General Bounds for Maximum Mean Discrepancy Statistics

HE Yulin(何玉林), HUANG Defa(黄德发)
DAI Dexin(戴德鑫), HUANG Zhexue(黄哲学)

(College of Computer Science & Software Engineering, Shenzhen University,
Shenzhen 518060, China)

Abstract: The classical maximum mean discrepancy statistics, i.e., $\text{MMD}_b(F, X, Y)$ and $\text{MMD}_u^2(F, X, Y)$, to test whether two samples $X = \{x_1, x_2, \cdots, x_m\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$ are drawn from the different distributions $p$ and $q$. $\text{MMD}_b$ and $\text{MMD}_u^2$ are two very useful and effective statistics of which the bounds are derived based on the assumption of $m = n$. This paper relaxes this assumption and provides the general bounds for these two statistics $\text{MMD}_b$ and $\text{MMD}_u^2$. The derived results show that the traditional bounds derived in previous study are the special cases of our general bounds.

Key words: Two-sample test; Maximum mean discrepancy (MMD); Reproducing kernel Hilbert space (RKHS); McDiarmid’s inequality

CLC Number: O242.2
AMS(2000)Subject Classification: 60B05; 65C50

Document code: A
Article ID: 1001-9847(2021)02-0284-05

1. Two MMD Statistics $\text{MMD}_b$ and $\text{MMD}_u^2$

In order to determine how to test the difference between two distributions $p$ and $q$ based on the independent and identical samples $X = \{x_1, x_2, \cdots, x_m\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$ drawn from them, where $m$ and $n$ are the numbers of sample belonging to $X$ and $Y$, respectively. Gretton, et al.[1] designed two $\text{MMD}_b$ and $\text{MMD}_u^2$ based on the maximum mean discrepancy (MMD) principle, where $\mathcal{F}$ is a class of smooth functions in a characteristic reproducing kernel Hilbert space (RKHS)[2]. $\text{MMD}_b$ and $\text{MMD}_u^2$ are the generalizations of $L_2$ statistic[3]. The calculations of $\text{MMD}_b$ and $\text{MMD}_u^2$ were provided as follows in [1], respectively:

$$\text{MMD}_b(\mathcal{F}, X, Y) = \left[ \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) \right]^\frac{1}{2}$$

(1.1)
and

\[ \text{MMD}_d^2(\mathcal{F}, X, Y) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(y_i, y_j) \]

\[ - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j), \] (1.2)

where \( k(\cdot) \) is a RKHS kernel function.

2. Traditional Bounds of MMD\(_b\) and MMD\(_u\) When \( m = n \)

Assume \( 0 \leq k(\cdot, \cdot) \leq K \), where \( K \) is the upper bound of kernel function. Corollary 9 and Corollary 11 in [1] gave the bounds of MMD\(_b\) and MMD\(_u\) based on the assumption of \( m = n \).

**Corollary 1**[1] A hypothesis test of level \( \alpha \) for the null hypothesis has the acceptance region

\[ \text{MMD}_b(\mathcal{F}, X, Y) < \sqrt{\frac{2K}{m}} \left( 1 + \sqrt{2\log \frac{1}{\alpha}} \right). \] (2.1)

**Corollary 2**[1] A hypothesis test of level \( \alpha \) for the null hypothesis has the acceptance region

\[ \text{MMD}_u(\mathcal{F}, X, Y) < \frac{4K}{\sqrt{m}} \log \frac{1}{\alpha}. \] (2.2)

3. General Bounds of MMD\(_b\) and MMD\(_u\) When \( m \neq n \)

Eq. (2.1) and Eq. (2.2) provide the useful and effective statistics for testing \( p = q \). However, the above-mentioned bounds of MMD\(_b\) and MMD\(_u\) are derived based on the assumption \( m = n \). In this section, we relax this assumption and derive the more general bounds for MMD\(_b\) and MMD\(_u\).

**Corollary 3** When \( m \neq n \), a hypothesis test of level \( \alpha \) for the null hypothesis \( p = q \) has the acceptance region

\[ \text{MMD}_b(\mathcal{F}, X, Y) < \sqrt{\frac{K(m+n)}{mn}} \left( 1 + \sqrt{2\log \frac{1}{\alpha}} \right). \] (3.1)

**Proof** When \( p = q \) and \( m \neq n \), we get

\[
E_{X,Y} [\text{MMD}_b(\mathcal{F}, X, Y)] = E_{X,Y} \left[ \sup_{f \in \mathcal{F}} \left[ \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right] \right]
\]

\[
= E_{X,Y} \left[ \left\{ \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j) \right\}^{\frac{1}{2}} \right] \leq \left[ \frac{1}{m^2} E_x [k(x, x)] + \frac{1}{mn} E_x [k(x, y)] \right]^{\frac{1}{2}} \leq \sqrt{\frac{K(m+n)}{mn}}. \] (3.2)

According to Theorem 7 in [1], we let

\[ \exp \left[ -\frac{\varepsilon^2 mn}{2K(m+n)} \right] = \alpha. \] (3.3)
Combining Eq. (3.2) and Eq. (3.3), the McDiarmid’s inequality\cite{4}

\[
\Pr_{X,Y} [\text{MMD}_b (F, X, Y) - E_{X,Y} [\text{MMD}_b (F, X, Y)] > \varepsilon] < \exp \left[ \frac{-\varepsilon^2 mn}{2K (m + n)} \right]
\]

(3.4)
for \(m \neq n\) is yielded. In Eq. (3.3), we derive

\[
\varepsilon = \sqrt{\frac{2K (m + n)}{mn}} \log \alpha^{-1},
\]

(3.5)
and then the bound

\[
\text{MMD}_b (F, X, Y) < E_{X,Y} [\text{MMD}_b (F, X, Y)] + \varepsilon
\]

\[
< \sqrt{\frac{K (m + n)}{mn}} + \sqrt{\frac{2K (m + n)}{mn}} \log \alpha^{-1}
\]

(3.6)
is obtained. This completes the proof.

**Corollary 4** When \(m \neq n\), a hypothesis test of level \(\alpha\) for the null hypothesis \(p = q\) has the acceptance region

\[
\text{MMD}_u^2 (F, X, Y) < \sqrt{\frac{8K^2 (m + n)}{mn}} \log \alpha^{-1}.
\]

(3.7)

**Proof** According to the definition of \(\text{MMD}_u^2 (F, X, Y)\) in Eq. (1.2), we calculate

\[
\text{MMD}_u^2 (F, X, Y) = \frac{1}{m (m - 1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k (x_i, x_j) + \frac{1}{n (n - 1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k (y_i, y_j)
\]

\[
- \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k (x_i, y_j)
\]

\[
= \frac{1}{m} \left[ \sum_{i=1}^{m} \left[ \frac{1}{m - 1} \sum_{j \neq i}^{m} k (x_i, x_j) - \frac{1}{n} \sum_{j=1}^{n} k (x_i, y_j) \right] \right]
\]

\[
+ \frac{1}{n} \left[ \sum_{i=1}^{n} \left[ \frac{1}{n - 1} \sum_{j \neq i}^{n} k (y_i, y_j) - \frac{1}{m} \sum_{j=1}^{m} k (y_i, x_j) \right] \right]
\]

\[
= f (x_1, x_2, \ldots, x_m) + g (y_1, y_2, \ldots, y_n).
\]

(3.8)
Then, we derive

\[
\sup_{X \in \mathcal{X}^m, x \in \mathcal{X}} \left[ f (x_1, x_2, \ldots, x_m) - f (x_1, \ldots, x_{i-1}, \bar{x}, x_{i+1}, \ldots, x_m) \right]
\]

\[
= \sup_{X \in \mathcal{X}^m, x \in \mathcal{X}} \frac{1}{m} \left[ \left[ \frac{1}{m - 1} \sum_{j \neq i}^{m} k (x_i, x_j) - \frac{1}{m - 1} \sum_{j \neq i}^{m} k (\bar{x}, x_j) \right] \right]
\]

\[
+ \left[ \frac{1}{n} \sum_{j=1}^{n} k (x_i, y_j) - \frac{1}{n} \sum_{j=1}^{n} k (\bar{x}, y_j) \right]
\]

\[
\leq \frac{4K}{m}
\]

(3.9)
and

\[
\sup_{Y \in \mathcal{Y}^n, y \in \mathcal{Y}} \left[ g (y_1, y_2, \ldots, y_n) - g (y_1, \ldots, y_{i-1}, \bar{y}, y_{i+1}, \ldots, y_n) \right]
\]
\[ \sup_{Y \in \mathcal{Y}, y \in Y} \begin{bmatrix} \frac{1}{n-1} \sum_{j \neq i} k(y_i, y_j) - \frac{1}{n-1} \sum_{j \neq i} k(\bar{y}, y_j) \\ \frac{1}{m} \sum_{j=1}^{m} k(y_i, x_j) - \frac{1}{m} \sum_{j=1}^{m} k(\bar{y}, x_j) \end{bmatrix} \leq \frac{4K}{n}. \]  

(3.10)

Based on the McDiarmid’s inequality\[^4\], we get

\[
\Pr_{X,Y} \left[ \text{MMD}^2_u(F, X, Y) - \text{MMD}^2(F, p, q) > \varepsilon \right] < \exp \left[ -\frac{2\varepsilon^2}{m(4K/m)^2 + n(4K/n)^2} \right] = \exp \left[ -\frac{\varepsilon^2 mn}{8K^2(m+n)} \right].
\]  

(3.11)

Letting \( \exp \left[ -\frac{\varepsilon^2 mn}{8K^2(m+n)} \right] = \alpha \), we derive

\[
\varepsilon = \sqrt{8K^2(m+n) \log \frac{1}{\alpha}}
\]  

(3.12)

and then the bound of \( \text{MMD}^2_u(F, X, Y) \) is obtained for the null hypothesis \( p = q \). This completes the proof.

We can find that the bounds of \( \text{MMD}_b(F, X, Y) \) and \( \text{MMD}^2_u(F, X, Y) \) when \( m = n \) are the special cases of Eq. (3.1) and Eq. (3.7), i.e.,

\[
\text{MMD}_b(F, X, Y) < \sqrt{\frac{K(m+n)}{mn}} \left( 1 + \sqrt{2 \log \frac{1}{\alpha}} \right)
\]  

(3.13)

and

\[
\text{MMD}^2_u(F, X, Y) < \sqrt{\frac{8K^2(m+n)}{mn} \log \frac{1}{\alpha}}
\]  

(3.14)

4. Conclusions and Future Works

This paper relaxes the assumption of \( m = n \) for the classical bounds of two statistics \( \text{MMD}_b \) and \( \text{MMD}^2_u \) and derives the general bounds based on \( m \neq n \). The yielded results show that the classical bounds derived in [1] are the special cases of our general bounds. The random sample partition (RSP)\[^5\] is a new big data representation model. In future, we will use the MMD statistics with general bounds to determine RSP for big data management and analysis. In addition, we will evaluate the complexity of RSP data block based on these general bounds.

References:

最大均方差异统计量的一般界

何玉林, 黄德发, 戴德鑫, 黄哲学
(深圳大学计算机与软件学院, 广东 深圳 518060)

摘要: 经典的最大均方差异统计量 $MMD_b(F, X, Y)$ 和 $MMD_u^2(F, X, Y)$ 基于等量假设(即 $m = n$)来检验两组样本 $X = \{x_1, x_2, \cdots, x_m\}$ 和 $Y = \{y_1, y_2, \cdots, y_n\}$ 是否来自不同的分布. 本文对样本等量假设进行了放松, 推广了经典的最大均方差异统计量界, 推导出当 $m \neq n$ 时统计量 $MMD_b(F, X, Y)$ 和 $MMD_u^2(F, X, Y)$ 的一般界. 结果表明经典的最大均方差异统计量界是本文推导的最大均方差异统计量一般界的特例.

关键词: 双样本检验; 最大均方差差异; 再生核希尔伯特空间; 马克迪尔米德不等式