Minería de datos

Practica sobre la comparación de varios métodos de aprendizaje sobre varios conjuntos de datos

Contenido

- Conjuntos de datos
- Algoritmos
- Test de signos
- Rankings

Conjuntos de datos

- Soybean
  - 683 instancias
  - 36 atributos (35 + clase)
  - 19 clases
- Vote
  - 435 instancias
  - 17 atributos (16 + clase)
  - 2 clases
- Labor
  - 57 instancias
  - 17 atributos (1+ clase)
  - 2 clases
- Ionosphere
  - 351 instancias
  - 35 atributos (34+clase)
  - 2 clases
- Diabetes
  - 768 instancias
  - 9 atributos (8 + clases)
  - 2 clases
- Glass
  - 214 Instancias
  - 10 atributos (9+clase)
  - 6 clases

Conjuntos de datos

- Segment-test
  - 810 instancias
  - 20 atributos (19+clase)
  - 7 clases
- Breast Cancer
  - 286 instancias
  - 10 atributos (9+clase)
  - 2 clases
- Credit-g
  - 1000 instancias
  - 21 atributos (20+clase)
  - 7 clases

- Tres conjuntos de datos a elegir en el repositorio de la UCI:

  http://archive.ics.uci.edu/ml/index.html

  - Tarea de clasificación
  - 50 < nº instancias < 10000
- Describir los tres conjuntos seleccionadas

Algoritmos

- SVM kernel lineal (SMO)
- 3-NN (Ibk)
- NB
- J48
- OneR

Test de signos
Comparación de 2 métodos, varios conjuntos

- Comparar NB y J48 sobre los 12 conjuntos de datos, estimando su tasa de error mediante validación cruzada con 10 particiones

- Utilizar WEKA Experimenter para estimar la tasa de error

Rankings

- Obtener los rankings promedio de los cinco algoritmos sobre los doce conjuntos de datos
- Estimar la tasa de error mediante validación cruzada repetida de 5 particiones con 5 repeticiones
- Estimar la tasa de error con Weka Experimenter
- Podéis obtener los rankings llevando los resultados a una hoja Excel
- Utilizar el test de Iman y Davenport para determinar si los rankings son significativamente diferentes del ranking medio
- Si son significativamente diferentes, realizar un test post-hoc para determinar cuáles son diferentes
- Discutir los resultados