BIOINFORMÁTICA

Artigo:

"A review of Feauture Selection techniques in bioinformatics"

Evan Saeys, Iñaki Inza and Pedro Larrañga

Jason Araújo

Problemas

- Variáveis de grande dimensão
- Número reduzido de observações

Técnicas de selecção de recursos

- Ferramenta crucial em aplicações de reconhecimento de padrões
- Não alteram a representação original das variáveis
- Existem duas formas de o fazer:
  - **Aprendizagem supervisionada (classificação)**
  - **Aprendizagem não supervisionada (clustering)**

Classificação

- Técnicas de selecção de recursos são organizadas em três categorias:
  - métodos de filtro;
  - métodos de embalagem;
  - métodos incorporados.

Técnica de filtro

- avalia a pertinência dos recursos, por ter em conta apenas as propriedades intrínsecas dos dados
- a relevância de cada recurso é calculada, sendo removidos os recursos que apresentam baixa pontuação
- subconjunto de recursos é apresentado como entrada para o algoritmo de classificação

Técnica de filtro

- Vantagens:
  - Aumentam facilmente o espaço dimensional para conjuntos de dados alta dimensão;
  - são computacionalmente muito simples e rápidos;
  - são independentes dos algoritmos de classificação

Técnica de filtro

- Desvantagens:
  - Normalmente ignora a interacção com os classificadores, ou seja, a procura no espaço do subconjunto de recursos é separada da procura no espaço de hipóteses;
  - cada recurso é avaliado em separado, o que significa que são ignoradas as dependências desse mesmo recurso, o que podem originar um desempenho pior na classificação

Técnica de embalagem

- incorpora a pesquisa do modelo de hipoteses, na pesquisa do subconjunto de recursos
- avaliação de um subconjunto específico de recursos é obtida, tratando e testando um modelo de classificação específica, adaptando esta abordagem a um algoritmo de classificação específico
- Para ser possível a pesquisa do espaço de todos os subconjuntos de atributos, um algoritmo de pesquisa é, então, "embrulhado" em torno do modelo de classificação

Técnica de embalagem

- Vantagens
  - Inclui a interacção entre a procura de um subconjunto de recursos e a selecção de modelos;
  - capacidade de ter em conta as dependências do recurso

Técnica de embalagem

- Desvantagens:
  - Têm um maior risco de "*overfitting*" do que as técnicas de filtragem ;
  - a construção do classificador tem um alto custo computacional

Técnica de incorparação

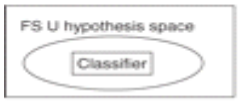
- A pesquisa de um subconjunto ótimo de recursos é criado na construção do classificador;
- pode ser visto como uma pesquisa no espaço combinado de subconjuntos recurso e hipóteses;
- as abordagens são específicas para um dado algoritmo de aprendizagem

Técnica de incorparação

- Vantagens:
  - Tem interação com o modelo de classificação, enquanto ao mesmo tempo, é muito menos intensivo computacionalmente do que o método de embalagem

Técnica de incorparação

- Desvantagens:
  - A seleção depende do classificador

Model search	Advantages	Disadvantages	Examples
Filter	Univariate		
	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	χ^2 Euclidean distance *i*-test Information gain, Gain ratio (Ben-Bassat, 1982)
	Multivariate		
	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation-based feature selection (CFS) (Hall, 1999) Markov blanket filter (MBF) (Koller and Sahami, 1996) Fast correlation-based feature selection (FCBF) (Yu and Liu, 2004)
Wrapper	Deterministic		
	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) (Kittler, 1978) Sequential backward elimination (SBE) (Kittler, 1978) Plus *q* take-away *r* (Ferri *et al.*, 1994) Beam search (Siedelecky and Sklansky, 1988)
	Randomized		
	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing (Skalak, 1994) Genetic algorithms (Holland, 1975) Estimation of distribution algorithms (Inza *et al.*, 2000)
Embedded	Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies	Classifier dependent selection	Decision trees Weighted naive Bayes (Duda *et al.*, 2001) Feature selection using the weight vector of SVM (Guyon *et al.*, 2002; Weston *et al.*, 2003)
			

Aplicação univariada vs multivariada

- Métodos de seleção univariada tem certas restrições e pode levar a uma solução menos correcta por parte dos classificadores, por exemplo, não tendo em consideração as interações gene-gene. Assim, os investigadores propuseram técnicas que tentam capturar essas correlações entre genes.

- A aplicação de métodos de filtro multivariados varia de interações simples para bivariada, é ideal para soluções mais avançadas que exploram interações de ordem superior, tais como seleção de características de correlação de base (CFS)

Aplicação univariada vs multivariada

- Selecção de Recursos utilizando os métodos de embrulho e o incorporado oferecem uma forma diferente de seleccionar um subconjunto de genes multivariado, incorporando o classificador

- O que oferece assim uma oportunidade de construir classificadores mais precisos

Conclusão

- Uma série de técnicas de selecção recursos foi a melhor solução que investigadores nas áreas de Bioinformática, Mineração de dados (*Data mining*) e Apredizam de máquina (*Machine learning*) para lidar com os problemas de grande número de entrada de dados e amostras de pequenas dimensões

Conclusão

- Embora alguns investigadores ainda pensem que as técnicas de filtro de Selecção de Recursos estão restritas apenas à análise de abordagens univariada. A proposta de algoritmos de seleção multivariada pode ser considerada como uma das linhas mais promissoras do futuro do trabalho para a comunidade de bioinformática.

Conclusão

- Outra linha de investigação futura é o desenvolvimento de um conjunto equipado de abordagens especificas de Selecção de Recursos para aumentar a robustez dos subconjuntos de recursos seleccionados

Questões?

Obrigado!