计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(10):331−335 [doi: 10.15888/j.cnki.csa.008105]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

谱聚类欠取样下自编码网络不平衡数据挖掘①

王舒梵, 严　涛, 姜新盈

(上海工程技术大学 数理与统计学院, 上海 201620)
通讯作者: 王舒梵, E-mail: 1814423428@qq.com

摘　要: 不平衡数据集的应用领域日益广泛, 需求也越来越高, 为提升整体数据集的分类准确率, 以谱聚类欠取样为前提条件, 构建一种自编码网络不平衡数据挖掘方法. 把聚类问题转换成无向图多路径划分问题, 通过无向图与标准化处理完成谱聚类, 经过有选择地欠取样处理多数类数据集, 获取分类边界偏移量, 利用学习过程是无监督学习的自编码网络, 升、降维数据, 获取各维度隐藏特征, 实现各层面的数据高效表示学习, 根据最大均值差异与预设阈值的对比结果, 调整自编码网络, 基于得到的分类界面, 完成不平衡数据挖掘. 选用具有不同实际应用背景的UCI 数据集, 从中抽取 10 组数据作为测试集, 经谱聚类欠取样处理与模拟实验, 发现所提方法大幅提升少数类分类精度与整体挖掘性能, 具有较好的适用性与可行性.

关键词: 谱聚类; 欠取样; 自编码网络; 不平衡数据; 分类边界; 聚类中心

引用格式: 王舒梵,严涛,姜新盈.谱聚类欠取样下自编码网络不平衡数据挖掘.计算机系统应用,2021,30(10):331–335. http://www.c-s-a.org.cn/1003-3254/8105.html

Unbalanced Data Mining of Self-Encoding Network under Spectral Clustering Undersampling

WANG Shu-Fan, YAN Tao, JIANG Xin-Ying

(School of Mathematics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: The application fields of unbalanced data sets are becoming increasingly extensive, and the demand for them is getting higher. Taking the spectral clustering undersampling as a prerequisite, this study develops an unbalanced data mining method based on a self-encoding network to improve the classification accuracy of the overall data set. The clustering problem is converted into the multi-path partition problem of an undirected graph, and the spectral clustering is completed depending on the undirected graph and standardized processing. The majority of data sets are processed through selective undersampling to yield the classification boundary offset. The learning process is a self-encoding network of unsupervised learning, based on which the dimensionality of data is increased or reduced so that hidden features of each dimension can be obtained and the efficient representation and learning of data are realized at all levels. The self-encoding network is adjusted according to the comparison between the maximum mean difference and the preset threshold. The unbalanced data mining is then completed with the obtained classification interface. UCI data sets with different practical application backgrounds are selected, from which 10 sets of data are extracted as test sets. After spectral clustering undersampling, the simulation experiments demonstrate that the proposed method greatly improves the classification accuracy of the minority and overall mining performance, which shows good applicability and feasibility.

Key words: spectral clustering; undersampling; self-encoding network; unbalanced data; classification boundary; clustering center

① 收稿时间: 2020-12-24; 修改时间: 2021-01-25; 采用时间: 2021-02-03

1 引言

信息化时代加快了数据量的增长速度, 各行各业的数据总数日渐庞大, 为在海量数据资源中挖掘出隐藏规律, 聚类算法应运而生且重要性日益显著. 在同一数据集中, 若某类别样本个数远超出余下类别样本个数, 则该数据集叫做不平衡数据[1]. 此类数据多用于故障诊断、目标检测等实际应用中, 但当前算法大部分都是以数据集均衡分布为前提的, 在处理不平衡数据时极易偏向多数类, 产生错分情况, 降低分类准度, 所以, 研究不平衡数据集的数据挖掘方法具有重要的实践意义.

向鸿鑫等人[2] 通过总结常用的不平衡数据预处理方法与挖掘算法, 从多维度梳理策略性能, 分析各应用领域的不平衡问题与解决方案后, 实现不平衡数据挖掘方法综述; 蔡莉等人[3] 构建出一种时空特征位置数据融合模型, 通过数据与算法层面, 解决不平衡数据的挖掘问题, 利用架构的综合评价指标, 反映聚类质量, 融合不平衡数据后, 完成热点区域挖掘; 文献 [4] 中许统德等人设计的多层级联式少数类聚类高精度数据挖掘算法中, 在聚类欠采样的前提下, 聚类多数类样本, 获取与少数类相同数量的质心, 架构新的平衡训练集, 采用合成少数类过采样技术 (Synthetic Minority Over-sampling TEchnique, SMOTE) 过采样, 级联 K-means 聚类与 C4.5 决策树算法, 改善分类决策边界.

鉴于上述文献方法在融合不平衡数据样本时存在一定的盲目性, 故基于谱聚类欠取样, 采用自编码网络来构架一种不平衡数据挖掘方法. 通过谱聚类方法聚类多数类数据, 在更改数据空间结构的基础上, 有选择地欠取样处理了多数类数据集, 通过选取代表性数据作为训练数据, 经过数据筛选, 使分类边界适当偏移, 提升划分准确率; 利用自编码器升、降维数据, 实现初始数据重构; 引入网络调整操作, 增加了目标领域网络的学习空间, 使其与目标领域样本特征表示更匹配.

2 谱聚类欠取样分类

谱聚类就是按照谱图理论[5] 完成数据分类, 将聚类问题转换成无向图多路径划分问题.

用 $G=(V, E)$ 界定无向图, 其顶点集合为 $V=\{v_1, v_2,\cdots,v_n\}$, 数据样本有 n 个, 若无向图 G 为一个加权图, 则顶点 v_i 与 v_j 的连边为 w_{ij}, 且 $w_{ij}>0$, 当 $w_{ij}=V$ 时, 说明两顶点之间不存在连接的边. 假设无向图加权连边矩阵的界定式是 $W'=\left(w_{ij}\right)$, 且加权连接边满足 $w_{ij}=w_{ji}$, 则顶点 $v_i\subset V$ 的度界定表达式为:

$$d'_i=\sum_{j=1}^{n}w_{ij} \tag{1}$$

采用下列公式界定无向图 G 的度矩阵:

$$D=\mathrm{diag}\left(d'_1,d'_2,\cdots,d'_n\right) \tag{2}$$

通过标准化处理无向图 G, 推导出下列拉普拉斯[6] 矩阵表达式:

$$L_{\mathrm{sym}}=I-D^{-\frac{1}{2}}W'D^{-\frac{1}{2}} \tag{3}$$

谱聚类算法流程具体描述如下:

(1) 输入聚类个数 k 与相似矩阵 $\sigma\in R^{n\times n}$;

(2) 利用相似矩阵 $\sigma\in R^{n\times n}$ 完成无向图 $G=(V,E)$ 的构建, 其加权连接矩阵用 W' 表示;

(3) 经过标准化处理建立拉普拉斯矩阵;

(4) 对拉普拉斯矩阵的前 k 个特征向量进行求解, 得到 $\mu_1,\mu_2,\cdots,\mu_k$;

(5) 将所得 $\mu_1,\mu_2,\cdots,\mu_k$ 特征向量当做列, 构建矩阵 $U\in R^{n\times k}$;

(6) 标准化处理矩阵 U 各行, 令特征向量 μ_{ij} 满足等式 $t_{ij}=\mu_{ij}\left(\sum_k\mu_{ij}^2\right)^{\frac{1}{2}}$, 完成矩阵 $T\in R^{n\times k}$ 架构;

(7) 利用 $t_i\in R^k$ 构成矩阵 T 的第 i 行向量, 其中, i=1, 2,⋯,n;

(8) 通过 K-means 算法[7] 聚类 $(t_i)_{i=1,2,\cdots,n}$ 所含数据, 组建为一个子集, 用 $C_1,C_2,\cdots,C_k$ 表示;

(9) 得到最终聚类结果 $A_1,\cdots,A_k$, 其中, $A_i=\left\{j\middle|t_j\in C_j\right\}$.

在不平衡数据挖掘过程中, 多数类数据通常会携带多个冗余数据信息与噪声数据, 导致分类边界偏移至少数类数据方向, 加大错分概率, 若想解决该问题, 就要对多数类数据实施相应处理, 即欠取样处理, 使分类边界偏移至多数类数据方向. 传统欠取样处理方法多为去除与边界距离较远的数据点, 或随机去除多数类数据, 这种不考虑数据信息的处理手段虽然均衡了不同类数据集, 但分类界限调整得并不够理想, 因此, 采用谱聚类方法聚类多数类数据, 在更改数据空间结构的基础上, 有选择地欠取样处理了多数类数据集, 通过选取代表性数据作为训练数据, 经过数据筛选, 获取分类边界偏移量.

3 基于自编码网络的不平衡数据挖掘

通过训练令网络输入与输出相等, 完成数据隐藏特征学习的一种神经网络模型就是自编码器 (Auto-Encoder, AE)[8], 作为深度学习网络的一种主要结构, 自编码网络在深度神经网络预训练中被广泛应用. 该网络即便不用带标签数据样本, 也能够达成训练目的, 也就是说, 其学习过程属于无监督学习. 自编码网络中的编码阶段是输入数据学习至高效表示特征, 解码阶段是以习得的隐藏特征为依据, 实现初始数据重构. 自编码器经过升、降维数据, 把提取出来的数据特征转换为适用、高效的隐藏特征后, 输送至有监督学习模型内, 即可实现挖掘目标. 图 1 所示为自编码器的基本框架形式, 由输入层、输出层以及隐含层组成, 近似于一个 3 层神经网络[9].

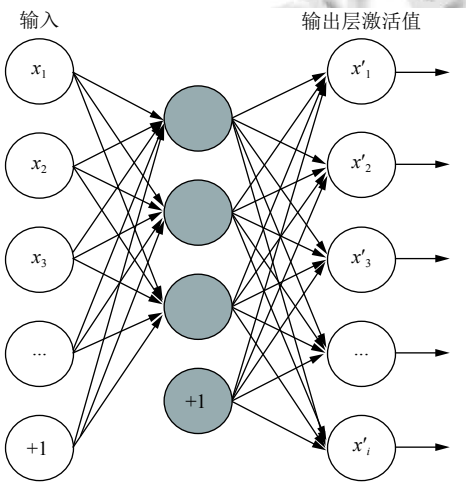

图 1 自编码器框架示意图

假设$(x_1, x_2, \cdots, x_i)$是一个输入样本, Sigmoid 激活函数[10] 用 S 表示, 输入层与隐含层间、隐含层与输出层间的权值分别为W_1与W_2, 则自编码器前向传播表达式如式 (4) 和式 (5) 所示.

$$d = S(W_1 x + b_1) \tag{4}$$

$$h_{w,b}(x) = S(W_2 d + b_2) \tag{5}$$

式中, 偏置项为 b_i, 输出层激活值为$h_{w,b}(x)$.

由于自编码器的训练标准期望是输入与输出相等, 所以, 采用下列表达式描述自编码器的最终学习结果:

$$h_{w,b}(x) \approx x \tag{6}$$

根据各隐藏单元数, 获取各维度隐藏特征, 升、降维处理初始数据, 通过堆叠多个自编码器, 结合约束条件, 实现各层面的数据高效表示学习.

利用无监督学习与有监督学习, 在谱聚类欠取样条件下架构用于挖掘不平衡数据的自编码网络. 因为无标签样本数据在源领域与目标领域中均可轻易取得, 因此, 当最大均值差异[11] 比预设阈值低时, 直接跳过网络调整阶段, 无监督训练目标领域数据; 反之, 当最大均值差异比预设阈值高时, 按照图 2 中所示的自编码网络形式进行调整, 并完成随机初始化. 网络调整操作增加了目标领域网络的学习空间, 使其与目标领域样本特征表示更匹配.

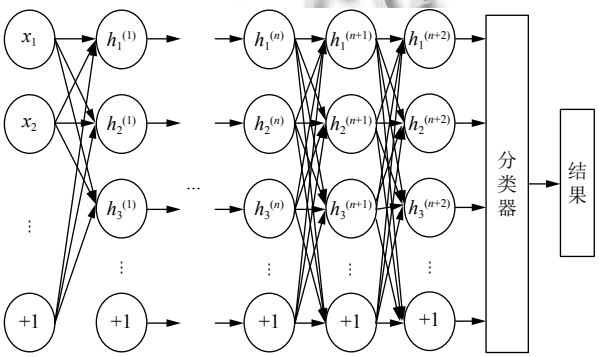

图 2 自编码网络结构示意图

在自编码网络中输入谱聚类欠取样处理的不平衡数据集合, 依照以下流程实现数据挖掘:

(1) 设定$MajorN = mMinorN$为待训练数据, 其中, 多数类数据是$MajorN$, 少数类数据是$MinorN$, 两类个数比例是 m.

(2) 若多数类数据样本有 n 个, 则高斯核[12] 相似矩阵表达式如下:

$$\sigma \in R^{n \times n} \tag{7}$$

(3) 利用谱聚类方法处理 n 个多数类数据样本, 获得聚类结果$A_1, A_2, \cdots, A_k$, $MajorN = k$.

(4) 根据各聚类结果以及聚类中心与少数类数据点的间距大小, 选取代表性数据点, 使分类界面偏移至多数类样本, 并最大程度删除多数类数据点的边界点. 各聚类结果中, 数据点选用数量随着多数类样本个数的增加而增多, 随着聚类中心与少数类数据点间距的增加而上升, 基于此, 采用下列选取公式, 筛选出有效数据点.

$$KDist_{il} = K(x_i, x_i) + K(x_l, x_l) - 2K(x_i, x_l) \tag{8}$$

$$IDist_i = \frac{1}{Ksize_i \times MinorN} \sum_{i=1}^{Ksize_i} \sum_{l=1}^{MinorN} KDist_{il} \tag{9}$$

$$Radio_i = \frac{Ksize_i}{\sum_{i=1}^{k} Ksize_i} \times \frac{IDist_i}{\sum_{i=1}^{k} IDist_i} \tag{10}$$

$$SSize_{MA}^{i} = MajorN \times \frac{Radio_i}{\sum_{i=1}^{k} Radio_i} \tag{11}$$

式里, $Ksize_i$表示聚类 A_i 的大小, $IDist_i$表示聚类 A_i 与少数类数据的间距均值, $SSize_{MA}^{i}$表示各聚类所选样本个数.

(5) 采用各聚类结果里与少数类数据点间距均值最小的前$SSize_{MA}^{i}$个数据点, 成为多数类样本中的训练子集数据.

(6) 训练上述多数类代表数据点与所有少数类数据, 将处理完的数据输入自编码网络, 在相同数据空间中, 实现其与谱聚类算法的无缝连接, 选取相同参数, 令网络和参数与谱聚类相似矩阵保持一致.

(7) 根据上述训练得出的分类界面, 完成不平衡数据挖掘.

4 不平衡数据挖掘模拟分析

4.1 数据集选取

选用具有不同实际应用背景的 UCI 数据集[13], 从中抽取 sonar、breast-w、vehicle、artificial、pendigits、letter、page-blocks、car、seg1、yeast5 等 10 组数据作为测试集 (如表 1 所示), 验证挖掘策略的有效性. 当数据包含多个类别时, 设定任意一类为少数类, 多数类则为其余各类别的合并结果, 所有不平衡数据集均经过谱聚类欠取样处理.

表 1 UCI 数据集具体信息统计表

数据集名称	样本个数	少数类数量	多数类数量	不平衡度
sonar	287	106	133	1.52
breast-w	725	291	515	2.65
vehicle	961	235	687	4.16
pendigits	11 538	1173	10 482	10.98
artificial	6284	824	5685	7.89
letter	28 238	806	20 573	29.54
page-blocks	6516	227	5565	51.33
seg1	23	474	2173	19.08
car	1832	78	1721	28.47
yeast5	11	49	1518	12.43

将表 1 中的不平衡度划分成下列等级表, 如表 2 所示.

sonar 与 breast-w 两个低度不平衡等级数据集的选取原因是验证挖掘方法在处理一般数据集时的有效性.

表 2 不平衡度等级表

不平衡度	等级
[1, 4)	低度不平衡
[4, 10)	中度不平衡
[10, +∞)	高度不平衡

4.2 性能评估指标

针对不平衡数据集, 采用合理的查全率*Recall*、查准率*Precision*、综合 *F-measure*、*AUC* (Area Under ROC Curve, ROC 曲线下方图面积) 值、*G-means* 等类别不平衡评估指标, 使少数类挖掘情况得以充分反映, 各指标均以表 3 中所示的混淆矩阵为依据完成创建.

表 3 类别混淆矩阵表

	预测正类	预测负类
实际正类	*TP*	*FN*
实际负类	*FP*	*TN*

其中, 具有描述少数类分类性能的指标为 *F-measure*, 是查全率与查准率的调和均值; *AUC* 作为不同判决阈值对应的分类性能反映指标, 性能随数值的增加而提升. 各评估指标表达式分别如下所示:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F\text{-}measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{14}$$

$$AUC = \frac{TP_{\text{rate}} + TN_{\text{rate}}}{2} \tag{15}$$

$$G\text{-}means = \sqrt{TP_{\text{rate}} \times N_{\text{rate}}} \tag{16}$$

式中, TP_{rate}表示多类样本准确率, TN_{rate}表示少类样本准确率.

4.3 不平衡数据挖掘效果

分别模拟文献 [2–4] 方法以及本文方法在挖掘 10 组不平衡数据集时的效果, 通过对比不同方法的评估指标数据, 验证方法的适用性与可行性. 对比结果如表 4–表 6 所示.

结合上列各表可以看出, 各方法少数类评估指标均随着不平衡度的增加而略有下降; 少数类样本数据

个数总量相对较少, 导致文献 [2–4] 方法的 *F-measure* 值整体偏低; 造成文献方法 *AUC* 值与 *G-means* 指标较低的原因是未考虑样本属性间的相关性, 忽略了监督判别性的类别标签信息; 而本文方法因引用了自编码网络, 根据各隐藏单元数, 获取各维度隐藏特征, 实现了各层面的数据高效表示学习, 通过对比最大均值差异比预设阈值, 完成了网络调整与随机初始化, 利用 K-means 算法与自编码网络, 充分结合了无监督学习与有监督学习形式, 因此, 取得了较为理想的少数类样本分类效果.

表 4 各方法 *F-measure* 实验数据结果比对表

数据集名称	文献[2]	文献[3]	文献[4]	本文方法
sonar	0.758	0.765	0.807	0.892
breast-w	0.716	0.753	0.797	0.854
vehicle	0.652	0.667	0.705	0.815
artificial	0.641	0.657	0.689	0.798
pendigits	0.635	0.643	0.687	0.795
yeast5	0.632	0.638	0.684	0.765
seg1	0.628	0.629	0.678	0.765
car	0.622	0.626	0.673	0.759
letter	0.617	0.616	0.665	0.751
page-blocks	0.613	0.615	0.648	0.732
均值	0.651	0.661	0.703	0.793

表 5 各方法 *AUC* 值实验数据结果比对表

数据集名称	文献[2]	文献[3]	文献[4]	本文方法
sonar	0.922	0.893	0.919	0.971
breast-w	0.905	0.868	0.892	0.967
vehicle	0.738	0.782	0.767	0.904
artificial	0.752	0.776	0.76	0.895
pendigits	0.749	0.772	0.754	0.892
yeast5	0.743	0.767	0.749	0.887
seg1	0.737	0.761	0.744	0.884
car	0.731	0.755	0.738	0.876
letter	0.726	0.743	0.732	0.87
page-blocks	0.712	0.736	0.721	0.862
均值	0.772	0.785	0.778	0.901

表 6 各方法 *G-means* 实验数据结果比对表

数据集名称	文献[2]	文献[3]	文献[4]	本文方法
sonar	0.785	0.812	0.792	0.898
breast-w	0.78	0.806	0.788	0.893
vehicle	0.776	0.8	0.781	0.882
artificial	0.771	0.792	0.777	0.876
pendigits	0.765	0.788	0.77	0.871
yeast5	0.759	0.783	0.765	0.868
seg1	0.752	0.776	0.754	0.863
car	0.747	0.774	0.748	0.854
letter	0.743	0.767	0.743	0.846
page-blocks	0.731	0.762	0.738	0.837
均值	0.761	0.786	0.766	0.869

5 结论

在多个实际应用数据里找到可用且易于用户理解的知识, 这一过程就叫做数据挖掘. 当挖掘的数据集内某类别样本个数与另外类别样本个数相差较大时, 该种数据集即为不平衡数据. 随着信息时代与大数据时代的来临, 网络入侵检测、文本分类、医疗诊断等各种领域中普遍存在不平衡数据, 一旦出现错分情况, 将引发极大损失, 因此, 本文以自编码网络为核心, 提出一种谱聚类欠取样下的不平衡数据挖掘方法. 由于时间限制, 方法未对运行时间展开针对性的改善, 准备将其作为下一步工作的研究重点, 结合创新型、组合型算法, 缩短挖掘时长; 谱聚类方法以图谱理论为基础, 因 KNN 图复杂度相对更低, 因此, 在今后的研究中需探索一种近似于 KNN 图的图构建方法, 减小复杂度.

参考文献

1 温雪岩, 陈家男, 景维鹏, 等. 面向不平衡数据集分类模型的优化研究. 计算机工程, 2018, 44(4): 268–273, 293. [doi: 10.3969/j.issn.1000-3428.2018.04.043]

2 向鸿鑫, 杨云. 不平衡数据挖掘方法综述. 计算机工程与应用, 2019, 55(4): 1–16. [doi: 10.3778/j.issn.1002-8331.1810-0420]

3 蔡莉, 李英姿, 江芳, 等. 面向城市热点区域的不平衡数据聚类挖掘研究. 计算机科学, 2019, 46(8): 16–22. [doi: 10.11896/j.issn.1002-137X.2019.08.003]

4 许统德, 赵志俊, 高俊文. 多层级联式少数类聚类高精度数据挖掘算法. 控制工程, 2018, 25(5): 829–834.

5 彭显刚, 郑凯, 林哲昊, 等. 基于谱图理论的居民用户非侵入式负荷分解. 电网技术, 2018, 42(8): 2674–2680.

6 王万良, 朱文博, 郑建炜. 基于 ADMM 的拉普拉斯约束表示型聚类算法. 浙江工业大学学报, 2018, 46(4): 363–368, 381. [doi: 10.3969/j.issn.1006-4303.2018.04.002]

7 唐东凯, 王红梅, 胡明, 等. 优化初始聚类中心的改进 K-means 算法. 小型微型计算机系统, 2018, 39(8): 1819–1823. [doi: 10.3969/j.issn.1000-1220.2018.08.033]

8 杨帅, 王鹃. 基于堆栈降噪自编码器改进的混合推荐算法. 计算机应用, 2018, 38(7): 1866–1871.

9 梁俊卿, 赵建视, 吕笑琳. 基于邻里支持和神经网络的 WSN 数据融合算法研究. 微电子学与计算机, 2019, 36(8): 87–91.

10 许赟杰, 徐菲菲. 基于 ArcReLU 函数的神经网络激活函数优化研究. 数据采集与处理, 2019, 34(3): 517–529.

11 孙俏, 凌卫新. 基于域间相似度序数的迁移学习源领域的选择. 科学技术与工程, 2020, 20(20): 8245–8251.

12 姜智涵, 朱军, 周晓锋, 等. 基于信息熵的混合属性数据谱聚类算法. 计算机应用研究, 2019, 36(8): 2256–2260.

13 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. 信息与控制, 2019, 48(3): 257–264.