**Lecture 30**

**Data Mining and KDD Presentation (2 of 4): Relevance Determination in KDD**

Monday, April 3, 2000

DingBing Yang
Department of Plant Pathology, KSU

Read:
"Irrelevant Features and the Subset Selection Problem"
George H. John; Ron Kohavi; Karl Pfleger

**Presentation Outline**

- **Objective**
  - Finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts

- **Overview**
  - Introduction
  - Relevance Definition
  - The Filter Model and The Wrapper model
  - Experimental results

- **References**

**Introduction**

- Why find a good feature subset?
  - Some learning algorithms degrade in performance (prediction accuracy) when faced with many features that are not necessary for predicting the desired output.
  - Decision tree algorithm: ID3, C4.5, CART; instance-based algorithms: ID3
  - Some algorithms are robust with respect to irrelevant features, but their performance may degrade quickly if correlated features are added, even if the features are relevant

- An example
  - Running C4.5, dataset is Monk1, there are 3 irrelevant features.
  - The induced tree has 15 interior nodes, five of them test irrelevant features, the generated tree has an error rate of 22.2%
  - If only the relevant features are given, the error rate is reduced to 11.1%

- What is a optimal feature subset?
  - Given an inducer I, and a dataset $D$ with features $X_1, X_2, \ldots, X_n$, from a distribution $D$ over the labeled instance space. An optimal feature subset is a subset of the features such that the accuracy of the induced classifier $C = I(D)$ is maximal.

**Incorrect Induced Decision Tree**

The tree induced by C4.5 for “Corral” dataset that has “correlated” features and irrelevant features

**Background Knowledge**

- **ID3 algorithm**
  - It is a decision tree learning algorithm. It constructs decision tree top-down.
  - Compute the information gain of each instance attribute among the candidate attributes. Select the attribute that has maximum IG value as the test at the root node of the tree.
  - The entire process is then repeated using the training example associated with each descendant node.

- **C4.5 algorithm**
  - It is an improvement over ID3. It is a rule post-pruning.
  - Infer the decision tree from the training set. Convert the learned tree into an equivalent set of rules.
  - Prune each rule by removing any precondition that result in improving its estimated accuracy.

- **K-Nearest neighbor Learning**
  - It is an instance-based learning. It just simply stores the training examples.
  - Generalization beyond these examples is postponed until a new instance must be classified.
  - Each time a new query instance is encountered, its relation to the previous stored examples is examined.
  - The target function value for a new query is estimated from the known values of the k nearest training examples.

- **Minimum Description Length (MDL) Principle**
  - Choosing the hypothesis that minimizes the description length of the hypothesis plus the description length of the data given the hypothesis.

- **Naïve Bayes classifier**
  - It incorporates the simplifying assumption that attributes values are conditionally independent, given the classification of the instance.
A heuristic search
- Each state in the search space specifies a subset of the possible features
- Each operator represents the addition or deletion of a feature

The four basic issues in the heuristic search process.
- Starting point:
  - forward selection, backward elimination, both of them.
- Search organization:
  - exhaustive search, greedy search, best-first search,
- Evaluate function:
  - prediction accuracy, structure size, induction algorithm
- Halting criterion:
  - when none of alternatives improves the prediction accuracy
  - until the other end of the search and then select the best

The type of heuristic search: Filter model and Wrapper model

Feature Subset Selection Algorithm

Input features → Feature subset selection → Induction algorithm

Filter Model

Input features → Feature subset search → Feature subset evaluation

Wrapper Model

Filter Approach
- Filter approach
  - FOCUS algorithm (min-features)
    - exhaustively examines all subsets of features
  - select the minimal subset of features that is sufficient to determine the label
  - problem: Sometimes the resulting induced concept is meaningless.
  - Relief algorithm
    - assign a relevant weight to each feature, which represents the relevance of the feature to the target concept.
    - It samples instances randomly from the training set and updates the relevance values based on the difference between the selected instance and the two nearest instances of the same and opposite class.
    - Problem: can’t remove many weakly relevant features.
  - Cardo algorithm
    - use a decision tree algorithm to select a subset of features for a nearest-neighbor algorithm.
- Example
  \[ |A \cap C| > |A \cap D| > |1 \cap A \cap B| \]
Filter Approach

- Totally irrelevant features
- Weakly relevant features
- Strongly relevant features

Relationship of filter approach and feature relevance

\[\text{FOCUS: all strong relevances and part of weak relevances.} \]
\[\text{Relief: both strong relevances and weak relevances.} \]

Wrapper Approach

- A wrapper search use the induction algorithm as a black box.
  - A search requires a state space, an initial state, a termination condition, and a search engine.
  - Each state represents a feature subset.
  - Operators determine the connectivity between the states. For example: operators that add or delete a single feature from a state.
  - The size of the search space for n features is \(O(2^n)\).
  - The goal of the search: find the state with the highest evaluation, using a heuristic function to guide it.

Subset Evaluation: Cross-validation (n-fold):
- The training data is split into \(n\) approximately equally sized partitions.
- The induction algorithm is then run \(n\) times, each time using \(n-1\) partitions as the training set and the other partition as the test set.
- The accuracy results from each of the \(n\) runs are averaged to produce the estimated accuracy.

Experimental Evaluation

- Datasets
  - Artificial datasets: CorrAL, Monk1*, Monk3*, Parity5+5
  - Real-world datasets: Vote, Credit, Labor
- Induction algorithm
  - ID3 and C4.5
- Feature subset selection approach
  - wrapper approach
- Cross validation
  - 25-fold
- Results
  - The main advantage of doing subset selection is that smaller structures are created.
  - Feature subset selection using the wrapper model did not significantly change generalization performance.
  - When the data has redundant feature, but also has many missing values, the algorithm induced a hypothesis which makes use of these redundant features.
  - Induction algorithms have a great influence on the performance of the FSS approach.

Summary

- Content Critique
  - Key Contribution - It presents a feature-subset-selection algorithm that depends on not only the features and the target concept, but also on the induction algorithm.
  - Strengths
    - It differentiates irrelevance, strong and weak relevance.
    - The wrapper approach works better on correlated features and irrelevant features.
    - Smaller structures are created: smaller trees allow better understanding of the domain.
    - Significant performance improvement is achieved on some datasets. (the error rate reduced)
  - Weaknesses
    - Its computational cost is expensive. Calling the induction algorithm repeatedly.
    - Overfitting. Overuse of the accuracy estimates in the feature subset selection.
    - Experiment only on the decision tree algorithm (ID3, C4.5). How about other learning algorithms (Naïve Bayesian classifier).
    - The performance is not always improved, just on some datasets.
  - Audiences: AI researchers and expert system researchers in all kinds of field.