文章编号:1671-4598(2019)04-0190-08 DOI:10.16526/j.cnki.11—4762/tp.2019.04.043 中图分类号:TP391.4 文献标识码:A

挖掘数据模式结构信息的混合数据分类方法

王惠宇[1], 顾苏杭[1,2]
(1. 常州轻工职业技术学院 信息工程学院,江苏 常州 213164;
2. 江南大学 数字媒体学院,江苏 无锡 214122)

摘要:数据集中数据之间往往相互关联,所有数据整体上呈现特定的模式结构,而传统分类方法(如支持向量机)忽略数据关联信息,仅仅利用数据的物理特征(如距离、相似性等)构建数据分类模型,并在分类阶段计算测试样本与所建立分类模型间的相似性来预测测试样本的标签类型;为了解决传统分类方法利用单一数据信息的问题,提出一种挖掘数据模式结构信息的混合数据分类方法;该方法融合了两种不同类型的分类技术,将使用单一数据物理特征的传统分类方法作为普通分类方法,将利用数据模式结构信息的分类方法作为高级分类方法;尤其是该方法不仅可有效地识别数据模式结构信息以提高数据分类性能,还能提高传统分类方法的泛化能力;在人造数据集和UCI真实数据集上的大量实验结果表明了该混合数据分类方法的有效性,其分类性能优于传统分类方法。

关键词:模式结构;复杂网络;高级分类方法;结构效率;节点重要性

A Hybrid Data Classification Method Based on Mining Information of Data Pattern Structure

Wang Huiyu[1], Gu Suhang[1,2]
(1. School of Information Engineering, Changzhou Vocational Institute of Light Industry, Changzhou 213164, China;
2. School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract: To the best of our knowledge, data are often correlated with other data in a dataset, and as a whole, a specific pattern structure is presented from all of the data. However, traditional classification methods (e.g., the support vector machine, SVM) do not take into account the correlation information between pair of data, and classification models are built just by taking advantage of the physical features (e.g., distance or similarity) of the input training data samples. Furthermore, data classification is realized by determining the similarities between the testing data samples and the built classification models in prediction phase. In order to solve the problem on the classification using the individual data information by traditional classification techniques, a hybrid data classification method based on mining the information of data pattern structure (HDCM) is proposed. The proposed classification method consists of two different types of classification techniques, on the one hand, the traditional classification methods based on using sole physical features of data are regarded as common classification methods, and on the other hand, the classification approach based on utilizing the information of data pattern structure is considered as advanced classification methods. In particular the proposed classification method not only has facility in effectively identifying the information of data pattern structure to enhance classification performance, but generalization ability of traditional classification approaches is promoted. A large number of experimental results on synthetic and UCI real-world datasets demonstrate the effectiveness of the proposed classification technique, and better classification performance can be obtained by the proposed classification technique in comparison to traditional classification methods.

Keywords: pattern structure; complex network; advanced classification approach; structural efficiency; node importance

0 引言

数据分类通过训练带有标签信息的样本生成分类模型以预测未标记样本的归属类别,是模式识别、机器学习、数据挖掘及统计学等领域最基本、最重要的问题之一。传统的数据分类方法,如支持向量机(Support Vector Machine, SVM)[1-3]、随机森林(Random Forest, RF)[4]、k近邻算法(k-Nearest Neighbor, kNN)[5]、决策树(C4.5)[6]以及朴素贝叶斯(Naive Bayesian, NB)[7]等,在训练阶段利用数据的物理特征(如距离、相似性等)构建数据分类模型,在分类阶段,通过确定测试样本与所建立数据分类模型之间的相似性预测测试样本的真实标签类型。在大多数情况下,传统的分类方法仅仅依靠数据之间的距离、相似度等物理特征信息构建数据分类模型,事实上,实际数据集中的每个数据并不是孤立的,数据之间存在关联,数据整体上都会呈现一定的模式结构,而且数据模式结构中蕴含着丰富的数据关联信息[8-10]。Thiago等[11]提出一种基于

收稿日期:2018-10-16; **修回日期**:2018-12-06。
基金项目:国家自然科学基金(81701793);常州市科技计划项目(CJ20160010);常州轻工职业技术学院博士基金(BSJJ13101010)。
作者简介:王惠宇(1977-),男,江苏常州人,硕士,讲师,主要从事计算机应用技术方向的研究。
通讯作者:顾苏杭(1989-),男,江苏盐城人,博士研究生,主要从事模式识别与人工智能、机器学习方向研究。

网络的高层次数据分类方法，该方法在建立的复杂网络中通过挖掘数据相互间的关联信息探索网络的同质性、聚集系数以及度等网络属性捕捉隐藏的数据拓扑结构信息，将数据拓扑结构信息与数据物理特征相结合形成一种智能分类方法；Sun 等[12]针对传统推荐系统并未考虑社交网络中各个用户之间的关系，提出社交正则化方法整合用户间的朋友等社交关系；Jiang 等[13]研究时尚、建筑及漫画等不同数据模式，针对现有大部分风格分类方法从数据局部模式中提取的鉴别特征过于多样化导致较差的分类性能，提出赋予不同特征相应权重的一致风格聚集自动编码策略学习鲁棒数据风格特征表示。

图 1 展示了传统分类方法用于实际数据分类过程中存在的不足。假设有一数据集包含三类数据 A、B 及 C，运用传统分类技术对这三类数据进行训练并构建数据分类模型。当向已建好的数据分类器输入测试样本 A1－t 时（图 1(b)），由于传统分类方法仅仅利用数据物理特征信息构建数据分类器，从颜色特征角度看，测试样本 A1－t 与 B1、C1 样本有着相同的颜色特征，它们之间有着极高的相似度，此时 A1－t 将被归为红色一类而不能获得真实的标签类型 A。如果在构建数据分类器的过程中还考虑到训练样本之间的模式结构关系，如从整体的角度看，A1、A2、A3 它们都是圆，共同组成圆类 A，它们之间的关联比较密切。将样本之间的关联信息用于数据分类模型的建立，构建的数据分类器将会正确地对测试样本 A1－t 进行分类。因此，将各种经典的分类技术用于实际数据分类时除了应考虑数据物理特征外还应有效地结合数据间的关联等这样一层模式结构关系，充分利用模式结构关系中数据间的关联作用信息，这样才能符合实际状况下数据分类并保证优越分类性能。

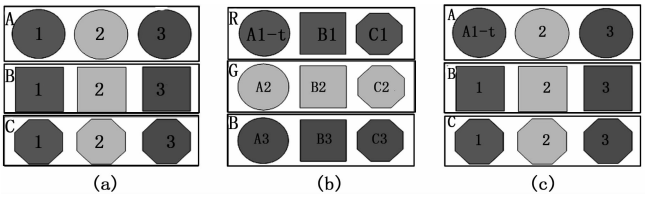

图 1　传统分类方法的分类过程

本文将仅仅利用数据物理特征信息的传统分类技术作为普通分类方法，将挖掘并采用数据关联信息的分类技术作为高级分类方法，基于这两种类型的分类方法，针对数据间相互关联的事实，提出一种挖掘数据模式结构信息的混合数据分类方法（HDCM）。HDCM 将输入的训练样本映射成复杂网络，在复杂网络中挖掘数据模式结构信息（网络节点效率、影响力）用于构建高级分类方法。使用任意一种传统分类方法以及高级分类方法分别计算测试样本对所有数据类型的隶属度，利用模糊分类技术将测试样本归为具有最大隶属度的数据类中，从而实现数据分类。由于 HDCM 考虑了数据关联信息，数据分类的泛化性能也有了明显提高。

1　高级分类模型描述

本文所提的数据分类模型由传统分类方法和高级分类方法混合而成，这里主要介绍构建高级分类模型的基础工作，包括构建 k 近邻复杂网络、确定有别于数据物理特征的数据模式结构特征：网络节点与子网络的效率以及节点影响力。

1.1　复杂网络

在建立复杂网络用于数据分类的所有方法中，基于 k 近邻算法的复杂网络是最常使用的方法[8,11,14]，且能够方便、简单地表达数据之间的关联，其过程可描述如下：对于输入的整个训练集 $X=\{x_1,x_2,\cdots,x_N\}$ 中某一样本 x_i，$x_i\in R^d$，选取与其距离最小的前 k 个样本 x_j，这里的距离为欧氏距离。如果样本 x_i 与样本 x_j 有相同标签，即 $L_{x_i}=L_{x_j}$，则样本 x_i 可关联于样本 x_j，记为 $x_i\rightarrow x_j$，对应于复杂网络则可建立节点 i 到节点 j 的有向边 e_{ij}，节点 i 为有向边 e_{ij} 的起始点，节点 j 为有向边 e_{ij} 的结束点。赋予复杂网络中不同有向边相应权重 ω_{ij}，使得当节点间的距离越小时权重 ω_{ij} 越大，权重 ω_{ij} 定义如下：

$$\omega_{ij}=\frac{1}{1+e^{d_{ij}}},1\leqslant i\leqslant N,1\leqslant j\leqslant k \tag{1}$$

其中：ω_{ij} 取值范围为（0，1），N 为复杂网络所有节点数，即训练样本总数，d_{ij} 为节点 i 与节点 j 之间的距离。

当输入的数据集包含 L 类数据，即 $C=\{c_1,c_2,\cdots,c_L\}$，由利用 k 近邻算法建立复杂网络的过程可知，建立的复杂网络包含 L 个子网络，即 $CN=\{cn_1,cn_2,\cdots,cn_L\}$，且子网络之间无关联，网络中每个节点 i 与样本 x_i 相对应。

1.2　模式结构效率特征

除了颜色、距离等物理特征信息外，数据的模式结构关系中蕴含着丰富的数据关联信息[15-17]，应该挖掘并将数据关联信息用于数据分类。如上述描述传统方法分类的例子中（图 1），如果仅依据颜色可将数据分为红、绿、蓝三类，建立的分类模型将不能正确分类测试样本 A1－t，若进一步考虑数据间的关联作用，可将数据分为圆、正方形、正六边形三类，按照 2.1 节可建立圆之间的连接、正方形之间的连接以及正六边形之间的连接三个子网络组成复杂网络，从而建立的分类模型可使得测试样本 A1－t 获得真实标签类型。赋予复杂网络中每个节点效率概念以区别网络中的其他节点，建立数据模式结构关系中的网络效率特征。社交网络中最常采用 PageRank 方法[18-19]计算网络节点的声誉，其基本思想是网络中某个节点连接其他节点数越多，说明该节点声誉越高；网络中其他节点连接某个节点越多，说明该节点声誉越高，本文复杂网络的节点效率计算方法正是源于 PageRank 方法。为了充分考虑节点之间的关联作用，对于复杂网络中节点 i 的效率定义如下：

$$\varepsilon_i^{out}=\begin{cases}\xi, & N_i=0\\ \dfrac{1}{N_i}\sum\limits_{i\rightarrow j}d_{ij}, & N_i>0\end{cases} \tag{2}$$

$$\varepsilon_i^{in}=\begin{cases}\xi, & N_k=0\\ \dfrac{1}{N_k}\sum_{k\to i}d_{ki}, & N_k>0\end{cases} \tag{3}$$

$$\varepsilon_i^{N_d}=\begin{cases}\xi, & N_d=0\\ \dfrac{1}{1+e^{\frac{1}{N_d}}}, & N_d>0\end{cases} \tag{4}$$

$$\varepsilon_i=\varepsilon_i^{out}+\varepsilon_i^{in}+\varepsilon_i^{N_d} \tag{5}$$

其中：N_i 代表以节点 i 为起始点的有向边个数，N_k 代表以节点 i 为结束点的有向边个数，N_d 代表节点 i 与其他节点相关联的有向边个数，即 $N_d=N_i+N_k$，ξ 为一较小值，赋予离群点或噪声点较小的效率，其对于分类样本所起的作用可忽略不计。

当计算出复杂网络每个节点效率后，与训练集每一类数据相对应的子网络 cn_l 效率便可确定，子网络效率定义如下：

$$\varphi_{cn_l}=\frac{1}{N_{cn_l}}\sum_{i\in cn_l}\varepsilon_i \tag{6}$$

其中：φ_{cn_l} 代表与训练集第 c_l 类数据相对应的子网络 cn_l 的效率，N_{cn_l} 为子网络 cn_l 包含的节点个数。复杂网络中节点及子网络的效率为基于挖掘数据模式结构信息的高级分类模型预测测试样本标签提供可靠依据，2.4 节将有详细内容介绍。

1.3 模式结构影响力特征

在利用数据模式结构信息建立高级分类模型的过程中，训练集中的每个数据样本对分类未标记测试样本所起的作用大小各不相同，有的数据样本对预测结果可能起决定性作用，有的数据样本影响力可能很弱[18-19]。这里定义复杂网络节点影响力如下：

$$In_j^{(h+1)}=\sum_{e_{ij}}\alpha In_i^h\omega_{ij}+(1-\alpha)\frac{1}{N} \tag{7}$$

其中：N 代表训练样本总数，$In_j^{(h+1)}$ 为复杂网络第 j 个节点在第 $h+1$ 次迭代过程中计算出的影响力大小，α 为复杂网络阻尼系数，根据文献 [20]，α 的最佳取值为 0.85，e_{ij} 代表利用 k 近邻算法构建复杂网络过程中建立的节点 i 到节点 j 的有向边。

公式（7）中 $1/N$ 表示训练样本是均匀分布的，而大多情况下实际数据集中的数据并不是均匀分布，每一个数据样本在一定距离范围内被不同个数的其他数据样本所包围[21]，类似的，复杂网络中的节点在一定距离范围内被不同个数的其他节点所包围，由此产生节点在整个网络中的浓度概念。复杂网络中第 i 个节点浓度定义为：

$$\rho_i=\frac{1}{N}\sum_j\chi(d_{ij}-dc) \tag{8}$$

其中：dc 代表截断距离，可根据实际的数据分类效果手动确定，或者使节点在 dc 距离范围内被占复杂网络节点总数 3%～5%的其他节点包围[21]，当 $d_{ij}-dc<0$ 时 $\chi(\cdot)=1$，否则 $\chi(\cdot)=0$。在复杂网络中以传播节点浓度的方式计算每个节点在整个网络中的真实影响力大小，定义如下：

$$\mathrm{In}_j^{(h+1)}=\sum_{e_{ij}}\alpha In_i^h\omega_{ij}+(1-\alpha)\rho_j \tag{9}$$

当满足以下迭代条件时计算节点真实影响力的迭代过程将会停止。

$$\sum_{j=1}^{N}\left\|In_j^{(h+1)}-In_j^h\right\|_2<\theta \tag{10}$$

其中：θ 的取值可根据实际数据集分类的效果手动选取，根据大量的实验结果表明 $\theta=10^{-4}$ 即可。

1.4 高级分类技术

经典的数据分类技术利用数据间的距离、相似性等物理特征实现数据分类，典型的方法如 SVM 及其改进方法。但是，实际数据集数据样本之间总会存在关联，当将数据集映射成复杂网络时这样的关联便显而易见，整体上数据样本具有一定的模式结构关系，并不是数据越靠近哪一类，它的标签就与该类相同，还应考虑数据的模式结构信息来确定数据的真实标签类型[8,22]。本文结合复杂网络在数据分类方面存在的优势，充分挖掘并利用蕴含在模式结构关系中的数据关联信息实现高级分类技术，定义如下：

$$\Lambda_{t,j}=\gamma\cdot\underset{j\in cn_l}{\varepsilon_{cn_l}}-d_{tj} \tag{11}$$

其中：ε_{cn_l} 代表子网络 cn_l 的效率，d_{tj} 为测试样本 t 与节点 j 间的欧氏距离，γ 为平衡系数，用于平衡数据物理特征和数据模式结构关系之间的作用，γ 越大则说明数据模式结构关系作用越大，反之则说明数据物理特征作用越大。

当输入一个未标记测试样本时，高级分类技术将依据 $\Lambda_{t,j}$ 确定未标记测试样本与每个子网络的连接集，定义如下：

$$\Omega_{cn_l}=\{j\mid j\in cn_l\&\Lambda_{t,j}>0\} \tag{12}$$

两种情况可将子网络 cn_l 中的节点 j 加入到连接集 Ω_{cn_l} 中：1）当测试样本与子网络 cn_l 中节点 j 的 $\Lambda_{t,j}$ 大于 0 时将节点 j 加入连接集 Ω_{cn_l} 中；2）当测试样本与每个子网络 cn_l 中节点的 $\Lambda_{t,j}$ 都小于 0 时，则将与最接近于 0 的 $\Lambda_{t,j}$ 对应的节点 j 加入到连接集 Ω_{cn_l} 中。高级分类模型将依据测试样本与子网络连接集影响力之和来判断测试样本标签类别，最大连接集影响力之和定义如下：

$$MAX_{\Omega_{cn_l}}=\operatorname{argmax}_{cn_l\in CN}\sum_{j\in\Omega_{cn_l}}In_j \tag{13}$$

高级分类模型将未标记测试样本归为与具有最大影响力之和的连接集所对应的类别中。

如图 2 所示演示了高级分类方法的详细分类过程。针对第 2 节高级分类模型的描述可知，高级分类方法涉及 3 个参数，即 k 近邻算法中的参数 k，截断距离 dc 以及平衡系数 γ。图 2 中 3 个参数分别设置为 $k=2$、$dc=3$ 及 $\gamma=0.3$。图 2（a）为利用 k 近邻算法建立的复杂网络，包含两个独立的子网络："■"类，标签为 0；"·"类，标签为 1。图 2（b）展示了节点的属性内容：部分节点之间的欧氏距离（如 $d_{12}=0.81$）及节点的度（如 $deg_2=3$），可用于计算节点的效率。图 2（c）为利用公式（2）～（5）计算出的节点效率（如 $\varepsilon_1=1.76$）及利用公式（6）计算出的子网络效率（如"■"类：$\varphi_0=1.57$）。图 2（d）展示了复杂网络中每个节点的影响力（如 $In_1=0.60$）；根据公式（11）可建

立测试样本（“▲”）与每个子网络的连接集，如图 2（e）所示。最终将测试样本归入到与具有最大连接集节点影响力之和对应的类中，如图 2（f）所示预测测试样本的标签类型为 0。

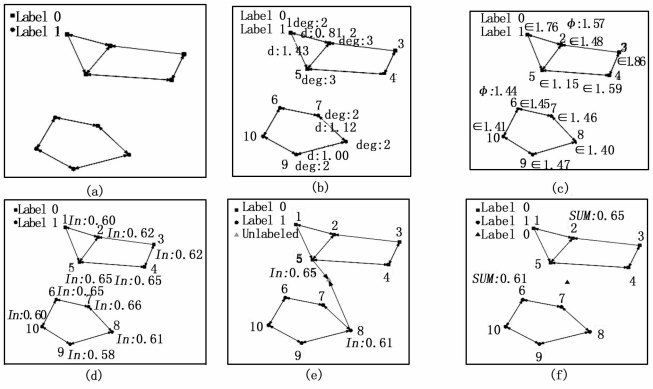

图 2　高级分类方法分类示例

2　混合数据分类方法

本文混合数据分类方法由普通分类方法和高级分类方法混合而成，一方面，普通分类方法（如 SVM、RF 及 kNN 等）依据数据的物理特征（如距离、相似性等）训练数据分类模型并预测测试样本的标签类型；另一方面，高级分类方法首先根据数据之间的关联作用将训练样本映射成复杂网络，在复杂网络中挖掘节点（每一个节点与数据样本相对应）的模式结构特征：节点及子网络效率和节点影响力，当输入一个测试样本时，根据高级分类技术（式(11)）建立测试样本与每个子网络的连接集，最终将测试样本归为与具有最大影响力之和的连接集相对应的类中。所提混合分类模型定义如下：

$$MC_t^{c_l} = \lambda E_t^{c_l} + (1-\lambda) F_t^{c_l} \tag{14}$$

其中：$MC_t^{c_l}$ 代表测试样本 t 对应于训练集中类 c_l 的隶属度，$E_t^{c_l}$ 代表采用任意一种传统分类方法（普通分类方法）所得的测试样本 t 对应于训练集中类 c_l 的隶属度，同样地，$F_t^{c_l}$ 代表采用高级分类方法所得的测试样本 t 对应于训练集中类 c_l 的隶属度。λ 为平衡系数，平衡普通分类方法与高级分类方法对分类结果所起的作用，当 $\lambda=1$ 时，普通分类方法对分类结果起绝对作用，反之亦然。当 λ 取值位于（0，1）时，为了能够取得最优的分类性能，λ 可由网格搜索结合交叉验证方法确定其取值。另外，$F_t^{c_l}$ 定义如下：

$$F_t^{c_l} = \frac{a_{cn_l} \sum_{j \in \Omega_{cn_l}} In_j}{\sum_{cn_l \in CN} a_{cn_l} \sum_{j \in \Omega_{cn_l}} In_j} \tag{15}$$

其中：a_{cn_l} 代表每个连接集影响力之和的重要性[11]，即连接集影响力之和越大，a_{cn_l} 的值越大，其取值范围为 [0，1]，且满足 $\sum_{cn_l=1}^{CN} a_{cn_l} = 1$，$CN$ 的大小等于训练集包含的类别数。由式（15）可知，采用高级分类方法所得的测试样本 t 对应于训练集中类 c_l 的隶属度 $F_t^{c_l}$ 即为测试样本 t 与训练集中某类 c_l 归一化的连接集影响力之和。

当输入一个未标记测试样本 t 时，可根据最大的隶属度 $MC_t^{c_l}$ 将 t 归入到相应的类 c_l 中，定义如下：

$$\overline{y_t} = \operatorname{argmax}_{c_l \in C} MC_t^{c_l} \tag{16}$$

其中：$\overline{y_t}$ 代表利用所提混合分类方法对测试样本 t 进行分类后预测的标签类型。

本文混合数据分类方法一方面能够在建立的复杂网络中探索并挖掘数据模式结构信息用于数据训练与分类；另一方面由公式（11）可知，从数据物理特征的角度，当一个测试样本的物理特征（如距离）与训练样本中的任何一类数据都不相似时，高级分类方法将起主要作用，从数据模式结构关系的角度，当一个测试样本的结构并不遵从训练样本中任何一类数据的结构关系时，普通分类方法将起主要作用。

3　实验与结果

为了验证所提混合数据分类方法的分类性能及其有效性，实验采用对比的方式将该方法与模糊 SVM[1]、模糊 C4.5[6]、加权的 kNN[23]、模糊分类方法 0－阶 TSK 及 1－阶 TSK[24-25] 分别在人造数据集以及 UCI 真实数据集上进行实验，通过实验结果与分析突出所提混合分类方法与传统分类方法的区别。其中，SVM 采用线性及高斯两种核类型的算法，为了公平起见，所有对比算法涉及的参数均采用网格搜索结合交叉验证的方法进行确定。所有对比算法均在 Matlab 软件平台上实现程序编写并在配置有处理器为 Intel（R）Core（TM）i3－3240、CPU 主频为 3.40 GHz、内存大小为 4.00 G、操作系统为 windows 7 ultimate system 的台式电脑上进行仿真。

3.1　高级分类方法

为了详细地了解所提高级分类方法的分类性能，组织 5 组高斯数据集实验，如图 3 所示，每组高斯数据集包含 3 类数据，具有各自的数据模式结构，3 类数据分别被标记为“·”类、“■”类及“▲”类，类之间有不同程度的交叉重叠，如图 3（e）所示的高斯数据集中 3 类数据的交叉程度已达到 80%，根据我们的知识和经验，这对于传统分类技术是一项十分具有挑战性的分类任务。

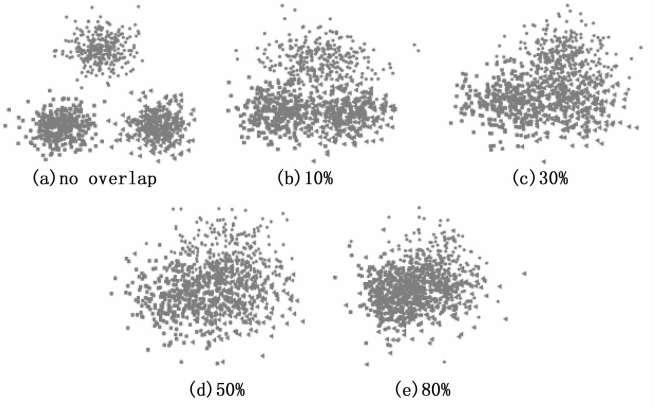

图 3　5 组高斯数据集

图 4 分别展示了利用高级分类技术对 5 组高斯数据集不同参数组合下的数据分类结果，其中，k 的取值范围为 [1，15][11]，截断距离 dc 使得复杂网络中每个节点被周围占节点总数 3%～5%的其他节点包围[21]，取值范围为 [0.01，0.1]，设定平衡系数 γ 的取值范围为 [0.1，1.5]。图中“Acc”代表分类精度，颜色条从下至上代表分类精度越来越高，所有实验结果均为运行程序 10 次后取得的平均结果。由图 4 实验结果可知，随着数据交叉程度的增加，数据分类精度逐渐降低，当数据交叉程度达到 80%，由于能够挖掘并利用数据模式结构信息，所提高级分类方法依然能够取得较高的分类精度（如图 4（e）所示的最高分类精度为 70%），充分彰显了所提高级分类方法鲁棒的分类性能。

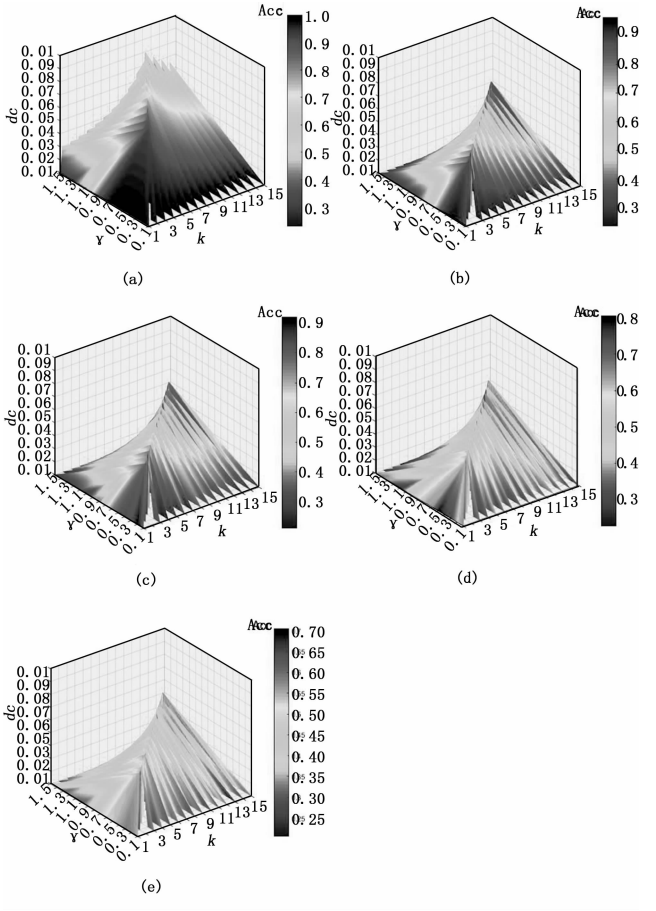

图 4 5 组高斯数据集不同参数组合下的分类结果

3.2 人造数据集仿真

挖掘并将数据模式结构信息用于数据分类的 HDCM 通过混合传统分类方法和高级分类方法两种类型的分类技术来弥补传统分类方法仅仅采用数据物理特征进行模型训练及分类的缺陷。HDCM 包含的两种不同类型分类技术在数据分类过程中所起的作用不同，如图 5 所示，当数据之间关联紧密，数据具有典型的模式结构时（蓝色“■”类），HDCM 在分类过程中将以高级分类方法为主导，即公式（14）中参数 λ 的取值偏大。这里将通过图 5 所示的数据集具体地演示参数 λ 如何平衡 HDCM 中两种不同类型分类器对数据分类所起的作用。图 5 所示的数据集“·”类包含 500 个样本，“■”类包含的样本数为 40，实验中选取广泛使用的 SVM 作为比较算法[1]，算法相关参数设置如下：对于线性 SVM，惩罚系数 $C=2^8$；高斯型 SVM 中惩罚系数 $C=2^8$，核宽度 $\sigma=2^{-3}$；混合分类方法中截断距离 $dc=1$，参数 $k=5$ 以及公式（11）中平衡系数 $\gamma=0.1$。表 1 记录了参数 λ 取不同值时采用不同分类方法计算的测试样本（“▲”）对于数据集中不同类数据的隶属度，其中，普通分类方法对应 Blue 列，HDCM 对应 Red 列。

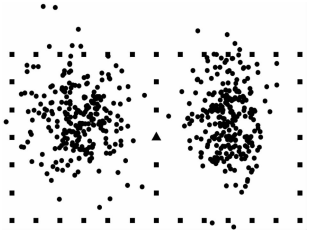

图 5 HDCM 的解释性示例

由图 5 可知，“·”类的样本数明显多于“■”类，且测试样本距离“·”类较近，如果使用传统分类方法，测试样本将被错误地归入到“·”类，即属于“·”类的模糊隶属度较大，如表 1 中当 $\lambda=0$。随着 λ 值逐渐变大，混合分类方法中传统分类方法的作用逐渐减弱，由于“■”类数据呈现明显的模式结构，且 HDCM 能够有效地挖掘数据之间的关联作用信息并用于数据分类，因此，HDCM 能够精确地预测测试样本的真实标签类型。结合图 5 和表 1 可知，当使用某种分类方法进行分类时，测试样本并不一定属于距离它较近的数据类，还应该考虑数据之间的关联。

表 1 不同 λ 值对分类的影响

Mthods	$\lambda=0$		$\lambda=0.5$		$\lambda=1$	
	Blue	Red	Blue	Red	Blue	Red
SVM(Linear)	0.07	0.93	0.48	0.52	0.02	0.98
SVM(Gaussian)	0.18	0.82	0.49	0.51	0.05	0.85

挖掘数据模式结构信息的混合数据分类方法在考虑数据物理特征的基础上，还通过构建复杂网络并探索数据的模式结构，并将数据模式结构信息用于数据分类。这里利用三组人造数据集来验证 HDCM 的数据分类性能。三组人造数据集分别为 Circles、Moons 以及 Rectangle，如图 6 所示，Circles 中三类包含的样本数分别为 2001、1001 及 601；Moons 中两类包含的样本数分别为 1001、501；Rectangle 中两类包含的样本数分别为 500、1000。每组数据集中的数据呈现明显的模式结构，分别为圆、月牙形以及长方形，不同数据类之间有重复交叉且包含不平衡样本数，即一类包含的样本数明显多于另一类，如 Moons 中左类样本数为 1001，而右类样本数只有 501，这样的数据集对于传统分类方法具有一定挑战性。

图 6　三组人造数据集

表 2　对比算法的人造数据集实验结果

数据集	方法	SVM(Linear) (C)	SVM(Gaussian) (C,σ)
Circles	单一	90.27±0.43 (2^{10})	89.58±0.11 (1,2^{12})
	混合	96.24±3.87 (4,0.2,0.9)	95.78±0.24 (8,0.2,0.9)
Moons	单一	92.33±1.98 (2^{11})	88.33±0.71 (2^{8},2^{10})
	混合	94.34±1.78 (5,0.1,0.8)	93.82±2.24 (6,0.1,0.8)
Rectangle	单一	95.00±1.52 (2^{9})	96.00±1.52 (2^{2},2^{9})
	混合	97.10±3.31 (4,0.2,0.9)	97.34±2.49 (5,0.3,0.8)

实验中，对于每一组人造数据集，随机选取样本总数的 80%作为训练样本，其余作为测试样本。仍然选取最为经典的分类方法 SVM 作为比较方法，这里使用模糊 SVM 方法[1]。针对 Circles、Moons 以及 Rectangle，HDCM 中截断距离 dc 大小具体设置为 0.7、0.1 及 0.2，算法涉及最优参数经网格搜索结合 5 折交叉验证的方法获得，具体参数设置如表 2 所示。实验所得数据为运行程序 5 次后的平均结果。

表 2 列出了所有对比算法在人造数据集上的详细数据分类结果，其中，“单一”表示只使用某一种传统方法进行数据分类，“混合”表示使用本文 HDCM 进行数据分类，分类精度及其标准差、算法最优参数分别表示为 * * ± * * （* *）。

由于图 6 三组人造数据集中的数据之间关联紧密，数据整体上呈现典型的模式结构，即使在发生明显数据重叠的情况下，使用本文所提的混合数据分类技术取得的分类结果普遍优于传统分类方法。人造数据集上的实验结果表明 HDCM 能够有效地挖掘数据之间的关联信息，也正因为将数据模式结构信息用于分类模型的训练及数据分类，使得 HDCM 具备良好的数据分类性能。

3.3　真实数据集仿真

除了人造数据集仿真实验，本文还将 HDCM 在 UCI 真实数据集[26]上进行实验，观察所提混合分类方法的实际分类性能。UCI 真实数据集的详细介绍如表 3 所示，其中，数据集中的样本数范围为 178 ～ 4174，最大和最小的数据特征维数分别为 3、18，数据集包含的类别数最小为 2，最大为 28。综上所述，所选取的真实数据集配置符合验证 HDCM 实际分类性能的需求。

表 3　UCI 真实数据集

数据集	样本数目	特征维数	类别数目
Contraceptive	1473	9	3
Abalone	4174	7	28
Wine	178	13	3
Haberman	306	3	2
Vehicle	846	18	4
Yeast	1484	8	10
Car	210	7	3

实验中，对于每一组真实数据集，随机选取样本总数的 80%作为训练样本，其余当作测试样本。所有对比算法参数设置作如下介绍：HDCM 算法共涉及四个参数，即高级分类方法中的 k、dc、γ 以及混合分类技术中用于平衡数据物理特征与模式结构关系特征作用的系数 λ 。由于截断距离 dc 使得复杂网络中的节点被占节点总数 3%～5%的其他节点包围，这里主要设置参数 k、γ 及 λ 。根据大量的实验结果，k、γ 及 λ 的取值可分别在 {1,2,…,14,15}、{0.1,0.2,…,2.9,3} 以及 {0,0.1,…,0.9,1} 范围内进行搜索，另外，针对参数 dc ，表 1 中的真实数据集从上往下分别设置为 3.3、0.08、2.9、4.1、0.6、0.2 以及 0.8。线性 SVM 中的惩罚系数 C 取值范围为 $\{2^{-3},2^{-2},\cdots,2^{11},2^{12}\}$ ，高斯型 SVM 的性能除了与惩罚系数 C 相关外，还与核宽度 σ 的设置有关，其取值范围为 $\{2^{-3},2^{-2},\cdots,2^{11},2^{12}\}$ 。加权的 k 近邻算法中参数 k 的设置与 HDCM 相同，其分类结果主要取决于测试样本与其所有近邻的加权之和，这里的权值大小为测试样本与其近邻之间欧氏距离的倒数。经典模糊分类方法 TSK 的数据分类性能主要与模糊规则数 R 及正则化参数 τ 相关，实验中这两个参数的取值搜索范围分别设置为 {5,10,…,195,200} 及 $\{10^{-5},10^{-4},\cdots,10^{4},10^{5}\}$ 。模糊 C4. 5[6]及对比算法的其他参数均采用默认设置。实验中的算法最优参数均由网格搜索结合 5 折的交叉验证方法确定，实验数据为运行程序 15 次后取得的平均结果，分类精度及其标准差、算法最优参数分别表示为 * * ± * * （* *）。表 4 给出的混合分类方法最优参数为 (k,γ,λ)，“－”代表参数的取值为空，表明 HDCM 中高级分类方法对分类结果未起作用。另外，为了探讨高级分类方法的实际分类性能，表 4 最后一列给出在 UCI 真实数据集上单一使用高级分类方法的分类效果，“－－－”表示无需使用 HDCM 进行分类。

如表 4 所示，通过对比算法在 UCI 真实数据集上的实验结果可得出以下几点分析：1）当传统分类方法与 HDCM 所取得的数据分类结果一致时，在混合分类技术分类过程中传统分类方法将起主导作用，HDCM 可智能地弱化高级分类方法的作用，即公式（14）中的参数 $\lambda = 0$ ，如高斯型 SVM 对于数据集 Vehicle、加权的 kNN 对于数据集 Contraceptive 等；2）当传统分类方法在真实数据集上所取得的分

表 4 UCI 真实数据集及人脸图像分类结果

数据集	方法	SVM(Linear) (C)	SVM(Gaussian) (C,σ)	Weighted kNN (k)	Fuzzy C4.5	0－阶 TSK (R,τ)	1－阶 TSK (R,τ)	高级方法 (k,γ)
Contraceptive	单一	49.58±3.26 (2^7)	53.40±1.24 ($1,2^8$)	57.82±4.36 (10)	48.97±2.44	45.92±0.73 ($15,10^3$)	45.58±0.56 ($55,10^4$)	54.82±1.57 (7,2.3)
	混合	56.15±2.27 (7,2.3,0.9)	59.48±0.63 (5,2.5,0.7)	57.82±4.36 (－,－,0)	55.68±0.46 (8,2.2,0.8)	53.71±0.44][(6,2.1,0.8)	53.62±0.24 (4,2.2,0.9)	－－－ －－－
Abalone	单一	25.62±2.88 (2^{10})	27.43±1.53 ($2^{-1},2^{12}$)	24.07±1.95 (13)	20.59±3.91	21.15±0.40 ($135,10^{-5}$)	26.19±0.67 (85,10)	30.56±1.18 (6,0.2)
	混合	31.72±5.84 (6,0.2,0.9)	33.57±2.74 (5,0.2,0.7)	34.25±0.56 (7,0.3,1)	35.46±2.15 (10,0.2,1)	32.66±0.87 (4,0.2,0.9)	36.29±1.06 (9,0.3,0.9)	－－－ －－－
Wine	单一	93.42±2.16 (2^{11})	95.58±0.94 ($2^2,2^{10}$)	93.85±4.83 (1)	97.14±1.22	52.38±1.35 ($15,10^{-1}$)	97.14±2.33 (55,10)	80.61±1.22 (5,1.5)
	混合	96.28±0.82 (5,1.6,0.9)	97.49±1.05 (6,2.2,0.8)	96.86±3.55 (8,1.6,0.6)	97.14±1.22 (－,－,0)	67.43±0.86 (11,1.8,0.9)	97.14±2.33 (－,－,0)	－－－ －－－
Haberman	单一	62.29±1.88 (1)	73.77±1.07 ($2^{-2},2^{-3}$)	76.32±5.77 (10)	55.73±6.72	66.67±1.55 (45,1)	81.97±0.64 (5,10)	80.98±0.27 (4,1.0)
	混合	73.63±0.12 (6,0.5,0.6)	78.53±0.01 (4,1,0.6)	80.33±2.12 (13,1.2,0.8)	69.02±0.05 (7,0.2,0.6)	67.21±0.21 (11,0.9,0.2)	83.61±0.96 (5,1.2,0.6)	－－－ －－－
Vehicle	单一	84.79±1.76 (2^9)	84.34±2.82 ($2^3,2^9$)	74.55±2.95 (1)	69.23±5.59	25.44±3.59 ($5,10^{-5}$)	75.15±1.12 ($5,10^{-3}$)	60.10±3.94 (3,0.8)
	混合	85.80±5.14 (9,0.8,0.2)	84.34±2.82 (－,－,0)	80.37±3.57 (3,0.7,0.9)	77.64±2.22 (7,0.8,0.9)	74.64±0.67 (7,0.9,1)	77.23±1.18 (5,0.5,0.4)	－－－ －－－
Yeast	单一	59.51±0.87 (2^7)	56.58±2.57 ($2^3,2^{11}$)	56.10±0.47 (15)	56.75±3.86	20.16±0.31 ($5,10^{-5}$)	39.53±8.64 ($250,10^{-2}$)	56.36±4.02 (7,0.3)
	混合	60.25±1.87 (5,0.2,0.7)	57.02±1.24 (7,0.2,0.8)	57.13±2.07 (4,0.2,0.4)	58.21±1.62 (9,0.2,0.7)	43.87±0.24 (6,0.2,1)	45.19±0.30 (3,0.2,0.7)	－－－ －－－
Seeds	单一	90.47±0.92 (2^5)	92.85±0.60 ($2^2,2^{11}$)	95.23±0.82 (1)	92.86±1.39	36.23±1.12 ($65,10^{-5}$)	95.24±2.13 ($140,10^{-5}$)	96.38±3.97 (10,1.9)
	混合	94.57±7.93 (7,2,0.4)	95.69±6.66 (8,1.8,0.3)	95.23±0.82 (－,－,0)	97.62±6.07 (4,2.2,0.3)	93.79±2.61 (5,2,1)	97.62±5.79 (3,1.7,0.5)	－－－ －－－
Face	单一	51.35±0.45 (2^{11})	53.41±0.52 ($2^5,2^{12}$)	52.32±1.97 (1)	46.89±2.68	17.35±0.75 ($125,10^{-3}$)	37.42±1.03 ($295,10^{-2}$)	54.30±1.12 (3,2.2)
	混合	63.21±0.13 (7,1.9,0.9)	64.48±0.11 (5,2.1,0.7)	63.70±2.89 (4,2.4,0.6)	63.79±1.12 (9,2.7,1)	61.17±1.16 (6,2.6,1)	63.88±1.57 (4,2.8,1)	－－－ －－－

类精度较低时，公式（14）中参数 λ 的值将等于或接近 1，HDCM 中的高级分类方法将对预测测试样本的标签类型起决定性作用，如线性 SVM 对于数据集 Abalone、加权的 kNN 对于数据集 Contraceptive、模糊 C4.5 对于数据集 Abalone 等；3）对于每一组真实数据集，混合分类方法都给出了不同的 γ 值，表明数据集中数据之间的确存在关联作用信息，且所提方法能够有效挖掘并利用这些不同于数据物理特征的数据信息来提高传统分类方法的分类性能；4）当单一使用高级分类方法时，通过与普通分类方法相比较，高级分类方法表现出了具有竞争力的分类性能，表明挖掘并使用数据模式结构信息确实能够有助于改善分类方法的性能。

表 5 给出了两种典型的传统分类器与所提分类技术在数据集 Wine、Contraceptive 以及 Haberman 上的算法运行时间对比。由表 2 结合表 4 可知 HDCM 分类精度均高于普通分类方法，但由于所提混合数据分类方法结合普通分类方法与高级分类方法，因此，从算法复杂度角度，HDCM 并不占明显优势。

表 5 算法运行时间分析

方法	方法	Wine	Contraceptive	Haberman
SVM(Linear)	单一	* (0.0000)	0.7762 (0.0094)	0.0062 (0.0135)
	混合	0.0236 (0.0079)	1.9268 (0.0112)	0.2002 (0.0137)
Fuzzy C4.5	单一	0.0086 (0.0110)	0.0305 (0.0287)	0.0174 (0.0148)
	混合	0.0293 (0.0095)	1.3828 (0.0235)	0.2156 (0.0200)

3.4 工业应用案例

本文还进行工业应用案例分析，将 HDCM 应用于人脸

识别。如图 7 所示，选取的 6 组人脸图像来自 Pointing’04 ICPR Workshop[27]，它所包含的人脸图像均为基准的人脸识别数据集。每一组人脸图像包含 15 幅序列图像，图像中的人脸姿势以 15°的间隔在［－90° 90°］范围内变化，实验中选取序列图像的前 7 或者后 7 幅图像组成人脸图像数据集。每一幅人脸图像的分辨率定为 80（120，且利用主成分分析法（Principle Component Analysis，PCA）对图像特征进行降维[28]，根据实验效果维度大小设置为 30。实验中选取每一组人脸图像的前 5 幅作为训练样本，其他图像作为测试样本。由图 7 可知，由于每个人脸的特征不同（如发型、面部表情等），且每个人脸姿势或朝右或朝左，因此，对应于每个不同人脸的数据整体上会呈现明显的模式结构，十分适合验证挖掘并利用数据模式结构信息的混合分类方法的有效性及其分类性能。实验中，HDCM 的参数 $dc = 6$，对比算法给出的所有最优参数均由网格搜索结合 5 折的交叉验证方法获得，实验数据为运行程序 15 次后所取的平均结果（表 4 最后一行数据）。

图 7　人脸识别数据集

由实验结果可知，SVM 等传统分类方法因在构建分类模型以及分类的过程中依赖单一的数据物理特征而忽略了数据之间存在关联信息的事实，在人脸识别数据集上的分类精度明显低于所提的混合分类方法，尤其当使用 0－阶 TSK 及 1－阶 TSK 模糊分类方法时实验对比效果更加明显。人脸识别数据集上的对比实验结果充分证明了 HDCM 不仅能够挖掘数据之间的关联信息、识别数据的模式结构关系，而且可有效地结合传统分类方法和高级分类方法两种不同类型的分类技术进行数据分类。

4　结束语

数据集中数据之间往往存在关联，数据并不是孤立的存在，在构建数据分类模型以及分类的过程中应考虑这样一种有别于数据物理特征的数据关联信息。本文所提的混合数据分类方法一方面兼顾了数据的物理特征，另一方还能够有效地识别数据的模式结构，并将数据之间的关联作用信息用于训练数据分类模型及数据分类。人造数据集及真实数据集上的仿真实验结果证明了 HDCM 的有效性，HDCM 实际分类性能优于传统的分类方法。实验中发现，HDCM 还能够解决数样本比例不平衡情况下的数据分类[29]，如人造数据集 Moons 及真实数据集 Yeast，样本比例分别为 2、2.46，因此，在今后的工作中将对此作进一步研究。另外，根据图论知识，一个复杂网络除了节点的度等常见属性外，还包含有同质性、聚类系数等[30]，如何将除了度之外其他属性结合进来探索复杂网络局部与全局特征作为数据分类的辅助信息[31]也将是今后的研究内容。

参考文献：

[1] Lin C F，Wang S D. Fuzzy support vector machines［J］. IEEE Transactions on Neural Networks，2002，13（2）：464－471.

[2] 张　进，丁　胜，李　波. 改进的基于粒子群优化的支持向量机特征选择和参数联合优化算法［J］. 计算机应用，2016，36（5）：1330－1335.

[3] 李　平，徐　新，董　浩，等. 利用可分性指数的极化 SAR 图像特征选择与多层 SVM 分类［J］. 计算机应用，2018，38（1）：132－136.

[4] Ho T K. The Random subspace method for constructing decision forests［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence，1998，20（8）：832－844.

[5] Keller J M，Gray M R，Givens J A. A fuzzy k－nearest neighbor algorithm［J］. IEEE Transactions on System，Man and Cybernetics，1985，SMC－15（4）：580－585.

[6] Olaru C，Wehenkel L. A complete fuzzy decision tree technique［J］. Fuzzy Sets and Systems，2003，138（2）：221－254.

[7] Bustamante C，Garrido L，Soto R. Fuzzy naive bayesian classification in robosoccer 3D：A hybrid approach to decision making［A］. RoboCup 2006：Robot Soccer World Cup X［C］. DBLP，2006：507－515.

[8] Thiago C S，Zhao L. High－level pattern－based classification via tourist walks in networks［J］. Information Science，2015，294：109－126.

[9] Madeleine S，Andreas M，Stefan K，et al. Extracting information from support vector machines for pattern－based classification［A］. Proceedings of the 29th Annual ACM Symposium on Applied Computing［C］. ACM，2014：129－136.

[10] Chu W T，Wu Y L. Image style classification based on learnt deep correlation features［J］. IEEE Transactions on Multimedia，2018（99）：1－13.

[11] Thiago C S，Zhao L. Network－based high level data classification［J］. IEEE Transactions on Neural Networks and Learning Systems，2012，23（6）：954－970.

[12] Sun Z B，Han L X，Huang W L，et al. Recommender systems based on social networks［J］. The Journal of Systems and Software，2012，99：109－119.

[13] Jiang S H，Shao M，Jia C C，et al. Learning consensus representation for weak style classification［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017（99）：1－14.

[14] Pedronette D C G，Guilherme I R. Unsupervised manifold learning through reciprocal kNN graph and connected components for image retrieval tasks［J］. Pattern Recognition，2018，75（3）：161－174.

（下转第 217 页）