The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients

I-Cheng Yeh, Che-hui Lien

**Contexto/introdução**:

No Taiwan, os portadores de cartões de créditos utilizaram-nos (de forma descontrolada e independentemente das suas capacidades financeiras) excessivamente, acumulando grandes dívidas (dinheiro e crédito).

A previsão de risco é uma ferramenta dos sistemas financeiros de gestão de crise que utiliza a informação financeira para prever o “business performance” ou o **risco de crédito** (i.e., a probabilidade de haver atrasos no pagamento do crédito concedido) dos clientes individuais e para reduzir o dano e a incerteza. Estes modelos podem ser desenvolvidos recorrendo a métodos estatísticos (como a análise discriminativa, a regressão logística, a classificação de Bayes e “nearest neighbor”) ou recorrendo a classificações que associam inteligência artificial e *machine learning* (redes neurais artificiais e *classification tree*).

Problema: A aproximação entre a *probability of default* estimada produzida a partir dos métodos de *data mining* e representar a *probability of default* “real” (**que é desconhecida**).

**Objetivo do paper**: propor um método, o “***Sorting Smoothing Method***”, capaz de deduzir a *“probability of default*” real.

**Técnicas de *Data mining***

*Data mining* é o processo de análise e exploração, por meios automáticos ou semi-automáticos, de grandes quantidades de dados de modo a descobrir padrões e regras importantes. É importante em várias áreas como o *scoring* de crédito.

1. *K-nearest neighbor classifiers (KNN)*

Baseado na aprendizagem por analogia. Ao atribuir uma amostra desconhecida, este classificador procura o padrão para o KNN mais próximo.

Principal vantagem: Não é necessário estabelecer um modelo predictivo antes de classificar.

Desvantagens: Não produz uma fórmula simples da classificação de probabilidade; A precisão da previsão é afetada pela distância medida e pela cardinalidade *k* da vizinhança.

1. *Logistic Regression (LR)*

*Fitted probability* do evento é uma função linear dos valores observados das variáveis explanatórias disponíveis.

Vantagem: Produção de uma fórmula simples de Classificação de probabilidade.

Desvantagem: Incapaz de lidar com a não linearidade e efeitos interativos das variáveis explanatórias.

1. *Discriminant Analysis (DA) ou Fisher’s rule*

Objetivo: aumentar a distância entre grupos diferentes e minimizar a distância dentro de cada grupo.

Baseia-se no seguinte: para uma cada classe de resposta variável, existe distribuição das variáveis explanatórias como uma distribuição normal multivariada (com uma matriz variância-covariância comum).

Vantagens e desvantagens: assemelham-se às de LR visto que DA é uma Alternativa ao LR.

1. *Naive Bayesian classifier (NB)*

Baseia-se na Teoria de Bayes: o efeito do valor de um atributo numa determinada classe é independente dos valores de outros atributos (- Independência de Classe Condicional).

Vantagem: fornecem justificações teóricas para outros classificadores (que não recorrem ao teorema de Bayes).

Desvantagem: a precisão de previsão está correlacionada com a assunção da Independência de Classe Condicional, que apesar de simplificar a computação nem sempre se verifica (pode existir dependência entre as variáveis).

1. *Artificial neural networks (ANNs)*

Utiliza Equações matemáticas não lineares para desenvolver relações importantes entre as variáveis de *input* e *output*.

Vantagem: permitem lidar facilmente com efeitos não lineares e interativos de variáveis explanatórias.

Desvantagem: não devolvem fórmulas simples de classificação probabilística.

1. *Classification trees (CTs)*

Simbolismo: Cada nó interno corresponde a um teste num atributo, cada ramo representa um resultado do teste e os nós das folhas representam as classes.

Aplicável: em variáveis de resposta qualitativa ou quantitativa discreta.

Objetivo: classificar as observações com base nas variáveis explanatórias e supervisão pela presença da variável de resposta.

O processo de segmentação ocorre utilizando uma variável explanatória de cada vez.

Baseia-se na minimização da impureza (medida de variabilidade dos valores de resposta das observações).

Vantagem: obtenção de regras de classificação simples; capaz de lidar com os efeitos não lineares e interativos das variáveis explanatórias.

Desvantagem: a natureza sequencial e complexidade algorítmica traduzem-se na dependência dos dados observados (E.g.: uma pequena alteração é capaz de gerar uma grande mudança estrutural na árvore); difícil de generalizar a estrutura de uma árvore a vários contextos.

**Aplicações das técnicas de *data mining* no score de crédito**

Decisões de interesse: extensão do crédito e qual o valor, quando iniciar a cobrança e que ação tomar.

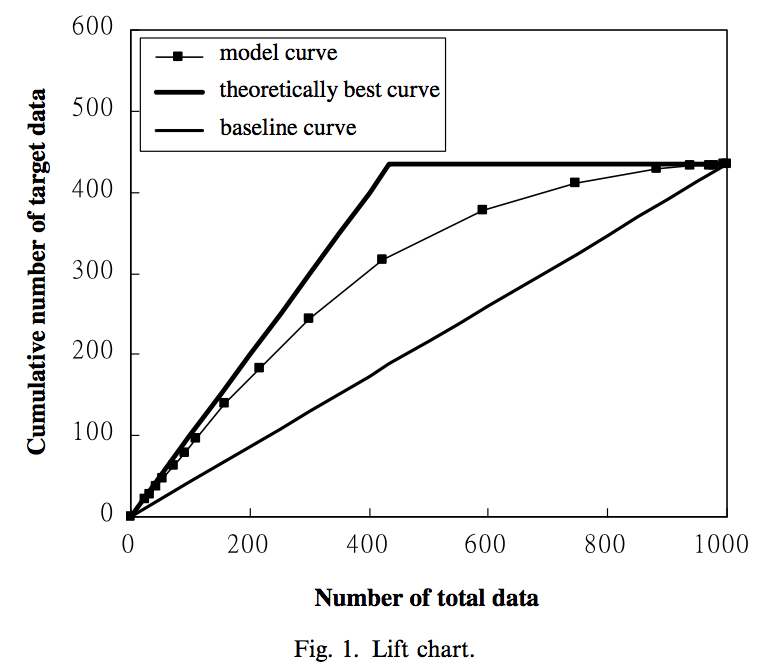
Credit scoring descreve os métodos estatísticos utilizados para classificar os candidatos ao crédito como risco “bom” ou “mau”.

Classificadores LS-SVM têm um desempenho muito bom em credit scoring. (Baesens *et al.*, 2003).

**Classificação da precisão entre as várias técnicas de *data mining***

**A variável de resposta é binária (Yes = 1, No = 0) e corresponde à falta de pagamento (*default payment*).**

Utilizaram 23 variáveis explanatórias (incluíam o género, a quantidade de crédito dada, educação, histórico de pagamentos, etc).

**Os dados foram divididos em dois grupos: um para treinar o modelo e o outro para o validar.** A taxa de erro não pode ser utilizada como medida da precisão da classificação do modelo porque é insensível à mesma. O rácio da área no *lift chart* permitem uma comparação mais adequada do desempenho dos diferentes modelos do que a taxa de erro. **Quanto maior a área entre a curva do modelo e a curva da linha de base, melhor o modelo.**

**Resultados**

KNN e CT têm a menor taxa de erro (supostamente significaria que possuem a maior eficácia). Além disso, CT tem o 2º maior rácio de área.

Conclusão

Na classificação da precisão entre as 6 técnicas verifica-se que existem poucas diferenças nas taxas de erro dos 6 métodos, mas que essas diferenças são grandes no rácio de área => rácio de área é mais sensível e um critério apropriado.

Em relação ao controlo de risco, é mais significativo estimar a *probability of default* do que classificar os clientes em *risky-non risky*.