Summary

Memo

Contents

1 Introduction

* 1. Restatement
  2. Our works

1. Assumptions and Notations
   1. Assumptions
   2. Notations
2. Model Construction and Analysis
   1. Momentum model for local performance analysis
      1. The calculation method of momentum
      2. Model testing
   2. Momentum-Score correlation test
      1. Visual observation
      2. Pearson correlation coefficient

3.3 Momentum-Swings prediction model

3.4 Promotion and universality test

4 Strengths and Weaknesses

4.1 Strengths

4.2 Weaknesses

5 Stability Test

6 Conclusion

References

Appendices for Code and Data

1 Introduction

1.1 Restatement

2023年温网男子决赛可谓精彩绝伦，两位选手不相上下的比分更是令人出乎意料，本来占据优势的德约科维奇出现了少见的波动，导致最终输掉了比赛，结束了自己的传奇表现。面对这种本不该发生的特殊情况，存在以下几个问题有待我们解决：

1. 描述比赛过程，确定在某一特定时间内某一球员的表现。
2. 反驳“动量并不能推动球员的成功，二者关系是随机的“言论。
3. 预测比赛中的波动，研究比赛何时开始倾向某一特定选手，并给出球员比赛的建议。
4. 测试预测的稳定性和推广性，并给教练写一封关于让球员做好应对复杂事件心理准备的信件。

1.2 Our works

1. 建立分析模型分析整场比赛的数据，得出选手双方的水平。
2. 量化“动量“概念，以此测算比赛中可能出现的波动情况，综合分析其与选手表现的联系。
3. 根据数据预测比赛开始倾斜的时间，并用更多比赛测试模型稳定性。
4. 将模型推广至更多比赛（如女子比赛）甚至其他比赛（如乒乓球，象棋等）。

2 Assumptions and Notations

Assumptions

1. 观众对球员的发挥不产生影响
2. 球场内自然环境对球员的发挥不产生影响
3. 球员身体状态良好
4. 比赛过程中没有任何突发事件

Notations

Ms 总动量

Mb 基础动量

Mc 变化动量

Sn 分局比

Wn胜局

Se 先手

Sg 分差

v 球速

Sp 每局得分

Wb 制胜球

Ac 制胜发球

Df 两次发球失误

Ue 非受迫性失误

3 Model Construction and Analysis

在我们的数据预处理步骤中，对异常值的处理尤为关键。我们选取所有数值类型的字段，计算出每个字段的均值和标准差。然后以三倍的标准差距离均值作为界定异常值的门槛。一旦识别出异常值，我们将它们替换为NA值，以避免数据损失。

尽管这造成了新的缺失值，但我们通过前向和后向填充策略来解决这一问题。这种策略在时序数据中常用，可以保持数据的整体趋势。这样处理后，我们成功地纠正了数据中的异常值，有助于提高模型的准确性。

3.1 局内表现分析动量模型

为了更好的分析某一时刻某一选手的“势头”，我们引入了一个名词“动量”，其字典含义是“strength or force gained by motion or by a series of events.”【1】，而在体育运动中，动量更多表现为运动员在比赛中充满信心和力量。显而易见的是，动量和选手的表现是息息相关的，且通常情况下是正相关，我们决定用动量来确定选手的表现。然而，这种动量实际上很难测量，因为它往往是瞬时的和心理上的。因此我们给出了一种模拟方法。

3.1.1 动量的计算方法

众所周知，进行一局比赛不仅是对比赛选手实力的考察，更是对选手心理素质的巨大考验。

因此选手的实力和心理都应当成为动量的一部分，相应的，动量也存在一些变化情况，比如当选手打出了制胜球，或者出现较大失误的时候，这不仅会影响心理，同时也可能对选手的实力造成某些影响，因此我们把这种情况归类为变化动量，而原来的动量称为基础动量。总动量是这二者的和。

Ms=Mb+Mc

基础动量

心理层面上，首先，当一位选手在经过较少对局就获得较多分数，或在经过较多对局获得较少分数的时候，他的心理必然会发生巨大变化，因此我们引入分局比的定义，可以一定程度上显示分数与局数的关系，侧面反映选手的一部分心理。其次，抛开分数来看，选手的胜利局数也会影响他的心理，使其更加渴望胜利，动量由此变化。再者，如果两位选手的分差较大，同样会使双方选手产生危机感。最后，如果一位选手是先手发球，他的自信一定高于另一位后手的选手。以上四个因素各占25%，共同构成了基础动量的心理部分，我们给这个部分赋予了α=45%的权重。

实力层面上，首先，最简单判断实力的方法是球员的每局得分，得分高者实力必然强。其次，如果选手能打出更多制胜球，也能表明其实力水平。最后，比较隐蔽的一点是，如果一位球员的实力更高，一般情况下他的平均球速会比其他人更高，这表明他对于网球和球拍有更高好的掌握能力。（配图）

这三个方面各占33.3%，共同构成了基础动量的实力部分，我们给它赋予了β=55%的权重。

Mb=α(Sn+Wn+Sp+Se)\*0.25 +β(v+Sp+Wb)\*0.333

α=45% β=55%

变化动量

比赛场上的情况不是一成不变的，有些时候选手的表现会在心理和实力的双重作用下产生一些变化，我们选取了其中具有代表性的几个方面，如制胜发球，两次发球失误和非受迫性失误，这些情况很难用选手的心理或实力解释，因此算作变化动量单独考虑。我们假设三者的初始值均为0，若发生一次则加2，表达式为：

Mc=Ac-Fb-Ue

通过以上计算，可以得到选手的总动量

Ms= α(Sn+Wn+Sp+Se)\*0.25 +β(v+Sp+Wb)\*0.333+(Ac-Fb-Ue)

以上数据综合为动量，基本体现了该选手的局内表现。

3.1.2 模型测试

我们举ID为1301的比赛作为本模型的例子，二位选手分别是Carlos Alcaraz和Nicolas Jarry。比赛过程中的动量如图所示（配图）

二者比赛过程中不同时段的表现情况一目了然，其中Alcaraz的发挥波动较大，但最高水平更高，二者反复争夺上风，最终Alcaraz略胜一筹拿下比赛胜利。

3.2 动量-分数相关性检验

一位教练不相信动量对选手的推动作用，反之，他认为波动和选手能否成功是完全随机的。这当然是完全错误的观点，我们给出了反驳的理由。

为了方便，我们仍然举ID为1301的比赛为例子。下图为3.1中的动量图附加一条分差折线，以便更容易的分析局势。（配图）

3.2.1 肉眼观察

放眼于两位选手的动量折线相差较大的地方，显然，对应的分差折线也有很大的起伏变化，如时间为50，200两处，人眼和直觉就可以告诉我们动量和比赛的局势是有很强的关联的，同时，动量的变化自然也是产生波动的必要条件之一。因此可以说，波动和选手能否成功是密切相关的。

3.2.2 Pearson correlation coefficient

我们知道，检验两组数据是否相关最简单的办法就是Pearson correlation coefficient

【2】。我们使用的是其中的[Permutation tests](https://en.wikipedia.org/wiki/Permutation_test)的P-values，即通过P-values判断数据相关性。我们利用Python的数学分析库计算了P-values，生成了下图，在上图的基础上添加了Pearson correlation coefficient的相关内容。（配图）

Pearson correlation coefficient比肉眼观察更加有力的反驳了教练的断言，由此可以明白，我们的模型和推断是完全正确的。

3.3 动量-波动预测模型

由于动量具有不稳定性，容易受到各种因素的影响，并且其对比赛有着至关重要的决定性作用，教练们想知道比赛局势在何时会出现起伏甚至逆转。由此我们设计了动量-波动预测模型。

5 Stability Test

我们采用了LSTM（Long Short-Term Memory）【3】测试来量化模型的稳定性。LSTM是一种长短期记忆网络，是一种特殊的RNN（循环神经网络）。与传统的RNN相比，LSTM更加适用于处理和预测时间序列中间隔较长的重要事件。标准LSTM结构包括了记忆细胞、输入门、输出门和遗忘门这四个部分。

* 记忆细胞（memory cell）: 它是 LSTM 的核心，负责保存重要的信息，并将这些信息传递给后面的网络层。
* 输入门（input gate）: 决定了当前输入信息是否写入记忆细胞，也就是说，能够控制输入信息对记忆细胞的影响。
* 遗忘门（forget gate）: 决定了记忆细胞中的信息是否被遗忘，也就是说，能够控制记忆细胞中保存的信息会不会消失。
* 输出门（output gate）: 决定了记忆细胞中的信息是否输出，也就是说，能够控制记忆细胞中保存的信息会不会对后面的网络层造成影响。

而我们则采用了简化后的GRU算法,去掉了输入门，综合了输入门和遗忘门的操作在重置门中完成。（配图）【4】

测试共100组数据，结果显示于（配图）

实际准确值与预测准确值比较接近，loss保持在较低水平，该模型成功通过稳定性测试。

1. https://www.merriam-webster.com/dictionary/momentum
2. https://en.wikipedia.org/wiki/Pearson\_correlation\_coefficient
3. Amazing Adoo，“Basic knowledge of LSTM，”https://zhuanlan.zhihu.com/p/598124702
4. LSTM, <https://bkimg.cdn.bcebos.com/pic/14ce36d3d539b600564bd957e550352ac75cb74d?x-bce-process=image/format,f_auto/resize,m_lfit,limit_1,h_506>