# 一种面向单个正例的 Fisher线性判别分类方法

尹军梅, 杨 明

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要] 提出了一种解决不平衡数据集中少数类只有一个样本的方法, 找出单个正例在负类中的 $k$ 个近邻, 按照一定规则依次在单个正例和它的各个近邻的连线上产生合成样本, 并把这些合成样本添加到原始的正类中, 用加权 Fisher线性分类方法对新的数据集进行训练. 实验结果表明该方法可有效地提高少数类的分类性能.

[关键词] 不平衡数据集, Fisher线性判别, 过抽样

[中图分类号] TP311 [文献标识码] A [文章编号] 1672-1292(2008)03-0061-05

# A Fisher Linear Discriminant Classification Approach Dealing With Single Positive Sample

Yin Junmei, Yang Ming

(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China)

**Abstract** An approach to dealing with imbalanced data set with only one positive sample is proposed. After finding out the K-Near-Neighbours(K-NN) of the single positive sample, according to certain rules, synthetic samples are produced in turn on the connected lines between the single positive sample and every near neighbour of it. Then the produced synthetic samples are added to the original positive classes. Further, the new data set is trained with the weighing Fisher linear discriminant classification approach. In the experiment, eight data sets are chosen from UCI, and the data sets are trained. The results show that this approach can improve the classification performance of the minority classes effectively.

**Key words** imbalanced data set; fisher linear discriminant(FLD); over-sampling

分类问题是机器学习领域的重要研究内容之一, 现有的一些分类方法都已经相对成熟, 用它们来对均衡数据进行分类一般都能取得较好的分类性能. 但在现实中数据往往都是不平衡的, 例如信用卡交易欺诈识别[1]、电信设备故障预测[2]、企业破产预测[3] 和雷达图像监测海洋石油污染[4] 等. 许多现有的分类器的设计都是基于类分布大致平衡这一假设的, 如果用这些方法来对不平衡数据进行分类就会导致分类器的性能下降, 因此不平衡数据的分类问题已成为机器学习领域的一个新的研究热点.

对于不平衡问题, 现有的解决方法有以下几种类型: 数据层的方法、算法层的方法以及集成的方法. 数据层的方法是在训练前对数据集进行重抽样, 用处理过的数据来训练产生分类器. 抽样方法可分为对少数类的过抽样[5] 和对多数类的欠抽样[6], 也可把两者结合起来[7]. 算法层的方法是对已有的算法进行改进, 使它能够适用于不平衡数据, 如已有对决策树 C4.5[8]、支持向量机 SVM[9]、非平衡数据集 Fisher线性判别模型[10] 等一些算法的改进. 已有的解决不平衡数据分类的集成的方法, 多数是把抽样方法和集成方法结合起来[11], 对原始训练集进行一系列抽样, 产生多个分类器, 然后用投票或合并的方式输出最终结果.

本文的工作是解决不平衡数据集的正类中只有一个样本时的分类问题. 已有人提出非平衡数据集 Fisher线性判别模型, 该方法能够较好地解决不平衡问题. 它是通过加权的办法来减小样本不平衡所造成的影响, 但该算法的前提是少数类的样本的数目至少为 2, 因为如果正类中只有一个样本, 此时该类的样本类内离散度矩阵 $S_1$ 为 0, 这时对式子 $S_w = S_1 + S_2$ 中的各类类内离散度阵 $S_i$ 分别进行加权是没有意义

收稿日期: 2008-06-18
基金项目: 国家自然科学基金(40771163)资助项目.
通讯联系人: 杨 明, 教授, 博士, 研究方向: 数据挖掘、机器学习与粗糙集理论及应用研究. E-mail: myang@njnu.edu.cn

的. 本文提出了一种解决办法, 找出单个正例在负类中的 $k$ 个近邻, 用合成的办法产生 $k$ 个合成样本, 将这 $k$ 个样本添加到正类中, 对抽样后的数据集用加权 Fisher线性分类方法 (WFLD) 来进行训练.

## 1　加权 *Fisher*线性判别

### 1.1　*Fisher*线性判别 *(FLD)*

Fisher线性判别分析的基本思想[12] 是通过寻找一个投影方向, 将高维问题降低到一维问题来解决, 并且要求变换后的一维数据具有如下性质: 同类样本尽可能聚集在一起, 不同类的样本尽可能地远. 假设有一个包含 $N$ 个样本的 $d$ 维样本集 $X = \{x_1, x_2, \cdots, x_n\}$, 其中 $N_1$ 个样本来自第一类样本 $X_1$, $N_2$ 个样本来自第二类样本 $X_2$; 将 $x_n$ 投影到一维空间上, 即 $y_n = \boldsymbol{w}^{\mathrm{T}} x_n$, 于是 $N$ 个 $d$ 维样本的样本集 $X$ 变换成包含 $N$ 个一维样本的样本集 $Y$, 并使得同类样本尽可能聚集在一起, 不同类的样本尽可能地远. 为求得最佳投影方向 $\boldsymbol{w}^*$ 及相应的 Fisher判别函数 $\boldsymbol{g}(x) = \boldsymbol{w}^{*\mathrm{T}} x - \boldsymbol{y}_0$, 提出这样一种解决方法, 寻找使得 Fisher准则函数 $\boldsymbol{J}_F(w) = \dfrac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{w}}$ 的值取为最大的 $w$, 即为所要求的最佳投影方向 $\boldsymbol{w}^*$. 由文献[12] 可得一种求解 $\boldsymbol{w}^*$ 和 $\boldsymbol{y}_0$ 的方法, 求得 $\boldsymbol{w}^* = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$, $y_0 = \dfrac{\boldsymbol{m}_1 + \boldsymbol{m}_2}{2}$. 其中, 在原始 $d$ 维空间中: $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$ 分别为第一类和第二类样本的均值; $\boldsymbol{S}_i$ 和 $\boldsymbol{S}_w$ 分别为样本类内离散度矩阵及总类内离散度矩阵, $\boldsymbol{S}_i = \sum_{x \in X_i} (x - \boldsymbol{m}_i)(x - \boldsymbol{m}_i)^{\mathrm{T}}$, $i = 1, 2$; $\boldsymbol{S}_w = \boldsymbol{S}_1 + \boldsymbol{S}_2$; $\boldsymbol{S}_b$ 为样本类间离散度矩阵, $\boldsymbol{S}_b = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\mathrm{T}}$. 在投影后的一维空间中, 各类样本均值 $\boldsymbol{m}_i = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{m}_i$; 样本类内离散度 $\boldsymbol{S}_i^2$和总类内离散度 $\boldsymbol{S}_w$ 分别为 $\boldsymbol{S}_i^2 = \sum_{y \in Y_i} (y - \boldsymbol{m}_i)^2$, $\boldsymbol{S}_w = \boldsymbol{S}_1^2 + \boldsymbol{S}_2^2$; 样本类间离散度 $\boldsymbol{S}_b$ 为 $\boldsymbol{S}_b = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{w}$.

对未标签的样本 $x$, 若 $\boldsymbol{g}(x) > 0$, 则该样本属于第一类; 若 $\boldsymbol{g}(x) \leqslant 0$, 则 $x$ 属于第二类.

### 1.2　加权 *Fisher*线性判别模型 *(WFLD)*[10]

因分类器性能是由投影方向 $\boldsymbol{w}^*$ 决定的, 而投影方向 $\boldsymbol{w}^*$ 由总类内离散度阵 $\boldsymbol{S}_w$ 和两类样本均值矢量之差共同决定. 在独立同分布假设下, 样本均值矢量与样本个数无关, 即两类样本均值矢量差与样本不平衡无关, 也即投影方向 $\boldsymbol{w}^*$ 由类内离散度阵 $\boldsymbol{S}_w$ 惟一决定. 设两类的样本协方差矩阵分别为 $\Sigma_1$ 和 $\Sigma_2$, 则 $\boldsymbol{S}_w = \boldsymbol{S}_1 + \boldsymbol{S}_2$ 变形为 $\boldsymbol{S}_w = \boldsymbol{S}_1 + \boldsymbol{S}_2 = N_1 \Sigma_1 + N_2 \Sigma_2$, 若两类样本个数不平衡 $(N_1 \ll N_2)$, $N_2 \Sigma_2$ 对 $\boldsymbol{S}_w$ 的贡献远远大于 $N_1 \Sigma_1$ 对 $\boldsymbol{S}_w$ 的贡献, 从而可能导致投影方向不利于少数类的分类. 为消除样本个数不平衡的影响, 对 $\boldsymbol{S}_w = \boldsymbol{S}_1 + \boldsymbol{S}_2$ 中的各类类内离散度阵 $\boldsymbol{S}_i$ 进行分别加权, $\boldsymbol{S}_w = N_2 \boldsymbol{S}_1 + N_1 \boldsymbol{S}_2 = N_1 N_2 (\Sigma_1 + \Sigma_2)$, 使两个类对 $\boldsymbol{S}_w$ 的贡献平衡.

## 2　面向单个正例的 *Fisher*线性判别分析

当正类只有一个样本时, 上面的加权模型就不再适用. 因为加权方式是这样: 原本 $\boldsymbol{S}_w = \boldsymbol{S}_1 + \boldsymbol{S}_2$, 加权后的模型为 $\boldsymbol{S}_w = N_2 \boldsymbol{S}_1 + N_1 \boldsymbol{S}_2$, 假设其中 $\boldsymbol{S}_1$ 为正类的样本类内离散度矩阵, $\boldsymbol{S}_2$ 为负类的样本类内离散度矩阵, $N_1$ 为正类的样本数目, $N_2$ 为负类的样本数目. 对于正类只有一个样本的数据集, 正类的样本类内离散度矩阵 $\boldsymbol{S}_1$ 为 0, 此时总类内离散度矩阵 $\boldsymbol{S}_w = \boldsymbol{S}_2$, 上述的加权模型对这类问题就没有意义. 针对以上问题, 本文提出了一种面向单个正例的 Fisher线性判别分类算法 FLDDWSPS, 其基本思想是: 如果正类只有一个样本时, 先对正类进行过抽样 (这样正类的类内离散度矩阵 $\boldsymbol{S}_1$ 就不再为 0, 从而加权模型才能适用), 然后对抽样后的数据集用加权 Fisher模型 (WFLD) 进行训练. FLDDWSPS算法的具体描述如下.

**算法 1**　FLDDWSPS(Fisher Linear Discriminant Dealing With Single Positive Sample)

输入: 训练集 $T = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, $x_i \in \mathbf{R}^d$, $y_1 = +1$, $y_j = -1$, $i = 1, 2, \cdots, n$, $j = 2, 3, \cdots, n$.

输出: $\boldsymbol{y} = \boldsymbol{w}^{\mathrm{T}} x - y_0$.

Step 1　对于 $x_1$, 找出它在另一类中的 $k$ 个近邻 neighbor$_t$, $t = 1, 2, \cdots, k$;

Step 2　For $t = 1, 2, \cdots, k$

Step 2.1　产生一个 $(0, 1)$ 之间的随机数 $r_t$;

Step 2.2 由 $rt_1 = \frac{2}{3}\left(rt - \frac{1}{2}\right)$ 可得一个 $\left(-\frac{1}{3}, \frac{1}{3}\right)$ 之间的随机数;

Step 2.3 由 $s_t = x_1 + rt_1 \times dif_t$ 得到一个新的正类样本, 其中 $dif_t$ 是 $x_1$ 与近邻 neighbor$_t$ 的矢量差;

Step 3 将得到的 $k$ 个正类样本加到正类集中, 正类集合变为 $\{x_1, s_1, s_2, \cdots, s_k\}$;

Step 4 求出正类的样本类内离散度矩阵 $S_1$ 和负类的样本类内离散度矩阵 $S_2$, 并求出总类内离散度矩阵 $S_w = (n-1)S_1 + (1+k)S_2$; 进一步, 可得最佳投影方向 $w^* = S_w^{-1}(m_1 - m_2)$ 及 $y_0 = \frac{m_1 + m_2}{2}$.

## 3 实验设计

### 3.1 评价准则

机器学习中对于两类问题经常使用混合矩阵来评价分类性能, 如表 1所示. 精确度 accuracy = $\frac{(TP + TN)}{(TP + TN + FP + FN)}$ 是分类问题中常用的评价标准, 它反映分类器对数据集的整体分类性能, 但不能正确反映不平衡数据集的分类性能. 因为如果少数类的识别率很低而多数类的识别率很高的话, 这时整体的识别率很高, 可大多少数类却不能被识别. 为此, 针对不平衡数据, 采用如下的评价准则将更加合理:

F-value = $\frac{(1+\beta^2) * recall * precision}{(\beta^2 * recall + precision)}$, 其中, recall 和 precision分别为查全率和查准率. β是可调参数, 通常取值为 1; 对于少数类, 它的 recall = $\frac{TP}{(TP + FN)}$, precision = $\frac{TP}{(TP + FP)}$, 当它的 recall和 precision值都大时 F-value值才会大, 因此 F-value值能很好地反映少数类的分类性能.

表 1 混合矩阵

Table 1 Confusion matrix

| | 被分为正类 | 被分为负类 |
|---|---|---|
| 实际为正类 | *TP* | *FN* |
| 实际为负类 | *FP* | *TN* |

### 3.2 数据集

实验的 8个数据集选自公用机器学习数据库 UCI. 如果一个数据集为两类数据集, 则将数目较少的一类标为正类, 数目较多的标为负类; 如果一个数据集为多类数据集, 则将其中某一类标为正类, 其余的都标为负类, 如表 2所示.

表 2 实验数据集

Table 2 Experimental data sets

| 数据集 | 正类标签 | 负类标签 | 正类样本数目 | 负类样本数目 |
|---|---|---|---|---|
| liver_disorder(2类, 简记为 Liv) | 1 | 2 | 145 | 200 |
| Glass(7类) | 2 | 其余 | 76 | 138 |
| Pima(2类) | 1 | 0 | 268 | 500 |
| Sonar(2类) | 1 | 2 | 97 | 111 |
| Wdbc(2类) | 2 | 1 | 212 | 357 |
| Diabetes(2类) | 1 | 2 | 268 | 500 |
| Vehicle(4类) | 1 | 其余 | 217 | 629 |
| Thyroid(3类) | 3 | 其余 | 30 | 185 |

### 3.3 实验结果

通过以上处理可得到 8个不平衡数据集, 接着再用以下的方法分别对这 8个数据集产生各自的训练集与测试集.

从表 2所示的不平衡数据集生成训练集 (由于本文要解决的是面向单个正类的不平衡问题, 所以训练集的正类只包含一个样本)和测试集的步骤:

(1) 生成一个随机数. 从表 2所示的数据集的正类中抽出一个样本作为训练集的正类, 其余作为测试集的正类.

(2) 生成 Reduce个随机数. 从表 2所示的数据集的多数类中抽出 Reduce个样本作为测试集的负类, 其余作为训练集的负类.

本文要对单个正类进行过抽样, 产生 $k$ 个合成样本. 在这个实验中为了比较不同的值的效果, 让 $k$ 分

别取 6, 9. 当 $k$取以上 2个不同值时, 分别对表 2中的 8个数据集按照上面所述的 2个步骤产生训练集和测试集. 由于随机抽样具有一定的随机性, 为了避免这个问题, 重复抽样 10次, 这样每个不平衡数据集都产生 10个训练集和 10个测试集. 分别用算法 FLDDWSPS对训练集进行训练, 产生 10个分类器, 再用相应的测试集进行测试, 求得正类的 10个 recall(简记为 R)、precision(简记为 P)、F-value(简记为 F)值, 分别求得它们的平均值, 结果见表 3.

表 3 实验结果

**Table 3 Experimental results**

| Datasets | $K=6$ | | | | | | $K=9$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FLD | | | FLDDWSPS | | | FLD | | | FLDDWSPS | | |
| | R | P | F | R | P | F | R | P | F | R | P | F |
| Liv | 0.206 | 0.884 | 0.324 | 0.545 | 0.883 | 0.670 | 0.249 | 0.908 | 0.376 | 0.518 | 0.863 | 0.636 |
| Glass | 0.349 | 0.888 | 0.476 | 0.578 | 0.872 | 0.693 | 0.327 | 0.870 | 0.465 | 0.621 | 0.849 | 0.715 |
| Pima | 0.195 | 0.906 | 0.316 | 0.402 | 0.880 | 0.544 | 0.186 | 0.928 | 0.296 | 0.468 | 0.883 | 0.605 |
| Sonar | 0.081 | 0.859 | 0.142 | 0.332 | 0.866 | 0.471 | 0.068 | 0.954 | 0.121 | 0.348 | 0.806 | 0.484 |
| Wdbc | 0.363 | 1.0 | 0.50 | 0.664 | 0.983 | 0.790 | 0.578 | 1.0 | 0.719 | 0.667 | 0.970 | 0.781 |
| Diabetes | 0.231 | 0.907 | 0.355 | 0.330 | 0.854 | 0.466 | 0.279 | 0.933 | 0.415 | 0.462 | 0.890 | 0.602 |
| Vehicle | 0.282 | 0.988 | 0.428 | 0.747 | 0.932 | 0.826 | 0.311 | 0.99 | 0.465 | 0.661 | 0.890 | 0.748 |
| Thyroid | 0.648 | 0.988 | 0.776 | 0.862 | 0.852 | 0.853 | 0.714 | 1.0 | 0.823 | 0.931 | 0.86 | 0.891 |

从表 3可以看出, 与 FLD算法相比, 本文提出的算法能够明显提高单个正类的分类性能. 从评价准则可得知, F-value值能很好地反映少数类的整体分类性能, recall值能够较好地反映少数类的识别率. 而本文提出的算法使 8个数据集的 F-value和 recall都有较大程度的提高, 其中分类性能提高最多的是数据集 Vehicle, 当 $k=6$时它的正类的 recall提高了 46.5%, F-value提高了 39.8%, precision却只降低了 5.6%. 而分类性能提高最少的 Wdbc数据集当 $k=9$时的 recall也提高了 8.9%, F-value提高了 6.2%, precision仅降低了 3.0%. 对以上多数数据集来说, 在 F-value和 recall得到很大程度提高的同时, precision却略微有所降低. 这是由于对单个正类进行过抽样, 生成 $k$个合成样本, 对正类来说, 增加了一些有用的样本信息, 所以正类的 recall值和 F-value值能够有所提高; 而对负类来说, 生成的 $k$个合成样本中有可能其中某些样本位于两个类的实际分类界上, 对负类样本来说是噪音数据, 这样就有可能会把少数未标签的负类样本误分为正类, 导致正类的 precision略微有所降低.

## 4 结语

本文通过对加权 Fisher线性判别模型 (WFLD)进行深入分析, 发现对于正类只有一个样本的不平衡数据集, 该方法就不再适用, 于是提出一种解决办法: 对正类中仅有的一个样本进行抽样, 在负类中找出它的 $k$个近邻, 依次对每个近邻按照一定的规则在它与单个正类的连线上合成新的样本, 把合成的 $k$个样本添加到原始不平衡数据集中, 再对数据集用 WFLD模型进行训练. 实验表明, 本文提出的方法能较好地解决 WFLD模型所不适用的面向单个正例的不平衡分类问题, 使单个正例的识别率有了较明显的提高.

[参考文献] (References)


[1] Chan P K, Stolfo S J. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection[C] // Proc of the Fourth International Conference on Knowledge Discovery and Data Mining(KDD-98). New York, 1998: 164-168.

[2] Weiss G M, Hirsh H. Learning to predict rare events in event sequences[C] // Proc of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). New York, 1998: 359-363.

[3] Atiya A F. Bankruptcy prediction for credit risk using neural networks: a survey and new results[J]. IEEE Trans on Neural Networks, 2001, 12(4): 929-935.

[4] Kubat M, Holte R C, Matwin S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning, 1998, 30(2): 195-215.

[5] Maloof M A. Learning when data sets are imbalanced and when costs are unequal and unknown[C] // ICML-2003 Workshop on Learning From Imbalanced Data Sets II, 2003.

[6] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection[C]//Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Press, 1997. 179-186.

[7] Chawla N, Bowyer K, Hall L, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[8] 周荃, 王崇骏, 王珺, 等. PC4.5: 用于不均衡数据集的 C4.5改进算法[J]. 计算机辅助工程, 2006, 15(3): 23-26.
Zhou Quan, Wang Chongjun, Wang Jun, et al. PC4.5: improved C4.5 algorithm applied in imbalanced dataset[J]. Computer Aided Engineering, 2006, 15(3): 23-26. (in Chinese)

[9] 肖健华, 吴今培. 样本数目不对称时的 SVM 模型[J]. 计算机科学, 2003, 30(2): 165-167.
Xiao Jianhua, Wu Jinpei. SVM model with unequal sample number between classes[J]. Computer Science, 2003, 30(2): 165-167. (in Chinese)

[10] 谢纪刚, 裘正定. 非平衡数据集 Fisher线性判别模型[J]. 北京交通大学学报, 2006, 30(5): 15-18.
Xie Jigang, Qiu Zhengding. Fisher linear discriminant model with class imbalance[J]. Journal of Beijing Jiaotong University, 2006, 30(5): 15-18. (in Chinese)

[11] Chawla N, Lazarevic A, Hall L, et al. SMOTEBoost: improving prediction of the minority class in boosting[C]//7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Croatia: Cavtat-Dubrovnik, 2003: 107-119.

[12] 边肇棋, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2001.
Bian Zhaoqi, Zhang Xuegong. Pattern Recognition[M]. Beijing: Tsinghua University Press, 2001. (in Chinese)

[责任编辑: 严海琳]

(上接第 46页)

## [参考文献] (References)

[1] Jeannot M A, Cantwell F F. Solvent microextraction into a single drop[J]. Anal Chem, 1996, 68(13): 2236-2240.

[2] Pedersen-Bjergaard S, Rasmussen K E. Liquid-liquid-liquid microextraction for sample preparation of biological fluid prior to capillary electrophoresis[J]. Anal Chem, 1999, 71(14): 2650-2656.

[3] de Jager L S, Andrew A R J. Preliminary studies of a fast screening method for cocaine and cocaine metabolites in urine using hollow fiber membrane solvent microextraction (HFMSME)[J]. Analyst, 2001, 126(8): 1298-1303.

[4] Jiang X, Lee H K. Solvent bar microextraction[J]. Anal Chem, 2004, 76(18): 5591-5596.

[5] Halvorsen T G, Pedersen-Bjergaard S, Rasmussen K E. Liquid-phase microextraction and capillary electrophoresis of citalopram, an antidepressant drug[J]. J Chromatogr A, 2001, 909(1): 87-93.

[6] Zhu L Y, Ee K H, Zhao L M, et al. Analysis of phenoxy herbicides in bovine milk by means of liquid-liquid-liquid microextraction with a hollow-fiber membrane[J]. J Chromatogr A, 2002, 963(1/2): 335-343.

[责任编辑: 严海琳]