南京大学
NANJING UNIVERSITY

基于SVM的企业财务风险识别

141292018 桑梓洲

选题动机

- 企业财务危机直接影响到企业管理层、投资者、债权者、证券监管者和政府等相关者的利益，其预警研究一直是国内外学术界的重要研究课题…
- 真实动机则包含以下几个方面
  - 问题–对从股市上赚钱比较感兴趣
  - 方法–最近刚好在接触支持向量机
  - 实现–上市公司财务数据方便获得
- 特点：相对侧重技术处理而非金融
  - 优点：较容易迁移到风险管理等其他领域
  - 缺点：缺乏金融理论根基和问题研究深度

文献综述

- 预警模型：定性和定量
- 定量模型：传统统计方法和人工智能专家系统
- 传统统计方法
  - 单变量：Fitzpatrick(1932)
  - 多元线性判别(MDA)：Altman(1968)
  - 逻辑回归(LR)：Martin(1977) Ohlson(1980)
- 人工智能专家系统
  - 人工神经网络(ANN)：Odam(1990)
  - 遗传算法(GA)：Varetto(1998)
  - 支持向量机(SVM)：Fan(2000)

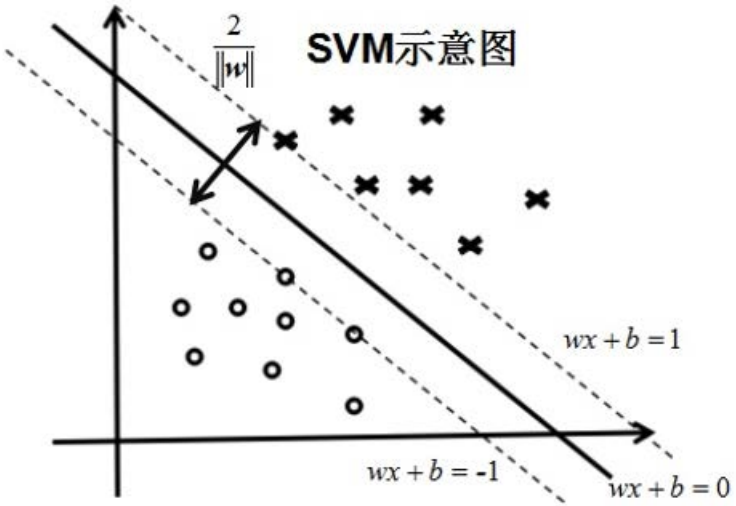
SVM摘要

- SVM两大类：分类机(SVC) | 回归机(SVR)
- 核心思想
  - SVC: 找到一组超平面，使两组样本间隔最大(分得最开)
  - SVR: 使所有样本落在平面$\pm\varepsilon$的弹性管道内(非线性回归)
  - 线性不可分$\mapsto$高维线性可分空间，核函数简化高维计算
  - 松弛变量和惩罚函数: 允许适当误差，防止过度拟合
  - 优化求解: 拉格朗日对偶问题与KKT条件(凸优化理论)
  - 优点: 效果最好*; 全局最优; 适合小样本、高维数据

*believed by many to be the best “off-the-shelf” supervised learning algorithm, Eric Xing, CMU

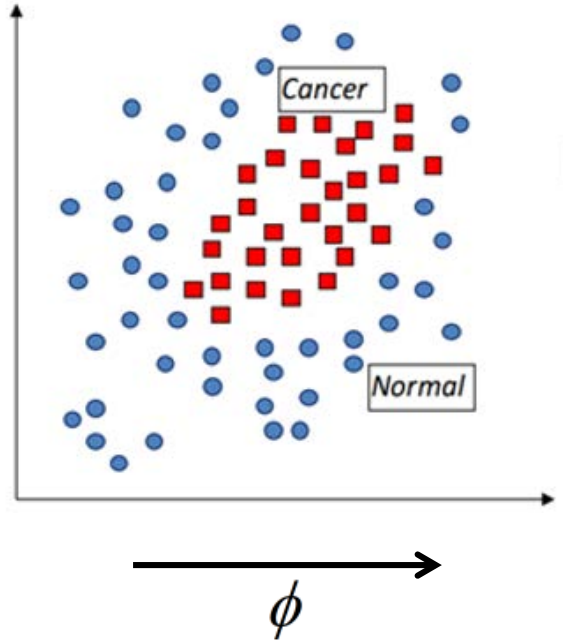
直观理解SVM

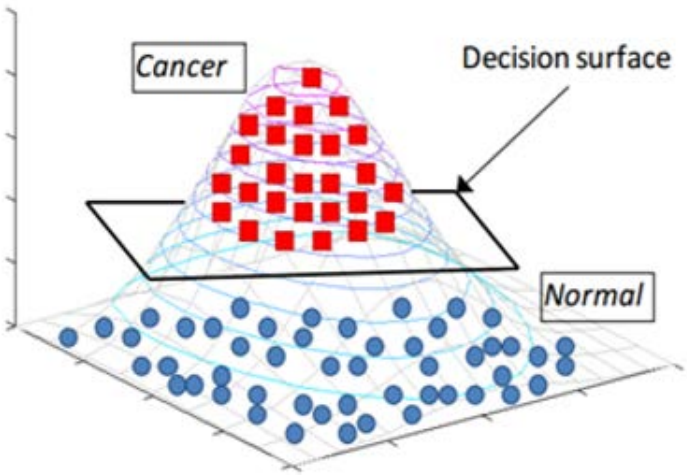
- SVM示意图



$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t.\, y_i(w^T x + b) \geq 1 - \xi_i, \xi_i \geq 0$$





实验构想及具体实施

- 实验构想: 基于SVM的问题公司识别
  - 选取偿债能力、资产结构、现金流量等股票指标若干
  - 将ST股票1年前和同时期正常公司样本作为训练集
  - 运用LibSVM进行训练和判别得到分类结果
  - 通过GA, PSO等算法包及random split方法进行参数调优
- 具体实施: 基于SVM的股票选取策略
  - ST股票非常非常少3年Wind数据难以进行训练
  - 目标变为如何通过SVM选取高收益率股票(涨跌幅)
  - 引入盈利估值等指标： ROE PE PB PCF等
  - 优点： 注重核心思想，短时间内适当简化工作量

Altman's Z Score

2014

1	证券代码	证券简称	Z值[报告期]2014年报
2	600145.SH	*ST新亿	-242.8060
3	600870.SH	厦华电子	-33.2546
4	002015.SZ	霞客环保	-8.9320
5	000155.SZ	*ST川化	-5.7090
6	000815.SZ	美利云	-1.0316
7	600707.SH	彩虹股份	-0.9154
8	600301.SH	ST南化	-0.8529
9	600242.SH	中昌数据	-0.7891
10	600163.SH	中闽能源	-0.7428
11	600291.SH	西水股份	-0.5874
12	600091.SH	ST明科	-0.3669
13	002306.SZ	*ST云网	-0.3226
14	600247.SH	ST成城	-0.3059
15	600793.SH	宜宾纸业	-0.2743
16	600644.SH	乐山电力	-0.2729
17	600725.SH	ST云维	-0.2538
18	000912.SZ	*ST天化	-0.2347
19	600319.SH	亚星化学	-0.1320
20	600722.SH	金牛化工	-0.1073

2015

1	证券代码	证券简称	Z值[报告期]2015年报
2	000155.SZ	川化股份	-9.9141
3	002608.SZ	江苏国信	-9.3170
4	600603.SH	广汇物流	-4.9747
5	600725.SH	ST云维	-2.4195
6	601005.SH	*ST重钢	-0.7655
7	600581.SH	八一钢铁	-0.7524
8	600339.SH	中油工程	-0.7009
9	000968.SZ	蓝焰控股	-0.5573
10	000717.SZ	韶钢松山	-0.5263
11	600319.SH	亚星化学	-0.5121
12	600117.SH	西宁特钢	-0.3782
13	600721.SH	百花村	-0.3122
14	600793.SH	宜宾纸业	-0.0471
15	600025.SH	华能水电	-0.0462
16	000932.SZ	*ST华菱	0.0011
17	600707.SH	彩虹股份	0.0194
18	000629.SZ	*ST钒钛	0.0484
19	000933.SZ	神火股份	0.1033
20	000959.SZ	首钢股份	0.1096

2016

1	证券代码	证券简称	Z值[报告期]去年年报
2	601558.SH	*ST锐电	-1.3529
3	601005.SH	*ST重钢	-1.2452
4	600815.SH	*ST厦工	-0.8953
5	000629.SZ	*ST钒钛	-0.7227
6	600707.SH	彩虹股份	-0.6777
7	600423.SH	*ST柳化	-0.4999
8	600432.SH	*ST吉恩	-0.3991
9	600397.SH	安源煤业	-0.2733
10	600871.SH	石化油服	-0.2402
11	000617.SZ	中油资本	-0.1592
12	600025.SH	华能水电	-0.1238
13	000755.SZ	*ST三维	-0.0733
14	600117.SH	西宁特钢	-0.0108
15	600877.SH	*ST嘉陵	0.0002
16	600691.SH	阳煤化工	0.0040
17	601918.SH	新集能源	0.0250
18	600793.SH	宜宾纸业	0.0437
19	600540.SH	*ST新赛	0.0980
20	000912.SZ	*ST天化	0.1155

数据获取与预处理

- Wind上获取14 15 16年全体A股财务数据
  - 估值 盈利 资本结构 偿债 营运 成长 Z值共17个指标
  - 选取沪市股票为实验对象, 收益率与大盘关系标注±1
  - Matlab进行归一化, 通过Java转成LibSVM格式

```
function [std_seq]=stdize(ori_seq,lbound,ubound)
std_seq=ori_seq;
left=quantile(ori_seq,0.05);
right=quantile(ori_seq,0.95);
```

```
A=xlsread('sh2014.xlsx');
B=zeros(size(A,1),size(A,2));
B(:,1)=A(:,1);
for i=2:size(A,2);
% the first column is label no standardization
    lbound=0;
    if min(A(:,i))<0
        lbound=-1;
    end
    B(:,i)=stdize(A(:,i),lbound,1);
end
```

LibSVM格式: 稀疏矩阵

```
label 1:value1 2:value2 N:valueN
```

最终处理结果

```
1 1:1 2:1 3:-0.28519
-1 1:1 2:-0.6109 3:-0.70615
-1 1:1 2:-0.30979 3:-0.79914
-1 1:1 2:-0.80589 3:-0.88579
-1 1:1 2:0.63288 3:-0.027944
1 1:1 2:0.28761 3:-0.58957 4
-1 1:1 2:-0.55191 3:-0.52066
```

遇到的bug及解决

-若LibSVM结果出现NaN说明数据中有NaN

SVM运行

南京大学 NANJING UNIVERSITY

- 不进行参数调优效果极差
  - 改进思路：14用作训练→15参数调优→16最终效果
  - 因训练集不用于调优，只能自己编写参数检验程序
  - 程序编写比较笨拙非常初级仅限于能用(半自动)

- 主要参数说明
  - -s: SVM 类型, 0为C-SVC, 3为ε-SVR (1号ν-SVC与0类似)
  - -t: 核函数类型, 0为线性, 1为多项式, 2为RBF
  - -c: cost, 惩罚系数, 越大则对误差要求越严格
  - -g: gamma, 核函数系数(核函数形式自带文档中给出)
  - -p: ε-SVR中的ε值, 越大则允许的回归管道宽度越大
  - -v: 交叉验证次数, 随机分成n份交互检验Acc/MSE
  - 填错或不填均按默认值处理，因此不必担心崩溃

```
for c=1:100:1000;
    for g=10:10:100;
        cmd=['-c ',num2str(c),' -g ',num2str(g)];
        % watch out for space!!
        model=svmtrain(oy,ox,cmd);
        [result,acc,~]=svmpredict(ty,tx,model);
        result(length(chg))=-1;
        % because java program always 1 line miss
        num=length(find(result==1));
        result(find(result==-1))=0;
        % only use +1 stock
        ret=result'*chg/num;
        % chg is the stock change sequence
        resultMatrix(k,:)=[c g acc(1) ret];
        k=k+1;
    end
end
```

中间结果展示

- 实验步骤及运行结果如下
  - 返回 c g 分类精度 收益率
  - 手动输入→程序批量检验
  - 参数先粗调再精调
  - Excel排序选出理想参数

c	g	Accuracy(%)	Return(%)
10	0.8	23.55	-7.64
100	0.8	26.80	-8.51
1000	0.8	27.29	-9.47
10	10	4.73	10.92
100	10	3.35	5.46
1000	10	3.44	5.46

14数据预测15-纯天然手工

c	g	Accuracy(%)	Return(%)	Acc*Ret
1	5	7.00	8.09	56.60
1	7	3.74	19.07	71.41
1	9	2.46	16.27	40.07
1	11	1.87	16.84	31.53
1	13	1.38	12.51	17.25
1	15	0.89	23.52	20.85
4	5	15.57	4.44	69.11
4	7	9.46	9.32	88.11
4	9	6.11	12.09	73.85
4	11	4.14	11.82	48.91
4	13	2.56	11.90	30.47
4	15	1.97	12.62	24.87
7	5	15.86	3.60	57.14
7	7	9.46	9.32	88.11
7	9	6.01	12.21	73.40
7	11	4.14	11.82	48.91
7	13	2.56	11.90	30.47
7	15	1.97	12.62	24.87

14数据预测15-半机械化

实验主要结论

南京大学 NANJING UNIVERSITY

一些结论

- c大选出股票数量多收益率低，g大股票少收益率高(相对)
- 最优g在5~9左右，c影响相对小些
- 14预测15的结果与14预测16的结果具有强相关性
- 想选出最高收益股票c在1以内较合适

c	g	Accuracy(%)	Return(%)
0.8	5	3.94	25.32
0.6	7	1.48	22.78
0.6	9	1.48	22.78
1	7	3.74	19.07
0.8	9	1.67	18.10
0.6	11	1.18	17.42
1	11	1.87	16.84
1	9	2.46	16.27
0.8	11	1.38	16.04
1.2	11	2.66	15.38

15.12.22-17.12.22上证收益率 -8.93%

c	g	Accuracy(%)	Return(%)
0.6	9	0.37	53.82
0.6	11	0.18	47.17
0.6	7	1.01	40.01
0.8	11	0.64	33.50
0.8	9	0.82	33.29
1	11	0.82	26.62
0.6	5	2.10	20.39
0.8	7	1.92	18.44
0.8	5	3.66	15.27
0.6	3	4.48	13.94

16.12.22-17.12.22上证收益率 6.25%

贡献不足与改进空间

- 不足及改进空间
  - 因数据时间关系只能用涨跌幅代替风险指标→偏题
  - 只用了上证数据而没来得及去做其他板块→缺乏可信性
  - 未来需找到更合适的风险指标(求老师指点)
  - 或许应分行业以实现可能更优的预测效果
- 贡献
  - 一次有趣的探索
- 参考文献

  - 彭静. *网络环境中企业财务危机预警研究*. Diss. 上海交通大学, 2008.
  - 施锡铨, and 邹新月. "典型判别分析在企业信用风险评估中的应用." *财经研究* 27.10(2001):53-57.

南京大学
NANJING UNIVERSITY

谢谢大家！
新年快乐~