Mineração de Dados

Mestrados MI/MEI

UCE - Sistemas de Suporte à decisão

Projecto

Ano Lectivo de 2009/2010

16 de Junho de 2010

Avaliação de Modelos de Previsão

O objectivo do projecto proposto para avaliação prática desta disciplina concentra-se num estudo de avaliação de modelos de previsão para classificação. Os modelos são derivados a partir de um conjunto de problemas dados seleccionados por cada grupo de trabalho.

Perante um conjunto de dados representativos de um problema, queremos desenvolver um estudo sobre os factores que influênciam a precisão dos modelos construídos a partir desses dados. Deve ser levado em conta factores como: tipo de dados, distribuição de valores, número de classes, qualidade dos dados, tipo de algoritmos, metodologias de combinação de modelos, etc.

Estudo

Cada grupo deve por sua iniciativa seleccionar um dataset, e consequentemente um problema, para usar neste projecto. O dataset escolhido tem de ter como dimensões mínimas 1000 registos e 7 attributos (para além do atributo objectivo).

A primeira etapa do estudo é a caracterização do problema (dataset) escolhido. Nomeadamente, esta caracterização deve conter:

- Breve descrição do problema que os dados representam
- Qual o atributo objectivo para o qual são geradas previsões, classes (diferentes valores para esse atributo) e seu significado, bem como cardinalidade do atributo.
- Distribuição de classes
- Número de atributos e sua caracterização (tipos de dados, distribuição, etc).
- A partição natural dos dados i.e. produzir o clustering associado aos dados quando ignoramos o atributo classe.

Nas etapas seguintes, os membros de cada grupo têm de desenhar um conjunto de experiências para avaliar a capacidade predictiva dos modelos gerados por 6 algoritmos diferentes. As experiências devem ser descritas em pormenor bem como todo o seu planeamento e que tipo de hipóteses estão a ser testadas. Devem ser seleccionados 6 algoritmos diferentes para avaliação. Destes 6, pelo menos dois devem implementar uma metodologia de combinação de modelos, onde os modelos simples são originados de um dos 4 algoritmos escolhidos. Um dos algoritmos devem originar modelos de árvores e.g. J48, e outro modelos de regras e.g. PART.

O estudo deve conter um tabela com o resumo dos resultados (em termos de valor de erro) obtidos na experimentação, onde deve ser usada validação cruzada. O estudo deve apontar respostas a questões como:

- genericamente qual é o melhor algoritmo?
- em termos de classes como devemos seleccionar os algoritmos? (use as várias medidas de avaliação por classe estudadas).
- as metodologias de combinação de modelos atingem mais valias em que classes e porquê?
- Apresente um estudo em termos de análise de curvas ROC para um classe específica
- Apresente 5 regras de associação geradas a partir dos dados usados justificando a sua escolha e o significado de cada uma. à sua escolha.

Os grupos tem a liberdade de escolha das ferramentas com as implementações dos algoritmos a usar e.g. WEKA, SQL Server, etc. No entanto, essa opção deve estar documentada.

Datas & Entregas

A documentação de cada projecto resume-se a uma apresentação em *Power Point*.

As apresentações são na tarde do dia 16 de Julho de 2010. Cada apresentação terá no máximo 15 minutos + discussão.

Conteúdos

A resolução deste problema deve ser apresentada na forma detalhada contendo a descrição das experiências, decisões, resultados e conclusões. Os resultados das experiências devem ser ilustrados na forma de tabelas de erro. Exemplo:

	Algo. A	**Algo. B**	**Algo. C**
data_1	0.0034 ± 1.20	0.0035 ± 1.39	0.0023 ± 0.30
data_2	0.1233 ± 2.11	0.0012 ± 1.31	0.0102 ± 3.13

Aqui, é apresentado o erro médio obtido da Validação cruzada (10 fold-CrossVal) e o desvio padrão. Deve-se usar 4 casas decimais de precisão no valor de erro e duas no valor do desvio padrão.