

実例-データマイニングによる”知識発見”のプロセス

VMStudio を用いて株価予測に挑戦!!

「データマイニングの宝箱」 石川 慎也

はじめに データマイニングにより知識を得ようとする場合、問題やデータにより柔軟な対応が求められるため画一的にマニュアルを作成することは困難であります。ここではケーススタディとして、Web サイト「データマイニングの宝箱」に読者から寄せられた”データマイニングで株価予測は可能か?”という要望に対する個人的な取組みを紹介しします。

株の世界を全く知らない素人が、VMStudio によってデータの中から知識を発見し、バブル崩壊の相場でも十分成果が出せるという検証結果のもと身銭を投じ、現在までの2年8ヶ月で平均月利+8.85%(年利+177%相当)で月別勝率87.5%(28勝4敗)と安定して高いパフォーマンスを出しております。また、数百人が参加する全日本株式投資選手権主催の投資レースにおいて史上初の完全優勝を達成(優勝賞金100万円)したことから、相対的に優れた知識を発掘したと捉えることができます。

成功の最大要因は、知識発見過程におけるVMStudioの活用です。VMStudioにより柔軟かつイメージ通りのマイニングシステムが構築可能となり、様々な視点で分析検証を繰り返すことにより、有用な”株価予測システム”が実現いたしました。

1 テーマの決定 (2004/1/1)

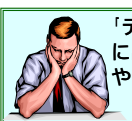
データマイニングのテーマは、本人の意思とは関係なく、漠然とまるで雲を掴むかのように降ってくる場合があります。上司から「この商品の良い販売戦略をみつめてくれ」やら「品質が悪いので原因をみつめてくれ」やら、どこから手をつけていいかわからない命令が下ることもしょなくありません。



読者「データマイニングで儲かる株を教えてください！」
わたし「……」
こうして、問題に立ち向かうのであった。

与えられるテーマは、無理難題とも思えるものも多く、マイニングを実行する人にとっては苦悩するところでもあります。ここでは「データマイニングをつかって、株で儲ける方法が発掘できるか」というテーマで取り組みます。

2 マイニングの方針決め (2004/1/2~)



「データマイニングはツールにかければ答えを勝手に出してくれる魔法の道具ではない、私は株なんかやったこともないのに、どうしろというのか……」

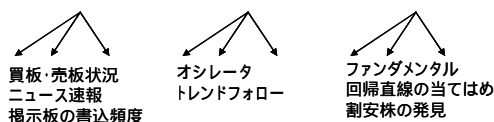
過度に期待される苦悩

マイニングの方針を決めるうえで、対象となる分野の専門知識はある程度必要になります。知識も経験もないデータマイナーは、資料をあつめ、過去の方法を検討するなりで、抽象的な目的を、明確な目的にしていく必要があります。



書店にならぶ株本のタイトルから……
「世の中には、株で大儲けするノウハウが溢れているぞ! これらを上手にマイニングすれば……」
しかし、ずいぶんうさんくさいなあ。

- 超短期予測(動きの観察とスピード判断が主体)
- 短期 予測(テクニカル分析が主体)
- 中長期予測(ファンダメンタル分析が主体)



目的を明確にし、結果につながるデータさえ準備できれば、株価予測でさえも解ける問題になります。

期待する結果が得られないということは……

- ・元のサンプルが少ない
- ・特殊な条件で集められた偏ったデータでマイニング
- ・異常なデータに影響を受ける
- ・解析のパラメータが適切でない
- ・結果に影響するデータが採取できない
- ・採取したデータだけで判断しようとしたがうまくいかない(視野が狭い)

- ・当たり前すぎるものしか抽出できない
- ・過剰に学習しすぎて実際は利用価値のない
- ・目標設定があいまいで曖昧な結果がでる
- ・マイニングの方向性が違う など

で「目的の設定」か「データ」の一方、あるいは両方に失敗要因が隠れています。

3 情報収集/既存手法の検証 (2004/1/4~)

期待する結果を得るためのデータが準備できるか落とし込んでいきます。まずは既存手法をベースに実現可能性を検証します。

超短期(秒・分単位)予測を目的として考えてみる

経験則に基づく。秒刻みのスピード判断が必要
そもそも、秒(分)刻みの過去データを集められない。
データがない。「データマイニングでは解けない問題」

中長期(数ヶ月・数年単位)予測を目的として考えてみる
集められるデータ

「会社の経営状態データ」「アナリストのレポート(格付)」「チャート」

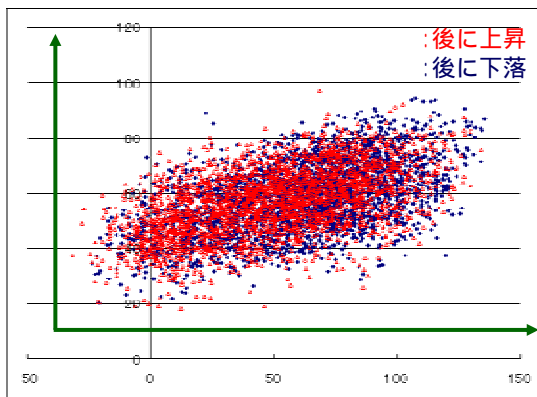
「現在の経営状況」と「未来の株価」を分析 因果関係はない。
「アナリスト予測」と「未来の株価」を分析検証 因果関係はない。
「チャートに回帰式を当てはめる(データ解析者)」 あてはまらない。

因果関係を見つけたぞ! 「来年の会社の経営状態」と「来年の株価」だ

未来の経営状態を予測できるのであれば株価を予測することは可能? これから起こり得ることにつながるデータが集められるか???

仮説による過程が多い。「解いたとしても信頼できない」

短期(日・週単位)予測を目的として考えてみる
データ : 日足四本値 + 出来高



なるほど! 分布に若干だが違いが見られるぞ。これを本格的に解析すれば面白いことが分かるかも?

短期予測で上昇株を発見する

4 目的を明確にするための予備調査 (2004/1/15~)

明確な目的を定める部分は、マイニングシステムが形になり始めると同時に、目的が決まった時点でシステムの価値がほぼ決定するともいえるところです。
辿り着いた明確な目的こそが、株価予測プロジェクト最大の知識であります。現在、実践で活用していることから、公開を控えさせていただきます。

予備調査 マイニングのタイプを決めるための予備調査



予備調査 明確な目的とするための予備調査

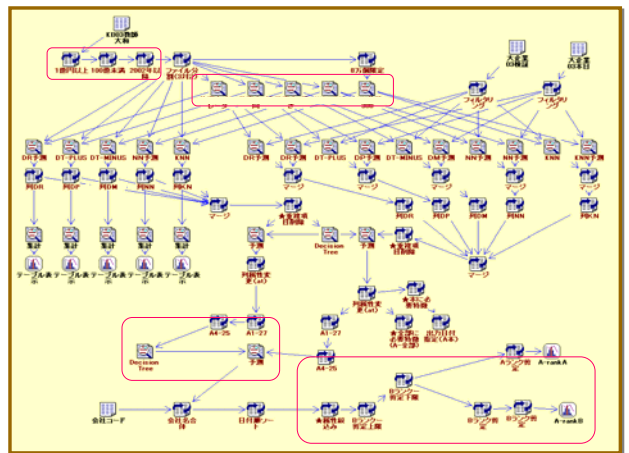
「目標の方向性は決まった。目標の具体的基準の決定は、色々なケースを試行して評価して決めよう!
Ummm...こうすれば良くなるのだが、何故だろう?」

さらに具体的に明確に目標の設定を行わなければならないが、ここまでくると「トライ&チェック」。ただやみくもに泥にまみれるのではなく、解析結果の意味づけを考えながら試行していくことが重要であります。

「目的を明確にする」と同様に、「データを準備する」ことがデータマイニングにとって重要。
この2つを怠っている問題は解くことができない。

5 マイニングモデルの作成 (2004/2/10~)

予備調査の結果をもとにマイニングモデルの試作版を作成しました。(図は試作時のもの)



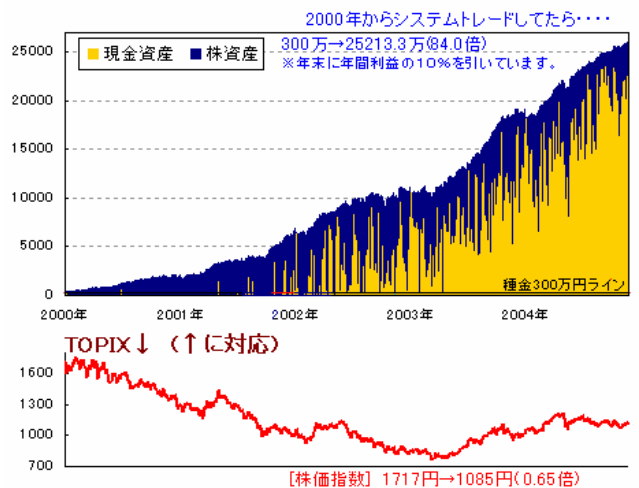
学習データ入力 検証データ入力 予測データ入力
学習データクリーニング部 予測モデル群
予測結果のとりまとめ 銘柄抽出 レポート

6 シミュレーション (2004/2/12~)

マイニングシステムの完成にともない、自分の投資コンセプトを踏まえ現実的に取引可能な環境となるよう計画を立て、シミュレーション(バックテスト)してみました。利益を追求するなら年間10倍にすることもできます。勝率を追求するなら勝率8割5分くらいにもできます。ですが、現実にお金を動かす人(自分)にとって耐えられないドラマが巻き起こりますし、相場の世界は自分の売買影響で値段も変わるわけですから、その影響をできるだけ小さくするようにシミュレーション設計しました。過去のデータを使って学習させたものを過去のデータで検証するわけですので過剰学習の影響をうけていること、その時代の背景をもって成り立っていることに注意を払わなければなりません。

ただし、マイニングモデル作成時から、汎用性があり、近い未来なら十分に対応できるように、色々な仕組みを折り込みました。シミュレーションは、目先の相場環境による影響が限定的であることを確認する意味もあります。

シミュレーション結果(バックテスト)



当時のモデルで2004年未までシミュレーション
(2004年1月以降はフォワードテスト)

評価尺度	算出方法	シミュレーション値
プロフィットファクター	総利益 ÷ 総損失 1以上になることが システムとして必須	2.167
最大ドローダウン	過去の運用資産最高 点に対する落ち込み 度で、最大のもの	15.50%
最大フラット期間	運用資産最高点の 更新にかかった期間 で、最長のもの	60日

学習モデルに含まれていない期間ということで、昔の年度でもシミュレーションしました。

運用資産の変化(すべて現物取引のみ)

1999年: 1000万円	6665万円
1998年: 1000万円	3943万円
1997年: 1000万円	1336万円
1996年: 1000万円	1945万円
1995年: 1000万円	2180万円
1994年: 1000万円	2118万円
1993年: 1000万円	2123万円
1992年: 1000万円	1514万円

細かいシミュレーションの設定を記述しておりませんが、未来のデータにも適用させてもよほど大きな問題が起きない限り破産の心配をせずに運用できるようにしております。

年度によりバラツキはあるものの1992年～2003年の12年間の検証において1年単位で見て年初の運用資産を割ったことはありません。バブル崩壊後で苦しい時期を含んでいるのにも関わらず、最低で年利30%、最高で年利500%くらいを叩き出しています。景気が1年を通じて横ばいなら年利300%になることがマイニングシステムのシミュレーション値です。(別途検証しておりますが運用資産1億円内に限ります)

時代が変わっても通用していることから、目的に沿って相場の本質に隠された知識がマイニングされていると考えます。



株の予測ってこんなに当たるのか？現実には色々と不都合も起きるが、やってみる価値はある(?)。
「データマイニングを使って株の予測ができる」という成功事例になるかもしれない。」

7 専門家との討論会 (2004/2/12～)

シミュレーションと平行して専門家を訪ね、ご意見をいただきました。「データ解析の専門家」と「意思決定者」と「対象分野の専門家」のコラボレーションは実行を視野にいれた知識発掘には重要な役割があります。

証券会社のディーラー

実に面白いことをするなあ。その解析目標に着目したことが素晴らしい。私も発想は同じかもしれない。

株式格付け担当者

なるほど！そんな見方があったのか！定量的な評価をしたことがなかった。

証券会社支店長

(具体的な話はできず) 予測できるわけがないケガしないうちにやめときな。今まで、失敗するのを何度もみてきたよ。

株の分析をやったことのある解析者

今は景気がいいから当たっているんだ。そのデータは信頼できない。また、実際に売買するとどうなるか試算されていない。

等。

想像通り、理解されない反応が多く、定量的な分析の視点で

の話し合いとはなりませんでした。実践で活動するにあたり運用面のアドバイスをいただきました。また、システムについてもいくつかアイデアもいただき有意義な議論となりました。

8 テスト運用による目標設定(2004/2/24～2004/4/30)

シミュレーションと現実にどれだけギャップがあり、どの程度利用できるのかを調査するために証券口座を開設し、現実の運用資金を用いてテスト運用を実施しました。

テスト運用して、あらためてはっきりしたこと

- 株(現物)は 失敗すればすべてなくなる といった代物ではなくて、1日で5%も変動すれば大きいほう。
- 数日間で20%も運用資産総額が動くことがあっても、無くなることは想定しなくていいということ
- 基本的にマイニングシステムがはじき出す通りの数値になる。等

テスト運用して、思った通りに行かなかったこと

- すぐ寄り付かないし、寄値で取引ができない。売ったお金で買い替える場合は反映を待たないといけない。
- 自分のお金で値が動く。取引量が少ない銘柄ほど。
- 精神が弱くてマイニングシステムの指示通りに売買できない。急落時は狼狽するし、暴騰時は舞い上がるし。
- 複数銘柄をできるだけ均等分散で買おうとしても、出来高や単元数の関係で大きくバラツク
- 相場の世界、独特の制度があった。入門者だけに知らないことだらけ。株式分割とかマーケットメイクとか色々。等

テスト運用の結果に基づいて本運用の目標を設定します。目標を設定することは、マイニングシステムが有効に機能しているか確認するために重要であり、システムを停止するための判断材料にもなります。

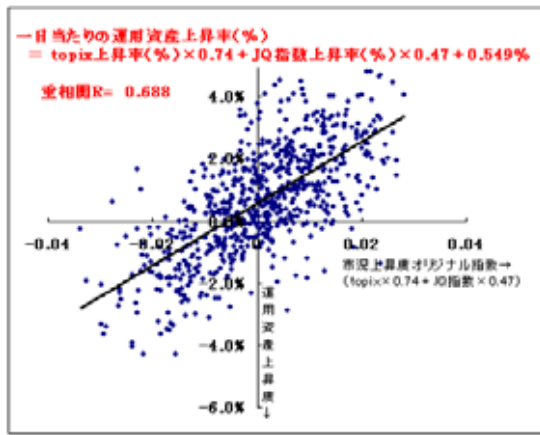
目標設定 : 年間の目標

シミュレーションでは、地合が横ばいなら年利300%程度が期待できることを算出しました。日利に直すと「平均利益 + 0.565% / 日(シミュレーション値)」ですが、計算通り行かない程度を考慮し、20%くらい引いて実取引では「平均利益 + 0.448% / 日(現実目標値)すなわち年利200%」を年間目標とします。

目標設定 : 1日あたりに期待できる利益

1年間の目標が経ったとしてもあくまでも1日1日の積み重ねが問われるシステムなので、毎日、的中具合を評価することが予測システムに求められます(ただし、1日1日で評価しても効果が見えにくい)。予測モデルで想定している取引スタイルは現物取引ですので、地合が良いときは利益も大きくいきますし、誰もが損しているような地合なら損失がでます。

景気指数を考慮したうえで、実際は毎日どの程度利益を上げられるか回帰分析を用いて考えてみます。(テスト期間はサンプル数が少ないため、シミュレーションに用いたデータを用いて重回帰分析を行いました)



この重回帰分析の結果では、「切片」が最も重要な意味をなしていると考えます。

切片から、TOPIX や JQ 指数といった市況が前日と変わらない状況なら1日当たり+0.549%が見込めます。また、損益分岐点を考えるには、x 軸との交点に着目し、市況(TOPIX、Jasdaq 指数)が-0.46%くらいの時であることがわかります。一日の市況指数と運用資産上昇率にはそれなりの相関関係があるため、市況から運用資産上昇率の理想値を出すことに意味があり、システムの精度評価にも利用できます。

グラフより市況の上昇率・下落率よりも、運用資産は平均的に1.2倍程度過剰に反応することがわかります。比較的値動きの大きい銘柄を選んでいることが関係しているようです。

以上のことを踏まえ、正常に機能しているかの判断材料にします。市況との分散具合も踏まえながら、一喜一憂し取り乱さないことにも有用な指標です。

計算値によると、運用資産上昇率は、一日あたりの平均が市況より+0.549%になります。目標設定 で記述したとおり計算値で売買できないため、+0.448%に落とし、「一日あたりに期待できる利益率 = TOPIX 上昇率 × 0.74 + Jasdaq 指数上昇率 × 0.47 + 0.448%」と定めます。

目標設定 : 運用停止の目安

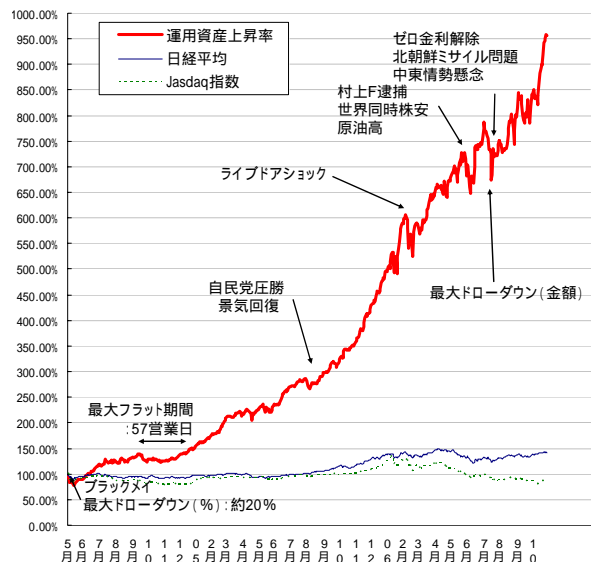
計算通りに機能しなくなったときは運用を停止するべきです。停止の判断ができることも本システムの大きなメリットです。

どんな状況に陥ったら停止させるのか・・・これまでの計算値から以下の状況であれば、想定外 & 未来のデータに通用しないシステムとみなし、運用を停止することにします。

- ・ドローダウン(運用資産の最高点からの下落率)
 - 15%で停止準備、20%停止。
 - (但し、市況が相応の暴落ならば25%まで容認する)
- ・フラット期間(運用資産の最高点の更新に要した期間)
 - 60営業日で停止準備、120営業日で停止。
- ・その他、理論値との大きなギャップが生じた場合。

9 本運用 (2004/5/1 ~ 2006/10/27 現在)

本運用開始と同時にブラックメイと呼ばれる大型の暴落に直面するという洗礼をうけましたが、その後2年半で運用資産を10倍に増やすことができました。年利+177%相当の実績は、設定した年間目標+200%の許容範囲であり、今なお高いレベルでシステムが機能しています。この間、延べ3682銘柄を購入し63.7%が上昇しました。月別の勝率は87.5%から安定していることもシステムの特筆すべき点であります。



	購入 点数	勝率	株資産 利回り /月	判定	[参考] 日経 上昇率 /月	[参考] JQ指数 上昇率 /月	備考
04年 3月	35	0.794	15.16%		6.10%	20.57%	テスト運用開始
4月	28	0.714	17.44%		0.40%	-17.85%	
5月	20	0.550	-10.95%	x	-4.47%	-10.48%	本運用開始
6月	61	0.850	32.37%		5.54%	9.57%	
7月	48	0.596	8.35%		-4.50%	-9.95%	
8月	57	0.582	2.86%		-2.15%	-1.19%	
9月	47	0.600	-1.50%		-2.33%	-6.37%	
10月	37	0.595	-2.62%		-0.48%	-3.37%	
11月	43	0.651	8.43%		1.19%	1.96%	
12月	63	0.694	12.71%		5.41%	7.90%	
05年 1月	97	0.755	12.54%		-0.88%	5.93%	
2月	122	0.802	19.25%		3.10%	-0.21%	
3月	138	0.594	6.16%		-0.61%	0.20%	
4月	107	0.552	2.78%		-5.66%	-0.10%	
5月	101	0.629	2.69%		2.43%	-2.45%	
6月	149	0.714	17.90%		2.73%	5.82%	
7月	137	0.639	5.35%		2.72%	2.31%	
8月	155	0.629	4.32%		4.32%	0.24%	
9月	122	0.678	7.95%		9.35%	1.12%	
10月	142	0.659	11.74%		0.24%	3.13%	
11月	118	0.786	21.81%		9.30%	5.77%	
12月	152	0.697	19.31%		8.33%	16.67%	
06年 1月	129	0.630	18.50%		3.34%	1.18%	
2月	120	0.518	0.49%		-2.67%	-8.47%	
3月	178	0.640	8.67%		5.27%	2.83%	
4月	194	0.565	4.94%		-0.90%	-6.28%	
5月	195	0.574	6.71%		-8.51%	-12.56%	
6月	182	0.561	6.45%		0.25%	-1.16%	
7月	163	0.590	-2.52%		-0.31%	-9.54%	
8月	206	0.618	8.13%		4.42%	2.82%	
9月	201	0.548	5.68%		-0.08%	-4.40%	
10月	135	0.692	16.57%		3.36%	-0.25%	
通算	3682	0.637	8.85%		1.84%	2.42%	
本運用の年間成績							
2004年通算	376	0.656	6.21%		-0.22%	-1.49%	
2005年通算	1540	0.677	10.98%		2.95%	3.20%	
2006年通算	1703	0.592	7.36%		0.42%	-3.58%	

まとめ

専門外の無理難題とも思えるマイニングテーマに取り組む約3年間の実績を紹介いたしました。個人の、極めて主観的な発想のもとでプロジェクトを遂行いたしました。未熟な点ばかりでございますが、ご参考いただける部分があれば幸いです。

得られた知識は「株価予測」以外にも多数あり、例えばニュースが与える株価の影響具合、投資家の銘柄を選ぶ基準のランク、経営情報と株価の関係など、面白い副産物が色々とできました。たとえ最終的に利益の得られるシステムが構築できなかつても、充実したマイニングができたことを確信しています。

最後に、VMStudioという素晴らしいツールに感謝いたします。今後どのような結末を迎えるか自分でもわかりませんが、素晴らしい知識を掘り当てた時の、あの瞬間を、あの喜びを求め、VMStudioと共に日々マイニングライフを堪能しております。

実践すると、コンピュータ上では見えない苦悩や、シミュレーション通りに実行できないこと、コンピュータが誤ってデータとして取り組んでいることが実に多い。そういうことを次は活かしていかないと。……さて、この後どうなるんでしょうか？

