用户人品预测大赛--就是gan队--竞赛分享

•发布于 2016-03-24 14:14 •作者 DataCastle运营 (/user/5451) •65 次浏览 •来自 微额借款用户人品预测大赛 (/?tab=148)

参赛队队名：就是 gan

竞赛报告书

一、参赛作品概述

此次赛题的目标是对测试样本获得尽可能准确的评分。训练集 15000 条测试集 5000 条无标签样本 50000 条。训练集大类小类比 8.7:1。如何有效的挖掘数据中的类别分布信息，提高对不平衡数据的分类性能是本赛题亟待解决的问题。

我们的处理包含数据预处理、加入特征分析、构建新特征组合、Gradient Boosting（Xgboost）、半监督学习、PSO 暴力集成。我们的算法大体流程如下图所示：

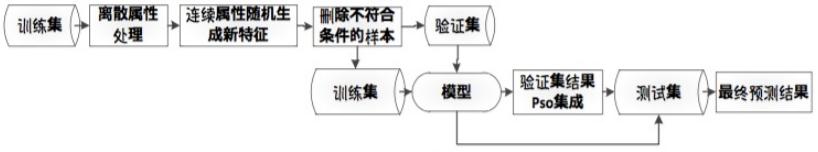


图 1. 算法大体流程图

（1）数据预处理

经过统计发现训练集中存在一些缺失值，为了提高分类精度，我们首先对数据进行预处理。

（2）特征组合

有效的特征是训练强分类器的关键，通过特征组合挖掘更多特征间的微弱关联关系。

（3）不平衡数据处理（过采样调整数据平衡度）

通过统计训练集中有 15000 条，大类小类比 8.7:1，不平衡度（多数类与少数类的比值）接近于 9，由于小类样本数量较少，在用传统机器学习方法进行分类时分类器会更倾向于大类，导致小类分类效果较差。因此我们结合邻域模型提出一种新的过采样方法，解决数据集不平衡的问题。

（4）半监督学习（对无标签数据打标）

本次比赛训练集有 15000 条，测试集有 5000 条无标签样本有 50000 条，有标签数据明显少于无标签数据为了提高分类器的精度，我们无标签数据进行学习，学习得到置信度高的不同层次的样本集，提高训练集的分类精度。

（5）模型选择（选择较优的分类器）

比赛中，我们尝试了很多分类器，例如，SVM,CART,C4.5,LR,RF 等，最后发现 xgboost 的效果要由于其他算法。因此，我们采用 xgboost 训练模型。

（6）模型集成（多分类器集成学习）

为了避免单模型的过拟合和大偏差，提高预测精度，我们采取 PSO 集成

关键技术：特征组合、不平衡数据处理、半监督学习、PSO 暴力集成

二、参赛作品技术路线

1.算法总思路：

整体思路是先对数据进行预处理，提高数据的质量，其次，调整数据的不平衡度，然后进行特征处理，特征组合；再次，半监督学习，得到不同层次的训练样本，采用 xgboost 建立不同的分类器，最后 PSO 暴力集成。

2.算法原理

2.1 特征预处理

(1)统计缺失样本，将缺失严重的训练样本进行删除；

（2）统计缺失属性，将缺失严重的训练属性进行删除；

（3）one-hot 编码，将离散型属性进行 one-hot 编码；

（4）删除无用属性，将值相同的属性删除；

2.2 特征组合

对于原始特征的连续值特征，随机抽取 30000 对（每对记为 x,y），计算 x*y、x^2+y^2、1/x+1/y，x/y 与标签列计算皮尔逊相关系数（Pearson correlation coefficient），取排名前 500 为新特征，将 null 值替换为-1。

2.3 不平衡数据处理

SMOTE 算法在合成小类样本时简便快捷，不会造成过拟合。但 SMOTE 采样后虽然扩大了小类的泛化空间，会同时缩小大类的泛化空间，以至于会降低对未知大类样本的预测准确率，即存在一定盲目性。我们可以通过一个例子从邻域粗糙集的角度来分析其采样不当造成新样本影响大类泛化空间的原因。

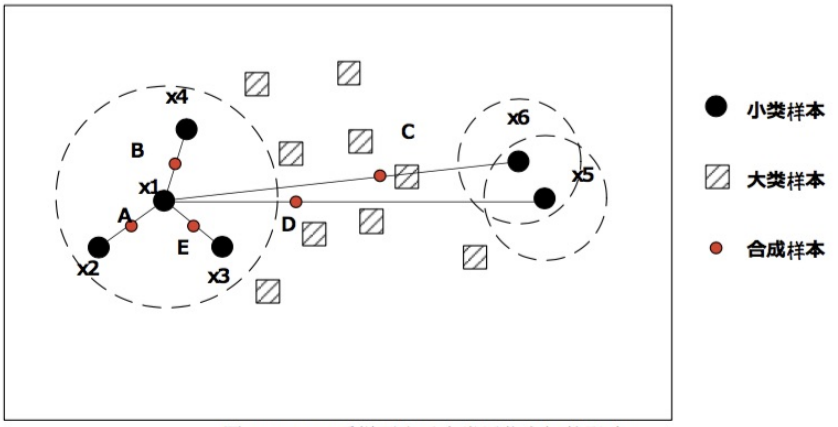


图 2　SMOTE 采样不当对大类泛化空间的影响

首先寻找小类样本x_1的同类 K 近邻（K=5），分别为{x2 □ x6}，其中x_1与x_5，x_6都为小类正域样本，且x_5，x_6距离x_1较远，若使用这些点作为近邻采样，则新生成的样本点 C,D 与多类样本混叠在一起，影响大类样本的正常分类。

在比赛中我们采用了一种新的过采样方法，利用样本集内部的分布特性，然后通过样本邻域内样本分布来确定样本性质，对于那些邻域范围内既还有大类样本又含有小类样本的小类样本，即边界域样本，需要进行过采样。最后针对合成的新样本，若其对大类正域无影响，即该合成样本不包含在任意正域大类样本的邻域内，保留该样本，否则进行二次采样。

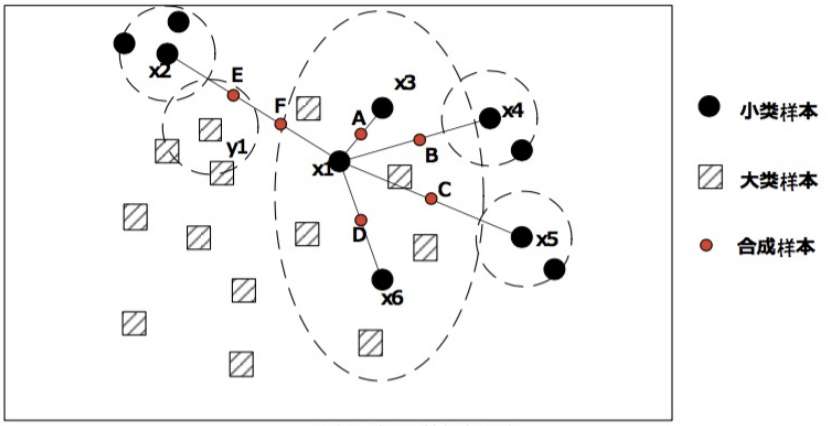


图 3 边界采样原理图

图 3 中，椭圆区域内的样本都属于邻域粗糙集的边界域，而椭圆区域外的样本则都属于决策正域。当我们需要对边界域小类样本x_1进行采样时，首先找到它的同类 K-近邻(k=5)，即图中的{x2 □ x6}，若选择了{x3 □ x6}其中的一个来合成新样本，可得到 A,B,C,D 四个合成样本其中的一个。可以看出，无论是哪个样本，都不会对决策正域内的大类样本产生影响，影响的只是边界域内的大类的泛化空间。当选择了x_2，若合成样本为 E，则可以看出$E \in \delta(y_1)$，则在预测大类样本的时候，会使样本y_1的规则覆盖范围变小，影响对未知大类样本的预测，导致错分。此时则需要进行二次采样，在x_1和x_2之间重新合成新样本，直到能够合成出样本 F，即$F \notin \delta(y_1)$（这里重采样次数 T 设置为 T=5）。

2.4 过采样步骤：

（1） 计算每个少数类样本的邻域半径。

$$\delta = \min\left(\square(x_i, s)\right) + w \times \text{range}\left(\square(x_i, s)\right), 0 \le w \le 1$$

这里$\min\left(\square(x_i, s)\right)$距离其最近的样本距离，$\text{range}\left(\square(x_i, s)\right)$表示在训练集中其距离的取值范围。

（2） 计算每个样本的邻域

$$\delta(x) = \{y \mid y \in U, \Delta(x, y) \le \delta\}$$

际问题中展示了其优越性。

PSO 算法的流程：

Step1：初始化一群粒子(群体规模为 m)，包括随机位置和速度；

Step2：评价每个粒子的适应度；

Step3：对每个粒子，将其适应值与其经过的最好位置（局部）pbest 作比较，如果较好，则将其作为当前的最好位置 pbest；

Step4：对每个粒子，将其适应值与群体所经过的最好位置（全局）gbest 作比较，如果较好，则将其作为当前的最好位置 gbest；

Step5：根据(2)、(3)式调整粒子速度和位置；

Step6：未达到结束条件则转 Step2，迭代终止条件为最大迭代次数 T。

在比赛中我们选取了 4 个模型进行了集成，分别拆分 1/10 的训练集(按比例随机拆分)作为验证集利用 PSO 计算三个模型的权重，并对三个模型的预测结果乘以相应的权重（对权重先归一化）相加得到最终的预测结果。其中，4 个模型下面图 4 介绍。

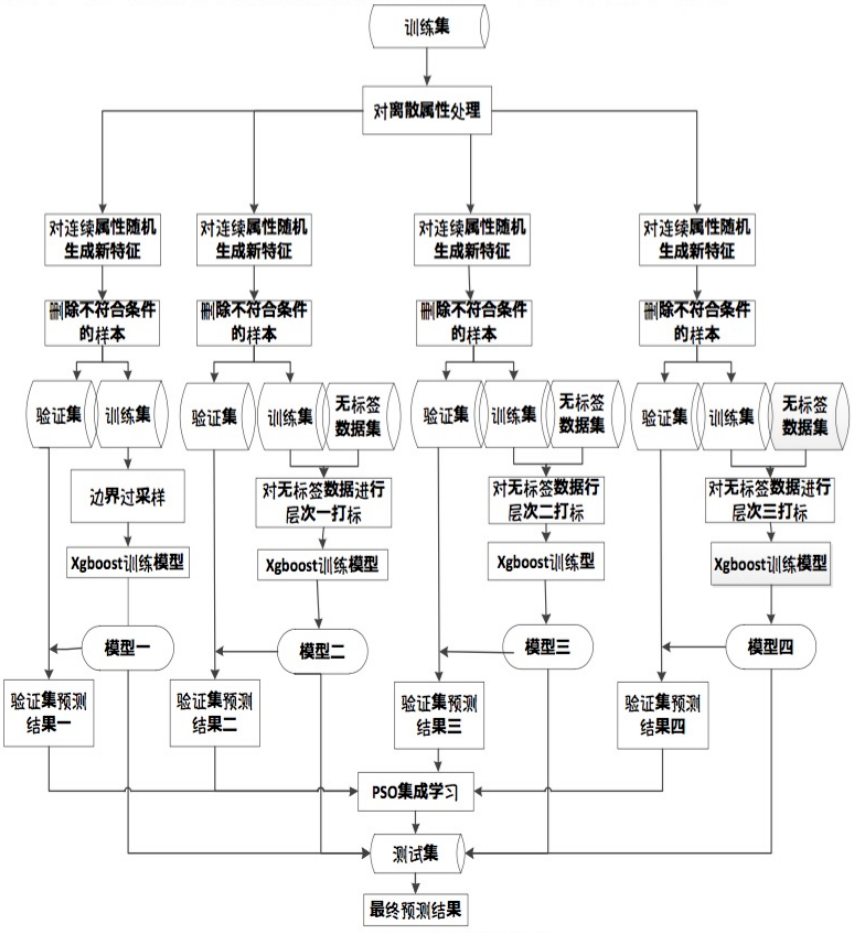


图 4 算法总流程图

比赛中单分类器（Xgboost）获得最好成绩 0.7218，加入新特征组合后最高获得成绩为 0.7252，解决数据不平衡性和采用半监督学习获得最高成绩 0.7304，集成获得最高成绩 0.7337。

三、作品总结

本算法先对数据进行预处理，提高数据的质量，其次，调整数据的不平衡度，然后进行特征处理，特征组合；再次，半监督学习，得到不同层次的训练样本，采用 xgboost 建立不同的分类器，最后 PSO 暴力集成。

1. 算法优势

（1）利用随机特征组合，挖掘潜在信息，提高分类器的性能；

（2）我们解决了数据的不平衡性，扩大了小类的泛化空间，有效避免了对大类泛化空间的影响，不会降低对未知大类样本的预测准确率。

（3）我们结合邻域模型，采取了一种自适应性的半监督学习方法，根据需求可以学习任意精度的样本集。

2. 可能的改进方向

模型复杂度高，需要大量的时间来训练模型。后面可以简化模型达到相当的分类精度。PSO 依赖于验证集的分布，当训练集的分布情况与验证集不同时，该方法的性能将会下降，应该采用更好的融合方式。

答辩PPT

微额借款用户人品预测大赛答辩

团队名：就是gan
线上排名：第三名

目录

1 赛题理解

2 特征与算法

3 多模型构造

4 总结

第一部分-赛题理解

1. 互联网金融领域无论是投资理财还是借贷放款，风险控制永远是业务的核心基础。
2. 在所有的互联网金融产品中，微额借款因其主要服务对象的特殊性，被公认为是风险最高的细分领域。
3. 本次比赛基于用户多维度行为数据的微额网贷风险预测。
4. 通过数据挖掘来分析“小额微贷”申请借款用户的信用状况，以分析其是否逾期。

数据分析

评价指标

- 本次比赛采用AUC来评价分类器的准确性

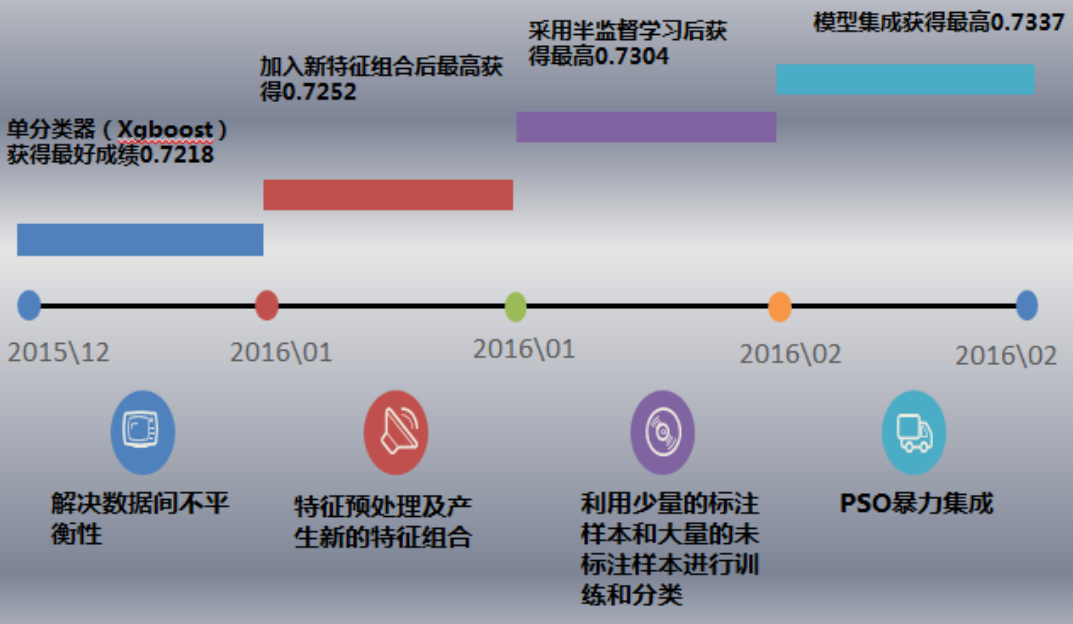
$$AUC = \frac{\sum_i S_i}{|p| * |N|}$$

- $|p|$为正样本个数，$|N|$为负样本个数，$|p|*|N|$为正负样本对的个数
- S_i 为第 个i 正负样本对的得分
- AUC 介[0,1]之间，越高越好。

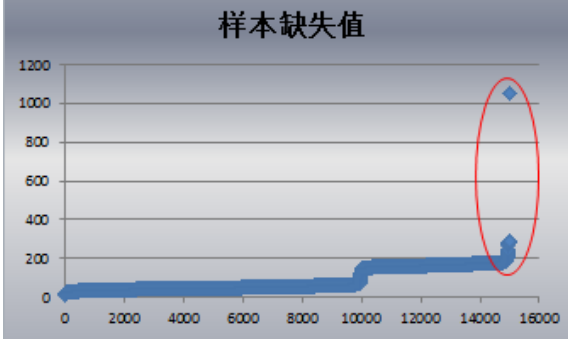
确立赛题目标

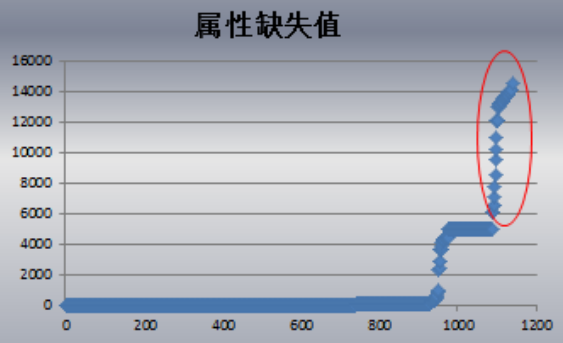
1. 经过数据预处理，特征选取，选取合适模型等处理提高AUC得分。
2. 通过改变数据的不平衡度，半监督学习等处理提高AUC得分。
3. 通过多模型融合的预测方法提高AUC得分。

第二部分-特征与算法



数据预处理

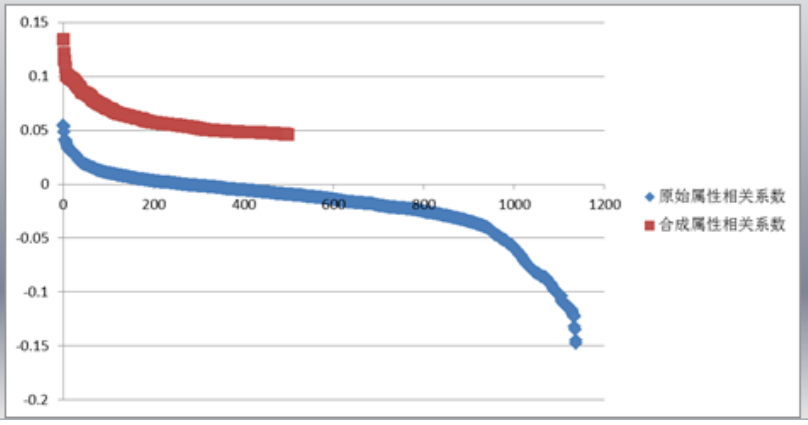




特征工程

1. 比赛原始特征分为数值型和离散型，对离散值属性进行one-hot编码

2. 对于原始特征的连续值特征，随机抽取30000对，计算多项式统计量，与标签列计算皮尔逊相关系数，取排名前500为新特征。



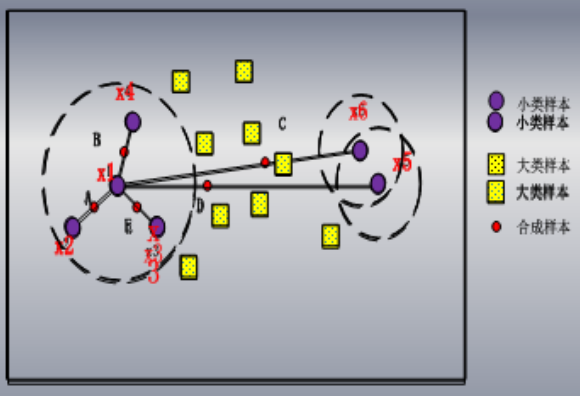
解决数据间的不平衡性

1. 由于小类样本数量较少，在用传统机器学习方法进行分类时分类器会更倾向于大类，导致小类分类效果较差。

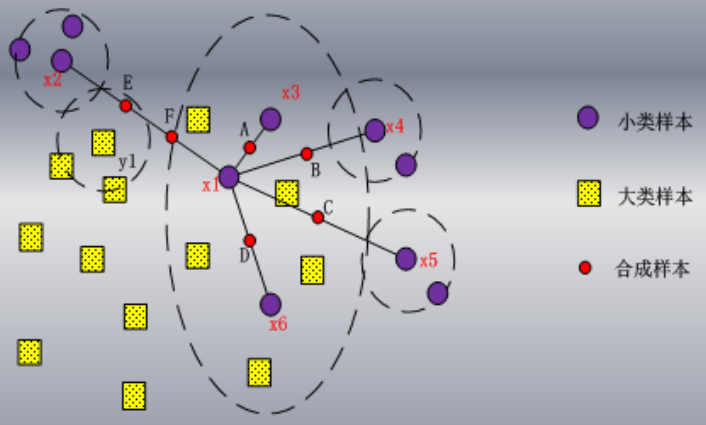
2. SMOTE采样后虽然扩大了小类的泛化空间，会同时缩小大类的泛化空间，以至于会降低对未知大类样本的预测准确率，即存在一定盲目性。

3. 我们可以通过一个例子从邻域粗糙集的角度来分析其采样不当造成新样本影响大类泛化空间的原因。

解决数据间的不平衡性



过采样



不平衡数据处理

计算每个样本的邻域

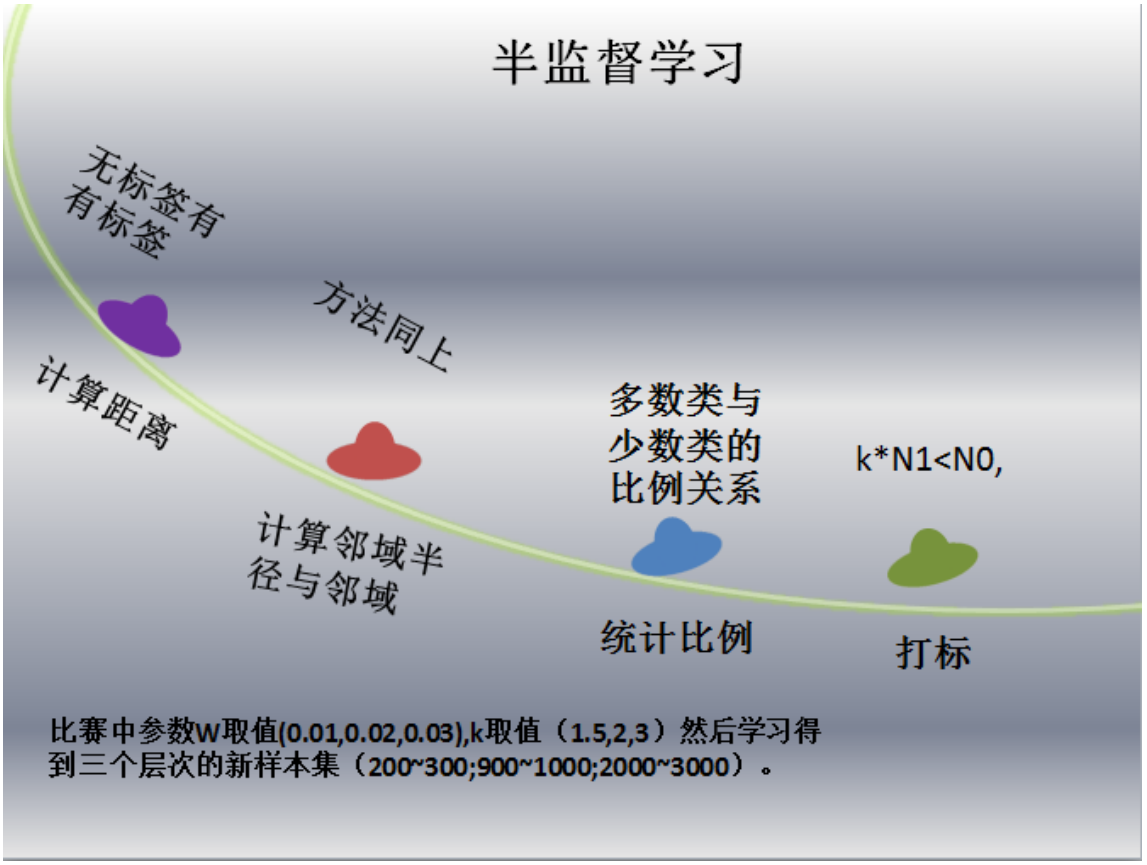
计算邻域半径

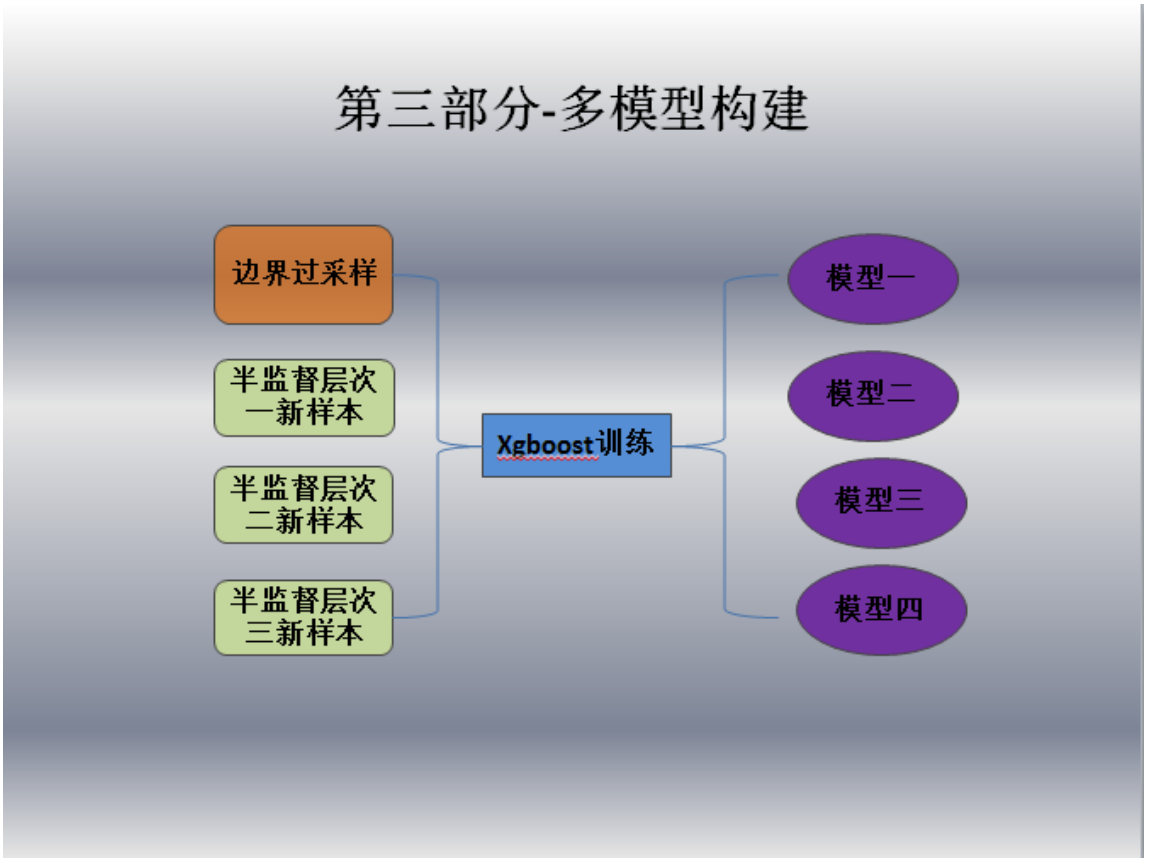
插值采样

$$\delta=\min\left(\Delta(x_i,s)\right)+w\times range\left(\Delta(x_i,s)\right),0\le w\le 1$$

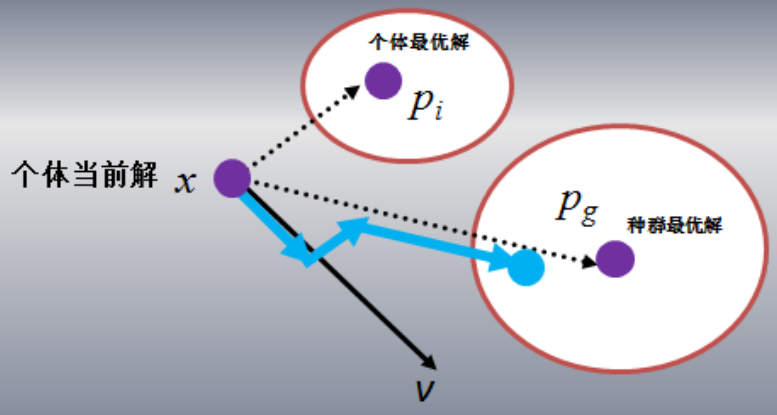
$$\delta(x)=\{y\mid y\in U,\Delta(x,y)\le\delta\}$$

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


半监督学习
无标签有有标签
计算距离
方法同上
计算邻域半径与邻域
多数类与少数类的比例关系
统计比例
k*N1<N0,
打标
比赛中参数W取值(0.01,0.02,0.03),k取值（1.5,2,3）然后学习得到三个层次的新样本集（200~300;900~1000;2000~3000）。


第三部分-多模型构建
边界过采样
半监督层次一新样本
半监督层次二新样本
半监督层次三新样本
Xgboost训练
模型一
模型二
模型三
模型四

PSO算法介绍

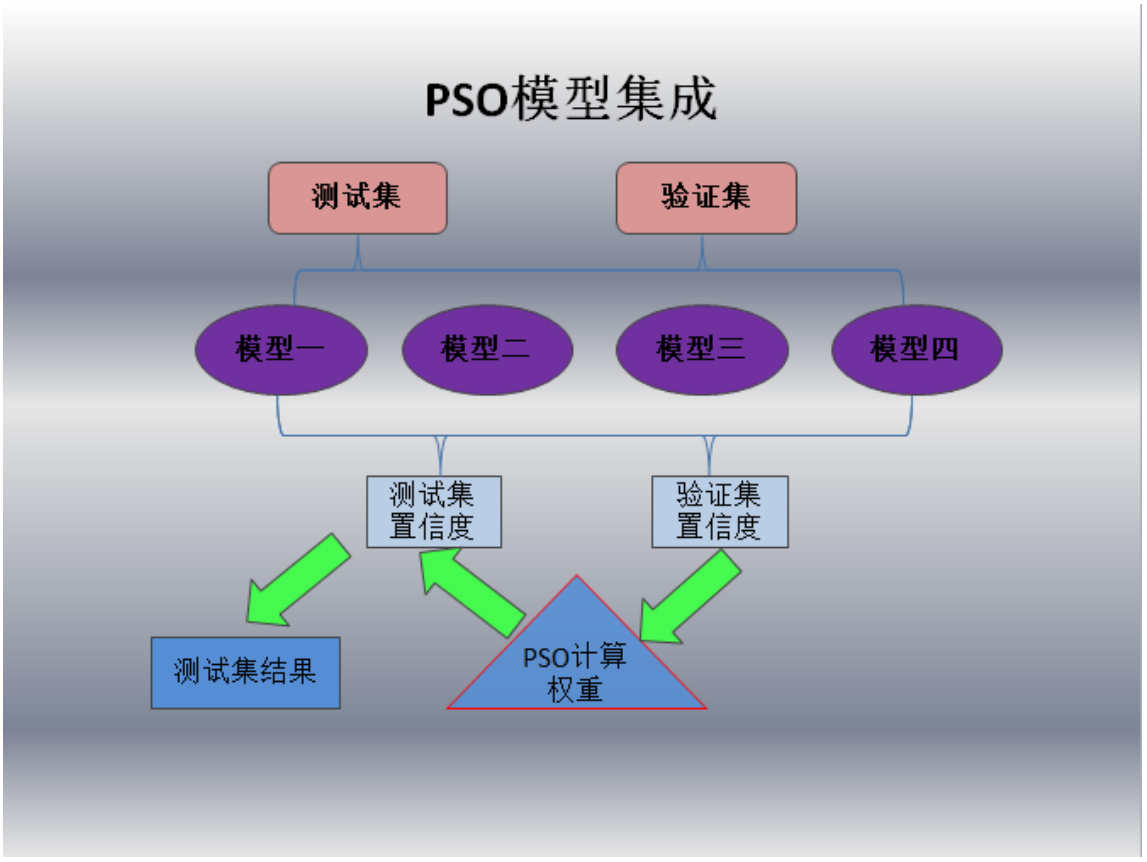


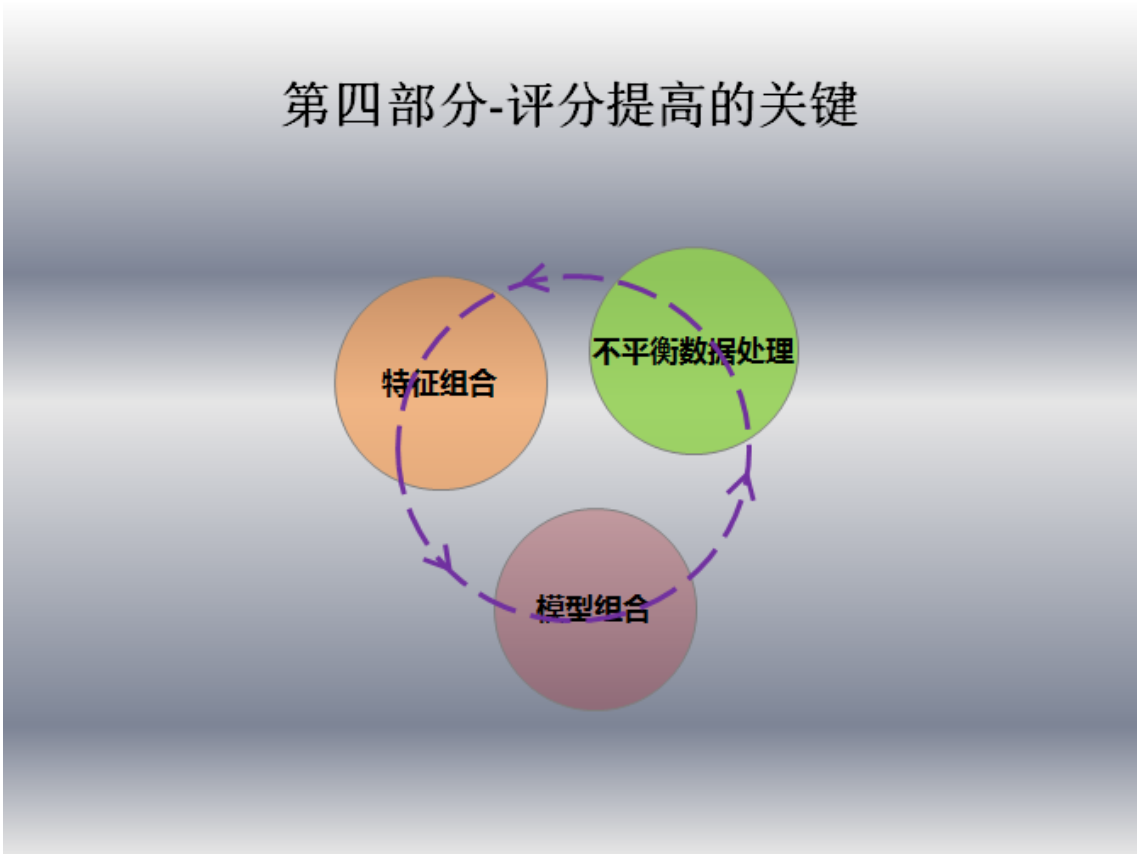
PSO算法介绍

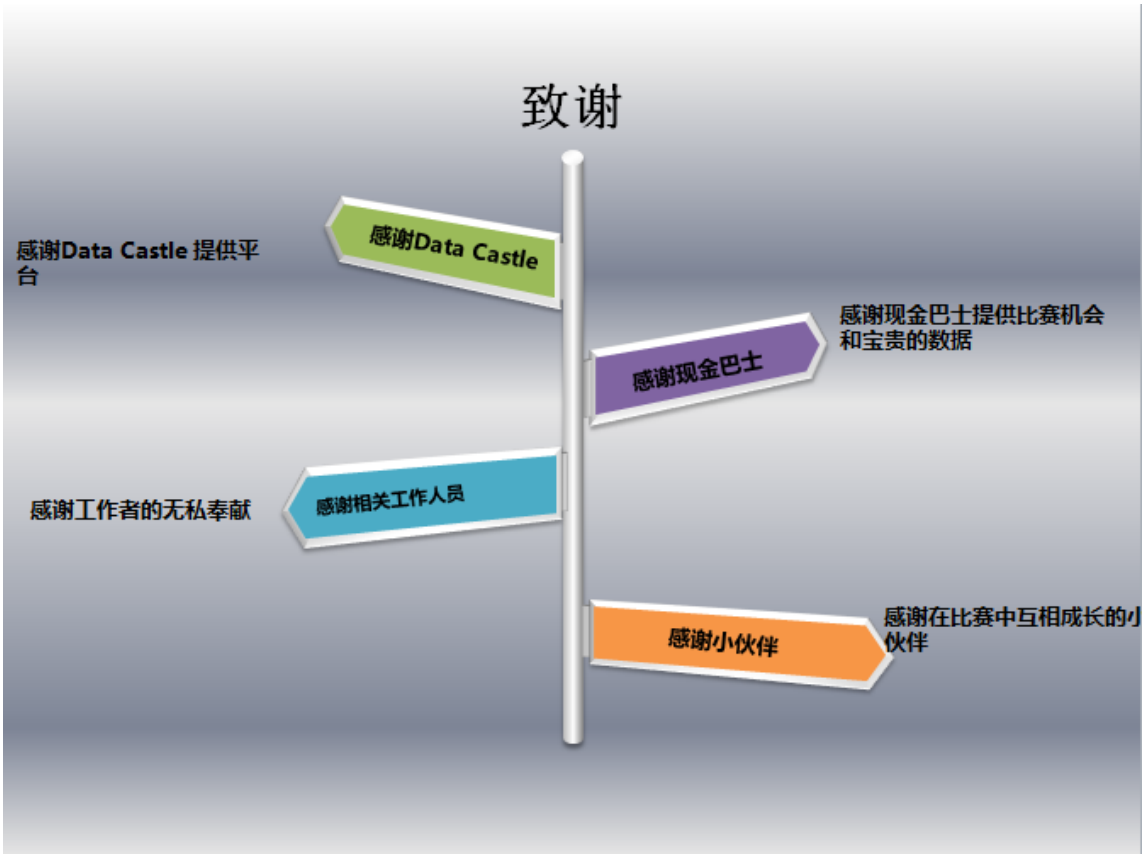
$$v' = w*v + c1*rand*(pbest - x) + c2*rand*(gbest - x)$$

$$x' = x + v'$$

pbest：每个个体曾经达到的最好位置
gbest：整个群体曾经达到的最好位置
w：惯性权重
c1，c2：学习因子


PSO模型集成
测试集
验证集
模型一
模型二
模型三
模型四
测试集
置信度
验证集
置信度
测试集结果
PSO计算
权重


第四部分-评分提高的关键
特征组合
不平衡数据处理
模型组合



0 回复

添加回复 **注:回复会奖励1点DC币，但被管理员删除回复，将扣除作者2DC币;可以使用@符号回复其他人**

回复

作者

(/user/5451) DataCastle运营 (/user/5451)

DC币: 428

无人回复话题

用户人品预测大赛--宝宝心里苦宝宝要说队--竞赛分享 (/topic/0f9a866d6d1e4f84bea2176db8237031.html)

用户人品预测大赛--wifi队--竞赛分享 (/topic/10078c08aecb44c18e1620686f0aa462.html)

用户人品预测大赛--火星人队--竞赛分享 (/topic/c5b1ce84f9ed42e7a933bbfcd2d6269a.html)

用户人品预测大赛--getmax队--竞赛分享 (/topic/cac927b5eff94193894f7dc588e1745a.html)

用户人品预测大赛--挖掘业务队--竞赛分享 (/topic/17416447cdab4bd5ad6a4bc00053f91e.html)

作者其他话题

用户人品预测大赛获奖团队分享 (/topic/58870500b2f84ddb9cbd4f6a45f180df.html)

用户人品预测大赛--宝宝心里苦宝宝要说队--竞赛分享 (/topic/0f9a866d6d1e4f84bea2176db8237031.html)

用户人品预测大赛--wifi队--竞赛分享 (/topic/10078c08aecb44c18e1620686f0aa462.html)

用户人品预测大赛--火星人队--竞赛分享 (/topic/c5b1ce84f9ed42e7a933bbfcd2d6269a.html)

用户人品预测大赛--getmax队--竞赛分享 (/topic/cac927b5eff94193894f7dc588e1745a.html)