文章编号: 1000-5277(2015)02-0011-06

半监督模糊聚类及其应用

杨昔阳, 李志伟
(泉州师范学院智能计算与信息处理福建省高等学校重点实验室, 福建 泉州 362000)

摘要: 提出了一种拓展的半监督模糊聚类模型, 给出求解这个模型的迭代公式. 这种半监督聚类能够合理、有效地利用部分已标识样本的类别信息对未标识样本产生影响, 从而提高半聚类算法的聚类效果. 其隶属度和聚类中心的迭代公式具有和 FCM 算法一样简洁的表示. 在黄瓜数据集上的聚类分析表明, 新提出的半监督聚类优于未改进的两种半监督算法、FCM 算法和线性判别方法.

关键词: 半监督算法; 模糊聚类; 叶片病害识别
中图分类号: O231 文献标志码: A

Semi-supervised Fuzzy Clustering and Its Application

YANG Xi-yang, LI Zhi-wei
(*Key Laboratory of Intelligent Computing and Information Processing of Fujian Province University, Quanzhou Normal University, Quanzhou* 362000, *China*)

Abstract: An extended form of semi-supervised fuzzy clustering algorithm is proposed, and its iterative solution is given. This new semi-supervised method uses class information of the labeled data effectively and reasonably to improve its classification ability. The iterative solutions of its membership degree and clustering centers have concise forms as those of FCM. Experiments on the cucumber data set show that the proposed algorithm is better than FCM, linear discriminant analysis and other two semi-supervised algorithms.

Key words: semi-supervised algorithm; fuzzy clustering; leave disease recognition

对于一个 m 维点集 $X=\{x_k\}$ 实施分类的算法通常称为聚类算法(Clustering), 比如 k-means, fuzzy c-means(FCM) 算法等. 若 x_k 的类别信息未知, 称这类算法为无监督算法. 如果还知道 x_k 的类别 f_k, 根据这些类别信息构造分类准则的算法往往称为判别分析(Clustering), 比如决策树算法, 支持向量机, 线性判别分析等. 若所有 x_k 的类别信息都已知, 称这类算法为监督算法. 所谓的半监督聚类, 就是结合一部分类别已知的样本点(称为标识样本) 和另一部分类别未知的样本点进行聚类的算法. 半监督聚类的示意图见图 1.

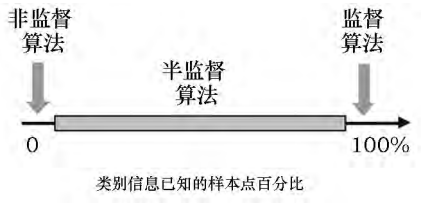

图 1 半监督算法示意图
Fig. 1 Illustration of semi-supervied algorithm

在现有的半监督聚类算法中, 一部分方法在原有半监督模型的基础上, 通过多种聚类算法的不断学习, 增加标识样本的比例[1], 从而提高半聚类的判别效果, 这类方法包括互训练(Co-training) 方法[2-3] 和自训练(Self-training) 方法[4].

另一些学者针对距离函数对半监督聚类进行改进, 在聚类过程使用加权的欧氏距离[5], 马氏距离[6], 基于图的距离[7], 或者基于度量(metric) 的距离[8], 使得各点之间的距离更能反映类别信息. 而采用核函数, 则可以使原本线性不可分的数据

收稿日期: 2014-10-13
基金项目: 福建省教育厅资助项目(JA12273, JK2013037); 泉州市科技计划资助项目(2012Z103)
通信作者: 李志伟(1965-), 男, 教授, 从事应用数学研究. wei2785801@qztc.edu.cn

集变成线性可分，从而提高聚类效果[9]. 但是这些变换涉及协方差矩阵的求逆运算，计算量比较大，此外部分参数也不容易设置.

Pedrycz 提出半监督 FCM 聚类算法是经典的 FCM 模型的推广[10]，文献[11]推广了这种模型，并在若干场合例举了一些成功的应用. 文献[12-13]针对某种改进的目标函数，推导出了在不同距离函数下的求解迭代公式. 本文结合几种模型的想法，设计了一种新的半监督模糊聚类的目标函数，使得在求解的迭代算法中，未标识样本到各个已标示样本的距离信息可以充分发挥作用. 本文推导出这个新模型的迭代公式，并将这种方法应用于叶片识别问题.

1 知识回顾

对于一个 m 维数据集 $X=\{x_k\}$，$x_k=(x_k^{(1)},x_k^{(2)},\cdots,x_k^{(m)})$，$k=1,\cdots,N$，如果 x_k 是标识样本，记其类别属性为 $f_k=(f_{1k},f_{2k},\cdots,f_{ck})$，并记 $f_{ik}=1$，如果 x_k 属于第 i 类. 否则记 $f_{ik}=0$. 在文献[1]中，Pedrycz 提出了如下半监督聚类算法(SFCM)：

$$Q=\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^2d_{ik}^2+\alpha\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik}-f_{ik}b_k)^2d_{ik}^2. \tag{1}$$

其中目标函数第一项和标准的 FCM 算法一致，它体现了数据集的类别结构. 第二项利用权重 $\alpha\in(0,+\infty)$ 给出了半监督学习的作用；标识向量 $b=[b_1,b_2,\cdots,b_N]^T$ 用来表示样本点 x_k 是否类别已知. f_{ik} 的定义同上，如果 x_k 是未标识样本，取 $f_k=(0,0,\cdots,0)$. u_{ik} 表示 x_k 属于类别 i 的隶属度，v_i 表示类别 i 的聚类中心，d_{ik} 表示 x_k 到 v_i 的距离.

Pedrycz 进一步说明了 SFCM 可以由迭代公式

$$u_{ik}=\frac{1}{1+\alpha}\left[\frac{1+\alpha(1-b_k\sum_{i=1}^{c}f_{ik})}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^2}+\alpha f_{ik}b_k\right] \tag{2}$$

和

$$v_i=\sum_{k=1}^{N}\frac{[u_{ik}^2+\alpha(u_{ik}-f_{ik}b_k)^2]x_k}{\sum_{i=1}^{c}[u_{ik}^2+\alpha(u_{ik}-f_{ik}b_k)^2]} \tag{3}$$

来求解.

在文献[12]中，笔者曾经对(1)式的 SFCM 模型进行改造，得到另一种半监督模型(改造的 SFCM)，

$$\begin{aligned}&\min Q=(1-\alpha)\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^2D_{ik}^2+\alpha\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik}-f_{ik})^2D_{ik}^2,\\&\text{s.t.}\sum_{i=1}^{c}u_{ik}=1,\end{aligned} \tag{4}$$

并且在文献[13]中将改造的 SFCM 模型中的欧氏距离推广成马氏距离，并说明它的迭代公式可以由

$$u_{ik}=(1-\alpha)u_{ik}^*+\alpha f_{ik} \tag{5}$$

和

$$v_i=\sum_{k=1}^{N}\frac{\Psi_{ik}x_k}{\sum_{i=1}^{c}\Psi_{ik}}\triangleq\sum_{k=1}^{N}\frac{[(1-\alpha)u_{ik}^2+\alpha(u_{ik}-f_{ik})^2]x_k}{\sum_{i=1}^{c}[(1-\alpha)u_{ik}^2+\alpha(u_{ik}-f_{ik})^2]} \tag{6}$$

确定. 其中 $u_{ik}^*=\left(\sum_{s=1}^{c}\frac{d_{ik}^2}{d_{sk}^2}\right)^{-1}$ 就是 FCM 中求解 u_{ik} 的迭代公式.

2　一种新的半监督模糊聚类

2.1　半监督模糊聚类的改进

注意到在改造的SFCM模型的第二项中,如果x_k是标识样本,那么u_{ik}将尽量靠近f_{ik},如果x_k是未标识的,那么u_{ik}将仅仅由(4)式的第一项确定. 也就是说,此时其他样本点x_j的类别信息f_{ij}对于u_{ik}的确定没有起到任何作用,即使x_k与x_j的关系非常密切. 迭代公式(5)也反映出了类似的结论. 事实上,一种更合理的做法是,如果类别未知的x_k与类别已知的x_j距离很近,那么u_{ik}也应该类似f_{ij}.

为了体现这种思想,修改目标函数如下:

$$\min Q = (1-\alpha)\sum_{j=1}^{l}\sum_{i=1}^{c}\sum_{k=1}^{N}\frac{u_{ik}^2 D_{ik}^2}{d_{kj}^2} + \alpha\sum_{j=1}^{l}\sum_{i=1}^{c}\sum_{k=1}^{N}\frac{(u_{ik}-f_{ij})^2 D_{ik}^2}{d_{kj}^2},$$
$$\text{s.t.}\sum_{i=1}^{c}u_{ik}=1, \tag{7}$$

其中$D_{ik}=\sqrt{(x_k^{(1)}-v_i^{(1)})^2+\cdots+(x_k^{(m)}-v_i^{(m)})^2}$表示$x_k$到$v_i$的欧氏距离. 类似地,$d_{kj}$表示$x_k$到$x_j$的欧氏距离. 规定$x_1, x_2, \cdots, x_l$为标识样本,$l$为标识样本的个数. 模型(7)与模型(4)的差别主要体现在第二项上,如果d_{kj}很小(x_k与x_j的关系密切),为了使得Q最小,$(u_{ik}-f_{ij})^2$也应该很小,这样类别未知的x_k的隶属度u_{ik}就越类似类别已知的x_j的隶属度f_{ij}. 类别已知的点的作用得到了加强. 为了达到与第二项对称的效果,式(7)对模型(4)的第一项也作了类似的修改,这样的修改将使得以下迭代公式的推导得到简化.

2.2　迭代求解公式

为了求解式(7)所表示的新模型,对于数据x_k,构造拉格朗日数函数

$$Q_k=(1-\alpha)\sum_{j=1}^{l}\sum_{i=1}^{c}\frac{u_{ik}^2 D_{ik}^2}{d_{kj}^2}+\alpha\sum_{j=1}^{l}\sum_{i=1}^{c}\frac{(u_{ik}-f_{ij})^2 D_{ik}^2}{d_{kj}^2}-\lambda\left(\sum_{i=1}^{c}u_{ik}-1\right). \tag{8}$$

令Q_k对u_{ik}的偏导为0,

$$\frac{\partial Q_k}{\partial u_{ik}}=(1-\alpha)\sum_{j=1}^{l}\frac{2u_{ik}D_{ik}^2}{d_{kj}^2}+\alpha\sum_{j=1}^{l}\frac{2(u_{ik}-f_{ij})D_{ik}^2}{d_{kj}^2}-\lambda=0, \tag{9}$$

并令$a_k=\sum_{j=1}^{l}\frac{1}{d_{kj}^2}$,$b=\sum_{j=1}^{l}\frac{f_{ij}}{d_{kj}^2}$,可得

$$u_{ik}=\frac{1}{2D_{ik}^2}\left(\sum_{j=1}^{l}\frac{1}{d_{kj}^2}\right)^{-1}\left(2\alpha D_{ik}^2\sum_{j=1}^{l}\frac{f_{ij}}{d_{kj}^2}+\lambda\right)=\frac{1}{2D_{ik}^2}(a_k)^{-1}(2\alpha D_{ik}^2 b_{ik}+\lambda). \tag{10}$$

将式(10)代入$\sum_{i=1}^{c}u_{ik}=1$,得

$$1=\sum_{i=1}^{c}u_{ik}=\sum_{i=1}^{c}\frac{1}{2D_{ik}^2}(a_k)^{-1}(2\alpha D_{ik}^2 b_{ik}+\lambda)=\alpha\sum_{i=1}^{c}\left(\frac{b_{ik}}{a_k}\right)+\frac{\lambda}{2a_k}\sum_{i=1}^{c}\frac{1}{D_{ik}^2}, \tag{11}$$

因此

$$\lambda=2\left(\sum_{i=1}^{c}\frac{1}{D_{ik}^2}\right)^{-1}\left(a_k-\alpha\sum_{i=1}^{c}b_{ik}\right). \tag{12}$$

将(12)代入(10)并注意到$\sum_{i=1}^{c}f_{ik}=1$,可得

$$u_{ik}=\frac{1}{2D_{ik}^2}(a_k)^{-1}\left(2\alpha D_{ik}^2 b_{ik}+2\left(\sum_{i=1}^{c}\frac{1}{D_{ik}^2}\right)^{-1}\left(a_k-\alpha\sum_{i=1}^{c}b_{ik}\right)\right)=$$

$$\alpha\left(\frac{1}{a_k}\sum_{j=1}^{l}\frac{f_{ij}}{d_{kj}^2}\right)+\frac{\frac{1}{D_{ik}^2}}{\left(\sum_{i=1}^{c}\frac{1}{D_{ik}^2}\right)}\left(1-\frac{\alpha}{a_k}\sum_{i=1}^{c}\sum_{j=1}^{l}\frac{f_{ij}}{d_{kj}^2}\right)=$$

$$\alpha\Big(\sum_{j=1}^{l}\frac{\frac{1}{d_{kj}^2}}{\sum_{j=1}^{l}\frac{1}{d_{kj}^2}}f_{ij}\Big)+\frac{\frac{1}{D_{ik}^2}}{\Big(\sum_{i=1}^{c}\frac{1}{D_{ik}^2}\Big)}\Big(1-\frac{\alpha}{a_k}\sum_{j=1}^{l}\frac{1}{d_{kj}^2}\Big)=$$

$$\alpha\Big(\sum_{j=1}^{l}\frac{\frac{1}{d_{kj}^2}}{\sum_{j=1}^{l}\frac{1}{d_{kj}^2}}f_{ij}\Big)+(1-\alpha)\frac{\frac{1}{D_{ik}^2}}{\Big(\sum_{i=1}^{c}\frac{1}{D_{kj}^2}\Big)}. \tag{13}$$

仍然记 $u_{ik}^{*}=\Big(\sum_{s=1}^{c}\frac{D_{ik}^2}{D_{sk}^2}\Big)^{-1}$,并记权重 $\omega_{kj}=\frac{1}{d_{kj}^2}\Big/\sum_{j=1}^{l}\frac{1}{d_{kj}^2}$,那么(13)式可写为

$$u_{ik}=\alpha\sum_{j=1}^{l}\omega_{kj}f_{ij}+(1-\alpha)u_{ik}^{*}. \tag{14}$$

(14)式表明,无论标识样本还是非标识样本,其隶属度迭代算法都可以写成 u_{ik}^{*} 与所有标识样本的 f_{ik} 的加权平均. 这个结论与(5)式在形式上类似,可以合理地体现数据分类结构,但更加突出了标识样本对非标识样本的隶属度的影响.

另一方面,为了得到聚类中心 $v_i=(v_i^{(1)},\cdots,v_i^{(m)})$ 的迭代公式,对(7)式中 v_i 的第 t 个分量 $v_i^{(t)}$ 求偏导,并令之为 0,

$$0=\frac{\partial Q}{\partial v_i^{(t)}}=(1-\alpha)\sum_{j=1}^{l}\sum_{k=1}^{N}\frac{u_{ik}^2 2(v_i^{(t)}-x_k^{(t)})}{d_{kj}^2}+\alpha\sum_{j=1}^{l}\sum_{k=1}^{N}\frac{(u_{ik}-f_{ij})^2 2(v_i^{(t)}-x_k^{(t)})}{d_{kj}^2},$$

由此可得

$$v_i^{(t)}\sum_{k=1}^{N}\sum_{j=1}^{l}\Big((1-\alpha)\frac{u_{ik}^2}{d_{kj}^2}+\alpha\frac{(u_{ik}-f_{ij})^2}{d_{kj}^2}\Big)=\sum_{k=1}^{N}\Big(\sum_{j=1}^{l}\Big((1-\alpha)\frac{u_{ik}^2}{d_{kj}^2}+\alpha\frac{(u_{ik}-f_{ij})^2}{d_{kj}^2}\Big)x_k^{(t)}\Big).$$

若令

$$\Psi_{ik}=\sum_{j=1}^{l}\Big((1-\alpha)\frac{u_{ik}^2}{d_{kj}^2}+\alpha\frac{(u_{ik}-f_{ij})^2}{d_{kj}^2}\Big), \tag{15}$$

则

$$v_i=\sum_{k=1}^{N}\frac{\Psi_{ik}}{\sum_{i=1}^{c}\Psi_{ik}}x_k, \tag{16}$$

它与(6)式类似,聚类中心的迭代可以表示成所有数据点的加权平均.

3 半监督模糊聚类的应用

3.1 黄瓜病害叶片的数据集

采集了 92 张不同程度患有霜霉病的黄瓜叶片,由人工判定这些叶片的患病程度,其中轻度患病(一类)叶片 36 张,中度患病(二类)叶片 33 张,重度患病(三类)叶片 23 张. 为了克服光照等无关因素的影响,采用 HIS(色调、亮度、饱和度)色彩系统来度量所有叶片,其中色调 $H\in\{0,1,\cdots,255\}$. 计算 H 的各种统计特征 $X_1,X_2,\cdots,X_7$ 作为叶片的分类指标,分别为色调均值、色调方差、色调偏度、色调峰度、色调能量、色调熵以及叶片病变面积比(通过分析,可以认为 $H\leqslant 52$ 的区域为患病区域). 对于某张叶片,这些特征可以通过下式计算而得[13]:

$$X_1=\bar{h}=\sum_{i=1}^{256}h_iP(h_i),\quad X_2=\sum_{i=1}^{256}(h_i-\bar{h})^2P(h_i),\quad X_3=\frac{1}{\sigma^3}\sum_{i=1}^{256}(h_i-\bar{h})^3P(h_i),$$

$$X_4=\frac{1}{\sigma^4}\sum_{i=1}^{256}(h_i-\bar{h})^4P(h_i),\quad X_5=\sum_{i=1}^{256}P^2(h_i),\quad X_6=-\sum_{i=1}^{256}P(h_i)\log_2P(h_i),\quad X_7=\sum_{i=1}^{52}P(h_i),$$

$P(h_i)$ 表示此张叶片中 $H=h_i$ 的频率. 中心化之后的特征数据集以及人工分类结果如表 1 所示. 简洁起

见,仍然记标准化之后的特征为 $X_1, X_2, \cdots, X_7$.

3.2　各种聚类方法的比较

设定各种不同的标识样本的百分比，根据这些百分比随机从数据集中选取某些样本作为标识样本，其余样本作为未标识样本. 采用(1)式表示的 SFCM 模型,文献[12]给出的改造的 SFCM 模型(如(4)式所示)和本文提出的新模型(如(7)式所示)对黄瓜患病叶片数据集进行半监督聚类(在这 3 个模型中,均取 $\alpha = 0.5$)；为加强比较，也采用监督聚类 FCM 和无监督的线性判别算法进行叶片识别. 比较结果列于表 2. 从表 2 可知，本文提出的新方法优于模型(4)和模型(1),而模型(4)和模型(1)的判别效果类似. 此外，这 3 种半监督聚类的判别效果均优于监督聚类的线性判别和无监督聚类的 FCM 算法.

表 1　叶片的特征数据集及其分类结果

Tab. 1　Leave features data set and their classification labels

序号	X_1	X_2	X_3	X_4	X_5	X_6	X_7	类别
1	0.36	−0.92	−0.78	−0.14	1.13	−1.09	−0.70	1
2	1.20	−0.98	−0.73	−0.09	1.26	−1.17	−1.00	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
37	0.78	−0.18	−0.35	−0.21	0.07	−0.08	−0.50	2
38	0.94	−0.29	−0.58	−0.32	0.21	−0.20	−0.59	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
70	−0.60	0.96	0.03	−0.49	−0.82	0.89	0.68	3
71	−0.83	1.23	0.37	−0.50	−1.18	1.29	1.18	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表 2　一致率结果比较

Tab. 2　Result comparison of concordance rate　　%

类别已知样本百分比	新模型	改造的 SFCM 模型	SFCM 模型	线性判别	FCM
10	72.83	70.65	68.48	42.39	67.39
20	81.52	73.91	72.83	75.00	67.39
30	82.61	77.17	77.17	75.00	67.39
40	86.96	81.52	82.61	73.91	67.39
50	88.04	85.87	85.87	78.26	67.39
60	94.57	91.30	91.30	76.09	67.39
70	95.65	93.48	93.48	79.35	67.39
80	98.91	96.74	96.74	80.43	67.39
90	98.91	96.74	96.74	80.43	67.39
100	100.00	100.00	100.00	81.52	67.39

4　小结

半监督模糊聚类算法是数据挖掘领域的研究热点之一，本文对一类半监督模糊聚类算法进行扩展，增强了标识数据在聚类过程中的作用，提出了一种新的半监督模糊聚类模型. 它的迭代公式形式简洁，易于解释. 在黄瓜病害叶片的程度识别问题中，新方法的识别结果优于其他几类半监督算法、FCM 算法和线性判别算法.

本文所提出的模型采用的是欧氏距离函数，如果采用马氏距离或者核函数，有可能进一步提高聚类效果，将在以后的研究中考虑距离的扩展问题.

参考文献:

[1] Stefan Fauβer, Friedhelm Schwenker. Semi-supervised clustering of large data sets with kernel methods [J]. Pattern Recognition Letters, 2014, 37: 78 – 84.

[2] Zhang Yihao, Wen Junhao, Wang Xibin, et al. Semi-supervised learning combining co-training with active learning [J]. Expert Systems with Applications, 2014, 41 (5): 2372 – 2378.

[3] 施伟民, 杨昔阳. 一种结合半监督算法和 SVM 的聚类方法 [J]. 泉州师范学院学报, 2013, 31 (6): 49 – 52.

[4] Gan Haitao, Sang Nong, Huang Rui, et al. Using clustering analysis to improve semi-supervised classification [J]. Neurocomputing, 2013, 101 (1): 290 – 298.

[5] 计华, 张化祥, 孙晓燕. 基于最近邻原则的半监督聚类算法 [J]. 计算机工程与设计, 2011, 32 (7): 2455 – 2458.

[6] Yin Xuesong, Shu Ting, Huang Qi. Semi-supervised fuzzy clustering with metric learning and entropy regularization [J]. Knowledge-Based Systems, 2012, 35: 304 – 311.

[7] 兰远东, 高蕾. 基于图的半监督学习的距离度量改进 [J]. 智能计算机与应用, 2014, 4 (2): 32 – 35.

[8] Yin Xuesong, Chen Songcan, Hu Enliang, et al. Semi-supervised clustering with metric learning: an adaptive kernel method [J]. Pattern Recognition, 2010, 43: 1320 – 1333.

[9] Mahdieh Soleymani Baghshah, Saeed Bagheri Shouraki. Kernel-based metric learning for semi-supervised clustering [J]. Neurocomputing, 2010, 73: 1352 – 1361.

[10] Witold Pedrycz. 基于知识的聚类——从数据到信息粒 [M]. 于福生, 译. 北京: 北京师范大学出版社, 2008: 79 – 87.

[11] 李春芳, 庞雅静, 钱丽璞, 等. 半监督 FCM 聚类算法目标函数研究 [J]. 计算机工程与应用, 2009, 45 (14): 128 – 132.

[12] 施伟民, 杨昔阳, 李志伟. 基于半监督模糊聚类的黄瓜霜霉病受害程度识别研究 [J]. 福建师范大学学报: 自然科学版, 2012, 28 (1): 33 – 37.

[13] 施伟民, 杨昔阳, 李志伟. 一种基于马氏距离的半监督模糊聚类方法及其应用 [J]. 厦门大学学报: 自然科学版, 2012, 51 (3): 311 – 315.

(责任编辑: 林 敏)