文章编号：2095-2163(2021)12-0043-05 中图分类号：TP393.4 文献标志码：A

# 基于代价敏感加权支持向量机的员工离职分类预测

万毅斌，王绍宇，秦彦霞
（东华大学 计算机科学与技术学院，上海 201620）

**摘 要：**企业员工在职及离职数据集往往具有高度非均衡的特点，因此使用传统支持向量机（Support Vector Machine，SVM）分类算法来对非平衡的企业员工数据集进行分类并进行离职预测时，往往会导致分隔超平面向少数类偏移，分类准确率不佳等情况。为解决以上问题，本文首先通过 SMOTE 过采样方法有效地减少数据集的非均衡性，针对 SMOTE 方法导致的过拟合问题，本文还提出了改进的代价敏感加权算法来 SVM 优化算法。通过某大型外企公司企业员工数据集进行的实验证明，相对于 SVM 及 SMOTE-SVM 算法，本文提出的改进算法在 G-mean 和 F-measure 上分别达到了 99.08%和 89.25%，分类准确度和性能都得到了较大提升，能有效地用于非均衡企业员工数据的分类及离职预测。

**关键词：**非平衡的企业员工数据集；代价敏感加权算法；SMOTE-SVM 算法

# Prediction of employee turnover based on cost sensitive weighted SVM

WAN Yibin，WANG Shaoyu，QIN Yanxia
（College of Computer Science and Technology，Donghua University，Shanghai 201620，China）

**【Abstract】**The on-the-job and turnover datasets of enterprise employees often have the characteristics of high imbalance. Therefore，when using the traditional support vector machine（SVM）classification algorithm to classify the imbalanced employee datasets，the separation hyperplane will often move to minority classes. In order to solve the above problem，the method proposed in this paper first reduces the imbalance of the data by smote oversampling method. Aiming at the over fitting problem caused by smote method，this paper also proposes an improved cost sensitive weighting algorithm to optimize the SVM algorithm. Experiment on a large foreign enterprise employee dataset shows that compared with SVM and smote-SVM algorithm，the improved algorithm proposed in this paper achieves 99.08% and 89.25% on G-mean and F-measure respectively. The classification accuracy and performance have been greatly improved and the proposed method can be effectively used in the classification and turnover prediction of imbalanced enterprise employee data.

**【Key words】**imbalanced employee datasets；cost sensitive weighting algorithm；smote-SVM algorithm

## 0 引 言

随着中国经济的高速发展，国内各类科技公司不断涌现，传统行业加速转型，许多省市都出台了各种各样的政策来吸引和留住人才，同样对于企业来说如何吸引人才和如何留住人才都是对企业的发展至关重要的。企业员工流失对于企业而言并不是简单人员流失，而会对企业的人事、财务、业务等多方面造成诸多影响，比如已投入费用的损失，流失员工所负责相关工作的临时性中断，流失员工可能会带走企业一些重要客户或关键技术，从而使企业承受巨大损失[1]。

随着员工需求和社会环境的不断变化，不同企业的员工所关心的点也不尽相同。根据调查统计，2020 年国内企业员工离职率为 19.8%，其中主动离职率达到了 13.4%，相较于过去有明显的升高。

目前很多企业在原有员工数据库的基础上，还通过统计、调查和问卷等方式建立了可用于预测员工离职倾向的数据集，以供人力资源等部门进行预警分类。目前，主流的分类算法，如基于结构风险最小化的 SVM 能克服传统分类器局部最优解、过拟合、维数灾难等缺点[2]。但对于企业员工离职倾向预测、疾病诊断、欺诈检测等不平衡数据集的处理上，SVM 在训练过程中由于自身的原因，以及数据集存在的界模糊，噪声污染等问题，导致对不平衡数据集的分类效果不佳[3]。为此，Bagging、Boosting、rotation forest 等一些组合算法提出，来解决分类问题中的数据不均衡问题[4]；董燕杰等提出的 Random-SMOTE 算法对小类样本进行上采样以平衡数据集，有效地解决了不平衡数据集中小类分类困难的

基金项目：国家自然科学基金（62006039）。

作者简介：万毅斌（1993-），男，硕士研究生，主要研究方向：大数据、机器学习；王绍宇（1973-），男，博士，副教授，主要研究方向：大数据、图像识别；秦彦霞（1985-），女，博士，讲师，主要研究方向：自然语言处理、信息抽取。

收稿日期：2021-09-09

问题[5];覃朗提出一种基于信息增益的超立方体顶点采样 SMOTE-SVM 算法,通过优化算法对改进后的 SMOTE-SVM 模型的参数进行自动寻优,进而增强了算法参数设置的合理性,提升了分类性能[6]。

现有方法大多使用 SMOTE 算法先对少数类样本进行合成,解决不平衡分类问题,但通常 SMOTE 与 SVM 结合是对少数类样本进行处理,没有结合 SVM 算法本身的特点,从而导致分类效果不够稳定。针对以上缺点,本文提出了一种改进的代价敏感算法,通过对合成的样本赋予错分代价,来增加通过 SMOTE 算法合成的数据集的合理性,减少了可能存在的过拟合风险,提升了对企业员工不平衡数据集的分类效果和稳定性。

## 1 基于 SMOTE-SVM 的企业员工离职分类

### 1.1 传统企业员工离职 SVM 分类算法

假设某企业员工信息数据样本集为 $\{(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_n,y_n)\}$,$i=1,\cdots,n$,其中 $n$ 代表该企业员工数量,$x_i \in R^m$,$m$ 表示该企业员工的信息维数,分类标签 $y_i=\{-1,+1\}$,其中 $-1$ 代表已经离职的员工,$+1$ 代表在职员工。本文使用的 SVM 算法通过在 $R^n$ 空间上寻找一个使分类边界 $\frac{1}{2}\|w\|^2$ 最小的实数函数 $g(x)=(W^{\mathrm{T}}x+b)$,从而确定企业员工是否离职的分类决策平面,使用决策函数 $f(x)=sgn(g(x))$ 来预测输入的任意一名新员工 $x$ 对应的是否可能离职分类类别 $y$。

对于一般的线性可分问题,SVM 通过求解下列二次规划问题得到最优分类超平面,式(1):

$$\min \frac{1}{2}\|w\|^2$$
$$\text{s.t.} y_i(W^{\mathrm{T}}x_i+b)\geqslant +1 \quad (1)$$
$$i=1,\cdots,n$$

对于这样的二次规划问题,通常转换成与其对应的 Lagrange 对偶问题来求解,该问题对应的 Lagrange 函数为式(2):

$$L(\alpha,w,b)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i[y_i(W^{\mathrm{T}}x_i+b)-1] \quad (2)$$

其中,$\alpha_i \geqslant 0$ 为 Lagrange 乘子。可利用 Lagrange 对偶方法将式(2)转化为对偶问题,式(3):

$$\max_{\alpha} L(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j$$
$$\text{s.t.}\sum_{i=1}^{n}y_i\alpha_i=0 \quad \alpha_i\geqslant 0 \quad (3)$$

求解后可得到分类决策的超平面函数,式(4):

$$f(x)=w^{\mathrm{T}}x+b=\sum_{i=1}^{n}\alpha_i y_i x_i^{\mathrm{T}}x+b \quad (4)$$

由于企业员工信息数据集的维度较高,分布不均匀,无法通过 SVM 对一般的线性可分问题的求解方法寻找分类超平面,因此本文使用核函数将企业员工样本数据集映射到高维空间,在高维空间求解分类超平面。通过核函数不仅无须知道高维变换的显示公式,还解决了高维数据带来的问题。对于给定的核函数 $K(x,y)=\varphi(x)\varphi(y)$,则非线性 SVM 的对偶问题可以写成式(5)形式:

$$\max_{\alpha} L(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)$$
$$\text{s.t.}\sum_{i=1}^{n}y_i\alpha_i=0\ \alpha_i\geqslant 0 \quad (5)$$

通过上述 SVM 算法的分类原理可知,SVM 算法分类的结果是由分隔超平面所决定,该超平面也就是最终的决策函数,通过两类样本中的少量样本点即支持向量所决定的,所以对于样本中其他的非支持向量数据,不会影响 SVM 算法的分类性能,算法的复杂性主要取决于支持向量的数量[7]。因此,传统 SVM 算法一般在数据集中正类与负类样本数量大致相同的情况下有较好的表现,而面对现实应用领域中数据集不平衡的特点,由于 SVM 算法决策平面偏移程度不足、支持向量分布不均匀等自身特点,其分类性能往往会大打折扣[8]。在企业员工离职倾向分类问题中,特别是对规模较大的企业,离职员工数量一般占员工总数的比例很小,但对于企业来说培养一名员工所投入的花费很大,为了避免和预防可能出现员工离职潮,能够提早发现员工离职倾向并采取措施是非常重要的。传统 SVM 在处理不平衡数据分类问题时,分类平面会向少数类偏移,即将更多的少数类样本错分为多数类,这样会导致企业对员工离职倾向判断不准确。

### 1.2 改进的 SMOTE-SVM 分类算法

为了解决上述问题,本文引入 SMOTE 算法,通过人工合成少数类样本,即离职员工的数据集,使离职员工数据数量与在职员工数据量达到均衡。具体操作是:首先找到离职员工样本 $x_i$ 的 $k$ 个邻近同类样本,在这 $k$ 个样本中随机选取一个 $x_j$,通过公式(6)合成新的样本:

$$x_{new}=x_i+rand(0,1)\times(x_i-x_j) \quad (6)$$

虽然 SMOTE 过采样方法已被证明在许多不平衡数据上表现良好，但是其对分类分布进行了假设。使用 SMOTE 算法对数据集进行过采样处理，会使得 SVM 算法存在一定的过拟合风险，同时 SMOTE 算法在合成数据的时候，并未考虑噪声的影响，会导致合成的数据增加了原始样本的噪声率，最终导致影响 SVM 算法的准确性[9]。

## 2　基于代价敏感加权的 SMOTE-SVM 方法

为减少 SMOTE 算法存在的过拟合问题，可通过改进的代价敏感算法对少数类、多数类以及合成实例进行不同的加权处理。改进的 SMOTE-SVM 的原始优化函数，式(7)：

$$\min_{w,b,\xi}\frac{1}{2}w^{\mathrm{T}}w + C\times k^{maj}\sum_{i=1}^{|S^{maj}|}\xi_i^{maj} + C\times k^{\min}\sum_{i=1}^{|S^{\min}|}\xi_i^{\min} + C\times k^{syn}\sum_{i=1}^{|S^{syn}|}\xi_i^{syn}$$

$$\text{s.t. } w^{\mathrm{T}}\varphi(x_i) + b \leqslant 1 - \xi_i^{maj} \quad \forall\ x_i \in S^{maj}$$

$$w^{\mathrm{T}}\varphi(x_i) + b \leqslant 1 - \xi_i^{\min} \quad \forall\ x_i \in S^{\min}$$

$$w^{\mathrm{T}}\varphi(x_i) + b \leqslant 1 - \xi_i^{syn} \quad \forall\ x_i \in S^{syn}$$

$$\xi_i^{maj},\xi_i^{\min},\xi_i^{syn} \geqslant 0 \tag{7}$$

其中，权重因子 $c^{maj}$、$c^{\min}$、$c^{syn}$ 控制了多数类、少数类和合成实例的错分代价。该方法通过对合成实例和原始的少数实例进行不同的加权，使得 SVM 能够更加精细地控制分离超平面。

通过(7)式求解，得到的 $a^*$ 来确定新样本实例 $a_{new}$ 的类别 $y$，式(8)：

$$y = sign\{\sum_{i=1}^{|S|}a_i^*\ y_iK(x,x_i) + \sum_{i=1}^{|S^{syn}|}a_i^*K(x,x_i^{pq}) + b\} \tag{8}$$

算法的主要流程如下：

代价敏感加权算法及 SMOTE-SVM 算法：

Input：企业员工数据集，其中 $S^{\min}$ 代表已离职员工，

$S^{maj}$ 代表在职员工。

Output：新员工离职倾向 SVM 预测模型

Initalization：少数类采样集合 $x_p$；随机邻近类集合 $x_q$；生成样本集合 $S^{syn}=\{\}$；

加权向量 $\boldsymbol{C}=\{\}$；$k=(S^{maj}-S^{\min})/P$，$P$ 为 SMOTE 循环次数

循环 1~ $P$ 次

从少数集 $S^{\min}$ 中随机取得一个样本 $x_p$

随机获取 $x_p$ 的邻居

使用 knn 算法获取邻近样本 $x_q$

把 $x_q$ 加进 $S^{syn}$ 集合

循环结束

初始化成本敏感加权值：$k^{maj}=N^{maj}/N$，$k^{\min}=N^{\min}/N$，$k^{syn}=N^{syn}/N$

循环 1 ~ $N$ 次，$N$ 为样本数据集数量

根据类别分别把 $k^{maj}$，$k^{\min}$ 加入 $\boldsymbol{C}$

循环结束

把 $k^{syn}$ 加入 $\boldsymbol{C}$

通过公式(7)求解得到结果

视同公式(8)返回企业员工离职倾向分类模型

## 3　实验结果及分析

### 3.1　数据来源

本文所使用的数据为某外企 2015~2020 年内的所有在离职员工数据集。该数据集包含了5 020 条员工的基础以及相关信息数据，包括员工的年龄、性别、职位等级、加班情况、旅游情况和公司满意度等 35 列特征信息，其中已离职员工数量为 434 名，在职员工数量为 4 586 名，离职员工数量与在职员工数量的比例为 1∶10，符合不平衡数据集的特点。

由于上述数据集中包含缺失值、噪音以及人工录入错误导致的异常值存在，不利于算法模型的训练。所以在训练实验前，首先要对数据集中的脏数据进行数据清理、集成和规约，使其能够达到标准，满足训练要求的数据集。

本文针对外企员工数据集中所存在的数据缺失、噪声及冗余等问题做了数据预处理。主要工作包括一是缺失值处理：对缺失率较高，且重要性较低的信息，比如亲属相关信息等直接删除变量，再使用随机插值法对小部分缺失值进行补充；二是冗余数据处理：由于企业员工数据信息数据集包含较多的属性信息，其中有部分属性与模型训练任务不相关，属于冗余数据，使用 matlab 的 scikit-learn 中的递归特征消除算法，由整个数据集开始，逐步删除尚在数据集中的最坏属性。

### 3.2　评价指标

本文采用基于混淆矩阵的评价方法，见表 1。其中 $TP$ 表示实际是正类且被正确分为正类的样本的数目；$FN$ 表示实际是正类但被错误分为负类的样本的数目；$FP$ 表示实际是负类但被错误分为正类的样本的数目；$TN$ 表示实际是负类且被正确分为负类的样本的数目。

表 1 评价混淆矩阵

Tab. 1 Evaluation confusion matrix

| 实际 | 预测 | |
|---|---|---|
| | 正类 | 负类 |
| 正类 | *TP* | *FN* |
| 负类 | *FP* | *TN* |

通过表 1 可计算出 5 种评估标准：

(1)查准率 *Precision*，表示预测正确的正类占总样本的比例，公式(9)：

$$precision = \frac{TP}{TP + FP} \tag{9}$$

(2)查全率 *Recall*，表示预测正确正类占所有正类的比例，公式(10)：

$$recall = \frac{TP}{TP + FN} \tag{10}$$

(3) *Overall Accuracy*(*OA*)，表示每个样本所分类的结果与检验数据类型一致的概率，公式(11)：

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{11}$$

(4) $F - measure$ 是查全率和查准率的调和值，是综合评价指标，公式(12)：

$$F - measure = \frac{(1 + \beta^2) \times recall \times precision}{\beta^2 \times recall + precision} \tag{12}$$

(5)G-mean 表示算法在正确正类和负类的平均性能，公式(13)：

$$G - mean = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} \tag{13}$$

其中，G-mean 考虑了两类样本的分类性能，只有分类平面不发生偏移，两类样本都有较大的查全率，$G - mean$ 值才会较大。$F - measure$ 考虑了少数类的查全率和查准率，任何一个值的变化都能影响 $F$ 的大小，因此能全面反映分类器对少数类样本的分类性能。

### 3.3 实验结果及分析

本文采用 Matlab 工具中的 LibSVM 工具箱在某公司员工信息数据集上进行实验，对比传统 SVM 和 SMOTE-SVM 两种模型，验证本文方法的有效性。实验采用 *RBF* 核函数，*gamma* 值取 1。由于企业员工信息数据中存在缺失和冗余数据，因此首先对原始数据集进行了预处理，然后利用 3 种模型进行学习，最后使用 $G - mean$ 和 $F - measure$ 衡量各方法的分类精确度，结果见表 2。

表 2 SVM、SMOTE-SVM 与本文方法比较

Tab. 2 Comparison between SVM, SMOTE-SVM and our method

| 实验次数 | SVM | | SMOTE-SVM | | MCS-SMOTE-SVM | |
|---|---|---|---|---|---|---|
| | $G - mean$/ % | $F - measure$/ % | $G - mean$/ % | $F - measure$/ % | $G - mean$/ % | $F - measure$/ % |
| 1 | | | 97.62 | 82.84 | 99.02 | 90.63 |
| 2 | | | 97.85 | 83.05 | 99.05 | 87.47 |
| 3 | | | 98.02 | 84.12 | 99.46 | 92.64 |
| 4 | | | 97.78 | 82.98 | 99.05 | 87.12 |
| 5 | | | 97.62 | 82.84 | 99.32 | 89.82 |
| 6 | | | 97.62 | 82.84 | 98.76 | 90.57 |
| 7 | | | 97.62 | 82.84 | 98.70 | 86.45 |
| 8 | | | 96.84 | 82.56 | 99.25 | 89.33 |
| 平均精度 | 83.28 | 81.95 | 97.62 | 83.01 | 99.08 | 89.25 |

从表 2 中的 3 种方法的比较结果可以看出，由于未考虑不平衡数据集的问题，传统 SVM 算法在三者中表现最差，$G - mean$ 和 $F - measure$ 分别只有 83.28% 和 81.95%；使用 SMOTE 算法对少数类样本进行新实例合成，多数类与少数类样本数量基本达到一致，分类精度有了明显的提升，$G - mean$ 和 $F - measure$ 分别达到了 99.08% 和 89.25%；本文方法对 SMOTE-SVM 方法增加了改进的代价敏感算法，对少数类样本、多数类样本和新合成的样本进行加权处理，实验精度有了进一步的提升。本文方法的 $G - mean$ 值比 SMOTE-SVM 略高，而 $F - measure$ 值均比 SMOTE-SVM 高 10%，充分证明了本文方法对企业员工离职倾向分析的有效性。

## 4 结束语

本文以某大型外资企业为例，针对 2015~2018 年 3 年的员工信息数据集，首先对原始数据集进行了数据预处理，包括缺失值补充，冗余数据处理等；针对传统 SVM 分类器在处理不平衡数据集时分类超平面会向少数类偏移的特点，以及使用 SMOTE 算法对数据样本集进行上采样后的合成数据会对 SVM 算法造成过拟合风险的问题，本文提出了一种改进的基于代价敏感算法与SMOTE-SVM算法，该

(下转第 53 页)