AP532- Preparação de Dados para Mineração de Dados

Lista de Exercícios 2: Abordagens para Seleção de Atributos

1. Carregar o arquivo **body.csv** no Weka e responder as seguintes questões:

   a) Quais os atributos que podem ser descartados desse dataset por meio da **Análise de Componentes Principais** (**PCA**)? Use o critério de corte sugerido por Jolliffe (1972).

   b) Que atributos podem ser removidos do dataset usando o teste do **Qui-quadrado**? Qual é o critério de corte?

   c) Usando o método **Wrapper** com árvore de decisão (**J48**), quais os atributos que são preservados (retidos)?

   d) Quais os atributos são retidos usando o método **CFS (Correlation Feature Selection)**? Compare o resultado com o método **Wrapper**.

   e) Determine os atributos que podem ser removidos usando os métodos **InfoGain** (**Ganho de Informação**) e **GainRatio** (**Taxa de Ganho**). Estabeleça um critério de corte baseado nos resultados apresentados pelos dois métodos.

   f) Existe alguma relação entre os métodos **InfoGain**, **GainRatio** e **Qui-quadrado** para esse dataset? Qual seria?

   g) Com base na análise dos métodos acima, qual seria o atributo mais importante desse dataset? Por quê?

2. Carregar o dataset **body.csv** no Weka e comparar a eficiência das abordagens de seleção de atributos (PCA, χ^2, Wrapper, InfoGain, GainRatio e CFS) com relação aos algoritmos de classificação apresentados na legenda abaixo. Cada célula da matriz abaixo deve conter a acurácia e a estatística Kappa do classificador. **Analise as melhores soluções em termos de acurácia e kappa**.

Abordagens	**Algoritmo I**		**Algoritmo II**		**Algoritmo III**		**Algoritmo IV**	
	Acurácia	**Kappa**	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**	**Acurácia**	**Kappa**
Sem seleção								
PCA								
χ^2								
Wrapper								
InfoGain								
GainRatio								
CFS								

- **Legenda dos algoritmos de classificação**:
  - ✓ **Algoritmo I:** J48 (C4.5) → Árvore de Decisão
  - ✓ **Algoritmo II:** IBk (kNN) → Classificador Lazy
  - ✓ **Algoritmo III:** NaiveBayes → Classificador Bayesiano
  - ✓ **Algoritmo IV:** SMO → Support Vector Machine

- **Guia para uso de Kappa em Epidemiologia e em Medicina:**
  - ✓ KAPPA $>$ 0,80 é considerado excelente.
  - ✓ KAPPA 0,60 – 0,80 é considerado bom.
  - ✓ KAPPA 0,40 – 0,60 é considerado regular.
  - ✓ KAPPA $<$ 0,40 é considerado ruim.

3. Faça uma análise da tabela acima e responda:

   a) Valeu a pena usar os métodos de seleção de atributos para esse dataset? Por quê?

   b) Qual foi o método mais eficiente de todos? Explique.

   c) E o método menos eficiente? Explique.