

R在金融数据挖掘的应用 ——预测股票收益率

新华社中经社控股集团

陈堰平

2011年11月12日



Why R?

1. 开源，免费，用户贡献自己的包
2. 灵活，可处理多种事务、**OOP**、泛型函数
3. 有大量的函数、包，不用再重新发明轮子
4. 先用**R**实现算法原型，再用**C++**、**C#**等语言开发系统

.....



传统的时间序列

Box-Jenkins: AR、MA、ARMA、ARIMA

异方差模型: ARCH类 ...

...



以交易为目的的预测

- 预测是为决策提供支持的
 - 预测模型与交易系统整合

- 评价标准是交易系统的收益，而不是预测的精确度



目录

- 一、数据导入
- 二、预测模型
- 三、从预测到决策
- 四、模型评价和选择

一、数据导入

1. 数据的结构：交易的日期、开盘价、最高价、最低价、收盘价、交易量、调整的收盘价
2. 为简单起见，用的是股票指数的数据
3. 处理时间序列的包 **zoo**, **xts**, 后者是前者的拓展
4. 表示的处理时间的类： **POSIXct/ POSIXlt**, **date**

```
Open High Low Close Volume AdjClose
1970-01-02 92.06 93.54 91.79 93.00 8050000 93.00
1970-01-05 93.00 94.25 92.53 93.46 11490000 93.46
1970-01-06 93.46 93.81 92.13 92.82 11460000 92.82
1970-01-07 92.82 93.38 91.93 92.63 10010000 92.63
1970-01-08 92.63 93.47 91.99 92.68 10670000 92.68
1970-01-09 92.68 93.25 91.82 92.40 9380000 92.40
```

5. 从CSV文件读数据

```
> GSPC <- as.xts(read.zoo("sp500.csv", header = T))
```

6. 从网络读取数据

```
> library(tseries)
```

```
> GSPC <- as.xts(get.hist.quote("^GSPC",start="1970-01-02",  
quote=c("Open", "High", "Low", "Close","Volume","AdjClose")))
```

7. quantmod包里的getSymbols()

```
> setSymbolLookup(IBM=list(name='IBM',src='yahoo'),
```

```
+ USDEUR=list(name='USD/EUR',src='oanda))
```

```
> getSymbols(c('IBM','USDEUR'))
```

8. 从数据库读数据：包RODBC，RMySQL

win下安装myodbc驱动使用RODBC，在linux下直接用RMySQL和DBI

目标变量

日平均价用下面的式子近似

$$\bar{P}_i = \frac{C_i + H_i + L_i}{3}$$

设 V_i 是接下来 k 天平均价对 t 时刻的变动百分比（算术收益率）

$$V_i = \left\{ \frac{\bar{P}_{i+j} - C_i}{C_i} \right\}_{j=1}^k$$

指标 T 定义为绝对值大于 $p\%$ 的那些变动的总和

$$T_i = \sum_v \{v \in V_i : v > p\% \text{ or } v < -p\% \}$$

T 为正且较大说明未来有若干天股价高于今天的收盘价，买入信号

T 为负且绝对值较大说明未来有若干天股价低于今天的收盘价，卖出信号



```
T.ind <- function(quotes,tgt.margin=0.025,n.days=10) {  
  v <- apply(HLC(quotes),1,mean)  
  r <- matrix(NA,ncol=n.days,nrow=NROW(quotes))  
  for(x in 1:n.days) r[,x] <- Next(Delt(v,k=x),x)  
  x <- apply(r,1,function(x) sum(x[x > tgt.margin | x < -tgt.margin]))  
  if (is.xts(quotes))  
    xts(x,time(quotes))  
  else x }  
}
```



用什么变量预测？

历史数据

技术指标：包TTR

有大量的指标，如何选择？

- 1.特征过滤（**feature filters**），不依赖于模型
- 2.特征封装（**feature wrappers**），依赖于模型，迭代的

候选特征

$$R_{i-h} = \frac{C_i - C_{i-h}}{C_{i-h}} \quad \text{变动}h: 1 \rightarrow 10$$

TTR的技术指标:

ATR (Average True Range), 衡量序列波动

SMI(Stochastic Momentum Index), 动量指标


Average Directional Movement Index (ADI)

Aroon指标, 捕捉起始趋势的; Bollinger Bands, 比较一段时期的波动率

Chaikin Volatility; EMV (Ease of Movement Value); MACD

MFI (Money Flow Index) . . .

先经过预处理, 产生单指标



```
myATR <- function(x) ATR(HLC(x))['atr']
mySMI <- function(x) SMI(HLC(x))['SMI']
myADX <- function(x) ADX(HLC(x))['ADX']
myAroon <- function(x) aroon(x[,c('High','Low')])$oscillator
myBB <- function(x) BBands(HLC(x))['pctB']
myChaikinVol <- function(x)
Delt(chaikinVolatility(x[,c("High","Low")]))[,1]
myCLV <- function(x) EMA(CLV(HLC(x)))[,1]
myEMV <- function(x) EMV(x[,c('High','Low')],x[, 'Volume'])[,2]
myMACD <- function(x) MACD(CI(x))[,2]
myMFI <- function(x) MFI(x[,c("High","Low","Close")], x[, "Volume"])
mySAR <- function(x) SAR(x[,c('High','Close')]) [,1]
myVolat <- function(x) volatility(OHLC(x),calc="garman")[,1]
```

特征选择

随机森林 *library(randomForest)*

把数据分两部分 (1) 构建交易系统 (2) 测试

```
library(randomForest)
```

```
data.model <- specifyModel(T.ind(GSPC) ~ Delt(CI(GSPC),k=1:10) +  
  myATR(GSPC) + mySMI(GSPC) + myADX(GSPC) + myAroon(GSPC)  
  + myBB(GSPC) + myChaikinVol(GSPC) + myCLV(GSPC) +  
  CMO(CI(GSPC)) + EMA(Delt(CI(GSPC))) + myEMV(GSPC) +  
  myVolat(GSPC) + myMACD(GSPC) + myMFI(GSPC) +  
  RSI(CI(GSPC)) + mySAR(GSPC) + runMean(CI(GSPC)) +  
  runSD(CI(GSPC)))
```

```
set.seed(1234)
```

```
rf <- buildModel(data.model,method='randomForest',  
training.per=c(start(GSPC),index(GSPC["1999-12-31"])), ntree=50,  
importance=T)
```

```
ex.model <- specifyModel(T.ind(IBM) ~ Delt(CI(IBM),k=1:3))
```

```
data <- modelData(ex.model,data.window=c('2009-01-01','2009-08-10'))
```

预测问题

1. 用解释变量来预测 T （回归问题），然后计算信号 $signal$

$$signal = \begin{cases} sell & \text{if } T < -0.1 \\ hold & \text{if } -0.1 \leq T \leq 0.1 \\ buy & \text{if } T > 0.1 \end{cases}$$

2. 用解释变量直接预测 $signal$ （分类）

问题!!!

sell和**buy**是少数，**hold**住是多数
原因：不平衡数据

评估准则

$$error.rate = \frac{1}{N} \sum_{i=1}^N L_{0/1}(y_i, \hat{y}_i)$$

		Predictions			
		sell	hold	buy	
True Values	sell	$n_{s,s}$	$n_{s,h}$	$n_{s,b}$	$N_{s,..}$
	hold	$n_{h,s}$	$n_{h,h}$	$n_{h,b}$	$N_{h,..}$
	buy	$n_{b,s}$	$n_{b,h}$	$n_{b,b}$	$N_{b,..}$
		$N_{.,s}$	$N_{.,h}$	$N_{.,b}$	N

$$Prec = \frac{n_{s,s} + n_{b,b}}{N_{.,s} + N_{.,b}}$$

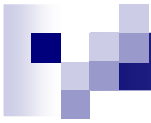
$$Prec_b = \frac{n_{b,b}}{N_{.,b}}$$

$$Rec = \frac{n_{s,s} + n_{b,b}}{N_{s,..} + N_{b,..}}$$

$$Rec_b = \frac{n_{b,b}}{N_{b,..}}$$

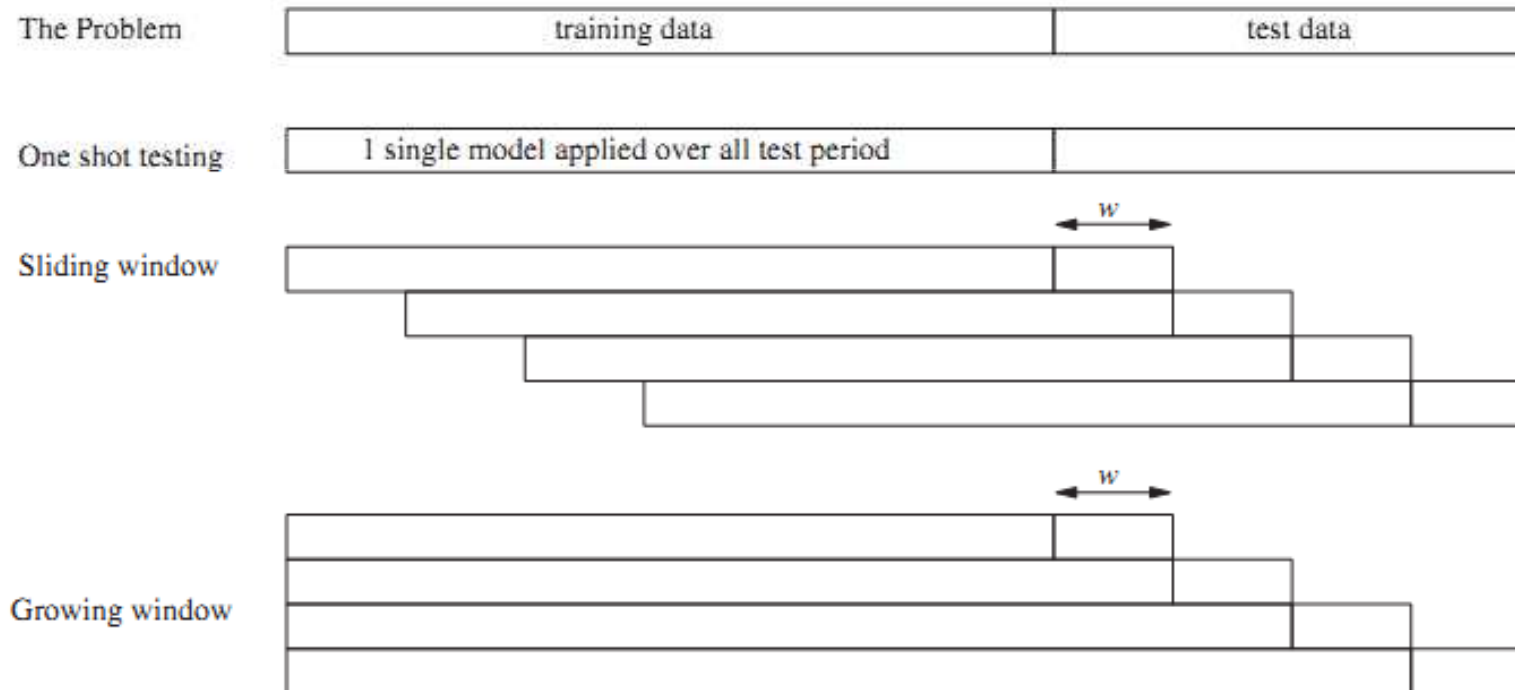
Prec和Recall通常合并到一起，成为单一的统计量，*F*-测度（Rijsbergen, 1979）

$$F = \frac{(\beta^2 + 1) \cdot Prec \cdot Rec}{\beta^2 \cdot Prec + Rec}$$



负责的金融时间序列通常会出现区域转移（**regime switch**）的情况

如果用训练集数据得到模型来预测接下来的时间区域，可能会发现区域转移——用较近的数据来扑捉最近区域（**regime**）的特性



预测模型

1. 人工神经网络

```
set.seed(1234)
library(nnet)
norm.data <- scale(Tdata.train)
nn <- nnet(Tform, norm.data[1:1000, ], size = 10, decay
  = 0.01,
  maxit = 1000, linout = T, trace = F)
norm.preds <- predict(nn, norm.data[1001:2000, ])
preds <- unscale(norm.preds, norm.data)
```

2. 支持向量机：包kernlab、e1071

```
library(e1071)
sv <- svm(Tform, Tdata.train[1:1000, ], gamma = 0.001,
          cost = 100)
s.preds <- predict(sv, Tdata.train[1001:2000, ])
```

```
library(kernlab)
data <- cbind(signals = signals, Tdata.train[, -1])
ksv <- ksvm(signals ~ ., data[1:1000, ], C=10)
ks.preds <- predict(ksv, data[1001:2000, ])
```

3. 多变量自适应回归样条

Multivariate Adaptive Regression Splines

$$mars(\mathbf{x}) = c_0 + \sum_{i=1}^k c_i B_i(\mathbf{x})$$

包mda的mars(), 包earth里的earth()

```
library(earth)
```

```
e <- earth(Tform, Tdata.train[1:1000, ])
```

```
e.preds <- predict(e, Tdata.train[1001:2000, ])
```

MARS只适用于回归问题，不可用于分类

三、从预测到决策

策略1

(1) t 时刻卖出信号

如果有头寸，信号被忽略


如果没有头寸，开一个空头头寸（价格 pr ），然后两个限价指令
一个买入指令 $pr-p\%$ ，一个买入指令 $pr+p\%$ ，用来止损。

(2) t 时刻买入信号

一个卖出指令 $pr+p\%$ ，一个卖出指令 $pr-p\%$ ，用来止损

策略2

只开一个头寸，等待收益达到预期，不设止损



与交易关联的评估准则

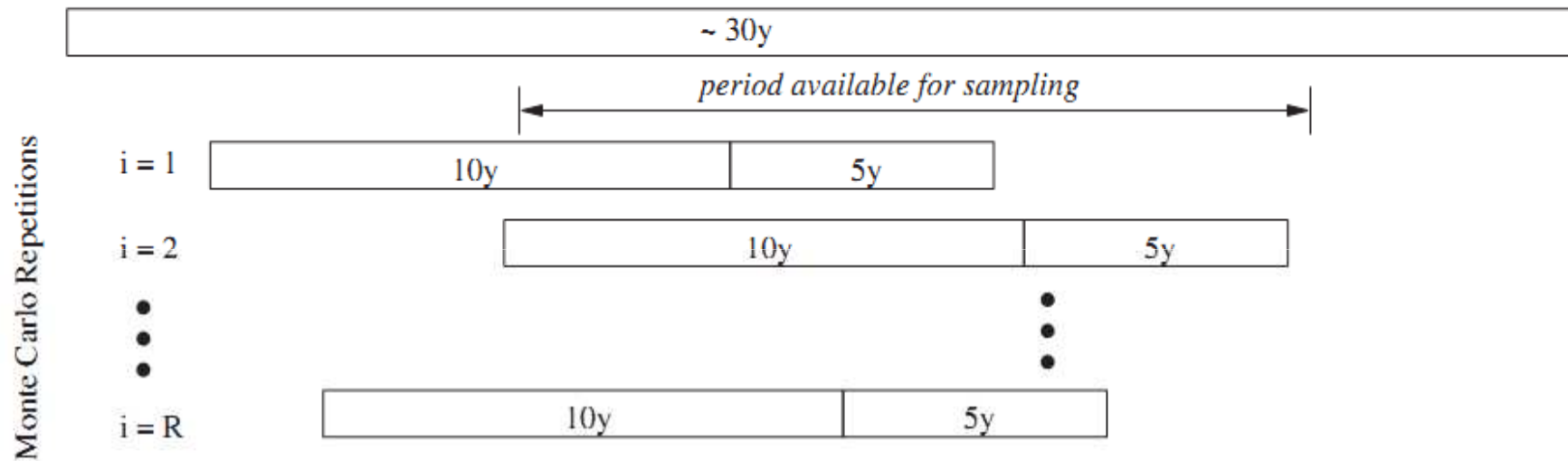
包PerformanceAnalytics

(1) 总体收益: Profit/Loss

(2) 风险相关的收益: Sharpe比率, 衡量单位风险的收益

四、模型评估和选择

1. Monte carlo模拟



2. 实验比较

3. 原因分析



参考文献

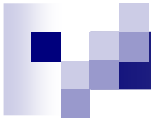
1. DataMining with R: Learning with Case Studies
2. 邓一硕整理的关于quantmod和PerformanceAnalytics的手册



Contact

yanping.chen@cos.name

<http://ypchen.inwake.com>



Thank you!