基于随机森林的不平衡特征选择算法*

尹 华，胡玉平
（广东财经大学信息学院，广东 广州 510320）

摘 要：数据高维不平衡是当前数据挖掘的挑战。针对传统特征选择方法基于类别平衡假设，导致在不平衡数据上效果不理想的问题，利用随机森林内嵌的变量选择机制，构造了一个新的不平衡随机森林特征选择算法 IBRFVS。IBRFVS 在平衡的取样数据上构造多样决策树，采用交叉验证方式获取单棵决策树的特征重要性度量值。各决策树的权重和特征重要性度量的加权平均决定了最终的特征重要性序列，其中，决策树的权重由该决策树与集成预测的一致性程度决定。在 UCI 数据集上的随机森林超参数选择和预处理对比验证实验中显示，四种超参数 K 经验取值中，当 K 的取值为特征数的平方根时，IBRFVS 性能较为稳定且优于传统特征选择算法。

关键词：不平衡数据；高维数据；特征选择；随机森林
中图分类号：TP181 **文献标志码**：A **文章编号**：0529－6579（2014）05－0059－07

An Imbalanced Feature Selection Algorithm Based on Random Forest

YIN Hua, *HU Yuping*
（School of Information, Guangdong University of Finance Economics, Guangzhou 510320, China）

Abstract: High-dimensional and imbalance data is a challenge for data mining. Balanced class distribution hypothesis leads to unsatisfied results of traditional feature selection algorithms on imbalanced data. For solving this problem, a new imbalanced feature selection algorithm IBRFVS, which uses the variable selection mechanism embedded in random forest, is constructed. IBRFVS construct vary decision trees on the balanced sampling data and get the feature importance measurements of individual decision tree by cross validation. The features importance list is decided by the weighted average of the decision tree weights and feature importance measurements, and the decision tree weights is decided by the consistent degree of the individual decision prediction and ensemble prediction. The random forest hyper parameter selection and preprocessing compare experiments on UCI dataset show that the performance of IBRFVS is more stable and prior than traditional feature selection algorithms when hyper parameter K is the square root of feature number, among four empirical parameters.

Key words: imbalance data, high dimensional data, feature selection, random forest

分类是数据挖掘中最常见的一项任务，通过分类建立预测模型，可提供对未知问题的准确预测。尽管目前成熟的分类算法已被成功应用于各领域，但是信息技术发展带来的数据复杂度的提升，也对分类算法提出了新的挑战[1]，其中具有大量属性的高维数据和类别分布不均匀的不平衡数据因在应用中的普遍性，成为目前研究的焦点。在诸如图像检索、入侵检测和生物信息挖掘等相关数据中都体现了高维和不平衡的双重特性。当分类高维数据时，由于特征空间大，产生的分类器复杂，数据容易过度拟合。特征选择可以减少数据维度，降低分类器的复杂度，使之更关注于提供丰富信息的特征

* **收稿日期**：2014－05－05
基金项目：国家自然科学基金资助项目（71202098）；广东省高校自然科学研究育苗工程资助项目（2013LYM0032）；广州市科学技术局珠江科技新星专项资助项目（2012J2200085）
作者简介：尹华（1981 年生），女；**研究方向**：数据挖掘；E-mail：yinhua@whu.edu.cn

向量。但传统特征选择方法大部分是以分布均衡的数据为研究对象，以优化总分类精度为基本目标，所以很少有方法在不平衡数据集得到理想的学习效果[2]。因此，针对不平衡数据设计有效的特征选择算法有其必要性。

尽管不平衡数据分类问题已经引起了研究界的重视，但是有关不平衡数据特征选择的研究仍然不够充分。文献［3］从统计的角度分析了含有丰富类别信息特征的分布特点，并提出了基于类别分布的特征选择框架。文献［4］从实验角度研究了不平衡问题上 IG 算法的困境，通过权重调整修改 IG 算法，使之能够较好应用于不平衡数据的特征选择。文献［5］综合考虑特征在正类和负类中的分布性质，结合四种衡量特征类别相关性的指标对文本中的特征词进行评分，使其能够更好解决传统特征选择方法在不平衡数据集上的不适应性。上述方法都是在文本分类的应用背景下研究不平衡特征选择问题，其解决思路是在已有常规特征选择方法的基础上，区别对待大类和小类的描述特征。刘天宇等在文献［6］中研究了故障诊断中的不平衡特征选择问题，在 EasyEnsemble 算法的基础上，结合基于预报风险误差的特征选择方法，提出了基于预报风险误差的 EasyEnsemble 算法 PREE。PREE 算法充分利用了集成算法的分类效果进行特征选择，但该方法得到的特征子集较大且冗余特征较多，且由于方法本身的计算特性，导致其不能有效地处理离散型特征。李霞等在文献［7］中采用先抽样，再特征选择，再集成投票的方法，提出了基于抽样集成的特征选择方法 EFSBS。EFSBS 可以获得较少的特征，但没有利用来自于分类算法的反馈。因此，应该设计新的能同时处理离散型和连续型特征，且可充分利用分类反馈的不平衡数据特征选择算法。

本文利用随机森林变量选择机制，针对不平衡问题，提出了一个不平衡随机森林变量选择方法（IBRFVS）。IBRFVS 在平衡 Bagging 的基础上，构建多个决策树（基分类器），针对每个基分类器，获取各特征重要性度量，通过对各基分类器所求得的特征重要性度量加权求和来获得最终的特征重要性序列。可在此序列上根据选择特征的数量由高至低进行特征选择。各基分类器所求得的特征重要性度量的权重由基分类器判定与投票评定的一致性决定。

1 随机森林

随机森林（Random Forest，RF）[8] 是由 Breiman 于 2001 年提出的一个集成学习算法框架，分为两个步骤：1）Bagging 获得多个子训练数据集；2）在各训练数据集上建立决策树，决策树中每个结点的特征选择不是在全部特征空间，而是在随机选择的固定数量的特征空间中选择。随机森林通过在实例和特征上引入双重随机性来保证所产生基分类器的多样性；尽管在构造决策树时是在随机子空间中选择最佳分裂点，但由于最佳属性的计算并非随机的，而是根据一定的属性选择原则，一定程度上确保了基分类器的准确性。

在决策树构建过程中，树的每个结点都是以一定原则从众多特征中选择出的“重要”特征，这一过程实际上就是一个显示的特征选择过程。随机森林是对决策树的集成，也相应继承了决策树选择“重要”特征的能力，所不同的是，随机森林采用一种隐式的方式进行特征选择。随机森林变量选择（RVS）[9] 是随机森林的一个副产品，是由 Breiman 提出的隐式的特征选择方法。当一个重要特征（对预测准确率有贡献）出现噪声时，预测的准确率应该明显减少。若此特征是不相关特征，则其出现噪声对预测准确率的影响应该不大。基于这一思想，在利用袋外数据（Out of Bag Data）预测随机森林性能时，若想获知某特征的重要程度，仅需随机修改该特征数值，而保持其他特征不变，由此获得的袋外数据预测准确率与原始袋外数据预测准确率之差体现了该特征的重要程度。

随机森林变量选择方法实际上是一种嵌入式的特征选择方法，充分利用了集成分类器构建过程所产生的分类模型。与 PREE 不同之处在于，PREE 利用的是在特定特征上的结构风险变化，PREE 在计算特定特征的 AUC 时，采用的是取特征平均值的方式；而随机森林变量选择方法基于的是无关特征对模型性能影响不大的思想，通过施加干扰来测试特征的准确程度。这种方法可以同时处理离散型数据和连续型数据，弥补了 PREE 的缺陷。同时在特征重要性度量计算时，充分利用了分类器的性能信息，相比于 EFSBS 仅考虑数据本身的特征相关性，在后续分类中，更易产生好的分类结果。

2 不平衡随机森林特征选择算法

高维数据处理的一种有效途径即通过特征选择降低特征维数，而不平衡数据处理的有效途径则是通过取样方法平衡数据。随机森林的两个步骤综合了此两项机制。由此可以考虑利用随机森林的算法机制设计合理的特征选择算法，用于同时处理高维

不平衡数据。不平衡随机森林特征选择算法（IBRFVS）受随机森林算法启发，利用随机森林的构造过程，对不平衡数据集进行特征选择。由于随机取样将造成数据集的差别，单个取样数据集仅能代表数据的某个侧面，可能导致特征选择的不准确。如果能够获得取样数据的不同侧面，则可以从不同角度来进行特征选择，然后根据不同侧面的权威性来决定最终的特征选择结果。

2.1　算法描述

假设原始训练数据集为 D，属性个数为 M，欠取样后数据集为 *UnderSamplingD*，个数为 N，算法框架如图 1 所示。

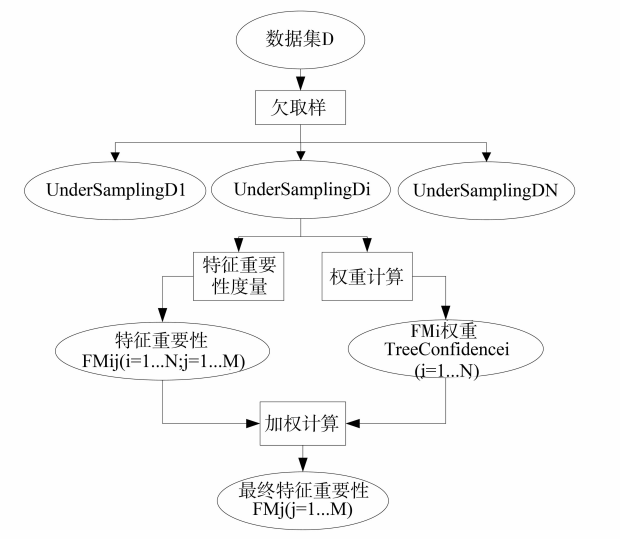

图 1　IBRFVS 算法框架

Fig. 1　IBRFVS framework

IBRFVS 的算法分为以下几个步骤：

1）用欠取样方法获得多个与小类实例数量相同的大类实例集；

2）在小类实例集上有放回取样获得与大类实例集相同数量的数据集；

3）将小类实例集与大类实例集组合获得多个平衡数据集 $UnderSamplingD_i(i = 1,\cdots,N)$；

4）For $UnderSamplingD_i(i = 1,\cdots,N)$

a. 计算各特征重要性度量 $FM_{ij}(j = 1,\cdots,M)$；

b. 计算 FM_i 的权重 $TreeConfidence_i$；

5）加权计算获得数据集的最终特征重要性度量值。

随机森林的 Bagging 步骤是对全部数据集进行 Boostrap 取样。但是对于不平衡数据集而言，为使多数类与少数类平衡，是对少数类 Boostrap 取样，对于多数类则采用欠取样方法。这样处理的好处是，构造了多个欠取样数据集，可以反映不同的欠取样结果，在这些不同的欠取样数据集上所做的特征选择比在单一欠取样数据集上包含更多的信息。在特征选择时，当特征数量特别多时，搜索空间将非常大，在随机子空间上构造决策树，是一种缩小特征空间的有效方法，而决策树算法计算分裂属性的过程也就是一个属性选择的过程，可以直接利用此过程选择重要特征。在每个 *UnderSamplingD* 数据集上都可以构造一棵在随机子空间中产生的决策树，也即获得一个特征重要性度量。N 个 *UnderSamplingD* 可以获得 N 个特征重要性度量。这些特征重要性度量体现了各特征在不同 *UnderSamplingD* 数据集上的重要程度。但是每个 *UnderSamplingD* 所获得特征重要性度量的可信度是不同的，在此体现为权重，也即可信度高的，所赋予的权重越高。因此 IBRFVS 算法的关键是特征重要性度量的计算和权重计算。

2.2　特征重要性度量

IBRFVS 采用随机森林变量选择方法（Ⅵ）来计算特征重要性度量值。Ⅵ的特征重要性度量的计算是基于 Out-of-Bag 样本的。基于 Out-of-Bag 样本测试算法性能或计算算法参数是当前常用的一种方法[10]。这种方法的好处是可以减少参数计算的时间。但是在 IBRFVS 中，由于采用欠取样方法平衡数据集中的类别，如果按照 Out-of-Bag 样本的获取方法，则会导致出现 Out-of-Bag 中的大类数据过多的情况。因此，IBRFVS 采用 k 层交叉验证的方法来获取特征重要性度量值。Ⅵ计算特征重要性度量值时，需要在每个特征上引入噪声之后进行交叉验证对比，以确定特征重要性程度。而随机森林需要构造多棵决策树，特征数量增加且 k 的层数增加时，算法运行时间将随之增加。因此在 k 的取值上采用最简单的交叉验证方法-holdout 方法的思路，即每次用于验证的数据量不超过总数据量的三分之一。由此将 k 取值为 3，将数据集随机分为 3 个不相交的子集，重复执行分类算法 3 次，取三次平均值作为最终的性能评价。单棵树的特征重要性度量值的计算方法如图 2 所示。

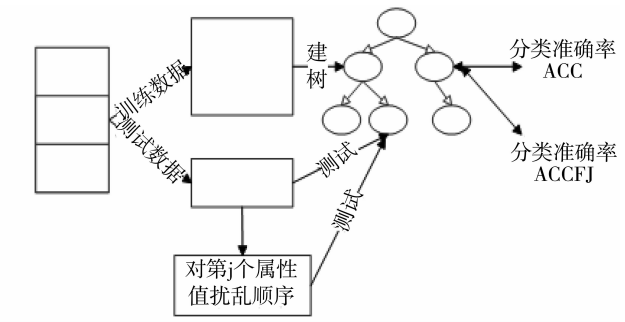

图 2 计算单棵树的特征重要性度量值

Fig. 2 Calculate the feature importance value of single tree

第 j 个属性的特征重要性度量值由 *Acc* 和 *AccFj* 的差值决定，其中 *Acc* 代表的是扰动属性值前的分类准确率，而 *AccFj* 代表的是扰动第 j 个属性的值后的分类准确率。由于采用三层交叉验证，每个 *UnderSamplingD* 分为三份，三份数据交叉作为测试数据集，因此，在同一 *UnderSamplingD* 数据集上，*Acc* 和 *AccFj* 需要计算三次，最终的特征重要性度量值 FM_{ij} 是由三次所产生的平均差值决定。

$$FM_{ij} = \frac{\sum_{k=1}^{3} (ACC_{ik} - ACCF_{ijk})}{3}$$

其中 i 代表第 i 个 *UnderSamplingD* 数据集，第 j 个特征，k 代表第 k 层数据。

2.3 权重度量

当大类数据与小类数据严重不平衡时，对大类数据欠取样可能产生差异性较大的 *UnderSamplingD* 数据子集。在此数据子集上所建立的树的准确率也将有所区别。假若把 N 个 *UnderSamplingD* 对于特征重要性的判断看做 N 个专家的判断，如 EFSBS 一样，可以考虑采用多数投票的方式。这种方式是基于各个专家的判断能力相同的假设上。实际上，由于 *UnderSamplingD* 的多样性，其准确性是不同的，由此，基于不同 *UnderSamplingD* 所做的特征重要性的判定能力也是不同的，应该赋予其不同的权重。BRFVS 认为与最终集成判定一致度高的基决策树应该具有更高的权重，其所获得的特征重要性度量值具有更好的可信度。假设存在一个实例数为 S 的测试数据集，有 N 个 *UnderSamplingD* 数据集产生的 N 棵决策树。根据决策树的预测结果可获得一个 $S \times (N+2)$ 的矩阵，矩阵的行代表要预测的实例，矩阵的前 N 列分别代表 N 棵决策树，第 $N+1$ 列代表集成投票的结果，在前 N 列中超过半数的判定，确定为第 N 列的最终结果，第 $N+2$ 列代表测试数据集的实例类标号。则第 i 棵决策树的判定可信度可通过下式计算

$$TreeConfidence_i = \frac{\sum_{j=1}^{S} I(Tree_{ij} = Ensemble_j)}{S} \times AccEnsemble$$

其中 $TreeConfidence_i$ 表示第 i 棵树的可信度，$Tree_{ij}$ 表示第 i 棵树对第 j 个实例的预测结果，$Ensemble_j$ 表示对第 j 个实例的集成预测结果。*AccEnsemble* 表示的是集成预测准确率，即 Ensemble 与 Original 的一致性程度。由于每棵树的 *AccEnsemble* 都是相同的，$TreeConfidence_i$ 与 *AccEnsemble* 相乘后所获得排序结果是一样的。之所以仍然需要加入这一影响因素，是为了缩小权重间的绝对差距，同时展现集成整体效果对于特征重要性度量的影响。表 1 为一个 $N=5$，$S=5$ 的一致性度量矩阵。

表 1 $N=5$，$S=5$ 的一致性度量矩阵

Table 1 Consistency measure matrix $N=5$ and $S=5$

	Tr1	Tr2	Tr3	Tr4	Tr5	Ensemble	Original
1	1	0	1	1	0	1	1
2	0	1	0	0	0	0	1
3	1	0	1	1	1	1	1
4	1	1	0	1	0	1	1
5	0	0	0	1	1	0	0

不考虑 *AccEnsemble* 因素，分别计算出 *Treei* 的可信度。

$TreeConfidence_1 = 1$，$TreeConfidence_2 = 0.4$，$TreeConfidence_3 = 0.8$，$TreeConfidence_4 = 0.8$，$TreeConfidence_5 = 0.4$

考虑 *AccEnsemble* 因素，分别计算出 *Treei* 的可信度。

$TreeConfidence_1 = 0.8$，$TreeConfidence_2 = 0.32$，$TreeConfidence_3 = 0.64$，$TreeConfidence_4 = 0.64$，$TreeConfidence_5 = 0.32$

特征权重计算时还需要考虑的一个问题就是测试数据集问题。由于特征权重计算的一个关键因素是集成性能。而集成性能是针对原始的整个数据集的。因此，可以借鉴 Out-of-Bag 的思想，构造欠取样后的 Out-of-Bag 数据集。欠取样后的 Out-of-Bag 数据集由两部分组成，一部分是小类 Bootstrap 后未取样的数据集，另一部分是大类欠取样后未取样的数据。

IBRFVS 用于计算特征重要性度量值所使用的数据集是交叉验证数据集，而在计算特征权重时所

使用的数据集是 Out-of-Bag 数据集。前者是平衡数据集，后者是不平衡数据集。区别对待的原因在于：特征重要性度量是在平衡数据集上获得，针对的是单棵决策树，此时用交叉验证的方法获得的是单棵决策树在某一种欠取样数据上对特征重要程度的一个判断。而基决策树的权重则是由该决策树对于集成所做的贡献决定的，随机森林分类的原始数据集是一个不平衡数据集，采用不平衡的 Out-of-Bag 数据测试当次集成的性能是合理的。通过每棵树确定所有特征的重要性度量值 FM 后，乘以各树的可信度 *TreeConfidence*，求平均即可获得最终的特征重要性度量值 $Final\text{Ⅵ}$。

$$Final\text{Ⅵ}_j = \frac{\sum_{i=1}^{N}(TreeConfidence_i \times FM_{ij})}{N}$$

3　实验验证

IBRFVS 算法的实验验证由两部分组成：算法参数选择和与预处理方法对比验证。实验数据集选自 UCI 数据集[11]。由于 UCI 数据集的数据不存在典型的不平衡特性，我们将多类问题转化为二分类问题，使数据集呈现了较为明显的不平衡特性。在二分类中，某一类比例低于 50% 的一定程度则展现了不平衡特性。在此，将不平衡程度分为四个级别：极高、高、中、低。其中正类比例在 0 ~ 5% 的范围属于极度不平衡，在 5% ~ 10% 的范围属于高度不平衡，在 10% ~ 25% 的范围则属于中度不平衡，在 25% ~ 40% 的范围属于低度不平衡，当正类比例超过 40% 时，数据类别达到基本平衡状态。由此构造了 7 个实验数据集，如表 2 所示。

表 2　高维不平衡实验数据集

Table 2　High-dimensional and imbalanced datasets

数据集	实例数	特征数	正类比例/%
SteelDirtiness	1 941	27	2. 8
SteelPastry	1 941	27	8. 1
Stalog4	4 435/2 000	36	9. 4
SemeionA - 8	1 593	256	9. 7
SemeionA - 3	1 593	256	20. 1
SteelBump	1 941	27	20. 7
Spambase	4 601	57	39. 4

不平衡特征选择的目的是为了更好的特征选择再取样，最终获得好的分类性能。因此可以采用预处理后的分类性能来评价 IBRFVS 算法的好坏。如果采用 IBRFVS 算法特征选择后取样的分类性能优于普通特征选择后取样的分类性能，则可认为该算法对于不平衡数据特征选择是有效的。因此，实验按照随机森林的不同超参数取值获得各数据集的特征选择结果后再取样得到预处理后的数据。在分类器评价方面，则选择不平衡数据分类中常用的 AUC 面积作为评价标准。对该数据分类后的 AUC 性能间接代表了 IBRFVS 算法的性能。

3. 1　RF 参数影响

随机森林涉及两个参数：K（随机选择的特征数）和 N（生成的决策树数量）。Breiman 在提出随机森林时，曾指出超参数 K 的取值对最终构建的集成学习算法有较大影响并建议 K 的取值为 1 或 $\log_2(M)+1$，其中 M 为原始数据集的实际特征数。在后续的多项研究中也形成了几个较好的 K 的经验值：1，$M/2$，$\sqrt{M}$ 和 $\log_2(M)+1$[12-13]。根据没有免费午餐的原理，某一参数未必能够适用于所有的问题，上述经验值在面临不同问题时，其实验效果可能有所区别。尽管基于 out-of-bag 的方法对所有的取值进行评估，确定最优的取值。但这种方法是一个穷举的方法，当特征维特别高的时候，计算量非常大。由于本文当前关注的是 IBRFVS 算法能否较好对不平衡数据特征选择，因此，并不专门研究最优参数的获取方法，而是研究了 IBRFVS 算法在采用不同的 K 参数设置时，对分类效果的影响。另一个重要的参数是决策树数量参数 N，在 Bernard 的实验研究中显示，决策树的数量设置为 100 是较为合理的设置[13]。因此，在 IBRFVS 算法的验证实验中，设置 $N=100$，K 为 1，$M/2$，$\sqrt{M}$ 和 $\log_2(M)+1$ 四种取值时，IBRFVS 算法对决策树 C4. 5 算法的性能影响。各数据集在不同 K 值下的 AUC 值对比如图 3 所示。其中横轴代表 K 的四种不同取值，纵轴代表 AUC 值，每一条曲线代表一个数据集。

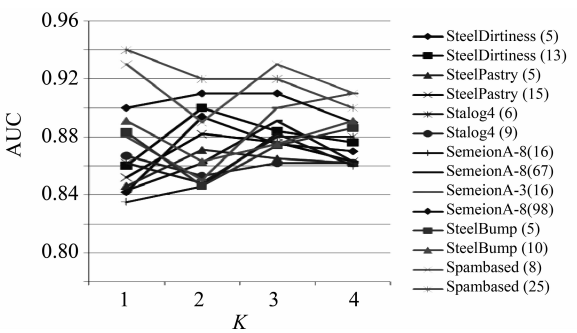

图 3　各数据集在不同 K 值下的 AUC 值对比图

Fig. 3　AUC of classifying different datasets with different K

从图 3 可以看出，在四种情况下，K 取值为 $\sqrt{M}$ 和 $log_2(M)+1$ 时，分类性能相对比较稳定，当 K 取值为 1 和 $M/2$ 时，各数据集的分类性能波动比较大。Steel 的两个数据集的特征数为 27，当 K 的取值为 $M/2$ 时，达到了最优性能。Semeion 数据集的特征数为 256，K 取 $M/2$ 并没有达到性能最优。原因在于当特征数比较少时，存在冗余特征的可能性也较小。当特征数较少时，K 取 $\sqrt{M}$ 与 K 取 $\log_2(M)+1$ 的效果差别不大，但当特征数增加时，K 取 $\sqrt{M}$ 的性能优于 K 取 $\log_2(M)+1$。图 3 还显示了一个特殊的现象，SteelBumps 数据的最佳分类性能出现在 K 取值为 1 的情况，当 K 取值为 1 时，代表所构造的决策树为纯随机树。尽管纯随机树增加了集成学习基分类器的多样性，但是基分类器的准确性未必能够保证，存在一定的偶然性。当树的数量取值较大时，可以考虑 K 取值为 1 的情况。通过对 K 取值的分析发现，当 K 取值为 $\sqrt{M}$ 时，IBRFVS 算法的性能相对稳定且较好。

3.2 预处理对比分析

在对比实验中，我们将 IBRFVS 算法与未特征选择、过滤式特征选择和封装式特征选择对比，其中过滤式特征选择采用较为常用的信息增益（Information Gain，IG）特征选择，封装式特征选择则采用融合遗传搜索的 Wrapper 特征选择算法。

由于通过 IBRFVS 算法进行特征选择后，数据集仍然是不平衡的，因此在特征选择后，对数据集欠取样，获得平衡数据集，再用 C4.5 算法分类此平衡数据集，获得最终的分类性能 AUC 评价。IBRFVS 算法输出的结果为数据集各特征重要性度量值。按照从大到小的顺序可获得最终的排列顺序。

在特征数的选择上，采用 IG 特征选择时，选择的特征数量为 $\sqrt{M}$，而用封装式特征选择方法时，直接采用选择的特征子集数 Q。为实现两者的对比，每个数据集用 IBRFVS 算法特征选择后，取两个数量的特征子集 $P=\sqrt{M}$ 和 Q。由于最佳优先搜索与遗传搜索的效果差别不大，尽管从时间角度考虑，遗传搜索的时间稍长，但是由于遗传搜索所获得特征数量相对多一些，为与 $\sqrt{M}$ 区别，Q 的值直接由遗传搜索后产生的数据子集个数决定。实验对比如图 4 所示。其中 SteelDirtiness，SteelPastry，SteelBumps，Stalog4，SemeionA－8，Semeion－3 和 Spambase 的 Q 值分别为：13，15，10，9，67，98，25。

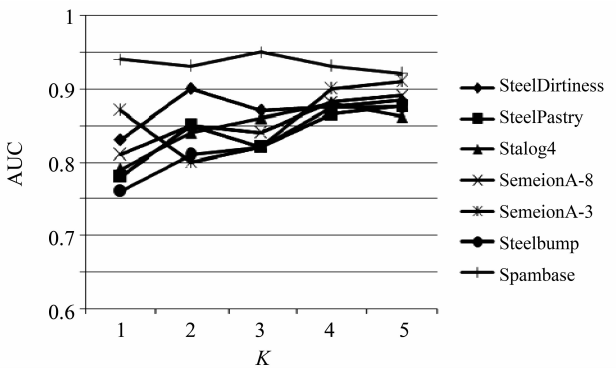

图 4 原始数据、IG、Wrapper、IBRFVS（$K=\sqrt{M},P$）和 IBRFVS（$K=\sqrt{M},Q$）处理后的 AUC 值

Fig. 4 AUC on original, preprocessing by IG feature selection, by Wrapper feature selection, by IBRFVS ($K=\sqrt{M},P$) and by IBRFVS ($K=\sqrt{M},Q$) datasets

图 4 的横轴代表未处理、IG 特征选择＋取样、Wrapper 遗传搜索＋取样和 K 取 $\sqrt{M}$ 时 IBRFVS 选择两种数量特征＋取样预处理方法。纵轴代表用 C4.5 算法分类不同预处理后数据的 AUC 性能。从图中可以看出，IBRFVS 的性能从大体趋势上优于 IG 和 Wrapper 特征选择再取样的分类性能。从 Spambase 数据集的曲线可看出，当数据集基本平衡时，采用 IBRFVS 算法和传统特征选择算法性能大致相当。

由于 IBRFVS 算法所获得的结果是特征重要性度量值的序列，因此需要选择一定数量的特征作为最终选择的特征。选择的特征数量是一个算法的参数值，实验过程中未专门研究参数的选择，但是为与传统过滤式和封装式算法对比，研究了两个数值：$\sqrt{M}$ 和封装式（遗传搜索）所获得的特征数 Q。当前的研究显示，两者之间的差别并不是特别大，选择 $\sqrt{M}$ 个特征的平均性能更优，相对比较稳定。

4 结 论

数据高维不平衡是现实世界数据的典型特性，现有的标准数据挖掘算法在解决实际问题中面临新的挑战。本文借鉴随机森林变量选择的思想，构造了一个新的不平衡随机森林变量选择算法 IBRFVS。IBRFVS 算法在平衡的取样数据上构造决策树，随机森林变量选择方法分别用于各决策树，产生特征重要性度量值，各决策树的权重则由该决策树与集成预测的一致性程度决定。一致性程度高，则说明其特征选择越准确，相应的特征度量分值则越高。

最终的特征度量值是由各决策树所产生的特征度量值的加权平均获得。在实验过程中，研究了 RF 参数对 IBRFVS 算法的影响。分别取 K 为 1，$M/2$，$\sqrt{M}$ 和 $\log_2(M)+1$ 四种取值，研究特征选择算法的效果。结果显示，在不同的数据集上，K 的取值不同对 IBRFVS 算法性能有一定影响，在当前的数据集中，当 K 取 $\sqrt{M}$ 和 $\log_2(M)+1$ 时性能大致相当，但 K 取值为 $\sqrt{M}$ 时，性能更为稳定。某些数据集在 K 取值为 1 时，算法也能取得好的效果，但这一结论不存在必然性。对比 IBRFVS 特征选择后再取样分类的效果和 IG 及 Wrapper（GA 搜索）特征选择后再取样分类的效果发现，IBRFVS 特征选择后再取样分类的性能优于后两种方法。当数据集的不平衡程度较低时，算法的性能在一些 k 取值情况下不如传统特征选择算法，也就是说，IBRFVS 较为适用具有一定不平衡程度的数据集。在随机森林的超参数选择上，我们选择的是经验数据值，在后续的研究中，将对参数影响进行深入分析研究。从算法的运行时间看，最快的是过滤式特征选择算法，IBRFVS 和采用遗传算法的封装式特征选择算法的运行时间都比较长，不适用于在线处理方式。根据循环次数计算，IBRFVS 的时间复杂度由输入的实例数和特征维数的乘积决定，即算法的时间复杂度为多项式时间。由于特征选择算法是预处理算法，通常可用离线方式执行。因此在提高精度的前提下，运行时间的影响不大。将 IBRFVS 改为并行算法，使之能在线执行是下一步的研究工作。

参考文献：

[1] YANG Q, WU X D. 10 challenging problems in data mining research [J]. International Journal of Inforamtion Technology & Decision Making, 2006, 5: 597 – 604.

[2] 蒋盛益,王连喜. 不平衡数据的无监督特征选择方法[J]. 小型微型计算机系统,2013, 34(1): 63 – 67.

[3] 靖红芳,王斌,杨雅辉,等. 基于类别分布的特征选择框架[J]. 计算机研究与发展, 2009, 46(9): 1586 – 1593.

[4] 尤鸣宇, 陈燕,李国正. 不均衡问题中的特征选择新算法: Im-IG[J]. 山东大学学报:工学版, 2010, 40(5): 123 – 128.

[5] 张玉芳,王勇,熊忠阳,等. 不平衡数据集上的文本分类特征选择新方法[J]. 计算机应用研究, 2012, 28(12): 4532 – 4534.

[6] 刘天羽, 李国正,尤鸣宇. 不均衡故障诊断数据上的特征选择[J]. 小型微型计算机系统, 2009, 30(5): 924 – 927.

[7] 李霞, 王连喜,蒋盛益. 面向不平衡问题的集成特征选择[J]. 山东大学学报:工学版, 2011. 41(3): .

[8] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5 – 32.

[9] GENUER R, POGGI J M, TULEAU-MALOT C. Variable selection using random forests [J]. Pattern Recognition Letters, 2010, 31(14): 2225 – 2236.

[10] 李毓,张春霞. 基于 out-of-bag 样本的随机森林算法的超参数估计[J]. 系统工程学报, 2011, 26(4): 566 – 572.

[11] ASUNCION A, NEWMAN D. UCI machine learning repository [G]. [2014 – 04 – 30]. http://archive. ics. uci. edu/ml/.

[12] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees [J]. Machine learning, 2006, 63: 3 – 42.

[13] BERNARD S, ADAM S, HEUTTE L. Using random forests for handwritten digit recognition [C]//Ninth International Conference on Document Analysis and Recognition, 2007, 2: 1043 – 1047.