Mastering chaos with cost-effective sampling

Mona Rahimi, Alexander Rasin, Daniela Stan Raicu, Jacob Furst

DATA MINING & PREDICTIVE ANALYTICS

DEPAUL UNIVERSITY
COLLEGE OF COMPUTING AND DIGITAL MEDIA

INTELLIGENT MULTIMEDIA PROCESSING LABORATORY

Abstract

Large datasets make it difficult to apply various data analysis techniques such as classification, clustering and prediction. Therefore, there is an immense interest in selecting a small subset of a dataset in a way that preserves the information contained within the original dataset thus making it easier to perform data analysis. A representative data sample is also beneficial for the purposes of visualizing large datasets; furthermore same techniques can help reduce the cost of knowledge discovery techniques. In this research our goal is to develop algorithms for selecting the most representative subset from a large dataset and use that subset to train the classifier while still maintaining a relatively high accuracy in classifying the data. We are planning to apply our methodology to Lung Cancer Database Consortium (LIDC) dataset, which includes Computed Tomography (CT) lung images, to select the most informative cases which are to be annotated by the radiologists manually. In long term, the proposed approach can reduce the number of cases to be interpreted by multiple radiologists, and therefore reducing the costs of medical diagnosis.

Methodology

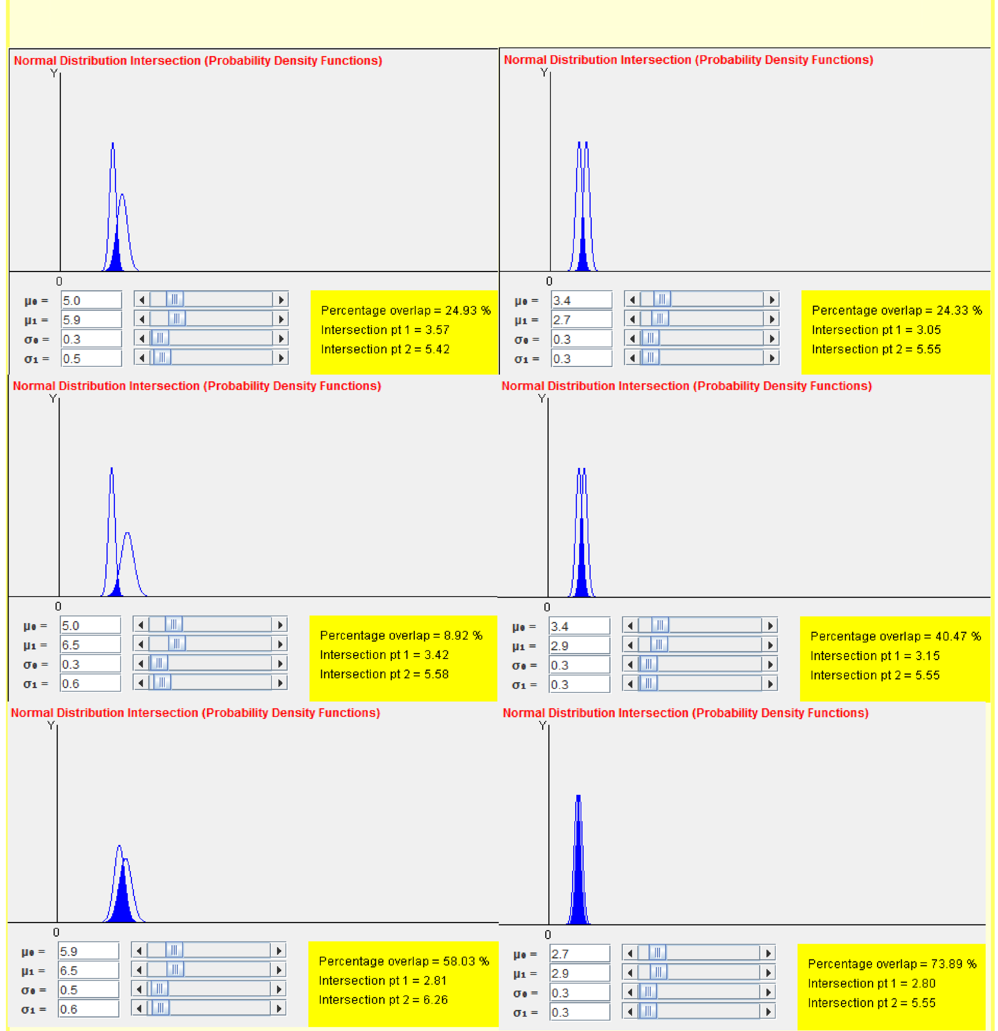
Dataset

- One specific example of a large dataset is Lung Cancer Database Consortium (LIDC) which provides a collection of CT lung images .
- Extracting these semantic contents from large volume dataset of images is expensive because analysis has to be manually performed by radiology experts.
- To Reduce the cost of acquiring radiologist annotations, we propose to reduce the amount of data that needs to be annotated
- Our final goal is to propose a new approach for selecting the most representative subsamples in LIDC without leaving out the important characteristics of data.
- Labeled subsamples can later serve to evaluate Computer-Aided Diagnosis (CAD) systems. Several CAD systems have been developed to help classify malignant nodules versus benign in lung cancer diagnosis.

Ratings and Boundaries across radiologists are different.
National Institutes of Health: "http://imaging.cancer.gov/programsandresources/informationsystems/lidc"

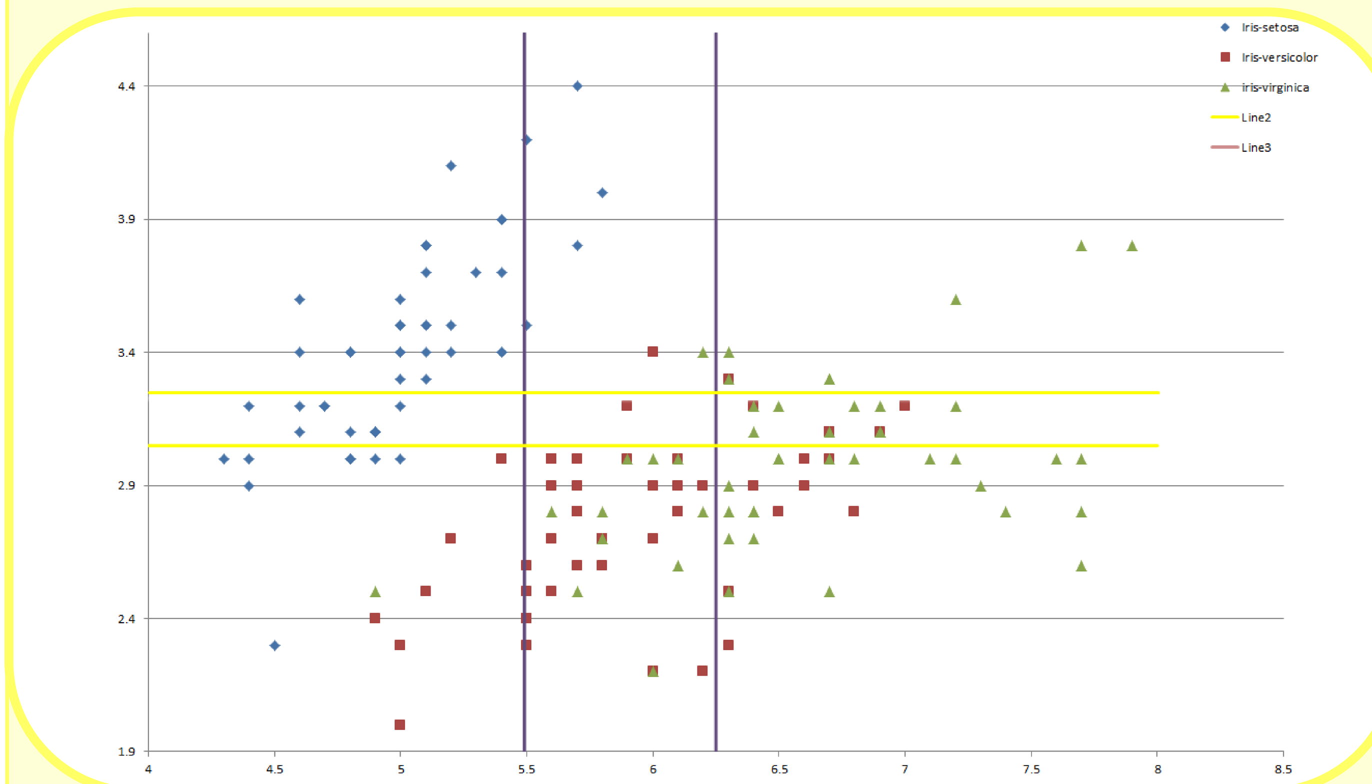
In phase1 of our research, we started with a smaller dataset and supervised classification method to select the most representative objects in dataset. Our approach is following the steps below:

Step 1: Projecting the distribution of samples in each class onto the basis of the attributes.

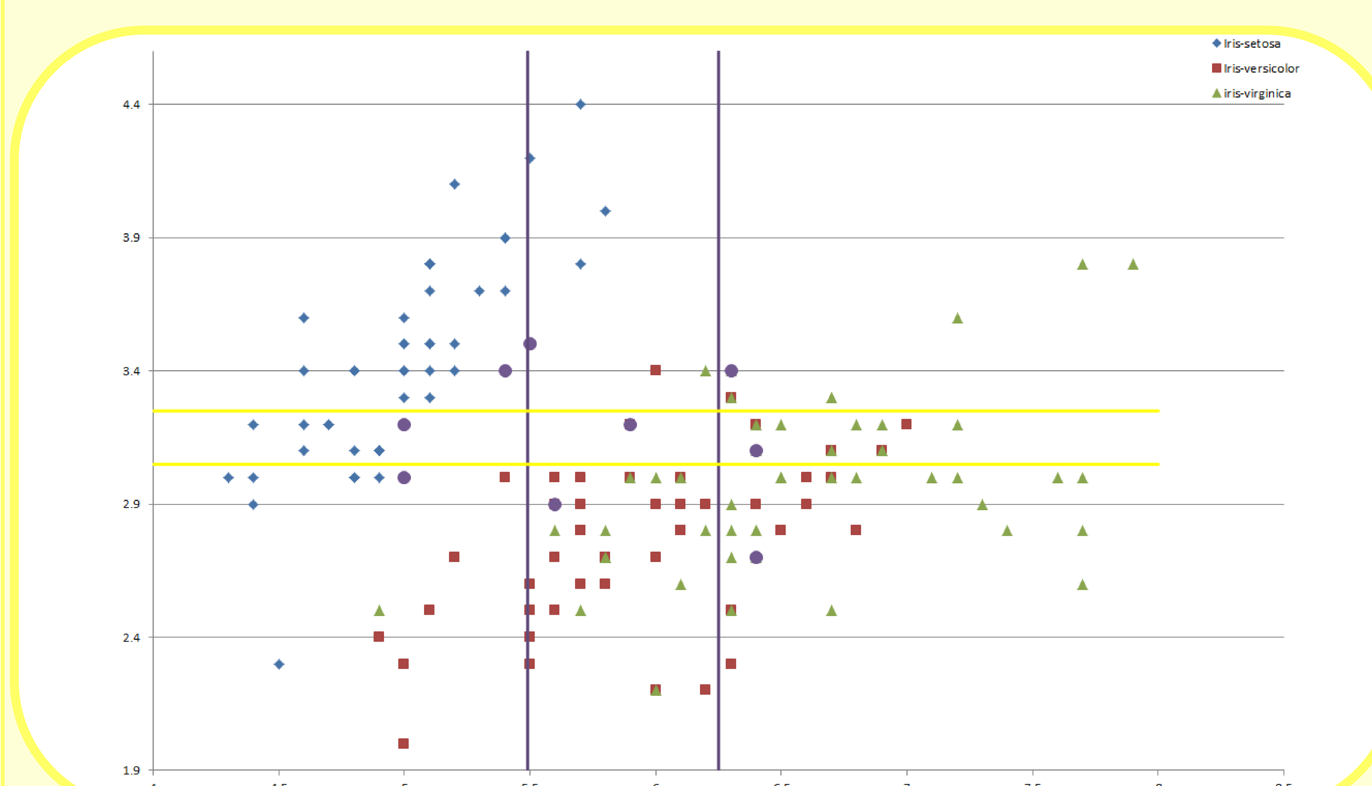


Normal Distribution Intersection of three classes Projected on attribute1 & attribute2

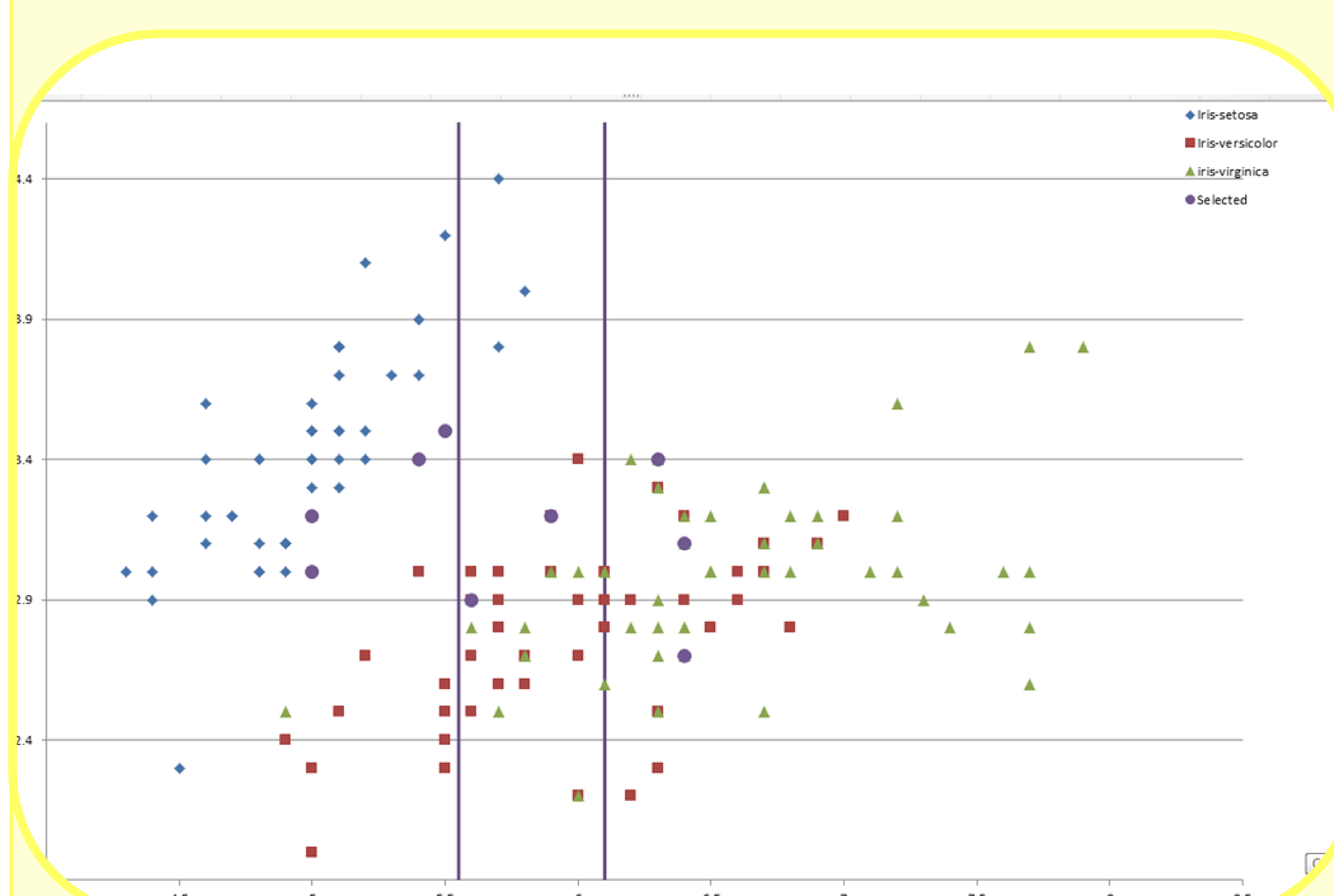
- **Step 2:** Selecting cut off values where pairwise class attribute distributions intersect and draw imaginary discriminant lines to partition the dataset.



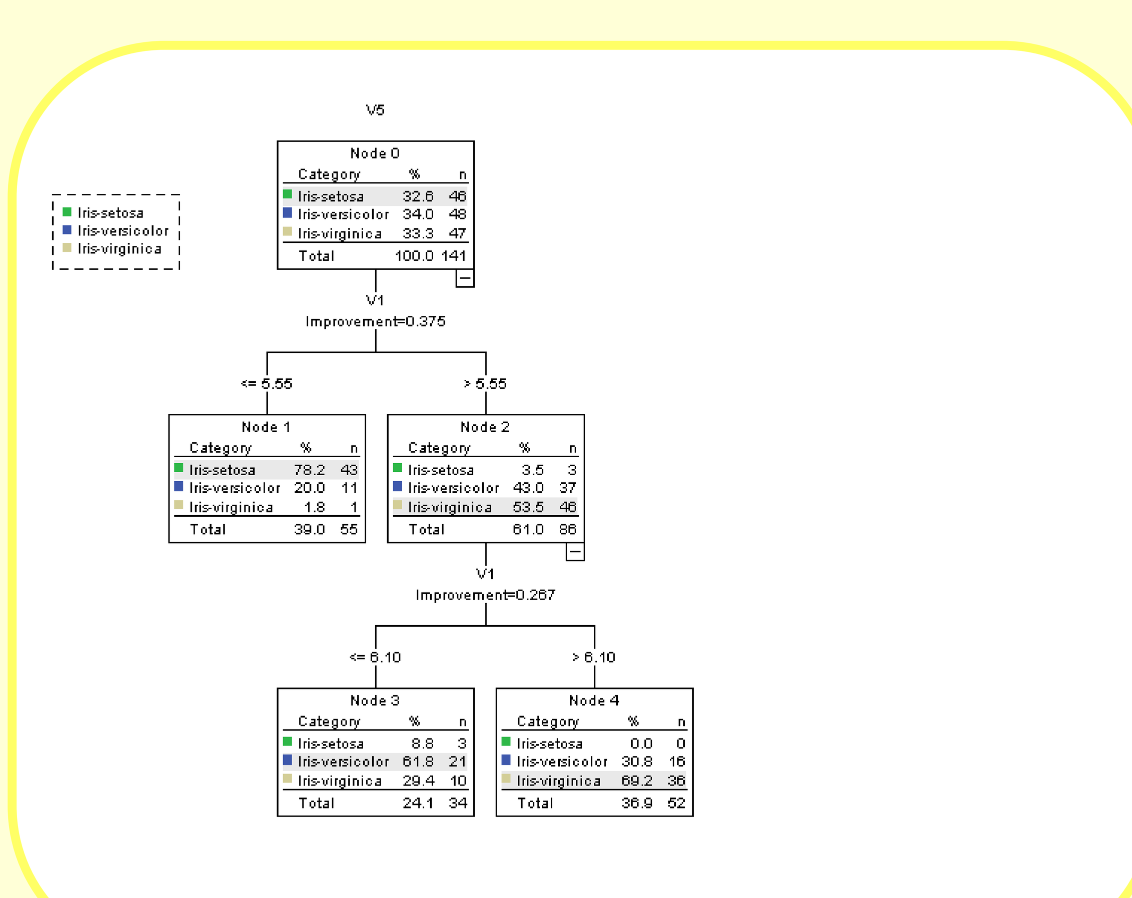
- **Step 3:** Selecting the samples closest to these discriminant lines, selecting an object based on the majority label of the samples in the partition.



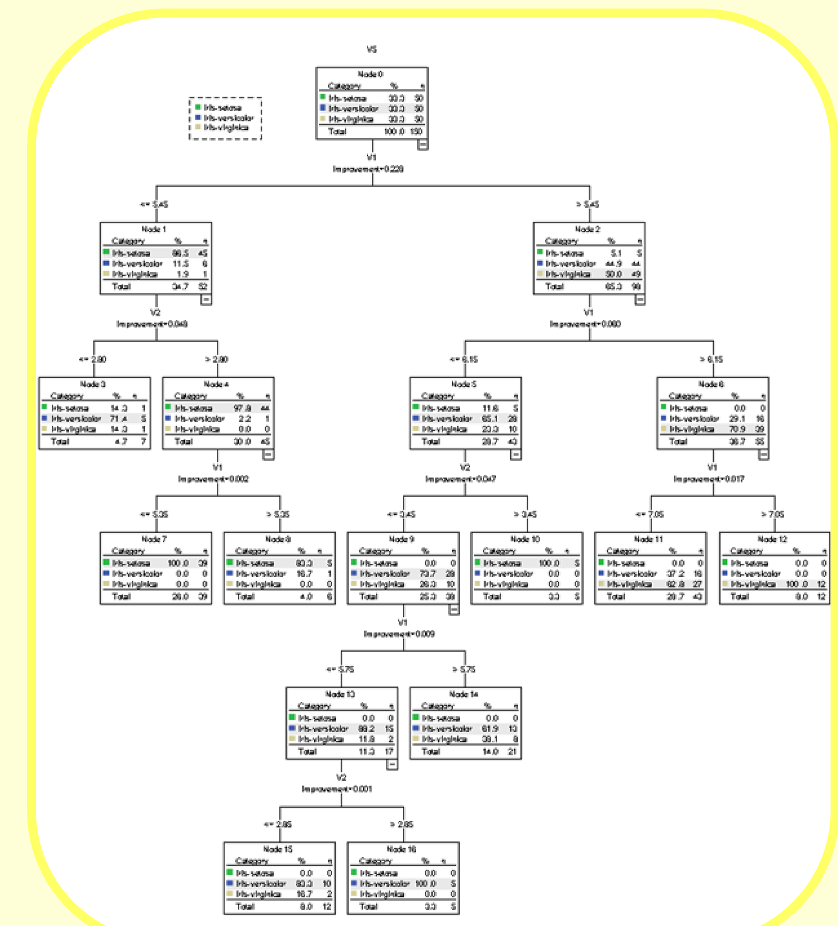
- **Step 4:** Training the classifier with the selected samples.



- **Step 5:** Testing the discriminative power of the classifier on the testing set.



Constructed tree using all 150 samples



- Validation method is cross validation with 10 folders, maximum number of parents set to be 10, maximum number of children is 5 and the depth of the tree is 5.

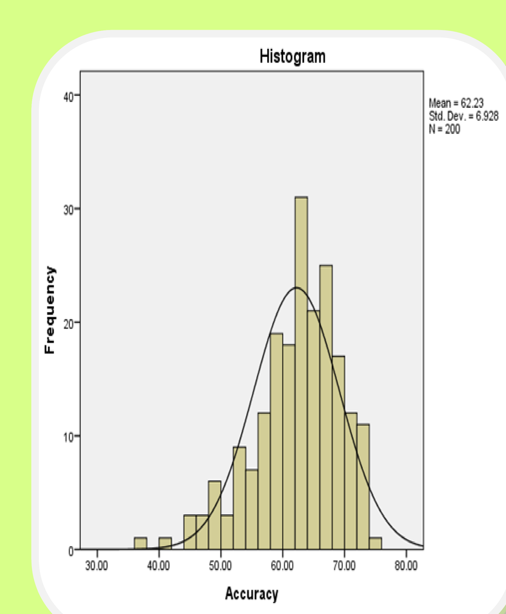
Observed	Predicted			
	Iris-setosa	Iris-versicolor	Iris-virginica	Percent Correct
Iris-setosa	49	1	0	98.0%
Iris-versicolor	1	33	16	66.0%
Iris-virginica	0	11	39	78.0%
Overall Percentage	33.3%	30.0%	36.7%	80.7%

Experimental Results

- **Step 6:** Comparing the results to cost effective number of different combinations of random subset selection of the same size.

Random

Distribution of accuracy for 200 different random selections



Number Of Samples	Number Of Runs	Min	Max	Mean
9	200	34.8%	74.5%	61.1%

Distribution of accuracy on testing data when the classifier is trained using 200 random subsets of size 9. As it is denoted in the plot, the mean accuracy on testing data for 200 different randomly selected samples of size 9.

Our Result

Classification

Sample	Observed	Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Percent Correct
Training	Iris-setosa	3	0	0	100.0%
	Iris-versicolor	0	3	0	100.0%
	Iris-virginica	0	0	3	100.0%
	Overall	33.3%	33.3%	33.3%	100.0%
Test	Iris-setosa	42	5	0	89.4%
	Iris-versicolor	1	30	16	63.8%
	Iris-virginica	0	11	36	76.6%
	Overall	30.5%	32.6%	36.9%	76.6%

The results are based on applying our algorithm to a smaller dataset of 150 data points, 4 features and 3 classes with same number of samples.

The selected classification method is CRT decision tree because no specific data distribution is necessary for these type of classifiers. We set the maximum number of parents to 2, maximum number of children to 1 and the depth of the tree to 5.

Conclusion

- The preliminary results indicate that our approach is able to perform better than the highest accuracy in cost effective number of different random subsets of the same size.
- The mean accuracy on testing data for 200 different randomly selected samples of size 9 is 61.1%. The minimum accuracy obtained is 34.8% while the maximum is 74.5%. Using our proposed approach, the accuracy of the classifier trained by our selected samples is 76.6% on the testing set.

Future Work

- Applying the approach to LIDC dataset;
- Compare our results with uniform and cluster-based design subset selection methods and publish the results in form of a paper.
- Extend the approach for unlabeled data;
- Efficient Implementation of Incremental Tree Induction algorithm and use it in subset selection.