基于相似性匹配和聚类的K线模式可盈利性研究

吕 涛 郝泳涛
(同济大学电子与信息工程学院 上海 200092)

摘 要 K线模式是股票短期投资中最常用的技术分析工具,但学术界却对K线模式的可盈利性存在争议。为了客观评价K线模式的可盈利性,提出从数据挖掘的角度出发,采用模式识别、模式聚类和模式知识挖掘的方法来对K线模式的盈利能力进行研究。为此,首先定义了K线序列的相似性匹配模型来解决K线模式的相似性匹配问题;然后,定义了K线序列的最近邻聚类算法来解决K线模式的聚类问题;最后,定义了K线模式盈利能力度量模型来对K线模式不同形态的利能力进行分析。实验采用近11年上证180指数成份股的数据作为测试数据集,对白三兵和黑三鸦这两个模式的盈利能力进行分析。实验结果表明:同一个K线模式的不同形态的盈利能力差别很大,有时甚至完全相反,这是K线模式可盈利性产生争议的一个主要原因。为了解决这一争议并提高基于K线模式的股票投资效果,亟需根据形态特征对现有的每一个K线模式做进一步分类,并提供更加严谨的模式定义。

关键词 K线,K线序列,K线模式,相似性匹配,聚类,可盈利性
中图法分类号 TP274+.5 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.03.029

Study on K-line Patterns' Profitability Based on Similarity Match and Clustering

LV Tao HAO Yong-tao
(College of Electronics and Information Engineering, Tongji University, Shanghai 200092, China)

Abstract K-line pattern is the most popular technical analysis method for short term stock investment. However, there are some disputes about the K-line patterns' profitability in academia. To resolve the debate, this paper used the method of pattern recognition, pattern clustering and pattern knowledge mining to study the profitability of K-line patterns. Therefore, firstly, the similarity match model was proposed for solving the problem of similarity match of K-line pattern. Secondly, the nearest neighbor-clustering algorithm was proposed for solving the problem of clustering of K-line pattern. Finally, the measurement model of K-line pattern's profitability was defined for measuring the profitability of a K-line pattern's different shapes. In the experiment, the profitability of three white soldiers pattern and three black crows pattern was analyzed with the testing dataset of the K-line series data of Shanghai 180 index component stocks over the latest 11 years. Experimental results show that the main reason for the debate is that the profitability of one pattern varies a great deal for different shapes and they are even opposite at sometimes. There is a need to further classify each of the existing K-line patterns based on the shape feature and give their strict mathematical definitions for improving the profitability and resolving the disputes.

Keywords Candlestick chart, K-line series, K-line pattern, Similarity match, Cluster, Profitability

1 引言

在股票投资技术分析领域,K线技术是股票投资(尤其是短期投资)中最流行的技术分析工具,其原因在于K线可以更加直观地反映股票价格的变动情况。以日K线为例(在本文中,若无特殊说明,所有的K线均指日K线),一根日K线记录的是某只股票在一天内的价格变动情况,它不仅可以记录股票当日的开盘价、收盘价、最高价和最低价等价格指标,还可以直观地反映价格之间(如收盘价与开盘价之间)的差值及大小等。

如果将每天的K线按时间顺序排列在一起,就组成了反映股票价格走势的数列,称为K线序列(K-line Series)。在K线序列中,若某段K线序列蕴含着可以用来进行股票预测的知识,则该段序列被称为K线模式序列(简称K线模式)。例如,如果某段K线序列具有股价涨跌预警功能,即当它出现之后,股票价格经常会出现增长或下跌的现象,那么该K线序列便是一种典型的模式序列。

K线是由18世纪的日本商人本间宗久发明的,最初主要应用于大米交易市场。本间宗久通过长期观察大米交易市场的K线序列,依靠人工方法发现了很多K线模式,并使用这

到稿日期:2017-02-27 返修日期:2017-06-05 本文受"十二五"国家科技支撑计划项目(2015BAF10B01),上海市科委基础研究项目(14JC1402203)资助。

吕 涛(1987-),男,博士,主要研究方向为数据挖掘、算法分析与设计,E-mail:superlvtao@163.com(通信作者);郝泳涛(1973-),男,博士,教授,主要研究方向为数据挖掘、人工智能。

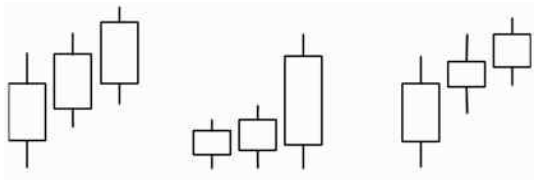
些 K 线模式来进行大米投资，最终赚取了巨额财富。Nison 于 1991 年将 K 线技术介绍给西方世界，并在其专著[1]中详细介绍了现有的一些 K 线模式及其特点，如“白三兵(Three White Soldiers，TWS)”“黑三鸦(Three Black Crows，TBC)”“十字星(Doji)”等模式。从此之后，K 线技术分析法在欧洲乃至整个世界开始变得越来越流行。

尽管 K 线模式已经广泛应用于股票投资，但学术界对于“基于 K 线模式的股票投资是否可以获利(即 K 线模式的可盈利性)”一直存在着一定的争议。不少文献[2-5]认为 K 线模式具有较好的盈利能力。例如，Caginalp 等[2]首次使用假设检验方法证明了在 K 线序列中某些 K 线模式(如 TWS 或 TBC)的出现将会增加之后股票价格上升或下降的概率。Goo 等[3]使用来自于台湾股票市场的 25 只股票在 1997 年－2006 年期间的日 K 线数据，对 K 线模式的盈利能力和持仓周期进行了研究。实验结果表明：一些 K 线模式具有很好的预测能力，不同的 K 线模式需要不同的持仓周期。Lu 等分别在台湾股票市场[4]和欧洲股票市场[5]上对 24 个候选二日 K 线模式的预测能力进行检验。实验表明：有 2 个候选模式在台湾股票市场可以实现盈利，有 3 个候选模式在欧洲股票市场可以盈利。

但也有不少文献[6-10]认为 K 线模式不具有盈利性。例如，Marshall 等[6]先使用道琼斯工业指数成份股在 1992 年－2002 年间的日 K 线数据对 28 个 K 线模式的盈利能力进行研究。实验结果表明：当持仓周期为 10 天时，基于 K 线模式的交易策略无法盈利。基于同样的方法，Marshall 等[7]又使用日本股票市场上市值最大的 100 只股票在 1975 年－2004 年期间的日 K 线数据对同样的 28 个 K 线模式的盈利能力进行研究，最终得到了同样的结论。Horton 等[8]基于 349 只股票研究了 4 对 3 日 K 线模式的盈利能力，最终得出结论：在股票交易时，K 线技术分析法没有用处，因此在买卖股票时，不推荐将“晨星”“乌鸦”或“十字星”等 K 线模式作为技术分析工具。Fock 等[9]和 Duvinage 等[10]分别构建了德国期货市场和美国股票市场的五分钟 K 线序列，并在五分钟 K 线序列上研究 K 线模式的预测能力。他们使用了不同的 K 线模式交易策略，但均得到了基于 K 线模式的交易策略并不能提高投资效益的结论。

针对这种现状，Lu[11]认为模式之前的股票价格走势可能会对 K 线模式的盈利能力产生影响。为此，他使用台湾股票市场 151 只股票在 1992 年－2009 年期间的日 K 线数据，对 12 只 1 日 K 线模式的盈利能力进行研究。实验结果表明：存在 2 个 1 日 K 线模式，处于下降趋势时是可以盈利的；另外存在 2 个 1 日 K 线模式，在上升趋势中可以盈利。在此基础上，Lu 等[12]认为趋势定义与持仓策略的区别很有可能是决定 K 线交易策略是否能够盈利的关键因素。为此，他们使用 3 种趋势定义和 4 种持仓策略，在道琼斯工业指数 30 成份股 1992 年－2012 年的日 K 线数据上，对 8 个 3 日 K 线模式的盈利能力进行了系统性分析。研究发现：持仓策略的类型严重影响着 K 线交易策略的有效性。例如：无论选择哪种趋势定义，当选择 Caginalp-Laurent[2]定义的持仓策略时，8 个 K 线模式均可以盈利；当选择 Marshall[6]定义的持仓策略时，这些模式均不能盈利。

通过梳理相关文献，本文认为产生这种争议的另外一个主要原因在于：现有文献对于 K 线模式缺乏严格的数学定义。例如：对于模式中 K 线的影线长度和实体大小等都缺乏严格的界定，从而使得同一个模式具有多种形态，而每种形态的模式的盈利能力可能不同。因此，在对 K 线模式盈利能力进行研究时，由于没有根据模式的形态差异对模式进行进一步的分类，而是把各种形态的模式作为一个大类统一进行研究，这很可能会影响最终的研究结果。例如：在图 1 中，同一个 TWS 模式具有 3 种不同的形态：形态 A、形态 B 和形态 C。其中形态 A 是 TWS 模式的常规形态，形态 B 和形态 C 是 TWS 模式的两种非常规形态。假设形态 A 具有盈利能力，形态 B 和形态 C 不具有盈利能力。在对 TWS 模式的盈利能力进行研究时，如果忽略这 3 个 TWS 模式在形态方面的差异，而将它们作为一个整体进行研究，就有可能得出错误结论：TWS 模式不具有盈利性。而如果将这 3 个 TWS 模式根据形态进一步分类并分别进行研究，就有可能得到正确结论：TWS 模式只有在形态 A 时才具有盈利性。



(a)形态 A　　(b)形态 B　　(c)形态 C

图 1　白三兵模式的 3 种形态

Fig. 1　Three kinds of shapes of TWS pattern

为了验证上述推断，本文提出从数据挖掘的角度出发，采用模式识别正确结论模式聚类和模式知识挖掘的方法来对 K 线模式的盈利能力进行研究。为此，本文首先定义了 K 线序列的相似性匹配模型来解决 K 线模式的相似性匹配问题；然后定义了 K 线序列的最近邻聚类算法来解决 K 线模式的聚类问题；最后定义了 K 线模式盈利能力度量模型来对 K 线模式不同形态的盈利能力进行分析。

2　K 线和 K 线序列聚类

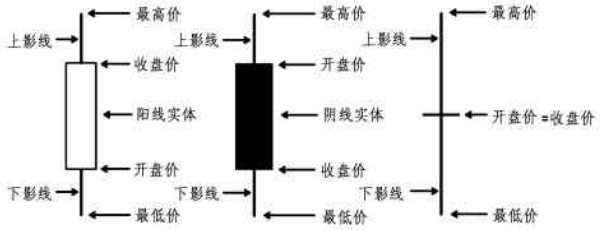
首先给出 K 线序列的数学定义。假设 KS^i 表示任意一只股票的第 i 段 K 线序列，则 $KS^i=[D_1^i,D_2^i,\cdots,D_{|KS^i|}^i)]$，其中 $|KS^i|$($|KS^i|\in N$)表示 KS^i 中元素的个数，D_t^i($1\leqslant t\leqslant|KS^i|$)表示 KS^i 第 t 天的 K 线。$D_t^i=\{C_t^i,O_t^i,H_t^i,L_t^i\}$，其中 C_t^i，O_t^i，H_t^i 和 L_t^i 分别表示 KS^i 第 t 天的收盘价、开盘价、最高价和最低价。

2.1　K 线

2.1.1　K 线定义

K 线是由开盘价、收盘价、最高价和最低价 4 个要素绘制而成的。其中，开盘价与收盘价之间的部分画成矩形实体，称为实体。最高价与矩形实体之间用细线连接，这条线称为上影线。最低价与矩形实体之间用细线连接，这条线称为下影线。由上影线、下影线和实体组成的这种非常具有个性化的线条称为 K 线，如图 2 所示。在 K 线中，如果开盘价小于收盘价，则 K 线被称为阳线，表示股市为牛市，实体通常用白色(或红色)表示，如图 2(a)所示；如果开盘价大于收盘价，则 K 线被称为阴线，表示股市为熊市，实体通常用黑色(或绿色)表

示，如图 2(b)所示；如果开盘价等于收盘价，则 K 线被称为十字星，表示股市稳定，如图 2(c)所示。在此需要说明的是，在中国股市与欧美股市中阴阳线的实体颜色有所不同，在 A 股市场阳线实体为红色，阴线实体为绿色，而在欧美股市阳线实体为绿色，阴线实体为红色。



(a)阳线 (b)阴线 (c)十字星线

图 2 K 线示意图

Fig. 2 K-line

2.1.2 K 线技术分析

假设 D_t 表示任意一只股票第 t 天的 K 线，首先给出 K 线技术分析的一些关键概念。

1)三日移动平均线(或均线)

股票价格在第 t 日的三日移动平均线的定义如下：

$$M_{\text{avg}}(t)=\frac{1}{3}\{C_{t-2}+C_{t-1}+C_t\} \tag{1}$$

其中，C_t 表示第 t 日的收盘价。

2)上升趋势/下降趋势

一般使用三日移动平均线来定义股票价格的走势。因此，第 t 天的股票价格走势为下降趋势(即 D_t 处于下降趋势)的定义公式如下：

$$M_{\text{avg}}(t-6)>M_{\text{avg}}(t-5)>\cdots>M_{\text{avg}}(t) \tag{2}$$

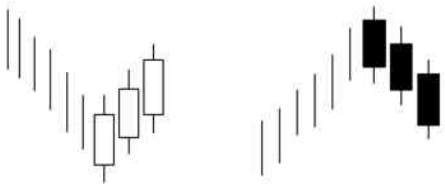
其中，至少有 5 个不等式成立。另外，D_t 处于上升趋势的定义与此类似，不再赘述。

2.1.3 K 线模式

现有的 K 线模式众多，限于论文篇幅，本文仅对 TWS 和 TBC 模式进行详细介绍。假设 $KS=[D_t,D_{t+1},D_{t+2}]$ 表示一段 3 日 K 线序列。

KS 可以成为 TWS 模式(见图 3(a))的条件如下：1)D_t 处于一个下降趋势；2)$C_m>O_m(m=t,t+1,t+2)$，$C_t<C_{t+1}<C_{t+2}$，即每日 K 线均为阳线，且收盘价连续上升；3)$C_t>O_{t+1}>O_t$，$C_{t+1}>O_{t+2}>O_{t+1}$，即每日开盘价都在前一日 K 线的实体范围内。现有文献对 TWS 模式的预测能力的评价为：TWS 是一个股票价格看涨的趋势反转模式，即 TWS 模式出现之后，股票价格很有可能从下降趋势转变为上升趋势，或者说股票市场很有可能从熊市转变为牛市，股价将会走高。

KS 可以成为 TBC 模式(见图 3(b))的条件如下：1)D_t 处于一个上升趋势；2)$C_m<O_m(m=t,t+1,t+2)$，$C_t>C_{t+1}>C_{t+2}$，即每日 K 线均为阴线，且收盘价连续下降；3)$C_t<O_{t+1}<O_t$，$C_{t+1}<O_{t+2}<O_{t+1}$，即每日开盘价都在前一日 K 线的实体范围内。现有文献对 TBC 模式预测能力的评价为：TBC 是一个股票价格看跌的趋势反转模式，即 TBC 模式出现之后，股票价格很有可能从上升趋势转变为下降趋势，或者说股票市场很有可能从牛市转变为熊牛，股价将会走低。



(a)TWS (b)TBC

图 3 TWS 和 TBC 的标准 K 线图

Fig. 3 Standard K-line series chart of TWS and TBC

2.2 K 线序列相似性匹配

目前，研究 K 线序列相似性匹配的文献并不多见，仅有文献[13]采用图像检索技术对 K 线图进行相似性匹配和搜索以进行股票预测。另外，文献[14]提出了一种基于传统欧氏距离的 K 线序列相似性度量方法。K 线序列之间的相似性指的是 K 线序列中 K 线走势(即股票价格走势)之间的相似性，而 K 线走势主要用开盘价、收盘价、最高价和最低价的涨幅度及开盘价与收盘价之间的大小关系来衡量。因此，要度量 K 线序列之间的相似性，应该比较 K 线价格涨幅度之间的相似性，而不应比较 K 线价格之间的相似性。但由于 K 线图并没有直接体现 K 线价格的涨幅度信息，因此文献[13-14]提出的相似性度量方法均使用 K 线价格之间的距离来代替 K 线涨幅度之间的距离，即均属于基于 K 线价格的 K 线序列相似性度量方法，而非基于 K 线涨幅度的 K 线序列相似性度量方法。因此，这些方法并不能准确地度量 K 线序列之间的相似性。

例如，假设 $C_t^i=10$，$C_{t+1}^i=10.5$，$C_t^j=20$，$C_{t+1}^j=21$，$RC_i^{t+1,t}$ 表示 KS^i 第 $t+1$ 日的收盘价涨幅度，其计算方法为 $(C_{t+1}^i-C_t^i)/C_t^i$，$SRC_{i,j}^{t+1,t}$ 表示 $RC_i^{t+1,t}$ 与 $RC_j^{t+1,t}$ 之间的相似度，则 $RC_i^{t+1,t}=RC_j^{t+1,t}=5\%$，$SRC_{i,j}^{t+1,t}=1$。若采用文献[13-14]定义的 K 线序列相似性度量方法，则均无法得到 $SRC_{i,j}^{t+1,t}=1$ 的正确结论。同理，对于开盘价、最高价和最低价之间的相似度，上述方法均存在此类问题。

因此，为了更好地度量 K 线走势之间的相似性，本文定义了一种新的基于 K 线价格涨幅度的 K 线序列相似性度量方法。在本文的方法中，K 线序列之间的相似性主要由两部分组成：1)K 线形态相似性，即两个序列中每两根相对应的 K 线之间在形态上的相似性；2)K 线位置相似性，即两个序列中每两根相对应的 K 线之间在序列位置上的相似性。因此，为了度量 K 线序列之间的相似性，本文将分别定义 K 线序列的形态相似性匹配模型和位置相似性匹配模型。最终，基于这两个相似性匹配模型，便可以得到整个 K 线序列的相似性匹配模型。假设现有两段 K 线序列 KS^i 和 KS^j($|KS^i|=|KS^j|$)需要进行相似性匹配，并记它们之间的相似度为 $Sim^{i,j}$。KS^i 和 KS^j 之间的相似性匹配模型的具体介绍如下。

2.2.1 K 线形态相似性

本文从 K 线的形态特征出发，使用形态距离来度量两个 K 线形态之间的相似性，称之为 K 线形态相似性度量方法。首先，根据 K 线的结构特征，提炼出组成 K 线形态的 3 个部分：上影线形态、实体形态、下影线形态；然后，分别定义这 3 个形态的相似性度量方法；最后，将这 3 个形态的相似度进行累加，从而获得整个 K 线的形态相似度。假设 D_t^i 和 D_t^j 分别

表示 K 线序列 KS^i 和 KS^j 第 t 天的 K 线，它们之间的形态相似性度量模型如下。

1）假设 D_t^i 的上影线长度为 $US^i[t]$，计算公式如式(3)所示：

$$US^i[t]=\begin{cases}\dfrac{H_t^i-O_t^i}{C_{t-1}^i*0.1}, & O_t^i\geqslant C_t^i\\ \dfrac{H_t^i-C_t^i}{C_{t-1}^i*0.1}, & O_t^i<C_t^i\end{cases} \tag{3}$$

其中，$C_{t-1}^i*0.1$ 主要用于实现上影线长度的归一化操作。因为中国 A 股市场规定，每日股价的涨跌度不能超过前一个交易日收盘价的 10%，所以可以使用 $C_{t-1}^i*0.1$ 对上影线、下影线和实体的长度进行归一化操作。

D_t^i 和 D_t^j 的上影线相似度为 $Sim_{US}^{i,j}(t)$，其计算公式如式(4)所示：

$$Sim_{US}^{i,j}(t)=\begin{cases}0, & US^i[t]*US^j[t]=0, US^i[t]\neq US^j[t]\\ \dfrac{\mathrm{Min}(US^i[t],US^j[t])}{\mathrm{Max}(US^i[t],US^j[t])}, & US^i[t]*US^j[t]>0\\ 1, & US^i[t]=US^j[t]=0\end{cases} \tag{4}$$

2）假设 D_t^i 的下影线长度为 $LS^i[t]$，其计算公式如式(5)所示：

$$LS^i[t]=\begin{cases}\dfrac{C_t^i-L_t^i}{C_{t-1}^i*0.1}, & O_t^i\geqslant C_t^i\\ \dfrac{O_t^i-L_t^i}{C_{t-1}^i*0.1}, & O_t^i<C_t^i\end{cases} \tag{5}$$

D_t^i 和 D_t^j 的下影线相似度为 $Sim_{LS}^{i,j}(t)$，其计算公式如式(6)所示：

$$Sim_{LS}^{i,j}(t)=\begin{cases}0, & LS^i[t]*LS^j[t]=0,\ LS^i[t]\neq LS^j[t]\\ \dfrac{\mathrm{Min}(LS^i[t],LS^j[t])}{\mathrm{Max}(LS^i[t],LS^j[t])}, & LS^i[t]*LS^j[t]>0\\ 1, & LS^i[t]=LS^j[t]=0\end{cases} \tag{6}$$

3）假设 D_t^i 的实体长度为 $B^i[t]$，其计算公式如式(7)所示：

$$B^i[t]=(C_t^i-O_t^i)/(C_{t-1}^i*0.1) \tag{7}$$

D_t^i 和 D_t^j 的实体相似度为 $Sim_{Body}^{i,j}(t)$，其计算公式如式(8)所示：

$$Sim_{Body}^{i,j}(t)=\begin{cases}0, & B^i[t]*B^j[t]<0\\ 0, & B^i[t]*B^j[t]=0, |B^i[t]|\neq|B^j[t]|\\ \dfrac{\mathrm{Min}(|B^i[t]|,|B^j[t]|)}{\mathrm{Max}(|B^i[t]|,|B^j[t]|)}, & B^i[t]*B^j[t]>0\\ 1, & B^i[t]=B^j[t]=0\end{cases} \tag{8}$$

4）假设 D_t^i 和 D_t^j 的形态相似度为 $Sim_S^{i,j}(t)$，其计算公式如式(9)所示：

$$\begin{cases}w_{Body}+w_{LS}+w_{US}=1\\ w_{Body}\geqslant 0\\ w_{US}\geqslant 0\\ w_{LS}\geqslant 0\\ Sim_S^{i,j}(t)=w_{Body}*Sim_{Body}^{i,j}(t)+w_{US}*Sim_{US}^{i,j}(t)+\\ \qquad\qquad w_{LS}*Sim_{LS}^{i,j}(t)\end{cases} \tag{9}$$

其中，w_{Body}，w_{US}，w_{LS} 分别表示 $Sim_{Body}^{i,j}(t)$，$Sim_{US}^{i,j}(t)$，$Sim_{LS}^{i,j}(t)$ 的权重。一般来说，在 K 线技术分析中，实体的重要性大于上影线和下影线的重要性。因此，在常规情况下，这些参数的权重可设置为 $w_{Body}=0.6$，$w_{US}=w_{LS}=0.2$。但是，对于具有特殊形态的 K 线（如锥形 K 线、十字星 K 线等），需要增加某些形态的权重才能更加准确地度量具有特殊形态的 K 线之间的相似性。

(5)假设 KS^i 和 KS^j 的形态相似度为 $SSim^{i,j}$，其计算公式如式(10)所示：

$$SSim^{i,j}=\sum_{t=1}^{n}Sim_S^{i,j}(t)*w_S^t \tag{10}$$

其中，$n=|KS^i|$，$\sum_{t=1}^{n}w_S^t=1$，w_S^t 表示 $Sim_S^{i,j}(t)$ 的权重。同理，一般来说，K 线序列中的每个 K 线的权重是相同的。但是，如果 K 线序列中存在着一些具有特殊形态的 K 线，则需要根据具体情况增加这些特殊 K 线的权重，这样才能更加准确地度量具有特殊形态的 K 线序列之间的相似性。

2.2.2　K 线位置相似性

当计算 K 线序列之间的相似性时，不仅需要考虑序列中 K 线之间的形态相似性，还需要考虑 K 线之间的位置相似性。这是因为，如果仅考虑 K 线之间的形态相似性，则会造成 K 线形态相同但序列位置不同的两个 K 线序列被误认为是相同的。例如，假设 KS^i 和 KS^j 的 K 线序列图如图 4 所示。通过分析基于 K 线序列的形态相似性匹配模型（见式(10)）可以发现，在 KS^i 和 KS^j 中，任意两根相对应的 K 线的形态都是完全相同的，即 KS^i 和 KS^j 的整体形态完全相似，$SSim^{i,j}=1$。但从图 4 可以很清晰地看出，虽然 D_1^i 和 D_1^j 在两个序列中的相对位置完全相同，但 D_2^i 和 D_2^j 的相对位置并不完全相同，因此这两个 K 线序列的整体 K 线走势并不完全相同，即 KS^i 和 KS^j 的相似度 $Sim^{i,j}<1$。但是，如果只考虑 K 线序列的形态相似度，则会得出 $SSim^{i,j}=Sim^{i,j}=1$ 的错误结论。

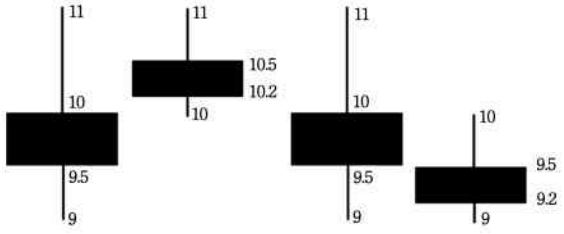


图 4　K 线序列

Fig. 4　K-line series

为了解决这个问题，本文引入 K 线坐标概念，通过定义 K 线在序列中的坐标来解决 K 线的位置匹配问题。本文将 K 线在 K 线序列中的顺序称为 K 线的横坐标；每日收盘价相对于前一日收盘价的涨幅称为 K 线的纵坐标，其中 K 线序列的第一个 K 线的纵坐标设置为 1。因此，$D_t^i(t=1)$ 的横坐标为 1，纵坐标为 1；$D_t^i(t=1)$ 的横坐标为 t，纵坐标为 $(C_t^i-C_{t-1}^i)/(C_{t-1}^i*0.1)$。基于 K 线坐标的 K 线序列位置相似性度量模型如下所示。

1）假设 (x_i^i,y_i^i) 表示 D_t^i 的坐标，如式(11)所示：

$$x_i^t=t, y_i^t=\begin{cases}1, & t=1\\ (C_t^i-C_{t-1}^i)/(C_{t-1}^i*0.1), & t>1\end{cases} \tag{11}$$

D_t^i 和 D_t^j 的位置相似度用 $Sim_P^{i,j}(t)$ 表示，如式(12)所示：

$$Sim_P^{i,j}(t)=\begin{cases}0, & y_t^i * y_t^j=0 \ , y_t^i \neq y_t^j \\ \dfrac{\mathrm{Min}(y_t^i,y_t^j)}{\mathrm{Max}(y_t^i,y_t^j)}, & y_t^i * y_t^j>0 \\ 1, & y_t^i=y_t^j=0\end{cases} \tag{12}$$

2)假设 KS^i 和 KS^j 的位置相似度为 $PSim^{i,j}$，其计算公式如式(13)所示：

$$PSim^{i,j}=\sum_{t=1}^{n} Sim_P^{i,j}(t) * w_P^t \tag{13}$$

其中，$n=|KS^i|$，$\sum_{t=1}^{n} w_P^t=1$，w_P^t 表示 $Sim_P^{i,j}(t)$ 的权重。同理，一般来说，K 线序列中每个 K 线的权重是相同的。但是，如果 K 线序列中存在着一些具有特殊坐标的 K 线，则同样需要根据具体情况增加这些特殊 K 线的权重，这样才能更加准确地度量具有特殊坐标的 K 线序列之间的相似性。

2.2.3 K 线序列相似性

基于 K 线序列的形态相似度和位置相似度，便可以获得整个 K 线序列的相似度。因此，KS^i 和 KS^j 的相似性匹配模型如式(14)所示：

$$Sim^{i,j}=SSim^{i,j} * w_S+PSim^{i,j} * w_P \tag{14}$$

其中，w_S 表示 K 线序列的形态相似度权重，w_P 表示位置相似度权重。一般来说，形态相似度大于位置相似度，因此它们的权重设置推荐为：$w_S=0.8$，$w_P=0.2$。但是，对于特殊的 K 线序列，需要根据具体情况对权重进行调整。

2.3 K 线序列聚类

基于 K 线序列的相似性匹配模型，采用最近邻聚类算法对 K 线模式进行聚类，便可以得到更加精确的 K 线模式分类结果。K 线序列的最近邻聚类算法(K-line series Nearest Neighbor Clustering Algorithm，KNNCA)如算法 1 所示。

算法 1 K 线序列的最近邻聚类算法

```
输入：KSet={KS1，KS2，…，KSn}；//K 线序列集合
      θ //相似度阈值
输出：CSet //簇集
Assign initial value for parameters: wS, wP, wBody, wUS, wLS, wSt,
wPt；// 初始化相关参数
m=1；
Qm={KS1}；//Qm 表示第 m 个簇
CSet={Qm}；
FOR i=2 TO n DO
{
  SimMax=0；
  FOR EACH Qitem IN CSet
    FOR EACH KSj IN Qitem
      Get Simi,j based on formula 14；
      If (Simi,j>SimMax)
      {
          SimMax=Simi,j；
          f=item；// f 表示与 KSi 最相似的 KSj 所属的簇号
      }
    End
  End
  IF (SimMax > θ)THEN
    Qf=Qf∪KSi；
  ELSE
  {
    m=m+1；
    Qm={KSi}；
  }
}
CSet={Q1，Q2，…，Qm}；
```

其中，相似度阈值 θ 的范围一般为 0.7～1。由于每个 K 线序列都要与已经在簇中的 K 线序列进行逐个比较，因此 KNNCA 算法的时间复杂度和空间复杂度均为 $O(n^2)$。

3 模式的盈利能力分析

基于 K 线模式定义和 K 线模式聚类算法，同一 K 线模式不同形态的盈利能力最终可以通过以下步骤进行挖掘和分析：

1)模式识别。根据 K 线模式定义，从数据集中识别出属于某一 K 线模式(如 TWS)的所有 K 线序列，并构成集合 $KSet$。

2)模式聚类。使用 KNSSC 算法对 $KSet$ 进行聚类，最终可得到簇集 $CSet$。在 $CSet$ 中，不同的簇代表同一模式的不同形态。

3)盈利能力分析。首先，本文使用主流文献中最常用的 K 线投资策略[3-7,11-12]来计算 K 线模式的收益率。具体如下：①以模式出现之后第一日的收盘价买入(或卖出)股票，该价格即为建仓价。②持有(或等待)一段时间再卖出(或买入)。该段时间即为持仓时间(或持仓周期)，记为 f。文献[1,15]均指出 K 线技术分析主要适用于短期预测，并且最有效的时间为 10 日以内。因此，一般情况下 $1\leqslant f\leqslant 10$。③以模式出现之后第 f 日的收盘价卖出(或买入)股票。该价格即为平仓价。④基于建仓价和平仓价计算 K 线模式持仓 f 日的收益率，记为 E_f。

令 $KS=[D_t,D_{t+1},D_{t+2}]$ 表示一个 3 日 K 线模式，PT 表示模式类型，CKS 表示 KS 的后续 K 线序列。则 E_f 的计算公式如下：

$$E_f=\begin{cases}(C_{t+2+f}-O_{t+3})/O_{t+3}, & PT=1 \\ (O_{t+3}-C_{t+2+f})/O_{t+3}, & PT=0\end{cases} \tag{15}$$

其中，O_{t+3} 表示 CKS 中第一日的开盘价，C_{t+2+f} 表示 CKS 中第 f 日的收盘价。$PT=1$ 表示看涨模式，$PT=0$ 表示看跌模式。

其次，计算每个簇内的 K 线模式在第 f 个持仓日(即持仓 f 日)收益为正的概率。记 $P(E_f^m>0)$ 表示 Q^m 中的 K 线模式在第 f 个持仓日收益为正的概率，如式(16)所示：

$$P(E_f^m>0)=|Q^{m,f}|/|Q^m| \tag{16}$$

其中，$|Q^{m,f}|$ 表示 Q^m 中满足条件 $E_f^m>0$ 的模式个数，$|Q^m|$ 表示 Q^m 中的元素个数。可以看出，$P(E_f^m>0)$ 的值越大表示该模式在第 f 个持仓日的收益为正的概率越大。若 $P(E_f^m>0)>0.5$，则表示该模式在第 f 个持仓日的收益倾向于正收益；若 $P(E_f^m>0)<0.5$，则表示该模式的收益倾向于负收益。

然后,计算每个簇内的 K 线模式在 n 个持仓日内的收益倾向于正收益的概率。记 $P_m^n \in [0,1]$ 表示 Q_m 中的 K 线模式在 n 日内的收益倾向于正收益的概率,如式(17)所示:

$$P_m^n = \sum_{f=1}^{n} \frac{R[P(E_f^m>0),0.5]}{n}$$

$$R[P(E_f^m>0),0.5] = \begin{cases} 1, & P(E_f^m>0)>0.5 \\ 0, & P(E_f^m>0)\leqslant 0.5 \end{cases} \tag{17}$$

其中,P_m^n 的值越大,表示其盈利能力越强,n 一般取值为 10。

最后,基于 K 线模式在第 f 日的盈利能力度量模型(见式(16))和 K 线模式 n 日内的盈利能力度量模型(见式(17)),对每个簇内 K 线模式的盈利能力(即同一 K 线模式不同形态的盈利能力)进行分析。

4　实验结果与讨论

4.1　实验数据及方法

由于雅虎公司提供了可以下载中国股市每天交易数据的 API——Finance stock API[16],因此基于 Finance stock API,通过编写程序可以获取中国 A 股市场任意时期的 K 线数据。为了使测试数据具有一定代表性,本文选择上证 180 指数成份股(2016-07-08 发布版本)所有股票近 11 年(2006-1-4—2016-8-24)的 K 线序列数据作为测试数据集。

由于现有 K 线模式较多,限于论文篇幅,本文仅对 TWS 和 TBC 模式在中国 A 股市场的盈利能力进行研究。其中 KNSSC 算法的相关参数设置如下:$\theta=0.7$,$w_S=0.8$,$w_P=0.2$,$w_{Body}=0.6$,$w_{US}=0.2$,$w_{LS}=0.2$,$w_S^t=w_P^t=1/|KS^i|$ $(t=1,2,\cdots,|KS^i|)$。

4.2　实验 1——TWS 模式的盈利能力分析

采用上文定义的模式盈利能力度量模型,对 TWS 模式在不同形态下的盈利能力进行分析。首先根据 TWS 模式的定义,从测试数据集中识别出 818 个 TWS 模式。接着使用 KNSSC 算法对这些模式进行聚类,最终可以得到 183 个簇。

首先选取簇元素最多的前 8 个簇进行统计分析,结果如表 1 所列。

表 1　TWS 的实验结果

Table 1　Experimental results of TWS

f	Q^m	Q^0	Q^1	Q^2	Q^3	Q^4	Q^5	Q^6	Q^7	Q^8
	$\lvert Q^m\rvert$	818	129	69	29	28	24	21	19	18
1		0.48	0.46	0.62	0.48	0.46	0.54	0.57	0.74	0.50
2		0.52	0.55	0.52	0.59	0.54	0.54	0.52	0.63	0.44
3		0.51	0.58	0.57	0.41	0.50	0.58	0.43	0.53	0.44
4		0.49	0.54	0.54	0.48	0.57	0.50	0.43	0.58	0.50
5		0.50	0.47	0.52	0.55	0.61	0.54	0.48	0.53	0.39
6	P	0.49	0.51	0.51	0.59	0.54	0.46	0.48	0.53	0.44
7	$(E_f^m>0)$	0.51	0.55	0.51	0.62	0.54	0.54	0.57	0.74	0.56
8		0.51	0.55	0.54	0.69	0.50	0.50	0.57	0.68	0.56
9		0.52	0.62	0.52	0.66	0.50	0.54	0.67	0.68	0.39
10		0.52	0.58	0.51	0.62	0.54	0.50	0.57	0.68	0.44
$P_m^{n=10}$		0.60	0.80	0.80	1.00	0.70	0.60	0.60	0.60	1.00

在表 1 中,Q^0 表示由 818 个 TWS 模式组成的簇。其 P_m^{10} 为 0.6,表明 TWS 模式在 10 日内的盈利能力一般。另外,$\forall f \in [1,10]$,$P(E_f^0>0)\approx 0.5$,表明 TWS 模式在每一个持仓日收益为正或为负的概率是基本均等的。这进一步表明 TWS 的盈利能力一般。这也是学术界对于 K 线模式盈利能力的评价产生冲突的根本原因。因此需要对 TWS 模式进行进一步的分类。对 TWS 模式进行聚类之后,从表 1 可以看出:1)Q^1,Q^2 和 Q^7 表现出了较好的盈利能力,其 P_m^{10} 均在 0.8 以上;2)Q^4,Q^5 和 Q^6 表现出的盈利能力一般,它们的 P_m^{10} 仅在 0.5～0.6 之间;3)Q^8 表现出了较差的盈利能力,其 P_m^{10} 仅为 0.2。

然后,将 Q^7 和 Q^8 的盈利能力进行对比,如图 5 所示。从图 5 可以看出,Q^7 和 Q^8 的盈利能力基本相反,其中 Q^7 倾向于正收益,Q^8 倾向于负收益。实验 1 的结果表明:不同形态的 TWS 模式的盈利能力的差别较大,这符合本文对于 K 线模式盈利能力的预期分析。

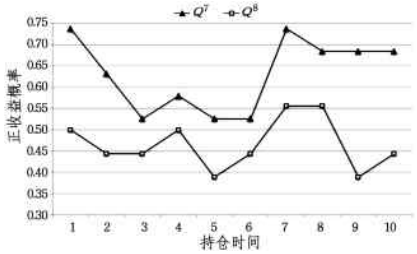


图 5　Q^7 和 Q^8 的盈利能力对比

Fig. 5　Comparison of profitability between Q^7 and Q^8

4.3　实验 2——TBC 模式的盈利能力分析

基于上文定义的模式盈利能力度量模型,对 TBC 模式在不同形态下的盈利能力进行分析。首先根据 TBC 模式的定义,从测试数据集中识别出 339 个 TBC 模式,然后使用 KNSSC 算法对这些模式进行聚类,最后得到 140 个簇。

同理,首先选取簇元素最多的前 8 个簇进行统计分析,如表 2 所列。

表 2　TBC 的实验结果

Table 2　Experimental results of TBC

f	Q^m	Q^0	Q^1	Q^2	Q^3	Q^4	Q^5	Q^6	Q^7	Q^8
	$\lvert Q^m\rvert$	339	24	16	15	15	15	12	9	9
1		0.38	0.54	0.56	0.40	0.47	0.33	0.44	0.22	0.44
2		0.45	0.46	0.63	0.33	0.53	0.50	0.67	0.44	0.78
3		0.45	0.46	0.56	0.33	0.67	0.75	0.22	0.33	0.89
4		0.48	0.42	0.50	0.60	0.80	0.75	0.33	0.44	0.78
5		0.47	0.42	0.44	0.47	0.87	0.75	0.22	0.44	0.56
6	P	0.48	0.42	0.44	0.60	0.80	0.83	0.33	0.56	0.44
7	$(E_f^m>0)$	0.49	0.42	0.56	0.60	0.73	0.83	0.33	0.56	0.33
8		0.49	0.50	0.56	0.60	0.80	0.67	0.33	0.56	0.56
9		0.50	0.50	0.63	0.53	0.80	0.67	0.67	0.56	0.33
10		0.52	0.42	0.50	0.53	0.73	0.75	0.56	0.67	0.44
$P_m^{n=10}$		0.10	0.10	0.60	0.60	0.90	0.80	0.30	0.50	0.50

类似地,在表 2 中,Q^0 表示由 339 个 TBC 模式组成的簇。其 P_m^{10} 为 0.1,表明 TBC 模式在 10 日内的盈利能力较差,这说明它很有可能是一个伪模式。但经过聚类之后,从表 2 可以看出:1)Q^4 和 Q^5 都表现出了较好的盈利能力,其 P_m^{10} 都在 0.8 以上;2)Q^2,Q^3,Q^7 和 Q^8 表现出的盈利能力一般,它们的 P_m^{10} 仅在 0.5～0.6 之间;3)Q^1 和 Q^6 表现出了较差的盈利能力,它们的 P_m^{10} 均小于 0.5,尤其是 Q^1,其 P_m^{10} 为0.1。这表明一些形态的 TBC 模式仍然具有较好的盈利能力,因此不能认为它是一个伪模式。

然后，将 Q^1 和 Q^4 的盈利能力进行对比，如图 6 所示。从图 6 可以看出，Q^1 和 Q^4 的盈利能力完全相反，其中 Q^1 倾向于负收益，Q^4 倾向于正收益。而且，除第一个持仓日外，两者在每一个持仓日的收益倾向均相反。实验 2 的结果表明：不同形态的 TBC 模式的收益能力的差别也较大，这同样也符合本文对 K 线模式盈利能力的预期分析。

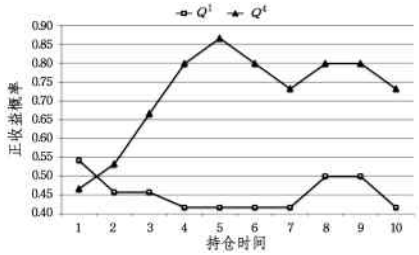


图 6 Q^1 和 Q^4 的盈利能力对比

Fig. 6 Comparison of profitability between Q^1 and Q^4

4.4 实验结论

通过上面的实验分析可以得出如下结论：1）对于同一个 K 线模式的不同形态，其盈利能力差别很大，有时甚至完全相反。以 TWS 和 TBC 模式为例，有一部分形态的 TWS/TBC 模式表现出了较好的盈利能力，但也有不少形态的 TWS/TBC 模式表现出了较差的盈利能力。因此，分析一个 K 线模式的盈利能力时应根据其具体形态进行具体分析，这样才能得到模式盈利能力的准确评价。2）现有文献在检验 K 线模式的盈利能力时，由于没有考虑同一 K 线模式不同形态之间的差异性，而是将它们作为一个大类统一进行研究，因此在对同一 K 线模式的可盈利性进行研究时，不同文献的实验数据中可能会包含该模式的不同形态，从而使得最终的实验结果有所不同，甚至是完全相反。这是 K 线模式可盈利性产生争议的一个重要原因。3）为了解决争议并提高基于 K 线模式的股票投资效果，亟需根据形态特征对现有的每一个 K 线模式做进一步分类，并提供更加严谨的模式定义。

结束语 作为股票短期投资的一种主要技术分析方法，基于 K 线模式的股票投资虽然在现实中的应用非常广泛，但学术界对其可行性一直存在争议。为了解决这些争议，本文采用模式识别、相似性匹配、聚类、统计分析等数据挖掘方法，对现有 K 线模式的盈利能力进行研究。实验结果表明，K 线模式的盈利能力之所以出现争议，主要是因为现有 K 线模式的定义比较开放且缺乏严谨的数学定义。基于形态特征对现有 K 线模式进行进一步的分类，便可以解决该问题。因此，要想提高股票投资收益，亟需对现有 K 线模式进行进一步分类，并选择盈利能力较好的模式作为技术分析指标。另外，本文提出的方法不仅可以对现有 K 线模式的盈利能力进行分析，还可以应用于 K 线模式挖掘和股票预测。因此，下一步的研究内容为：1）由于现有模式主要是通过人工方式挖掘出来的，其中很可能存在一些伪模式，因此基于本文方法从现有模式中识别出可能存在的伪模式，将是一件非常必要而又有意义的工作。2）在本文方法的基础上，研究出一种自动化的模式挖掘方法，以挖掘出更多、更好的 K 线模式进行股票预测。

参 考 文 献

[1] NISON S. Japanese Candlestick Charting Techniques: A Contemporary Guide to the Ancient Investment Technique of the Far East [M]. New York: New York Technique of the Far East, 1991: 30-40.

[2] CAGINALP G, LAURENT H. The Predictive Power of Price Patterns [J]. Applied Mathematical Finance 1998, 5(3-4): 181-205.

[3] GOO Y J, CHEN D H, CHANG Y W. The application of Japanese candlestick trading strategies in Taiwan[J]. Inverstment Management and Financial Innovations, 2007, 4(4): 49-67.

[4] LU T H, SHIU Y M. Tests for Two Day Candlestick Patterns in the Emerging Equity Market of Taiwan [J]. Emerging Markets Finance & Trade, 2012, 48(1): 41-57.

[5] LU T H, CHEN J. Candlestick charting in European stock markets [J]. Jassa: The Finsia Journal of Applied Finance, 2013(2): 20-25.

[6] MARSHALL B R, YOUNG M R, ROSE L C. Candlestick technical trading strategies: Can they create value for investors? [J]. Journal of Banking & Finance, 2006, 30(8): 2303-2323.

[7] MARSHALL B R, YOUNG M R, CAHAN R. Are candlestick technical trading strategies profitable in the Japanese equity market? [J]. Review of Quantitative Finance & Accounting, 2008, 31(2): 191-207.

[8] HORTON M J. Stars, crows, and doji: The use of candlesticks in stock selection [J]. Quarterly Review of Economics & Finance, 2009, 49(2): 283-294.

[9] FOCK J H, KLEIN C, ZWERGEL B. Performance of Candlestick Analysis on Intraday Futures Data [J]. Journal of Derivatives, 2005, 13(1): 28-40.

[10] DUVINAGE M, MAZZA P, PETITJEAN M. The intra-day performance of market timing strategies and trading systems based on Japanese candlesticks[J]. Quantitative Finance, 2013, 13(7): 1059-1070.

[11] LU T H. The profitability of candlestick charting in the Taiwan stock market [J]. Pacific-Basin Finance Journal, 2014, 26(26): 65-78.

[12] LU T H, CHEN Y C, HSU Y C. Trend definition or holding strategy: What determines the profitability of candlestick charting?[J]. Journal of Banking & Finance, 2015, 61: 172-183.

[13] TSAI C F, QUAN Z Y. Stock Prediction by Searching for Similarities in Candlestick Charts[J]. Acm Transactions on Management Information Systems, 2014, 5(2): 1-21.

[14] CHMIELEWSKI L, JANOWICZ M, KALETA J, et al. Pattern Recognition in the Japanese Candlesticks[C]// Advances in Intelligent Systems and Computing. Miedzyzdroje: Springer Press, 2015: 227-234.

[15] MORRIS G L, LITCHFIELD R. Candlestick charting explained: timeless techniques for trading stocks and futures 3rd Ed [M]. New York: McGraw-Hill, 2006.

[16] Yahoo!. Finance stock API[EB/OL]. [2017-06-28]. http://blog.sina.com.cn/s/blog12b66a6db0102w934.html.