Summary: Toward optimal Feature Selection using Ranking Methods and Classification Algorithms

Name: Dean D'souza

H.U. ID: 168424

Summary

The paper presents a comparison between feature ranking methods on two real datasets and highlights the importance of selection of ranking methods for classification accuracy. It also recommends the use of multiple different indices to be certain that a subset of features gives the highest accuracy.

Feature selection is an active field in computer science and is a process which involves choosing a minimum subset of features from all the original features available, to reduce the feature space and is done based on an evaluation criteria. This is done as for most target concepts a small subset of features can be almost equally relevant and gives us the advantages of reduced dimensionality and removal of redundant, irrelevant or noisy data.

There are two kinds of feature selection algorithms:

1. Wrappers: In which filter methods evaluate quality of the selected features without the use of the classifier while wrapper methods (which requires application of a classifier trained on a given feature subset) evaluate quality.
2. Embedded Approach: In which feature selection is performed during learning of optimal parameters.

The paper also gives a general architecture for feature selection which involves Subset Generation (a search procedure to generate subsets), Subset Evaluation (which involves evaluating the performance with each subset iteratively), Stopping Criterion (such as if a predefined number of features or iterations are reached, or if the optimal subset based off the evaluation criterion has been generated) and Result Validation (where the selected best features are validated through different tests).

The paper goes on to discuss about the practical usefulness of the commonly used methods for ranking. These methods include entropy-based methods like Information Gain (IG) attribute evaluation, Gain Ratio (GR) attribute evaluation and Symmetrical Uncertainty (SU) attribute evaluation. It also includes statistical methods like Relief-F (RF) attribute evaluation, One-R (OR) attribute evaluation and Chi-Squared (CS) attribute evaluation.

Some issues with these methods include the IG criterion favoring features with more values which may not be informative and GR and SU methods favoring variables with fewer values. The OR method is also the simplest way of handling feature selection statistically, while the RF method is mainly used for two-class problems and requires extra efforts to be applied for multi-class problems.

The paper also discusses about some of the most widely used classification algorithms, including:

1. IB1: It is a nearest neighbor classifier whose advantage is that they can learn quickly from a very small dataset. It works based on the intuition that classification of one instance is most likely to be similar to the classification of other instances, nearby in the vector space. However, calculations can be quite expensive for large data and hence stresses the need for Principal component analysis and IG-based feature ranking.
2. Naïve Bayes: Based on the Bayes theorem, this classifier assumes that features are independent and has the advantage of requiring a small amount of training data to estimate parameters for classification. The nature of the method can itself be a disadvantage as not all features may necessarily be independent.
3. C4.5 decision tree: This algorithm recursively partitions the training data set per test on potential of feature values in separating classes. The feature tests are chosen one at a time in a greedy manner but are dependent on results of previous tests. The advantages include being simple to understand, requires little data preparation, is robust and performs well with large datasets in less time. A disadvantage of this method is that it would require very large datasets to produce reliable results.
4. Radial Basis Function (RBF) network: It a popular type of feed-forward network and has two layers, excluding the input layer. The advantages include requiring less formal statistical training, ability to detect nonlinear relationships between dependent and independent variables as well as interactions between predictor variables. A disadvantage of this method is that it gives the same weight for every feature (as all are treated equally in distance computation) and hence cannot properly deal with irrelevant features.

The paper goes on to describe the experiments conducted by the authors and the results obtained from two data sets, “Statlog (Australian Credit Approval)” (which has a good mix of features as well as missing values) and “Statlog (German Credit Data)” of the UCI repository of machine learning databases. The experiment involved using all the mentioned feature selection methods and classification algorithms on both the datasets and a ten-fold cross validation was done to estimate accuracy. A note made by the authors at this point was that for real world applications ranking and classification should be done separately to avoid overfitting.

Out of all the ranking methods, only OR ranking method seemed to perform poorly with all classifiers (especially in the cases of IB1 and C4.5 decision trees) while other ranking methods gave good results for balanced accuracy. Additionally, it was observed that for the Australian credit approval dataset, classification accuracy using the RF ranking method with RBF networks and Naïve Bayes classifier were high but were low in the case of IB1 classification.

The paper concludes by emphasizing that different ranking indices should be used for real-world applications and comments on possible improvements such as algorithms and datasets being chosen on more precise criteria, features with the lowest ranking values of various indices in all cross validations being safely rejected and remaining features being analyzed using methods that eliminate redundant and correlated features.

References

[1] “TOWARD OPTIMAL FEATURE SELECTION USING RANKING METHODS AND CLASSIFICATION ALGORITHMS”, Received: April 2009 / Accepted: March 2011, Jasmina NOVAKOVIĆ (Faculty of Computer

Science, Megatrend University, Serbia jnovakovic@megatrend.edu.rs), Perica STRBAC, Dusan BULATOVIĆ, Yugoslav Journal of Operations Research