代表类间总变差, $E_1 = \frac{1}{c} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m$ $\|V_i - V_0\|^2$, 这里的 $V_0 = \frac{1}{c} \sum_{i=1}^c V_i$ 是指各个聚类中心的平均值. $\text{Sep}(c;U)$ 的值越大,各聚类中心的距离就越大,类间分离越明确.

3.3 新的重叠度

由于上面新的分离度定义是根据类中心之间的距离衡量的,如图 1 所示,如果根据类中心的距离衡量分离度,则类 A 中心到类 C 中心的距离和类 B 中心到类 C 中心的距离是相等的,但是明显的类 B 和类 C 之间的分离更加明确,类 A 和类 C 之间有一些样本点是重叠的、模糊的. 所以说仅仅通过计算类中心之间的距离是无法正确衡量类间的分离程度,并且无法考虑到整个类的形状的. 根据这一点,本文提出了一个重叠度的衡量方法.

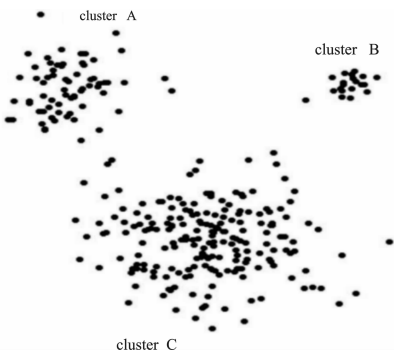


图 1 类 A,B,C 三个类的分布图

定义 1 第 j 个样本点相对于类 p 与类 q 的相似度,定义如下:

$$\text{overlap}_{pj}(c,U) = 1 - |u_{pj} - u_{qj}| \quad (7)$$

式中, u_{pj}, u_{qj} 分别代表第 j 个数据对第 p 类和第 q 类的隶属度. 当某个样本点同时

定义 2 第 p 个类与第 q 个类之间的相似度,即两个类的重叠度:

$$\text{overlap}_{pq}(c,U) = \frac{1}{n} \sum_{j=1}^n \text{overlap}_{pj}(c,U) * s(x_j) \quad (8)$$

这里 $s(x_j) = -\sum_{i=1}^c u_{ij} \log_a u_{ij}$ 是用于调节交叠部分数据点的权重指数,使得类间重叠度的计算对不同的聚类大小和形状鲁棒.

定义 3 通过定义 2,重叠度定义如下:

$$\text{overlap}(c;U) = \min_{p \neq q} (\text{overlap}_{pq}(c,U)) \quad (9)$$

$\text{Overlap}(c,U)$ 描述聚类结果基于数据点的所有类别之间重叠度最小的. 一个模糊集与它本身的重叠度是最大的,所以,在一个模糊划分中, $\text{overlap}(c;U)$ 的值越小,则表明在重叠部分的数据点的分离度就越小,重叠程度越低,这也表明类间的划分更清晰明确.

3.4 新的有效性指标

由于紧致度、分离度和重叠度有不同的标量,故需要归一化处理. 其结果可表示为

$$\text{com}^N(c;U) = \frac{\text{com}(c;U)}{\text{com}_{\max}} \quad (10)$$

$$\text{sep}^N(c;U) = \frac{\text{sep}(c;U)}{\text{sep}_{\max}} \quad (11)$$

$$\text{overlap}^N(c;U) = \frac{\text{overlap}(c;U)}{\text{overlap}_{\max}} \quad (12)$$

式中,

$$\text{com}_{\max} = \max_c \text{com}(c;U), \text{sep}_{\max} = \max_c \text{sep}(c;U)$$

$$\text{overlap}_{\max} = \max_c \text{overlap}(c;U)$$

由式(10)、(11)和(12)定义新的指标如下:

$$\text{CSO}(c;U) = \frac{\text{com}^N(c;U)}{\text{sep}^N(c;U)} + \text{overlap}^N(c;U) \quad (13)$$

利用紧致度和分离度来反映数据集的分布特征,再结合重叠度提出了一个新的聚类有效性指标 CSO.

新的有效性指标中的紧致度 com 反映的是类内总变差,分离度 sep 反映的是类间总变差,类内数据的总变差和类间距离可以反映数据集的分离特征,其中类内总变差越小,说明类内的数据相似性就越大,类间总变差越大,说明类间数据划分越明确. 这里的重叠度反映了聚类的形状,其值愈小,则表明划分越明确. 同时,它可以作为校正值使用. 因此, CSO 指标的最小值对应最佳聚类结果.

4 实验结果与分析

考虑到 FCM 算法对初始值(包括初始的聚类

中心矩阵和隶属度矩阵)值的敏感性,以下实验对每一组数据集,运行多次,选择所有结果的平均值作为最终的输出结果.

为了验证该指标的可行性与有效性,本文选 5 组数据集进行验证,分别来自人工合成与 UCI 的标准数据集库.其中有 4 组人工数据集和 1 组真实数据集.将新提出的指标与其他传统的模糊指标进行比较,验证该指标的有效性 with 准确性.同时,为了验证所提出指标的稳定性与鲁棒性,我们将 m 分别设置为 $m = 1.5, m = 2.0, m = 2.5$ 进行验证.

4.1 实验数据集

实验用到的人工数据集高斯正态随机分布的数据集.数据集 data_1_3 由 18 个二维数据组成,分为 3 类,每类有 60 个样本,三个类间相对分离明确.数据集 data_2_5 也是包含 150 个二维数据,分为 5 类,边上的 4 类都与中间的类有一定的交叠.数据集 data_3_3 包含了 150 个三维数据,其中一类单独,另外两类有重叠.数据集 data_4_5 是包含 250 个二维数据,分为 5 类,类与类之间有更多的交叠部分,产生了许多交叠子类.

真实数据集 data_IRIS 由 150 个 4 维数据组成,公认为 3 类 setosa, versicolor 和 virginica,每类包含 50 个样本,第一类与其他两类是完全分离的,后面两类之间存在一定的交叠.

4.2 结果

实验是利用传统的 FCM 算法对每一组数据集进行实验测试,其参数设置如下: $m = 2.0, \epsilon = 1.0 \times 10^{-6}, C_{\min} = 2$.对于 $C_{\max}$ 目前还没有确定其值,根据大量的实验证明,取 $C_{\max} = \sqrt{N}$,其中 N 是数据集

样本的个数,但是当 N 对应值越大时,则上述方法的工作量将十分巨大,因此,本文将 $C_{\max}$ 设定为 $2\ln N$,因为 $\sqrt{N}$ 的增长速度比 $\ln N$ 要快,这样就减少了计算的工作量.

为了测试新提出指标的运行效率,将其与 5 种常用指标进行实验对比,并对每一组数据集进行 30 次实验.下面用到的 5 种传统指标分别是 PC, PE, XB, MPC, Kwon. 聚类数的划分结果是根据聚类有效性指标的最优值进行选取的.表 2 给出了每一组数据集对应的最佳聚类数及各指标对应每组数据集的均值.其中 ON 代表了由聚类有效性指标得到的最佳聚类数,Mean 代表了各指标的均值.

由表 2 可以看出,利用 CSO 指标,传统的 FCM 聚类算法能够在 5 组数据集上 100%收敛得到正确聚类数.可以正确聚类出明确分离的数据集 1 以及类间有较多重叠的数据集 4;且对于明确分离的聚类中含有交叠子类的数据集(数据集 2,数据集 3, I-RIS 数据集),新提出的 CSO 指标均能够实现正确的聚类.

Kwon 指标仅对数据集 1 做出了正确的判断,而仅指出其他 4 组数据集相对明确的部分,说明了 Kwon 指标在考虑类间有交叠的情况是有缺陷的.

MPC 指标对数据集 1,数据集 2,数据集 4 做出了正确的判断,但未能有效地指出有交叠的数据集的最佳聚类数. XB 指标和 PE 指标都仅对数据集 1 做出了正确的分类,PC 在数据集 1 和数据集 2 得出了正确聚类数,这同时说明了,仅仅考虑聚类中心的距离和隶属度矩阵,而不考虑数据集的结构计算指标,未必能够得到正确的聚类数.

表 2 基于 FCM 聚类算法在各数据集上获得最佳聚类数(ON)及对应指标值(Mean)

数据集	最佳聚类数	PC		PE		XB		MPC		Kwon		CSO	
		ON	Mean	ON	Mean	ON	Mean	ON	Mean	ON	Mean	ON	Mean
数据集 1	3	3	0.606 5	2	0.517 1	3	0.040 8	3	0.409 8	3	10.463 8	3	0.711 3
数据集 2	5	2	0.469 8	2	0.598 2	5	0.043 7	5	0.282 4	4	11.399 4	5	0.856 5
数据集 3	3	2	0.899 5	2	0.181 0	2	0.047 6	2	0.799 1	2	7.412 5	3	0.874 5
数据集 4	5	2	0.745 0	2	0.405 9	4	0.089 7	5	0.653 5	4	23.340 4	5	0.849 2
IRIS	3	2	0.892 0	2	0.066 5	2	0.054 2	2	0.784 0	2	8.140 4	3	1.017 1

4.3 M 指数的对比

Pal 和 Bezek 讨论过权重指数对指标的影响,并且指出了 m 的取值范围[1.5, 2.5](引用文献[5])为了证明 CSO 指标的鲁棒性.下面将用三个典型的 m 的参数值进行验证.分别取 $m = 1.5, m = 2, m =$

2.5,图 2 展示的 CSO 指标分别在 $m = 1.5, m = 2, m = 2.5$ 下,不同聚类数所对应的指标值.由图 2 可以看出不论 m 取何值,新的聚类有效性指标 CSO 总是能够准确地找到每一个数据集的最佳聚类数.由这组实验可以证明新提出的聚类有效性指标具有鲁

棒性,而且新的指标有更好的稳定性。

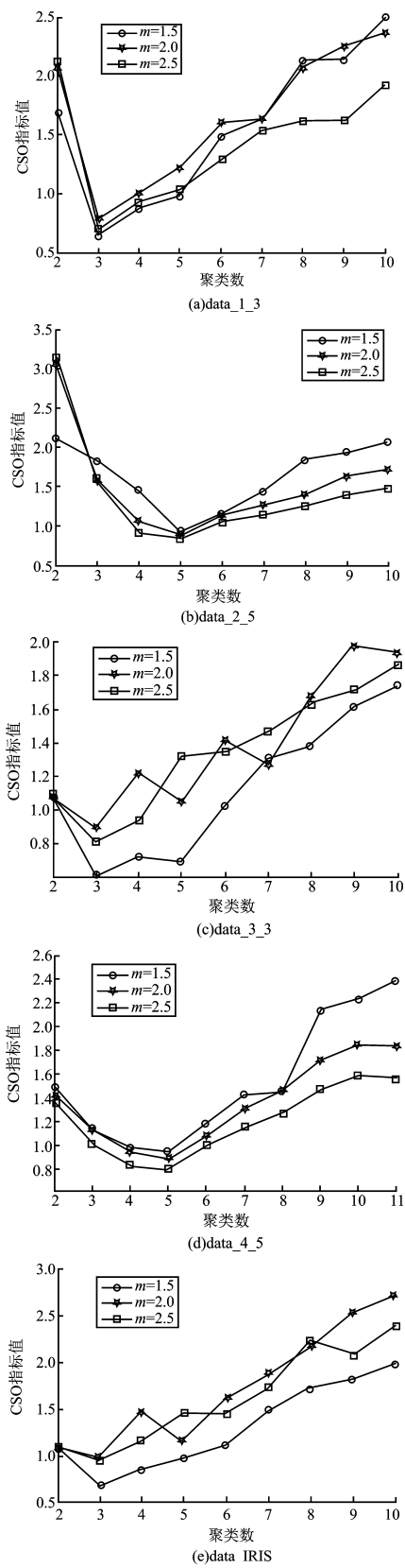


图 2 在不同 m 值下 CSO 指标值与聚类数的关系

为了测试聚类有效性指标与模糊加权指数 m

的关联程度,在 m 区间 $[1.5, 2.5]$ 中分别取 m 为 1.5,1.8,2.0,2.2,2.5 重复上述实验,表 3 给出了上述 6 个聚类有效性指标函数对所有实验数据集判定最佳聚类数的正确率对比。通过表 3 可以分析 PC 指标和 PE 指标性质类似,随着 m 指标的增加,正确率不断下降,当 $m = 2.5$ 时,判断力最低,与 m 指数相关性较小。

XB 和 Kwon 指标具有相似的断能力, $m \leq 2.2$ 较好, $m > 2$ 时,稍有下降,聚类有效性函数的性质与 m 相关性较小,MPC 指标的判断能力一直很稳定,但是其正确率相对比较低,与 m 的相关性比较小。CSO 指标与 m 的相关性较小,对所有的 m 的值其正确率保持不变,且保持较高的准确率。所以通过实验分析可知,CSO 的性能优良,能够较准确地判断出数据集的聚类数,性质比较稳定。

表 3 聚类有效性函数对实验数据集最佳聚类数判定正确率对比

m	PC	PE	XB	MPC	KWON	CSO
1.5	3/5	3/5	2/5	3/5	2/5	5/5
1.8	3/5	1/5	2/5	3/5	1/5	5/5
2.0	2/5	1/5	1/5	3/5	1/5	5/5
2.2	1/5	1/5	1/5	3/5	1/5	5/5
2.5	1/5	0/5	2/5	3/5	1/5	5/5
平均正确率	0.40	0.24	0.28	0.60	0.24	1.00

5 结束语

本文借助于 FCM 算法,提出了一个基于紧致度、重叠度和分离度的聚类有效性指标 CSO,可以有效确定最佳聚类数和最优化分。由实验数据分析知,不仅可以正确有效地判断类间无交叠分类明确的数据集的最佳聚类数,而且对于类间有交叠子类的情况也能作出正确的判断。而且也实现了聚类的无监督学习过程。因此可以得出新提出的聚类有效性指标具有良好的稳定性、有效性及鲁棒性。

在以后指标的研究过程中,可以向海量高维数据集和具有某些特殊形状数据集的分类展开更加深入的研究。

参考文献:

[1] Stocker A J, Siddle D R, Warrington E M. Comparison between the measurement and prediction of HF radio signals propagating along the mid-latitude trough[C]// 2003 9th International Conference on HF Radio Systems and Techniques. [s. l.]: IEEE, 2003: 145-149.

[2] J C Bezdek. Numerical taxonomy with fuzzy sets[J]. Journal of Mathematical Biology, 1974, 1(1): 57-71.

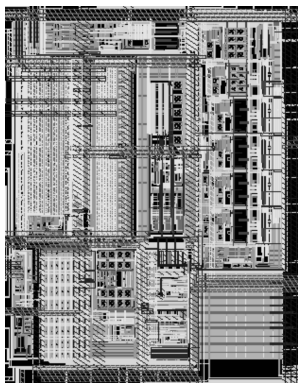
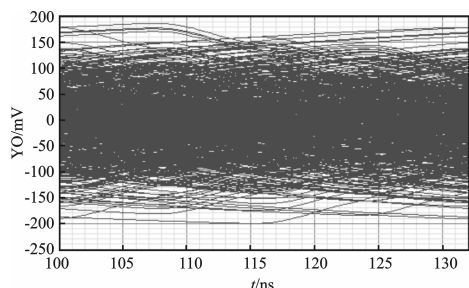
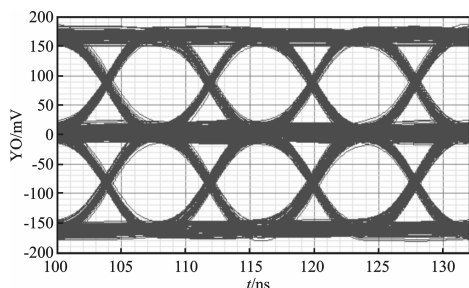


图 6 以太网收发器版图



(a)接收端均衡前眼图



(b)接收端均衡后眼图

图 7 100 m 传输线缆的收发眼图

5 结束语

本文设计了一款 100Base-TX 以太网收发器芯片,在接收端采用了连续时间线性均衡和判决式均衡的混

合均衡技术,并利用 LMS 算法实现了自适应均衡.文中提出一种可编程连续时间线性均衡电路结构,有效减少了功耗和面积.根据仿真结果,在 100 mUTP-5 线缆模型和 125 Mb/s 传输速率下,该以太网收发器的误码率小于 11,满足 IEEE802.3 标准规定.

参考文献:

- [1] Media Access Control (MAC) Parameters, Physical Layer, Medium Attachment Units, and Repeater for 100 Mb/s Operation, Type 100BASE-T, IEEE Standard 802.3u[s]. 1995.
- [2] J Everitt, J F Parker, P Hurst, et al. A CMOS Transceiver for 10MB/s and 100 MB/s Ethernet[J]. IEEE J of Solid State Circuits, 1998,33(11):2169-2177.
- [3] Jong-SangChoi, Moon-Song Hwang, Deog-kyoon-Jeong. A 0.18-um CMOS 3.5-Gb/s Continuous-Time Adaptive Cable Equalizer Using Enhanced Low-Frequency Gain Control Method[J]. IEEE J of Solid State Circuits, 2004,39(3):419-425.
- [4] E Lee, D Messerschmitt Digital Communication[M]. Boston, MA:Kluwer Academic, 1988.
- [5] Scott Huss, Mark Mullen. A DSP Based 10BaseT/100BaseTX Ethernet Transceiver in a 1.8V, 0.18um CMOS Technology[C]// IEEE Custom Integrated Circuits Conference. [s. l.]:IEEE,2001:135-138.
- [6] Mueller K H, Muller M. Timing Recovery in Digital Synchronous Data Recivers[J]. IEEE Transactions on Communications, 1976,24(5):516-531.

作者简介:

王晋男,(1973-),博士,高级工程师.研究方向为模拟集成电路设计. E-mail:wangjin0215@163.com

田泽男,(1965-),博士,研究员.研究方向为 SoC 设计,嵌入式系统设计,VLSI 设计.

(上接第 125 页)

- [3] Wang Weina, Zhang Yunjie. On fuzzy cluster validity indices[J]. Fuzzy sets and systems, 2007, 158(19): 2095-2117.
- [4] 王丽娜,马晓晓.一种改进的模糊聚类有效性指标[J].微电子学与计算机, 2014, 31(4): 68-74.
- [5] 杨燕,靳蕃,KAMEL Mohamed. 聚类有效性评价综述[J]. 计算机应用研究, 2008, 25(6): 1630-1632.
- [6] 普运伟,金炜东,朱明,等.核模糊 C 均值算法的聚类有效性研究[J]. 计算机科学,2007,34(2):207-210.

作者简介:

赵娜娜女,(1990-),硕士研究生.研究方向为数据挖掘. E-mail:znn909090@163.com.

钱雪忠男,(1967-),副教授,硕士研究生导师.研究方向为数据挖掘及其应用、数据库技术及其应用、WEB 服务与网络计算、计算机网络应用.

冯振华男,(1990-),硕士研究生.研究方向为数据挖掘.