文章编号：193

面向高考阅读理解鉴赏题语言风格判别方法*

陈鑫[1]，王素格[1,2]，李德玉[1,2]，谭红叶[1,2]，陈千[1,2]，王元龙[1,2]

（1. 山西大学计算机与信息技术学院，山西 太原 030006；

2. 山西大学计算智能与中文信息处理教育部重点实验室，山西 太原 030006）

摘要：语言风格是高考阅读理解中的重要考察内容，然而，不同考察方式所需的分类层次不尽相同，本文将语言风格鉴赏转化为层次分类问题。在类别标签指导下，利用图分割算法，获取与特定类别相对应的原始簇。基于原始簇，利用层次聚类获取语言风格类别层次结构，之后结合层次结构训练 SVM 层次分类器。在解答语言风格鉴赏题过程中，依据阅读理解题干确定所需分类层次，利用 SVM 层次分类器完成对阅读材料语言风格判别，最后结合知识库生成语言风格鉴赏题答案。实验结果表明，基于层次结构的语言风格判别方法，可以为高考鉴赏问题的解答提供技术支撑。

关键词：层次分类；语言风格；阅读理解；鉴赏题

中图分类号：TP391　　**文献标识码：**A

An Approach of Language Style Discrimination for Reading Comprehension and Appreciation in College Entrance Examination

Chen Xin[1], Wang Suge[1,2], Li Deyu[1,2], Tan Hongye[1,2], Chen Qian[1,2], Wang Yuanlong[1,2]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China; 2.Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: Language style discrimination plays a significant role in reading comprehension and appreciation in the college entrance examination. However, the hierarchy of classification varies with different exam points. The task of language style discrimination and appreciation is regarded as a hierarchical classification problem. Guided by class labels, original clusters corresponding to the specific categories are acquired by using graph segmentation algorithm. Furthermore, hierarchical clustering is applied to generate a hierarchy of language style, with which a hierarchical classifier based on SVM is trained. During answering appreciation question, the hierarchy of classification, which is determined by the question stem, is used to discriminate the language style of reading materials. Finally, the answers are generated based on the combination of language style and knowledge base. The experiments show the effectiveness of our proposed method which can provide technical support for reading comprehension and appreciation.

Key words: hierarchical classification; language style; reading comprehension; appreciation questions

1 引言

在信息革命的浪潮中，人工智能应运而生并蓬勃发展，极大推动计算机语音识别、图像分析及文本语义理解能力。为了检验计算机对文本语义深层理解效力，国家 863“超脑计划”牵头研制“高考机器人”，即利用人工智能程序模拟高考生，参与高考。对于高考语文中考察的内容，不仅考察考生对文本理解的能力，还检验其对文本的鉴赏能力，其中语言风格是比较重要的考察内容。由于语言风格是说话者个人语言情感的流露，其情感色彩相比任何其

*收稿日期：　　定稿日期：

基金项目：国家“八六三”高技术项目(2015AA015407);国家自然科学基金资助项目(61573231, 61632011, 61672331);

作者简介：陈鑫（1992——），女，博士研究生，主要研究方向为中文信息处理；王素格（1964——），女，教授，主要研究方向为中文信息处理；李德玉（1965——），男，教授，主要研究方向为数据挖掘。

它语言现象更为丰富[1]。例如，语言风格中“明朗”一般较多使用在情感色彩比较鲜明的词语情感表达中，而“含蓄”语言风格则使用描绘性辞格进行情感表达[2]。因此，语言风格类别的判别既能为鉴赏题解答技术提供支撑，也能为分析阅读材料作者的情感奠定基础。

由于语言风格体系复杂，类别标签繁多，传统的二元分类器（如 SVM）对多分类问题解决效果都不尽人意。利用语言风格的层级化系统[3]，研究基于层次结构的语言风格判别，既缓解多分类对二元分类器带来的挑战，也可以灵活选择分类的层次，以满足高考对语言风格不同考察方式。例如：

题目 1：以③④段为例，简要分析本文语言的两个主要特点。

题目 2：本文的细节描写细腻而生动，从多个角度抒发着作者的生命感悟。请选择一个最打动你的细节进行语言特色分析。

题目 1 未提及特定的语言风格，为提高判别准确率，可进行粗粒度分类。而题目 2 针对语言风格“细腻”考察，则需进行细粒度分类。

通常层次分类依赖的类别层次结构可由专家编制，也可通过聚类生成[4]。为了克服专家编制的类别层次结构主观性，Tang 等[5]提出一种动态结构调整方法，该方法具有较高的时间复杂度，随后，Nitta[6]对其时间开销进行改进，但调整结构受限于最初层次结构。为了减小结构生成过程对专家知识的依赖性，Phongwattana 等[7]基于欧氏距离，利用层次聚类获取类别层次结构，但欧式距离仅能刻画簇间空间距离，并未对其语义距离度量。另外，Karypis 等[8]提出一种动态的层次聚类算法。首先利用 K 近邻算法构建图，然后，基于快速图分割算法 METIS[9]，将数据图划分为多个子簇，最后，基于簇间相对互连性与相对相似性，对簇进行迭代合并，得到最终层次聚类结果。此层次聚类方法可对形状各异、大小不一子簇进行动态聚合，被应用到文本、图像及高铁故障检测任务中[10]，并取得理想的效果。

本文综合多名学者对语言风格的类别划分结果[1-3,11-16]，结合高考对考生考察要求，研究语言风格类别标签的判别问题。为了实现高效的语言风格的类别判断，将语言风格鉴赏转化为分类任务，并利用识别结果辅助语言风格鉴赏题解答。

本文第 2 节将确定语言风格层次结构；第 3 节展现基于层次聚类的类别层次结构获取算法、基于层次分类的语言风格识别及面向高考语言风格鉴赏题解答流程；实验数据和评价指标在第 4 节呈现；第 5 节对实验结果进行详细的分析；最后一节给出全文的结论与展望。

2 语言风格类别的层次划分

由于语言风格体系复杂，语言学家研究粒度存在差异。宗世海[11]从多个角度划分粒度，从篇幅划分，粒度可为单篇文档、多篇文档；从作品集角度，可为单个作者作品、某类作者作品、一个语体。丁金国[12]认为语言风格粒度具有层级化，可分为语体-文体-语篇 3 个层次，其中最小粒度的语篇可为一个句群、一个段落、一篇文章等。而高考对语言风格鉴赏是面向单篇文档或单个段落，因此本文的研究粒度设定为单个段落。

由于同一时期的不同学者对语言风格定义迥异，而同一学者在不同时期的定义也不完全相同[13]，因此，语言学家对语言风格的类别划分差异较大。依据文献 [1-3]和文献[11-16]，我们将语言风格的表达方式分为平面划分、对立划分、层次划分，其具体划分结果见表 1。

根据表 1，综合多名学者对语言风格的类别划分结果[1-3,11-16]，结合高考对考生考察要求，将语言风格划分为 12 个类别，分别为幽默诙谐、细腻隽永、朴素自然、华丽典雅、含蓄深沉、简洁明快、雄浑豪放、清新婉约、率性旷达、严谨工整、舒缓和平、急骤猛烈。

由于语言风格中存在对立类别，为了防止层次聚类中对立类别簇聚合，本文参考语言学家的对立划分结果[3,11,13,15,16]，建立对立集 R，即：{雄浑豪放—清新婉约，雄浑豪放—细腻隽永，急骤猛烈—舒缓和平，华丽典雅—朴素自然，含蓄深沉—简洁明快，率性旷达—含蓄深沉，率性旷达—严谨工整}。另外，依据丁金国[12]定义的类别层次结构，见图 1，结合本

文确定类别标签，修改后的类别层次结构 MH 见图 2。

表 1 语言风格划分结果

划分方式	学者	划分结果
平面划分	丁金国[1]	朴素平实、简明通达，准确简练、严密典雅，严谨准确、简洁刚健，浓艳绮丽，淡雅清新，雄健豪放，高古典雅，含蓄委婉，幽默诙谐，肃穆深沉
	郑荣馨[2]	朴素，华丽，简练（简约），繁丰，明朗，含蓄，雄浑（雄健），柔婉，通俗，庄重（庄严），幽默，疏放，空灵，清奇，飘逸，旷达，流动
	陈继民[14]	奔腾激越、豪迈奔放，芬芳清新、似出山泉，字字含情、诗意浓郁，朴素真实、深刻隽永
对立划分	黎运汉[3]	豪放—柔婉，繁丰—简约，蕴藉—明快，藻丽—朴实，幽默—庄重，文雅—通俗，疏放—谨严
	宗世海[11]	庄重—幽默，刚健—柔婉，明快—含蓄，平实—藻丽，简约—繁丰
	黎运汉[13]	豪放—柔婉，繁丰—简约，蕴藉—明快，藻丽—朴实，幽默—庄重，文雅—通俗，疏放—缜密
	宋振华等[15]	简约—繁丰，朴素—华丽，庄严—幽默，文雅—通俗，谨严—疏放，豪放—柔婉
	戈娟[16]	简约—繁丰，刚健—柔婉，平淡—绮烂
层次划分	丁金国[12]	见图 1

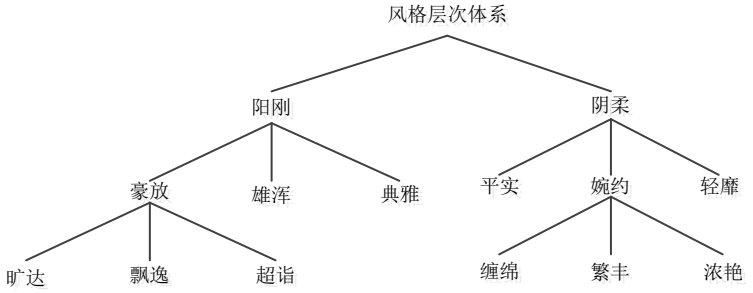

图 1 专家编制层次结构图

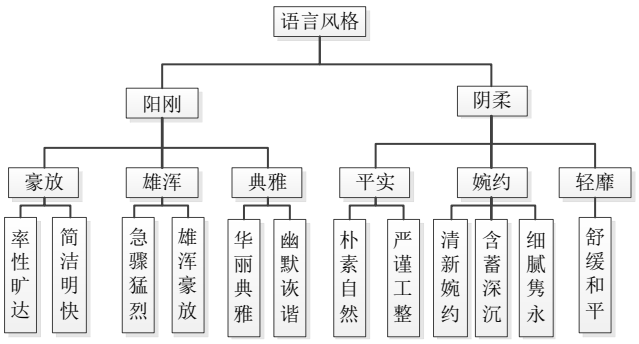

图 2 修改后专家编制类别层次结构 MH

3 基于层次结构的语言风格判别方法

为了适应高考不同考查要求，本文利用层次分类方法判别语言风格，其分类策略可划分为全局处理策略、化繁为简策略、分而治之策略[4]。全局处理策略基于整个层次结构优化分类器，有较大的时间开销。化繁就简策略首先筛选与待分类样本相关的候选类别，然后利用对应分类器进行分类，虽可以灵活选择分类类别及分类器，但计算开销较大。分而治之策略依据层次结构逐层分类，虽存在错误累计问题，但有较小时间开销。因此，本文采用分而治之的层次分类策略，用于语言风格的类别判别。

基于层次结构的语言风格判别，主要由获取类别的层次结构、判别语言风格两部分组成，具体流程见图 3。

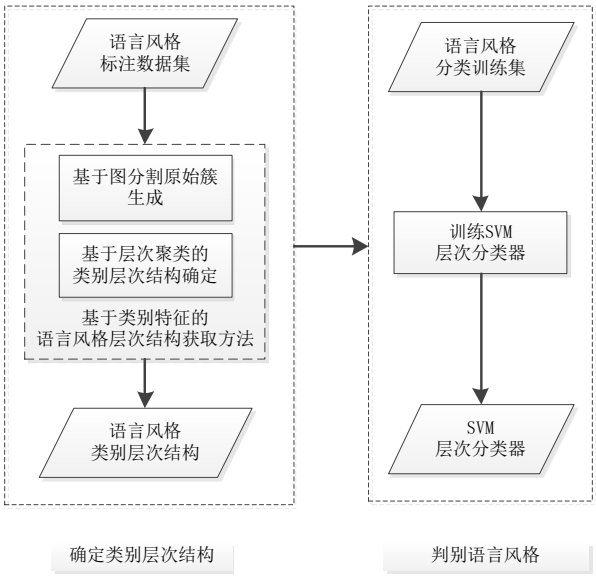

图 3 基于层次结构语言风格判别方法流程图

3.1 语言风格类别层次结构获取算法

在语言风格类别层次结构确定过程中，为减少对专家知识的依赖，本文利用层次聚类方法[8]获取语言风格类别层次结构。

设语言风格数据集 $D=\{d_1,d_2,...,d_n\}$，其类别标签为 $C=\{c_1,c_2,...,c_m\}$，对立集 $R=\{r_1,r_2...,r_k\}$，（其中，$r_i=(c_t,c_h)$，（$1\le k1\le th\le mt$ with, $N\in$ ）。在构建类别层次结构过程中，数据类别标签 C 将有助于划分层次聚类原始簇，使每个原始簇分别刻画一个类别，与此同时，可以利用 R 阻止对立类别合并成新簇。为了充分利用类别标签 C 及其对立集 R 的信息，本文在 D 的原始特征集 $F=\{f_1,f_2,...,f_{|F|}\}$ 基础上，将类别标签 C 增加为新的特征 f_c，即 $FC=F\cup\{f_C\}=\{f_1,f_2,...,f_{|F|},f_C\}$。另外，为了增强类别特征区分性，将每个类别特征赋予一个特征值，即 $V_{f_C}=\{1,2,...\}m$，并利用 R 中的对立信息设计 V_{f_C} 与 C 间映射关系 $M_{f_C}=\{v_{c_o}=1,v_{c_p}=2,...,v_{c_q}=m\}$（$1\le o,p,q\le m;o,p,q\in N$）。由于 R 中每对 r_i 为对立类别，其特征值差值越大，其类别区分能力越强，因此，在确定特征值过程中，综合 R 中所有对立信息，即保证 R 中所有对立类别间特征值差距之和最大，则最优映射关系 $M_{f_C}^{opt}$ 的计算过程见公式（1）。

$$M_{f_C}^{opt}=\arg\max_{M_{f_C}}\Big(\sum_{i=1,(c_t,c_h)\in r_i}^{k}|\#_{c_t}(M_{f_C})-\#_{c_h}(M_{f_C})|\Big) \tag{1}$$

其中，$\#_{c_t}(M_{f_C})$ 为 M_{f_C} 映射关系中 c_t 的特征值。

在层次聚类过程中，本文采用 Karypis[8]提出算法，综合簇间相对互连性（见公式 2）、相对近似性（见公式 3）度量簇间相似性（见公式 4），迭代完成簇间合并。

$$RI(sc_i,sc_j)=\frac{2*\left|EC_{\{sc_i,sc_j\}}\right|}{\left|EC_{sc_i}\right|+\left|EC_{sc_j}\right|} \tag{2}$$

其中，sc_i,sc_j 代表两个簇，$EC_{\{sc_i,sc_j\}}$ 为簇 sc_i,sc_j 的连接边，EC_{sc_i} 为簇 sc_i 的二等分极小割边。

$$RC(sc_i,sc_j)=\frac{\bar{S}EC_{\{sc_i,sc_j\}}}{\frac{|sc_i|}{|sc_i|+|sc_j|}\bar{S}EC_{sc_i}+\frac{|sc_j|}{|sc_i|+|sc_j|}\bar{S}EC_{sc_j}} \tag{3}$$

其中，$\overline{SEC}_{\{sc_i,sc_j\}}$ 为 sc_i, sc_j 连接边平均权重。

$$Sim_{\{sc_i,sc_j\}} = RI(sc_i, sc_j) * RC(sc_i, sc_j)^{\alpha} \quad (4)$$

其中，α 代表比例参数，用来度量簇间相似度计算过程中相对互联性与相对相似性重要程度。

依据特征集 *FC*，将语言风格样本表征为向量，采用 KNN 算法构造样本图，并利用图分割算法获取样本标签原始簇，最后利用层次聚类确定类别层次结构，具体见算法 1。

算法 1：语言风格类别层次结构获取算法

输入：语言风格标注数据集 *D*；对立类别标签集 *R*；KNN 算法中 *k* 值
输出：语言风格层次类别结构 AH

1 利用公式（1）确定 f_c 特征值 V_c
2 采用 *FC* 表征 *D*
3 基于欧式距离，采用 KNN 方法构建稀疏图
4 利用图分割算法 METIS,获取原始簇集 *CL*（其中 $|CL|=|C|=m$）
5 While(|*CL*| > 2)
6 For $cl_i \in CL$，$cl_j \in CL$ $(i \neq j)$
7 利用公式（2），计算其相对互连性 $RI(cl_i, cl_j)$
8 利用公式（3），计算相对近似性 $RC(cl_i, cl_j)$
9 利用公式（4）度量 cl_i, cl_j 相似性
10 End For
11 合并相似性最大的 cl_i, cl_j，更新 H，更新 *CL*
12 End While
13 返回 AH

3.2 基于 SVM 层次分类的语言风格识别方法

为了对文本语言风格实时、高效的判别，并将类别层级结构信息保留于判别结果，本文基于 3.1 节确定的语言风格类别层次结构，采用“分而治之”的层次分类方法识别语言风格。另外，SVM 作为一个以间隔最大化为学习策略的二元分类器，与 3.1 节中确定的二叉语言风格层次结构相吻合。因此，本文基于 SVM 层次分类，实现对语言风格识别，具体流程见图 4。

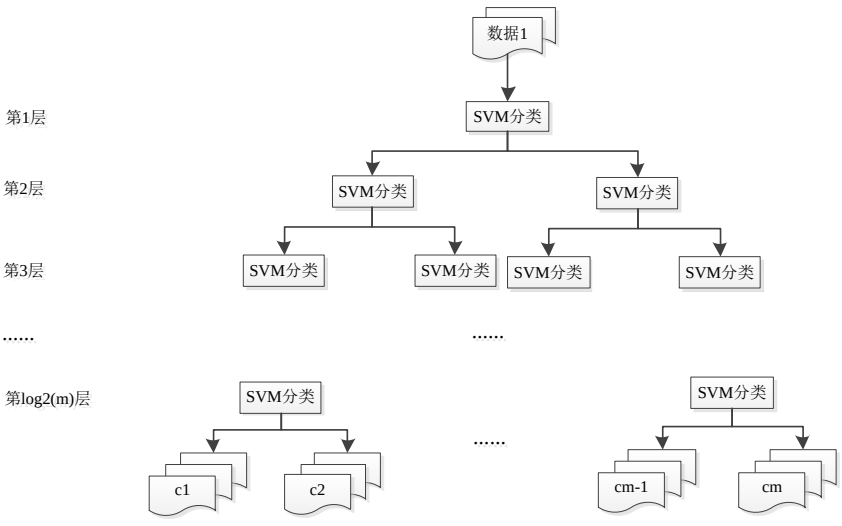

图 4 基于 SVM 层次分类流程图

层次分类过程中，首先利用第 1 层分类器对数据集 D 进行分类，获得分类结果；然后依据分类结果，寻找对应 SVM 分类器，进行第 2 层分类后；……；直到获取最终的语言风格标签类别 c_k（$1 \le k \le m$）。

3.3 基于语言风格识别的鉴赏题解答

为了应对高考对语言风格的考察，本文将利用 3.2 节中训练的层次 SVM 分类器，完成对文本语言风格的识别。在高考鉴赏题解答过程中，根据题干选择分类层次，即若题干包括特定的语言风格，则确定分类层次为叶节点；如果题干未提及具体的语言风格，为提高识别准确率，则分类层次确定为叶节点的父节点。然后，基于 3.1 节确定的类别层次结构 AH，利用 3.2 节中 SVM 层次分类器识别阅读材料段落语言风格，并结合语言风格作用知识库，生成答案，具体流程见图 5。

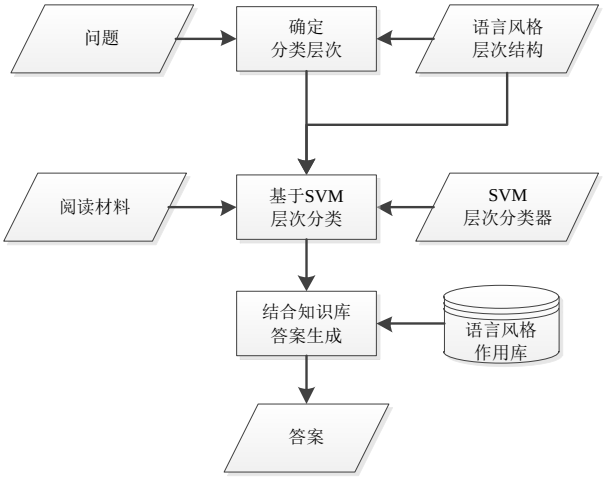

图 5 面向高考阅读理解的语言风格鉴赏题解答流程

4 实验数据集及评价指标

4.1 实验数据集

数据集 1：收集人教版高中课文、全国高考（2002-2016）阅读理解材料，共计 484 篇，6646 段。利用第 2 节确定的类别标签，进行人工标注，12 种类别具体比例见表 2。

表 2 语言风格标注语料类别占比

类别	数据数量	比例	类别	数据数量	比例	类别	数据数量	比例
朴素自然	2148	32.32%	简洁明快	580	8.73%	清新婉约	139	2.09%
细腻隽永	831	12.50%	急骤猛烈	577	8.68%	幽默诙谐	138	2.08%
含蓄深沉	826	12.43%	舒缓和平	303	4.56%	雄浑豪放	131	1.97%
严谨工整	763	11.48%	率性旷达	174	2.62%	华丽典雅	36	0.54%

数据集 2：为了避免数据不平衡性对类别层次结构获取造成影响，从数据集 1 中 12 个类别标注数据中分别选取 36 条数据，共计 432 条，作为类别层次结构确定方法验证数据。

4.2 评价指标

语言风格判别整个过程由类别层次结构获取、基于 SVM 层次分类两部分构成。类别层次获取过程中类别原始簇利用熵、纯度度量；层次分类结果则采用正确率（accuracy）、准确率（precision）、召回率（recall）及 F 值度量。

（1）生成原始簇的评价指标

类别分布度量：利用熵度量图分割生成的类别原始簇在各个类别分布情况。假设 p_{ij} 为

簇 i 中成员属于类 j 的概率，即 $p_{ij}=\frac{u_{ij}}{u_i}$，其中，$u_i$ 代表簇 i 中样本数，u_{ij} 为簇 i 中数据类 j 的个数。聚类簇 i 的熵值 e_i 见公式（5），整个聚类划分的熵值见公式（6）。

$$e_i=-\sum_{j=1}^{|C|}p_{ij}\log_2 p_{ij} \tag{5}$$

$$e=\sum_{i=1}^{l}\frac{u_i}{u}e_i \tag{6}$$

其中，l 代表簇的个数，u 代表整个聚类划分样本数。

其熵值越大，说明原始簇分布在各个类别越均匀，则原始簇对类别刻画能力越弱。

簇的纯度度量：簇的纯度为簇中最大类别所占比值，即纯度值越大，簇对单个类别刻画能力越强。聚类簇 i 的纯度计算见公式（7），整个聚类划分的纯度见公式（8）。

$$p_i=\max_j(p_{ij}) \tag{7}$$

$$p=\sum_{i=1}^{l}\frac{u_i}{u}p_i \tag{8}$$

（2）层次分类的评价指标

正确率（accuracy）为测试集正确分类的样本数与测试集总样本数占比，其刻画层次分类总体分类准确性。除此之外，本文还利用准确率（precision）、召回率（recall）及 F 值度量每个类别的分类效果。

5 实验结果与分析

本节针对语言风格判别过程中的类别层次结构生成、基于 SVM 语言风格层次分类进行实验，用于验证本文语言风格判别的有效性。

实验 1：语言风格类别层次结构的获取

语言风格类别是由多种因素决定的，其中，词汇表达占有重要的地位[1,17]。例如：“丢掉、拿手、脑袋”这些词为口语词语，却体现出“朴素自然”语言风格，而书面词语“遗弃、擅长、头颅”则能表现出“华丽典雅”的语言风格。因此，我们选取词袋特征作为其表征单元，使用 3.1 节的层次聚类，设计了 3 组特征表征实验方案，用于获取类别层次结构，具体如下：

方案 1：仅使用词袋模型表征文本，记作 baseline；

方案 2：在词袋模型的基础上，增加 12 维 one-hot 类别特征，指导层次结构生成；

方案 3：在词袋特征基础上，增加 1 维类别特征，并从公式（1）计算得出的 11520 最优特征值中随机选取一组进行实验，其各个类别 $M_{f_C}^{opt}$ ={率性旷达=1，雄浑豪放=2，急骤猛烈=3，简洁明快=4，朴素自然=5，幽默诙谐=6，严谨工整=7，华丽典雅=8，细腻隽永=9，舒缓和平=10，清新婉约=11，含蓄深沉=12}。

利用 3.1 节算法，基于图分割的原始簇生成结果见表 3、表 4，层次聚类结果见图 6。

观察表 3 和表 4，随着将类别信息加入到特征后，图聚类生成的原始簇的熵值降低，纯度增加；并且方案 3 比方案 2 熵值更低，纯度更高。则说明，类别特征对图聚类原始簇生成有指导作用，并且一维特征优于“one-hot”方式。分析其中原因：

（1）方案 1 图分割原始簇生成过程，由于缺少类别标签的指导，每个原始簇中包含多个类别，且各个类别比例差异不大，熵值大，纯度低，即初始簇不能刻画语言风格特定类别。

（2）一维特征比 one-hot 特征区分类别能力强。

由于方案 1 及方案 2 图分割生成原始簇有较高的熵值、较低的纯度，皆无法明确簇与类别间对应关系。因此，将方案 3 生成的类别层次结构（图 6）AH 作为之后层次分类依赖的

层次结构。

表 3 聚类原始簇熵值

	簇 0	簇 1	簇 2	簇 3	簇 4	簇 5	簇 6	簇 7	簇 8	簇 9	簇 10	簇 11	Total
方案 1	3.36	3.09	3.30	3.50	3.15	3.11	2.64	3.14	2.69	3.36	3.29	3.23	**3.15**
方案 2	3.14	3.01	3.25	2.24	3.15	2.88	2.91	3.12	2.88	3.28	3.22	2.61	**2.97**
方案 3	1.73	0.74	1.24	1.66	1.30	1.33	1.04	1.69	1.77	1.96	1.49	0.75	**1.39**

表 4 聚类原始簇纯度

	簇 0	簇 1	簇 2	簇 3	簇 4	簇 5	簇 6	簇 7	簇 8	簇 9	簇 10	簇 11	Total
方案 1	0.17	0.18	0.19	0.14	0.19	0.19	0.36	0.21	0.35	0.15	0.16	0.22	**0.21**
方案 2	0.18	0.16	0.19	0.51	0.17	0.41	0.30	0.24	0.35	0.17	0.19	0.30	**0.27**
方案 3	0.46	0.84	0.51	0.59	0.70	0.62	0.76	0.50	0.49	0.34	0.49	0.78	**0.60**

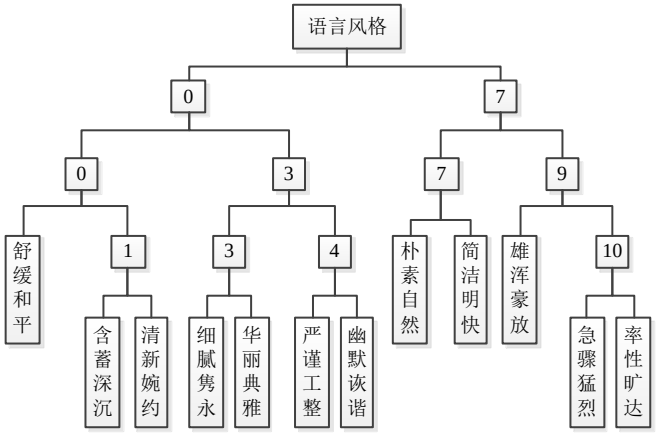

图 6 自动生成语言风格类别层次结构 AH

实验 2：基于类别层次结构的层次 SVM 分类

在数据集 1 上，选取词袋为特征，词频为特征值，分别基于专家编制层次结构 MH（图 2）、自动生成层次结构 AH（图 6）、平面结构（即一层结构，baseline），采用 5 次交叉验证对语言风格进行判别。针对实验结果，本文从结点分类、整体分类 2 个角度分析实验结果。

（1）结点分类结果

为了验证层次分类过程中结点分类效果，又鉴于分而治之策略层次分类方法有错误累计的缺点，本文利用正确率（accuracy）度量层次结构中每个结点的分类效果，具体结果见图 7、图 8。

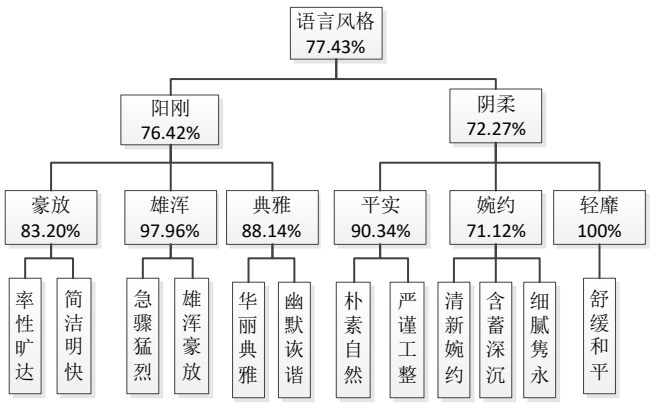

图 7　MH 结点 SVM 分类正确率

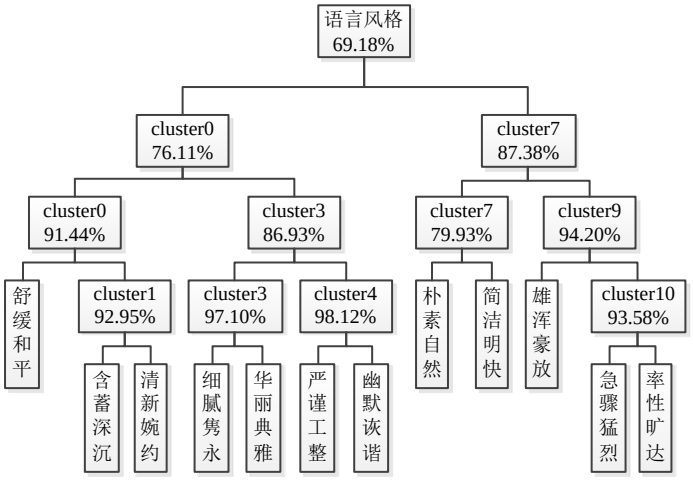

图 8　AH 结点 SVM 分类正确率

对比图 7、图 8 中每个结点 SVM 分类正确率，除第一层外，AH 最低正确率为 76.11%，最高正确率为 98.12%，而 MH 中最低正确率为 71.12%，最高正确率为 90.34%。从而证明 AH 细粒度分类效果优于 MH。另外，从图 8 中发现，“简洁明快”与“朴素自然”的分类正确率低，只有 79.93%，这是由于两种语言风格用词一致性高造成的。

（2）整体分类结果

为了验证层次分类过程中，结构对整体分类效果影响，本文将利用正确率（accuracy）、宏准确率（Macro-Precision）、宏召回率（Macro-recall）及宏 F 值（Macro-F）对分类结果进行评价，具体结果见表 5。

表 5 语言风格分类 accuracy、Macro-Precision、Macro-recall 及 Macro-F

结构	Macro-Precision	Macro-recall	Macro-F	第一层 accuracy	第二层 accuracy	第三层 accuracy	第四层 accuracy
平面结构（baseline）	45.72%	36.48%	40.25%	50.55%	-	-	-
专家编制结构 MH	38.57%	34.52%	35.99%	77.43%	56.51%	48.47%	-
自动生成结构 AH	38.97%	35.59%	37.68%	69.18%	57.25%	48.71%	47.60%

观察表 5 可以看出：

① AH 的 Macro-Precision、Macro-recall、Macro-F 均超过 MH，即证明自动生成层次结构过程中，本文方法对语言风格类别间关联认识优于专家知识，说明本方法的层次结构划分是由具体数据决定，可以根据数据的不同实现层次结构的动态调整。

② 语言风格识别过程中，类别层次结构确定与层次分类独立实现，未考虑两个子任务关联关系，造成 AH 的 Macro-Precision、Macro-recall、Macro-F 都低于平面结构。但 AH 结构具有层次性，语言风格识别过程中，能自由选择分类的层次，如第 1 节题目 1，为提高准确率，分类过程中可以将其分类至叶节点上一层。从表 5 中结果可以看出，AH 第三层之前正确率均高于平面分类。

③最终叶节点的分类的正确率，MH 略优于 AH。但在第二、三层分类正确率中，AH 高于 MH。结合图 7、图 8，AH 中细粒度的分类结果也高于 MH。

实验 3：基于语言风格识别高考语文鉴赏题解答

利用 3.3 节中语言风格鉴赏题解答流程，针对第 1 节题目 1 分别基于平面结构、基于 MH、基于 AH 解答语言风格鉴赏题，记为方案 1（baseline）、方案 2、方案 3。为了验证 AH 层次信息在答题过程有效性，设计方案 4、方案 5，即分别在 MH 第二层、AH 的第三层完成语言风格分类，实验结果见表 6。

表 6　2012 年山东卷高考试题解答结果

【问题】2012 年山东高考-21. 以③④段为例，简要分析本文语言的两个主要特点。	
方案 1 答案	a）语言具有“**华丽典雅**”的特点，语言字字珠玑，词句优美，言简意深，凝炼有力。
方案 2 答案	a）语言具有“**简洁明快**”的特点，语言直接、明朗的表达情感，单刀直入，一语中的，一针见血，痛快淋漓。
方案 3 答案	a）语言具有“**细腻隽永**”的特点，语言有“曲、细、柔”，曲径通幽，情调缠绵，表达感情细如抽丝，富有感染力。 b）语言具有“**含蓄深沉**”特点，语言意在言外，不直接叙述，字里行间总是留着启人联想、开人悟性的“空白”。
方案 4 答案	**【方案 2 答案】** b）语言具有“率性旷达”的特点，语言疏狂不羁，通脱豁达，潇洒飘逸，高洁特立。
方案 5 答案	**【方案 3 答案】** c）语言具有“清新婉约”特点，语言比较通俗，新颖独到，不落俗套，拥有泥土气息、生活气息。 d）语言具有“**华丽典雅**”的特点，语言字字珠玑，词句优美，言简意深，凝炼有力。

从表 6 结果看出，方案 1 与方案 2，分别从两段话中识别出一种正确的语言风格，方案 3 识别出“细腻隽永”与“含蓄深沉”两种正确语言风格，效果优于平面结构及 MH 结构。

方案 4、5 分别相对于方案 2、3 扩充识别语言风格的兄弟节点，然而方案 4 扩充的语言风格是错误的，方案 5 扩充的两种语言风格中的“华丽典雅”为正确的语言风格。从而说明 AH 结构优于 MH，且 AH 较平面分类能自由选择分类的粒度。

6 结论与展望

语言风格作为高考重要考察点，为应对高考不同考察方式所需分类层次差异，缓解多分类对二分类器带来的挑战，本文利用层次分类方法识别语言风格，并结合知识库，完成语言风格鉴赏题的解答。实验证明，层次分类比平面分类具有较强灵活性，并且基于自动获取结构分类效果好于专家编制结构。但层次分类叶节点的准确率低于平面分类，这是由层次结构获取与基于层次结构分类独立进行，未考虑其关联性造成。未来我们将综合考虑结构获取与层次分类，完成层次多分类任务，进一步提高语言风格识别效果。

参考文献

[1] 丁金国. 关于语言风格学的几个问题[J]. 河北大学学报(哲学社会科学版), 1984，(3):45-57.

[2] 郑荣馨. 语言表现风格论:语言美的探索[M]. 安徽大学出版社，1999.

[3] 黎运汉. 语言风格系统论[J]. 渤海大学学报(哲学社会科学版)，1996，(3):100-105.

[4] 何力, 贾焰, 韩伟红,等. 大规模层次分类问题研究及其进展[J]. 计算机学报, 2012, 35(10):2101-2115.

[5] Tang L, Zhang J, Liu H. Acclimatizing taxonomic semantics for hierarchical content classification[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006:384-393.

[6] Nitta K. Improving taxonomies for large-scale hierarchical classifiers of web documents[C]// Proceedings of the ACM Conference on Information and Knowledge Management, 2010:1649-1652.

[7] Phongwattana T, Engchuan W, Chan J H. Clustering-based multi-class classification of complex disease[C]// Proceedings of the International Conference on Knowledge and Smart Technology. IEEE, 2015:25-29.

[8] Karypis G, Han E H, Kumar V. CHAMELEON: a hierarchical clustering algorithm using dynamic modeling[J]. Computer, 1999, 32(8):68-75.

[9] Karypis G, Kumar V. A fast and high quality multilevel scheme for partitioning irregular graphs[J]. Siam Journal on Scientific Computing, 2006, 20(1):359-392.

[10] Xiao W, Yang Y, Wang H, et al. Semi-supervised hierarchical clustering ensemble and its application [J]. Neurocomputing, 2016, 173:1362-1376.

[11] 宗世海. 论言语风格的分类[J]. 语文研究, 2003，(3):42-46.
[12] 丁金国. 语言风格的研究平面[J]. 烟台大学学报(哲学社会科学版), 1991，(4):65-73.
[13] 黎运汉. 1949 年以来语言风格定义研究述评[J]. 语言文字应用, 2002，(1):100-106.
[14] 陈继民. 品鉴散文的语言风格[J]. 中文自修, 1995，(12):17.
[15] 宋振华, 吴士文, 张国庆,等. 现代汉语修辞学[M]. 天津人民出版社, 1963.
[16] 戈娟. 初中现代散文语文教学研究[D]. 杭州师范大学, 2016.
[17] 马琳. 论以语言要素为手段的语言风格构建[J]. 长江师范学院学报, 2004, 20(6):48-50.

作者联系方式：
作者一：陈鑫，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006,18734838988, 1315614497@qq.com；
作者二：王素格，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006,13934649855，wsg@sxu.edu.cn。
作者三：李德玉，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006,15834168298，lidy@sxu.edu.cn。