文章编号：1672－4348(2006)04－0436－04

拆分特征选择及其在企业信用评估中应用

凌健，林成德

（厦门大学自动化系，福建 厦门 361005）

摘要：评估指标体系的选取是企业信用评估的首要问题，它是一个特征选择问题。文章提出了一种针对 SVM 组合技术的拆分特征选择方法，其主要思想是对 SVM 组合中的各个分类器分别进行特征选择，再采用不同的特征子集作为各子分类器的输入，进行组合建模与预测。文章从 filter 和 wrapper 相结合的思想出发，进行了子分类器的特征选择；之后，针对企业信用评估问题的特点，采用了二叉树结构作为SVM 的组合策略。实验表明，拆分特征选择方法能选出规模较小、具有一定差异的关键指标集，提高了模型的分类性能，并且具有计算简单，运行快速的优点。

关键词：特征选择；支持向量机(SVM)；企业信用评估

中图分类号：TP391；F830 文献标识码：A

Separating feature selection and its application to enterprise credit assessment

Ling Jian, Lin Chengde

(Automation Department, Xiamen University, Xiamen 361005, China)

Abstract: The selection of the evaluating index system is the key to enterprise credit assessment, which is essentially a feature selection problem. A separating feature selection approach concerning the combination of SVMs (support vector machine) is proposed, whose basic idea is to execute feature selection on each SVM in the combination and use the different selected feature subsets as the inputs. The composite feature selection based on filter and wrapper is used in the selection process of each SVM. The SVMs are then combined using binary tree structure adapted to the characteristic of enterprise credit assessment. Experiment shows that separating feature selection can select feature subsets with small scale and diversity and improve the classification ability of the model and reduce the computing time and complexity.

Keywords: feature selection; support vector machine (SVM); enterprise credit assessment

1 概述

信用评估是银行贷款业务中关键的一环，银行通过对企业各财务指标的评估，得到企业的信用等级，以此判断贷款与否。近年来，支持向量机(Support Vector Machine，SVM)等人工智能技术逐步被应用到信用评估领域[1]，然而由于企业财务指标数量繁多、关系复杂，使得智能分类模型的训练时间、复杂度及预测精度都受到影响，达不到较好的效果。于是，选取数量较少、分类精度较高的关键指标，成为采用智能方法进行企业信用评估需要解决的首要问题。

财务指标的选取是特征选择问题，目前应用的特征选择方法可分成 2 大类：filter 方法和 wrapper 方法[2]。filter 方法在算法中不包含分类器，具有较高的选择效率；而 wrapper 方法将分类器的准确率作为特征选择的评价准则，具有较高的精度。当我们采用SVM（二分类器）的组合来解决像企

收稿日期：2006－07－04

第一作者简介：凌健(1983－)，女(汉)，福建莆田人，硕士研究生，主要研究方向为智能计算方法及其在信用评估中的应用。

业信用评级这种典型的多分类问题时，传统的特征选择方法，都存在着一定的缺陷。因为这些方法都是作用于整个 SVM 组合的，同一个选出特征子集应用于组合中的所有 SVM，忽略了各个分类器之间的差异，一定程度上限制了 SVM 组合性能的提高；且这些方法由于涉及到多类问题(filter)或是分类器的组合(wrapper)，算法复杂，运算效率不高。针对这一问题，本文提出一种拆分特征选择方法，其基本思想是对分类器组合中的各个学习器分别进行特征选择，采用不同的特征子集作为输入，进行组合建模与预测。这样，不仅可以简化特征选择算法，加快特征选择速度，且各个选出特征子集的针对性强，使得单个分类器具有较高的分类精度，从而提高组合整体的分类性能。

为使选出特征子集具有较高的分类能力同时兼顾算法的运行速度，我们提出一种基于 filter 和 wrapper 的组合特征选择方法对单个分类器进行特征选择；且为了适应信用数据分布不平衡的特点、进一步提高建模速度，我们采用二叉树结构作为SVM 的组合策略。对福建省某银行企业数据的实验表明，拆分特征选择方法能选出规模较小、具有一定差异的关键指标集，提高了模型的分类性能，并且具有计算简单、运行快速的优点。

2　支持向量机

支持向量机(Support Vector Machine，SVM)，是一种基于小样本学习理论的通用学习算法，由于能在理论上保证它的泛化性能，加之求解简易的优点，已经成为机器学习和数据挖掘领域的标准工具[3]。SVM 的基本思想是将数据映射到高维空间，在此高维空间中找到以最大间隔将 2 类正确分开的超平面，即最优分类超平面。在它的基本形式中，SVM 学习了这样的决策规则 $f(\boldsymbol{x})=\mathrm{sig}\{\omega^{\mathrm{T}}\boldsymbol{x}+b\}$，其中 ω和b 为决策函数中的权向量和阈值。设给定的训练集为$\{(\boldsymbol{x}_k, y_k)\}$，$k=1,\cdots,m$，其中样本 $\boldsymbol{x}_k\in\mathscr{R}^n$，它所属于的类别 $y_k\in\{-1,1\}$，则 SVM 分类算法可归结为下列二次规划问题：

$$\begin{aligned}&\min_{\omega,b,\xi}\quad \frac{1}{2}\omega^{\mathrm{T}}\omega+C\sum_{k=1}^{m}\xi_k,\\ &s.t.\quad y_k(\omega^{\mathrm{T}}\boldsymbol{x}_k+b)\geqslant 1-\xi_k,\qquad (1)\\ &\xi_k\geqslant 0,\ k=1,\cdots,m.\end{aligned}$$

其中：$\xi_k\geqslant 0$ 为松弛项，表示错分样本的惩罚程度；C 为常数，用于控制对错分样本惩罚的程度，实现在错分样本数与模型复杂性之间的折衷。在实际应用中，我们采用核函数 $k(\boldsymbol{x},\boldsymbol{x}')$ 来计算高维空间中的内积以解决原空间中的线性不可分问题。本文选择最常用的径向基核函数(RBF)来训练SVM 模型[4]。

3　拆分特征选择算法

拆分特征选择算法分为 2 部分：①对单个 SVM 进行特征选择；②采用一定的策略组合 SVM 用于实际预测。在①中，为了使特征选择算法具有filter 方法的高效率和 wrapper 方法的高性能，我们提出一种基于 filter 和 wrapper 的组合特征选择方法对单个 SVM 进行特征筛选；在②中，为了适应信用数据分布不平衡的特点、进一步提高建模速度，我们采用二叉树结构作为 SVM 的组合策略。

3.1　单个 SVM 的组合特征选择

首先，我们采用 Relief 评估[5](filter)计算各个输入特征的类别区分能力(权重)，但并不进行特征筛选，仅仅进行特征排序；之后，根据排序的结果，再用 wrapper 方法提取精度高的特征子集。Relief 是一种基于距离准则的特征权重算法，其基本思想是：好的特征应该使同类的样本接近，不同类的样本彼此远离。它是一种二分类特征选择方法，对数据类型无限制，对噪音不敏感，是一种较好的 filter 式特征评估算法[6]。在后续的 wrapper 方法中，SVM 五重交叉验证(5—fold Cross Validation，5—fold CV)的平均准确率被用作特征集分类性能的评价准则，并在选择过程中综合考虑了特征集的维数。特征选择具体步骤如下。

1)用 Relief 方法计算原始特征集中各个特征的权重并进行降序排序；同时计算原始特征集的 SVM 5—fold CV 平均准确率，记为 acc_1。

2)去掉 Relief 排序中权重最小的 2 个特征。

3)用剩余的特征对训练集再进行 Relief 评估，对其权重降序排列；同时计算该特征集的 SVM 5—fold CV 平均准确率，记为 acc_j。

4)重复步骤 2)～3)直到剩余 2 个特征为止。

5)在 1)～4)得到的一列SVM 5—fold CV 平均准确率中找到它的最大值，记为maxacc；为折衷考虑特征集的维数及其分类性能，寻找 SVM 5—fold

CV 平均准确率满足条件 $\max acc - acc_j < 0.5\%$ 的维数最小的特征子集，作为特征选择的结果，用于后续的建模与预测。

3.2 SVM 的二叉树多分类方法

我们采用二叉树结构作为 SVM 的组合策略[7]。该方法采用了一种“one against others”的分类器构造，以3类问题为例（分别记为A类、B类、C类），其结构图如图1所示。

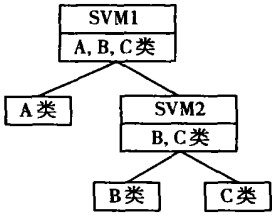

图1 基于 SVM 的二叉树多分类法结构图

Fig. 1 Structure of the multi-classifying binary tree based on SVM

其中，SVM1 采用整个训练样本集进行训练，将A作为一类，B、C作为另一类。这样用于测试时，SVM1 可以先将 A 类的样本从总体中区分出来，剩下的 B 类与 C 类样本采用 SVM2 进行区分。二叉树 SVM 组合的好处是各级 SVM 的训练数据逐渐减少，可加快建模速度；所用的分类器个数较少，可节省训练和预测的时间；且银行提供的企业数据，其规模随着等级的降低而不断缩小，二叉树算法可明显改善数据不平衡的现象，提高模型的预测能力。我们将 3.1 中特征选择后的各个 SVM 进行二叉树组合，应用于企业信用评估问题中。

4 仿真实验

4.1 数据预处理

本文中的实验数据来自于福建省某商业银行2003年度的客户资料，其中共有 2 890 家企业的财务数据，每条数据包含定量财务指标及银行给予的信用等级（共有 5 个等级：AAA 级、AA 级、A 级、B 级及 C 级，表征从好到差的信用程度）。根据财务专家的建议，本文选取财务指标中的 24 个财务比率作为原始特征集并将其编号，编号及其具体含义如表 1 所示。

表1 财务比率指标编号及其具体含义

Tab. 1 ID of the financial ratio indexes

编号	指标名称	编号	指标名称	编号	指标名称
1	资产负债率	9	净资产收益率	17	速动比率
2	流动比率	10	净资产增长率	18	固定资产/总资产
3	流动资产周转次数	11	或有负债率	19	息税前收益/总负债
4	销售利润率	12	产品销售率	20	息税前收益/营运资本
5	总资产报酬率	13	销售收入归行份额与贷款份额比	21	销售收入/总资产
6	利息保障倍数	14	存贷比	22	流动负债/净资产
7	营运资本比率	15	利息偿还率	23	存货/销售收入
8	经营性现金比率	16	到期信用偿还率	24	净现金流量/总负债

企业数据中，B 级与 C 级的数据特别少，够不成SVM 学习的有效模式，将其与 A 级合成一类（下文记为 ABC 级）作为 SVM 的输入。这样信用评估变为一个 3 类问题，在二叉树结构中仅需建立2 个 SVM 模型。数据中，行业跨度很大。不同行业的企业其财务特征存在着差异，即使财务比率相同，反映的财务意义也会有所不同。所以我们分行业进行财务指标选取，并选择其中数据量较多的轻工业和商业数据作为本文的实验数据。从实验数据的每个等级中，随机抽取 75%作为训练样本以选择指标，剩下的 25%作为测试样本，验证选取指标的预测效果。抽样后数据的分布情况为：①轻工业：全体训练/测试数据数为 857/286，其中 AAA 级训练/测试数为 437/138、AA 级训练/测试数为 347/124、ABC 级训练/测试数为 73/24；②商业：全体训练/测试数据数为 511/172，其中 AAA 级训练/测试数为 249/83、AA 级训练/测试数为 225/76、ABC 级训练/测试数为 37/13。

4.2 实验及结果分析

将采用 AAA 级、AA 级—ABC 级训练样本建模的 SVM 记为 SVM_1；采用 AA 级、ABC 级建模的记为 SVM_2；它们的二叉树组合记为 SVM—BT。则按照 3 中的步骤进行特征选择，SVM_1 与 SVM_2 的选出特征子集及其与原始特征集的 5—fold CV 平均

准确率比较如表 2 所示，其中 CVAcc－SS 与 CVAcc－OS 分别表示选出特征子集和原始特征集的 5－fold CV 平均准确率。特征选择之前与特征选择之后的 SVM－BT 测试准确率对比如表 3 所示。

表 2　选出特征子集及其与原始特征集的 5－fold CV 平均准确率对比

Tab. 2　Comparison of the performances between the selected feature subsets and the original set

		选出特征的编号	CVAcc－SS/%	CVAcc－OS/%
轻工业	SVM_1	2, 5, 6, 8, 9, 10, 12, 13, 14, 18, 19, 23(12 个)	81.533 1	80.487 8
	SVM_2	1, 2, 3, 4, 5, 6, 8, 10, 12, 13, 18, 23 (12 个)	89.510 5	89.510 5
商业	SVM_1	4, 7, 8, 9, 10, 11, 12, 13, 14, 21, 22, 23 (12 个)	76.744 2	73.385 5
	SVM_2	3, 4, 5, 7, 11, 15, 16, 18, 21, 22 (10 个)	85.893 3	86.098 8

表 3　特征选择之前与特征选择之后的 SVM－BT 测试准确率对比

Tab. 3　Comparison of the performances of SVM－BT before and after feature selection

	轻工业测试准确率/%	商业测试准确率/%
特征选择前	75.26	68.66
特征选择后	77.00	73.84

从表 2 中可以看出，拆分特征选择后，单个分类器的输入特征个数大大减少，两组数据各个 SVM 的输入特征数都降到了原始特征集的一半左右，这极大地减少了模型的复杂程度和建模时间；且选出特征集的 5－fold CV 平均准确率都高于或接近原始特征集，说明该特征选择方法在减少特征个数的同时，保证了单个分类器的泛化性能。从表 3 中可以看出，选出特征子集测试准确率比原始特征集高。可见拆分特征选择能通过提高单个分类器的性能，使得组合算法的整体分类性能也得以提高。从表 2 中还可以看出，同个行业中不同模型的特征子集，既具有共性也具有差异性。这可直观地解释为相同地特征是信用评级的公共关键指标，而不同的特征则表示其更擅长于分辨哪两个等级。这为财务专家从指标的经济意义上解释特征选择的结果提供了依据。

5　结语

企业信用评估是一个多分类问题，我们采用 SVM 的组合来进行建模和预测。针对组合 SVM 的特征选择，传统方法存在着算法复杂、运行速度慢、选出特征子集对单个分类器针对性差的缺点。对此我们提出一种拆分特征选择方法，其主要思想是对 SVM 组合中的各个分类器分别进行特征选择，再采用不同的特征子集作为子分类器的输入，进行建模与预测。基于 filter 和 wrapper 相结合的思想，我们采用 Relief 评估(filter)对特征进行排序，并以 SVM 5－fold CV 平均准确率作为 wrapper 方法的特征集分类性能评价准则，综合考虑特征集的维数，进行了子分类器的特征选择；之后，针对企业信用评级问题的特点，采用二叉树结构作为 SVM 的组合策略。实验表明，拆分特征选择方法能选出规模较小、具有一定差异的关键指标集，提高了模型的分类性能，并且具有计算简单、运行快速的优点。

参考文献：

[1] 刘 闽，林成德. 基于支持向量机的商业银行信用风险评估模型[J]. 厦门大学学报，2005，44(1)：29－32.

[2] Kohavi R，John G. Wrappers for feature subset selection[J]. Artificial Intelligence，1997，12：273－324.

[3] Cristianini N，Shawe－Taylor J. 支持向量机导论[M]. 李国正，王猛，曾华军，译. 北京：电子工业出版社，2004.

[4] Chen Yiwei，Lin Chihjen. Combining SVMs with various feature selection strategies[EB/OL]，http：//www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

[5] Kira K，Rendell L A. The feature selection problem：Traditional methods and a new algorithm[C] // Proceedings of the Tenth National Conference on Artificial Intelligence，Menlo Park：AAAI Press/The MIT Press，1992：129－134.

[6] 张丽新，王家厥，赵雁南，等. 基于 Relief 的组合式特征选择[J]. 复旦学报，2004，43(5)：893－897.

[7] 马笑潇，黄席樾，柴毅. 基于 SVM 的二叉树多类分类算法及其在故障诊断中的应用[J]. 控制与决策，2003，18(3)：272－277.