一种基于邻域粗糙集的多标记专属特征选择方法

孙 林[1,2,3] 潘俊方[4] 张霄雨[1] 王 伟[1] 徐久成[1,2]
(河南师范大学计算机与信息工程学院 河南 新乡 453007)[1]
(河南师范大学生命科学学院生物学博士后流动站 河南 新乡 453007)[2]
(计算智能与数据挖掘河南省高校工程技术研究中心 河南 新乡 453007)[3]
(电子科技大学基础与前沿研究院 成都 610054)[4]

摘 要 在多标记学习中,数据降维是一项重要且具有挑战性的任务,而特征选择又是一种高效的数据降维技术。在邻域粗糙集理论的基础上提出一种多标记专属特征选择方法,该方法从理论上确保了所得到的专属特征与相应标记具有较强的相关性,进而改善了约简效果。首先,该方法运用粗糙集理论的约简算法来减少冗余属性,在保持分类能力不变的情况下获得标记的专属特征;然后,在邻域精确度和邻域粗糙度概念的基础上,重新定义了基于邻域粗糙集的依赖度与重要度的计算方法,探讨了该模型的相关性质;最后,构建了一种基于邻域粗糙集的多标记专属特征选择模型,实现了多标记分类任务的特征选择算法。在多个公开的数据集上进行仿真实验,结果表明了该算法是有效的。

关键词 多标记学习,邻域粗糙集,专属特征,特征选择
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.01.030

Multi-label-specific Feature Selection Method Based on Neighborhood Rough Set

SUN Lin[1,2,3] PAN Jun-fang[4] ZHANG Xiao-yu[1] WANG Wei[1] XU Jiu-cheng[1,2]
(College of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan 453007, China)[1]
(Post-doctoral Mobile Station of Biology, College of Life Science, Henan Normal University, Xinxiang, Henan 453007, China)[2]
(Engineering Technology Research Center for Computing Intelligence and Data Mining of Henan Province, Xinxiang, Henan 453007, China)[3]
(Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China)[4]

Abstract Dimensionality reduction of data is a significant and challenging task under multi-label learning, and feature selection is a valid technology to reduce the dimension of vector. In this paper, a multi-label-specific feature selection method based on neighborhood rough set theory was proposed. This method ensures theoretically that there exists a strong correlation between the obtained label-specific features and the corresponding labels, and then reduction efficiency can be improved well. Firstly, a reduction algorithm of rough set theory is applied to reduce redundant attributes, and the label-specific features are obtained while keeping the classification ability unchanged. Then, the concepts of neighborhood accuracy and neighborhood roughness are introduced, the calculation approaches to dependence and attribute significance based on neighborhood rough set are redefined, and the related properties of this model are discussed. Finally, a multi-label-specific feature selection model based on neighborhood rough set is presented, and the corresponding feature selection algorithm for multi-label classification task is designed. The experimental results under some public datasets demonstrate the effectiveness of the proposed multi-label-specific feature selection method.

Keywords Multi-label learning, Neighborhood rough set, Label-specific feature, Feature selection

1 引言

多标记学习是模式识别、机器学习、数据挖掘及数据分析等领域的一个研究热点[1-2]。在传统的监督学习中,每个样本都具有清晰、单一的语义标记,被称为单标记学习问题[3]。然而,在现实生活中,样例通常具有多义性。例如:一篇新闻报道可以同时有政治、经济、文化等多个类别标记;一张风景图可能同时有大海、湖泊、小岛等多个类别标记;一首诗歌可同

到稿日期:2017-05-08 返修日期:2017-09-16 本文受国家自然科学基金项目(61772176,61402153,61370169,61602158),中国博士后科学基金项目(2016M602247),河南省科技攻关项目(162102210261),新乡市科技攻关计划项目(CXGG17002),河南师范大学博士科研启动费支持课题(qd15132)资助。

孙 林(1979-),男,博士,副教授,CCF会员,主要研究方向为粒计算、数据挖掘、生物信息学等,E-mail:sunlin@htu.edu.cn(通信作者);潘俊方(1994-),女,硕士生,主要研究方向为多标记学习、数据挖掘等;张霄雨(1993-),女,硕士生,主要研究方向为粒计算;王 伟(1975-),男,博士,讲师,主要研究方向为生物信息学;徐久成(1964-),男,博士,教授,CCF高级会员,主要研究方向为粒计算、数据挖掘、生物信息学等。

时具有多种感情色彩，如欢快、深沉等。这类数据的分类被称为多标记学习问题。因此，多标记学习在现实生活中被广泛应用，并逐渐引起了研究人员的关注[1-6]。

在研究多标记分类的过程中面临很多难题：一方面，每个样本可能同时具有多个类别标记，并且这些标记之间存在一定的关联性，如一幅图片标记有“荒岛”，那么它同时被标记为“荒漠”的可能性就要比它被标记为“海水”的可能性大。因此，在多标记问题的研究过程中，需要考虑标记之间的相关性。另一方面，与单标记学习一样，多标记分类通常也存在高维数据的情况。在多标记数据中，数据的高维性严重干扰了多标记分类器的分类性能[6]。高维数据中存在着许多冗余或不相关的特征，不仅会消耗较多的计算时间和空间，还容易给分类带来很多不利之处。因此，数据降维是机器学习中一项重要且具有挑战性的任务。而特征的降维技术可以在减少特征维数的基础上提高分类任务的效率和性能。数据降维主要分为特征提取和特征选择[3]。前者通过转换或映射方法将原始高维特征转换到一个新的低维特征空间。后者根据一定的评价准则，从原始特征空间中选取一组最优的特征子集。常见的评价准则有 Hamming Loss，Ranking Loss，One Error，Coverage 和 Average-Precision。

目前，已有很多种多标记数据的特征选择方法。例如：LDA[7]将多标记问题转化为单标记问题，但没有考虑到标记之间的相关性，将单标记数据的降维方法直接用来处理多标记数据；CCA[8]基于原始单标记数据降维方法，将多标记分类学习中的特征和标记作为看待样本的两个视角；PLS[9]类似于 CCA，通过使用核矩阵方法，在维度降低后获得与原始数据相同维度的新矩阵，但不能获得一个新的特征子空间；MDDM[10]通过映射降维和子空间降维两种映射策略进行降维，每种策略可以分别使用线性核和非线性核，并且该映射仍然使用核矩阵；MLNB[11]在主成分分析法和遗传算法的基础上采用贝叶斯分类法实现特征提取；MEFS[12]以预报风险的嵌入式特征选择方法为基础，通过对每个特征进行评价，最终获得最优特征子集，但是该方法与分类器和评价指标密切相关，从而导致时间复杂度高、降维效率低；MLFSIE[13]计算每一个特征与标签集合的信息增益值，并设定阈值以删除不相关的特征，但忽略了特征之间的相互关系[3]。

粗糙集理论在近年来受到了广泛的关注，目前已被应用于特征选择、模式识别、机器学习、数据挖掘及分类器设计等领域[14]。邻域粗糙集是经典粗糙集理论的扩展和延伸，由于粗糙集算法不能很好地处理数值型数据，并且常规的邻域粗糙集通常采用前向贪心算法，这使得它们不能很好地处理条件属性个数远多于样本数据的情况。近年来，单标记学习的特征选择（属性约简）得到了充分的研究[3-4,15]，但是多标记特征选择的研究成果相对较少。

胡清华等[15-16]重新定义了邻域粗糙集的下近似和依赖度，采用前向搜索算法进行属性约简，得到了特征子空间，但是这些算法无法保证得到的特征与相应标记间的相关性。鉴于这一问题，本文在邻域粗糙集理论的基础上提出了一种多标记邻域粗糙集模型及其专属特征选择算法，这样不仅从理论上确保了得到的专属特征与相应的标记具有较强的相关性，而且有效地提高了约简效率。该方法首先运用粗糙集约简理论来减少冗余属性，在保持分类能力与之前相比不变的条件下获得标记的专属特征，然后给出了邻域精确度和邻域粗糙度的概念，并重新定义了邻域粗糙集的属性依赖度和重要度的计算方法，提出了一种基于邻域粗糙集的多标记专属特征选择算法，以实现多标记分类任务的特征选择。在多个实例和公开数据集上的仿真实验验证了该算法的有效性。

2 邻域粗糙集

粗糙集是由波兰学者 Pawlak 于 1982 年提出的一种刻画不确定信息分类的数学工具，它不需要任何附加条件即可找到一个最小的、与全部属性具有相同区分能力的属性子集，通过属性约简可以大幅提高运算速度[17]。但是它在属性约简之前需要对连续型数据进行离散化处理，而在离散化的过程中会产生误差，改变原数据的属性特征，这在一定程度上影响了原属性集的信息表达，造成了信息的部分缺失，从而导致分类效果降低。胡清华等人提出了邻域粗糙集模型，该模型能够直接处理连续型数据，不再需要离散化处理，从而能避免在离散化过程中丢失信息。下面在文献[3-5]的基础上简要介绍邻域粗糙集模型的相关概念。

定义 1 $\langle U,\Delta\rangle$ 是非空度量空间，$x\in U$，$\delta\geqslant 0$，称点集 $\delta(x)=\{y\mid\Delta(x,y)\leqslant\delta,y\in U\}$ 为 x 的 δ 邻域。

定义 2 设 $U=\{x_1,x_2,\cdots,x_n\}$ 是由全部样本构成的集合，$A=\{a_1,a_2,\cdots,a_N\}$ 是描述样本的条件属性集，$D=\{l_1,l_2,\cdots,l_d\}$ 是分类决策属性集，给定 $\langle U,A,D\rangle$，如果 A 生成一组邻域关系，则称 $\langle U,A,D\rangle$ 为邻域决策系统。

定义 3 在邻域决策系统 $\langle U,A,D\rangle$ 中，D 将 U 划分为 N 个等价类 $\{X_1,X_2,\cdots,X_N\}$，$B\subseteq A$ 生成 U 上的邻域关系 N_B，那么决策 D 关于 B 的邻域下近似和上近似分别表示为：

$$\underline{N_B}D=\{\underline{N_B}X_1,\underline{N_B}X_2,\cdots,\underline{N_B}X_N\}$$

$$\overline{N_B}D=\{\overline{N_B}X_1,\overline{N_B}X_2,\cdots,\overline{N_B}X_N\}$$

于是，决策边界可表示为：

$$BN(D)=\overline{N_B}D-\underline{N_B}D$$

其中，$\underline{N_B}X_i=\{x\mid\delta(x)\subseteq X_i\}$，$\overline{N_B}X_i=\{x\mid\delta(x)\cap X_i\neq\varnothing\}$，$i=1,2,\cdots,N$。

3 基于邻域粗糙集的多标记专属特征选择

3.1 多标记学习框架

假定 $X=R_N$ 表示 N 维样本空间，$U=\{x_1,x_2,\cdots,x_n\}$，$L=\{l_1,l_2,\cdots,l_m\}$ 表示类别标记集合，$T=\{(x_i,y_i)\mid i=1,2,\cdots,n\}$ 表示在标记上的映射关系。每个样本 x 表示 N 维属性向量，$y=\{y^1,y^2,\cdots,y^m\}$ 表示样本 x 相对应的标记集合，当 x 含有标记 l_j 时，$y^j=1$，否则 $y^j=0$。下面给定一个实例，如表 1 和表 2 所列。在表 1 中，样本 x_1 具有 l_1 和 l_3 两个标记，在表 2 中的就可以标记 l_1 和 l_3 为 1，这表明样本 x_1 有两个标记。同理，表 1 中的样本 x_2 具有 l_1 和 l_2 两个标记，在表 2 中 l_1 和 l_2 下可以被标记为 1，但在 l_3 下被标记为 0，则表示样本 x_2 没有这个标记，x_3 与 x_4 同理。

表 1　多标记分类示例

Table 1　An example of multi-label task

U	y
x_1	l_1,l_3
x_2	l_1,l_2
x_3	l_2,l_3
x_4	l_1

表 2　多标记分类任务的二值表示

Table 2　Binary representation of multi-label task

U	l_1	l_2	l_3
x_1	1	0	1
x_2	1	1	0
x_3	0	1	1
x_4	1	0	0

3.2　多标记邻域粗糙集模型

单标记学习中,在邻域决策系统中邻域粗糙集的下近似通过借用邻域概念体现出属性集可以将样本进行分类;在多标记学习中,属性集可清晰地将样本分在每一类标记中,该能力通过多标记学习中邻域粗糙集的下近似形式进行表示[3]。下面给出多标记邻域粗糙集模型的相关概念及性质。

定义 4　在多标记邻域决策系统 $MNDT=\langle U,C\cup D\rangle$ 中,标记集合 $L=\{l_1,l_2,\cdots,l_m\}$,D^j 表示具有类别标记 l_j 的样本集合,D_i 表示样本 x_i 所具有的标记集合,给定 $B\subseteq C$,多标记邻域粗糙集的近似空间表示为:

$$\underline{N_B}D=\{x_i\in U\mid \forall\, l_j\in D_i,\delta_B(x_i)\subseteq D^j\}$$

$$\overline{N_B}D=\{x_i\in U\mid \forall\, l_j\in D_i,\delta_B(x_i)\cap D^j\neq\varnothing\}$$

$$BN_B(D)=\overline{N_B}D-\underline{N_B}D$$

定义 5　在多标记邻域决策系统 $MNDT=\langle U,C\cup D\rangle$ 中,决策属性 D 对论域 U 的划分记为 $U/D=\{X_1,X_2,\cdots,X_N\}$,对于任意条件属性子集 $B\subseteq C$,U/D 相对于 B 的邻域精确度表示为:

$$\rho=\frac{|\underline{N_B}D|}{|\overline{N_B}D|}$$

于是,U/D 相对于 B 的邻域粗糙度表示为:

$$rough(D)=1-\rho=1-\frac{|\underline{N_B}D|}{|\overline{N_B}D|}=\frac{|BN_B(D)|}{|\overline{N_B}D|}$$

粗糙度是一种重要的描述粗糙性的方法,反映了知识的不完备程度。

定义 6　在多标记邻域决策系统 $MNDT=\langle U,C\cup D\rangle$ 中,决策属性 D 对条件属性子集 $B\subseteq C$ 的依赖度可以表示为:

$$\gamma_B(D)=(1-\rho)\frac{|\underline{N_B}D|}{|U|}$$

性质 1　显然,$0\leqslant\gamma_B(D)\leqslant 1$。于是有:

(1)当 $\gamma_B(D)=1$ 时,D 对 B 是强依赖的;

(2)当 $0<\gamma_B(D)<1$ 时,D 对 B 是弱依赖的;

(3)当 $\gamma_B(D)=0$ 时,D 对 B 是完全不依赖的。

在多标记邻域决策系统中,上述依赖度的定义反映了决策属性对条件属性的重要程度。它不仅可以考查结果分类属性对条件属性的依赖程度,而且有助于发现对分类起决定作用的关键属性,从而达到特征选择和发现最小特征子集的目的。

定义 7　在多标记邻域决策系统 $MNDT=\langle U,C\cup D\rangle$ 中,$\forall\, a\in B\subseteq C$,若 $\gamma_B(D)\neq\gamma_{B-a}(D)$,则称 a 在 B 中相对决策属性 D 是必要的,否则是不必要的。

定义 8　在多标记邻域决策系统 $MNDT=\langle U,C\cup D\rangle$ 中,$B\subseteq C$,$a\subseteq B$。若:

(1)$\gamma_B(D)=\gamma_C(D)$

(2)$\gamma_{(B-a)}(D)<\gamma_B(D)$,$\forall\, a\in B$

则称 B 是 C 的一个属性约简。式中,$\gamma_B(D)$ 表示决策属性 D 对条件属性 B 的依赖度。

定义 9　在多标记邻域决策系统 $MNDT=\langle U,C\cup D\rangle$ 中,$B\subseteq C$,属性 $a\in C-B$ 在条件属性 B 上相对于决策属性 D 的重要度表示为:

$$sig(a,B,D)=\gamma_{B\cup\{a\}}(D)-\gamma_B(D)$$

定义 9 表明,属性 a 对于条件属性子集 B 的重要程度可以通过在 B 中添加 a 后引起的依赖度的变化来度量,变化越大,则 a 对于 B 就越重要。

例 1　在邻域决策系统中,x_1,x_2,x_3 和 x_4 是 4 个样本,假设它们关于 B 的邻域分别为:$\delta_B(x_1)=\{x_1,x_2\}$,$\delta_B(x_2)=\{x_1,x_2,x_3\}$,$\delta_B(x_3)=\{x_3\}$,$\delta_B(x_4)=\{x_2\}$,那么在单标记实例(如文献[3]的表 1 所列)学习中,根据定义 3 可以得到:

$$\underline{N_B}D=\{x_3\}\bigcup\{x_4\}=\{x_3,x_4\}$$

在多标记实例(见表 2)学习中,以 x_1 为例,有 $\delta_B(x_1)=\{x_1,x_2\}$,$y_1=\{l_1,l_3\}$,$y_2=\{l_1,l_2\}$,$y_3=\{l_2,l_3\}$,$D^1=\{x_1,x_2,x_4\}$,$D^2=\{x_2,x_3\}$,$D^3=\{x_1,x_3\}$。由于 $\delta_B(x_1)\not\subset D^2\wedge\delta_B(x_1)\not\subset D^3$,因此 $x_1\notin\underline{N_B}D$。

同理,可得 $y_2=\{l_1,l_2\}$,$\delta_B(x_2)\not\subset D^1\wedge\delta_B(x_2)\not\subset D^2$,故 $x_2\notin\underline{N_B}D$;$y_3=\{l_2,l_3\}$,$\delta_B(x_3)\subseteq D^2\wedge\delta_B(x_3)\subseteq D^3$,故 $x_3\in\underline{N_B}D$;$y_4=\{l_1\}$,$\delta_B(x_4)\subseteq D_1$,故 $x_4\in\overline{N_B}D$,因此有 $\underline{N_B}D=\{x_3,x_4\}$。

定义 10　在多标记邻域决策系统 $MNDT=\langle U,C\cup D\rangle$ 中,$B\subseteq C$,$\underline{N_B}D$ 也可称为属性 B 所给的知识水平下多标记分类的正域,记为 $POS_B(D)$。由此,多标记分类的依赖度可表示为:

$$\gamma_B(D)=(1-\rho)\frac{|\underline{N_B}D|}{|U|}=(1-\rho)\frac{|POS_B(D)|}{|U|}$$

于是,条件属性 $a\in C-B$ 在条件属性 B 上相对于决策属性集 D 的重要度可表示为:

$$sig(a,B,D)=\gamma_{B\cup\{a\}}(D)-\gamma_B(D)$$

从属性的依赖度角度来看,属性的重要度可以提供一种有效的特征选择方法。易证 $sig(a,B,D)\geqslant 0$,如果 $sig(a,B,D)=0$,则称属性 a 是多余的。在以下两种情况下认为属性是多余的:1)属性 a 与当前的分类任务无关;2)属性 a 所包含的分类信息已被包含在其他属性中,此时也称该属性是冗余的。

3.3　基于邻域粗糙集的多标记专属特征选择算法

在上述多标记邻域粗糙集模型的基础上,设计基于邻域粗糙集的多标记专属特征选择算法(NRS-MLSFS)。该算法包含两个主要步骤:1)使用文献[5]的正域约简启发式算法计算每个标记的正域约简,该约简对应于标记的专属特征;2)在多标记邻域粗糙集模型下,依次对每一个标记进行判断。

算法 1 基于邻域粗糙集的多标记专属特征选择算法

输入:多标记邻域决策系统⟨U,C∪D⟩和邻域半径 δ

输出:约简子集 red

Step1 预处理:对给定的数据集做插值拟合,进行特征值约简,对数据做归一化处理以消除数据的数量级差异;

Step2 对于决策系统⟨U,C∪D⟩,设定邻域半径 δ 以及约简条件属性子集相对于决策属性的重要度下限 efc=0.005;

Step3 初始化 Ø→red;

Step4 For $\forall a_k \in C-red$

{取 $\forall x_i \in U, \forall l_j \in D_i$;

If $\delta_{red}(x_i) \subseteq D^j$,then 计算 $\overline{N_{red}}D$;

If $\delta_{red}(x_i) \cap D^j \neq \emptyset$,then 计算 $\underline{N_{red}}D$;

计算属性的依赖度:

$$\gamma_{red}(D)=\frac{|BN(D)|}{|\overline{N_{red}}D|}\frac{|\underline{N_{red}}D|}{|U|}=(1-\rho)\frac{|POS_{red}(D)|}{|U|}$$

计算属性重要度 $sig(a_k,red,D)$,选择带有 $\max_k(sig(a_k,red,D))$ 的属性 a_k;

If $sig(a_k,red,D)>0$,then

$red\cup\{a_k\}\rightarrow red$;}

Step5 If 大于设定的重要度下限 efc,then 输出 red,else 返回 Step4。

假设多标记邻域决策系统中有 n 个样本、L 个标记和 N 个特征,特征子集有 k 个属性,则 NRS-MLSFS 算法的时间复杂度计算过程如下:Step1 预处理的时间复杂度为 $O(n)$;由文献[3]可知计算样本邻域的时间复杂度为 $O(n\log n)$,于是计算属性的依赖度的时间复杂度为 $O(Ln^2\log n)$,Step4 的最坏时间复杂度为 $O(NLn^2\log n)$。由此可计算出 NRS-MLSFS 算法总的最坏时间复杂度为 $O(n+NLn^2\log n)$,空间复杂度为 $O(n+L)$。

4 实验分析

4.1 数据集与实验环境

本节从 Mulanlibrary[1)] 上下载了 4 个多标记的分类任务,即 Emotions,Enron,Birds 和 Bibtex 数据集,来进行算法的仿真测试。表 3 给出了 4 个数据集的描述。实验采用的硬件配置为 CPU Intel i5-4200M 2.50GHz 和 4GB 内存,软件环境为 MatlabR2010b 和 Weka3.6.11。所有实验均采用十折交叉验证进行仿真测试。

表 3 数据集描述

Table 3 Description of datasets

数据集	训练集	测试集	特征	标记数
Emotions	391	202	72	6
Enron	1123	579	1001	53
Birds	322	323	260	19
Bibtex	4880	2515	1836	159

4.2 实验结果分析

不同算法可以使用多种分类模型,这里选择 Lazy,Rules,Trees 和 Bayes 4 种分类器,将 NRS-MLSFS 算法与文献[4]的 MDMR 算法、文献[15]的 FARNeMF 算法进行对比实验。实验首先使用特征降维算法对数据集进行降维,然后选择分类器,通过十折交叉验证对所选特征进行评价。利用邻域粗糙集理论去除初选特征子集中的冗余属性,在这里对邻域半径 δ 以及算法终止条件 efc 进行优化,根据多次实验结果的比较分析,在计算各样本的邻域时,设置 $\delta=0.15$ 和 $efc=0.005$。

为了验证 NRS-MLSFS 算法的有效性,表 4 列出了 3 种算法在不同数据集上选择的特征个数和分类精度。表 5 列出了 FARNeMF 算法在 3 种不同分类器上选择的特征个数和分类精度。表 6 列出了 NRS-MLSFS 算法在 3 种不同分类器上选择的特征个数和分类精度。表 7 列出了两种不同算法在相同分类器上的分类精度。

表 4 3 种算法在不同数据集上选择的特征个数和分类精度的比较

Table 4 Comparison of the selected gene number and classification accuracy of three algorithms on different datasets

数据集	MDMR		FARNeMF		NRS-MLSFS	
	特征个数	分类精度/%	特征个数	分类精度/%	特征个数	分类精度/%
Emotions	26	72.357	3	93.861	16	94.799
Enron	10	75.586	7	93.855	4	95.785
Birds	32	71.363	2	88.021	60	91.316
Bibtex	49	87.245	8	91.919	19	99.296

表 5 FARNeMF 算法在 3 种分类器上选择的特征个数和分类精度的比较

Table 5 Comparison of the selected gene number and classification accuracy of FARNeMF algorithm with three classifiers

数据集	Lazy		Rules		Trees	
	特征个数	分类精度/%	特征个数	分类精度/%	特征个数	分类精度/%
Emotions	3	92.5831	3	94.8849	3	94.1176
Enron	7	93.5886	7	92.2529	7	95.7257
Birds	2	90.7671	2	84.532	2	88.763
Bibtex	8	87.5637	8	93.6013	8	94.593

表 6 NRS-MLSFS 算法在不同分类器上所选的特征个数和分类精度的比较

Table 6 Comparison of the selected gene number and classification accuracy of NRS-MLSFS algorithm with different classifiers

数据集	Rules		Trees		Bayes	
	特征个数	分类精度/%	特征个数	分类精度/%	特征个数	分类精度/%
Emotions	16	94.3734	16	95.1407	16	94.8849
Enron	4	95.7257	4	96.0819	4	95.5476
Birds	60	88.9743	60	90.0621	60	94.912
Bibtex	19	99.2623	19	99.2828	19	99.3443

表 7 两种算法在相同分类器上的分类精度的比较/%

Table 7 Comparison of classification accuracy of two algorithms with the same classifier/%

数据集	Rules		Trees	
	FARNeMF	NRS-MLSFS	FARNeMF	NRS-MLSFS
Emotions	94.8849	94.3734	94.1176	95.1407
Enron	92.2529	95.7257	95.7257	96.0819
Birds	84.532	88.9743	88.763	90.0621
Bibtex	93.6013	99.2623	94.593	99.2828

由表 4 可知,MDMR 算法虽然可以得到较高的分类精度,但选择的特征数量过大;FARNeMF 算法可以有效地去除不相关的特征子集,但是在去除冗余特征的过程中也剔除了

1) http://mulan.sourceforge.net/datasets.html

与分类相关的一些特征,从而导致分类精度低于 NRS-MLSFS 算法;而较为理想的特征选择方法不仅可以获得规模较小的特征子集,而且具有较高的分类精度。例如,在 Bibtex 数据集上,虽然选择的特征个数略多,但是分类精度达到了 99.296%。从表 4 的结果可知,虽然 NRS-MLSFS 算法约简后的特征个数不少于 FARNeMF 算法的特征个数,但是 NRS-MLSFS 算法在 4 个数据集上均得到了较高的分类精度。

从表 5 的实验结果可以看出,同一种算法在不同分类器上的分类精度略有不同。例如,FARNeMF 算法在 Birds 数据集上通过 Lazy 分类器得到的分类精度比其他两种算法略高;而在 Emotions 数据集上,3 种分类器得到的分类精度差别很小;在 Enron 数据集上,Trees 分类器得到的分类精度最高。同样,从表 6 的实验结果可以看出,Trees 分类器在不同数据集上表现出的分类性能各有高低。例如:在 Bibtex 数据集上,Trees 分类器得到的分类精度最高;而在 Enron 数据集上,Bayes 和 Rules 得到的分类精度基本相同。于是可以得出,NRS-MLSFS 算法在不同数据集上的分类精度均略高于 FARNeMF 算法。

从表 7 的实验结果可以看出,针对同一数据集,采用不同的算法在同一分类器上得到的分类精度也是有差异的。例如,在 Enron 数据集上,两种算法在 Trees 分类器上得到的分类精度大致相同,而在 Rules 分类器上 NRS-MLSFS 算法得到的分类精度高于 FARNeMF 算法;在 Emotions 数据集上,NRS-MLSFS 算法在 Rules 分类器上得到的分类精度略低于 FARNeMF 算法。同时,每种算法都有各自适应的分类器,不能因为一组数据而否定整个算法,只要算法在各种分类器上的总体性能优于其他算法即可。分析表 7 的实验结果可知,NRS-MLSFS 算法在 Rules 和 Trees 分类器上的分类精度总体优于 FARNeMF 算法。因此,本文算法在选择特征子集数量和分类精度上均能取得较好的效果。

综上所述,基于多标记邻域粗糙集模型的专属特征选择算法能够选择出特征数量较少的特征子集,而且其分类精度也高于其他相关算法。这说明 NRS-MLSFS 算法能够选择出信息含量高的特征子集,也能有效地减少所选特征子集的无关性,有助于解决多标记分类任务的高维数、高冗余问题,提高多标记分类问题的精度和效率。

结束语　本文在胡清华等人提出的邻域粗糙集的基础上,设计了一种多标记专属特征选择模型及其算法。该模型不仅从理论上确保了所得到的专属特征与相应的标记具有较强的相关性,而且也有效地改善了约简效果。同时,由粗糙集约简理论获得的专属特征是原始特征集合的一个子集,有效反映了专属特征的直观意义;其次,重新定义了邻域粗糙集的依赖度和重要度概念,有助于选择较优的特征子集。本文通过实例和一系列的对比实验验证了 NRS-MLSFS 算法的有效性。同时,针对多标记之间的相关性,下一步将在考虑标记之间关联性的基础上深入研究专属特征的多标记学习模型及其分类算法。

参 考 文 献

[1] LI F, MIAO D Q, PEDRYCZ W. Granular multi-label feature selection based on mutual information[J]. Pattern Recognition, 2017, 67(C): 410-423.

[2] HYUNKI L, JAESUNG L, DAE-WON K. Optimization approach for feature selection in multi-label classification[J]. Pattern Recognition Letters, 2017, 89(C): 25-30.

[3] DUAN J, HU Q H, ZHANG L J, et al. Feature selection for multi-label classification based on neighborhood rough sets[J]. Journal of Computer Research and Development, 2015, 52(1): 56-65. (in Chinese)
段洁,胡清华,张灵均,等. 基于邻域粗糙集的多标记分类特征选择算法[J]. 计算机研究与发展,2015,52(1):56-65.

[4] LIN Y J, HU Q H, LIU J H, et al. Multi-label feature selection based on max-dependency and min-redundancy[J]. Neurocomputing, 2015, 168(C): 92-103.

[5] LI H, LI D Y, WANG S G, et al. Multi-label learning with label-specific features based on rough sets[J]. Journal of Chinese Computer Systems, 2015, 36(12): 2730-2734. (in Chinese)
李华,李德玉,王素格,等. 基于粗糙集的多标记专属特征学习算法[J]. 小型微型计算机系统,2015,36(12):2730-2734.

[6] LIU J H, LIN M L, WANG C X, et al. Multi-label feature selection algorithm based on local subspace[J]. Pattern Recognition & Artificial Intelligence, 2016, 29(3): 240-251. (in Chinese)
刘景华,林梦雷,王晨曦,等. 基于局部子空间的多标记特征选择算法[J]. 模式识别与人工智能,2016,29(3):240-251.

[7] SUN L, JI S W, YE J P. Multi-Label Dimensionality Reduction [M]. Florida: CRC Press, 2013: 20-22.

[8] FISHER R A. The use of multiple measurements in taxonomic problems[J]. Annals of Human Genetics, 1936, 7(2): 179-188.

[9] WOLD H. Estimation of principal components and related models by iterative least squares[J]. Multivariate Analysis, 1966 (1): 391-420.

[10] ZHANG Y, ZHOU Z H. Multi-label dimensionality reduction via dependence maximization[J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(3): 14-20.

[11] ZHANG M L, PENA JOSÉ M, ROBLES V. Feature selection for multi-label naïve Bayes classification[J]. Information Sciences, 2009, 179(19): 3218-3229.

[12] GE L, LI G Z, YOU M Y. Embedded feature selection for multi-label learning[J]. Journal of Nanjing University (Natural Sciences), 2009, 45(5): 671-676. (in Chinese)
葛雷,李国正,尤鸣宇. 多标记学习的嵌入式特征选择[J]. 南京大学学报(自然科学),2009,45(5):671-676.

[13] ZHANG Z H, LI S N, LI Z G, et al. Multi-label feature selection algorithm based on information entropy[J]. Journal of Computer Research and Development, 2013, 50(6): 1177-1184. (in Chinese)
张振海,李士宁,李志刚,等. 一种基于信息熵的多标签特征选择算法[J]. 计算机研究与发展,2013,50(6):1177-1184.

[14] SUN L, LIU R N, ZHANG X Y, et al. A fuzzy biclustering approach based on rough mean square residue[J]. Journal of Henan Normal University (Natural Science Edition), 2017, 45(5): 93-100. (in Chinese)

孙林,刘弱南,张霄雨,等. 一种基于粗糙均方残基的模糊双聚类方法[J]. 河南师范大学学报(自然科学版),2017,45(5):93-100.

[15] HU Q H,ZHAO H,YU D R. Efficient symbolic and numerical attribute reduction with neighborhood rough sets[J]. Pattern Recognition & Artificial Intelligence,2008,21(6):732-738. (in Chinese)
胡清华,赵辉,于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法[J]. 模式识别与人工智能,2008,21(6):732-738.

[16] HU Q H,YU D R,LIU J F,et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences,2008,178(18):3577-3594.

[17] XUE Z A,WANG N,SI X M,et al. Research on multi-granularity rough intuitionistic fuzzy cut sets[J]. Journal of Henan Normal University (Natural Science Edition),2016,44(5):131-139. (in Chinese)
薛占熬,王楠,司小朦,等. 多粒度粗糙直觉模糊截集的研究[J]. 河南师范大学学报(自然科学版),2016,44(5):131-139.

(上接第 147 页)

所有实验均在经过词语级分词处理的语料上进行,基于分类的模型得到了较好的结果。

结束语 本文构建了基于分类的文本摘要模型,该模型将基于递归神经网络的编码-解码机构与分类器相结合,并在大量的语料下同时训练优这两部分,从而在文本摘要任务中取得了优异的性能。但在模型中低频词的问题仍然存在,特别是在词语级的实验中,本文通过使用字符级的分词处理来解决该问题。近年来,研究者们提出了神经图灵机(Neural Turing Machines,NTM)[22]的概念,其已在许多问题中表现出了极佳的性能,在今后的工作中将尝试把神经图灵机融入到文本摘要模型中,以获得更好的文本摘要性能。

参 考 文 献

[1] GAMBHIR M,GUPTA V. Recent automatic text summarization techniques: a survey [J]. Artificial Intelligence Review,2017,47(1):1-66.

[2] LUONG M T,PHAM H,MANNING C D. Effective Approaches to Attention-based Neural MachineTranslation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon:ACL, 2015:1412-1421.

[3] GRAVES A,MOHAMNED A R,HINTON G. Speech recognition with deep recurrent neural networks IEEE[C]// International Conference on Acoustics,Speech and Signal Processing. IEEE,2013:6645-6649.

[4] GAMBHIR M,GUPTA V. Recent automatic text summarization techniques: a survey[J]. Artificial Intelligence Review,2017,47(1):1-66.

[5] RUSH A M,CHOPRA S,WESTON J. A Neural Attention Model for Abstractive Sentence Summarization [C]// Proceedings of NAACL. 2016.

[6] BENGIO Y,SCHWENK H,SENÉCAL J,et al. Neural Probabilistic Language Models[J]. Journal of Machine Learning Research,2006,3(6):1137-1155.

[7] HU B,CHEN Q,ZHU F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset [C]// Procee-dings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon:ACL,2015:1967-1972.

[8] LOPYREV K. Generating News Headlines with Recurrent Neural Networks[J]. Computer Science,2015.

[9] CHOPRA S,AULI M,RUSH A M. Abstractive Sentence Summarization with Attentive Recurrent [C]// Neural Networks Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016:93-98.

[10] NALLAPATI R,ZHOU B,SANTOS C N D,et al. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond[J]. CoNLL,2016,1(1):280-290.

[11] SURHONE L M,TENNOE M T,HENSSONOW S F. Long Short Term Memory [C]// Betascript Publishing. 2010.

[12] CHO K,VAN M B,GULCEHRE C,et al. Learning Phrase Representations using RNN Encoder—Decoder for Statistical Machine Translation[J]. Computer Science,2014,1(1):43-66.

[13] SUTSKEVER I,VINYALS O,LE Q V. Sequence to Sequence Learning with Neural Networks[J]. Advances in Neural Information Processing Systems,2014,4:3104-3112.

[14] BENGIO S,VINYALS O,JAITLY N,et al. Scheduled sampling for sequence prediction with recurrent Neural networks[C]// International Conference on Neural Information Processing Systems. MIT Press,2015:1171-1179.

[15] BAHDANNAU D,CHO K,BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]// Proceedings of Neural Information Processing Systems 2014. NIPS,2014.

[16] VINYALS O,KAISER L,KOO T,et al. Grammar as a foreign language[J]. Eprint Arxiv,2014,1(1):2773-2781.

[17] LIN C Y,HOVY E. Automatic evaluation of summaries using N-gram co-occurrence statistics[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics,2003:71-78.

[18] Google. Tensorflow(Version 1.2)[OL]. http://www.tensorflow.org.

[19] KOEHN P,HOANG H,BIRCH A,et al. Moses:Open source toolkit for statistical machine translation[C]// Proceedings of ACL. 2007:177-180.

[20] CHOPRA S,AULI M,RUSH A M,et al. Abstractive sentence summarization with attentive recurrent neural networks[C]// Proceedings of NAACL. 2016.

[21] ELMAN J L. Finding structure in time[J]. Cognitive Science,1990,14(2):179-211.

[22] GRA VES A,WAYNE G,DANIHELKA I. Neural Turing Machines[J]. Computer Science,2014,1(1):89-95.