DOI: 10.11992/tis.201807027
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190525.1801.002.html

代价敏感数据的多标记特征选择算法

黄琴[1,2], 钱文彬[1,2], 王映龙[1], 吴兵龙[2]
（1. 江西农业大学 计算机与信息工程学院, 江西 南昌 330045; 2. 江西农业大学 软件学院, 江西 南昌 330045）

摘　要: 在多标记学习中, 特征选择是提升多标记学习分类性能的有效手段。针对多标记特征选择算法计算复杂度较大且未考虑到现实应用中数据的获取往往需要花费代价, 本文提出了一种面向代价敏感数据的多标记特征选择算法。该算法利用信息熵分析特征与标记之间的相关性, 重新定义了一种基于测试代价的特征重要度准则, 并根据服从正态分布的特征重要度和特征代价的标准差, 给出一种合理的阈值选择方法, 同时通过阈值剔除冗余和不相关特征, 得到低总代价的特征子集。通过在多标记数据的实验对比和分析, 表明该方法的有效性和可行性。

关键词: 特征选择; 属性约简; 代价敏感; 粗糙集; 粒计算; 多标记学习; 信息熵; 正态分布
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)05−0929−10

中文引用格式: 黄琴, 钱文彬, 王映龙, 等. 代价敏感数据的多标记特征选择算法 [J]. 智能系统学报, 2019, 14(5): 929–938.
英文引用格式: HUANG Qin, QIAN Wenbin, WANG Yinglong, et al. Multi-label feature selection algorithm for cost-sensitive data[J]. CAAI transactions on intelligent systems, 2019, 14(5): 929–938.

Multi-label feature selection algorithm for cost-sensitive data

HUANG Qin[1,2], QIAN Wenbin[1,2], WANG Yinglong[1], WU Binglong[2]
(1. School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China; 2. School of Software, Jiangxi Agricultural University, Nanchang 330045, China)

Abstract: In multi-label learning, feature selection is an effective means to improve multi-label learning classification performance. Aiming at the problem that the existing multi-label feature selection methods have high computation complexity and do not consider the cost of data acquisition in real-world applications, this paper proposes a multi-label feature selection algorithm for cost-sensitive data. The algorithm first analyzes the relevance between the feature and label based on information entropy, and redefines a criterion for feature significance by employing feature test cost; it then gives a reasonable threshold selection method on the basis of the standard deviation of feature significance and feature cost that obey normal distribution. At the same time, the algorithm derives the feature subsets with low total cost by removing redundant and irrelevant features according to a threshold. Finally, the effectiveness and feasibility of the proposed algorithm are verified by the comparison and analysis of the experimental results on a multi-labeled dataset.
Keywords: feature selection; attribute reduction; cost-sensitive; rough sets; granular computing; multi-label learning; information entropy; normal distribution

随着物联网及信息技术的发展, 数据资源呈海量特征。在数据量不断增大的同时, 数据标注结构的复杂度也在增加, 传统的单标记学习已不能满足现实应用的需求, 因此多标记学习的重要性逐渐突显。在多标记学习中, 每个样本在一个特征向量下, 可能同时隶属于多个类别标记。近年来, 多标记学习问题已成为机器学习、数据挖掘和模式识别等领域的研究热点之一[1-4]。

波兰数学家 Pawlak 教授于 1982 年提出的粗糙集理论是一种用于处理不精确、不完全和不相

收稿日期: 2018−07−26.　**网络出版日期:** 2019−05−27.
基金项目: 国家自然科学基金项目 (61502213, 61662023); 江西省自然科学基金项目 (20161BAB212047); 江西省教育厅科技项目 (GJJ180200).
通信作者: 钱文彬. E-mail: qianwenbin1027@126.com.

容知识的数学工具[5], 近年来, 该理论在机器学习和数据挖掘领域得到了广泛的应用[6-7]。属性约简, 又称特征选择, 是粗糙集理论的核心内容之一, 其目的是在保持分类能力不变的条件下, 删除不相关或冗余特征。与单标记学习一样, 多标记学习也面临着“维数灾难”的挑战。高维数据不仅影响算法的执行效率, 也降低了分类器的分类性能, 而特征降维技术是解决维数灾难的有效方法。目前, 针对单标记数据特征降维技术的研究较为广泛, 而针对多标记数据特征降维技术的研究相对较少。因此, 基于多标记学习特征选择的研究具有重要的理论和应用意义。另外, 在现实应用领域中, 数据特征的获取往往需要花费一定的代价, 为此从代价敏感的视角研究多标记特征选择问题显得尤为重要。

1　相关工作

近年来, 在多标记特征提取方面已经取得一些有意义的研究成果。如 Sun 等[8] 提出的多标记降维方法 (LDA), 其直接将单标记特征降维的方法应用于多标记特征降维中, 忽略了标记之间的相关性。Zhang 等[9] 采用核矩阵进行映射降维, 设计了一种最大化依赖度的多标记特征降维方法 (MDDM)。Yu 等[10] 提出了一种有监督的多标记潜在语义索引降维方法 (MLSI)。多标记特征提取能够实现特征降维的效果, 但由于其忽略了标记之间的关联以及损失了原始特征的物理含义, 这对多标记学习问题的研究造成了较大的困难。

多标记特征选择通过设计特征度量准则从原始特征中剔除冗余或不相关特征, 得到一组相对最优的特征子集, 从而可有效降低特征空间的维数, 提升算法的分类性能。特征选择的结果能够保持原始特征的物理含义, 使得多标记学习的研究更容易理解。目前许多研究人员针对多标记特征选择开展研究, 段洁等[11] 重新定义了多标记邻域粗糙集的下近似和依赖度的计算方法, 在此基础上, 设计了一种基于邻域粗糙集的特征选择算法 (ARMLNRS)。王晨曦等[12] 从每个标记对样本不同分组的角度出发, 提出了基于信息粒化的多标记特征选择算法 (MFIG)。Lin 等[13] 在乐观、中立和悲观这 3 种不同的视角下, 通过 3 种基于邻域互信息准则进行特征选择。刘景华等[14] 通过引入局部子空间模型, 构建了一种基于局部子空间的多标记特征选择算法 (MFSLS)。上述算法的计算复杂度相对较大。后来 Lee 等[15] 通过特征信息熵之差最大化和正向搜索的方法选择特征子集, 设计时间复杂度较低的特征选择算法, 但其没有给出和分析的信息熵阈值对特征子集的影响。张振海等[16] 利用信息增益下的阈值选择设计了一种多标记特征选择算法 (MLFSIE)。综上所述, 这些多标记特征选择算法并未考虑到特征的代价敏感问题。

在许多实际应用领域中, 获取和采集数据是需要花费代价的, 因此从代价敏感的视角研究多标记学习具有重要的意义。针对当前多标记特征选择算法的计算复杂度较大且未考虑特征代价的问题, 提出了一种面向代价敏感数据的多标记特征选择算法。首先, 该方法计算出特征与标记集合之间的信息增益, 在此基础上重新定义了特征重要度的计算方法, 并根据服从正态分布的特征重要度与特征代价的标准差之间的差值, 提出了一种合理的阈值选择方法, 从而实现对冗余或不相关特征的剔除, 同时能得到总代价较低的特征子集。为了验证算法的有效性, 利用 Mulan 平台上的真实多标记数据集进行实验比较和分析, 通过实验结果进一步验证算法的有效性和可行性。

2　基本知识

2.1　多标记学习

在粒计算理论中, 多标记数据可表示成一个多标记决策表 $\mathrm{MDT}=(U,A\cup D,V,f)$, 其中: U 为样本集 $\{x_1,x_2,\cdots,x_n\}$, 也称为论域; A 为条件特征集 $\{a_1,a_2,\cdots,a_m\}$; D 为多标记决策特征 $\{l_1,l_2,\cdots,l_k\}$, 且 $A\cap D=\varnothing$; V 为全特征集的值域, 其中 $V=\cup V_a$, $a\in A\cup D$, V_a 表示特征 a 的值域; f 是 $U\times(A\cup D)\to V$ 的信息函数。

定义 1　给定多标记决策表 $\mathrm{MDT}=(U,A\cup D,V,f)$, 对于 $\forall a\in A$, 特征 a 的等价关系 R_a 为

$$R_a=\{(x_i,x_j)\in U\times U, f(x_i,a)=f(x_j,a)\}$$

定义 2　给定多标记决策表 $\mathrm{MDT}=(U,A\cup D,V,f)$, 对于 $\forall l_t\in D$, 标记 l_t 的等价关系 R_{l_t} 为

$$R_{l_t}=\{(x_i,x_j)\in U\times U, f(x_i,l_t)=f(x_j,l_t)\}$$

2.2　信息熵

基于条件信息熵下的特征选择是研究者从信息观视角对高维数据进行特征选择, 该方法可有效地度量信息的不确定性程度。

定义 3　给定多标记决策表 $\mathrm{MDT}=(U,A\cup D,V,f)$, 对于任意特征子集 $B\subseteq A$, 根据特征子集 B 的等价关系 R_B 可得 $U/B=\{X_1,X_2,\cdots,X_q\}$, 则特征子集 B 的信息熵为

$$H(B) = -\sum_{i=1}^{q} p(X_i)\log p(X_i)$$

当信息熵 $H(B)$ 的值越大, 说明特征子集 B 的不确定性越大。

定义 4　给定多标记决策表 MDT = $(U, A\cup D, V, f)$, 对于任意标记子集 $L\subseteq D$, 根据标记子集 L 的等价关系 R_L 可得 $U/L=\{Y_1, Y_2, \cdots, Y_p\}$, 则在特征子集 B 下标记子集 L 的条件熵为

$$H(L|B) = -\sum_{i=1}^{q} p(X_i)\sum_{j=1}^{p} p(Y_j|X_i)\log\ p(Y_j|X_i)$$

由定义 4 可知, 当 $H(L|B)$=0 时, 说明标记子集 L 完全依赖于特征子集 B, 当 $H(L|B)$=$H(L)$ 时, 表明标记子集 L 独立于特征子集 B。

定义 5　给定多标记决策表 MDT = $(U, A\cup D, V, f)$, 对于任意特征子集 $B\subseteq A$, 则标记子集 L 在特征子集 B 上的信息增益为

$$\mathrm{IG}(L|B) = H(L) - H(L|B)$$

信息增益 IG($L|B$) 值用于衡量特征子集 B 与标记子集 L 的相关程度, IG($L|B$) 值越大, 说明其特征子集 B 与标记子集 L 的相关程度越大。

为了使得各个特征与标记之间的信息增益值在同一量纲下比较, 需先对信息增益的值进行归一化处理:

$$\mathrm{NIG}(L|B) = \frac{\mathrm{IG}(L|B)}{H(B)+H(L)}$$

3　代价敏感下的多标记学习

3.1　基于特征代价的信息熵模型

在机器学习和数据挖掘领域, 代价敏感学习是十大最具有挑战性问题之一[17]。因此, 将特征代价引入到多标记特征选择具有重要的意义。

定义 6　给定基于测试代价的多标记决策表 CMDT = $(U, A\cup D, V, f, c)$, 其中 $c: A\to R^+\cup\{0\}$ 为测试代价函数, 对于任意特征 $\forall a_i, a_j\subseteq A$, 标记集合 D 在特征 a_i 上的特征重要度为

$$\mathrm{CSIG}(D|a_i) = \mathrm{NIGS}(D|a_i)^* - \mathrm{Cost}(a_i)^*$$

由定义 5 可得, 标记集合 D 在特征 a_i 上的信息增益为

$$\mathrm{NIGS}(D|a_i) = \sum_{t=1}^{k}\mathrm{NIG}(l_t|a_i)$$

为了获取合理的阈值, 使得信息增益的值服从正态分布:

$$\mathrm{NIGS}(D|a_i)^* = \frac{\mathrm{NIGS}(D||a_i) - \mu}{\sigma}$$

式中: μ 表示特征与标记集合的信息增益均值; σ 表示特征与标记集合的信息增益标准差, 其公式分别为

$$\mu = \frac{1}{k}\left(\sum_{t=1}^{k}\mathrm{NIG}(l_t|a_i)\right)$$

$$\sigma = \sqrt{\frac{1}{k}\sum_{t=1}^{k}[\mathrm{NIG}(l_t|a_i)-\mu]^2}$$

在计算测试代价下的标记集合下特征重要度之前, 需先将特征代价进行归一化处理:

$$\mathrm{Cost}(a_i)^* = \frac{\mathrm{Cost}(a_i) - \min(\mathrm{Cost}(a_j))}{\max(\mathrm{Cost}(a_j)) - \min(\mathrm{Cost}(a_j))}$$

式中: $\max(\mathrm{Cost}(a_j))$ 表示特征的测试代价最大值; $\min(\mathrm{Cost}(a_j))$ 表示特征的测试代价最小值。

定义 7　给定基于测试代价的多标记决策表 CMDT = $(U, A\cup D, V, f, c)$, 其阈值 δ 定义为

$$\delta = \frac{1}{m}\sum_{i=1}^{m}|\mathrm{CSIG}(D|a_i)| - \mathrm{Cost}.\sigma$$

式中: Cost.σ 表示所有特征的测试代价标准差, 其公式为

$$\mathrm{Cost}.\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}[\mathrm{Cost}(a_i) - \mathrm{Cost}.\mu]^2}$$

$$\mathrm{Cost}.\mu = \frac{1}{m}\left(\sum_{i=1}^{m}\mathrm{Cost}(a_i)\right)$$

式中: Cost.μ 表示所有特征的测试代价均值。

3.2　基于特征代价的信息熵模型可行性分析

性质 1　若特征 a 与标记 l_t 相互独立, 则特征 a 与标记 l_t 之间的信息增益取最小值; 若标记 l_t 完全依赖于特征 a, 则特征 a 与标记 l_t 的信息增益取最大值。

证明　由信息论理论结合定义 4 和定义 5 可推导出, $H(L|B)\geqslant 0$, 且 IG($L|B$) $\leqslant H(L)$。当 $H(L|B)$=0 时, IG($L|B$)=$H(L)$, 信息增益的值最大; 当 $H(L|B)$=$H(L)$ 时, IG($L|B$)=0, 此时信息增益的值最小, 同时可知, 信息增益 IG($L|B$) 值具有非负性。

对于任意特征 $a\in A$, $l_t\in D$, 由定义 5 可知, IG($l_t|a$) = $H(l_t) - H(l_t|a)$, 由定义 3 和定义 4 可推导出, $\mathrm{IG}(l_t|a) = \sum_{i=1}^{q} p(X_i)\log p(X_i) - \sum_{i=1}^{q} p(X_i)\sum_{j=1}^{p} p(Y_j|X_i)\log p(Y_j|X_i)$, 且由上述推导可知, 当 $H(l_t|a)$=$H(l_t)$ 时, IG($l_t|a$) = 0, 信息增益的值最小, 此时 $\log p(X_i) = \sum_{j=1}^{p} p(Y_j|X_i)\log p(Y_j|X_i)$, 表明标记 l_t 独立于特征 a。同理, 当 $H(l_t|a)=0$ 时, 此时 IG($l_t|a$)=$H(l_t)$, 信息增益 IG($l_t|a$) 最大, 即 $\log p(X_i) - \sum_{j=1}^{p} p(Y_j|X_i)\log p(Y_j|X_i)$ 最大, 当 $\sum_{j=1}^{p} p(Y_j|X_i)|\log p(Y_jX_i)$=0 时, 表明标记 l_t 完全依赖于特征 a。

性质 2　标记集合 D 在特征 a_i 上的特征重要

度具有单调性，即标记 D 在特征 a_i 上的信息增益随单个标记在特征 a_i 上的信息增益的增大而增大，且标记与特征之间的相关程度越大。

证明　由性质 1 可得，若单个标记与特征 a_i 的相关性越大，则信息增益值越大，即 $\mathrm{IG}(l_t|a)$ 越大。由定义 6 可知，$\mathrm{NIGS}(D|a)=\sum_{t=1}^{k}\mathrm{NIG}(l_t|a_i)$，因此 $\mathrm{NIG}(l_t|a_i)$ 越大，则 $\mathrm{NIGS}(D|a)$ 的值越大，而特征代价的值 $\mathrm{Cost}(a_i)^*$ 是固定的，此时可得$\mathrm{CSIG}(D|a_i)=\mathrm{NIGS}(D|a_i)^*-\mathrm{Cost}(a_i)^*$ 越大，即标记 D 在特征 a_i 上的信息增益随单个标记在特征 a_i 上的信息增益的增大而增大。由性质 1 可知，标记与特征之间信息增益越大，则其相关程度也越大。由此可得，标记集合 D 在特征 a_i 上的信息增益具有单调性。

性质 3　阈值 δ 具有单调性，即阈值随标记集合 D 在特征 a_i 上的信息增益值的增大而增大。

证明　由性质 2 和定义 7 可知，特征代价的标准差 $\mathrm{Cost}.\sigma$ 的值是固定的，$\mathrm{CSIG}(D|\mathrm{a}_i)$ 的值越大，则 $\frac{1}{m}\sum_{i=1}^{m}|\mathrm{CSIG}(D|a_i)|-\mathrm{Cost}.\sigma$ 越大，即阈值 δ 的值越大。

4　多标记特征选择算法

4.1　算法描述

根据上述分析可知，在多标记学习算法中，一个特征不仅与某一标记具有相关性，也可能同时与多个标记具有相关性，因此需要计算单个特征与标记集合之间的相关性。在此基础上，从代价敏感学习的视角，提出了一种基于测试代价的特征重要度；然后根据服从正态分布的特征重要度以及特征代价的标准差设计出一种合理的阈值选择方法；最后，通过计算的阈值删除冗余或不相关的特征。

本文提出的代价敏感数据的多标记特征选择算法 (CSMLFSIE) 具体步骤如下：

算法　代价敏感数据的多标记特征选择算法 (CSMLFSIE)

输入　多标记决策表 $<U,A\cup D,V,f>$；

输出　特征子集 Red。

1) 初始化 $\mathrm{Red}\leftarrow\varnothing$；

2) 对于 $\forall a\in A$, $\forall l_t\in D$, 执行操作：

①计算在特征集 A 下每个特征的信息增益 $H(a)$；

②每个特征相对于每个标记的条件信息熵 $H(l_t|a)$；

3) 对于 $\forall a\in A$, $\forall l_t\in D$分别计算每个特征相对于每个标记的信息增益 $\mathrm{IG}(l_t|a)$；

4) 对于 $\forall a\in A$, 执行操作：

①计算标记集合下每个特征的重要度 $\mathrm{CSIG}(D|a)$；

②计算阈值 δ；

5) 对于 $\forall a\in A$, 执行操作：

若 $\mathrm{CSIG}(D|a)>\delta$, 则 $\mathrm{Red}\leftarrow\mathrm{Red}\cup\{a\}$

6) 输出特征子集 Red, 算法结束。

4.2　时间复杂度分析

代价敏感数据的多标记特征选择算法中：步骤 1) 初始化一个变量存放特征选择后的特征子集，其时间复杂度为 $O(1)$；步骤 2) 中①需利用基数排序[18] 计算等价类，则整个条件特征集每个特征的信息熵的时间复杂度为 $O(|A||U|)$，步骤 2) 中②计算每个特征的条件信息熵的时间复杂度为 $O(|U/A||U||D|)$，可知步骤 2) 的时间复杂度最坏为 $O(|A||U||D|)$；步骤 3) 分别计算每个标记与每个特征之间的信息增益，其时间复杂度为 $O(|A||U||D|)$；步骤 4) 中①计算标记集合下每个特征重要度，其时间复杂度为 $O(|A||U||D|)$，步骤 4) 中②计算阈值的时间复杂度为 $O(|A||U|)$，因此步骤 4) 的时间复杂度最坏为 $O(|A||U||D|)$；步骤 5) 根据阈值进行特征选择其时间复杂度为 $O(|A|)$。因此本文算法的时间复杂度为 $O(|A||U||D|)$。

为了分析本文算法在计算复杂度上的优越性，将本文算法分别与 CSMLPA[19] 算法和 MLDM 算法进行比较。CSMLPA 算法是基于文献 [20] 的正区域模型设计的，并且考虑了测试代价的多标记特征选择算法，算法采用的是向前启发式搜索策略，其计算复杂度主要消耗在计算加入单个特征到特征子集后的正域大小，时间复杂度为 $O(|A|^2|U||D|)$。MLDM 算法是基于文献 [21] 的差别矩阵方法改进的多标记特征选择算法，该算法主要耗时在对实例进行两两比较，其时间复杂度为 $O(|A||U|^2|D|)$。本文算法与 CSMLPA 算法和 MLDM算法相比，时间复杂度由非线性 $O(|A|^2|U||D|)$ 和 $O(|A||U|^2|D|)$ 降低至线性 $O(|A||U||D|)$。由此可知，本文算法在计算复杂度上具有显著的优越性。

5　实验结果与分析

为了验证本文的 CSMLFSIE 算法的性能，从 Mulan 数据集中选取了 Emotions、Birds 和 Yeast 这 3 个真实数据集进行实验测试和分析。实验将算法 CSMLFSIE 与 MLFSIE、CSMLPA、MLPA 和 MLDM 进行对比分析，其中，MLFSIE[16] 是一类基于信息熵的多标记特征选择算法，CSMLPA 算法

是基于文献 [20] 的正区域模型设计的考虑了测试代价的多标记特征选择算法, MLPA 是一种利用文献 [20] 中的正区域模型改进的多标记特征选择算法, MLDM 算法是基于文献 [21] 的差别矩阵方法改进的多标记特征选择算法。最后通过 IBLR-ML 多标记分类器验证上述算法特征选择结果的分类性能。

实验过程中首先采用以上 5 种特征选择算法分别对 3 个数据集进行特征降维, 然后使用分类算法对降维后的数据集采用 10 倍交叉验证法验证算法的有效性。本实验的测试环境: CPU 为 Inter(R) Core(TM) i5-4590s (3.0 GHz), 内存 8.0 GB, 算法编程语言为 Python 和 Java, 使用的开发工具分别是记事本和 Eclipse 4.7。

5.1　数据集

实验中选取的 3 个真实数据集的相关信息如表 1 所示, 表中 Yeast[22] 数据集描述的是酵母菌的基因功能分类, Emotions[23] 数据集是来自于某音乐学院的音频剪辑, Birds[24] 数据集通过鸟叫声的记录来区分鸟的种类。其中, Yeast 数据集涉及的是生物信息领域, 而 Emotions 和 Birds 数据集涉及的是音频信息领域。表 1 中对数据集中的实例个数、特征数、标记数、标记基数和总代价进行了描述, 其中, 标记基数用于统计训练集中实例的平均标记个数, 总代价是指利用正态分布函数为数据集中的所有特征生成的代价总和。

表 1　多标记数据集

Table 1　Multi-label datasets

数据集	实例数	特征数	标记数	标记基数	总代价
Yeast	2 417	103	14	4.24	10 033
Emotions	593	72	6	1.87	8 013
Birds	645	260	19	1.01	26 390

5.2　评价指标

文中选用了代价约简率以及平均分类精度 (average precision, AP)、汉明损失 (Hamming loss, HL)、覆盖率 (Coverage)、1 错误率 (one error, OE)、排序损失 (ranking loss, RL) 这 5 种多标记评价性能指标来评价算法性能。给定一组多标记对象集合 (x_i, Y_i), $i = 1,2,\cdots,m$, m 表示对象大小, Y_i 表示多标记分类器预测测试对象 x_i 具有的标记集合, Y_i 表示多标记分类器预测测试对象 x_i 具有的标记集合, $Y_i \subseteq L$, $L=\{\lambda_j : j = 1,2,\cdots,q\}$, L 表示所有标记集合, Z_i 表示测试对象 x_i 实际的标记集合, $\overline{Y_i}$ 表示 Y_i 的补集, $r_i(\lambda)$ 为标记 λ 的排序。

1) 代价约简率是考虑特征代价的特征子集 B 的代价占全特征集 A 总代价的比率:

$$\mathrm{PC} = \frac{\mathrm{Cost}_B(D)}{\mathrm{Cost}_A(D)}$$

2) 平均分类精度 (AP) 是指在标记预测序列中, 排在相关标记之前的标记仍是相关标记的比率:

$$\mathrm{AP}=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|}\sum_{\lambda\in Y_i}\frac{|\{\lambda'\in Y_i : r_i(\lambda')\leqslant r_i(\lambda)\}|}{r_i(\lambda)}$$

3) 汉明损失 (HL) 是指预测出的标记与实际标记的平均差异值:

$$\mathrm{HL}=\frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \Delta Z_i|}{M}$$

其中 Δ 为 Y_i、Z_i 两个集合之间的对称差。

4) 覆盖率 (Coverage) 是指所有对象实际包含的所有标记所需最大的排序距离:

$$\mathrm{Coverage}=\frac{1}{m}\sum_{i=1}^{m}\max_{\lambda\in Y_i} r_i(\lambda)-1$$

5) 1 错误率 (OE) 是指预测出的标记排序最靠前的标记不在实际对象中的比率:

$$\mathrm{OE}=\frac{1}{m}\sum_{i=1}^{m}\delta(\arg\min_{\lambda\in Y_i} r_i(\lambda))$$

若 $\arg\min_{\lambda\in Y_i} r_i(\lambda)\notin Y_i$ 条件满足时, 则 $\delta(\arg\min_{\lambda\in Y_i} r_i(\lambda))=1$, 否则为 0。

6) 排序损失 (RL) 是指预测出的标记中实际不包含的标记比实际包含的标记排序高的比率:

$$\mathrm{RL}=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\|Y_i\|\left\|\overline{Y_i}\right\|}\times$$
$$|\{(\lambda_a,\lambda_b) : r_i(\lambda_a) > r_i(\lambda_b)\ (\lambda_a,\lambda_b)\in Y_i\times\overline{Y_i}\}|$$

平均分类精度越大说明分类性能越好, 代价约简率、汉明损失、覆盖率、1 错误率、排序损失越小说明分类性能越好。

5.3　实验结果及比较

5.3.1　离散参数 k 的选择

由于本文所选择的 3 个多标记数据集的特征值都包含连续型数据, 但 CSMLFSIE 算法处理的是离散型特征变量, 因此对于多标记数据集的处理需要对特征值进行离散化处理。在实验过程中发现, k 的步长取值为 5 时, 降维后的特征子集的分类性能差别较为明显。因此, 本文将 k 以步长 5 从 5 增加到 50 进行实验分析与比较。下面以 Emotions 数据集为例, 讨论离散化参数 k 的选择对多标记分类性能的影响, 图 1 ~ 5 给出了 Emotions 数据集的 5 项评价指标随着离散化参数 k 的值增加的变化曲线。CSMLFSIE 曲线、MLFSIE 曲线、CSMLPA 曲线、MLPA 曲线和 MLDM 曲线分别为这几种多标记特征选择算法的性能。

由图 1 ~ 5 可知, CSMLFSIE 和 MLFSIE 算法

的 5 项分类性能随离散化参数增加变化较为平缓，CSMLPA、MLPA 和 MLDM 算法的 5 项分类性能随离散化参数增加变化较为显著，其中，变化较显著的是 MLPA 算法，5 项分类性能随离散化参数 k 取值的增加变化趋势较明显。针对 CSMLPA 算法，当 k 的取值在 [5, 10] 这个区间时，HL、OE、Coverage 和 RL 的值随离散化参数 k 的取值增加而增大，同时，AP 的值随离散化参数 k 的取值增加而减小，5 项多标记性能指标的值变化较为明显。当 k 的取值在 [40, 50] 区间时，HL、OE、Coverage、RL 和 AP 这 5 项多标记分类性能指标的值变化较小。针对 MLDM 算法，当离散化参数值从 5 变化至 20 时，HL、OE、Coverage 和 RL 的值呈上升趋势，AP 的值呈下降趋势，降维后的特征子集的分类性能变化较显著。当离散化参数取值从 25 变化至 35 时，HL、OE、Coverage、RL 和 AP 的 5 项性能指标的值变化较为平缓，由此可知，降维后的特征子集的分类性能在这个区间的稳定性较强。另外，通过实验结果可知，当离散化参数 k 的值为 25 时，CSMLFSIE 和 MLFSIE 算法所取得的特征子集的分类性能最优；CSMLPA 和 MLDM 算法在离散化参数 k 为 5 时，其降维后的特征子集分类性能最优；当 $k=35$ 时，MLPA 算法得到的特征子集的分类性能最优。

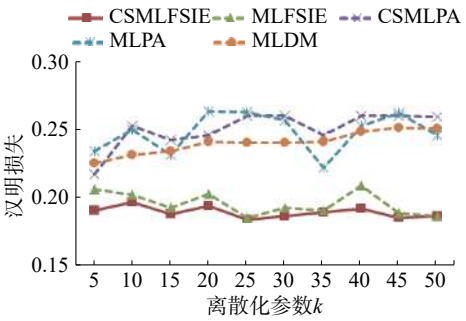

图 1　汉明损失随着离散化参数增加的变化曲线

Fig. 1　Variation of Hamming loss with increase in the discretization parameter

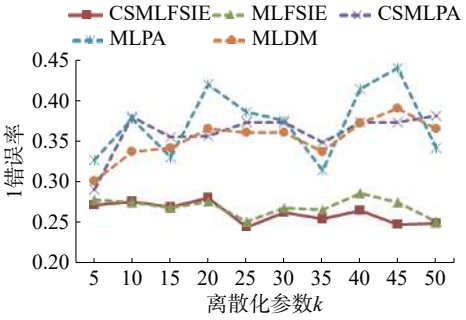

图 2　1 错误率随着离散化参数增加的变化曲线

Fig. 2　Variation of one error rate with increase in the discretization parameter

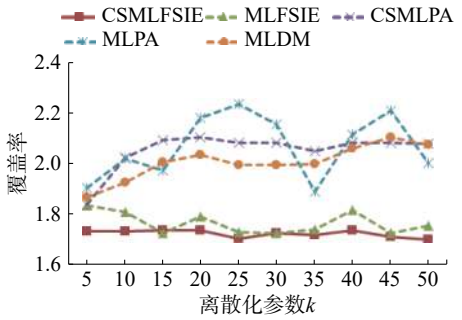

图 3　覆盖率随着离散化参数增加的变化曲线

Fig. 3　Variation of coverage with increase in the discretization parameter

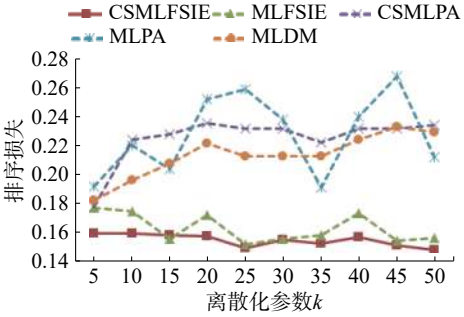

图 4　排序损失随着离散化参数增加的变化曲线

Fig. 4　Variation of ranking loss with increase in the discretization parameter

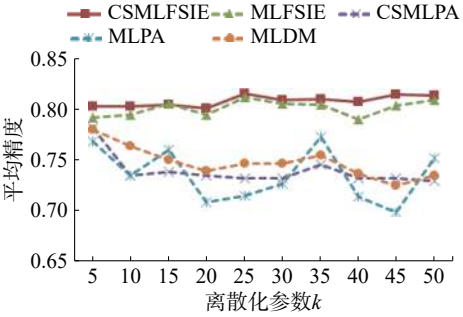

图 5　平均精度随着离散化参数增加的变化曲线

Fig. 5　Variation of average precision with increase in the discretization parameter

综上所述，与其他 4 种算法相比，随离散化参数增加，CSMLFSIE 算法的 5 项分类性能变化最为平缓，即离散化参数的变化对 CSMLFSIE 算法影响最小，因此 CSMLFSIE 算法的稳定性和健壮性更优。

5.3.2　实验对比

实验过程中将训练数据集和测试数据集相结合，采用 10 倍交叉验证法来验证算法的有效性，实验结果采用评价指标的平均值和标准差表示。另外，由于 Mulan 数据集自身并不含测试代价，因此本文采用正态分布函数为每个特征生成测试代价，其中，正态分布函数的取值以 100 为期望，以 30 为标准差。

表 2 ~ 4 中表示在正态分布函数下, 用 5 种多标记特征选择算法分别对 Yeast、Emotions 和 Birds 这 3 个数据集进行特征降维, 并用 IBLR-ML 分类算法验证降维后的特征子集的分类性能, 同时, 与原始数据集的分类性能进行对比。表 2 ~ 4 中给出的数据是 AP 取最优值时, 所对应的 PC、HL、OE、Coverage、RL 和 k 的值。另外, 各项评价指标的最优值用黑体标注, ↓表示该项指标值越小算法的分类性能越好, ↑表示该项指标的值越大算法的分类性能越好。

表 2　Yeast 数据集的实验结果比较

Table 2　The comparisons of Yeast datasets

性能指标	原始数据集	CSMLFSIE算法	MLFSIE算法	CSMLPA算法	MLPA算法	MLDM算法
PC(↓)/%	100	**2.34**	3.26	16.34	15.81	15.02
HL (↓)	0.193 4±0.011 8	0.193 2±0.012 1	**0.192 6±0.011 5**	0.208 2±0.009 2	0.210 9±0.010 2	0.209 4±0.008 6
OE (↓)	0.226 3±0.031 6	**0.219 7±0.033 8**	0.227 1±0.027 5	0.237 9±0.034 4	0.245 3±0.030 6	0.244 1±0.035 6
Coverage(↓)	6.192 7±0.174 7	**6.156 7±0.202 8**	6.166 2±0.182 2	6.481 1±0.156 4	6.597 2±0.209 7	6.572 6±0.180 3
RL (↓)	0.163 5±0.011 7	**0.161 9±0.012 2**	0.162 4±0.011 7	0.180 8±0.013 4	0.188 2±0.014 8	0.184 7±0.014 7
AP (↑)	0.768 7±0.020 0	**0.772 4±0.021 5**	0.771 1±0.019 4	0.747 0±0.022 1	0.736 8±0.021 2	0.739 8±0.022 9
k	—	45	45	5	5	5

表 3　Emotions 数据集的实验结果比较

Table 3　The comparisons of Emotions datasets

性能指标	原始数据集	CSMLFSIE算法	MLFSIE算法	CSMLPA算法	MLPA算法	MLDM算法
PC(↓)/%	100	**6.47**	18.96	21.83	**6.47**	17.08
HL (↓)	0.188 3±0.023 9	0.186 4±0.023 7	**0.185 3±0.019 9**	0.202 4±0.020 0	0.221 8±0.025 8	0.224 9±0.021 4
OE (↓)	0.258 1±0.065 6	0.258 2±0.073 2	**0.251 4±0.056 1**	0.278 4±0.056 0	0.315 3±0.048 7	0.302 0±0.066 0
Coverage(↓)	1.708 7±0.130 3	**1.690 5±0.154 0**	1.729 2±0.142 0	1.803 1±0.128 2	1.890 8±0.150 8	1.867 4±0.132 1
RL (↓)	0.149 6±0.028 0	**0.148 8±0.031 5**	0.151 9±0.026 5	0.168 3±0.021 5	0.191 4±0.024 3	0.182 6±0.026 7
AP (↑)	0.812 6±0.034 9	**0.813 5±0.040 4**	0.812 7±0.032 4	0.795 0±0.028 3	0.773 4±0.031 0	0.780 3±0.033 6
k	—	25	25	5	35	5

表 4　Birds 数据集的实验结果比较

Table 4　The comparisons of Birds datasets

性能指标	原始数据集	CSMLFSIE算法	MLFSIE算法	CSMLPA算法	MLPA算法	MLDM算法
PC(↓)/%	100	11.65	46.54	21.83	**2.77**	2.95
HL (↓)	**0.050 1±0.006 5**	0.051 2±0.007 5	0.050 2±0.007 7	0.052 0±0.007 5	0.053 3±0.008 4	0.055 7±0.008 7
OE (↓)	0.718 9±0.044 6	0.703 4±0.043 7	**0.702 0±0.028 8**	0.725 2±0.034 9	0.740 6±0.030 1	0.737 5±0.040 3
Coverage(↓)	2.622 7±0.537 1	**2.445 0±0.555 3**	2.485 6±0.565 9	2.657 1±0.532 2	2.464 2±0.400 7	2.611 2±0.542 2
RL (↓)	0.091 5±0.021 8	**0.084 1±0.020 2**	0.084 8±0.021 7	0.092 0±0.020 3	0.088 3±0.014 6	0.093 0±0.019 4
AP (↑)	0.591 4±0.045 8	**0.617 1±0.057 7**	0.613 7±0.050 2	0.590 5±0.056 4	0.574 6±0.042 6	0.573 4±0.047 2
k	—	45	35	5	30	25

由表 2 ~ 4 中的 5 种多标记分类性能评价指标的结果可以看出, CSMLFSIE 算法总体优于其他 4 种算法, 较为明显的有 Coverage、HL 和 AP 这 3 项性能指标。同时, 通过 CSMLFSIE 算法进行特征降维后, 特征子集的分类性能优于原始数据集, 其中, 最为突出的是在 PC 这项指标上。另外, 由表 2 ~ 4 可知, 各个算法分类性能最优时, 所对应的离散化参数 k 的取值也存在差异。由表 2

中的 Yeast 数据集可知，由本文 CSMLFSIE 算法降维后的特征子集的分类性能较优，其降维后的特征子集 PC 的值与 MLFSIE、CSMLPA、MLPA 和 MLDM 算法相比，分别减少了 0.92%、14%、13.47% 和 12.68%；同时，AP 分别提高了 0.13%、2.54%、2.54% 和 3.26%，且 OE、Coverage 和 RL 值相对较优。另一方面，CSMLFSIE 算法降维之后的特征子集与原始数据集相比，PC 的值为 2.34%，比原始数据集减少了 97.66%，且其他 5 项分类性能评价指标的值更优，其中，AP 的值提高了 0.37%，HL、OE、Coverage、RL 值分别降低了 0.02%、0.66%、3.6% 和 0.16%。

针对 Emotions 数据集，由本文 CSMLFSIE 算法选择的特征子集的分类性能总体较优，根据表 3 中各项性能指标的结果可知，除 HL 和 OE 性能指标之外，其他 3 项多标记分类性能指标的值最优，且总测试代价 PC 的值最小。CSMLFSIE 算法与 MLPA 算法相比，PC 的值同为 6.47%，但 AP 的值却提高了 6.26%，同时，HL、OE、Coverage 和 RL 的值都显著降低。由此可知，CSMLFSIE 算法要优于 MLPA 算法。CSMLFSIE 算法与 MLFSIE 算法相比，PC 的值减少了 12.49%，另外，Coverage、RL 和 AP 这 3 项性能指标相对更优，因此 CSMLFSIE 算法总体优于 MLFSIE 算法。此外，CSMLFSIE 算法与 CSMLPA 和 MLDM 算法相比，PC 的值分别减少了 15.36% 和 10.61%，其 HL、OE、Coverage、RL 和 AP 这 5 项性能指标的值更优。

从表 4 中的 Birds 数据集可以看出，CSMLFSIE 算法选择后的特征子集的分类性能与 Raw Data 相比，除 HL 这项性能指标之外，其他 4 项性能指标的值都更优，PC 的值也减少了 88.65%，因此 CSMLFSIE 算法选择后的特征子集的分类性能总体优于 Raw Data。CSMLFSIE 算法与 MLFSIE 算法相比，Coverage、RL 和 AP 的值较优，PC 的值减少了 35.89%。CSMLFSIE 算法与 CSMLPA 算法相比，PC 的值减少了 10.18%，AP 的值由 59.05% 提高至 61.71%，HL、OE、Coverage 和 RL 的值分别降低了 0.08%、2.18%、21.21% 和 0.79%。CSMLFSIE 算法与 MLPA 算法和 MLDM 算法相比，AP 的值分别提高了 4.25%、4.37%，HL、OE、Coverage 和 RL 的值有所下降，但 PC 的值分别增加了 8.88%、8.7%。由此可知，由 CSMLFSIE 算法选择的特征子集的分类性能总体较优。

另外，由表 2 ~ 4 可知，针对 CSMLPA 算法，3 个数据集的离散化参数 k 取值为 5 时，其降维后特征子集的分类性能最优；针对 CSMLFSIE 算法进行离散化处理时，参数值为 45 时，Yeast 数据集和 Birds 数据集降维后的特征子集的分类效果较佳，而对于 Emotions 数据集来说，离散化参数取 25 较优；针对 MLFSIE 算法和 MLPA 算法，在 3 个数据集降维后的特征子集的分类性能最优时，所对应的离散化参数的取值也不同；针对 MLDM 算法，对于 Yeast 数据集和 Emotions 数据集，离散化参数取 5 时，降维后的特征子集的分类性能较优，在 Birds 数据集中，离散化参数取值为 25 较好。由此可知，各个算法的分类性能与离散化参数 k 的取值相关，降维后的特征子集影响着分类器的分类性能。

综上所述，原始数据集中存在大量冗余和不相关特征，且这些特征直接影响了分类器的分类性能，综合各项性能评价指标可知，CSMLFSIE 算法总体优于其他 4 种算法，达到了较好的特征降维的效果。

6　结束语

传统的基于多标记的特征选择算法往往忽略了每个特征获取和采集所需花费的代价问题，为此，本文提出了一种代价敏感数据的多标记特征选择算法，该算法利用信息熵分析特征与标记之间的相关性，利用均匀分布函数和正态分布函数为特征生成测试代价，从代价敏感的研究视角，构建一种新特征重要度准则。然后，根据服从正态分布的特征重要度和特征代价的标准差设置阈值，通过阈值剔除冗余和不相关特征。通过对 3 个真实数据集实验结果的分析与比较，验证了本文算法的有效性和高效性。但是，该算法并未充分考虑标记之间的相关性以及误分类代价的问题，这也是我们下一步的研究工作。

参考文献：

[1] ZHANG Minling, ZHOU Zhihua. A review on multi-label learning algorithms[J]. IEEE transactions on knowledge and data engineering, 2014, 26(8): 1819–1837.

[2] TSOUMAKAS G, KATAKIS I, VLAHAVAS I. Random - Labelsets for multilabel classification[J]. IEEE transactions on knowledge and data engineering, 2011, 23(7): 1079–1089.

[3] 郑伟, 王朝坤, 刘璋, 等. 一种基于随机游走模型的多标签分类算法 [J]. 计算机学报, 2010, 33(8): 1418–1426.
ZHENG Wei, WANG Chaokun, LIU Zhang, et al. A multi-label classification algorithm based on random walk model[J]. Chinese journal of computers, 2010, 33(8): 1418–1426.

[4] 李宇峰, 黄圣君, 周志华. 一种基于正则化的半监督多标记学习方法 [J]. 计算机研究与发展, 2012, 49(6): 1272–1278.
LI Yufeng, HUANG Shengjun, ZHOU Zhihua. Regularized semi- supervised multi-label learning[J]. Journal of computer research and development, 2012, 49(6): 1272–1278.

[5] PAWLAK Z. Rough sets[J]. International journal of computer and information sciences, 1982, 11(5): 341–356.

[6] PAWLAK Z, SO-Winski R. Rough set approach to multi-attribute decision analysis[J]. European journal of operational research, 1994, 72(3): 443–459.

[7] 刘清. Rough 集及 Rough 推理 [M]. 北京: 科学出版社, 2001.

[8] SUN Liang, JI Shuiwang, YE Jieping. Multi-label dimensionality reduction[M]. Florida: CRC Press, 2013: 20–22.

[9] ZHANG Yin, ZHOU Zhihua. Multi-label dimensionality reduction via dependence maximization[C]//Proceedings of the 23rd National Conference on Artificial Intelligence. Chicago, Illinois, 2008: 1503−1505.

[10] YU Kai, YU Shipeng, TRESP V. Multi-label informed latent semantic indexing[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil, 2005: 258−265.

[11] 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.
DUAN Jie, HU Qinghua, ZHANG Lingjun, 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.

[12] 王晨曦, 林耀进, 唐莉, 等. 基于信息粒化的多标记特征选择算法 [J]. 模式识别与人工智能, 2018, 31(2): 123–131.
WANG Chenxi, LIN Yaojin, TANG Li, et al. Multi-label feature selection based on information granulation[J]. Pattern recognition and artificial intelligence, 2018, 31(2): 123–131.

[13] LIN Yaojin, HU Qinghua, LIU Jinghua, et al. Multi-label feature selection based on neighborhood mutual information[J]. Applied soft computing, 2016, 38: 244–256.

[14] 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.
LIU Jinghua, LIN Menglei, WANG Chenxi, et al. Multi-label feature selection algorithm based on local subspace[J]. Pattern recognition and artificial intelligence, 2016, 29(3): 240–251.

[15] LEE J, LIM H, KIM D W. Approximating mutual information for multi-label feature selection[J]. Electronics letters, 2012, 48(15): 929–930.

[16] 张振海, 李士宁, 李志刚, 等. 一类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.
ZHANG Zhenhai, LI Shining, LI Zhigang, et al. Multi-label feature selection algorithm based on information entropy[J]. Journal of computer research and development, 2013, 50(6): 1177–1184.

[17] YANG Qiang, WU Xindong. 10 challenging problems in data mining research[J]. International journal of information technology & decision making, 2006, 5(4): 597–604.

[18] 徐章艳, 刘作鹏, 杨炳儒, 等. 一个复杂度为 max($O(|C||U|)$,$O(|C|^2|U/C|)$) 的快速属性约简算法 [J]. 计算机学报, 2006, 29(3): 391–399.
XU Zhangyan, LIU Zuopeng, YANG Bingru, et al. A quick attribute reduction algorithm with complexity of max($O(|C||U|)$,$O(|C|^2|U/C|)$)[J]. Chinese journal of computers, 2006, 29(3): 391–399.

[19] WU Binglong, QIAN Wenbin, HUANG Qin, et al. Cost-Sensitive multi-label feature selection algorithm based on positive approximation[C]//Fuzzy Systems and Data Mining IV-Proceedings of FSDM 2018. Bangkok, Thailand, 2018: 381−386.

[20] QIAN Yuhua, LIANG Jiye, PEDRYCZ W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory[J]. Artificial intelligence, 2010, 174(9/10): 597–618.

[21] WEI Wei, WU Xiaoying, LIANG Jiye, et al. Discernibility matrix based incremental attribute reduction for dynamic data[J]. Knowledge-based systems, 2018, 140: 142–157.

[22] ELISSEEFF A, WESTON J. A kernel method for multi-labelled classification[C]//Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, Canada, 2001: 681−687.

[23] TROHIDIS K, TSOUMAKAS G, KALLIRIS G, et al. Multi-label classification of music into emotions[C]//Proceedings of the 9th International Society for Music Information Retrieval Conference. Philadelphia, PA, 2008: 325−330.

[24] BRIGGS F, HUANG Yonghong, RAICH R, et al. The 9th annual MLSP competition: new methods for acoustic classification of multiple simultaneous bird species in a noisy environment[C]//Proceedings of 2013 IEEE International Workshop on Machine Learning for Signal Processing. Southampton, UK, 2013: 22−25.

作者简介：

黄琴，女，1993 年生，硕士研究生，主要研究方向为粒计算与机器学习。取得计算机软件著作权 2 项，发表学术论文 3 篇。

钱文彬，男，1984 年生，副教授，博士，主要研究方向为粒计算、知识发现与机器学习。主持完成国家青年科学基金项目和江西省青年科学基金项目各 1 项。发表学术论文 20 余篇。

王映龙，男，1970 年生，教授，博士，主要研究方向为知识发现与数据挖掘。参与国家自然科学基金项目 2 项，先后主持江西省自然科学基金项目 3 项。发表学术论文 20 余篇。

CCAI2019 中国人工智能大会

中国人工智能大会由中国人工智能学会创办于 2015 年，每年举办一届。该会是我国最早发起举办的人工智能大会，目前已经成为我国人工智能领域规格最高、规模最大、影响力最强的会议之一。

2019 年中国人工智能大会（CCAI 2019）由中国人工智能学会、青岛市政府共同主办，胶州市政府、马上科普承办，旨在搭建人工智能前沿技术探索桥梁、打造高端交流平台和引领科技创新发展。

CCAI 2019 将延续过去四届的强大阵容，设置 1 个主论坛、1 个大会论坛和 6 个分论坛。该大会将邀请全球人工智能领域顶尖科学家和企业家，围绕当前热点话题、核心技术以及国家和社会关注的热点问题进行重点探讨，并着重关注如何认识我国当前人工智能发展态势。

会议官网：http://ccai.cn/#05

会议日期：2019 年 9 月 21—22 日

会议地点：中国青岛胶州方圆体育中心