Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(2), 177-184
Published Online May 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.132019

有序多类ROC超曲面下体积的快速无偏估计

吴海平，朱鸿斌，肖芷菁，徐维超*

广东工业大学自动化学院，广东 广州

收稿日期：2024年1月29日；录用日期：2024年4月9日；发布日期：2024年4月17日

摘　要

接收机工作特性曲线(ROC)分析在科学和工程领域中特别时在机器学习中处理二分类问题时被广泛应用。然而，在实际情况下，多分类问题常常出现。为解决这个问题，学者引入了多类ROC超曲面下的体积VUHS概念，尽管已有学者提出了连续样本下计算VUHS的快速算法，但对离散样本下的VUHS研究仍显不足。本文提出了一种新的方法：基于动态规划(DP)的VUHS快速无偏估计算法。该算法旨在提高计算效率并确保无偏性，可同时处理连续及离散母体样本下的问题。通过实验验证了该算法的无偏性和计算效率，证实了其在处理多分类问题中的有效性和可靠性。

关键词

机器学习，接收机工作特性曲线(ROC)，曲面下面积(AUC)，多分类，(超)曲面下体积(VUHS)

Fast and Unbiased Estimation of Volume under Ordered Multi-Class ROC Hyper-Surface (VUHS) with Discrete Measurements

Haiping Wu, Hongbin Zhu, Zhijing Xiao, Weichao Xu*

School of Automation, Guangdong University of Technology, Guangzhou Guangdong

Received: Jan. 29th, 2024; accepted: Apr. 9th, 2024; published: Apr. 17th, 2024

*通讯作者。

文章引用: 吴海平, 朱鸿斌, 肖芷菁, 徐维超. 有序多类 ROC 超曲面下体积的快速无偏估计[J]. 人工智能与机器人研究, 2024, 13(2): 177-184. DOI: 10.12677/airr.2024.132019

Abstract

Receiver Operating Characteristic (ROC) analysis is extensively utilized in scientific and engineering domains, particularly when dealing with binary classification problems in machine learning. However, multiclass classification issues frequently arise in practical scenarios. To tackle this issue, scholars have introduced the concept of the Volume under the Hypersurface of the multi-class ROC (VUHS); although fast algorithms for computing VUHS have been proposed under continuous sample distributions, research on VUHS for discrete sample cases remains insufficient. This paper presents a novel approach: a Fast and Unbiased Estimation Algorithm for VUHS based on Dynamic Programming (DP). This algorithm aims to enhance computational efficiency while ensuring unbiasedness, capable of addressing problems derived from both continuous and discrete parent populations. The experimental validation confirms the algorithm's unbiasedness and computational efficiency, substantiating its effectiveness and reliability in handling multiclass classification problems.

Keywords

Machine Learning, Receiver Operating Characteristic (ROC) Curve, Area under the Curve (AUC), Multiclass Classification, Volume under the (Hyper) Surface (VUHS)

1. 引言

接收机工作特性(Receiver Operating Characteristic，简称 ROC)分析起源于二战时期的雷达探测理论，最初主要用于评估雷达探测性能[1]。如今，这一工具已广泛应用于医学、心理学、生物信息学、信号处理以及机器学习等诸多科学与工程研究领域。上世纪 90 年代，学者们将 ROC 曲线引入至机器学习领域[2]，用以评判二类分类器的性能。自此以后，ROC 曲线在机器学习、计算机视觉等领域的算法评估与优化工作中发挥了重要作用。

ROC 分析是根据样本隶属类别的先验知识，可以根据不同的阈值设置绘制出假阳性率和真阳性率的二维曲线图，通过计算曲线下方的面积(AUC)来评价二元分类器的总体性能，首先它是非参数的，不需要预先知道样本的分布，同时具对数据中类别分布和误分类代价不敏感的优良特性，若正负分类比例发生变化，ROC 曲线不受影响[3]。

原来的 ROC 分析框架只适用于两类的情况，随着人工智能等领域的迅猛发展，研究者们的关注点已不再局限于两类问题，而是逐渐转向解决三类甚至多分类问题。Alonz 等人提出 ROC 曲面下体积 VUS，将 ROC 分析拓展到三类背景下[4]，LIU 等人提出了基于动态规划的 VUS 的快速算法[5] [6]。Nakas 等人对多类问题进行了研究，并提出了多类 ROC 超曲面下面积 VUHS 的无偏估计量[7]，ZHU 等人在这个基础上提出了连续样本下 VUHS 的快速算法[8]。尽管学者提出了 VUHS 概念以应对多类问题，并进行了一系列相关研究，但目前大部分研究主要集中在连续样本条件下对 VUHS 的估算上；相比之下，针对离散样本条件下 VUHS 的无偏估计研究则相对较少。同时现有对此类问题的算法存在计算复杂度高、偏差较大的问题。为此，本文提出了一种线性算法，用于对多个有序连续或离散测量的 VUHS 进行均值的无偏估计。

2. 利用动态规划计算 VUHS 的无偏估计量

2.1. VUHS 的概率解释

令 $X_1,\cdots,X_r$ 是 r 类问题中的 r 个随机变量。假设 $\left\{X_{1,i_1}\right\}_{i_1=1}^{n_1},\cdots,\left\{X_{1,i_r}\right\}_{i_r=1}^{n_r}$ 是分别取自累积分布函数为 $F_1\left(x_1\right),\cdots,F_r\left(x_r\right)$ 的离散分布的独立同分布样本集。如 Nakas 等人[7]的论文中所示，以下概率的线性组合

$$\theta_r = Pr\left(X_1 \le \cdots \le X_r\right) \tag{1}$$

可以解释为在单位 r 立方体内的 r 类 ROC 超曲面(VUHS)下的体积，当样本母体分布为连续时，式(1)自动退化成连续样本下的 VHUS 形式 $\theta_r = Pr\left(X_1 < \cdots < X_r\right)$。以三分类离散样本情况下为例，可获得以下概率公式：

$$\begin{aligned}\theta &= Pr\left(X_1 \le X_2 \le X_3\right)\\ &= Pr\left(X_1 < X_2 < X_3\right) + \frac{1}{2}Pr\left(X_1 = X_2 < X_3\right)\\ &\quad + \frac{1}{2}Pr\left(X_1 < X_2 = X_3\right) + \frac{1}{6}Pr\left(X_1 = X_2 = X_3\right)\end{aligned} \tag{2}$$

从式(2)中不难发现，当 $F_1\left(x_1\right)=F_2\left(x_2\right)=F_3\left(x_3\right)$ 时，$\theta = 1/3! = 1/6$，这说明统计模型采用了随机预测的策略，是一个随机选择分类器；而当 $\theta = 1$ 时，则意味着从左到右 X_1,X_2,X_3 是完全可分的。因此，可以从式(2)中得出，VUHS 可以表征多类样本的分离程度。

2.2. VUHS 的无偏估计量

为了方便起见，首先引入一些符号。设 $R=\{1,2,\cdots,r-1\}$ 为 $X_1,\cdots,X_{r-1}$ 的下标序列，Ω 为序列 R 的递增子序列的集合，其元素个数为 $S = C_{r-1}^0 + \cdots + C_{r-1}^{r-1} = 2^{r-1}$，$\omega_j$ 表示 Ω 中的第 j 个元素，使用 ω_j 标记有序序列 $X_1,\cdots,X_r$ 内等号开始的下标位置，设 ω_j 序列共有 k 个连续子序列，每个子序列的长度分别为 b_i

$$\begin{aligned}R &= \left\{1,2,\cdots,\underbrace{n_1,\cdots,n_1+b_1-1}_{b_1},\cdots,\underbrace{n_k,\cdots,n_k+b_k-1}_{b_k},\cdots,r-1\right\}\\ \omega_j &= \left\{\underbrace{n_1,\cdots,n_1+b_1-1}_{b_1},\cdots,\underbrace{n_k,\cdots,n_k+b_k-1}_{b_k}\right\}\end{aligned} \tag{3}$$

其中 $n_i \ge 1$，$n_i + b_i \le r$，$b_i \ge 1$。由式(1)可以构造出 VUHS 的非参数估计量，即：

$$\hat{\theta}_r = \frac{1}{n_1\cdots n_r}\sum_{j=1}^{2^{r-1}}\left[Weight_{\omega_j}\cdot\sum_{i_1=1}^{n_1}\cdots\sum_{i_r=1}^{n_r}I_{\omega_j}\left(X_{1,i_1},X_{2,i_2},\cdots,X_{r,i_r}\right)\right] \tag{4}$$

其中，$Weight_{\omega_j}$ 表示 ω 对应的事件的权重，根据 ω 序列内连续子序列区块的长度求得：

$$Weight_{\omega_j} = \begin{cases}1 & k=0\\ \prod_{i=1}^{k}\dfrac{1}{\left(b_i+1\right)!} & k>0\end{cases} \tag{5}$$

对于式(3)定义的标记序列，若满足

$$x_1 < \cdots < x_{n_1} = \cdots = x_{n_1+b_1} < \cdots < x_{n_k} = \cdots = x_{n_k+b_k} < x_r$$

则 $I_\omega(\bullet)$ 的值等于 1，否则为 0，例如，当 $r=3$ 时，$\Omega = \left\{\omega_1,\omega_2,\omega_3,\omega_4\right\} = \left\{\{\},\{1\},\{2\},\{1,2\}\right\}$，对应的

权重 $Weight_{\omega_j}$ 依次为 1，1/2，1/2，1/6。当满足 $X_{1,i_1} < X_{2,i_2} < X_{3,i_3}$ 时，$I_{\omega_1}(\bullet)$ 等于 1，否则为 0；当满足 $X_{1,i_1} = X_{2,i_2} < X_{3,i_3}$ 时，$I_{\omega_2}(\bullet)$ 等于 1，否则为 0；当符合 $X_{1,i_1} < X_{2,i_2} = X_{3,i_3}$ 时，$I_{\omega_3}(\bullet)$，为 1，否则为 0。当符合 $X_{1,i_1} = X_{2,i_2} = X_{3,i_3}$ 时，$I_{\omega_4}(\bullet)$ 等于 1，否则为 0。

2.3. VUHS 估计量的快速算法

直接基于式(4)计算 $\hat{\theta}_r$ 十分低效，例如，当所有的样本量相等时，即 $n_1 = \cdots = n_r = m$，则时间复杂度为 $O(m^r)$，然而我们可以采用动态规划的方法进行快速实现。首先需要对 $I(\cdot)$ 括号内的事件进行分类，每个标记序列 ω 对应一个事件，符合事件条件的事件数量为：

$$S = \sum_{i_1=1}^{n_1} \cdots \sum_{i_r=1}^{n_r} I_{\omega_j}\left(X_{1,i_1}, \cdots, X_{r,i_r}\right) \tag{6}$$

例如，当 $r=4$ 时，其全部事件如下表所示，结合式(4)和式(6)便可以计算出 4 类的 VUHS 的估计量 $\hat{\theta}$：(表 1)

$$\hat{\theta} = \frac{1}{n_1 n_2 n_3 n_4}\left(S_1 + \frac{1}{2}S_2 + \frac{1}{2}S_3 + \frac{1}{2}S_4 + \frac{1}{6}S_5 + \frac{1}{4}S_6 + \frac{1}{6}S_7 + \frac{1}{24}S_8\right) \tag{7}$$

Table 1. The quantities required fbr quickly estimating the Variance of the estimator of VUHS
表 1. 四类样本情况下计算 VUHS 估计值所需事件及对应参数

事件	标记序列 ω	满足(·)关系的事件数	事件权重
S_1	{}	$\varepsilon(x_4 > x_3 > x_2 > x_1)$	1
S_2	{1}	$\varepsilon(x_4 > x_3 > x_2 = x_1)$	1/2
S_3	{2}	$\varepsilon(x_4 > x_3 = x_2 > x_1)$	1/2
S_4	{3}	$\varepsilon(x_4 = x_3 > x_2 > x_1)$	1/2
S_5	{1, 2}	$\varepsilon(x_4 > x_3 = x_2 = x_1)$	1/6
S_6	{1, 3}	$\varepsilon(x_4 = x_3 > x_2 = x_1)$	1/4
S_7	{2, 3}	$\varepsilon(x_4 = x_3 = x_2 > x_1)$	1/6
S_8	{1, 2, 3}	$\varepsilon(x_4 = x_3 = x_2 = x_1)$	1/24

接下来将介绍如何通过动态规划获得所需要的基础事件的实际数量。

2.3.1. DP 计算矩阵

令 $Z_1, \cdots, Z_N$, $N = n_1 + n_2 + \cdots + n_r$ 是由集合 $X_1, \cdots, X_r$ 合并组成的联合序列。对这个序列进行升序排列，得到该序列的统计量：

$$\underbrace{Z_{(1)} = \cdots = Z_{(1)}}_{Block_1} < \cdots < \underbrace{Z_{(J)} = \cdots = Z_{(J)}\left(= Z_i\right)}_{Block_J} < \cdots < \underbrace{Z_{(K)} = \cdots = Z_{(K)}}_{Block_K} \tag{8}$$

假设区块 $Block_J$ 中的所有元素的值都等于 Z_i。令 $\delta_{1,i} \cdots \delta_{r,i}$ 为当 $i = 1, \cdots, K$ 时，样本集 $X_1, \cdots, X_r$ 中分别等于 Z_i 的元素的数量。那么对于 r 类向量集可得到 r 个计数向量，分别表示为 $C_{X_1} = \left[\delta_{1,1}, \cdots, \delta_{1,K}\right], \cdots, C_{X_r} = \left[\delta_{r,1}, \cdots, \delta_{r,K}\right]$，向量长度均为 K。将这些向量按序号自上而下堆叠，便得到了用于动态规划的计算矩阵。

2.3.2. 4 类问题

下面首先探讨当$r=4$情况下，标记序列$\omega_3=\{2\}$时，S_3的计算过程，进一步解释动态规划的计算结构。由上表可得：

$$\begin{aligned} S_3 &= \varepsilon\left(X_4 > X_3 = X_2 > X_1\right) \\ &= \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3} I\left(X_{4,i_3} > X_{3,i_2} = X_{2,i_2} > X_{1,i_1}\right) \\ &= \sum_{i_3=1}^{K}\sum_{i_2=1}^{i_3-1}\sum_{i_1=1}^{i_2-1} \delta_{1,i_1}\delta_{2,i_2}\delta_{3,i_2}\delta_{4,i_3} \end{aligned} \tag{9}$$

式(9)可以通过动态规划计算结构实现，首先将之前介绍的计数向量$C_{X_1},C_{X_2},C_{X_3},C_{X_4}$自下而上进行堆叠，构成计数矩阵$C_{4\times k}$，然后进一步把$C_{[3,1]},C_{[2,1]},C_{[1,2]}$初始化为 0，如图 1 所示：

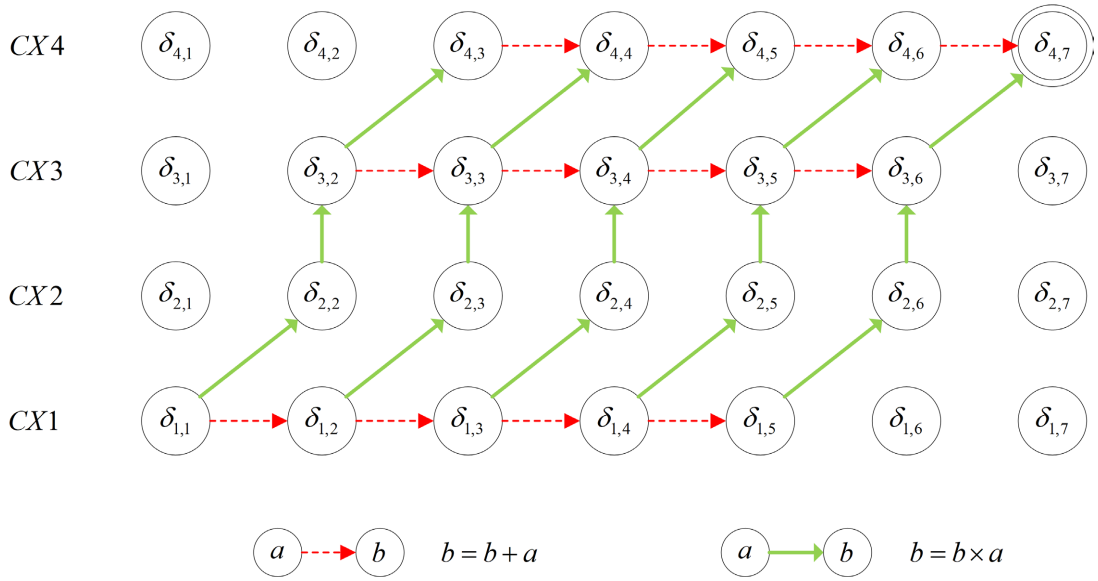

Figure 1. Diagram for computing S_3 defined in (9), where $K=7$ in order to facilitate visualization

图 1. 计算式(9)中定义的S_3流程图，其中$K=7$为了方便可视化

规划路径在线性时间($O(4K)$)内就可以从计数矩阵的左下角更新到右上角，其中的更新规则为：

$$C_{[I,J]} = \begin{cases} C_{[I,J]} + C_{[I,J-1]} & I=4, J\in[2,K-2] \\ C_{[I,J]} \cdot C_{[I+1,J-1]} & I=3, J\in[2,K-1] \\ C_{[I,J]} \cdot C_{[I+1,J]} + C_{[I,J-1]} & I=2, J\in[2,K-1] \\ C_{[I,J]} \cdot C_{[I+1,J-1]} + C_{[I,J-1]} & I=1, J\in[3,K] \end{cases} \tag{10}$$

最后当路径移动到右上角后，我们想要得到的S_3的值将会存储在矩阵元素$C_{[1,K]}$中。其他的事件均可以由该 DP 计算矩阵获得，进而由式(7)得到$\hat{\theta}$的值。

2.3.3. r 类问题

为了将动态规划方法推广到类问题，需要把以上算法的更新规则进行进一步的处理，首先将更新规则，分为累乘计算与累加计算，再根据标记序列与索引关系选择不同的更新规则，具体算法的伪代码如下：

Algorithm 1. Calculating the number of events S
算法 1. 计算各事件的个数 S

输入：大小为 $r\times K$ 的计数矩阵 C 和标注等号起始位集合的 ω
输出：通过 C 和 ω 计算出的 S

```
begin
    L ← ω 集合的元素个数
    StartIdx ← 1
    EndIdx ← K − r + 1 + L
    for i = r to 1 do
        if r − i + 1 ∈ ω then
            for j = StartIdx to EndIdx do
                C_{i,j} ← C_{i,j} × C_{i+1,j}
        else
            If i ≠ r then
                StartIdx ← StartIdx + 1
                EndIdx ← EndIdx + 1
                for j = StartIdx to EndIdx
                    C_{i,j} ← C_{i,j} × C_{i+1,j−1}
        if r − i + 2 ∈ ω then
            for j = StartIdx + 1 to EndIdx do
                C_{i,j} ← C_{i,j} + C_{i,j−1}
    S ← C_{1,K}
```

3. 实验及分析

为了验证本文所介绍方法的无偏性和快速性，首先将基于动态规划的 VUHS 算法(用 $\hat{\theta}_{DP}$ 表示)与基于式(4)的估计算法(用 $\hat{\theta}_{SLOW}$ 表示)以及基于图论(用 $\hat{\theta}_{GRA}$ 表示)的方法进行比较。我们生成了基于泊松分布的 r 个独立同分布连续样本集，$\left\{X_{k,i_k}\right\}_{i_k=1}^{n_k}\sim P(\lambda), k\in[1,r]$，在进行无偏性实验时，利用均值相对误差(*REM*)作为指标来评估算法的无偏性：

$$REM \triangleq \frac{E\left(\hat{\theta}_{\varsigma}-\hat{\theta}_{SLOW}\right)}{\hat{\theta}_{SLOW}} \tag{11}$$

其中，$\varsigma=\{DP,GRA\}$，实验设置了 $r=4, r=5$ 两组样本进行比较，为了使得到的结果更加稳定，实验中每个算法运行 1000 次再取其平均值。我们实验结果如下图所示。

从图 2 中我们可以看到，由于 $\hat{\theta}_{DP}$ 算法中需要对数据进行预处理，$\hat{\theta}_{DP}$ 的运算时间要略慢于同样是线性对数量级的 $\hat{\theta}_{GRA}$，但由于差别实在太小，可以认为这两个算法在计算效率上是一致的。反观 $\hat{\theta}_{SLOW}$ 在样本类别及样本量增加时，其运算时间均会出现飞速的增长，因为它的算法时间复杂度是 $O\left(n^r\right)$。

如图 3 所示，在无偏性上显然 $\hat{\theta}_{DP}$ 的表现优于 $\hat{\theta}_{GRA}$，$\hat{\theta}_{DP}$ 的均值相对误差完全拟合 $REM=0$ 的直线，而 $\hat{\theta}_{SLOW}$ 是 VUHS 的无偏估计，因此证明了 θ_{DP} 算法的无偏性。另一方面，$\hat{\theta}_{GRA}$ 的均值相对误差，随着样本数量的增加，在某个固定值附近波动，当样本类别增加时，误差会进一步增大，说明了 $\hat{\theta}_{GRA}$ 在处理离散样本时是有偏的。

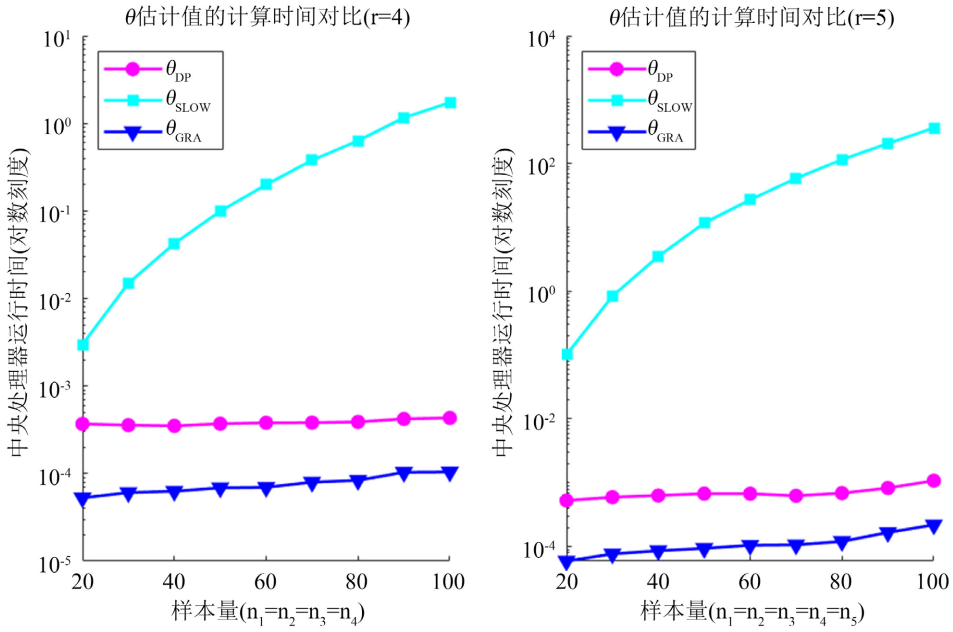

Figure 2. Comparison of CPU running time when calculating VUHS point estimation by three algorithms
图 2. 三种算法计算 VUHS 点估计时 CPU 运行时间对比结果

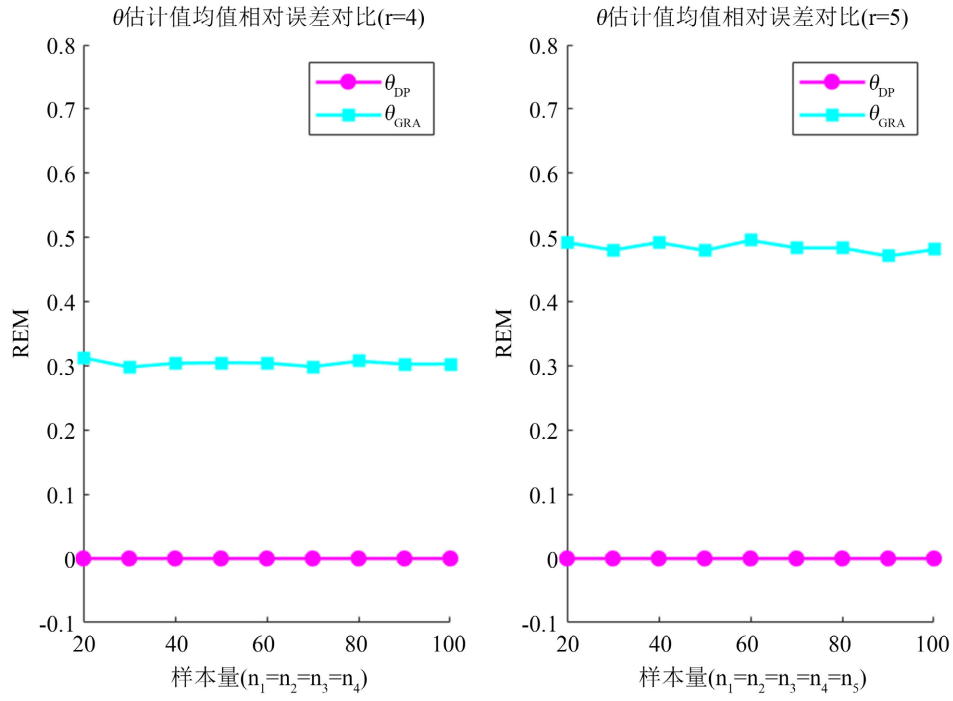

Figure 3. The unbiased comparison between the algorithm based on DP and the method based on GRA
图 3. 基于动态规划的估计算法与基于图论的方法的无偏性比较结果

4. 结论

本文基于动态规划提出了一种 VUHS 的快速无偏估计算法，首先对 VUHS 的估计值表示方法进行优

化，并通过建立动态计算矩阵，将算法时间复杂度降低至线性对数级，其次将 VUHS 点估计快速算法拓展到连续及离散样本下，并设计了蒙特卡洛实验进行了检验。实验结果表明，本文设计的 DP 算法矩阵可以有效的提升 VUHS 的计算效率，该算法相较于基于图论的方法有更好的无偏性，相较于 SOLW 算法有更好的快速性，特别是在应对类别多和样本量大的机器学习模型的应用背景下。因此，本文的方法在 VUHS 的研究上有一定的理论意义及技术价值。

基金项目

本文研究工作由国家自然科学基金项目(62171141, 61771148)资助。

参考文献

[1] Hanley, J.A. (1989) Receiver Operating Characteristic (ROC) Methodology: The State of the Art. *Critical Reviews in Diagnostic Imaging*, **29**, 307-335.

[2] Spackman, K.A. (1989) Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning. *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann Publishers, Burlington, MA, 160-163. https://doi.org/10.1016/B978-1-55860-036-2.50047-3

[3] 王彦光, 朱鸿斌, 徐维超. ROC 曲线及其分析方法综述[J]. 广东工业大学学报, 2021, 38(1): 46-53.

[4] Alonzo, T.A., Nakas, C.T., Yiannoutsos, C.T., *et al.* (2009) A Comparison of Tests for Restricted Orderings in the Three-Class Case. *Statistics in Medicine*, **28**, 1144-1158. https://doi.org/10.1002/sim.3536

[5] Liu, S., Sun, X., Xu, W., Zhang, Y. and Dai, J. (2018) Null Distribution of Volume Under Ordered Three-Class ROC Surface (VUS) with Continuous Measurements. *IEEE Signal Processing Letters*, **25**, 1855-1859. https://doi.org/10.1109/LSP.2018.2877930

[6] Liu, S., Zhu, H., Yi, K., Sun, X., Xu, W. and Wang, C. (2020) Fast and Unbiased Estimation of Volume Under Ordered Three-Class ROC Surface (VUS) with Continuous or Discrete Measurements. *IEEE Access*, **8**, 136206-136222. https://doi.org/10.1109/ACCESS.2020.3011159

[7] Nakas, C.T. and Yiannoutsos, C.T. (2004) Ordered Multiple-Class ROC Analysis with Continuous Measurements. *Statistics in Medicine*, **23**, 3437-3449. https://doi.org/10.1002/sim.1917

[8] Zhu, H., Liu, S., Xu, W., *et al.* (2022) Fast and Unbiased Estimation of Volume under the Ordered Multi-Class ROC Hyper-Surface with Continuous Measurements. *Digital Signal Processing*, **126**, Article ID: 103500. https://doi.org/10.1016/j.dsp.2022.103500