Análise de técnicas de selecção de atributos em Bioinformática

Rui Mendes
100378011

Bioinformática 10/11 DCC

Artigo Base

Yvan Saeys, Inaki Inza and Pedro Larranaga . “A review of feature selection techniques in bioinformatics ”. *Bioinformatics (2007) 23(19): 2507-2517 first published online*

Técnicas de selecção de atributos

- Seleccionam um subconjunto dum conjunto inicial de variáveis
- Preservam a semântica original das variáveis
- Maior facilidade de interpretação

Técnicas de selecção de atributos - objectivos

- Melhorar desempenho dum modelo de classificação
- Melhorar a detecção de clusters
- Obter um maior conhecimento sobre os processo que geram os dados

---

- Desvantagem ☹:

  Camada adicional de complexidade

Técnicas de selecção de atributos - tipos

- Filtragem
- “Wrapper”
- Embebidas

Técnicas de Filtragem

- Avaliam a relevância dos atributos tendo em conta apenas as propriedades inerentes dos dados
- Removem os atributos de baixa relevância
- Subconjunto de atributos é dado como input a um algoritmo de classificação

Técnicas de Filtragem - vantagens

- Grande escalabilidade
- Computacionalmente simples e rápidas
- Independentes do algoritmo de classificação

Técnicas de Filtragem - desvantagens

- Ignoram interacção com o classificador
- Grande parte são univariadas

Técnicas “Wrapper”

- Incorporam a pesquisa de hipóteses de modelos de classificação
- Gerem vários subgrupos de atributos e avaliam os mesmos
- A avaliação é obtida formando e testando um modelo de classificação concreto

Técnicas “Wrapper”- vantagens

- Interacção entre escolha de atributos e modelo de classificação
- Permite-nos ter em conta as dependências dos atributos na escolha do classificador

Técnicas “Wrapper”- desvantagens

- Maior risco de “overfitting”
- Muito intensivas computacionalmente

Técnicas Embebidas

- A procura de um subgrupo óptimo é feita dentro da construção de um classificador
- Pesquisa no espaço combinado de subgrupos de atributos e hipóteses
- Vantagens:
  - Incluem interacção com o classificador
  - Menos intensivos computacionalmente que os métodos “wrapper”

Aplicações na Bioinformática

- Análise de Sequências
- Análise de Microarranjos
- Análise de espectrometria de massa

Análise de sequências – Porquê selecção de atributos?

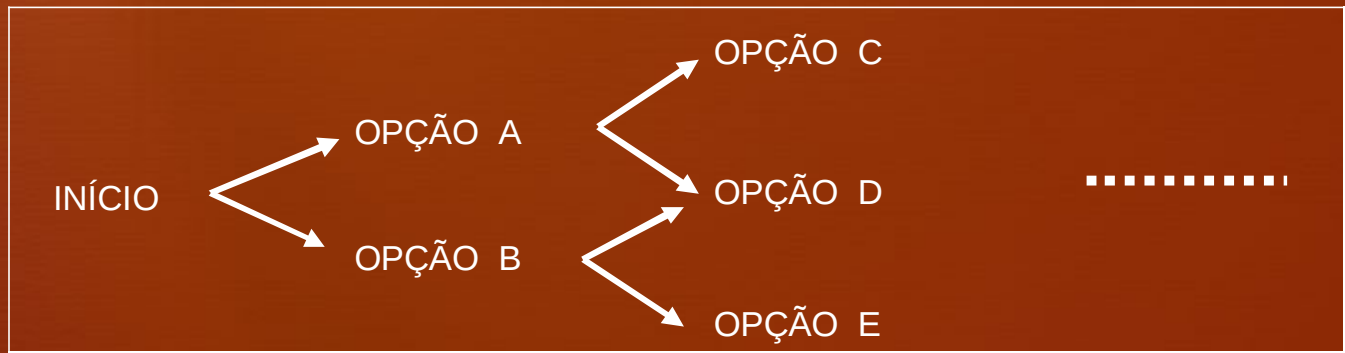
- Número enorme de atributos
- Muitos deles podem ser irrelevantes ou redundantes
- Necessário a criação de um subgrupo de atributos realmente relevantes

Análise de sequências

- Análise de conteúdo:
  - Foca-se nas características gerais de uma sequência
  - Exemplo: Tendência para codificar proteínas; execução de uma determinada função biológica
- Análise de sinal:
  - Foca-se na identificação de motivos importantes na sequência
  - Exemplo: elementos estruturais do gene; ou elementos reguladores

Análise de sequências – exemplos

- *Interpolated Markov Model (IMM)*
  - Usa a interpolação entre diferentes ordens do modelo de Markov para tratar pequenas amostras
  - Usa um método de filtragem para seleccionar apenas os atributos relevantes
- *Interpolated Context Model (ICM)*
  - Atravessa uma árvore de decisão Bayesiana
  - Juntamente com um método de filtragem que avalia a relevância do atributo



Análise de Microarranjos – o que é?

“ Procurar um gene dentro de um genoma humano é comparável a procurar uma pessoa sem sobrenome numa casa sem endereço numa rua desconhecida numa cidade não identificada de um país estrangeiro.”

Solange Farah

Análise de Microarranjos – o que é?

- Arranjos que contêm um grande número de genes roboticamente distribuídos
- Permite-nos comparar expressão de um grande número de genes em simultâneo
- Podem chegar aos 400.000 pontos
- Exemplos de utilização:
  - Detecção de mutações genéticas
  - Sequenciação de ADN
  - Genotipagem

Técnica de clonagem de genes em laboratório da Unesp

Análise de Microarranjos – porquê selecção de atributos?

- Grande dimensão de genes
- Pequenos tamanhos das amostras
- Necessário técnicas de selecção de atributos para reduzir a dimensão dos dados

Análise de Microarranjos

- Domínio das técnicas de filtragem univariadas
  - Output intuitivo e fácil de entender
  - Desconhecimento da existência de técnicas de análises de dados para seleccionar genes duma forma multivariada
  - Tempo de computação extra das técnicas multivariadas

Análise de Microarranjos – exemplos

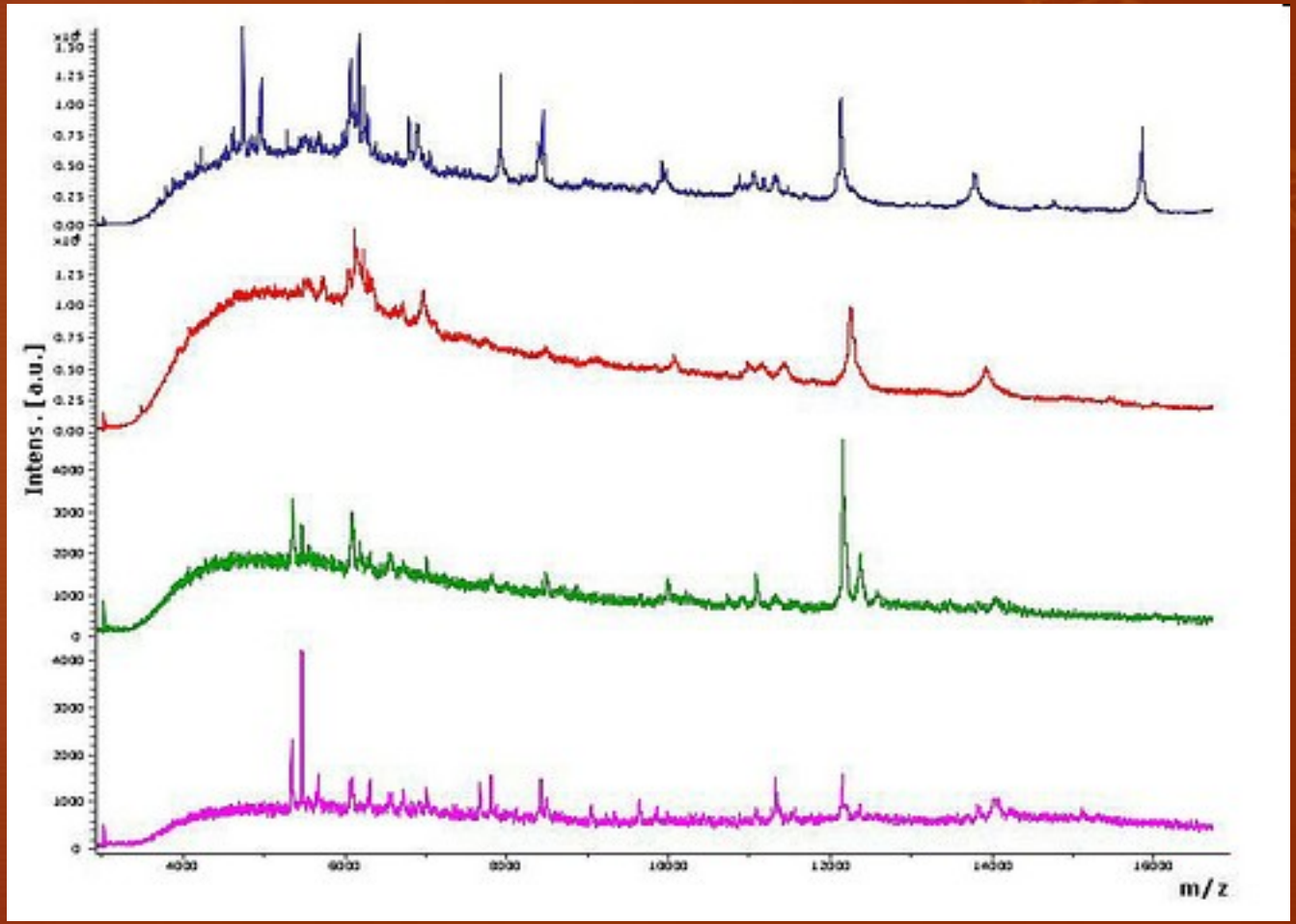
- *Thresholding*
  - Colocar um limite nas diferenças entre estados já observadas anteriormente
  - Detecção de um ponto de threshold que minimiza o número de más classificações

- *Métodos paramétricos*
  - Assumem uma distribuição dada a partir da qual os modelos foram gerados
  - Os modelos *t-test* e ANOVA são os mais utilizados

Análise de espectrometria de massa – o que é?

- Um espectrómetro de massa bombardeia uma substância com electrões para produzir iões
- Os íons atravessam um campo magnético
- O campo separa os íons em um padrão chamado espectro de massa
- A massa e a carga dos iões podem ser medidas pela posição no espectro
- Cientistas identificam assim os elementos e isótipos presentes na amostra

Análise de espectrometria de massa – o que é?



Análise de espectrometria de massa – o que é?

- Permite diagnosticar doenças e criar perfis de biomarcadores baseados em proteínas
- Amostras contêm milhares de diferentes rácios massa/carga (m/z), com um valor de intensidade de sinal associado
- Um perfil de baixa resolução pode conter até 15 500 pontos de dados no espectro
- Número de pontos aumenta com uso de instrumentos de alta resolução

Análise de espectrometria de massa – exemplos

- Técnicas de filtragem univariadas são as mais utilizadas
  - Realizam, à priori, um procedimento de extracção agressivo usando como base detecção de picos e técnicas de alinhamento
  - Diminuem bastante a dimensão de dados
  - Classificadores começam, normalmente com menos de 500 variáveis

Análise de espectrometria de massa – exemplos

- Abordagem embebida
  - Classificadores descartam atributos como input
  - Variação de um método originalmente proposto para expressão genética
  - Usa os pesos das variáveis guardados numa máquina de vectores de suporte
  - Elimina assim os atributos com pesos baixos

Conclusões

- Técnicas de filtragem univariadas são as mais utilizadas
- Necessidade de técnicas “wrapper” ou de filtragem multivariada
- Técnicas de selecção de atributos são e continuarão a ser prática comum na análise de dados em Bioinformática

Perguntas