Mineração de Dados Biológicos

PROJETO – Segundo Semestre – 2010

Prof. Ricardo J. G. B. Campello

O presente trabalho consiste da aplicação e avaliação de algoritmos de seleção de atributos e classificação no contexto de dados de expressão gênica. As bases de dados que serão utilizadas são provenientes de diversos experimentos associados à detecção de câncer. Nestas bases de dados, os objetos são amostras de tecidos pertencentes a diferentes classes enquanto os atributos correspondem a genes para os quais os níveis de expressão foram mensurados. Para tal avaliação, as implementações dos algoritmos vistos em aula e disponíveis no software de mineração de dados WEKA devem ser utilizadas. Dois diferentes cenários de aplicação e avaliação devem ser considerados, como descrito a seguir[1].

1. Classificadores

Neste cenário os algoritmos de classificação de dados devem ser aplicados e avaliados **sem** a utilização prévia de métodos de seleção de atributos. Os seguintes algoritmos devem ser considerados para avaliação:

- kNN
- Naïve Bayes
- 1Rule
- J48 (implementação do C4.5 disponível no WEKA)

Para o algoritmo kNN, varie o número de vizinhos de 1 à 10. Plote um gráfico na forma *número de vizinhos x acurácia de classificação* e avalie a influencia que o número de vizinhos possui na acurácia do método. Escolha o melhor dentre os 10 classificadores kNN avaliados para posterior comparação com os demais métodos. Para tanto, plote gráfico(s) que permitam a comparação entre os métodos: melhor kNN, Naïve Bayes, 1Rule e J48. Além da acurácia, considere também ao menos a medida F-measure. Por fim comente e apresente sua interpretação para os resultados obtidos.

2. Seleção de Atributos + Classificadores

Neste cenário, algoritmos de seleção de atributos devem ser aplicados previamente aos classificadores[2]. Os algoritmos de seleção de atributos que devem ser avaliados são:

- ReliefF
- CFS
- LVF

Após a aplicação dos filtros de seleção de atributos, cada um dos classificadores anteriormente mencionados devem ser avaliados. Quanto ao algoritmo ReliefF, este gera uma ordenação de atributos, de forma que a escolha de um número de atributos é necessária após sua aplicação. Considerando ainda o algoritmo ReliefF, varie o número de atributos selecionados por este algoritmo em um intervalo determinado (cabe a você escolher o intervalo de valores) e plote gráficos para os classificadores na forma *número de atributos x acurácia,* de forma a mostrar o impacto do número de atributos selecionados nos métodos de classificação. Observe as correlações entre os atributos selecionados e verifique se há presença de atributos redundantes nos ranks de atributos gerados (utilize a parte de visualização do WEKA – aba *visualize*). Compare os resultados de acurácia obtidos entre os pares algoritmo de seleção – algoritmo de

[1] Para ambos os cenários de avaliação utilize o procedimento de validação cruzada de 10 pastas estratificado.

[2] Aplique os métodos de seleção somente nas pastas de treinamento e não utilizando todo o conjunto de dados.

classificação e comente os resultados obtidos. Por fim, compare os resultados obtidos após a seleção de atributos com aqueles obtidos somente com a aplicação dos classificadores (sem seleção de atributos). Comente sobre os impactos e efeitos da seleção de atributos em cada um dos classificadores utilizados. Além disso, verifique se todos os classificadores obtiveram as mesmas melhoras com a aplicação dos métodos de seleção de atributos, justificando os fatos observados.

Além dos gráficos solicitados, podem ser gerados gráficos e tabelas adicionais a fim de facilitar a comparação e obtenção de conclusões. Sua análise e conclusões devem ser apresentados em forma de um relatório com no máximo 5 páginas. Relatórios contendo somente gráficos e/ou tabelas sem análises profundas e fundamentadas receberão nota **zero**.

- **Bases de Dados**

A cada aluno foi atribuída uma base de dados única. A distribuição das bases de dados é apresentada na tabela abaixo. Note que uma porção de genes irrelevantes já foram removidos das bases. Caso ache necessário aplique técnicas de pré-processamento aos dados (normalização, discretização, ...) e comente sobre seu impacto nos resultados obtidos.

Nome	Base de Dados	Nome	Base de Dados
Alessandra Figueiredo	*alizadeh-2000-v1*	Gilson Campani Junior	*lapointe-2004-v1*
Amanda Souza Camara	*alizadeh-2000-v2*	Gustavo M. Alvares de Lima	*lapointe-2004-v2*
Americo Tavares Ranzani	*alizadeh-2000-v3*	Heloisa dos Santos Muniz	*liang-2005*
Andre Zamith Selvaggio	*armstrong-2002-v1*	Jaqueline Yu Ting Wang	*nutt-2003-v1*
Andrea Rodrigues Cardoso	*armstrong-2002-v2*	Jessica Baleiro Okado	*nutt-2003-v2*
Anielle Coelho Ranulfi	*bhattacharjee-2001*	Leticia Marchese	*nutt-2003-v3*
Bruno L. S. Paula de Mello	*bittner-2000*	Luis Felipe Santos Mendes	*pomeroy-2002-v1*
Bruno Yasui Matsuyama	*bredel-2005*	Luma Godoy Magalhaes	*pomeroy-2002-v2*
Caio M. Ramos de Oliveira	*dyrskjot-2003*	Mariana R. Bunoro Batista	*ramaswamy-2001*
Cesar Maschio Fioravanti	*garber-2001*	Mira Melke de Oliveira	*risinger-2003*
Eduardo Cocca Padovani	*golub-1999-v1*	Pedro Michelao Neuber	*shipp-2002-v1*
Evandro Jose Mulinari	*golub-1999-v2*	Ricardo Simionato Boffa	*singh-2002*
Fernando Takeshi Tanouye	*gordon-2002*	Thais Panhan Merlo	*su-2001*
Flavio A. C. da Silva Andrade	*khan-2001*	Wagner Rafael Correr	*tomlins-2006-v1*
Francesco B. Teixeira	*laiho-2007*	Waldomiro Thiago Corsi	*tomlins-2006-v2*

As bases de dados em formato ARFF encontram-se disponíveis no site da disciplina no arquivo bases.tar. Informações e referências adicionais para cada uma das bases de dados podem ser encontradas no site: http://algorithmics.molgen.mpg.de/Static/Supplements/CompCancer/datasets.htm.

Informações Importantes

Os trabalhos podem ser realizados individualmente ou em dupla. Caso seja realizado em dupla, somente uma das bases de dados atribuídas aos alunos da dupla deve ser utilizada durante a realização do trabalho. O relatóio deverá ser entregue **pessoalmente** de forma impressa ao estagiário PAE da disciplina. O estagiário PAE pode ser consultado nos horários de atendimento para sanar dúvidas quanto à realização do trabalho. A **data limite de entrega** do relatório (**pessoalmente**) ao estagiário PAE será 29/11/2010 . No momento da entrega o aluno (ou a dupla) deverá escolher a data e horário da argüição em grade de disponibilidade pré-estabelecida. A preferência das datas e horários na grade será por ordem de entrega dos trabalhos.