基于$k-$means半监督聚类集成模型确定亚盘投资策略

中国科学院数学与系统科学研究院

研究背景及意义

背景

足球博彩是指针对尚未有结果的足球比赛进行猜测结果，并投注金钱赌输赢的合法或非法的赌博行为。

在足球博彩领域，对盘口的预测一直是一个热门话题，盘口某种程度上显示了庄家对比赛结果的预测。

意义

对盘口趋势做出合理地预测，可为其投资及管理等提供科学的决策参考。

亚盘与大小球盘

亚盘

亚洲盘口简称亚盘，它通过让球盘口使两个实力相差悬殊的球队能在一个相对公平的平台上比赛，拉平了交战双方客观上的实力差距，使得比赛双方各有50%的获胜几率。

亚盘由交战球队、交战双方的获胜赔率和让球数（即盘口）这三部分组成。

主队水位	盘口	客队水位
1.05	两球	0.75

1.05 代表购买了 1 元巴西赢盘，若赢了则净赢 1.05 元，本利共返回 2.05 元。

0.75 代表购买 1 元韩国赢盘，若赢则净赢 0.75 元，本利共返回 1.75 元。

图 1: 巴西VS韩国的盘口

亚盘与大小球盘

大小球盘

大小球以“让球”的方式来投注90分钟后比赛的总比分。如果投注者认为总比分能超过比分让球，那就俗称为“大球”，反之，则称为“小球”。

表 1: 大小球盘及竞猜结果

大小球盘口	比赛总进球	买大球结果	买小球结果
2	两球	走盘	走盘
	一球及一球以下	全输	全赢
	三球及三球以上	全赢	全输
2.25	三球及三球以上	全赢	全输
	两球	输一半	赢一半
	一球及一球以下	全输	全赢
2.5	三球及三球以上	全赢	全输
	两球及两球以下	全输	全赢
2.75	两球及两球以下	全输	全赢
	三球	赢一半	输一半
	四球及四球以上	全赢	全输

数据来源

本文研究的是BET365公司亚盘和大小球盘从开盘到闭盘的盘口数据，盘口数据随着时间在不断变化，本质上是一个时间序列。

表 1: 切尔西VS南安普敦的部分盘口数据

主胜赔率	客胜赔率	让球数	大球赔率	小球赔率	分界值	时间
1.91	2.01	-1.75	2.03	1.87	3.25	2013-11-29 15:16:37
1.93	1.99	-1.75	2.03	1.87	3.25	2013-11-29 15:31:53
1.93	1.99	-1.75	1.98	1.92	3.25	2013-11-29 15:36:09
1.86	2.05	-1.50	1.83	2.07	3.00	2013-11-29 20:43:15
1.87	2.04	-1.50	1.83	2.07	3.00	2013-11-29 20:45:36
1.91	2.00	-1.50	1.83	2.07	3.00	2013-11-29 20:47:49
1.89	2.03	-1.50	1.86	2.04	3.00	2013-11-29 22:31:13
1.86	2.06	-1.50	1.86	2.04	3.00	2013-11-30 01:20:48
1.87	2.04	-1.50	1.86	2.04	3.00	2013-11-30 02:13:33
1.81	2.12	-1.50	1.80	2.10	3.00	2013-11-30 02:25:15
1.90	2.02	-1.50	1.86	2.02	3.00	2013-11-30 02:31:30
1.86	2.06	-1.50	1.87	2.02	3.00	2013-11-30 02:31:55
1.79	2.14	-1.50	1.86	2.04	3.00	2013-11-30 02:33:44
2.00	1.91	-1.75	1.86	2.03	3.00	2013-11-30 02:35:47
1.91	2.00	-1.75	1.87	2.01	3.00	2013-11-30 03:42:34

市场期望进球数的反演

1. 在获得了市场各种玩法的赔率后，可以计算市场的去水价格。在其他参数给定的前提下，我们构造由主客队期望进球数计算Dixon模型相应的主队胜去水价格函数和大球去水价格函数，称为模型主队胜去水价格和模型大球去水价格。

2. 市场期望进球数的反演算法的核心思想是：在给定亚盘和大小球盘让球数、赔率的条件下，求主队和客队的期望进球数(x_1^*, x_2^*)，使得：

$$\begin{cases} \text{模型主队胜去水价格} = \text{市场主队胜去水价格} \\ \text{模型大球去水价格} = \text{市场大球去水价格} \end{cases}$$

期望进球差的趋势

亚盘的让球数体现了交战双方的实力差距，传统上以主队和客队表示的交战双方，此时用强队和弱队区分交战双方更为合理。如果亚盘的让球数为负值，则主队为强队，客队为弱队，反之则主队为弱队，客队为强队。

1. t时刻强队与弱队的期望进球差x_t^{*-}为

$$x_t^{*-} = \begin{cases} x_1^* - x_2^*, & \text{盘口为负值}, \\ x_2^* - x_1^*, & \text{其他}. \end{cases}$$

2. t时刻期望进球差的趋势x_t^-定义为

$$x_t^- = \begin{cases} 1, & x_t^{*-} - x_{t-1}^{*-} \geq 0 \\ -1, & x_t^{*-} - x_{t-1}^{*-} < 0 \end{cases}$$

研究数据及目标

研究数据

数据集为$D=\{x_1,x_2,\cdots,x_n\}$, x_i表示第i场比赛强队与弱队的期望进球差的涨跌趋势序列，$R=\{(z_1,wr_1,lr_1),(z_2,wr_2,lr_2),\cdots,(z_n,wr_n,lr_n)\}$, z_i表示比赛的赛果，$z_i=1$表示强队获胜，$z_i=-1$表示弱队获胜，wr_i表示投强队的收益，lr_i表示投资弱队的收益。

研究目标

利用期望进球差的趋势序列确定亚盘闭盘时的投资策略，最大化投资收益。

过滤式特征选择

特征选择

在聚类学习任务中，先进行特征选择，此后再训练学习器。特征选择是一个重要的“数据预处理”过程，进行特征选择的原因有两个：首先是由于属性过多，经常会遇到维数灾难问题；其次去除不相关特征往往会降低学习任务的难度，提高模型的学习性能。

理论基础

Relief是一种著名的过滤式特征选择方法：设计一个相关统计量来度量特征的重要性，每个分量分别对应于一个初始特征。相关统计量的分量越大，对应属性的分类能力就越强。

特征选择

距离度量

样本之间的距离度量方式采用马氏距离，马氏距离的定义如下：

$$\mathrm{dist}(x_i, x_j) = (x_i - x_j)' \hat{\Sigma}^{-1} (x_i - x_j) \tag{1}$$

其中$\hat{\Sigma}$表示协方差的估计.

相关统计量

Relief依据距离度量式(1)在x_i的同类样本中寻找其最近邻$x_{i,nh}$，称为“猜中近邻”,再依据式(1)从x_i的异类样本中寻找其最近邻$x_{i,nm}$，称为“猜错近邻”。相关统计量对应于属性j的分量为

$$\delta^j = \sum_i -\mathrm{diff}(x_i^j, x_{i,nh}^j)^2 + \mathrm{diff}(x_i^j, x_{i,nm}^j)^2 \tag{2}$$

$$\mathrm{diff}(x_b^j, x_d^j) = \begin{cases} 1, & x_b^j \neq x_d^j \\ 0, & x_b^j = x_d^j \end{cases} \tag{3}$$

特征选择

相关统计量

从式(2)可看出，若x_i与其猜中近邻$x_{i,nh}$在属性j上的距离小于x_i与其猜错近邻$x_{i,nm}$的距离，则说明属性j对区分同类与异类样本是有益的，于是增大属性j所对应的统计量分量；反之，若x_i与其猜中近邻$x_{i,nh}$在属性j上的距离大于x_i与其猜错近邻$x_{i,nm}$的距离，则说明属性j起负面作用，于是减小属性j所对应的统计量分量。相关统计量的分量越大，则对应属性的分类能力就越强。

特征选择结果

我们选取欧洲五大联赛2012-2013赛季共计2000场比赛的期望进球差趋势序列进行特征选择，选取了相关统计量最大的12个特征。作为稳健性检验，我们使用欧洲五大联赛2013-2014赛季的数据重复过滤式特征选择过程。结果显示，选出的特征与赛季时间保持一致性，选出了和上述使用2012-2013数据相同的特征。为简单起见，下文以x_i表示经过特征选择后的期望进球差趋势序列。

聚类思想

1. 聚类算法的思想是：将样本集$D=\{x_1,x_2,\cdots,x_n\}$划分为若干个不相交的簇$\{C_l \mid l=1,2,\cdots,k\}$，每个簇对应于一些潜在的类别。

2. k-means算法针对聚类所得簇划分$\mathscr{C}=\{C_1,C_2,\cdots,C_k\}$最小化(4)式

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\mathrm{dist}(x,\boldsymbol{\mu}_i) \tag{4}$$

$$\boldsymbol{\mu}_i=\frac{1}{|C_i|}\sum_{x\in C_i}x, \tag{5}$$

$$\mathrm{dist}(x,\boldsymbol{\mu}_i)=(x-\boldsymbol{\mu}_i)'\hat{\Sigma}^{-1}(x-\boldsymbol{\mu}_i) \tag{6}$$

$\boldsymbol{\mu}_i$是簇C_i的均值向量，$\hat{\Sigma}$表示协方差的估计。

3. 直观看来，式(4)刻画了簇内样本距离簇均值向量的距离，E值越小则簇内样本相似度越高。k-means算法可以发现k个不同的簇，每个簇的中心由簇中所含样本的特征的均值表示。

半监督的聚类

1. 训练集中的比赛已知赛果z_i、投资强队的收益wr_i、投资弱队的收益lr_i。引入$\lambda_j \in \{-1,0,1\}$表示簇C_j的决策标签，属于簇C_j的比赛做相同的投资决策。

$$\lambda_j = \begin{cases} -1, & \frac{1}{|C_j|}\sum_{x_i \in C_j} lr_i \geq \max(cr, \frac{1}{|C_j|}\sum_{x_i \in C_j} wr_i) \\ 1, & \frac{1}{|C_j|}\sum_{x_i \in C_j} wr_i \geq \max(cr, \frac{1}{|C_j|}\sum_{x_i \in C_j} lr_i) \\ 0, & \text{其他} \end{cases}$$

2. 预测集上比赛的投资策略由其所属簇的决策标签决定。

改进的k-means算法

1. k均值算法初始聚类个数k的选取至关重要，改进k-means算法，使其能自动确定聚类个数k的大小。

2. 训练集中的比赛已知赛果z_i，记样本x_i所属簇的决策标签为λ_i^*，根据决策标签的准确率$accurate$选择最优的k。

$$a = |SS|, \quad SS = \{(z_i, \lambda_i^*) | z_i = \lambda_i^*\}$$
$$b = |SB|, \quad SB = \{(z_i, \lambda_i^*) | \lambda_i^* \neq 0\}$$
$$accurate = \frac{a}{b}$$

聚类集成

1. 聚类集成通过对多个聚类学习器进行集成，能有效降低聚类过程中的随机性、聚类假设与真实聚类结构不符等因素带来的不利影响。

2. 采用改进的k-means算法，运行算法N次，得到N个聚类结果。再利用多数投票策略，样本的决策由出现次数最多的决策标签决定。

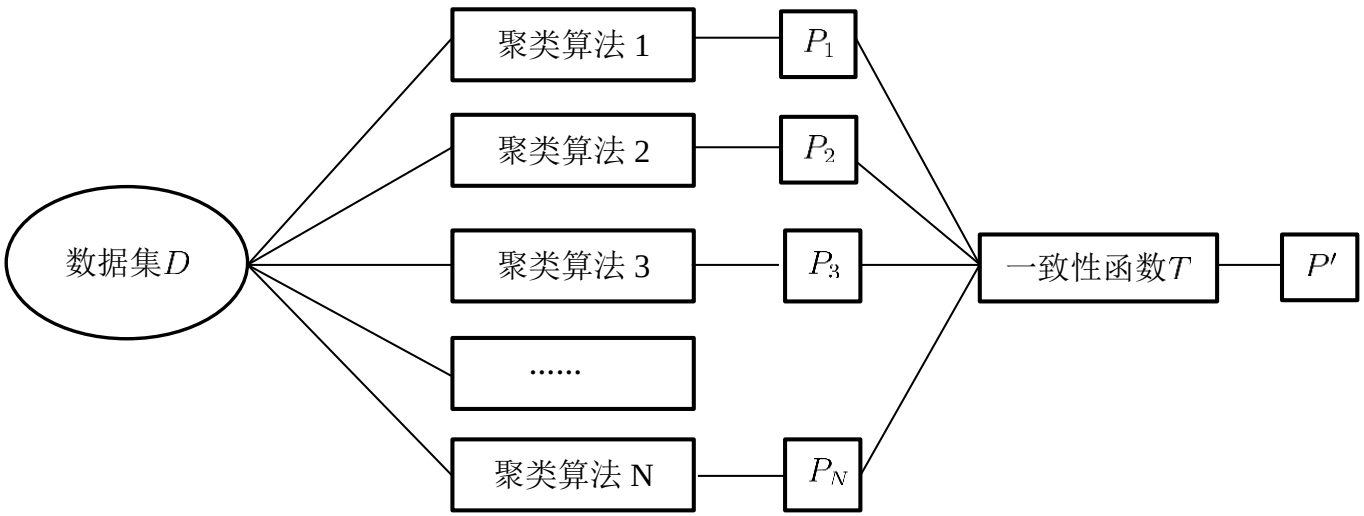


图 1: 聚类集成过程

模型建立

研究设计的算法基于MATLAB2016b，训练集的数据是欧洲五大联赛2012-2014赛季共计4000场比赛，改进的$k-$means聚类集成中簇的数目k，算法运行的次数N，回报率的阈值cr，多数投票的阈值c是决定模型最终结果的参数，通过5次5折交叉验证选择最优参数，交叉验证的部分结果如表1所示。

表 1: 交叉验证阶段模型效果对比

k	N	cr	回报率			准确率		
			c=0.55	c=0.60	c=0.65	c=0.55	c=0.60	c=0.65
10	500	0.01	1.86%	1.94%	1.82%	54.65%	54.70%	54.61%
	500	0.02	1.95%	1.96%	2.00%	54.69%	54.70%	54.72%
	1000	0.01	1.85%	2.00%	2.10%	54.66%	54.72%	54.91%
	1000	0.02	2.25%	2.35%	2.41%	54.90%	54.94%	55.20%
12	500	0.01	1.86%	1.92%	1.92%	54.64%	54.68%	54.67%
	500	0.02	1.87%	1.95%	1.92%	54.67%	54.71%	54.67%
	1000	0.01	1.92%	1.95%	1.86%	54.68%	54.71%	54.64%
	1000	0.02	1.88%	1.88%	1.99%	54.66%	54.68%	54.72%
15	500	0.01	1.83%	1.91%	1.57%	54.65%	54.73%	54.53%
	500	0.02	1.86%	1.74%	1.49%	54.70%	54.64%	54.48%
	1000	0.01	1.54%	1.71%	1.96%	54.51%	54.65%	54.74%
	1000	0.02	1.93%	1.76%	1.61%	54.73%	54.65%	54.57%

模型分析

从表1可以看出：

1. 固定k,N,cr，回报率不一定随着投票的阈值的增大而提高。
2. 预测准确率随参数变化的波动比回报率的波动小，相对比较稳健。
3. 参数$k=10,N=1000,cr=0.02$时，回报率和预测准确率均最高。

由此可见模型参数$k=10,N=1000,cr=0.02$时，模型是最优的。

预测

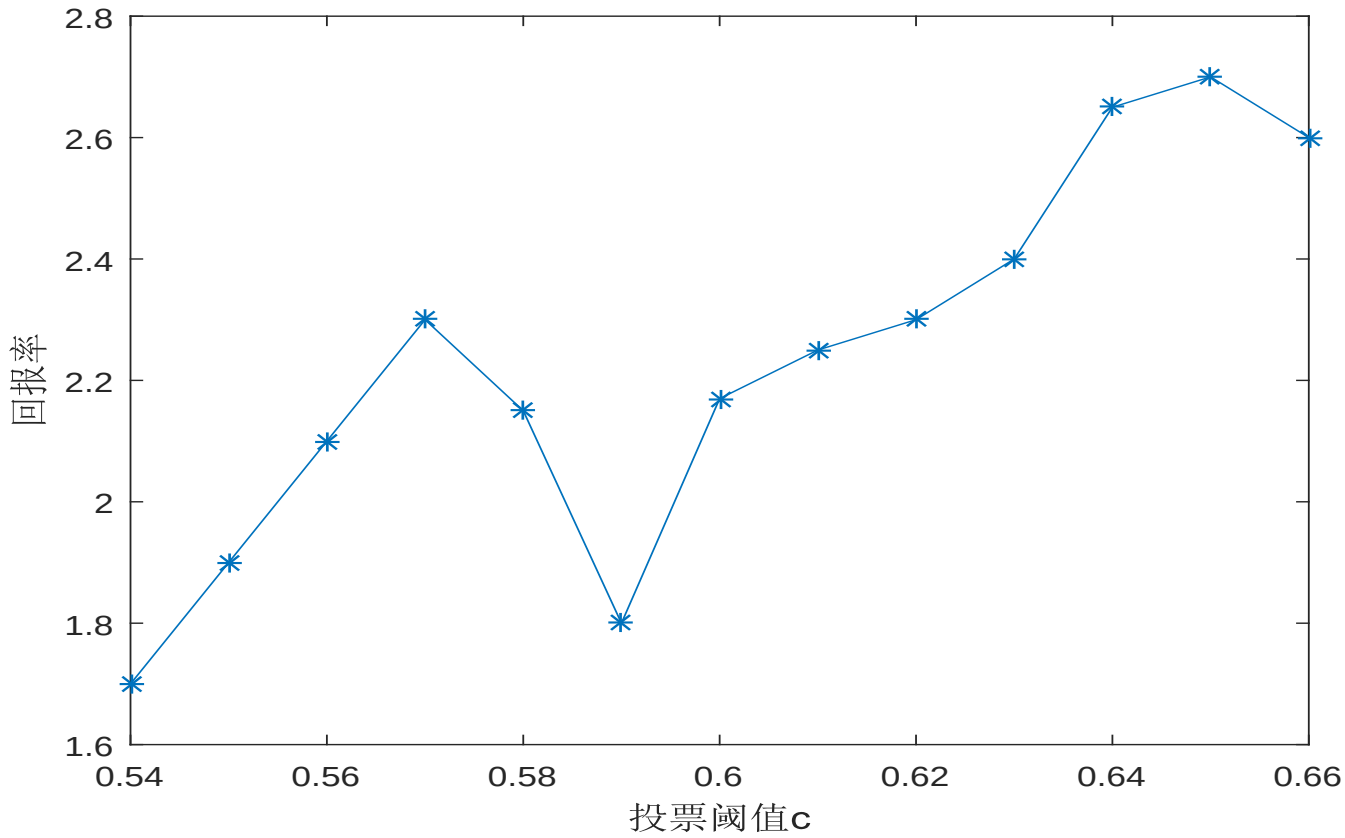
1. 数据集：测试集的数据是欧洲五大联赛2014-2015赛季共计2000场比赛。
2. 更新策略：采用滚动窗口策略不断更新数据集。
3. 模型参数：$k = 10, N = 1000, cr = 0.02$。
4. 表2刻画了不同投票阈值c对应的预测效果。

表 2: 不同投票阈值c对应的预测准确率

投票阈值c	投注的比赛总数	预测准确的比赛总场数	预测准确率
0.55	1905	1045	54.86%
0.56	1888	1037	54.93%
0.57	1866	1026	54.98%
0.58	1848	1014	54.87%
0.59	1834	1005	54.80%
0.60	1818	998	54.90%
0.61	1800	990	54.98%
0.62	1780	979	55.00%
0.63	1757	967	55.04%
0.64	1731	955	55.17%
0.65	1710	946	55.32%

预测分析

1. 投注的比赛的数量随着投票阈值c的提高而减小。
2. 预测效果比较稳健，对投票阈值c不敏感，预测准确率稳定在55%左右，回报率稳定在2.5%左右。
3. 回报率随投票阈值c的变化曲线如下图所示。



结论

1. 本文研究分析了欧洲五大联赛2012-2015共计3个赛季6000场比赛的亚盘和大小球的价格数据，采用了过滤式特征选择，选择了相关统计量最大的12个特征。

2. 我们提出了基于$k-$means聚类集成的半监督模型，以确定亚盘闭盘时的投资策略为目标，以预测准确率和回报率作为两种度量性能的准则。模型建立阶段，采用交叉验证选择模型参数，参数$k=10, cr=0.02, N=1000$时模型是最优的，最优模型在预测集上的预测准确率稳定在55%左右，回报率稳定在2.5%左右。

3. 模型证实了有一定的正收益，在特征选择方面可能还有很大的改进空间，可以把更多的特征（如期望进球差原始序列、最大投注量、联赛类别等）列入候选特征进行特征选择，这是我们未来工作的方向和侧重点。

Chen Yue and **SHI Jian** (2019).
A Cluster Ensemble Strategy for Asian Handicap Betting.
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v 11607 LNAI, p 28-37, 2019, ***Trends and Applications in Knowledge Discovery and Data Mining*** - *PAKDD 2019 Workshops, BDM, DLKT, LDRC, PAISI, WeL, Revised Selected Papers*.
Springer Nature, Switzerland AG 2019

References

1. Karen, D.: Sport matters: sociological studies of sport, violence, and civilization by eric dunning. Br. J. Sociol. **20**(4), 756–757 (2000)
2. Constantinou, A.C., Fenton, N.E., Neil, M.: Profiting from an inefficient association football gambling market: prediction, risk and uncertainty using Bayesian networks. Knowl.-Based Syst. **50**(3), 60–86 (2013)
3. Asian handicap. https://en.wikipedia.org/wiki/Asian handicap. Accessed 21 Dec. 2018
4. Gandar, J., Zuber, R., O'Brien, T., Russo, B.: Testing rationality in the point spread betting market. J. Finan. **43**(4), 995–1008 (1988)
5. Pope, P.F., Peel, D.A.: Information, prices and efficiency in a fixed-odds betting market. Economica **56**(223), 323–341 (1989)
6. Dixon, M.J., Coles, S.G.: Modelling association football scores and inefficiencies in the football betting market. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) **46**(2), 265–280 (1997)
7. Cain, M., Law, D., Peel, D.: The favourite-longshot bias and market efficiency in UK football betting. Scott. J. Polit. Econ. **47**(1), 25–36 (2000)
8. Goddard, J., Asimakopoulos, I.: Forecasting football results and the efficiency of fixed-odds betting. J. Forecast. **23**(1), 51–66 (2004)
9. Forrest, D., Goddard, J., Simmons, R.: Odds-setters as forecasters: the case of English football. Int. J. Forecast. **21**(2), 551–564 (2005)
10. Zhou, Z.H., Tang, W.: Clusterer ensemble. Knowl.-Based Syst. **19**(1), 77–83 (2006)
11. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: a K-means clustering algorithm. J. R. Stat. Soc. Ser. C (Appl. Stat.) **28**(1), 100–108 (1979)
12. Wing, C.K., Yi, K.L.: The use of profits as opposed to conventional forecast evaluation criteria to determine the quality of economic forecasts. Differ. Uravn. **18**(7), 1164–1170 (2007)
13. Williams, L.V.: Weak form information efficiency in betting markets. Leighton Vaughan Williams **51**(1), 1–30 (2005)
14. Hamerly, G., Elkan, C.: Alternatives to the k-means algorithm that find better clusterings. In: 11th International conference on Information and knowledge management, pp. 1–2. ACM Press, Vancouver (2002)