doi:10.6041/j.issn.1000-1298.2018.09.003

基于半监督主动学习的菊花表型分类研究

袁培森[1]　任守纲[1,2]　翟肇裕[3]　徐焕良[1,2]

(1. 南京农业大学信息科学技术学院, 南京 210095; 2. 国家信息农业工程技术中心, 南京 210095;
3. 马德里理工大学技术工程和电信系统高级学院, 马德里 28040)

摘要: 鉴于人工和专家分类模式的局限性,基于表型的菊花分类存在效率低下的问题。本文采用基于半监督主动学习技术,在已分类菊花数据的基础上,利用未标号菊花样本数据提供的信息,建立了菊花表型分类模型,提升了分类质量和效率。该模型可以不依赖外界交互,利用未标号样本来自动提升菊花分类的质量。为了训练学习模型,本文收集了菊花的表型特征数据,标注了菊花表型类别,并研究了菊花分类属性特征的编码技术。在此数据集上,采用基于图标号传播的半监督学习技术对未标号的菊花数据进行建模,为了提升半监督分类的有效性,在标号传播的基础上使用主动学习技术,采用熵最大策略来选择难以识别的样本,以改进分类质量。在该数据集上进行了试验验证,并进行了试验对比和分析,试验结果表明,本文方法能够较好地利用未标号菊花样本提升分类的精度,随着标号百分比从 6.25% 升至 23%,识别精度达到 0.7 以上,标号百分比在 81.25% 时,平均识别精度和召回率分别达到 0.91 和 0.88。

关键词: 菊花表型分类; 半监督学习; 图模型; one - hot 编码; 主动学习; 熵最大化

中图分类号: TP391　　**文献标识码**: A　　**文章编号**: 1000-1298(2018)09-0027-08

Chrysanthemum Phenotypic Classification Based on Semi-supervised Active Learning

YUAN Peisen[1]　REN Shougang[1,2]　ZHAI Zhaoyu[3]　XU Huanliang[1,2]

(1. *College of Information Science and Technology*, *Nanjing Agricultural University*, *Nanjing* 210095, *China*
2. *National Engineering and Technology Center for Agriculture*, *Nanjing* 210095, *China*
3. *Superior School of Technical Engineering and Telecommunication Systems*, *Technical University of Madrid*, *Madrid* 28040, *Spain*)

Abstract: Phenotype-based classification plays an essential role in plant research. Chrysanthemum flower has great momentous economic value and medicinal value, and has feature of morphological and genetic diversity as well. Due to the limitations of the artificial classification model by expert and the characteristic of genetic diversity, phenotype-based classification has been facing great challenges for its research. At present, the technologies and applications of machine learning and artificial intelligence are developing rapidly. With the vehicle of machine learning, the semi-supervised learning technology was employed to provide an effective way for improving the classification performance. This method was based on label propagation of graph model as well as active learning technique. According to this method, a small number of classified chrysanthemum data as well as a large amount of unlabeled chrysanthemum samples were exploited to improve the classification accuracy. This method can automatically make use of the unlabeled samples to improve the quality of chrysanthemum classification without relying on external interactions. The chrysanthemum phenotypic data was collected to train the learning model, and manually annotate the chrysanthemum category information. For exploiting the categorical attribute, the coding skill was studied as well. The label propagation of graph model was utilized by the semi-supervised learning skill for the unlabeled chrysanthemums. In order to improve the effectiveness of semi-supervised classification, active learning technique was applied, which was based on the entropy maximization strategy to select difficult-to-identify samples to improve classification performance further. Extensive

收稿日期: 2018 - 03 - 24　修回日期: 2018 - 05 - 05

基金项目: 国家自然科学基金项目(61502236)和中央高校基本科研业务费专项资金项目(KYZ201752、KJQN201651)

作者简介: 袁培森(1980—),男,讲师,博士,主要从事数据挖掘和海量数据处理与分析研究,E-mail: peiseny@njau.edu.cn

通信作者: 徐焕良(1963—),男,教授,博士,主要从事农业信息化与大数据技术研究,E-mail: huanliangxu@njau.edu.cn

experiments were conducted and comparisons were made. The experimental results showed that the unlabeled chrysanthemum samples can improve the classification accuracy remarkably, with the labeled ratio increasing from 6.25% to 23%, the recognition accuracy rapidly reached 0.7, the average recognition accuracy and recall rate can reach 0.91 and 0.88, respectively, when the labeled ratio was 81.25%. In conclusion, semi-supervised based learning for the intelligent identification and effective management of chrysanthemum flowers had great significance in theory and application for the studying of chrysanthemum phenotype.

Key words: chrysanthemum phenotype classification; semi-supervised learning; graph model; one - hot encode; active learning; entropy maximum

0 引言

植物表型的研究在植物保护、育种等领域具有重要的研究价值[1]。形态学性状(表型)最易观察与测定,并且最有生态适应性,至今仍是形态学研究最重要的基本方法。

菊花作为一种重要的花卉,具有种类多、品种形态变异丰富、分布广泛等特点,这些为菊花品种资源调查、品种分类与鉴定等带来了困难[2]。目前世界范围内有 20 000 ~ 30 000 个菊花品种,大量关于菊花分类的研究见文献[3 - 6]。菊花表型性状和各种遗传标记之间具有强关联性,其研究可为菊花品种分类向更为精确的数量分类方向发展奠定基础。

花卉品种分类是以形态学为基础,诸多分类研究也均以形态学为基础[4-7]。菊花的形态学分类和管理主要通过观察和测量表型性状,该方法直接、简单,但是面对庞大的菊花品种和数量,效率不高。

型态分类一般需要使用特殊设备或通过耗时的试验进行,但是其中存在海量的无类标签的样例[8]。此类应用具有类标签的样本少、无类标签样本过剩的特征,如何利用这些无标签数据提升分类的效率成为信息领域研究重点。半监督学习(Semi-supervised learning,SSL)[9-13]是利用未标记样本提升分类模型泛化能力的分类技术。半监督学习在少量带有标记的数据基础上,利用大量未标记的数据进行训练。半监督学习将未标记数据与少量标记数据结合,显著地提高学习效果,从而降低获取未标记数据标签的代价。鉴于半监督学习对利用未标记样本的有效性,其研究和应用成为近年来机器学习领域的热点[8,10,14-15]。

目前,研究者在植物表型方面的研究及应用等方面做了很多有意义的工作。CRIMINISI 等[16]提出了基于机器学习技术的模型,用于表型识别,其中就涵盖了半监督方法。BEAULIEU-JONES 等[15]用半监督学习技术进行表型分层研究。SINGH 等[17]使用半监督度量学习识别核表型。DOOSTPARAST 等[18]提出基于半监督学习的基因组数据整合和表型分类研究。SCHARR 等[19]研究了基于机器学习的植物表型分叶技术。ZHANG 等[14]使用半监督学习技术用于植物叶片识别。LEVATI 等[8]指出,半监督学习在表型特征预测中的研究中具有重要作用。AGARWAL 等[20]使用带噪声的标签数据学习表型统计模型。

本文基于主动学习[21]和图半监督学习技术[18,22-23],首先利用标签数据和未标签数据构建图模型,并构建标记矩阵;在主动学习阶段通过计算熵最大化样本作为标记样本,重新进行训练,提升分类效果,并在菊花数据集上进行试验和对比分析。

1 菊花特征及编码

1.1 菊花特征

菊花数据集的表型特征包括:株高(cm)、花瓣类型、花径(cm)、舌状数和花期。其中,花期包括夏菊、秋菊和寒菊。花瓣共分为 5 种类型:平瓣、匙瓣、管瓣、桂瓣和畸瓣。5 种菊花类型示例图如图 1 所示。

1.2 one - hot 编码

one - hot 编码,又称为一位有效编码,主要针对离散特征进行编码[24-25]。它采用 t 位状态寄存器对 t 个状态进行编码,每个状态都有独立的寄存器

图 1 5 种菊花类型示例图

Fig. 1 Illustrations of five chrysanthemum type images

位，并且在任意时候只有一位有效。

one－hot 编码是把分类变量作为二进制向量来表示[24－25]。它将分类值映射到整数值，然后，每个整数值被表示为二进制向量，除了整数的索引被标记为 1，其余都为 0。

例如，对于花型类别取值“平瓣”、“匙瓣”、“管瓣”、“桂瓣”、“畸瓣”，其编码如图 2 所示。每一行对应一个类别，第 1 行的编码[1 0 0 0 0] 为“平瓣”类，依此类推。

$$\begin{bmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&0&0&1\\ 0&0&1&0&0\\ 0&0&0&1&0 \end{bmatrix}$$

图 2　5 种菊花类型的 one－hot 编码

Fig. 2　Illustration of one－hot code of five chrysanthemum type

one－hot 编码后的离散特征可以作为连续特征进行计算。该方法的优点是能够处理非连续型数值特征，在一定程度上扩充了特征，使得模型具有更好的鲁棒性。

2　半监督学习和主动学习

2.1　半监督学习

半监督学习的基本设置是给定一个未知分布的样本集 $S=D_L\cup D_U$，其中有标记样本集 $D_L=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l)\}$ 以及一个未标记样本集 $D_U=\{\boldsymbol{x}_{l+1},\boldsymbol{x}_{l+2},\cdots,\boldsymbol{x}_{l+u}\}$，求解模型 $f:X\rightarrow Y$ 可以准确地对示例 $\boldsymbol{x}$ 预测其标记 y。这里 $\boldsymbol{x}_i\in X$ 为 d 维向量，$y_i\in Y$ 为样本 $\boldsymbol{x}_i$ 的标记，l 和 u 分别为 D_L 和 D_U 所包含的样本数。

半监督学习是在样本集 S 上寻找最优的分类器，目的是利用带标签数据和未标记数据学习设计分类模型，使得该模型比仅使用带标签数据分类性能更好。

MILLER 等[26]从理论方面定性地指出可以利用未标记示例来提升模型分类性能，并从数据分布估计角度给出了直观的分析。为了利用未标记样本，需要将未标记样本的数据分布信息与已标记类别相联系的假设。最常见的假设包括聚类假设和流形假设[27]。其中聚类假设的数据倾向于形成离散的集群，并且同一集群中的点更可能共享类标签。流形(Manifold)假设是假设数据分布在一个流行结构上，邻近的样本具有相似的输出值。

目前，半监督分类学习方法包括生成式方法、半监督支持向量机、基于图的半监督等[27]。

基于图(Graph)的半监督学习方法[22－23]使用图建模，每个标记和未标记的示例都采用结点表示，样本之间的相似度或相关性用图的边表示。该图可以使用领域知识或样本的相似性来构建，具有较强的鲁棒性[23]，适用于本文的菊花多类分类问题。

2.2　主动学习策略及熵最大化

主动学习(Active learning)[21,28]是半监督学习的特例，其中学习算法能够交互式地利用其他信息源以在新数据点处获得期望输出。主动学习目的是通过尽量少的标签提高准确率。此外，在多标号分类应用中，异常点的标号对分类器准确率带来了很大的挑战[29]。本文采用基于图模型的半监督学习与主动学习相结合的方法，提升菊花分类的性能，并借助于信息理论领域熵的概念和性质，提升菊花多标签分类的鲁棒性[21,29]。

熵(Entropy)表示随机变量的不确定性，变量的不确定性越大，熵也越大。设随机事件 R，它有 r_1，$r_2,\cdots,r_m$ 共 m 个不同的结果，每个结果出现的概率分别为 $p_1,p_2,\cdots,p_m$，那么 R 的信息熵表示为

$$H(E)=-\sum_{i=1}^{m}p_i\mathrm{lb}p_i \tag{1}$$

对于一个分类系统来说，假设类别 C 可能的取值为 $c_1,c_2,\cdots,c_k$(k 为类别总数)，每一个类别出现的概率分别是 $p_1,p_2,\cdots,p_k$。此时，该分类系统的熵可以表示为

$$H(C)=-\sum_{i=1}^{k}p(c_i)\mathrm{lb}p(c_i) \tag{2}$$

在主动学习阶段，通过计算传播标签分布使得式(2)熵最大化的样本，即选择分类器最难分类的样本，远离标记的样本作为标记样本放入训练集，提升分类的性能。

3　基于半监督主动学习过程

3.1　处理过程

基于图模型的半监督学习(Graph-based semi-supervised learning, GSSL)[22－23]采用标签传播(Label propagation)原理，根据样例之间的几何结构构造图模型，用图的结点(Vertice)表示样例，利用图结点的邻接关系将类标签从已标签的样本向无类标签的样本传播来实现分类。

本文采用 k-最近邻法 (k-nearest neighbor, k-NN)构建图模型，每个点与最近的 k 个点相连。近邻矩阵基于归一化图拉普拉斯算子的近邻矩阵以及在标签上进行软锁定[27]，系统处理过程如图 3 所示。

首先，给定菊花标号数据集 $D_L=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l)\}$ 和未标号菊花数据集 $D_U=$

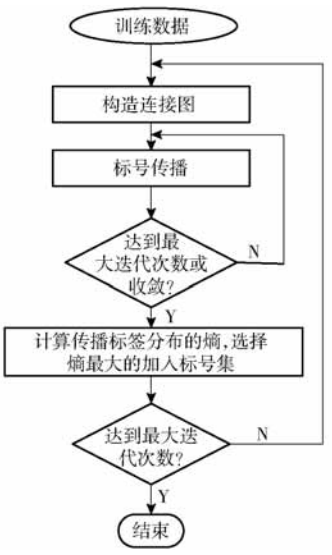

图 3 系统处理过程

Fig. 3 System processing procedure

$\{\boldsymbol{x}_{l+1},\boldsymbol{x}_{l+2},\cdots,\boldsymbol{x}_{l+u}\}$,$l<u$,$l+u=M$。假定 $y_i\in Y$,首先基于 $D_L\cup D_U$ 构建图 $G=(V,E)$,其中 V 表示结点集,E 表示边集,$V=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l,\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_{l+u}\}$,$w_{ij}\in E$ 表示顶点 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 之间的相似性距离,计算公式为

$$w_{ij}=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/(2\sigma^2)) \tag{3}$$

其中 $i,j\in\{1,2,\cdots,m\}$,$\sigma>0$ 是高斯带宽函数常数,$\|\cdot\|$表示向量范数。图 G 可以表示为一个相似矩阵 $\boldsymbol{W}_{M\times M}$,相似矩阵的每一项通过式(3)定义的高斯带宽函数计算,标签传播原理是通过节点之间的边传播类标签。边的权重越大,表示两个节点越相似,那么标签越容易传播过去。定义一个 $M\times M$ 的概率转移矩阵 $\boldsymbol{P}$,计算公式为

$$P_{ij}=\boldsymbol{P}(i\rightarrow j)=\frac{w_{ij}}{\sum_{k=1}^{M}w_{ik}} \tag{4}$$

其中,P_{ij}表示从节点 i 转移到节点 j 的概率。假设有 C 个类和 l 个已标号样本,定义标号矩阵 $\boldsymbol{L}_{l\times C}$,给 u 个未标号样本一个标号矩阵 $\boldsymbol{U}_{u\times C}$。合并两个矩阵可得到矩阵 $\boldsymbol{F}=[\boldsymbol{L}\ \boldsymbol{U}]$,$\boldsymbol{F}$ 大小为 $M\times C$,其中 $M=l+u$。基于图正则化的优化目标函数可以表示为

$$J=\min_f\sum_{i\in\boldsymbol{L}}(y_i-f_i)^2+\lambda\sum_{i,j\in\boldsymbol{L},\boldsymbol{U}}w_{ij}(f_i-f_j)^2 \tag{5}$$

其中,假设样本具有相似的预测函数,f 是菊花数据集 $D_L\cup D_U$ 上的预测函数。式(5)中第 1 项为最小化标签数据上的损失,第 2 项为确保在标号数据和未标号数据集上的标号光滑性[27]。

对角矩阵 $\boldsymbol{D}=\mathrm{diag}(d_1,d_2,\cdots,d_{l+u})$ 的对角元素 $d_i=\sum_{j=1}^{M}W_{ij}$。定义一个 $M\times|Y|$ 的非负标记矩阵 $\boldsymbol{F}=[\boldsymbol{F}_1\ \ \boldsymbol{F}_2\ \ \cdots\ \ \boldsymbol{F}_M]^{\mathrm{T}}$,其第 i 行元素 $\boldsymbol{F}_i=(F_{i1},F_{i2},\cdots,F_{i|Y|})$ 为示例 $\boldsymbol{x}_i$ 的标记向量,则相应的分类规则为

$$y_i=\arg\max_{1\leqslant j\leqslant|Y|}F_{ij} \tag{6}$$

对 $i=1,2,\cdots,M$ 和菊花类别 $j=1,2,\cdots,C$,将 $\boldsymbol{F}$ 初始化为

$$\boldsymbol{F}(0)=Y_{ij}=\begin{cases}1 & (1\leqslant i\leqslant l\ \text{且}\ y_i=j)\\ 0 & (\text{其他})\end{cases} \tag{7}$$

基于 $\boldsymbol{W}$ 构造一个标记传播矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$,其中 $\boldsymbol{D}^{-\frac{1}{2}}=\mathrm{diag}(d_1^{-\frac{1}{2}},\cdots,d_{l+u}^{-\frac{1}{2}})$,通过迭代计算可得

$$\boldsymbol{F}(t+1)=\alpha\boldsymbol{L}\boldsymbol{F}(t)+(1-\alpha)\boldsymbol{Y} \tag{8}$$

其中 $\alpha\in(0,1)$ 为用户指定的夹紧因子,用于平衡标记传播项 $\boldsymbol{L}\boldsymbol{F}(t)$ 与初始化项 $\boldsymbol{Y}$ 的重要性。这个夹紧因子可以放宽,对于 $\alpha=0.1$,即保留原始标签分布比例为 90%。基于式(8)迭代至收敛最终可得

$$\boldsymbol{F}^*=\lim_{t\rightarrow\infty}\boldsymbol{F}(t)=(1-\alpha)(I-\alpha S)^{-1}\boldsymbol{Y} \tag{9}$$

由式(9)中的 $\boldsymbol{F}^*$ 可获得未标号数据集 D_U 中样本的标记为 $(y_{l+1},y_{l+2},\cdots,y_{l+u})$。

3.2 处理算法

本文算法是在基于图模型半监督学习基础上[30]增加了熵最大化策略。首先输入标记样本集 $D_L=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l)\}$、未标记样本集 $D_U=\{\boldsymbol{x}_{l+1},\boldsymbol{x}_{l+2},\cdots,\boldsymbol{x}_{l+u}\}$、构图参数 σ、参数 α、半监督迭代次数 S_i、收敛阈值 τ 和主动学习迭代次数 A_i。算法步骤如下:

```
while(A_i > 0){
  计算相似矩阵 W;
  基于相似矩阵 W 构造标记传播矩阵
    L = D^{-1/2} W D^{-1/2};
  根据式(7)初始化 F(0);
  t = 0;
  while(S_i > 0 && |F(t+1) - F(t)| > τ){
    F(t+1) = αLF(t) + (1-α)Y;
      t++; S_i--;
  }
for i = l+1, l+2, ..., l+u do
y_i = arg max_{1≤j≤C} (F*)_ij;
计算未分类中样本熵并排序,选取熵最大化的 E 个样本加入已标号序列;
  A_i++;
}
```

最后,算法输出未标记样本的预测结果 $(y_{l+1},y_{l+2},\cdots,y_{l+u})$。

4 试验结果与分析

4.1 试验平台

试验平台为 Windows 10 系统,16 GB 内存,

QuadCore Intel Core i7 - 7700，4.2 GHz。算法采用 Python 3.6 实现。

测试菊花表型数据集共有 160 条，菊花特征包括：株高(cm)、花型、花径(cm)、舌状数和花期。其中，花型包括平瓣、匙瓣、管瓣、桂瓣和畸瓣，类别标号分别记为 1、2、3、4、5。花期包括夏菊、秋菊和寒菊[4]。

数据集分布如图 4 和图 5 所示。图 4 是数据集中菊花株高、花径和舌状数概率密度分布图。图 5 是数据集中不同菊花类别的株高与舌状数分布散点图。

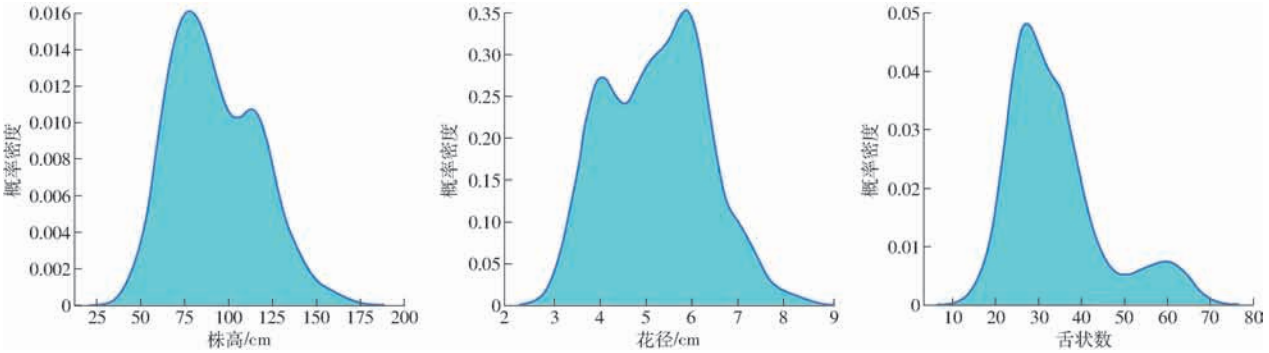

图 4　菊花株高、花径和舌状数概率密度分布图

Fig. 4　Probability density distribution of flower height, diameter and ligule number of dataset

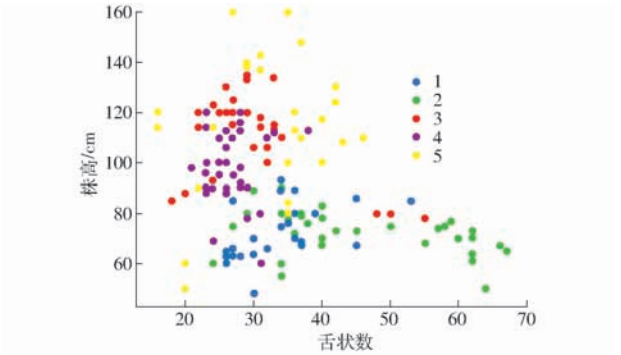

图 5　不同菊花类别的株高与舌状数分布散点图

Fig. 5　Scatter plot of plant category with plant height and ligule number

参数设置如下：夹紧因子 α 为 0.2，即始终保留原始标签分布的比例为 80%。参数 λ 的默认值是 $1/C$（C 是菊花类别个数），共 5 类菊花，因此设置为 0.2。k-NN 中构建图的近邻数 k 设置为 5。半监督迭代次数 S_i 设置为 6，收敛阈值 τ 设置为 0.01，主动学习迭代次数 A_i 设置为 10。

4.2　分类精度指标

试验从半监督分类预测结果的质量分类效率进行了系统的研究和分析。定义 T_P（True positive）为模型预测为正的正样本，F_P（False positive）为模型预测为正的负样本，F_N（False negative）为模型预测为负的正样本。分类结果从 3 方面进行度量：加权精度（Weighted precision）、加权召回率（Weighted recall）和加权 F_1 度量，定义分别为

$$p_w = \frac{\sum_{i=1}^{C} s_i \rho_i}{\sum_{i=1}^{C} s_i} \tag{10}$$

$$r_w = \frac{\sum_{i=1}^{C} s_i r_i}{\sum_{i=1}^{C} s_i} \tag{11}$$

$$F_1 = \frac{\sum_{i=1}^{C} s_i f_i}{\sum_{i=1}^{C} s_i} \tag{12}$$

其中

$$\rho_i = \frac{T_P^i}{T_P^i + F_P^i}$$

$$r_i = \frac{T_P^i}{T_P^i + F_N^i} \quad f_i = 2\frac{\rho_i r_i}{\rho_i + r_i}$$

式中　s_i——第 i 个类别的支持度

i——菊花类别个数，$i = 1, 2, \cdots, C$，即每一类的真实样本出现次数

ρ_i——第 i 个类别的精度

T_P^i——第 i 类的 T_P 值

r_i——第 i 个类别的召回率

F_P^i——第 i 类的 F_P 值

F_N^i——第 i 类的 F_N 值

4.3　参数对分类质量的影响

本部分试验的标记样本比例设定为 31%。首先测试了构件图模型中 k-NN 参数 k 对分类质量的影响，试验中 k 取 3、4、5、6，结果如图 6 所示。

从图 6 可以看出，在菊花数据集上，k-NN 参数 k 对分类的质量有一定的影响，试验中 k 的取值为 3、4、5、6，分类结果的精度、召回率和 F_1 度量增大之后又有所降低，在 k 为 5 时 3 个评价指标取得最大值。试验结果表明针对此菊花数据集的半监督近邻参数 k 设置为 5 比较合适。

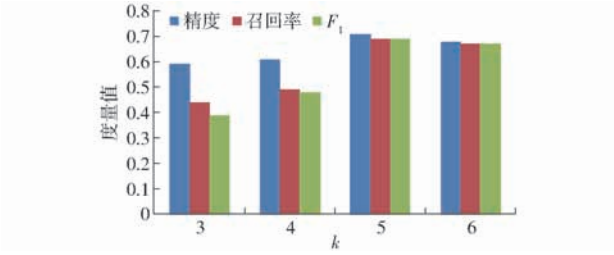

图 6 k-NN 中参数 k 对分类质量的影响

Fig. 6 Influence of parameter k of k-NN on classification quality

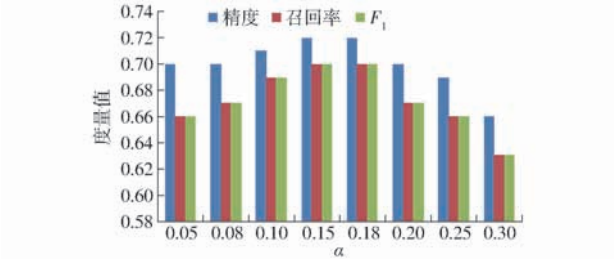

图 7 参数 α 对分类质量的影响

Fig. 7 Influence of parameter α on classification quality

图 7 是夹紧因子 α 对分类质量的影响。图 7 的试验结果表明,参数 α 对结果的影响较大,参数 α 为 0.15 时,半监督学习模型在菊花数据集上 3 个指标较好;随着 α 的增加,分类预测的质量增加达到一个最高值之后随之降低。从 0.15 至 0.18,3 个指标均在 0.7 以上,超过 0.2 之后,3 个指标都有所下降。因此本文参数 α 的选取范围为 0.15 ~ 0.18。

图 8 是熵最大化样本个数 E 对分类质量的影响,从图 8 的试验结果可见,参数 E 为 5 时,对菊花数据具有最好的分类效果,3 个指标都在 0.7 以上。因此参数 E 设置为 5。

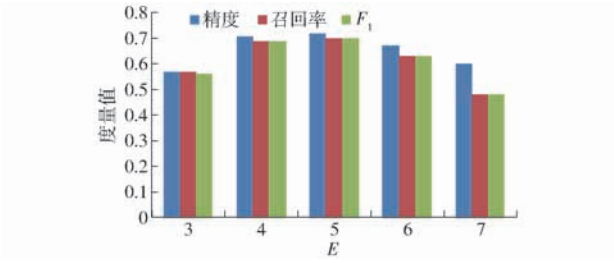

图 8 熵最大化的样本个数对分类质量的影响

Fig. 8 Influence of samples with maximum entropy on classification quality

4.4 分类预测精度

对菊花的标号比例进行了试验,测试了分类的精度、召回率和 F_1 与已标号样本百分比的关系,试验中已标号样本的百分比为 6.25% ~ 81.25%,试验结果如图 9 所示。从图 9 可以看出,随着样本标号百分比的增加,本文方法的分类精度逐步提升,从 0.45 提升至 0.9 以上。在已标号百分比为 6.25% ~ 23% 之间,精度提升较快,精度从 0.45 快速增加到 0.7,在已标号样本百分比为 23% 时分类精度达到 0.7 以上。标号百分比为 81.25% 时,精度为 0.91。

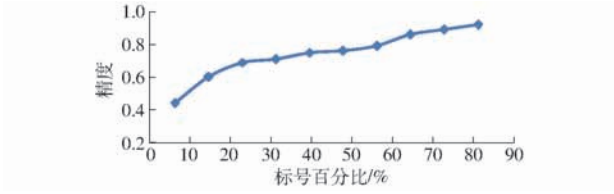

图 9 标号百分比与精度关系

Fig. 9 Relationship of labeled percentage and precision

图 10 是分类的召回率与已标号样本百分比之间的关系。从图 10 可以看出,随着标号百分比的增加,本文方法的分类召回率在逐步提升,从 0.4 提升至 0.88 左右,在标号百分比为 6.25% ~ 23% 时,召回率增加较快,从 0.4 增加到 0.62 以上。在样本标号百分比为 23% 时分类的召回率达到 0.62 以上。标号百分比为 81.25% 时,召回率为 0.88。

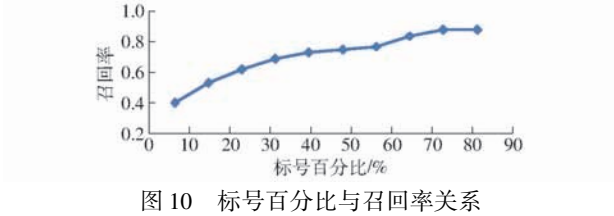

图 10 标号百分比与召回率关系

Fig. 10 Relationship of labeled percentage and recall rate

图 11 是分类结果 F_1 与样本标号百分比之间的关系。从图 11 可以看出,随着样本标号百分比的增加,本文方法的分类 F_1 值在逐步提升,从 0.36 提升至 0.88 左右,在标号百分比为 6.25% ~ 23% 时,F_1 增加较快,从 0.36 增加到 0.63 左右。在样本标号百分比为 23% 时,F_1 达到 0.63 以上。标号百分比为 81.25% 时,F_1 为 0.88。

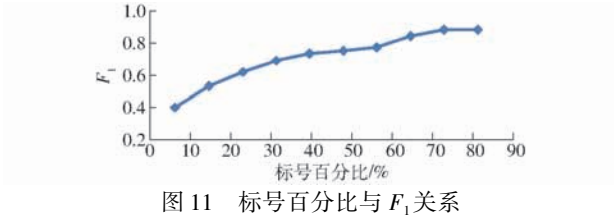

图 11 标号百分比与 F_1 关系

Fig. 11 Relationship of labeled percentage and F_1

图 12 是标号百分比为 56% 时,半监督菊花分类归一化混淆矩阵(Confusion matrix)图,图 12 显示了本文算法在菊花集上分类错误分布情况,行代表的是实际类别,列代表半监督模型的预测类别。从图 12 可以看出,类别 1、2、4 和 5 分类准确率较高,类别 3 被错分的概率较大,可能原因是数据集分布不均衡,这一点也可以从图 5 的类别分布上看出,类别 3 和类别 2、4 之间的分类困难较大。通过提升标号百分比,可以进一步改善分类的质量。

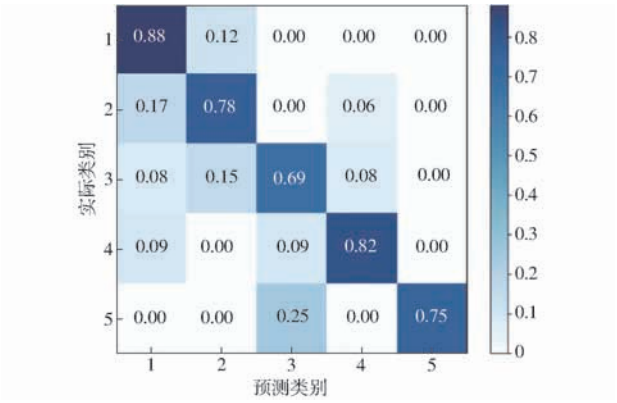

图 12　半监督菊花分类混淆矩阵

Fig. 12　Confusion matrix of semi-classification on chrysanthemum dataset

4.5　分类性能

图 13 为算法的分类时间与菊花样本标号百分比之间的关系图。由图 13 可以看出算法的分类性能与标号百分比关联性不大，分类的时间基本稳定在 4.0 ms 左右，主要时间消耗在图构建过程。表明本文方法的分类性能比较稳定。

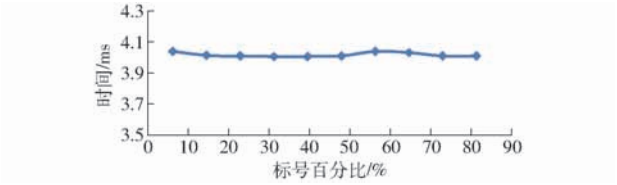

图 13　标号百分比与菊花半监督分类时间关系

Fig. 13　Performance of semi-classification with labeled percentage

4.6　试验对比

把基于图半监督分类无熵策略[27,30]作为基线，对比了分类结果的 3 个指标，试验结果如图 14 所示。基于图半监督方法用无熵表示，本文加入主动学习熵最大化策略方法用加入熵表示。图 14 中横轴百分数表示标号百分比，试验中选取标号百分比分别为 20%、30% 和 37%。从图 14 的试验结果可知，随着标号百分比增加，两种方法的精度、召回率和 F_1 都在增加；加入熵最大化之后，3 个指标都有较大提升，平均提升了 25% 左右；加入熵最大化策略之后，系统的 3 个指标都在 0.63 以上，3 个分类评价指标相对比较稳定。

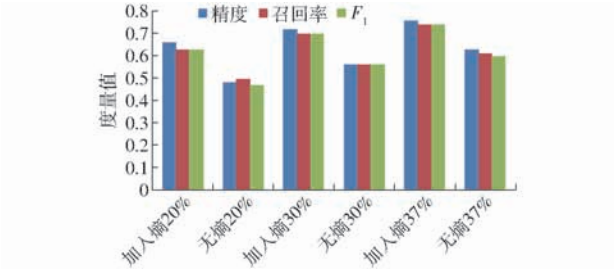

图 14　有无主动学习熵最大化策略分类质量对比

Fig. 14　Classification quality comparison with and without active learning entropy maximization strategy

5　结论

(1)针对部分已标号的菊花数据，利用半监督学习算法，通过主动学习熵最大化策略来提升菊花表型分类效果。

(2)在菊花数据集上对系统参数进行了详细的试验验证，并进行了对比和分析。试验结果表明：本文方法能够较好地利用未标号菊花样本提升分类精度，标号百分比在 23% 时，识别精度快速达到 0.7 以上，标号百分比在 81.25% 时，平均识别精度和召回率分别达到 0.91 和 0.88。

参　考　文　献

1　SINGH A, GANAPATHYSUBRAMANIAN B, SINGH A K, et al. Machine learning for high-throughput stress phenotyping in plants[J]. Trends in Plant Science, 2016, 21(2): 110 – 124.

2　SHEN Z, MAO Y, WU D, et al. Comparative analysis of morphologic traits of 50 large-flowered chrysanthemum varieties[J]. Agricultural Science & Technology, 2016, 17(2):317 – 322.

3　洪艳，白新祥，孙卫，等. 菊花品种花色表型数量分类研究[J]. 园艺学报, 2012, 39(7): 1330 – 1340.
HONG Yan, BAI Xinxiang, SUN Wei, et al. The numerical classification of chrysanthemum flower color phenotype[J]. Acta Horticulturae Sinica, 2012, 39(7): 1330 – 1340. (in Chinese)

4　孙文松. 菊花品种起源及形态学分类研究[J]. 黑龙江农业科学, 2013(9): 58 – 60.
SUN Wensong. Varieties origin and morphological classification of chrysanthemum[J]. Heilongjiang Agricultural Sciences, 2013(9): 58 – 60. (in Chinese)

5　伏静，戴思兰. 基于高光谱成像技术的菊花花色表型和色素成分分析[J]. 北京林业大学学报, 2016, 38(8):88 – 98.
FU Jing, DAI Silan. Analysis of color phenotypic and pigment contents of chrysanthemum based on hyperspectral imaging[J]. Journal of Beijing Forestry University, 2016, 38(8):88 – 98. (in Chinese)

6　ZHANG F, CHEN S, CHEN F, et al. Genetic analysis and associated SRAP markers for flowering traits of chrysanthemum (*Chrysanthemum morifolium*)[J]. Euphytica, 2011, 177(1):15 – 24.

7　张莉俊，戴思兰. 菊花种质资源研究进展[J]. 植物学报, 2009, 44(5):526 – 535.
ZHANG Lijun, DAI Silan. Research advance on germplasm resources of *Chrysanthemum* × *morifolium* [J]. Chinese Bulletin of

Botany, 2009, 44(5):526 – 535. (in Chinese)
8 LEVATI J, BRBI M, PERDIH T S, et al. Phenotype prediction with semi-supervised classification trees[C]// International Workshop on New Frontiers in Mining Complex Patterns. Springer, 2017: 138 – 150.
9 刘建伟,刘媛,罗雄麟. 半监督学习方法[J]. 计算机学报, 2015, 38(8): 1592 – 1617.
LIU Jianwei, LIU Yuan, LUO Xionglin. Semi-supervised learning methods[J]. Chinese Journal of Computers, 2015, 38(8): 1592 – 1617. (in Chinese)
10 ZHU X, GOLDBERG A B. Introduction to semi-supervised learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, 3(1): 1 – 130.
11 YU Z, LU Y, ZHANG J, et al. Progressive semisupervised learning of multiple classifiers[J]. IEEE Transactions on Cybernetics, 2017, 48 (2): 689 – 702.
12 ZHU S, SUN X, JIN D. Multi-view semi-supervised learning for image classification[J]. Neurocomputing, 2016, 208(C): 136 – 142.
13 YANG Z, COHEN W W, SALAKHUTDINOV R. Revisiting semi-supervised learning with graph embeddings[C]// International Conference on International Conference on Machine Learning. JMLR. org, 2016:40 – 48.
14 ZHANG S, LEI Y, ZHANG C, et al. Semi-supervised orthogonal discriminant projection for plant leaf classification[J]. Pattern Analysis and Applications, 2016, 19(4):1 – 9.
15 BEAULIEU-JONES B K, GREENE C S. Semi-supervised learning of the electronic health record for phenotype stratification[J]. Journal of Biomedical Informatics, 2016, 64:168 – 178.
16 CRIMINISI A, SHOTTON J, KONUKOGLU E. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning[J]. Foundations & Trends in Computer Graphics & Vision, 2011, 7(2 – 3): 81 – 227.
17 SINGH S, JANOOS F, PÉCOT T, et al. Identifying nuclear phenotypes using semi-supervised metric learning[C]// International Conference on Information Processing in Medical Imaging. Inf Process Med Imaging, 2011: 398 – 410.
18 DOOSTPARAST T A, PETZOLD L R. Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification[J]. Journal of the American Medical Informatics Association Jamia, 2018, 25(1): 99 – 108.
19 SCHARR H, MINERVINI M, FRENCH A P, et al. Leaf segmentation in plant phenotyping: a collation study[J]. Machine Vision and Applications, 2016, 27(4): 585 – 606.
20 AGARWAL V, PODCHIYSKA T, BANDA J M, et al. Learning statistical models of phenotypes using noisy labeled training data [J]. Journal of the American Medical Informatics Association, 2016, 23(6): 1166 – 1173.
21 DU B, WANG Z, ZHANG L, et al. Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion[J]. IEEE Transactions on Image Processing, 2017, 26(4): 1694 – 1707.
22 WIDMANN N, VERBERNE S. Graph-based semi-supervised learning for text classification[C]// ACM SIGIR International Conference on Theory of Information Retrieval. ACM, 2017:59 – 66.
23 ZHA Z J, MEI T, WANG J, et al. Graph-based semi-supervised learning with multi-label[C]// IEEE International Conference on Multimedia and Expo. IEEE, 2008:1321 – 1324.
24 DENOORD A V, KALCHBRENNER N, VINYALS O, et al. Conditional image generation with pixel CNN decoders[C]// Advances in Neural Information Processing Systems, 2016: 4790 – 4798.
25 KELLEY D R, SNOEK J, RINN J L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks[J]. Genome Research, 2016, 26(7): 990 – 999.
26 MILLER D J, UYAR H S. A mixture of experts classifier with learning based on both labelled and unlabelled data[J]. IEEE Transactions on Medical Imaging, 1997, 9(5): 571 – 577.
27 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.
28 CHEN X, WANG T. Combining active learning and semi-supervised learning by using selective label spreading[C]// IEEE International Conference on Data Mining Workshops. IEEE Computer Society, 2017:850 – 857.
29 FENG Y, HUAMG X, SHI L, et al. Learning with the maximum correntropy criterion induced losses for regression[J]. Journal of Machine Learning Research, 2015, 16(1): 993 – 1034.
30 ZHOU D, HOFMANN T, SCHÖLKOPF B. Semi-supervised learning on directed graphs[C]// Advances in Neural Information Processing Systems, 2005: 1633 – 1640.