Background and Analysis of Experimental Results

Adam Nelson

January 1, 2010

Contents

1 Building the Experiment

To construct the experiment, certain aspects were first determined to be pertinent in the final selection of top, actionable attributes in the data. The following represents brief explanations of each method used. Results obtained from a combination of which are then analyzed.

1.1 Number of Attributes

An attribute in the data could be something such as GPA, or ZIPCODE. The number of attributes to select is crucial in the analysis of the data, because it allows us to conclude how many of the attributes selected we should concentrate on. This is central in selecting actionable attributes. For example, suppose a data set consists of 1000 attributes, but the results from experimentation find that only 15 of the 1000 are actually important. The bulk of subsequent attention could then be spent on what actions to take based on the 15 found, as opposed to the rest of the 985.

In this experiment, we chose n to be the number of attributes selected in increments of 5. Thus, with a maximum of 103 attributes in each data set used in the experiment, 20 different intervals of n were chosen by our feature subset selectors (described below).

1.2 Classifiers

Classifiers are used in data mining by employing machine learning techniques in order to learn patterns in data. Once these patterns are learned, we can then begin to attempt to predict outcomes in the data by reflecting on data that has already been examined. We can also determine how well a classifier predicts for the data. This is done by learning on a certain portion of the data, and reflecting on how well predictions are made by another portion of the data that has not yet been seen in the learning process. By examining overall performance, we can make a statement about how much better one classifier predicts on a specific data set than another.

- Naive Bayes - A naive Bayes classifier is a simple and fast probabilistic classifier that uses Bayes' theorem to classify training data. Bayes' theorem, as shown in Equation 1, determines the probability P of an event H occurring given an amount of evidence E. The classifier also assumes feature independence; the algorithm examines features independently to contribute to probabilities, as opposed to the assumption that features depend on other features. Surprisingly, even though feature independence is an integral part of the classifier, it often outperforms many other learners [11].

$$Pr(H|E) = \frac{Pr(E|H) * Pr(H)}{Pr(E)} \tag{1}$$

- C4.5 - C4.5 [10] is a type of classifier known as a decision tree, and is an extension to the ID3 [9] algorithm. A decision tree [7] (shown in Figure 1) is constructed by first determining the best attribute to make as the root node of the tree. ID3 decides this root attribute by using one that best classifies training examples based upon the attribute's information gain (described below). Then, for each value of the attribute representing any node in the tree, the algorithm recursively builds child nodes based on how well another attribute from the data describes that specific branch of its parent node. The stopping criteria are either when the tree perfectly classifies all training examples, or until no attribute remains unused. C4.5 extends ID3 by making several improvements, such as the ability to operate on both continuous as well as discrete attributes, training data that contains missing values for a given attribute(s), and employ pruning techniques on the resulting tree.

- One-R - One-R, described in [6], builds rules from the data by iteratively examining each value of an attribute and counting the frequency of each class for that attribute-value pair. An attribute-value is then assigned the most frequently occurring class. Error rates of each of the rules can then be calculated, and the best rules can be ranked based on the lowest error rates.

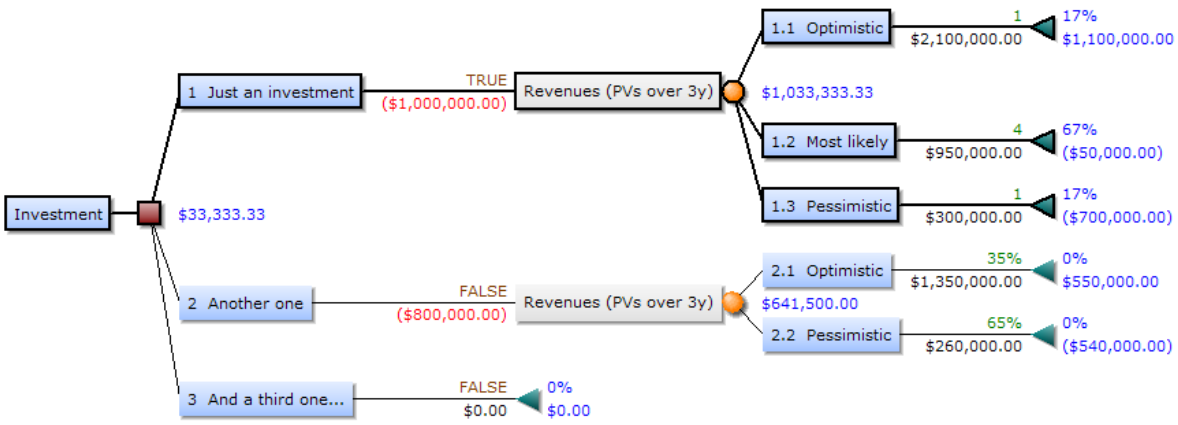

Figure 1: A decision tree consists of a root node and descending children nodes who denote decisions to make in the tree's strucure. This tree, for example, was constructed in an attempt to optimize investment portfolios by minimizing budgets and maximizing payoffs. The top-most branch represents the best selection in this example.

- Zero-R - Often used to evaluate the success of other classification algorithms, Zero-R is an extremely simple algorithm that gives the majority class from the training data.

- Alternating Decision Trees - ADTrees [3] are decision trees that contain both decision nodes, as well as prediction nodes. Decision nodes specify a condition, while prediction nodes contain only a number. Thus, as an example in the data follows paths in the ADTree, it only traverses branches whose decision nodes are true. The example is then classified by summing all prediction nodes that are encountered in this traversal. ADTrees, however, differ from binary classification trees, such as C4.5, in that in those trees an example only traverses a single path down the tree.

- Bayesian Network - Bayesian networks, illustrated in Figure 2, are graphical models that use a directed acyclic graph (DAG) to represent probabilistic relationships between variables. As stated in [5] Bayesian networks have four important elements to offer:

  1. Incomplete data sets can be handled well by Bayesian networks. Because the

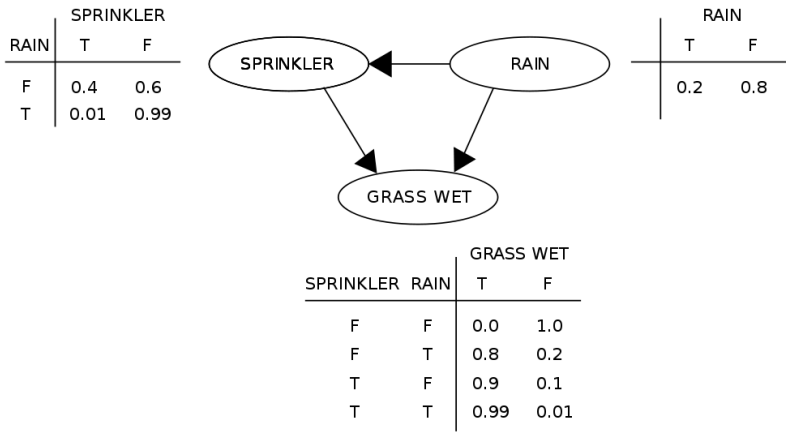

Figure 2: In this simple bayesian network, the variable *Sprinkler* is dependent upon whether or not its raining; the sprinkler is generally not turned on when it's raining. However, either event is able to cause the grass to become wet - if it's raining, or if the sprinkler is caused to turn on. Thus, Bayesian networks excel at investigating information relating to relationships between variables.

networks encode a correlation between input variables, if an input is not observed, in will not necessarily produce inaccurate predictions, as would other methods.

2. Causal relationships can be learned about via Bayesian networks. For instance, if an analyst wished to know if a certain action taken would produce a specific result, and also to what degree.

3. Bayesian networks promote the amalgamation of data and domain knowledge by allowing for a straightforward encoding of causal prior knowledge, as well as the ability to encode causal relationship strength.

4. Bayesian networks avoid over fitting of data, as "smoothing" can be used in a way such that all data that is available can be used for training.

- Radial Basis Function Network - A radial basis function network (RBFN) [1] is a type of network called an artificial neural network (ANN). However, RBFNs are specialized in that they utilize a radial basis function as an activation function. An ANN's activa-

tion function is used in order to offer non-linearity to the network. This is important for multi-layer networks containing many hidden layers, because their advantages lie in their ability to learn on non-linearly separable examples.

1.3 Feature Subset Selectors

Feature Subset Selection (FSS) methods provide ways to determine how important the attributes (or features) are in the data set, and how we can keep the best scoring ones, and throw out the rest. However, we must experiment with varying FSS procedures, because each method can return strikingly different results. Thus, just by experimenting with attributes selected from a handful of FSS, we are not left with a sense of how well attributes were selected from a data set compared to other feature selection tools.

A brief overview of the FSS methods used in this study were as follows:

- CFS - Correlation-Based Feature Selection [4] begins by constructing a matrix of feature to feature, and feature-to-class correlations. It then uses a best first search by expanding the best subsets until no improvement is made, in which case the search falls to the unexpanded subset having the next best evaluation until a subset expansion limit is met.

- Information Gain - Information Gain works by using a concept from information theory known as entropy. Entropy measures the amount of uncertainty, or randomness, that is associated with a random variable. Thus, high entropy can be seen as a lack of purity in the data. Information gain, as described in [8] is an expected reduction of the entropy measure that occurs when splitting examples in the data using a particular attribute. Therefore an attribute that has a high purity (high information gain) is better at describing the data than one with a low purity. The resulting attributes are then ranked by sorted their information gain scores in a descending order.

- Chi-squared - Attributes can also be ranked using the chi-squared statistic. The

chi-squared statistic [2] is used in statistical tests to determine how distributions of variables are different from one another. Note that these variables must be categorical in nature. Thus, the chi-squared statistic can evaluate an attribute's worth by calculating the value of this statistic with respect to a class. Attributes can then be ranked based on this statistic.

- One-R - One-R (as described above), can also be used to deliver top-ranking attributes. Since each rule contains one attribute and a corresponding value, we can evaluate attributes by sorting them based on the error rate of the rule associated with that attribute-value pair. Using this, top attributes are those whose rules result in the lowest error rates.

1.4 Cross-Validation

In the process of experimentation, it is crucial to determine a method's performance. Using performance criteria, further analysis can be conducted on experimental results to aid in the search for an optimal solution. Cross-validation provides the ability to discover how well a classifier performs on any given data set or a treatment of that data set. This is conducted by randomly partitioning the data into two subsets, called the training set, and the testing set. Specifically for this experiment, the data prior to partitioning has been reduced given n attributes selected using an FSS method.

In the learning phase, only the training subset is used by the classifier. The testing set is then used to determine how well the concepts learned from the training phase can be applied to unseen data. However, to reduce variability, the partitioning of the data and reclassification of resulting subsets is generally conducted multiple times. In this experiment, for example, a 5 X 5 cross-validation was performed. This means that five times we partitioned the data into a testing set consisting of $\frac{1}{5}$-th of the data, and a training set of $\frac{4}{5}$-ths of the data. After the five rounds, median values of the validation results are examined, and are assigned to a particular combination of the above facets.

2 Analysis of Experimental Results

2.1 Evaluation Metrics

The evaluation metrics used in this experiment are standard in data mining to measuring the performance of a method. These are represented as probability of detection (PD), probability of false alarm (PF), and variance. PD denotes the probability that the classifier will predict correctly for a given class, given both its correct and incorrect predictions. Thus, PD values should be maximized. PF, on the other hand, is the probability that the classifier will predict incorrectly for a given class, also given its correct and incorrect predictions. For this reason, PF results should be minimized.

Variance was also used in the experiment based on PD and PF values independently as an extra means of determining performance. Variance in these values provides insight into how much reliability a classifier supports on the data. For example, if a method's PD values range from very low to very high, we can determine that the particular method is not consistent in its probabilities of detection. Therefore, it is desired to have a very small variance in both PD and PF values.

2.2 Visualizing the Results

Figures 3, 4, and 5 show the PD and PF median results for first, second and third year retention against the variance of these values. Each point represents a specific combination of the number of attributes selected, the feature subset selector used to select them, and the classifier used to train on the resulting data. For example, one point on a graph could be seen as 50/Information Gain/Naive Bayes, where 50 denotes the number of attributes used. The color of each point shows the number of attributes used for that particular combination representing that point.

The horizontal line segmenting the PD graphs are given as a baseline reference designated by the already existing retention rates in the data. Thus, to predict for retention

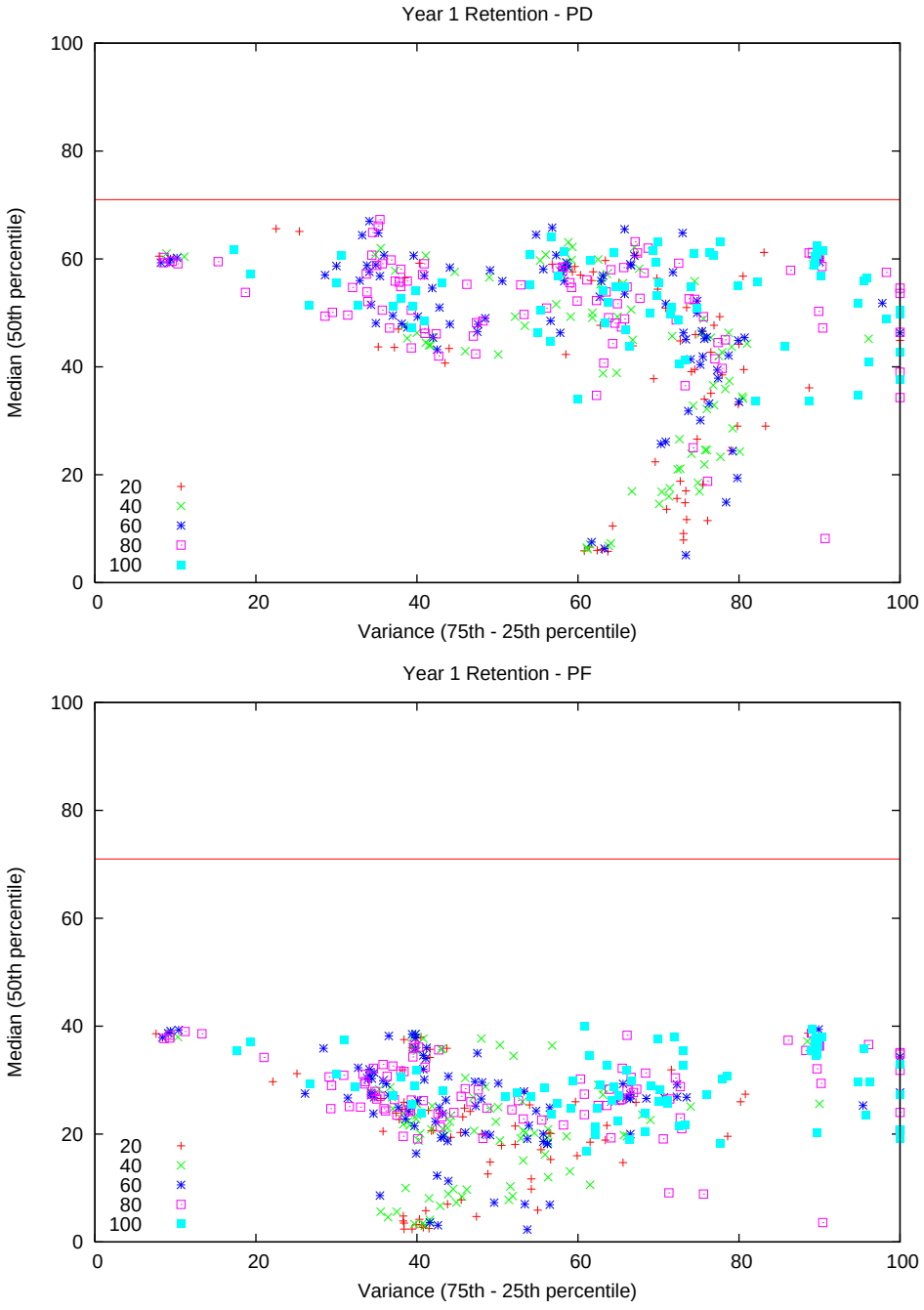

Figure 3: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for first year retention.

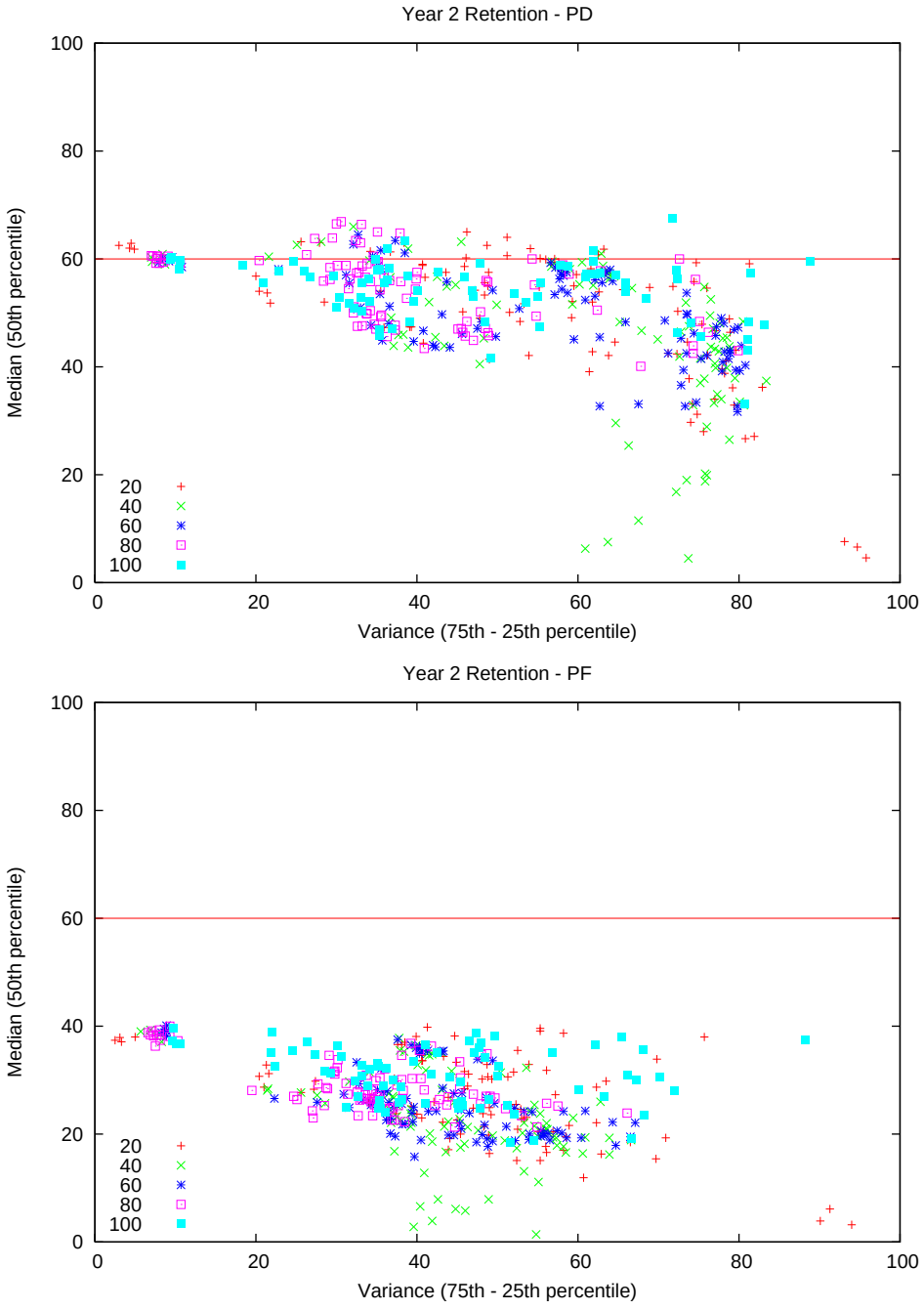

Figure 4: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for second year retention.

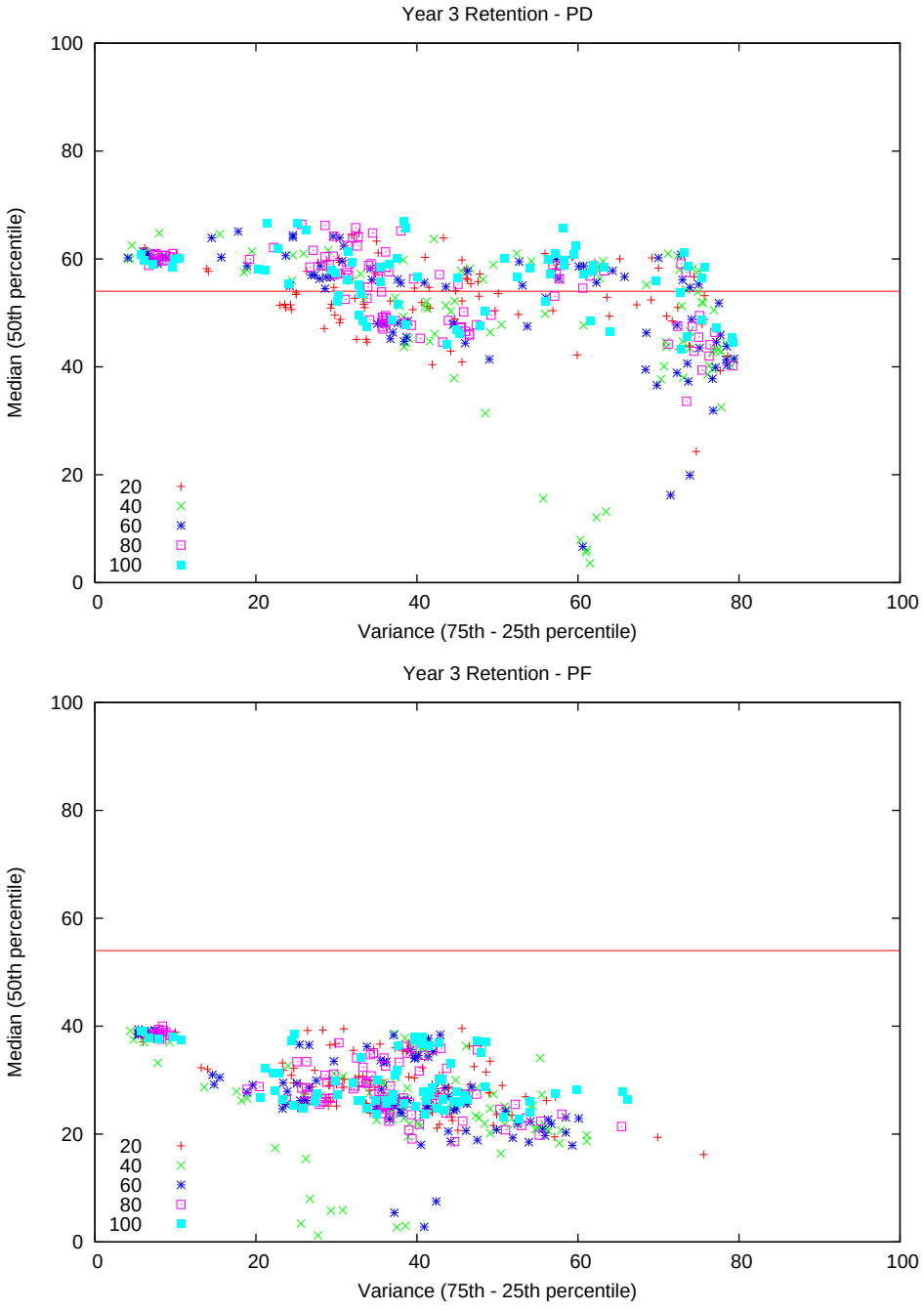

Figure 5: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for third year retention.

in a given year, it is desirable to yield results higher than the baseline. As can be seen in the figures, the median probability of detection of retention values for the first year do not meet the baseline, and therefore we can assume that first year retention cannot accurately be predicted for using our methods. Second year retention provides better results than first year retention, but these results are hardly significant. For example, most of the points lie at or below the baseline. For this reason, second year retention is also not considered in further analysis. Lastly, third year PD values successfully exceed the baseline, and so require more thorough examination.

2.3 Narrowing the Search

From the visualizations described above, we can narrow our space of possible combinations to examine for third year retention. The graphs for PD and PF medians show that the range of number of attributes that maximizes PD and minimizes PF values while maintaining minimal variance is approximately 20 to 60. This is significant, as it allows filtering of the results so that concentration can be placed on only treatments whose attribute numbers lie in this range.

2.3.1 Ranking with the Mann-Whitney Test

At the moment of pruning the results based on attribute ranges, we are left with many combinations to be analyzed. In order to rank each combination, we performed a statistical Mann-Whitney test at 95% confidence in order to rank a treatment. A rank is determined by how many times a combination wins compared to another. The method that won the most number of times is then given the highest rank. The table in Figure 6 shows the top ten ranking combinations based on a PD performance measure. Note that identical ranks are given to those treatments whose win value is equal in magnitude.

Rank	Number of Attributes	FSS	Classifier
61	30	oneR	bnet
61	50	cfs	adtree
57	50	oneR	adtree
56	30	oneR	adtree
55	30	cfs	adtree
52	50	oneR	bnet
51	30	infogain	adtree
51	30	cfs	bnet
48	50	infogain	adtree

Figure 6: The top ten ranking treatments for third year retention. Ranks represent how many times a particular treatment wins over all other treatments in the experiment.

2.4 Selected FSS and Classifier

Figure 6 shows the top-most ranking combination of FSS and classifier is obtained by either using 30 attributes, or 50. Since, the two numbers of attributes (along with their own FSS and classifier) result in the same Mann-Whitney rank, we can make the statement that the two are statistically similar, and thus by focusing on only 30 attributes selected, we can concentrate on approximately 1/3 of the original data. Thus, our analysis of the results show that 30 attributes selected using One-R as the feature subset selection method are the most critical to third year retention.

References

[1] Adrian Bors. *Introduction of the Radial Basis Function (RBF) Networks.*

[2] William Notz David Moore. *Statistics: concepts and controversies.* 2006.

[3] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *In Machine Learning: Proceedings of the Sixteenth International Conference*, pages 124–133. Morgan Kaufmann, 1999.

[4] Mark A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. pages 359–366. Morgan Kaufmann, 2000.

[5] David Heckerman. A tutorial on learning with bayesian networks. 1996.

[6] R.C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63, 1993.

[7] Tom M. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997.

[8] Tom M. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997.

[9] J. Ross Quinlan. *Induction of decision trees.* 1 edition, 1986.

[10] J. Ross Quinlan. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning).* Morgan Kaufmann, 1 edition, January 1993.

[11] Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI-01 workshop on "Empirical Methods in AI".*