



Curso de Engenharia de Sistemas e Informática - 5º Ano

Ficha T. Prática n.º 5

Objectivo:

Reflectir acerca dos dados utilizados em análise e predição com especial ênfase na problemática da amostragem.

Proposta de Trabalho

Numa idade em que o armazenamento de terabytes de dados é comum, a importância de redução de casos aumenta. Serão todos os casos que residem numa base de dados, necessários para um mining efectivo? A eficácia de grandes amostras de treino é dependente da aplicação, e o tamanho apropriado da amostra depende da complexidade dos conceitos que se pretendem extrair dos dados. Se os dados são todos ruído aleatório, então não se extrairá qualquer conceito útil dos dados, não importa qual o tamanho destes. Similarmente, quando o conceito é relativamente simples, os resultados dificilmente se alterarão mesmo com casos adicionais. Por outro lado, para algumas aplicações, os resultados continuam a melhorar com casos adicionais

Será o treino com todos os casos possíveis melhor do que utilizar uma amostra aleatória? Os dados suplementares devem proporcionar mais informação para minerar conceitos dos dados e, em teoria, os programas de predição não deverão fazer pior. Na prática, os resultados de treino em todos os casos serão quase óptimos quando comparados com subconjuntos aleatórios de tamanhos diversos. Não precisam de ser sempre melhores, podendo ser um pouco piores devido ao muito maior espaço de soluções que são examinadas, algumas das quais poderão parecer promissoras, mas falham em casos novos.

Aumentando o número de casos tende a aumentar a complexidade das soluções induzidas. Com maior número de casos, as estimativas de erro podem melhorar marginalmente, acompanhadas por grandes incrementos na complexidade da solução (lembramo-nos que a solução deverá responder a um número muito maior de excepções encontradas na amostragem maior). A maior complexidade poderá ser de interesse ou não. Se a aplicação é relativa a investimentos, maior complexidade será o preço a pagar pelo melhor desempenho predictivo. Mas, em outras aplicações (ex. conhecer os factores que influenciam as vendas), soluções complexas implicam menor interpretabilidade da solução. Soluções menos complexas tendem a ter menor variação no desempenho em novos casos.

Embora respostas quase óptimas sejam atingidas tipicamente através do treino com todos os casos, um número de razões foram citadas para a utilização de subconjuntos de todos os casos. A razão usual para a redução de casos é de importância crucial: o número de caso pode ser demasiado grande para processar num intervalo de tempo aceitável. Mesmo quando a exequibilidade está assegurada, a obtenção de um instantâneo rápido do desempenho potencial é usualmente preferido à espera de respostas do treino em todos os casos.

A abordagem principal à redução de casos é a amostragem aleatória. Em vez de minerar todos os casos, são retiradas amostras aleatórias. A forma de obter as amostras deverá

assegurar a sua não polarização (ou enviesamento), através de técnicas apropriadas.

A utilização de técnicas de mineração de dados nas amostras poderá, por outro lado conhecer duas formas denominadas de amostragens incrementais e de média.

Na primeira técnica são mineradas amostras sucessivamente maiores e verificados, de cada vez, qual o ganho em desempenho e aumento de complexidade, até se verificarem aumentos desprezíveis de um face a grandes incrementos de outra. Mesmo uma análise de complexidade versus erro é muitas vezes interessante.

Quanto à segunda técnica, são utilizadas N amostragens que são mineradas, sendo a solução a média ou voto entre as diversas soluções obtidas.

Como se comporta a solução média assim calculada face à solução única provinda de todos os casos? A resposta pode ser surpreendente. Estas soluções têm normalmente menos erros do que as providas de todos os casos.

*Adaptado de Predictive Data Mining: a practical guide
Sholomon M. Weiss e Nitin Indurkha*

1. Porque julga que a complexidade do conceito procurado determinará a dimensão dos dados?
2. Porque é que, em alguns casos, a solução extraída de um número maior de casos tem um menor desempenho do que em amostras maiores?
3. O texto, parece revelar uma relação directa entre a complexidade da solução e a dimensão da amostra e fazer uma avaliação do binómio complexidade versus desempenho. Comente.
4. Qual a razão primeira da amostragem?
5. Sugira possíveis soluções para a não amostragem. Serão desejáveis em todas as situações?
6. Fala-se em polarização de amostragem. Em que consiste e como evitá-la?
7. No final do texto, referem-se duas formas de mining em amostragens: incremental e de média. Parece surpreendente que especialmente a segunda conduza a melhores resultados, face ao obtido em mineração de todos os casos. Tente encontrar uma razão para o facto.