基于社交网络的双知识表达分类方法

顾苏杭[1,2†], 王士同[1]

(1. 江南大学 数字媒体学院，江苏 无锡 214122； 2. 常州轻工职业技术学院 信息工程与技术学院，江苏 常州 213164)

摘 要: 针对实际数据集中的每一类数据都潜在或显著地包含独有的数据风格信息,提出一种挖掘数据风格信息的双知识表达分类方法. 在训练阶段,利用K近邻(KNN)算法构建社交网络以表达数据点之间的组织架构,并利用社交网络属性挖掘数据点及每一类数据整体风格信息. 在分类阶段,用双知识表达约束所提出方法的分类行为,即赋予测试样本标签时既要使该样本物理上与所建分类模型最相似,也要使该样本风格上与分类模型最相似. 与其他对比分类方法相比,所提出方法在不包含或包含不显著风格的数据集上至少能够取得竞争性的分类性能,在包含明显风格的数据集上能够取得优越性的分类性能.

关键词: 分类算法；双知识表达；社交网络；数据风格信息

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0141

引用格式: 顾苏杭,王士同. 基于社交网络的双知识表达分类方法[J]. 控制与决策, 2020, 35(11): 2653-2664.

Double knowledge representations based classification method from perspective of social networks

GU Su-hang[1,2†], *WANG Shi-tong*[1]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. School of Information Engineering and Technology，Changzhou Institute of Industry Technology，Changzhou 213164，China)

Abstract: Since the distinguished style information of data may latently or obviously present in each data class in a given real-world dataset, a double knowledge representations based classification method (DKR-CM) from the perspective of social networks is proposed. In the training stage, a social network corresponding to all data samples in a dataset is easily built using the on-hand KNN method. In addition, style information of each data sample and each data class are respectively exploited in the social network. In the prediction stage, the proposed double knowledge representations (DKR) is utilized to improve the classification hebaviors of the DKR-SCM. In other words, each data sample is classified into the data class which it approaches to as far as possible from the perspectives of both physical features and style information of data. Experimental results demonstrate that the DKR-CM is at least comparative to the compared classification methods on the datasets with no or inapparent style information and outperforms them on the datasets with obvious style information.

Keywords: classification algorithms；double knowledge representation；social networks；style information of data

0 引 言

传统的模糊/非模糊分类方法,如支持向量机(support vector machine, SVM)[1-3]、K近邻算法(K-nearest neighbors, KNN)[4-5]、随机森林算法(random forest, RF)[6-7]、朴素贝叶斯算法(Naïve Bayes, NB)[8]、决策树C4.5算法[9]以及Takagi-Sugeneo-Kang(TSK)[10-12]等,利用数据点的物理特征(如距离、颜色或相似性)训练数据分类模型,并通过数据点的物理特征判断数据点与所建立分类模型之间的相似性,赋予数据点相应标签类型. 然而,大多数实际数据集中的每一类数据潜在或明显地具有数据风格特征. 例如:癫痫脑电信号识别[13-14],正常人群的脑电信号波形明显不同于患有癫痫的病人;手写体识别[3,15],作者之间的手写体风格互不相同;元音识别[3,16],英文中的每一个元音有着独特的发音且互不相同. 因此,单一地利用数据点物理特征训练数据分类模型,进而对数据点进行分类,并不符合实际数据集中每一类数据拥有独特数据风格的事实.

收稿日期: 2019-01-29；**修回日期:** 2019-05-25.

基金项目: 国家自然科学基金项目(61572236, 61300151)；常州工业职业技术学院博士基金项目(BSJJ13101010)；常州工业职业技术学院新一代信息技术团队项目(YB201813101005)；常州市科技计划项目(CJ20190016).

责任编委: 薛建儒.

†通讯作者. E-mail: gusuhang09@163.com.

文献[3,16]认为,来源于同一类的数据点有着相同的数据风格,而不同类之间的数据风格相互区别. 利用风格矩阵探索数据点组织结构特征并挖掘数据点风格信息可以用于区别不同类数据点. 建筑、漫画以及时尚等图片呈现各自领域的画面风格特征,利用这些不同风格信息计算生成的目标判别式能够有效地区分风格分类(建筑分类、漫画分类以及时尚分类)中的不同对象[17],将具有相同风格的对象归为一类. 文献[15]提出一种风格约束的文本分类方法,即通过可供选择的风格假设赋予文本中每个类(即每个风格)不同高斯密度,用以区别不同类中文本样本的风格特征. 该方法在文本数据集上的分类效果明显优于基于语义、词典(单一的数据物理特征)的文本分类方法. 上述基于风格的分类方法虽然能够很好地表示且将每一类数据风格信息用于分类分析,然而,却忽略了数据点的基本物理特征.

针对上述分析,本文基于社交网络提出一种双知识表达约束的风格分类方法(DKR-CM). 鉴于数据集中的每一类数据暗含或显著具有风格特征,利用社交网络[18-20]映射所有数据点的组织架构并在该架构下利用社交网络属性挖掘数据点以及每一类数据的风格信息. 在此基础上,利用数据点的双知识表达预测数据点的真实标签类型,即数据点的风格信息和物理特征.

1 动 机

传统分类方法[1-12]单纯考虑数据物理特征(距离、颜色或相似性)来训练分类模型,即以单纯计算机的思维规划分类. 但从人类思维角度出发,除了考虑数据物理特征之外还会考虑数据的组织结构. 尤其当数据集包含明显的风格信息时应该挖掘并利用这种风格信息. 例如在社交网络中,人们的言行易受身边亲人或者朋友的影响以及声誉较好的人更容易影响其他人的行为[18,20];关系亲密或者较好的人群更容易具备相同的爱好[21-22];人们在购物时更倾向于选择评价较好的商品[19-20]等. 通过社交网络能够反映实体之间微妙的关系,可以抽象地表达实体之间暗含的组织结构关系,进而可以表示暗含在数据集每一类数据中的风格信息. 因此,本文从人类思维的角度构建数据分类模型,当进行分类时使测试样本点在物理特征和风格特征层面同时无限接近所属的真实数据类,所构建的分类模型能够同时实现以下两种情况的数据分类:

1) 通过基于物理特征的传统分类方法实现数据集不包含或包含不明显数据风格情况下的分类(在分类过程中数据物理特征起决定作用);

2) 通过基于物理特征和数据风格信息的分类方法实现数据集包含明显数据风格情况下的分类(在分类过程中数据风格信息起决定作用).

大量的科学研究已经证明KNN算法[4-5,23-25]是一种高效的分类方法,简单且易于执行. 因此,本文选择KNN算法来实现上述情况1)下的分类. 对于上述情况2)下的分类,由于数据风格信息隐藏在数据点的组织结构中,如何真实地表达数据点的组织结构,进而全面地挖掘暗含的数据风格信息并用于约束分类行为成为关键步骤. 文献[3]在目标函数中引入风格矩阵,通过迭代优化的方式挖掘每一类数据风格信息;文献[17]通过线性变换和非线性激活函数不断地将输入特征映射成隐藏的数据风格知识表达;文献[15]分析了数据集中每一类数据风格的相互独立性,从数学统计的角度以概率的形式预测数据点的标签类型. 显然,这些挖掘数据风格信息的方法并不能简便地表达数据点的组织结构关系.

社交网络致力于表达网络中各个实体之间相互作用的微妙关系,进而为决策系统提供基于实体真实组织结构信息的参考意见[18-22]. 而基于KNN的社交网络[5,23-25]由于其灵活性且易于实现,越来越受到研究人员的青睐. 对于给定数据集$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_N]$中的两个数据点$\boldsymbol{x}_i$和$\boldsymbol{x}_j$(其中:$N$为数据点总数,$\boldsymbol{x}_i\in\boldsymbol{R}^d$),如果同时满足下述两个条件:1) 数据点$\boldsymbol{x}_j$为$\boldsymbol{x}_i$的$k$个邻居之一;2) 数据点$\boldsymbol{x}_i$和$\boldsymbol{x}_j$具有相同的标签. 则基于KNN的社交网络将会建立起点为$\boldsymbol{x}_i$、节点为$\boldsymbol{x}_j$的有向边$\boldsymbol{e}_{ij}$. 于是,通过选择KNN中不同的k值便可将数据集$\boldsymbol{X}$灵活、简便地映射成不同的社交网络SN. SN包含C个带有标签信息且相互独立的子网络并与数据集$\boldsymbol{X}$中的C个数据类一一对应,即$\mathrm{SN}=\{\mathrm{sn}_1,\mathrm{sn}_2,\dots,\mathrm{sn}_C\}$. 通过在SN中探索数据点之间的微妙作用关系,挖掘数据点风格信息以区别数据集中的每一个数据类,为分类决策提供不同于数据物理特征的额外数据信息,以实现上述情况2)下的数据分类.

2 DKR-CM

2.1 数据风格信息

本文通过KNN算法构建社交网络SN,在SN中探索并挖掘每一个数据点以及数据类的风格信息(为便于理解,每一个数据类以及SN中每个子网络统一标识为$\mathrm{sn}_c,1\leqslant c\leqslant C$). 数据风格信息包括数据点的权威性及影响力、数据类的权威性[19,26-27]. sn_c的权威性定义为

$$\varepsilon_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\varepsilon_c^j. \tag{1}$$

其中: N_c 代表 sn_c 中的数据点个数, ε_c^j 代表 sn_c 中第 j 个数据点 $\boldsymbol{x}_j$ 的权威性. ε_c^j 定义为

$$\varepsilon_c^j = \frac{1}{N_i}\sum_{\boldsymbol{e}_{ij}} d_{ij}. \tag{2}$$

其中: $\boldsymbol{e}_{ij}$ 代表起点为 $\boldsymbol{x}_i$、节点为 $\boldsymbol{x}_j$ 的有向边; N_i 代表以 $\boldsymbol{x}_i$ 为起点的有向边个数; d_{ij} 代表 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 间的距离函数且定义为

$$d_{ij} = \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right\}, \tag{3}$$

参数 σ 为距离函数 d_{ij} 的宽度, 可控制 d_{ij} 的衰减速度. 通过式 (3), 本文将在 2.3 节中分析得到每一个数据类的权威性等同于 Parzen 窗口估计. 数据点影响力来源于因特网中的 PageRank 概念, 即对于社交网络中的数据点 $\boldsymbol{x}_i$, 与之相关联的数据点越多, 说明 $\boldsymbol{x}_i$ 在整个社交网络中越重要[28-29]. 数据点影响力定义为

$$\mathrm{In}_j^{(h+1)} = \lambda\sum_{\boldsymbol{e}_{ij}}\frac{\mathrm{In}_i^h}{\deg_i} + (1-\lambda)\frac{1}{N}. \tag{4}$$

其中: In_i^h 代表数据点 $\boldsymbol{x}_i$ 在第 h 次迭代过程中的影响力; λ 代表社交网络阻尼系数, 本文采用文献 [28] 推荐值 $\lambda = 0.85$; $\deg_i$ 代表以数据点 $\boldsymbol{x}_i$ 为起点的有向边个数; N 代表数据点总数. 由式 (4) 可知, 本文迭代地计算社交网络中每个数据点的影响力, 其中第 2 项 $(1-\lambda)\dfrac{1}{N}$ 也反映了每个数据点有着相同的初始状态. 然而, 由文献 [30] 可知, 社交网络中的每个数据点在一定范围内被一定数量的邻居节点所包围, 每个数据点有着不同的局部浓度. 因此, 式 (4) 中每个数据点有着相同的初始状态并不符合数据集的实际分布情况. 本文将式 (4) 中每个数据点的初始状态替换为数据点的局部浓度, 在迭代过程中计算更符合数据集实际分布情况的数据点影响力. 由此, 式 (4) 将变为

$$\mathrm{In}_j^{(h+1)} = \lambda\sum_{\boldsymbol{e}_{ij}}\frac{\mathrm{In}_i^h}{\deg_i} + (1-\lambda)\rho_j, \tag{5}$$

其中 ρ_j 代表第 j 个数据点的局部浓度, 其计算公式如下:

$$\rho_j = \frac{1}{N}\sum_l \chi(d_{jl} - \mathrm{dc}). \tag{6}$$

这里: d_{jk} 代表数据点 $\boldsymbol{x}_j$ 与 $\boldsymbol{x}_k$ 间的距离; dc 代表截断距离, 其值可根据实验分类效果人为地设置, 或根据文献 [30] 使得每个数据点的周围邻居数据点个数占数据点总数的 $1\% \sim 2\%$; $\chi(\cdot)$ 代表距离判断函数, 当 $d_{jk} - \mathrm{dc} < 0$ 时, $\chi(\cdot) = 1$, 反之 $\chi(\cdot) = 0$.

当迭代次数达到设置的最大值 H 或满足以下条件时, 式 (5) 将终止迭代过程:

$$\sum_{j=1}^{N}\|\mathrm{In}_j^{(h+1)} - \mathrm{In}_j^h\| < \theta. \tag{7}$$

其中: $\|\cdot\|$ 代表 2 范数; θ 代表迭代停止阈值, 其值可根据实际分类效果人为地设置, 大量的实验结果表明 $\theta = 10^{-4}$ 能够满足大部分情况下的分类需求.

评论 1　式 (5) 表明, 数据点影响力基于数据点在数据集中的分布情况, 即在每一次迭代过程中都传播数据点局部浓度. 因此, 每个数据点的影响力具有动态特性, 这种动态特性决定了每次迭代过程中的数据点影响力将逐渐趋向于数据点的实际分布情况. 因此, 本文所挖掘的数据风格信息 —— 数据点影响力, 更能反映数据集中所有数据点的实际组织结构关系.

2.2　基于双知识表达的分类

建立起关于数据风格信息的知识表达后便可对数据点进行预测分类. 如第 1 节中所述, 本文从人类思维的角度出发考虑数据点分类问题, 一方面保证数据点的风格与其真实所属的数据类相同, 另一方面保证数据点的物理特征与其真实所属的数据类相同. 与传统分类方法相似, DKR-CM 包含训练和预测两个阶段, 分别描述如下.

训练阶段: DKR-CM 采用基于 KNN 的社交网络构建方法将训练集映射成一个较大网络, 并在该网络中挖掘数据风格信息, 即数据点和数据类的权威性、数据点的影响力.

预测阶段: DKR-CM 利用数据风格信息和数据物理特征两个数据知识表达约束 DKR-CM 的分类行为, 并精确地预测数据点的真实标签类型.

针对测试集中某个测试数据点 $\boldsymbol{t}$, DKR-CM 的具体分类行为描述如下: 当 $\boldsymbol{t}$ 嵌入所建立的社交网络 SN 中时, 首先保证数据点的物理特征与其真实所属的数据类尽可能相同, 这里执行 KNN 算法, 确定 $\boldsymbol{t}$ 的 k 个邻居数据点组成邻居数据集 $\upsilon_{\boldsymbol{t}}$, 即

$$\upsilon_{\boldsymbol{t}} = \mathrm{KNN}(\boldsymbol{t}). \tag{8}$$

其次, 保证数据点的风格与其真实所属的数据类尽可能相同. 这里通过数据集 $\upsilon_{\boldsymbol{t}}$ 根据数据风格信息将测试数据点 $\boldsymbol{t}$ 归为与其风格最相似的数据类, 即

$$c^* = \arg\max_c \psi_c. \tag{9}$$

其中: $\psi_c (1 \leqslant c \leqslant C)$ 与数据类的权威性以及数据点的影响力相关, 具体定义为

$$\psi_c = \varepsilon_c \sum_{\upsilon_c^i \in \upsilon_{\boldsymbol{t}}} \mathrm{In}_{\upsilon_c^i}. \tag{10}$$

υ_c^i 代表数据集 $\upsilon_{\boldsymbol{t}}$ 中的第 i 个数据点且其标签类型与

第c个数据类相同. 按照式(10)计算每一个ψ_c,将测试数据点$\boldsymbol{t}$归为与最大ψ_c(即c^*)相对应的数据类中.

2.3 算法及复杂度分析

真实数据集暗含或明显地具有数据风格特征且显著不同于数据物理特征[3,15,17],而传统分类方法在训练模型和分类阶段并未涉及数据风格特征,针对这种局限性,本文提出一种双知识表达的分类方法. 所提出的DKR-CM兼顾数据物理特征和数据风格信息,打破了传统分类方法在训练和预测阶段单一依赖数据物理特征的局限性. 2.1节详细描述了DKR-CM如何有效地探索并挖掘隐藏的数据风格信息,2.2节描述了DKR-CM如何预测数据点的真实标签类型. 因此,DKR-CM的算法流程可描述如下.

step 1: 给定当前训练集$\boldsymbol{X}_{\rm train} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^{\rm T}(\forall \boldsymbol{x}_j \in R^d, 1 \leqslant j \leqslant N)$,与$\boldsymbol{X}_{\rm train}$相对应的标签集$\boldsymbol{Y}_{\rm train} = [y_1, y_2, \ldots, y_N]^{\rm T}$,测试集$\boldsymbol{X}_{\rm test} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_M]^{\rm T}(\forall \boldsymbol{t}_m \in R^d,\ 1 \leqslant m \leqslant M)$,用于构建社交网络SN的参数$k_1$,用于执行基于数据物理特征知识表达的参数$k_2$,距离函数中的宽度$\sigma$,循环迭代最大次数$H$以及迭代停止阈值$\theta$.

训练阶段:

step 2: 利用KNN算法将训练集$\boldsymbol{X}_{\rm train}$映射成社交网络SN.

step 3: 利用式(1)~(3)分别计算所有$\boldsymbol{x}_j$的权威性ε_c^j以及所有$\rm sn_c$的权威性ε_c.

step 4: 利用式(6)计算所有$\boldsymbol{x}_j$的局部浓度ρ_j.

step 5: 设置$h = 0, {\rm In}_j^0 = \rho_j$.

step 6: 利用式(5)计算所有$\boldsymbol{x}_j$的影响力${\rm In}_j^{(h+1)}$.

step 7: $h = h + 1$,直到h达到最大迭代次数H,或满足$\sum\limits_{j=1}^{N} \|{\rm In}_j^{(h+1)} - {\rm In}_j^h\| < \theta$.

预测阶段:

step 8: 设置$m = 1$.

step 9: 利用式(8)确定$\boldsymbol{t}_m$的邻居数据集$\upsilon_{\boldsymbol{t}_m}$.

step 10: 利用式(10)和(9)确定c^*.

step 11: 将第c个数据类的标签确定为$\boldsymbol{t}_m$的真实标签.

step 12: $m = m + 1$,直到m达到M.

评论2 利用式(3),可将式(1)转换成

$$\varepsilon_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{1}{N_i}\sum_{\boldsymbol{e}_{ij}} d_{ij} = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{1}{N_i}\sum_{\boldsymbol{e}_{ij}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right\} \propto \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{1}{hN_i}\sum_{\boldsymbol{x}\in\boldsymbol{e}_{ij}} \phi(\boldsymbol{x} - \boldsymbol{x}_i, h). \tag{11}$$

其中: $\phi(\cdot)$代表窗口函数; h为该窗口函数的宽度,可用于控制窗口函数的平滑度. 由式(11)可知,此时,每个数据类的风格信息——权威性,本质上等同于Parzen窗口估计[31].

评论3 由DKR-CM算法流程可知,对于训练阶段,所提出的DKR-CM分类方法在所构建的社交网络中可以探索每个数据点之间的微妙作用关系,挖掘数据风格信息. 因此,在训练阶段,DKR-CM不需要根据给定的训练集训练相关分类模型,这与传统分类方法必须训练分类模型存在明显的区别.

评论4 在分类阶段,DKR-CM预测$\boldsymbol{t}_m$真实标签类型时,通过双知识表达来约束DKR-CM的分类行为. 具体地,通过式(8)可使$\boldsymbol{t}_m$在距离上无限接近于所属的数据类,通过式(10)和(9)可使$\boldsymbol{t}_m$在风格上无限接近于所属的数据类. 式(1)中的权威性反映了每个数据类的整体组织结构,式(5)中的影响力具体反映了数据类中每个数据点的动态特性. 因此,本文挖掘的数据风格信息用于分类决策时以概率分布为基础,即将式(10)中ε_c类推到$p(c)$并将$\sum\limits_{\upsilon_c^i\in\upsilon_{\boldsymbol{t}}} {\rm In}_{\upsilon_c^i}$类推到$p(\boldsymbol{x}|c)$,此时,式(10)可转换为

$$c^* = \arg\max_c \varepsilon_c \sum_{\upsilon_c^i\in\upsilon_{\boldsymbol{t}}} {\rm In}_{\upsilon_c^i} \propto \arg\max_c p(c)p(\boldsymbol{x}|c) \propto \arg\max_c p(c|\boldsymbol{x}). \tag{12}$$

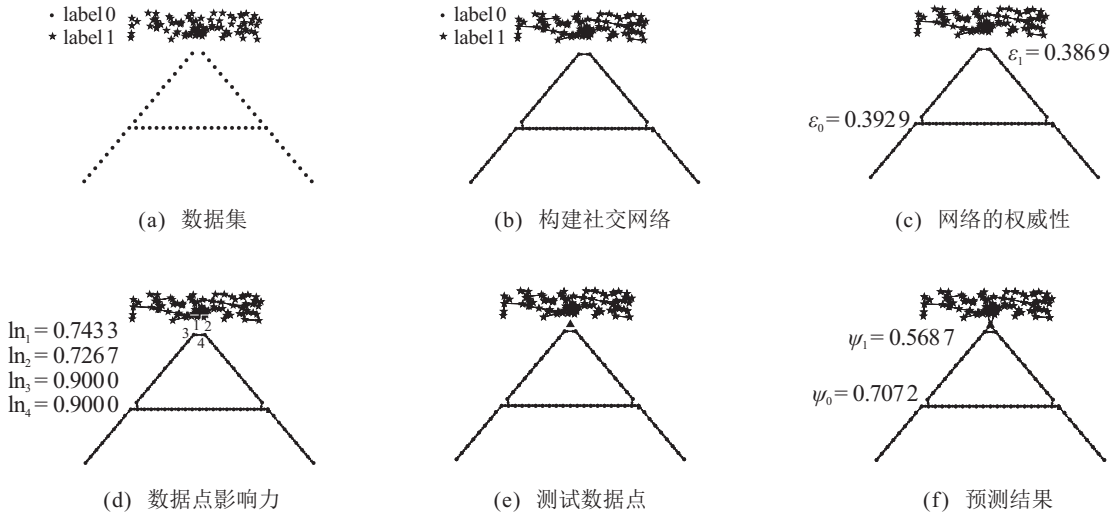
由式(12)可知,可将DKR-CM挖掘数据风格信息用于分类决策的行为类推到贝叶斯决策理论[32],即可将式(9)看作贝叶斯决策规则.

针对DKR-CM的详细算法流程,可给出以下算法复杂度分析. 对比训练阶段与测试阶段,DKR-CM的主要复杂度集中于训练阶段,这里主要分析训练阶段的算法复杂度. 将给定包含N个数据点的训练集$\boldsymbol{X}_{\rm train}$映射成社交网络SN,需要利用式(3)计算任意两个数据点之间的距离,因此,step 2的复杂度为$O(N^2)$. step 3需要$O(N)$复杂度计算每个数据点的权威性,并且需要$O(1)$复杂度计算每个数据类的权威性,考虑到$k_1 \ll N$,因此,step 3的复杂度为$O(N)$. 根据式(6)可得出step 4需要$O(N)$复杂度计算每个数据点的局部浓度. 由于迭代地计算每个数据点的影响力,需要确定每一次迭代的复杂度. 由式(5)可知,step 6的复杂度为$O(N)$. 当考虑最大迭代次数H时,迭代计算每个数据点影响力的最大复杂

度为$O(HN)$. 对于给定的训练集$\boldsymbol{X}_{\text{train}}$, DKR-CM在训练阶段的复杂度为$O(N^2+N+N+HN)\approx O(N^2)$. 因此, DKR-CM的算法复杂度与$N^2$成线性关系, 表明DKR-CM非常适合数据点个数适中的数据分类问题.

2.4　流程展示

图1详细展示了DKR-CM算法的分类过程. 其中, 图1(a)展示了包含两个数据类的数据集, 标签类型分别为0(标志为"•")和1(标志为"⋆"). 标签为0的数据类中所有数据点为随机分布, 并不包含明显的风格; 标签为1的数据类呈现典型的形状, 意味着该数据类显著地拥有数据风格, 暗含着丰富的数据风格信息. 图1(b)展示了利用KNN算法构建的社交网络, 其中参数$k_1=2$, 两个子网络相互独立. 图1(c)展示了两个子网络的权威性, 分别为$\varepsilon_0=0.3929$、$\varepsilon_1=0.3869$, 其中, 式(3)中的参数$\sigma=1$. 图1(d)展示了部分数据点的影响力大小, 其中, 式(5)中参数λ取文献[28]的推荐值, 即$\lambda=0.85$, 式(6)中的截断距离$\mathrm{dc}=0.5$. 图1(e)展示了未标记测试数据点"▲". 当将测试数据点"▲"嵌入到所构建的社交网中时, 首先根据式(8)得到测试数据点"▲"的邻居数据集υ_t, 此时KNN算法中的参数$k_2=4$, υ_t包含4个数据点分别为图1(d)中的数据点1～数据点4; 然后, 根据式(9)和(10), 因为$\psi_0=0.7072>\psi_1=0.5687$, 所以, DKR-CM最终将测试数据点"▲"归为"•"类.



(a) 数据集　(b) 构建社交网络　(c) 网络的权威性

(d) 数据点影响力　(e) 测试数据点　(f) 预测结果

图1　DKR-CM工作流程展示

3　实验研究

本节针对人造数据集、真实数据集以及典型案例研究3个方面将所提出的DKR-CM与对比方法的分类性能进行比较. 对比方法包括传统分类方法和DKR-CM的简易版本. 其中, 传统分类方法包括模糊和非模糊两种类型, 模糊分类方法选择0-order TSK和1-order TSK, 非模糊分类方法选择线性SVM(L-SVM)和高斯型SVM(G-SVM)、KNN、RF、NB以及C4.5. 为了有效验证本文利用数据点局部浓度迭代计算数据点真实影响力, 将DKR-CM的简易版本DKR-CM_0作为对比方法, 即该方法并不考虑数据点的局部浓度, 直接利用式(4)计算每个数据点的影响力.

所有对比算法的参数设置如下: 0-order TSK和1-order TSK的分类性能主要取决于模糊规则数R以及正则化参数τ两个参数, 其中, 参数R的搜索范围为[5, 300], 搜索间隔为5, 参数τ的搜索范围为$\{10^{-5},10^{-4},\ldots,10^{4},10^{5}\}$; L-SVM和G-SVM中的惩罚参数$c$的搜索范围为$\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$, 另外, G-SVM中高斯核宽度参数$\sigma$的搜索范围为$\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$; KNN算法中参数$k$的搜索范围为$\{1,2,\ldots,29,30\}$; RF算法的性能与叶子节点个数$T$相关, 根据文献[7], 将$T$设置为$T=2^7$; NB和C4.5两种算法采用默认配置.

对于本文所提出的DKR-CM, 由详细算法流程可知, 其分类性能主要与以下参数相关: step 2中的参数k_1, step 3中参数σ(见式(3)), step 4中的截断距离dc(见式(6))以及step 9中的参数k_2(见式(8)). 其中, 截断距离dc的设置应使每个数据点周围邻居数据点的个数占数据点总数的1 %～2 %[30], 实验中不再具体列出其值. 另外, 参数σ的搜索范围为$\{2^{-6},2^{-5},\ldots,2^{5},2^{6}\}$, 参数$k_1$的搜索范围为$\{1,2,\ldots,9,10\}$, 较大

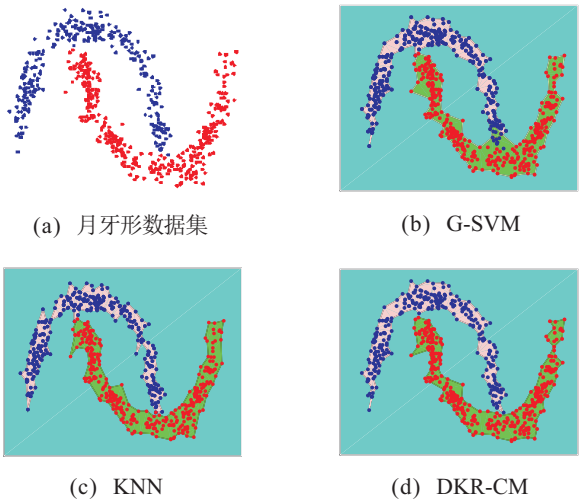
的k_1值将使所建立的社交网络变得复杂. 文献[33]的相关研究结果表明, 当KNN中的邻居个数达到25时, KNN算法性能不再显著提升, 因此, DKR-CM中参数k_2的搜索范围为$\{1,2,\ldots,24,25\}$. 关于对比算法DKR-CM_0, 其相关参数设置参照DKR-CM.

实验中, KNN算法运行30次后取其平均结果, 其他算法运行10次后取其平均结果. 所有算法最优参数均经网格搜索结合10折交叉验证的方法获得. 所有算法均在Matlab软件平台下实现编程.

3.1 人造数据集

本节将在人造数据集上从视觉上展示所提出的DKR-CM分类方法与传统分类方法的区别. 如图2(a)所示为一包含两个数据类(分别标志为"•"和"•")的月牙形数据集, 其中, "•"类包含275个数据点, "•"类包含335个数据点. 从图2(a)可知, "•"类和"•"类分别对应于分布不同的月牙形, 表明这两个数据类各自具有显著的数据风格. 另外, 由于在两个数据类相近的地方, 其中一个数据类的数据点分布稀少而另一个数据类的数据点分布稠密, 因此, 实现图2(a)所示数据集的精确分类是一个考验. 为了便于理解, 此处仅考察G-SVM、KNN以及DKR-CM在图2(a)所示月牙形数据集上的分类性能.

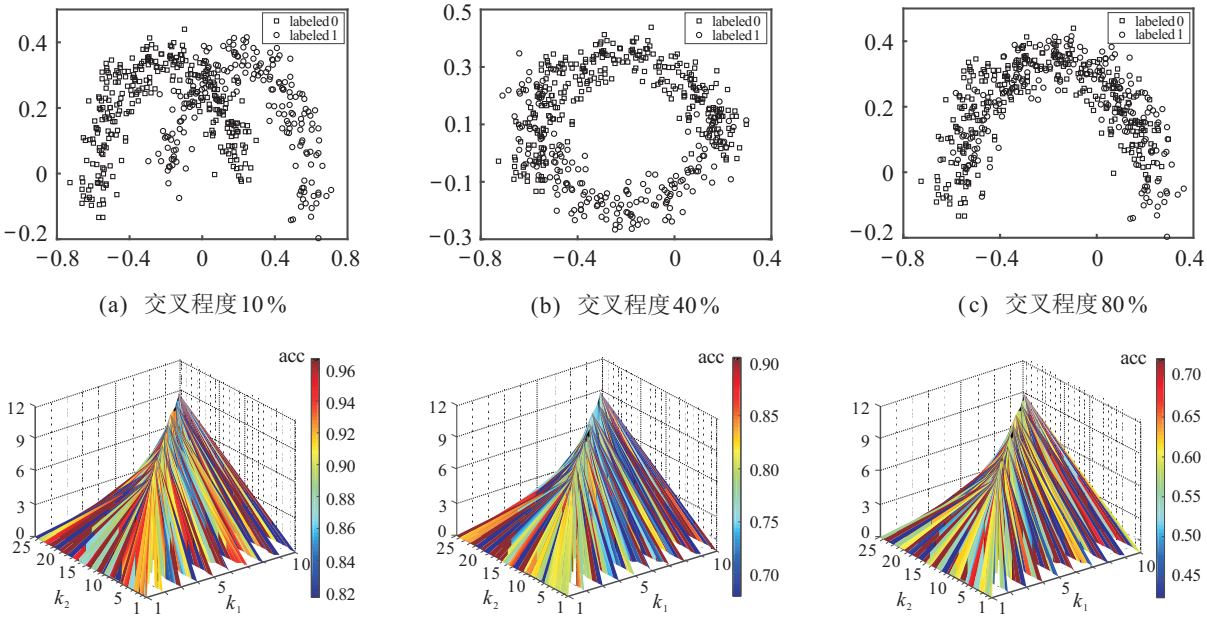
图2(b)、图2(c)以及图2(d)分别从视觉上展示了G-SVM、KNN、DKR-CM的决策边界. 3种算法的参数设置如下: G-SVM中参数为$c=2^4,\sigma=2^5$; KNN中参数$k=4$; DKR-CM中参数为$k_1=5,\sigma=1,k_2=6$. 另外, 在训练阶段, 选取80%的数据点作为训练集. 由于月牙形数据集中两个数据类相近的地方数据点分布不一致, G-SVM和KNN在这两个数据类上的决策边界相互渗透, 不能很好地将两类数据完全隔开; 而本文所提出的DKR-CM能够完全将两类数据区别开来, 尤其在两个数据类相近的地方能够精确地识别每一类数据中的数据点.



(a) 月牙形数据集 (b) G-SVM (c) KNN (d) DKR-CM

图2 3种算法的决策边界比较

为进一步了解DKR-CM, 在人造数据集(如图3(a)～图3(c)所示)上对参数k_1、σ以及k_2的敏感性进行分析. 如图3(a)～图3(c)所示数据集, 每个数据集中的每一类数据对应于分布不同的半月牙形, 因此, 这3个数据集包含典型的数据风格. 另外, 图3(a)～图3(c)所示的数据集中, 不同类数据之间的交



(a) 交叉程度10% (b) 交叉程度40% (c) 交叉程度80%

(d) 参数敏感性结果(1) (e) 参数敏感性结果(2) (f) 参数敏感性结果(3)

图3 DKR-CM参数敏感性分析

叉程度分别为10 %、40 %、80 %, 不同数据类之间的交叉程度越高, 精确分类的难度也相应提高. 实验中, 当分析某一参数敏感性时, 另外两个参数保持最优参数值, 所有参数最优值均经网格搜索结合10折交叉验证的方法获得. DKR-CM参数敏感性分析结果如图3(d) ∼ 图3(e)所示, 其中参数σ的优化范围为$\{2^{-6}, 2^{-5}, \ldots, 2^{5}, 2^{6}\}$, 分别对应于图3(d) ∼ 图3(e)中轴刻度值$\{0, 1, \ldots, 11, 12\}$. 另外, 图3(d) ∼ 图3(e)中纵向颜色条的不同颜色代表不同的分类精度, 从颜色条底部到顶部代表分类精度越来越高.

由图3(d) ∼ 图3(e)可得到以下观察: 1) 当图3(a) ∼ 图3(c)所示数据集中不同类数据之间交叉程度达到80 %时, DKR-CM至少能取得的分类精度为0.70, 这说明本文所提出的挖掘数据风格信息的双知识表达在数据集难以分类的情况下能够保证DKR-CM取得较为满意的分类结果; 2) 通过颜色变化观察, 参数k_1和k_2对DKR-CM的分类影响较大, 因此, 在实验过程中这两个参数的最优值必须通过网格搜索结合交叉验证的方法获取; 3) 数据点及数据类权威性基于式(3)定义的距离函数, 当参数σ取不同值时, DKR-CM取得分类精度变化较大, 因此, 在实验过程中参数σ的最优值必须通过网格搜索结合交叉验证的方法获取, 以保证数据点及数据类权威性贴近实际数据集包含的数据风格.

3.2　真实数据集

本节将在真实数据集上将所提出的DKR-CM与对比分类方法进行实际分类性能比较, 所选数据集不包含或者包含不明显的数据风格. 所有真实数据集均来自UCI数据库[34], 表1详细列出了真实数据集的配置信息. 实验中, 从数据集中随机选取60 %的数据点作为训练样本, 其余作为测试样本.

表 1　人造数据集详细配置

数据集	数据点个数	维数	类别数
flare(FLA)	1 066	11	6
haberman(HAB)	306	3	2
monks1(MON)	432	6	2
phoneme(PHO)	5 404	5	2
seeds(SEE)	210	7	3
seismic_bumps(SEI)	2 584	18	2
titanic(TIT)	2 201	3	2
vehicle(VEH)	846	18	4
wisconsin(WIS)	683	9	2
zoo(ZOO)	101	16	7

表2详细列出了各分类方法在真实数据集上的分类结果. 其中: “acc”代表分类精度, “opt”代表每个算法取得最佳分类结果情况下的参数值, “rank”代表利用统计测试方法按照各分类方法分类精度的排序值, “average rank”代表各分类方法的平均排序值, “–”代表分类方法采用默认配置, “(–)”代表相关分类性能指标的标准差小于10^{-4}, 各分类方法中最好的分类结果已用黑体标出.

根据表2可得出以下观察:

1) 就分类精度而言, 所提出的DKR-CM和DKR-CM_0在大部分真实数据集上取得了最好的分类结果. 另外, 作为典型的模糊分类方法, 由于具有优越的泛化性, 1-order TSK在HAB、MON、SEE以及TIT上取得了最好的分类结果, 表现出较好的分类性能; 同时, DKR-CM和DKR-CM_0在大部分真实数据集上的分类性能明显优于0-order TSK.

2) 由于考虑了数据集中数据点的实际分布情况, DKR-CM在真实数据集上的分类性能总体上优于DKR-CM_0, 从而验证了在迭代挖掘数据风格信息 —— 数据点影响力的过程中传播数据点局部浓度的有效性.

3) 在不包含或包含不明显数据风格真实数据集上的分类结果并不能显著地说明对比分类方法优于所提出的DKR-CM分类方法, 即不能明显地表明在分类过程中应该仅考虑数据物理特征还是考虑数据物理特征结合数据风格信息来训练数据分类模型.

为了进一步比较DKR-CM与其他对比分类方法的区别, 本文利用文献[35]中的统计测试方法对所有分类方法在真实数据集上的分类性能进行分析. 该统计测试方法主要包括以下3个步骤.

step 1: 首先根据表2列出的分类精度对所有对比分类方法进行排序. 例如: 对于WIS, 因DKR-CM取得了最好的分类精度0.979 9, 故其排名为1, 即$\mathrm{rank} = 1$; 相应地, 1-order TSK的排名为$\mathrm{rank} = 2$. 依此类推, 可得出各对比分类方法在真实数据集上的排序, 如表2中列出的“rank”值.

step 2: 在进行统计测试前, 假设各对比分类方法在真实数据集上具有相同的分类性能. 由表2根据F-分布检验该空假设是否成立. 首先, 根据文献[35], 利用Friedman测试确定$F((N_c-1),(N_c-1)(N_d-1))$, 其中$N_c$代表各对比分类方法个数, N_d代表真实数据集个数. 由表2有$N_c = 10$和$N_d = 10$. 因此, 根据F-分布有$F(9, 81) \approx 2.00$, 其中显著性水平α选择文献[35]推荐值$\alpha = 0.05$. 其次, 根据式(13)和(14)计算确定F_F, 只要F_F值大于$F(9, 81) \approx 2.00$, 则step 2中的空假设即被否定. F_F的计算公式为

$$F_F = \frac{(N_d - 1)\chi_F^2}{N_d(N_c - 1) - \chi_F^2}, \tag{13}$$

表 2 各分类方法在真实数据集上的分类结果

数据集	指标	0-order TSK (R,τ)	1-order TSK (R,τ)	L-SVM (c)	G-SVM (c,σ)	KNN (k)	RF	NB	C4.5	DKR-CM_0 (k_1,σ,k_2)	DKR-CM (k_1,σ,k_2)
FLA	acc	0.331 9	0.682 3	0.717 4	0.729 8	0.732 4	0.745 5	0.637 6	0.732 4	**0.755 9**	0.744 1
		(0.039 2)	(0.012 3)	(0.023 9)	(0.013 2)	(0.014 6)	(0.017 0)	(0.031 8)	(0.016 8)	(−)	(−)
	opt	$(30.5,10^{-5})$	$(295,10^{-2})$	(2^5)	$(2^5,2^6)$	(13)	–	–	–	(5,1,3)	$(5,2^{-2},3)$
	rank	10	8	7	6	4	2	9	4	1	3
HAB	acc	0.775 4	**0.778 7**	0.726 2	0.731 1	0.772 1	0.676 2	0.741 0	0.679 5	0.762 3	**0.778 7**
		(0.041 3)	(0.024 1)	(0.034 6)	(0.025 3)	(0.026 6)	(0.033 4)	(0.025 2)	(0.036 0)	(0.000 9)	(0.001 1)
	opt	(115.5,10)	$(245.5,10^{-5})$	(2^4)	$(2^6,2^5)$	(26)	–	–	–	$(3,2^{-1},16)$	$(10,2^{-5},20)$
	rank	3	1	8	7	4	10	6	9	5	1
MON	acc	0.5419	**0.558 1**	0.380 3	0.470 5	0.441 9	0.200 6	0.446 5	0.278 4	0.517 4	0.529 1
		(0.031 9)	(0.021 8)	(0.027 0)	(0.015 6)	(0.031 4)	(0.015 7)	(0.020 6)	(0.013 4)	(0.000 5)	(0.000 3)
	opt	$(270,10^{-4})$	$(75,10^2)$	(2^6)	$(2^{-6},2^{-6})$	(24)	–	–	–	$(2,2^{-3},20)$	$(2,2^{-1},15)$
	rank	2	1	8	5	7	10	6	9	4	3
PHO	acc	0.706 1	0.880 8	0.804 8	0.804 4	0.892 2	**0.900 1**	0.754 3	0.851 8	0.844 5	0.850 1
		(0.008 2)	(0.005 8)	(0.011 6)	(0.013 2)	(0.065 8)	(0.006 1)	(0.010 3)	(0.006 7)	(−)	(−)
	opt	$(145.5,10^3)$	$(280,10^{-3})$	(2^5)	$(2^4,2^4)$	(1)	–	–	–	$(5,2^{-4},8)$	$(5,2^{-3},7)$
	rank	10	3	7	8	2	1	9	4	6	5
SEE	acc	0.395 2	**0.940 5**	0.916 7	0.906 0	0.921 4	0.922 6	0.900 0	0.919 0	0.924 6	0.940 5
		(0.044 8)	(0.035 0)	(0.028 2)	(0.026 3)	(0.017 8)	(0.014 3)	(0.024 5)	(0.023 7)	(0.000 2)	(0.000 1)
	opt	$(255,10^4)$	$(110,10^{-1})$	(2^3)	$(2^{-1},2^3)$	(11)	–	–	–	$(2,2^{-6},6)$	$(5,2^{-1},2)$
	rank	10	1	7	8	5	4	9	6	3	1
SEI	acc	0.942 1	0.940 9	0.935 4	0.934 3	0.941 5	0.932 7	0.102 0	0.896 2	**0.943 9**	0.940 0
		(0.004 3)	(0.005 2)	(0.070 1)	(0.053 7)	(0.066 7)	(0.004 2)	(0.049 7)	(0.010 3)	(−)	(0.000 1)
	opt	$(95.5,10^{-2})$	$(110,10^5)$	(2^4)	$(2^6,2^{-1})$	(27)	–	–	–	$(1,2^{-3},17)$	$(2,2^2,18)$
	rank	2	4	6	7	3	8	10	9	1	5
TIT	acc	0.723 0	**0.798 1**	0.785 5	0.783 4	0.787 7	0.780 3	0.780 5	0.784 7	0.790 3	0.791 5
		(0.034 5)	(0.005 7)	(0.042 7)	(0.010 7)	(0.017 3)	(0.010 2)	(0.010 1)	(0.012 8)	(−)	(−)
	opt	(10,10)	$(10.5,10^{-4})$	(2^5)	$(2^6,2^6)$	(8)	–	–	–	$(4,2^{-3},17)$	$(6,2^{-2},12)$
	rank	10	1	5	7	4	9	8	6	3	2
VEH	acc	0.279 3	0.685 4	**0.821 3**	0.815 7	0.700 3	0.752 7	0.469 8	0.684 3	0.698 2	0.699 7
		(0.012 8)	(0.010 9)	(0.018 1)	(0.015 4)	(0.016 0)	(0.007 8)	(0.024 4)	(0.023 7)	(−)	(0.000 1)
	opt	$(155,10^{-5})$	$(165,10^{-3})$	(2^6)	$(2^6,2^6)$	(6)	–	–	–	(1,2,1)	(1,2,1)
	rank	10	7	1	2	4	3	9	8	6	5
WIS	acc	0.965 8	0.978 0	0.969 2	0.971 1	0.971 8	0.969 2	0.957 5	0.944 0	0.972 5	**0.979 9**
		(0.015 3)	(0.007 9)	(0.033 6)	(0.010 2)	(0.010 2)	(0.009 3)	(0.005 0)	(0.013 1)	(−)	(−)
	opt	$(255,10^5)$	(160,10)	(2^6)	$(2^4,2^2)$	(5)	–	–	–	(1,1,1)	$(8,2^6,4)$
	rank	8	2	6	5	4	6	9	10	3	1
ZOO	acc	0.508 3	0.950 0	0.975 0	0.947 5	0.947 5	0.920 0	0.820 0	0.880 0	**0.987 5**	0.950 0
		(0.031 2)	(0.020 4)	(0.015 8)	(0.034 4)	(0.028 4)	(0.035 0)	(0.005 6)	(0.036 7)	(0.000 2)	(0.003 9)
	opt	$(120,10^{-5})$	$(20,10^{-1})$	(2^3)	$(2^5,2^4)$	(1)	–	–	–	$(3,2^{-6},1)$	(1,2,2)
	rank	10	3	2	5	5	7	9	8	1	3
average rank		7.50	3.10	5.70	6.00	4.20	6.00	8.40	7.30	3.30	**2.90**

其中χ_F^2可进一步定义为

$$\chi_F^2=\frac{12N_d}{N_c(N_c+1)}\Big[\sum_{j=1}^{N_c}\bar{R}_j^2-\frac{N_c(N_c+1)^2}{4}\Big],\quad(14)$$

$\bar{R}_j$代表第j个对比分类方法在真实数据集上的平均排序. 至此,可计算得出$F_F\approx 4.79$. 显然,F_F值大于$F(9,81)\approx 2.00$,step 2中的空假设不成立,即各个对

比分类方法在真实数据集上的分类性能存在本质区别.

step 3: 根据 step 2 的结果, 可进一步利用 Bonferroni-Dunn 测试检验对比分类方法之间详细的分类性能区别. 根据文献 [35], 通过表 2 可以确定 CD ≈ 3.75. CD 的计算公式为

$$\mathrm{CD} = q_a\sqrt{\frac{N_c(N_c+1)}{6N_d}}, \tag{15}$$

其中 $q_a = 2.773$ 且 $\alpha = 0.05$[35].

根据表 2, 只要任意两个对比分类方法之间的平均排名值之差大于 CD ≈ 3.75, 即表明这两个对比分类方法之间存在显著区别. 根据统计测试结果可得出以下结论:

1) DKR-CM 及 DKR-CM_0 与对比分类方法 NB 之间的平均排名差值分别为 5.50、5.10, 排名差值均大于 CD ≈ 3.75, 因此, DKR-CM 及 DKR-CM_0 与对比分类方法 NB 之间存在显著区别.

2) 由于 DKR-CM 在迭代计算数据点影响力的过程中动态传播数据点局部浓度, DKR-CM 在真实数据集上的平均排名优于 DKR-CM_0, 这与上述实验分析中的 2) 保持一致.

3) DKR-CM 具有最小的平均排名值 2.90, 代表所提出的方法在真实数据集上的平均分类性能最好. 但由于 DKR-CM 与其他对比分类方法 (除 NB 以外) 之间平均排名差值均小于 CD ≈ 3.75, DKR-CM 与其他对比分类方法之间不存在显著区别, 代表 DKR-CM 在不包含或者包含不明显数据风格数据集上至少能够取得竞争性的分类性能.

3.3　典型案例研究

本节选取癫痫脑电信号识别以及手写体识别作为典型案例来验证所提出的 DKR-CM 在具有典型数据风格数据集上的优秀分类性能. 每一个典型案例中的每一个数据类拥有明显的数据风格, 且数据类之间的风格互不相同.

3.3.1　癫痫脑电信号识别

对于癫痫脑电信号识别[13-14], 其原始信号和经核化主成分分析 (kernel principal component analysis, K-PCA) 降维后的特征分别如图 4 和图 5 所示. 通过观察图 4 和图 5 可知, 正常人群的脑电信号 (A 组和 B 组) 明显区别于患癫痫人群的脑电信号 (C 组、D 组和 E 组). 另外, 即使属于同一人群, 所表现出的脑电信号也相互区别, 如图 4 和图 5 中 A 组和 B 组所示, 以及 C 组、D 组和 E 组所示. 实验中分别将 A 组和 C 组, A 组、C 组和 E 组, C 组、D 组和 E 组组成 EEG 信号数据集并分别命名为 EEG-D1、EEG-D2、EEG-D3. 3 个 EEG 信号数据集的详细配置信息如表 3 所示.

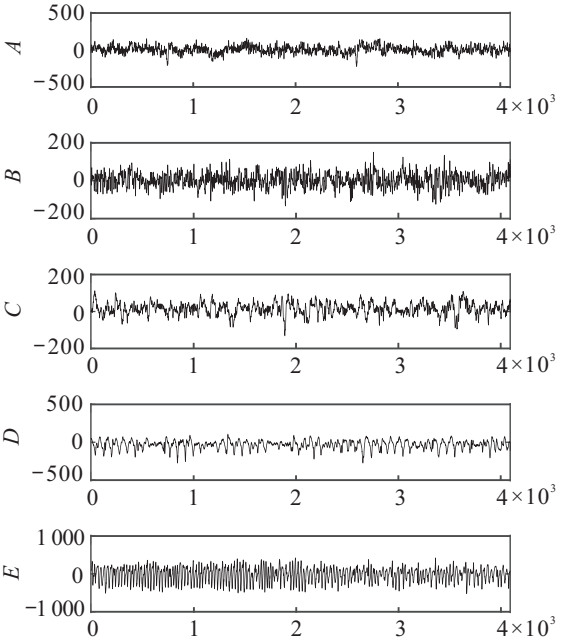


图 4　EEG 原始信号

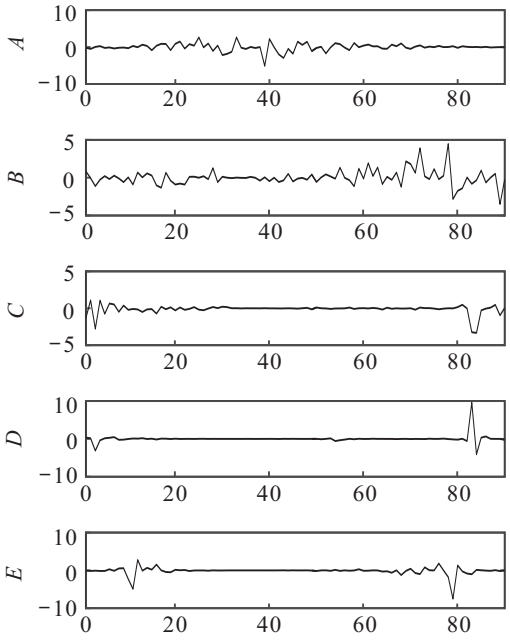


图 5　部分手写体数据展示

表 3　典型案例数据集配置信息

数据集	数据点个数	维数	类别数
flare(EEG-D1)	200	50	2
haberman(EEG-D2)	300	50	3
monks1(EEG-D3)	300	50	3
phoneme(HW1)	10 000	6	5
seeds(HW2)	20 000	6	10
seismic_bumps(HW3)	30 000	6	15

3.3.2 手写体识别

对于手写体识别[3,15], 实验中选取Chinese Academy of Science Institute of Automation (CASIA)官网公布的手写体数据集[36], 其中部分数据如图6所示.

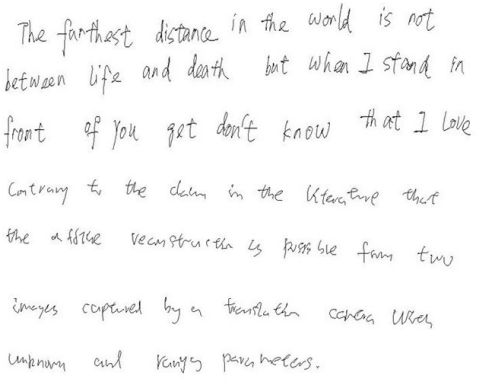


图6 经K-PCA特征降维后的部分EEG信号

由此可知, 出自每一位作者的手写体均相互区别, 因此, 相应的手写数据风格也相互区别. 参照文献[3], 实验中随机选取5位、10位以及15位作者, 并从每一位作者的手写数据中随机选取2 000条数据样本组成手写体数据集HW1、HW2、HW3. 3个手写体数据集的详细配置信息如表3所示.

3.3.3 各分类方法在典型案例上的分类结果及分析

表4详细地列出了各分类方法在典型案例上的分类结果. 由表4可知, 所提出的DKR-CM在所有具有典型数据风格的数据集上均取得了最好的分类性能. 作为DKR-CM的简易版本, 由于考虑了数据风格信息以及数据物理特征, DKR-CM_0取得的分类性能优于大部分对比算法(除DKR-CM以外). 另外, 作为优秀的传统分类方法, RF在典型案例研究中同样取得了较好的分类结果.

表4 各分类方法在典型案例上的分类结果

数据集	指标	0-order TSK	1-order TSK	L-SVM	G-SVM	KNN	RF	NB	C4.5	DKR-CM_0	DKR-CM
		(R,τ)	(R,τ)	(c)	(c,σ)	(k)				(k_1,σ,k_2)	(k_1,σ,k_2)
EEG-D1	acc	0.550 0	0.785 0	0.825 0	0.823 8	0.815 0	0.862 5	0.807 5	0.702 5	0.881 3	**0.907 5**
		(0.027 0)	(0.024 6)	(0.017 7)	(0.040 1)	(0.035 2)	(0.041 8)	(0.035 9)	(0.041 0)	(0.003 2)	(0.001 4)
	opt	$(40.5,10^5)$	$(125,10^{-2})$	(2^6)	$(2^6,2^6)$	(1)	–	–	–	$(7,2^3,1)$	$(3,2^{-1},1)$
	rank	10	8	4	5	6	3	7	9	2	1
EEG-D2	acc	0.375 0	0.761 7	0.871 7	0.867 5	0.809 2	0.890 8	0.866 7	0.815 8	0.887 5	**0.915 8**
		(0.036 0)	(0.021 5)	(0.026 6)	(0.028 5)	(0.034 9)	(0.032 7)	(0.026 4)	(0.027 8)	(0.000 1)	(0.001 1)
	opt	$(55,10^{-4})$	$(70.5,10^{-3})$	(2^6)	$(2^6,2^6)$	(1)	–	–	–	$(8,2^5,1)$	$(5,2^{-5},3)$
	rank	10	9	4	5	8	2	6	7	3	1
EEG-D3	acc	0.372 2	0.801 7	0.830 0	0.770 5	0.845 8	0.834 2	0.760 0	0.755 8	0.855 0	**0.865 0**
		(0.030 7)	(0.035 9)	(0.027 0)	(0.029 4)	(0.030 6)	(0.023 7)	(0.033 1)	(0.050 0)	(0.004 4)	(–)
	opt	(30.5, 10)	$(100,10^{-4})$	(2^5)	$(2^{-6},2^{-6})$	(1)	–	–	–	$(8,2^4,1)$	$(4,2^{-3},2)$
	rank	10	6	5	7	3	4	8	9	2	1
HW1	acc	0.206 4	0.894 7	0.920 4	0.919 2	0.912 7	0.949 7	0.836 1	0.930 6	0.923 4	**0.960 1**
		(0.002 2)	(0.006 8)	(0.043 7)	(0.031 1)	(0.055 1)	(0.002 3)	(0.008 8)	(0.022 4)	(–)	(–)
	opt	$(40,10^{-5})$	$(205.5,10^{-2})$	(2^6)	$(2^5,2^3)$	(1)	–	–	–	$(7,2^2,1)$	$(5,2^2,1)$
	rank	10	8	5	6	7	2	9	3	4	1
HW2	acc	0.103 5	0.422 6	0.770 7	0.769 2	0.799 8	0.812 7	0.620 8	0.806 2	0.807 5	**0.853 7**
		(0.002 5)	(0.006 1)	(0.239 5)	(0.035 9)	(0.040 7)	(0.039 0)	(0.004 7)	(0.047 3)	(–)	(0.000 1)
	opt	(90.5, 1)	$(290.5,10^{-4})$	(2^2)	$(2^2,2^5)$	(1)	–	–	–	$(6,2^2,1)$	(6, 2, 1)
	rank	10	9	6	7	5	2	8	4	3	1
HW3	acc	0.106 8	0.282 6	0.686 1	0.683 2	0.740 7	0.788 6	0.531 8	0.769 9	0.783 9	**0.799 2**
		(0.002 2)	(0.002 3)	(0.029 2)	(0.054 5)	(0.046 3)	(0.041 9)	(0.004 7)	(0.053 6)	(–)	(–)
	opt	$(60,10^3)$	$(295.5,10^{-5})$	(2^3)	$(2^6,2^3)$	(1)	–	–	–	$(3,2^2,1)$	$(8,2^{-1},1)$
	rank	10	9	6	7	5	2	8	4	3	1
average rank		10.00	8.17	5.00	6.17	5.67	2.50	7.67	6.00	2.83	**1.00**

为了进一步分析各分类方法在典型案例上的分类性能,同样采取上述统计测试方法[35]. 首先,根据表4,利用Friedman测试以及F-分布确定$F(9,45)\approx 2.10$. 其次,根据式(13)和(14)可计算确定$F_F\approx 26.46$,因此,可以否定空假设,即所有分类方法在典型案例研究中的分类性能有着本质区别. 进一步,可执行Bonferroni-Dunn测试检验,根据式(15)可确定$\mathrm{CD}\approx 4.85$. 因此,任何两个分类方法之间的平均排名差值大于$\mathrm{CD}\approx 4.85$,从而可以确定这两个分类方法之间存在本质区别.

根据表4可得出以下结论:

1) DKR-CM在典型案例研究中的分类性能优于DKR-CM_0,再次验证了结合数据点在数据集中的实际分布情况计算数据点影响力的有效性.

2) 由于DKR-CM与0-order TSK、1-order TSK、G-SVM、NB以及C4.5之间平均排名差值分别为9.00、7.17、5.17、6.67、5.00,因此,DKR-CM与这些对比方法之间存在本质区别. DKR-CM与L-SVM、KNN之间的平均排名差值虽然小于$\mathrm{CD}\approx 4.85$,但DKR-CM具有最好的分类性能.

3) 典型案例研究结果表明,在数据分类过程中不应仅利用数据物理特征训练数据分类模型,还应结合数据风格信息.

4 结　论

本文在与给定数据集相匹配的社交网络中挖掘数据点以及每一类数据的数据风格信息,即数据点权威性、影响力以及数据类权威性. 特别地,本文根据数据集中数据点的实际分布情况计算每一个数据点浓度,并在迭代过程中传播每一个数据点浓度,动态地计算更符合数据点实际分布情况的影响力. 在社交网络中挖掘数据风格信息决定了本文所提出的DKR-CM并不需要训练分类模型,在分类阶段直接利用数据物理特征以及数据风格信息预测数据点的标签类型,即利用双知识表达约束DKR-CM的分类行为. 大量实验结果表明,DKR-CM能够在不包含或者包含不明显数据风格的数据集上取得与传统分类方法相当的分类性能,在包含显著数据风格的数据集上取得优于传统分类方法的分类性能.

在未来研究中,将进一步研究社交网络属性以及如何将基于社交网络的双知识表达分类模型推广到数据聚类分析中.

参考文献(References)

[1] 史荧中, 邓赵红, 王士同, 等. 基于共享矢量链的多任务概念漂移分类方法[J]. 控制与决策, 2018, 33(7): 1215-1222.
(Shi Y Z, Deng Z H, Wang S T, et al. Multi-task concept drift classification method based on shared vector chain[J]. Control and Decision, 2018, 33(7): 1215-1222.)

[2] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27.

[3] Huang K, Jiang H, Zhang X. Field support vector machines[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2017, 1(6): 454-463.

[4] Zhang S, Li X, Zong M, et al. Efficient kNN classification with different numbers of nearest neighbors[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1774-1785.

[5] Roberto J, Zhao L, Motta R, et al. A nonparametric classification method based on K-associated graphs[J]. Information Sciences, 2011, 181(24): 5435-5456.

[6] Wang Y, Xia S, Tang Q, et al. A novel consistent random forest framework: Bernoulli random forests[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(8): 3510-3523.

[7] Oshiro T M, Perez P S, Baranauskas J A. How many trees in a random forest?[C]. Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition. Berlin, 2012: 154-168.

[8] Russell S, Norvig P. Artificial intelligence: A modern approach[M]. 2nd ed. Upper Saddle River: Prentice Hall, 2003: 597.

[9] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[10] 李滔, 王士同. 适合大规模数据集且基于LLM的0阶TSK模糊分类器[J]. 控制与决策, 2017, 32(1): 21-30.
(Li T, Wang S T. Zero-order TSK fuzzy classifier based on LLM for large-scale data sets[J]. Control and Decision, 2017, 32(1): 21-30.)

[11] Zhou T, Chung F L, Wang S. Deep TSK fuzzy classifier with stacked generalization and triplely concise interpretability guarantee for large data[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(5): 1207-1221.

[12] Zhang Y, Hisao I, Wang S. Deep Takagi-Sugeno-Kang fuzzy classifier with shared linguistic fuzzy rules[J]. IEEE Transactions on Fuzzy Systems, 2018, 26(3): 1535-1549.

[13] Yang C, Deng Z, Wang S, et al. Transductive domain adaptive learning for epileptic electroencephalogram recognition[J]. Artificial Intelligence in Medicine, 2014, 62(3): 165-177.

[14] Xie L, Deng Z, Wang S, et al. Generalized hidden-mapping transductive transfer learning for recognition of epileptic electroencephalogram signals[J].

IEEE Transactions on Cybernetics, 2018, 49(6): 2200-2214.

[15] Sarkar P, Nagy G. Style consistent classification of isogenous patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(1): 88-98.

[16] Zhang X, Huang K, Liu C. Pattern field classification with style normalized transformation[C]. Proceedings of the International Joint Conference on Artificial Intelligence. Barcelona, 2011: 1621-1626.

[17] Jiang S, Shao M, Jia C, et al. Learning consensus representation for weak style classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2906-2919.

[18] Shang F, Liu Y, Yan D, et al. Fuzzy double trace norm minimization for recommendation systems[J]. IEEE Transactions on Fuzzy Systems, 2018, 26(4): 2039-2049.

[19] Golbeck J, Hendler J. Accuracy of metrics for inferring trust and reputation in semantic web-based social networks[C]. Proceedings of the International Conference on Knowledge Engineering and Knowledge Management. Northamptonshire: Whittlebury Hall, 2004: 116-131.

[20] Ma H, Lyu M R, Zhou D, et al. Recommender systems with social regularization[C]. Proceedings of the International Conference on Web Search and Web Data Mining. Hong Kong, 2011: 287-296.

[21] McPherson M, Smith-Lovin L, Cook J. Birds of a feather: Homophily in social networks[J]. Annual Review of Sociology, 2001, 27(1): 415-444.

[22] Tang J, Gao H, Liu H. mTrust: Discerning multi-faceted trust in a connected world[C]. Proceedings of the International Conference on Web Search Data Mining. Washingtion, 2012: 93-102.

[23] Ozaki K, Shimbo M, Komachi M, et al. Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data[C]. Proceedings of the International Conference on Computational Natural Language Learning. Oregon, 2011: 154-162.

[24] Chen J, Fang H R, Saad Y. Fast approximate KNN graph construction for high dimensional data via recursive lanczos bisection[J]. Journal of Machine Learning Research, 2009, 10: 1989-2012.

[25] Carneiro M G, Zhao L. Organizational data classification based on the importance concept of complex networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(8): 3361-3373.

[26] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC, 2003: 137-146.

[27] Kuter U, Golbeck J. SUNNY: A new algorithm for trust inference in social networks using probabilistic confidence models[C]. Proceedings of the National Conference on Artificial Intelligence. Vancouver, 2007: 1377-1382.

[28] Boldi P, Santini M, Vigna S. PageRank as a function of the damping factor[C]. Proceedings of the International Conference on World Wide Web. Chiba, 2005: 557-566.

[29] Liu Q, Xiang B, Yuan N J, et al. An influence propagation view of PageRank[J]. ACM Transactions on Knowledge Discovery from Data, 2017, 11(3): 30.

[30] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344: 1492-1496.

[31] Kwak N, Choi C. Input feature selection by mutual information based on Parzen window[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(12): 1667-1671.

[32] Duda R O, Hart P E, Stork D G. Pattern classification[M]. 2nd ed. Wiley-Interscience, 2000: 20.

[33] Liu Q, Liu C. A novel locally linear KNN method with applications to visual recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(9): 2010-2021.

[34] Frank A, Asuncion A. UCI Machine Learning Repository[EB/OL]. Available: http://archive.ics.uci.edu/ml.

[35] Demsar J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of Machine Learning Research, 2006, 7: 1-30.

[36] Chinese Academy of Sciences Institute of Automation (CASIA). Handwriting database[EB/OL]. http://biometrics.idealtest.org/.

作者简介

顾苏杭(1989−), 男, 博士生, 从事人工智能与模型识别、机器学习的研究, E-mail: gusuhang09@163.com;

王士同(1964−), 男, 教授, 博士生导师, 从事人工智能与模式识别、机器学习、深度学习等研究, E-mail: wxwangst@jiangnan.edu.cn.

(责任编辑：李君玲)