Computer-Implemented Method for Data Classification and Hierarchical Clustering

Technology #m07-091

“Lead Inventor: Hassan Haider Malik

Classifying Large Amounts of Data Is Fundamental Machine Learning Problem: Classification is a fundamental machine learning problem that has broad application to areas that depend on being able to sub-group data that may be presented in a variety of forms (numeric, text, etc.). For the applications in internet search, customized advertising, email filtering and computational biology, an efficient method of categorizing large amount of data information is needed. This technology presents an highly efficient classification algorithm, identifying features that have a large information value, which is defined in the training mode. It has shown promise on highly sparse or unbalanced data as well, with unmatched accuracy on web page classification in comparison to other methods.

Data Classification Software Uses Hierarchical Clustering, Improves Internet Searching, Customized Advertising: This technology presents a unique approach to the classification problem, which calculates a score for every pair of features in the training instance, considering global, local and class-based importance. A novel score adjustment scheme is applied and test instances are classified using this metric. Data is split into training and testing sets for k-fold cross validation. Training instances are traversed and a score is calculated for each feature (item) pair, based on frequency and “global interestingness.” The top scoring item sets are selected and placed in a class-item set tree, and scores are adjusted using a scheme empirically identified to have the best performance.

Applications:
- Text categorization – particular useful over the internet
- Customized Internet advertising based on the above
- Filtering spam from non-spam email
- Identifying credit card transactions which are fraudulent from those which are valid
- Authentication based on face, speech or handwriting recognition
- Computational Biology – splitting disease from non-disease patients

Advantages:
- Efficient, processing each feature only once, and stores this knowledge in a compact form
- Provides a pattern-based hierarchical clustering technique that can build a cluster hierarchy without requiring mining for globally significant patterns

Licensing Status: Available for Licensing and Sponsored Research Support

Inventors

John Ronald Kender