A Genetic Algorithm-Based Approach for Building Accurate Decision Trees

by

Z. Fu, Fannie Mae
Bruce Golden, University of Maryland
S. Lele, University of Maryland
S. Raghavan, University of Maryland
Edward Wasil, American University

Presented at INFORMS National Meeting
Pittsburgh, November 2006

A Definition of Data Mining

- Exploration and analysis of large quantities of data
- By automatic or semi-automatic means
- To discover meaningful patterns and rules
- These patterns allow a company to
  - Better understand its customers
  - Improve its marketing, sales, and customer support operations

Data Mining Activities

- Discussed in this session
  - Classification
    - Decision trees
  - Visualization
    - Discrete models
    - Sammon maps
    - Multidimensional scaling

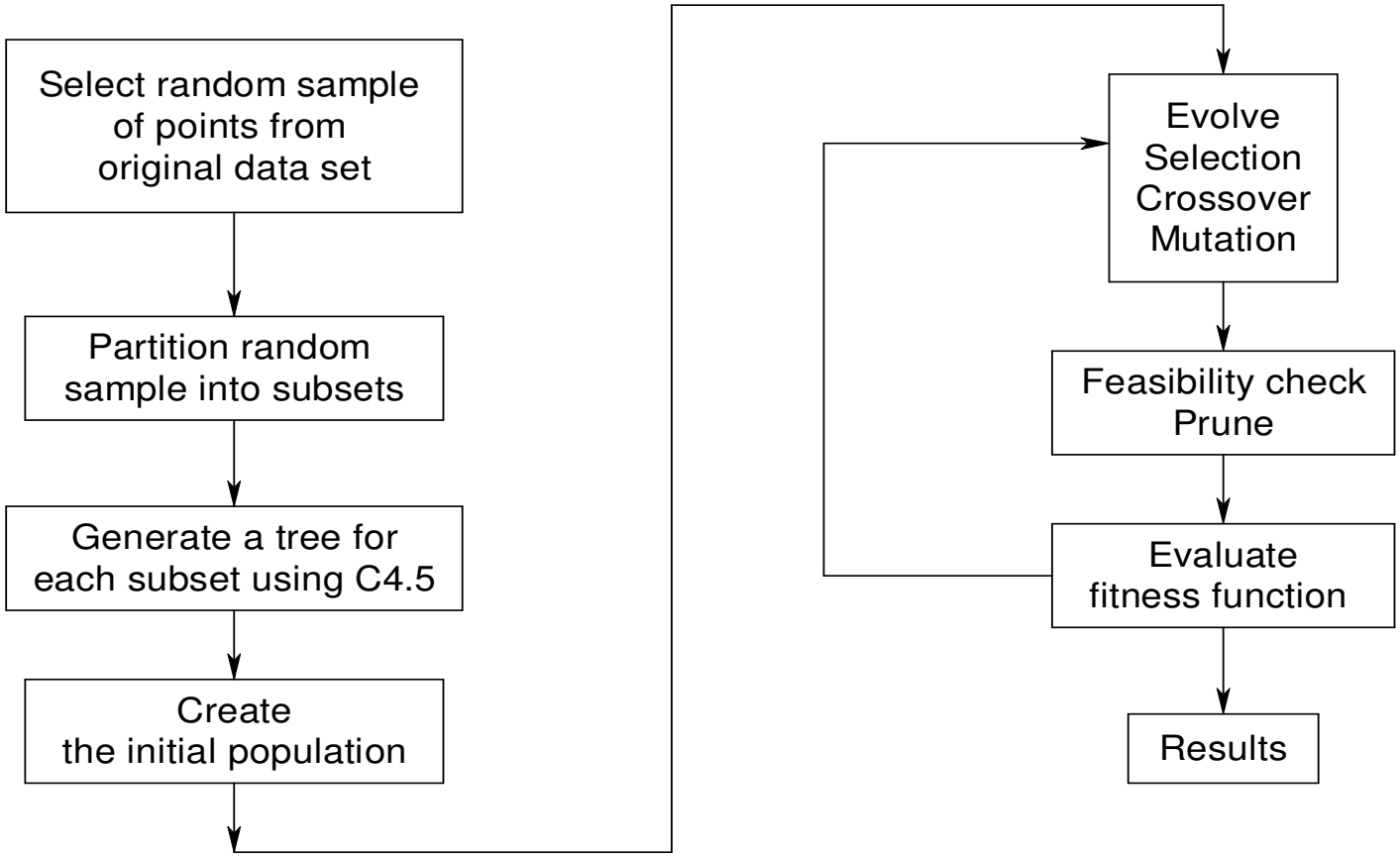
Classification

- Given a number of prespecified classes
- Examine a new object, record, or individual and assign it, based on a model, to one of these classes
- Examples
  - Which credit applicants are low, medium, high risk?
  - Which hotel customers are likely, unlikely to return?
  - Which residents are likely, unlikely to vote?

Background

- A decision tree is a popular method for discovering meaningful patterns and classification rules in a data set
- With very large data sets, it might be impractical, impossible, or time consuming to construct a decision tree using all of the data points
- Idea: combine C4.5, statistical sampling, and genetic algorithms to generate decision trees of high quality
- We call our approach GAIT- a Genetic Algorithm for Intelligent Trees

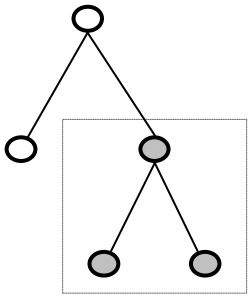
Flow Chart of GAIT



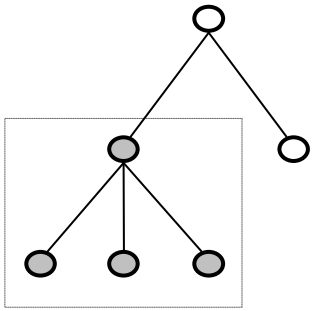
GAIT Terminology

- The initial population of trees is generated by C4.5 on the subsets from the partitioned random sample
- Trees are randomly selected for the crossover and mutation operations, proportional to tree fitness
- Fitness = the percentage of correctly classified observations in the scoring data set
- Crossover and mutation are illustrated next

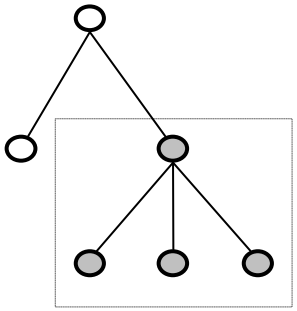
Crossover Operations



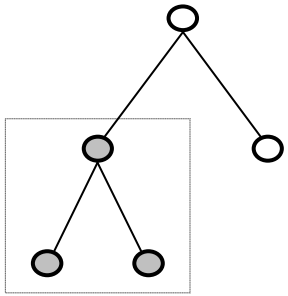
Parent 1



Parent 2

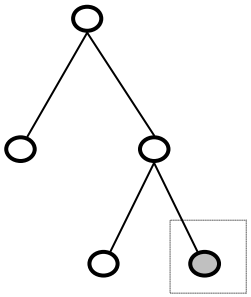


Child 1

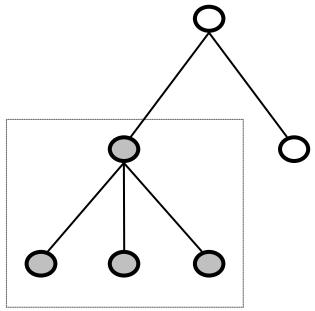


Child 2

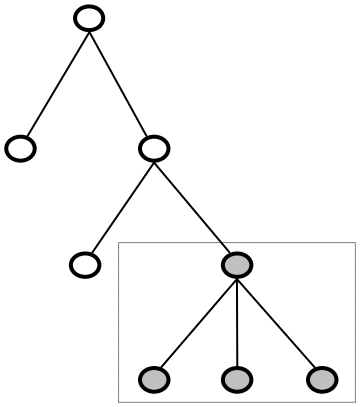
Subtree-to-subtree Crossover



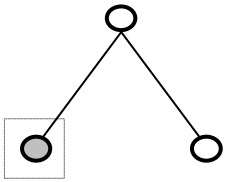
Parent 1



Parent 2



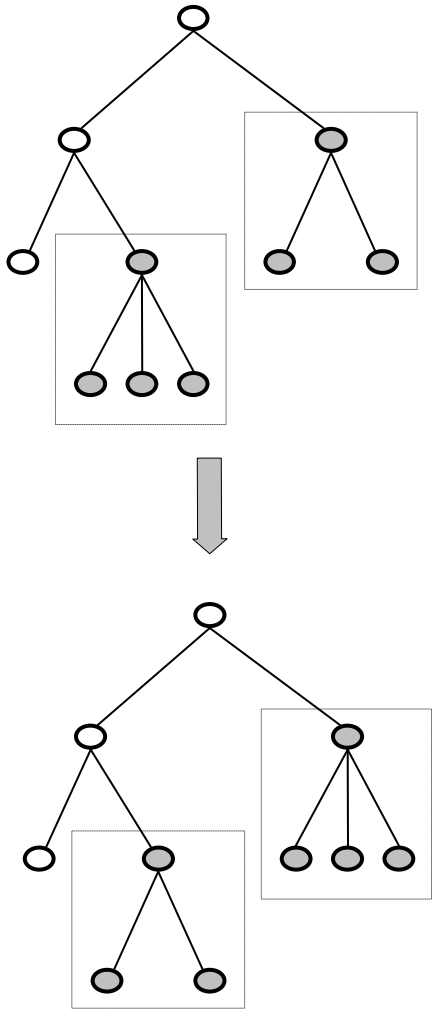
Child 1



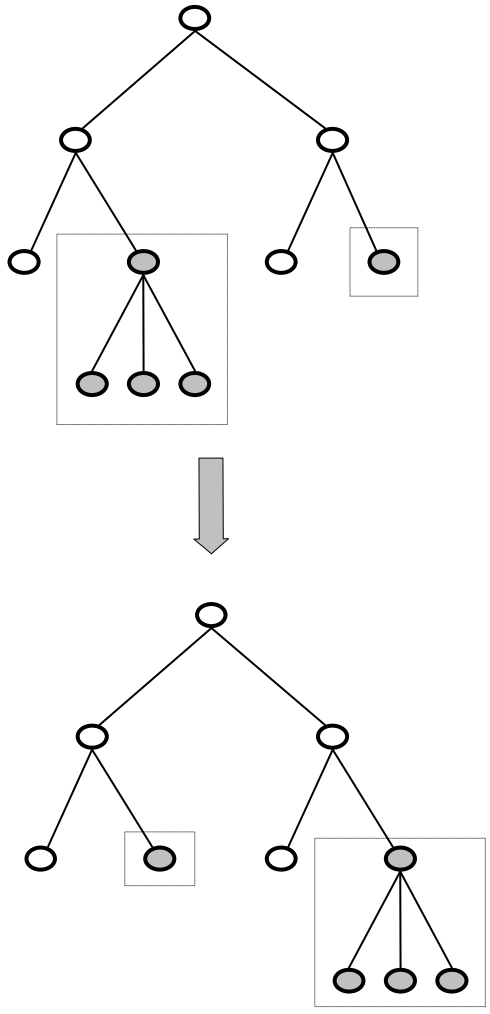
Child 2

Subtree-to-leaf Crossover

Mutation Operations

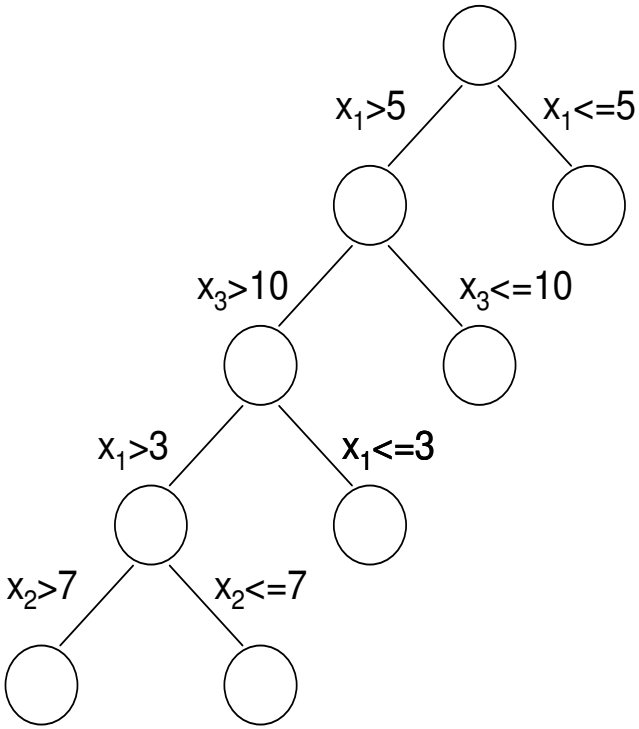


Subtree-to-subtree Mutation

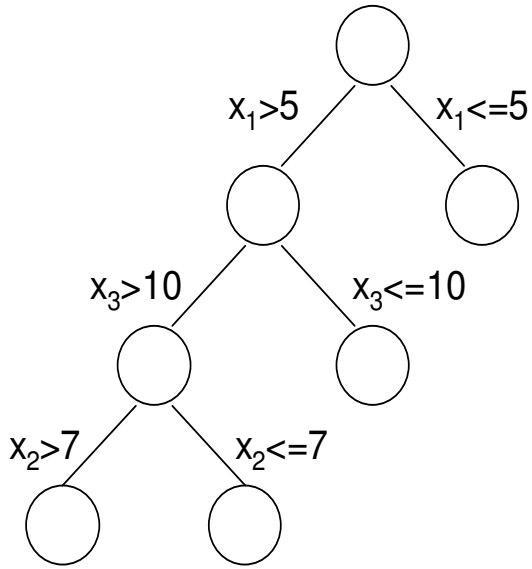


Subtree-to-leaf Mutation

Feasibility Check

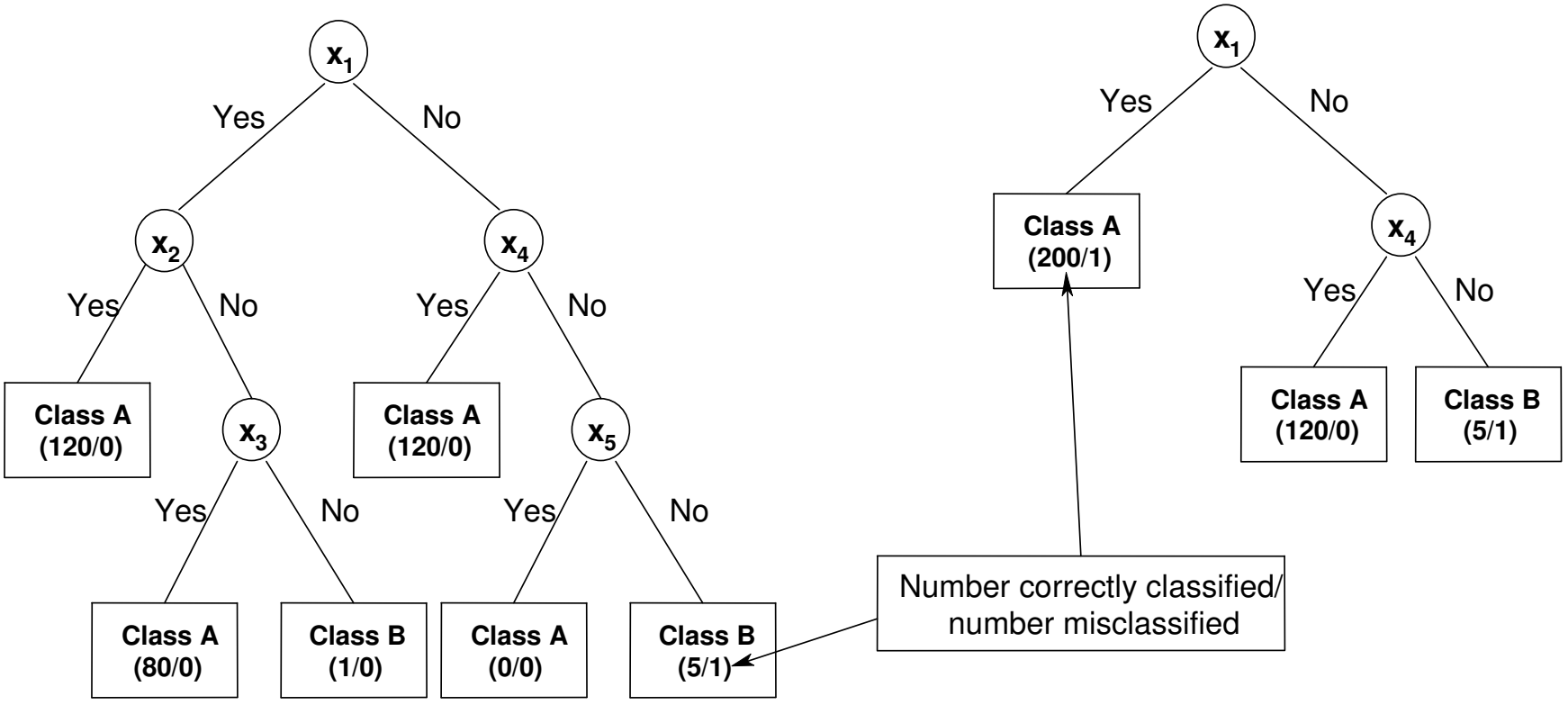


Before feasibility check



After feasibility check

Pruning



Before pruning

After pruning

Computational Experiments

- Real-life marketing data set in transportation services industry
- Approx. 440,000 customers with demographic and usage information
- Key question: Do the customers reuse the firm's services following a marketing promotion?
- The goal, then, is to predict repeat customers
- In the data set, 21.3% are repeat customers

Computational Experiments - - continued

- We focused on 11 variables
- We seek to identify customers who will respond positively to the promotion
- The primary performance measure is the overall classification accuracy
- We coded GAIT in Microsoft Visual C++ 5.0
- Experiments were run on a Windows 95 PC with a 400 MHz Pentium II processor and 128 MB of RAM

Experimental Design

- Training set of 3,000 points
- Scoring set of 1,500 points
- Test set of 1,000 points
- The test set is not available in the development of any classifiers
- The combined size of the training and scoring sets is approx. 1% of the original data
- We designed three different experiments

Experiment 1

- Partition the 3,000 points in the training set into 50 subsets of 60 points each
- From each subset, obtain a decision tree using C4.5
- The GA runs for 10 generations
- Save the best 50 trees at the end of each generation
- The GAIT tree is the one with the highest fitness at the end
- Finally, we computed the accuracy of the best GAIT tree on the test set

Experiments 2 and 3

- Experiment 2
  - Fewer initial trees
  - 10 subsets of 60 points each ⇒ 10 initial trees
  - In the end, we compute the accuracy of the best generated decision tree on the test set
- Experiment 3
  - Low-quality initial trees
  - Take the 10 lowest scoring trees (of 50) from the first generation of Experiment 1
  - In the end, we compute the accuracy of the best generated decision tree on the test set

Details of Experiments

- Two performance measures
  - Classification accuracy on the test set
  - Computing time for training and scoring
- Five different decision trees
  - Logistic Regression (SAS)
  - Whole-Training
  - Best-Initial
  - Aggregate Initial
  - GAIT

Decision Tree Classifiers

- The Whole-Training tree is the C4.5 tree generated from the entire training set
- The Best-Initial tree is the best tree (w.r.t. scoring set) from the first generation of trees
- The Aggregate-Initial classifier uses a majority voting rule from the first generation of trees
- The GAIT tree is the result of our genetic algorithm

Classification Accuracy on the Test Set

Size	Method	Experiment 1	Experiment 2	Experiment 3
500	Logistic Regression	0.7563	0.7412	0.7228
	Whole-Training	0.7612	0.7491	0.7226
	Best-Initial	0.7525	0.7392	0.6767
	Aggregate-Initial	0.7853	0.7784	0.7687
	GAIT	0.7903	0.7854	0.7787
1000	Logistic Regression	0.7627	0.7432	0.7317
	Whole-Training	0.7595	0.7457	0.7316
	Best-Initial	0.7557	0.7394	0.6778
	Aggregate-Initial	0.7849	0.7775	0.7661
	GAIT	0.7910	0.7853	0.7784
1500	Logistic Regression	0.7598	0.7478	0.7305
	Whole-Training	0.7603	0.7495	0.7312
	Best-Initial	0.7527	0.7398	0.6791
	Aggregate-Initial	0.7830	0.7790	0.7691
	GAIT	0.7898	0.7844	0.7756

Note: Average accuracy from ten replications. The left-most column gives the size of the scoring set.

Computing Time (in Seconds) for Training and Scoring

Size	Method	Experiment 1	Experiment 2	Experiment 3
500	Logistic Regression	0.94	0.53	0.57
	Whole-Training	2.07	1.23	1.35
	Best-Initial	1.34	0.54	0.59
	Aggregate-Initial	2.17	1.33	1.35
	GAIT	16.70	8.15	8.14
1000	Logistic Regression	0.95	0.55	0.59
	Whole-Training	2.12	1.29	1.40
	Best-Initial	1.39	0.57	0.63
	Aggregate-Initial	2.17	1.19	1.44
	GAIT	31.26	14.38	14.40
1500	Logistic Regression	1.02	0.54	0.61
	Whole-Training	2.14	1.37	1.45
	Best-Initial	1.44	0.59	0.68
	Aggregate-Initial	2.07	1.28	1.51
	GAIT	45.70	20.77	20.74

Note: Average time from ten replications. The left-most column gives the size of the scoring set.

Computational Results

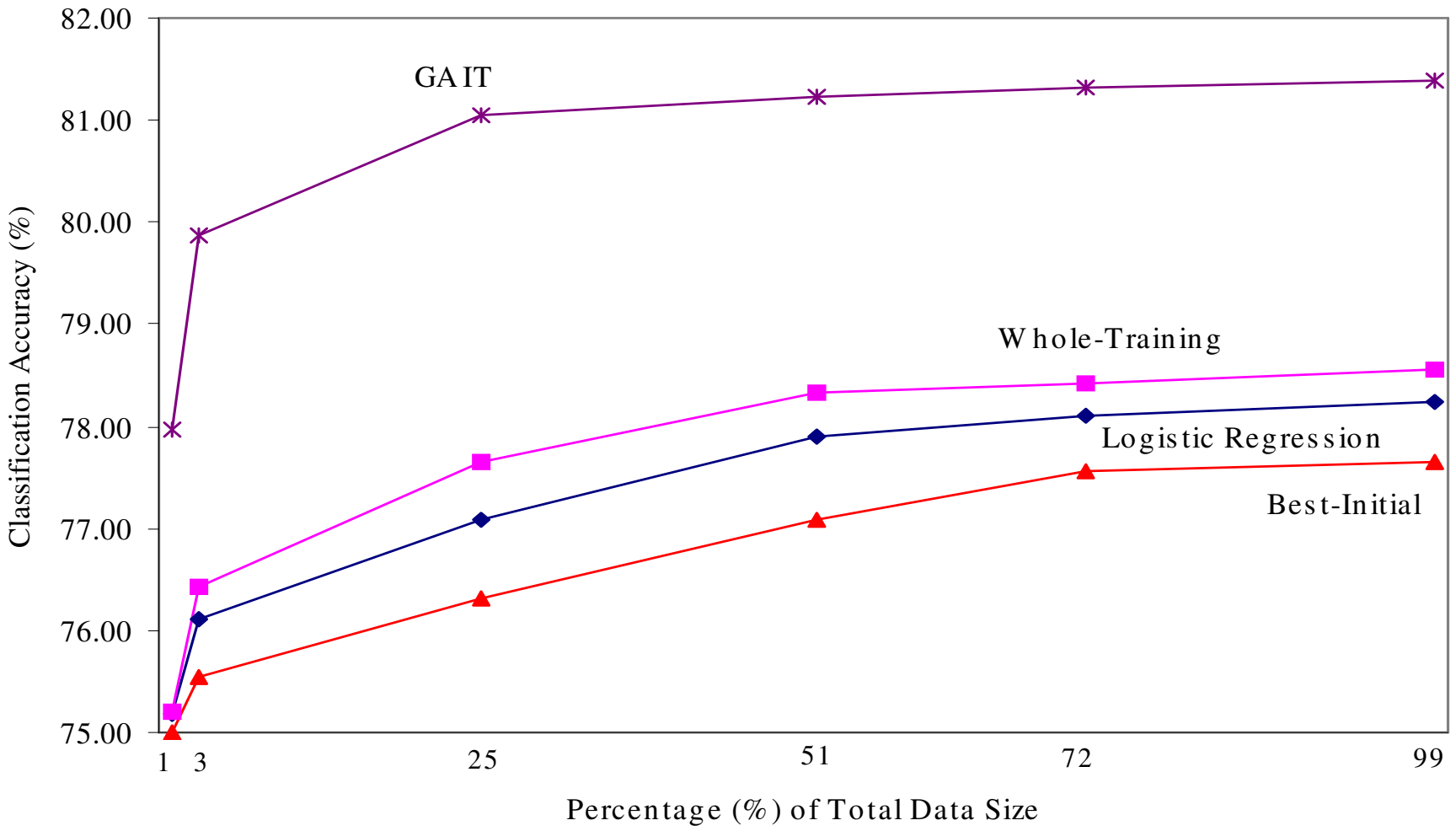
- In general, GAIT outperforms Aggregate-Initial which outperforms Whole-Training which outperforms Logistic Regression which outperforms Best-Initial
- The improvement of GAIT over non-GAIT procedures is statistically significant in all three experiments
- Regardless of where you start, GAIT produces highly accurate decision trees
- We experimented with a second data set with approx. 50,000 observations and 14 demographic variables and the results were the same

Scalability

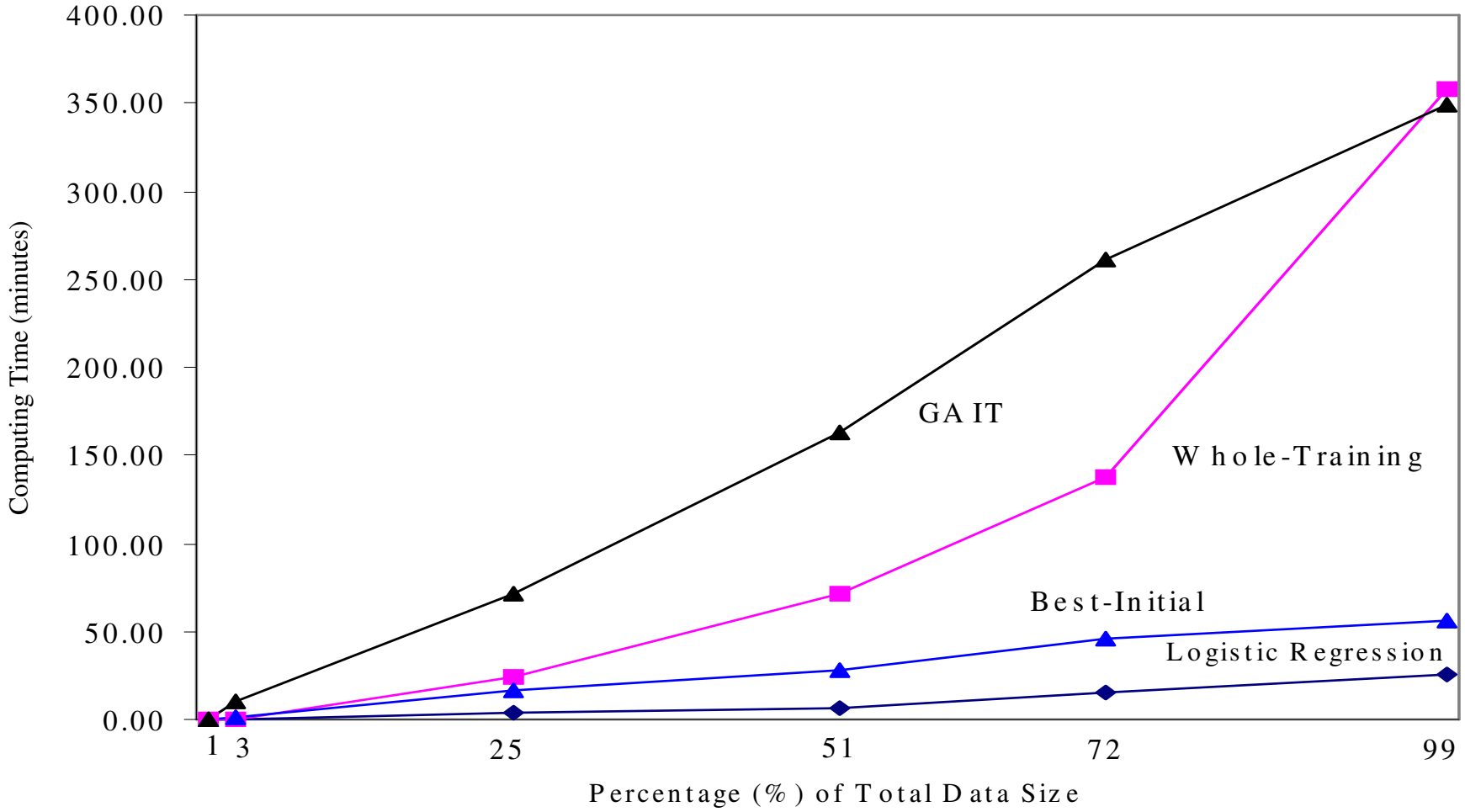
- We increase the size of the training and scoring sets while the size of the test set remains the same
- Six combinations are used from the marketing data set

Percent (%)	Size	
	Training Set	Scoring Set
99	310,000	124,000
72	240,000	96,000
51	160,000	64,000
25	80,000	64,000
3	10,000	4,000
1	3,000	1,500

Classification Accuracy vs. Training/Scoring Set Size



Computing Time for Training and Scoring



Computational Results

- GAIT generates more accurate decision trees than Logistic Regression, Whole-Training, Aggregate-Initial, and Best-Initial
- GAIT scales up reasonably well
- GAIT (using only 3% of the data) outperforms Logistic Regression, Best-Initial, and Whole Training (using 99% of the data) and takes less computing time

Conclusions

- GAIT generates high-quality decision trees
- GAIT can be used effectively on very large data sets
- The key to the success of GAIT seems to be the combined use of sampling, genetic algorithms, and C4.5 (a very fast decision-tree package from the machine-learning community)
- References
  - INFORMS Journal on Computing, 15(1), 2003
  - Operations Research, 51(6), 2003
  - Computers & Operations Research, 33(11), 2006