社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－MPS－62 (4)
2006－BIO－7 (4)
2006／12／21

バンプ探索における解の精度

行實 隆広† 廣瀬 英雄‡ 大井 伸哉† 宮野 英次‡

†‡九州工業大学情報工学部 〒820-8502 福岡県飯塚市川津 680-4

E-mail: †{yukizane, ohi}@ume98.ces.kyutech.ac.jp, ‡{hirose, miyano}@ ces.kyutech.ac.jp

要旨 多次元空間内に特徴量を持ち２値(0/1)反応をとるN個の点の中から反応1を示す点が他に比べて密な領域（バンプ，ホットスポット）を探索する問題を考える．これまでに，探索結果を予測に使いやすくするためには決定木を用いたバンプ探索法が有効であり，また最適値を求めるためには確率的探索法（GA）に加え，極値統計を用いる方法を新しく提案した．しかし，得られた結果がどのような精度を持っているかはまだ分かっていなかった．ここでは，学習データとテストデータを使うことで，この問題がいかに深刻であるかを指摘し，次に最適に探索された結果の精度について述べる．テストサンプル法とブートストラップのメリットを併せ持つ方法も提案する．

キーワード 決定木，遺伝的アルゴリズム，交差検証法，ブートストラップ法，ブートストラップ的テストサンプル法，バイアス

Accuracy of the Solution in the Bump Hunting

Takashi YUKIZANE† Hideo HIROSE‡ Shin-ya OHI† and Eiji MIYANO‡

†‡Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology
680-4 Kawazu Iizuka, Fukuoka, 820-8502 Japan

E-mail: †{yukizane, ohi}@ume98.ces.kyutech.ac.jp, ‡{hirose, miyano}@ ces.kyutech.ac.jp

Abstract Suppose that we are interested in searching for denser regions showing response 1 with many feature variables in a z-dimensional space, where each point is assigned response 1 or response 0 as its target value; such a region is called the bump or the hot-spot. In a series of the previous study, we have shown that the bump hunting using the decision tree is useful in the ease-of-use and the prediction capability view points, and have developed a new bump hunting method using probabilistic (GA) and statistical (extreme-value statistics) methods. However, the accuracy of the estimated maximum capture rate was assessed by using the simple bootstrap method without correction formula. We have not thought seriously of the bias and the variance to the predicted estimate; we are, however, now aware of that we should treat the value of the predicted estimate very carefully. Thus, we have proposed a new method to assess the prediction error in the bump hunting problem, where the test sample method and the bootstrap method are nicely combined.

Keyword decision tree, genetic algorithm, cross-validation, bootstrap, bootstrapped test sample method, bias.

1. まえがき

多量の顧客データを有するデータベースから優良顧客の特徴を取り出し，顧客の将来の行動予測を行うような問題を考える．このようなデータ[11]は図 1 に示すように実際には優良顧客（反応 1）とそうでない顧客（反応 0）とが雑然と混在しているため単純な分類方法では解決できないことが多い．そこで，次善の策として優良顧客が比較的密集している部分の特徴量を探し出すことを考える[4]．このようにデータベースなどから興味の対象ができるだけ密に存在している領域を探し出すことをバンプ探索という[2,3]．

将来の行動予測のためには，例えば「女性で

ある」とか「一度に購入する商品の個数」とかルールとして使いやすい形が望ましいので，比較的単純なルールで明示的に表すことのできるアルゴリズムを利用できることが好ましい．そこで，決定木を使う．今，通常の決定木アルゴリズムを用いてでき上がった木において，上から順に下りてきて，あるノードで反応 1 に占める割合（これを純度と呼ぶ）が一定値 p 以上であれば，そのノード（つまりルール）を採用することにして，これらのノードにおける反応 1 の総数 m（これを獲得数と呼ぶ）を集計する．反応 1 をとるデータ総数 M に占める獲得数の割合を獲得率 c と呼ぶとき，p と c との間にはトレードオフの関係が生まれ，顧客獲得の戦略にはこのトレードオフを表す最適な曲線を知ることが重要となる[5,6].

通常の決定木アルゴリズムでは，特徴量の選定と最適分岐点とを木の上位から順次決定していき，適切な剪定を行うことで最終的な木を作るが，上のトレードオフ曲線を最適に求めるためには，この通常のアルゴリズムでは不十分であり，特徴量の選定と最適分岐点とについて全ての組み合わせを行わなければ得られない[5]．この場合，計算量の爆発が起こるため，筆者らはまず遺伝的アルゴリズムのような確率的アルゴリズムを使って近似解を求めることを試みた[6]．しかし，木に遺伝的アルゴリズム（以下 GA）を適用させると初期値依存度が高くなることがあり，このため極値統計を用いて最適値の予測を行ってトレードオフの最適曲線を得る方法を提案した[7,12]．また，特徴量の選定には GA を用いるにしても分岐点の最適選択までをそれに頼るには計算量が大きすぎるため，それには Gini の方法を用いても可能であることを示している[8].

その際，推定値の信頼度はブートストラップ法を用いて求めていたが，対象となるデータに最適にフィットした木に依存したブートストラップ法であるため，学習データだけを使ったことになり，過大評価を行っている可能性がある[1,9,10]．そこで今回，顧客データのような分類困難なデータの場合，このことによる影響は小さくないことをまず再確認し，推定値の精度を正しく把握するためのいくつかの方法について調べた．ここでは，ブートストラップとテストサンプル法とを組み合わせた方法を提案している．

2. テストデータを用いたナイーブな方法によるトレードオフ(p,m)

ここではまず，学習データだけによるトレードオフ曲線と，想定したテストデータによるそれとの違いを示す．一般的にその差を議論するのは難しいため，[7,12]で用いた実際の顧客データを模擬した仮想データを再度用いることとする．例えば，p=0.45 をあらかじめ指定したとき，データ総数 N=1000 で，M=200 の場合，20 個の初期値を用いた場合の GA の最適値には獲得数 138 が得られ，極値統計を用いた最適値の推定値には 144 が得られていた．

実際のデータではテストデータは得られないが，ここでは背後の確率分布を指定して仮想データを乱数により生成しているので，M=200 のテストデータを作成することができる．そこで，学習データから得られたバンプ獲得のために得られた最適なルールに，このようにして作られたテストデータを 10 ケース適用して獲得数を調べてみた．その結果，113，97，104，94，73，1，86，0，95，77 を得た．最も多い場合で 113，少ない場合で 0，平均は 74.0 である．獲得数を数え上げる方法は，学習データのときに使用した方法と全く同じである．つまり，学習データの場合に純度が p 以上となったノードにおいて，テストデータの場合でもそこでの純度が p 以上となったノードでの獲得数をすべて数え上げている．これをナイーブな方法と呼ぶ．平均 74.0 は学習データの場合に得られた 138 と比較してかなり大きく低減した値である．テストデータによっては獲得数がわずか 0 となる悲惨な結果となることもあり，将来予測のためのルールとしては全く使い物にならないということになる．GA を行う場合に 20 ケースの初期値から出発して最終値を得るまでに生存した 20 ケースを記録しているので，それらのケースのルールを用いて同様なことを行っても同じような結果が得られたので，ナイーブな方法を用いることは破綻している．

3. テストデータを用いた緩和法によるトレードオフ(p,m)

学習データから得られた最適ルールにおける純度 p 以上を示すノードにテストデータを適用した場合のノード毎の純度を調べてみると，p をわずかに下回るノードが見られる．ナイーブ

な方法による場合，このノードの獲得数は0とカウントされるので，学習データの場合のこのノードの反応1のデータ数が多ければ，テストデータによる獲得率が大きく下がる原因になる．しかし，そこでの純度はpからそれほど離れていないこともある．例えば，先のケースの場合，純度p以上を示すノードでの学習データ（lと表記）の場合の重み付き平均純度p_lを計算すると0.48であり，各ノードで0.45以上の純度を満たすナイーブな方法による獲得数m_lは113となっているが，テストデータ（tと表記）の場合，学習データから得られた純度p以上を示すノードすべてのノードでの獲得数m_tを求めると122となり，この結果重み付き平均純度p_tは0.48となりp=0.45からあまり離れていない．このように学習データから得られた純度p以上を示すノードすべての反応1の獲得数をカウントする方法を緩和法と呼ぶことにする．図2に，20ケースのGA最終ケースによるルールを10ケースのテストデータに適用した場合の緩和法から得られたトレードオフ（p_t, m_t）を，トレーニングデータを用いた場合のトレードオフ（p_l, m_l）と併せて示す．ここではp=0.45, 0.5, 0.6, 0.7に設定している．図から，テストデータの獲得数は学習データからの獲得数よりも小さくなる傾向はあるものの，ナイーブな方法のときに見られた極端な獲得数の落ち込みは確認されない．以降，この緩和法を用いる．

4. バイアス補正法

緩和法によっても学習データによる獲得率とテストデータによるそれとの間にはある程度のバイアスが発生することが分かった．実データの場合，上に示したような背後の確率分布は分かっていないので乱数データを作成することができず，従って豊富なテストデータが用意されている訳ではない．手元にあるデータをもとにして，学習データとトレーニングデータを分けて使うか，あるいは，手元のデータ数が少ない場合何らかのバイアス補正法が必要となる．ここではこのバイアス補正法について述べる．

ブートストラップ法では元のデータをリサンプルする場合63.2%のデータを重複して見ることになり，テストデータによる結果は学習データの結果とあまり変わらない[1]．つまり，甘い結果が出る．筆者らが信頼度の計算に用いた方法はあまり勧められないことになる．ただ，いくつものケースについて推定値を計算できるので信頼度の計算には向いている．

元データを学習データとテストデータに2分するテストサンプル法では，両データは全く異なるので，比較的辛めの結果が出る．ただし，推定値は出るがその信頼度は得られない．

交差検証法では，元データをいくつか（例えば10個）に分けて，一つをテストデータとして，残りを学習データとして使って，それぞれの推定値を求め，また信頼度も同時に得る方法であり，比較的良い結果が出るので良く用いられている．

ここで取り扱っている問題のように，分類しにくく，しかもN=1000, M=200のような大きさの場合には，テストサンプルのような方法は可能であるとしても，交差検証法を行うにはあまりにもテストデータの数が少なく実用的でない．そこで，筆者らは次のようにして獲得数のバイアスを推定した．

4.1. ブートストラップ的テストサンプル法

元データからその数だけのデータをリサンプルする．その半分のデータで学習データとしてルールを作り，残り半分をテストデータとして作られたルールにあてはめる．これを適当な回数繰り返す．それぞれに得られた推定値の2倍をもとにして平均や標準偏差を出し，最初に出していた元データすべてを学習データとして求めた獲得数との間のバイアスを求める．

図3はこのようにして求めた得られたトレードオフ（p_t,m^*_t）である．p=0.45のときの，学習データだけから求めたm_lの平均は138.6, m^*_tの平均は107.9であり，その差は背後分布を仮定して学習データとテストデータとで求めた獲得数の差とそれほど違いがないため，この方法が実用的であることを示している．

5. まとめ

実際の顧客データベースのように優良顧客を分類しにくいデータベースの中でその割合が高いと思われる領域とそこでの顧客獲得数を求めるバンプ探索法において，求められた優良顧客の割合pとその獲得数の割合cとのトレードオフの関係の推定精度について考察した．

将来のデータに対しての獲得顧客数を，データベースから得られた最適ルールから得られる

獲得顧客数からどの程度割り引いて考えればよいのかは，提案するブートストラップ的テストサンプル法を用いることで推定できることを示した．

文　献

[1] Efron, B. (1983), Estimating the error rate of a prediction rule: improvements in cross-validation, J. American Statist. Assoc., 78, pp.316-331.

[2] Friedman, J.H. and Fisher, N.I. (1999), Bump hunting in high-dimensional data, Statistics and Computing, 9, pp.123-143.

[3] Gray, J.B. and Fan, G (2003), Target: Tree analysis with randomly generated and evolved trees, Technical report, The University of Alabama

[4] Hirose, H. (2005a), A method to discriminate the minor groups from the major groups, 2005 Hawaii International Conference on Statistics, Mathematics and Related Fields.

[5] Hirose, H. (2005b), Optimal boundary finding method for the bumpy regions, IFORS2005 Triennial 2005 Conference, FD-19-3.

[6] Hirose, H., Yukizane,T. and Miyano, E. (2006a), The bump hunting method using the genetic algorithm and the extreme-value statistics with application to a messy customer database, 2006 Hawaii International Conference on Statistics, Mathematics and Related Fields.

[7] Yukizane, T., Ohi, S., Miyano, E. and Hirose, H. (2006), The bump hunting method using the genetic algorithm with the extreme-value statistics, IEICE Trans. Inf. & Syst., E89-D, pp.2332-2339 (Invited Paper from New Horizons in Computing).

[8] Hirose, H., Yukizane,T. and Miyano, E. (2006b), Boundary detection for bumps using the Gini's index in messy classification problems, CITSA 2006, pp.293-298.

[9] Kohavi, R. (1995), A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, IJCAI.

[10] Hastie, T., Tibshirani, R. and Friedman, J.H. (2001), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag.

[11] 小林，内尾，廣瀬: 顧客データベースのデータマイニング，日本計算機統計学会第16回大会論文集, pp.48-51 (平成 14.5.17).

[12] 廣瀬，行實，大井，宮野: Bump hunting 問題における極値統計の応用，日本計算機統計学会第 19 回シンポジウム論文集，pp.55-58 (平成 17.10.20)

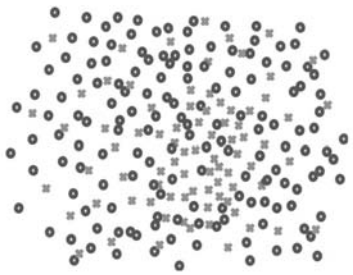

図 1　分類困難な 2 値データ

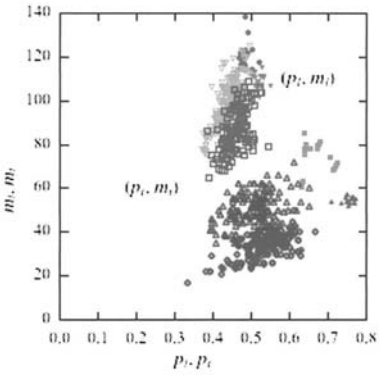

図 2　緩和法によるトレードオフ

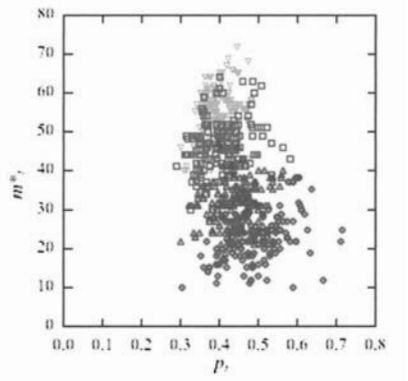

図 3　ブートストラップ的テストサンプル法によるトレードオフ