两类非平衡大数据分类中的几个研究工作

翟俊海

Hebei University

March 30, 2019

Outlines

我们组近期的工作：

- 基于负类大数据样例约简的两类非平衡大数据分类；
- 聚类分析与集成学习相结合的两类非平衡大数据分类；
- 正类样例生成与集成学习相结合的两类非平衡大数据分类。

我们组近期的工作：

- 基于负类大数据样例约简的两类非平衡大数据分类；
- 聚类分析与集成学习相结合的两类非平衡大数据分类；
- 正类样例生成与集成学习相结合的两类非平衡大数据分类。

我们组近期的工作：

- 基于负类大数据样例约简的两类非平衡大数据分类；
- 聚类分析与集成学习相结合的两类非平衡大数据分类；
- 正类样例生成与集成学习相结合的两类非平衡大数据分类。

符号说明：

- 两类非平衡大数据集合$S = S^- \cup S^+$，其中S^-表示负类样例集合，它是一个大数据集，S^+表示正类样例集合。
- $S^- = S_1^- \cup S_2^- \cup \cdots \cup S_p^-$；
- $R_i^-(1 \le i \le p)$表示负类样例子集S_i^-的约简子集。
- $\boldsymbol{x}$表示给定的样例，$\boldsymbol{x}'$表示生成的样例。

符号说明：

- 两类非平衡大数据集合$S = S^- \cup S^+$，其中S^-表示负类样例集合，它是一个大数据集，S^+表示正类样例集合。
- $S^- = S_1^- \cup S_2^- \cup \cdots \cup S_p^-$；
- $R_i^-(1 \le i \le p)$表示负类样例子集S_i^-的约简子集。
- $\boldsymbol{x}$表示给定的样例，$\boldsymbol{x}'$表示生成的样例。

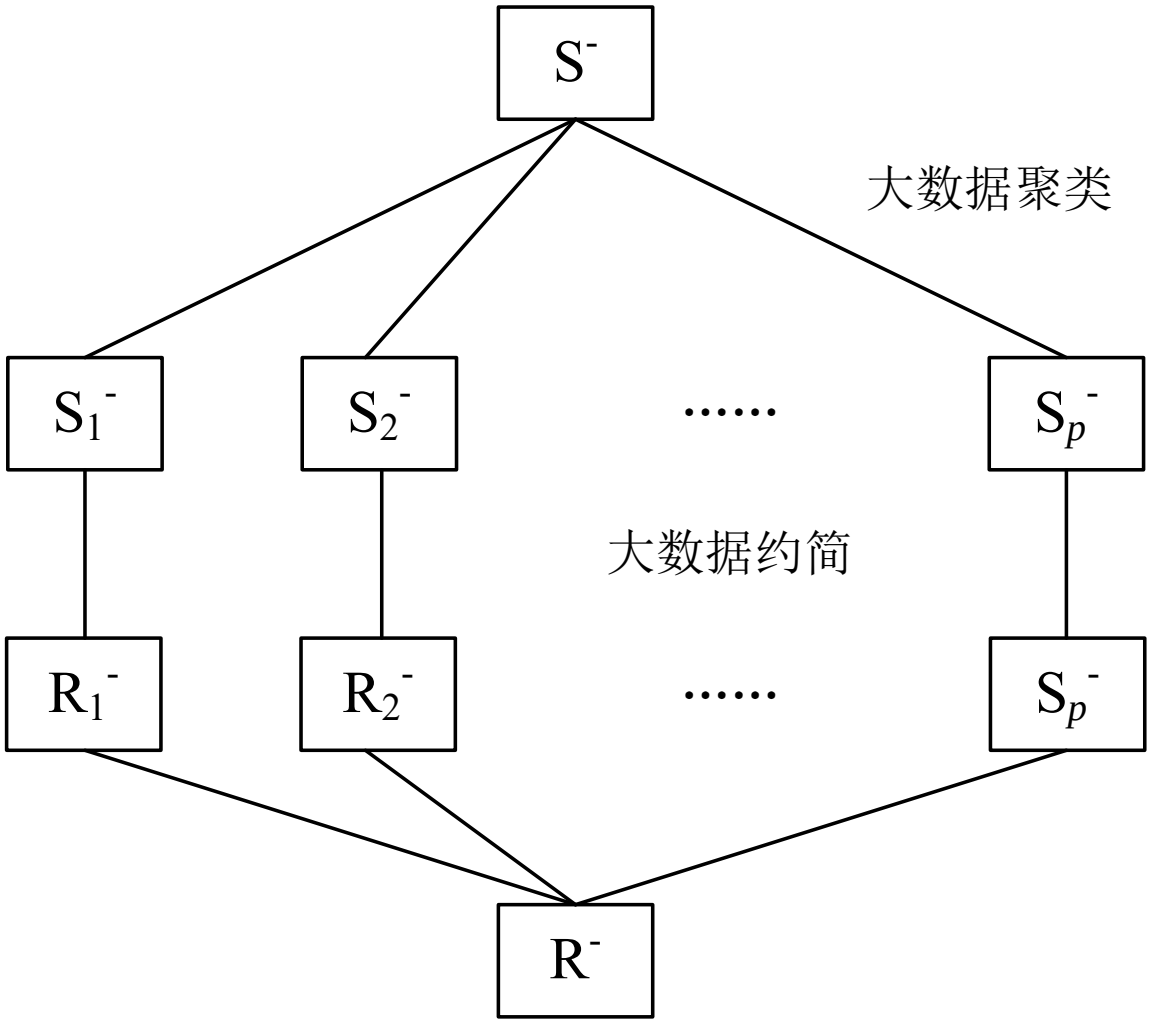
符号说明：

- 两类非平衡大数据集合$S = S^- \cup S^+$，其中S^-表示负类样例集合，它是一个大数据集，S^+表示正类样例集合。
- $S^- = S_1^- \cup S_2^- \cup \cdots \cup S_p^-$；
- $R_i^-(1 \leq i \leq p)$表示负类样例子集S_i^-的约简子集。
- x表示给定的样例，x'表示生成的样例。

符号说明：

- 两类非平衡大数据集合$S = S^- \cup S^+$，其中S^-表示负类样例集合，它是一个大数据集，S^+表示正类样例集合。
- $S^- = S_1^- \cup S_2^- \cup \cdots \cup S_p^-$；
- $R_i^-(1 \leq i \leq p)$表示负类样例子集S_i^-的约简子集。
- $\boldsymbol{x}$表示给定的样例，$\boldsymbol{x}'$表示生成的样例。

Outlines


S-
大数据聚类
S1-
S2-
……
Sp-
大数据约简
R1-
R2-
……
Sp-
R-

负类大数据约简后，可能会出现两种情况：

- 第一种情况：S^+和R^-的规模大致相同。针对这种情况，直接构造平衡的训练集；
- 第二种情况：S^+和R^-的规模是非平衡的。针对这种情况，首先对正类类样例集合S^+参照R^-进行上采样，然后构造平衡的训练集。也可以参照每一个负类约简子集$R_i^-(1 \leq i \leq p)$进行上采样，然后构造p平衡的训练集，训练p分类器，最后用集成的方法集成这些分类器。

负类大数据约简后，可能会出现两种情况：

- 第一种情况：S^+和R^-的规模大致相同。针对这种情况，直接构造平衡的训练集；
- 第二种情况：S^+和R^-的规模是非平衡的。针对这种情况，首先对正类类样例集合S^+参照R^-进行上采样，然后构造平衡的训练集。也可以参照每一个负类约简子集$R_i^-(1 \leq i \leq p)$进行上采样，然后构造p平衡的训练集，训练p分类器，最后用集成的方法集成这些分类器。

Outlines

基本思想：聚类分析与集成学习相结合的两类非平衡大数据分类分为两个阶段：

- 第一个阶段将负类大数据S^-自适应地聚类成p个簇（p个负类子集），即$S^- = S_1^- \cup S_2^- \cup \cdots \cup S_p^-$；
- 构造p个平衡的训练集，训练p个分类器，用模糊积分对它们进行集成，并用于未见数据的分类。

基本思想：聚类分析与集成学习相结合的两类非平衡大数据分类分为两个阶段：

- 第一个阶段将负类大数据S^-自适应地聚类成p个簇（p个负类子集），即$S^- = S_1^- \cup S_2^- \cup \cdots \cup S_p^-$；
- 构造p个平衡的训练集，训练p个分类器，用模糊积分对它们进行集成，并用于未见数据的分类。

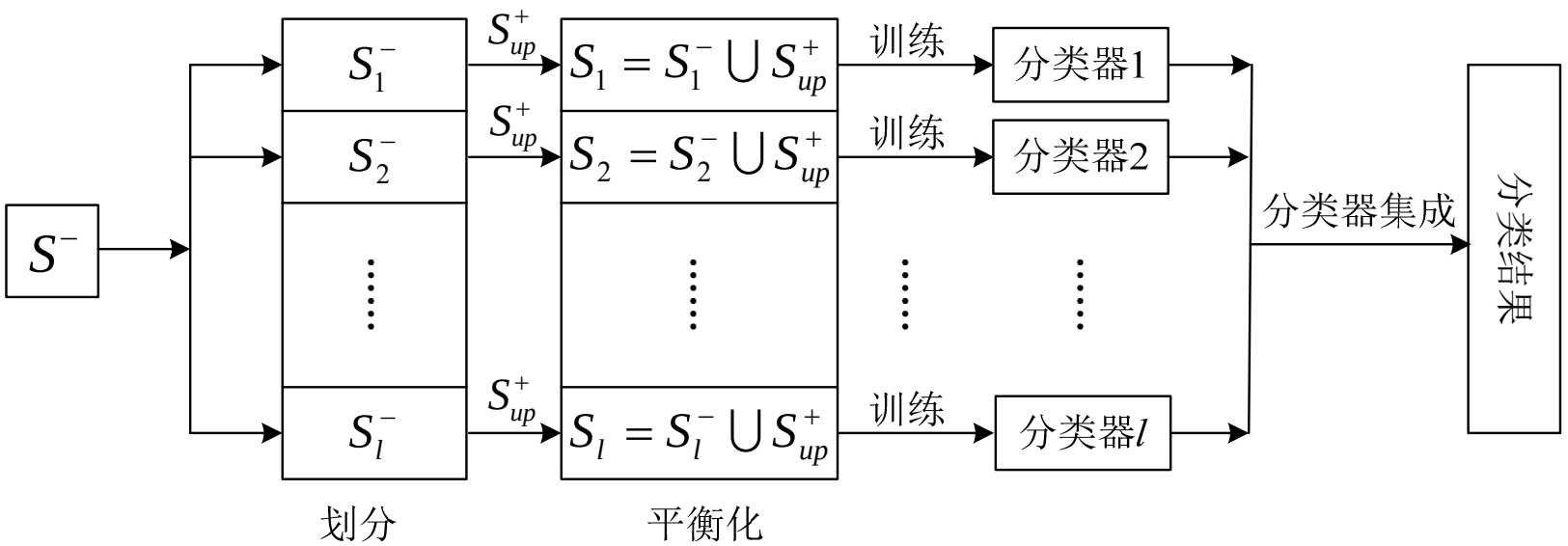
在正类样例集合S^+和p个负类子集构造平衡的训练集时，分两种情况：

- 第一种情况：S^+和$S_i^-(1 \leq i \leq p)$的规模大致相同。针对这种情况，直接构造p个平衡的训练集；
- 第二种情况：S^+和$S_i^-(1 \leq i \leq p)$的规模是非平衡的。针对这种情况，首先对正类类样例集合S^+参照每一个负类样例子集S_i^-进行上采样，然后构造p个平衡的训练集。

在正类样例集合S^+和p个负类子集构造平衡的训练集时，分两种情况：

- 第一种情况：S^+和$S_i^-(1 \leq i \leq p)$的规模大致相同。针对这种情况，直接构造p个平衡的训练集；
- 第二种情况：S^+和$S_i^-(1 \leq i \leq p)$的规模是非平衡的。针对这种情况，首先对正类类样例集合S^+参照每一个负类样例子集S_i^-进行上采样，然后构造p个平衡的训练集。

聚类分析与集成学习相结合的两类非平衡大数据分类技术路线图：



Outlines

正类样例生成的两种基本思路：

- 基于深度生成模型的正类样例上采样，上采样准则是类内散度最大化。
- 基于深度对抗扰动学习的正类样例上采样，上采样准则是对抗扰动学习最大化准则。

正类样例生成的两种基本思路：

- 基于深度生成模型的正类样例上采样，上采样准则是类内散度最大化。
- 基于深度对抗扰动学习的正类样例上采样，上采样准则是对抗扰动学习最大化准则。

类内散度最大化准则：

$$\max_{\boldsymbol{x}' \in S^{+}_{up}} \left\{ \frac{1}{|S^{+}|} \sum_{\boldsymbol{x} \in S^{+}} (\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^{T} + \frac{1}{|S^{+}_{up}|} \sum_{\boldsymbol{x}' \in S^{+}_{up}} (\boldsymbol{x}' - \boldsymbol{m}')(\boldsymbol{x}' - \boldsymbol{m}')^{T} \right\}$$
$$s.t. class(\boldsymbol{x}') = class(\boldsymbol{x}) \tag{1}$$

深度生成模型可采用生成对抗网络和分自动编码器。

对抗扰动学习最大化准则：

$$\begin{aligned} &\max_{\boldsymbol{r}\sim D}\{\boldsymbol{x}+\boldsymbol{r}\} \\ &s.t. class(\boldsymbol{x}) = class(\boldsymbol{x}+\boldsymbol{r}) \end{aligned} \tag{2}$$

Thanks!