MIP group

Multi-paradigm Inductive Programming

Dept. de Sistemes Informàtics i Computació
Universitat Politècnica de València

Organización

- *Presentación del grupo MIP*
- *Exposición del artículo:* Re-designing cost-sensitive decision tree learning

Grupo MIP

Composición y líneas de investigación

- El MIP inició sus actividades en 1997 en el seno del grupo ELP/GPLIS. (http://www.dsic.upv.es/users/elp/elp.html)
- Consta de 4 personas (3 profesores, un becario FPI,2 alumnos de doct.), dirigiéndose 4 tesis.
- Las líneas de investigación
  - **Programación multiparadigma inductiva.**
  - **Aprendizaje automático.**
  - **Extracción e intercambio de conocimiento.**
  - **Minería de datos.**
  - **Depuración inductiva.**
  - **Análisis ROC y evaluación de modelos para la toma de decisiones.**
  - **Aplicaciones de la extracción automática de conocimiento.**

Grupo MIP

- ***Colaboración con otras grupos o redes nacionales o internacionales***
  - Los contactos establecidos con la comunidad científica internacional se concretan en
    - **Estancias de investigación en:**
      - Universidad de Udine, Kiel, Bristol, Viena
    - **Proyectos**
      - Acciones integradas
      - CICYT
      - Autonómicas
      - Universidad

Grupo MIP

- Integración del grupo en
  - ***Network of Excellence in Inductive Logic Programming ILPnet2.***
  - **Propuesta internacional *RuleML* para el desarrollo de un intercambio de reglas basado en XML.**
- Expresiones de interés (VI Programa Marco U.E.)
  - ***Network of Excellence on Relational Data Mining ReDaM*, para la difusión de la minería de datos relacional.**
  - ***RISEN: Rules In a Semantic Web Environment.***

Grupo MIP

- ***Capacidad de difusión y transferencia de resultados***
  - El sistema FLIP
    - **entorno para la inducción de programas lógico-funcionales a partir de hechos.**
    - **extensión del campo ILP.**
      - http://www.dsic.upv.es/~flip/smiles
  - El sistema SMILES
    - **sistema de aprendizaje automático.**
    - **integración de diversas técnicas y paradigmas de aprendizaje.**
    - **extensión del aprendizaje de árboles de decisión.**
      - http://www.dsic.upv.es/~flip/smiles

Re-designing Cost-Sensitive Decision Tree Learning

V. Estruch, C. Ferri, J. Hernández-Orallo, M.J. Ramírez-Quintana

Dept. de Sistemes Informàtics i Computació

Universitat Politècnica de València

WS-Aprendizaje y minería de datos
Iberamia (Sevilla 2002)

Motivation

- New applications (*web and data mining*, *knowledge discovery*, etc.) are demanding the integration of several features in machine learning methods.
  - **Output of comprehensible models.**
  - **Efficient management of huge volumes of data.**
  - **Context sensitiveness.**
    - misclassification cost, **ROC** analysis
  - **High accuracy.**
- The properties above aren't orthogonal.
  - **Often, $\uparrow\uparrow$ accuracy $\Rightarrow$ $\downarrow\downarrow$ comprehensibility.**

Motivation

- The applicability of machine learning methods are hampered by several costs.
  - **Outer costs**
    - Data cleaning and data transformation.
  - **Inner costs**
    - Generation costs: computational cost.
    - Application costs: interpretation, model inaccuracies, misclassification and test cost, . . .
- Generation and application costs are **not** orthogonal

Tree ensembles (forest)

- Decision tree learning
  - efficient (eager strategy)
  - comprehensible
  - accuracy depends on the right choice of the splits
- The combination of a set of trees improves the accuracy. But:
  - $\uparrow\uparrow$ memory
  - $\uparrow\uparrow$ time
  - comprehensibility is **LOST** !!

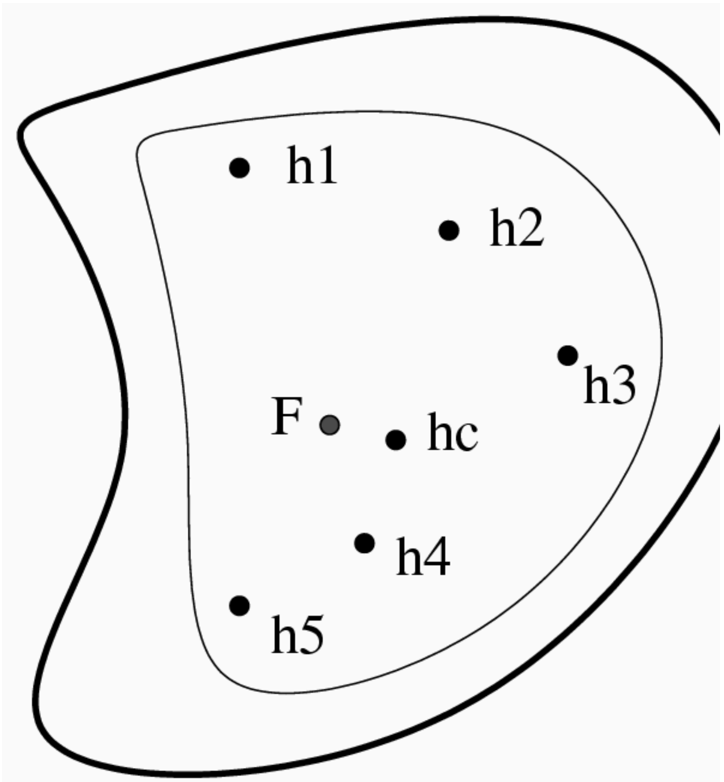
Shared ensembles

- One way to overcome the two first limitations could be to share the common parts of several trees.
- Multi-tree:
  - unselected splits are stored in a queue of suspended nodes.
  - further trees can be obtained by *"wakening"* stored splits.
  - an **AND/OR** tree is generated.
- Multi-tree is a set of models with shared structure.
  - we can

    select **one** model (Occam, expected error, etc.)
    select **n** models
    combine them locally

Archetype

- The **single hypothesis** which is the **most similar** to the combined one according to a measure of similarity.



archetype

Learning cost-sensitive decision trees

- Traditionally, accuracy (percentage of number of instances that are correctly classified) has been used as a measure of the quality of a classifier.
- Misclassification costs must be taken into account.
  - **e.g. medical decision making, diagnosis, etc.**

Learning cost-sensitive decision trees

- Depending on whether costs are known or not, we can design a cost-sensitive learning algorithm.
  - **Costs are known**
    - Incorporating cost in to the splitting criterion.
      - † Do not always lead to better cost-sensitive decision trees.
    - Assigning the less *expensive* class at each node.
  - **Costs are unknown (ROC analysis)**
    - AUC can be used as an alternative measure to accuracy.
    - Splitting criterion based on AUC.

Experimental evaluation

- Experiments were performed within the SMILES system.
- The datasets belong to the **UCI** *dataset repository*.
- Multi-tree vs. *boosting* and *bagging*
  - $\uparrow\uparrow$ number of iterations $\Rightarrow$ better results
  - Computational time required by multi-tree is **always** lower.
- *Archetype* vs. combination
  - When the number of iterations is increased, we obtain a better *archetype* solution, which is increasingly closer to the combined solution.

Conclusions

- Multi-tree as a structure which makes a suitable use of **computational resources** and gives **better** results than *boosting* and *bagging*.
- Archetype classifier in order to get a trade-off between **comprehensibility** and **accuracy**.
- AUC splitting criterion in order to design a **cost-sensitive** learning algorithm.