Ensemble based Classification

Objective

Understand the working of ensemble based classification by implementing

- Ensemble of same model applied on subsets of records(Bagging)
- Ensemble of different Models applied on same data
- Ensemble of same model applied on subset of fields (Random Forest)

Ensemble Node

The Ensemble node combines two or more model nuggets to obtain more accurate predictions than can be gained from any of the individual models. By combining predictions from multiple models, limitations in individual models may be avoided, resulting in a higher overall accuracy. Models combined in this manner typically perform at least as well as the best of the individual models and often better.

Ensemble of same model applied on subsets of records (Bagging)

1. Read the given metadata file and read in data using appropriate source node. Take a Partition node, connect source node to it and specify the partition ratio as 90:10. Generate Training and testing select nodes and connect partition node to both of them.
2. Connect training select node to *ten* Sampling nodes. In all nodes select *Complex* sample method, specify the Sample type as *Random* and Sample size as *Fixed (0.1)*. Make sure that random seed in all of them are different.
3. Take *five* Append Nodes and connect *one* pair of sampling nodes to each one of them. Connect testing select node to all the append nodes. Connect each append node to a type node and choose appropriate measurement type and role for each filed.
4. Connect each type node to a classification node (preferably C5.0) and configure it to generate best possible model.
5. Save the project (stream). Run all the streams. Connect each model nugget to a select node and specify the condition as to select only testing records.
6. Connect all testing select nodes to a Merge node. Configure the merge node to merge the data on the basis of *Order*. Include whole input field set of one source and predicted fields from all sources, rename predicted fields as appropriate. Filter out rest of the fields.
7. Connect this merge node to an *Ensemble* node from Field Ops tab in nodes palette. Select appropriate *Target Field* and *Ensemble method* as Voting.
8. Connect the ensemble node to an *Analysis* node and Run the stream from merge node
9. Carefully analyze the output. See the accuracy of each model and the accuracy of ensemble of these models. Try to tune each model individually in order to get most accurate model.

Ensemble of models applied on same data

10. In place of connecting all streams to same classification node in Step 5 above, now connect each stream to a different classification model (C5.0, KNN, SVM, C&R Tree, etc.).
11. Change the field names on merge node accordingly. Run all the streams.
12. Carefully analyze the output. See the accuracy of each model and the accuracy of ensemble of these models. Try to tune each model individually in order to get most accurate model.

Additional Questions

1. Increase the number of base classifiers and find the most appropriate number of classifiers giving the best accuracy
2. Apply ensemble by first manipulating the classes of the dataset.
3. Using random vector select different features for different base classifiers and find the accuracy using ensemble of appropriate base classifiers.