Confusion Matrix-based Feature Selection

Sofia Visa
Computer Science Department
College of Wooster
Wooster, OH
svisa@wooster.edu

Brian Ramsay
Sentry Data Systems, Inc.
Fort Lauderdale, FL
brian.ramsay@gmail.com

Anca Ralescu
Computer Science Department
University of Cincinnati
Cincinnati, OH
anca.ralescu@uc.edu

Esther van der Knaap
Ohio Agricultural Research
and Development Center
The Ohio State University
Wooster, OH
vanderknaap.1@osu.edu

Abstract

This paper introduces a new technique for feature selection and illustrates it on a real data set. Namely, the proposed approach creates subsets of attributes based on two criteria: (1) individual attributes have high discrimination (classification) power; and (2) the attributes in the subset are complementary - that is, they misclassify different classes. The method uses information from a confusion matrix and evaluates one attribute at a time. **Keywords:** classification, attribute selection, confusion matrix, k-nearest neighbors;

Background

In classification problems, good accuracy in classification is the primary concern; however, the identification of the attributes (or features) having the largest separation power is also of interest. Even more, for very large data sets (such as MRI images of brain), the classification is highly dependent on feature selection. This is mainly because the larger the number of attributes, the more sparse the data become and thus many more (exponential growth) training data are necessary to accurately sample such a large domain. In this sense, the high dimensional data sets are almost always under represented. This problem is also known in literature as "the curse of dimensionality". For example, a 2-attribute data set having 10 examples in the square defined by the corners (0,0) and (1,1) covers the domain acceptably. If the domain to be learned is the cube defined by the corners (0,0,0) and (1,1,1), 10 points will not cover this 3-D domain as effectively.

Reducing the number of attributes for a classification problem is a much researched field. The brute force approach in finding the best combination of attributes for classification requires the trial of all possible combinations of the available n attributes. That is, consider one attribute at a time, then investigate all combinations of two attributes, three attributes, etc. However, this approach is unfeasible because there are $2^n - 1$ such possible combinations for n attributes and, for example, even for n=10 there are 1,023 different attribute combinations to be investigated. Additionally, feature selection is especially needed for data sets having large numbers of attributes (e.g. thousands). Examples of such data domains with many features include text categorization and gene expression analysis. In the first example, each document is described by the most frequent words, leading to 20,000 words or more. In working with expressed genes in order to separate healthy from cancer patients, for example, the number of attributes may grow as high as 30,000 (Guyon and Elisseeff 2003). Another example of a challenging domain is the microarray data found in (Xin, Jordan, and Karp 2001), where a hybrid of filter and wrapper approaches is employed to successfully select relevant features to classify 72 data examples in a 7,130 dimensional space.

In addition to reducing the data dimensionality, selecting fewer attributes may improve classification and may give a better understanding of the underlying process that generated that data. Here we propose an attribute-selection technique based on a confusion matrix with the two-fold objective of better classification and better data understanding.

Depending on where the feature selection module is placed in relation to the classification module, there are two classes of methods for feature selections (Jain and Zongker 1997):

- **Filter methods** (Pudil, Novovicova, and Kittler 1994) rank features (or feature subsets) independently of the predictor. These methods investigate irrelevant features to be eliminated by looking at correlation or underlying distribution. For example, if two attributes have the same probability distribution, then they are redundant and one of them can be dropped. Such analysis is performed regardless of the classification method. Another filtering method ranks attributes based on the notion of nearest hit (closest example of same the class) and nearest miss (closest example of a different class) (Kira and Rendell 1992). The i^th feature ranking is given by the score computed as the average (over all examples) of the difference between the distance to the nearest hit and the distance to the nearest miss, in the projection of the i^th dimension (Guyon and Elisseeff 2003).
- **Wrapper methods** (Kohavi and John 1997) use a classifier to assess features (or feature subsets). For example, the decision tree algorithm selects the attributes having

Table 1: The confusion matrix for two-class classification problem.

	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	a	b
ACTUAL POSITIVE	c	d

the best discriminatory power and places them closer to the root. Hence, besides the classification tree, a ranking of attributes results.

Another classification of attribute selection methods considers the search technique of the feature subsets. There are two main greedy search strategies: forward selection and backward elimination. Both techniques yield nested subsets of features. The forward selection starts with one attribute and continues adding one attribute at a time if the newly formed subset gives better classification. During backward elimination, unpromising attributes are progressively eliminated (Guyon and Elisseeff 2003). This greedy search technique is often used in system identification. The result, in either case, is not guaranteed to yield the optimal attribute subset (Sugeno and Yasukawa 1993).

In this research we investigate the use of the confusion matrix (Kohavi and Provost 1998) (which contains information about actual and predicted classifications) for attribute selection. In the context described above, this approach is a wrapper method because it uses a classifier to estimate the classification power of an attribute (or subset of attributes).

The Confusion Matrix and Disagreement Score

A confusion matrix of size n x n associated with a classifier shows the predicted and actual classification, where n is the number of different classes. Table 1 shows a confusion matrix for n = 2, whose entries have the following meanings:

- a is the number of correct negative predictions;
- b is the number of incorrect positive predictions;
- c is the number of incorrect negative predictions;
- d is the number of correct positive predictions.

The prediction accuracy and classification error can be obtained from this matrix as follows:

$$Accuracy = \frac{a+d}{a+b+c+d} \quad (1)$$

$$Error = \frac{b+c}{a+b+c+d} \quad (2)$$

We define the disagreement score associated with a confusion matrix in equation (3). According to this equation the disagreement is 1 when one of the quantities b or c is 0 (in this case the classifier misclassifies examples of one class only), and is 0 when b and c are the same.

$$D = \begin{cases} 0 & \text{if } b = c = 0; \\ \frac{|b-c|}{max\{b,c\}} & \text{otherwise.} \end{cases} \quad (3)$$

The attribute selection methodology proposed here selects attributes that not only have good discrimination power on their own, but more importantly are complementary to each other. For example, consider two attributes A1 and A2, having similar classification accuracy. Our approach will select them as a good subset of attributes if they have a large disagreement in terms of what examples they misclassify. A large disagreement is indicated by D values closer to 1 for both attributes, but distinct denominators in equation (3).

Algorithm for Confusion Matrix-based Attribute Selection

The pseudocode outlined below shows the steps to perform confusion matrix-based attribute selection for a 2-class classification problem. This method basically constructs attribute-subsets that: (1) have attributes with good individual classification power, and (2) have attributes that are complementary (i.e. they disagree in their misclassifications).

Note that the algorithm may lead to several subsets of attributes to be further investigated, i.e. further the subset yielding higher classification accuracy may be selected.

Also, the algorithm does not account for the possibility that two individually lower ranked attributes may combine in a high classification accuracy subset due to their high complementarity.

Algorithm 1 Pseudocode for Confusion Matrix-based Attribute Selection Algorithm

Require: 2-class data of n attributes
Require: classification technique
Require: k - number of member subset
Ensure: Output k-attribute subset as tuple $S_k = (A_1, A_2, ..., A_k)$
Compute classifier C_i based on feature A_i, $i = 1..n$
Obtain: $Accuracy(C_i)$ and $ConfMatrix(C_i)$
Rank A_i according to $Accuracy(C_i) \Rightarrow R_A$
for $i = 1$ ... n **do**
  Compute disagreement based on $ConfMatrix(C_i)$ as: $D_i = \frac{|b-c|}{max\{b,c\}}$
end for
Rank A_i according to $D_i \Rightarrow R_D$
Select top k (according to R_A) attributes having large D (according to R_D) but in different classes: $\Rightarrow S_k = (A_1, A_2, ..., A_k)$

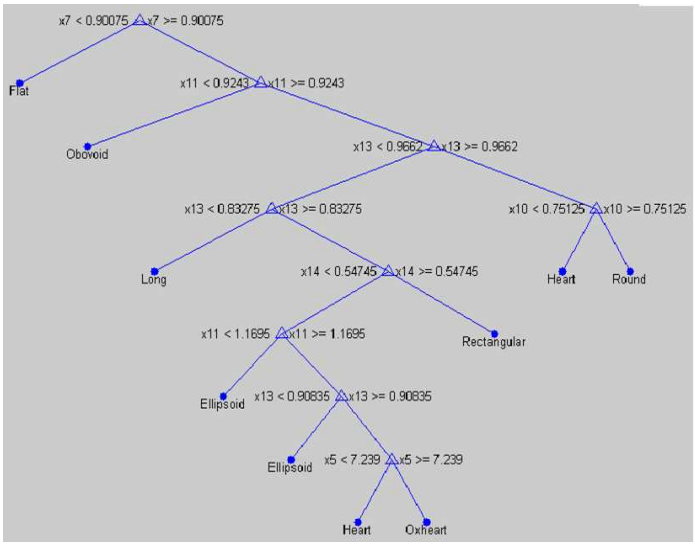

Figure 1: Decision tree obtained with CART for all data and all attributes.

Table 2: Data distribution across classes. In total, there are 416 examples each having 34 attributes.

Class	Class label	No. of examples
1	Ellipse	110
2	Flat	115
3	Heart	29
4	Long	36
5	Obvoid	32
6	Oxheart	12
7	Rectangular	34
8	Round	48

The Tomato Fruit Data Set

The data set used in the experimental part of this research consists of 416 examples having 34 attributes and distributed in 8 classes (the class-distribution is shown in Table 2). This set was obtained from the Ohio Agricultural Research and Development Center (OARDC) research group led by E. Van Der Knaap (Rodriguez et al. 2010) and the classification task is to correctly label a tomato fruit based on morphological measurements such as width, length, perimeter, circularity (i.e. how well a transversal cut of a tomato fits a circle), angle at the tip of the tomato, etc.

The data set was collected as follows: from the scanned image of a longitudinally section of a tomato fruit the 34 measurements are extracted by the Tomato Analyzer Software (TA) (Rodriguez et al. 2010) developed by the same group. For a complete description of the 34 tomato fruit measurements and the TA software see (Gonzalo et al. 2009).

In addition to tomato classification, of interest here is to find which attributes have more discriminative power and further to find a ranking of the attributes.

Data Classification and Attribute Selection

In this section we show the decision tree classification of the tomato data set, then we illustrate our attribute selection algorithm (in combination with a k-nearest neighbor classifier) on two (out of 8) classes. These two classes (1 and 7) are identified by both, decision trees and k-nearest neighbors, as highly overlapping.

Classification with Decision Trees - CART

We used the Classification and Regression Trees (CART) method (Breiman et al. 1984) because it generates rules that can be easily understood and explained. At the same time, classification trees have a built-in mechanism to perform attribute selection (Breiman et al. 1984) and we can compare our set of selected attributes, obtained from the confusion matrix and k-nearest neighbors analysis, with the set of attributes identified by CART. However, we anticipate that these sets will not perfectly coincide, which only means that the two approaches quantify the importance of a given attribute (or subset of attributes) differently and that the two methods learn the data differently.

The pruned decision tree obtained using CART is shown in Figure 1. The train and test error associated with this tree are 11.54% and 18.27%, respectively. As it can be seen from this figure, 10 rules can be extracted. In addition, CART selects the following 8 attribute as best in classification (listed in decreasing order of their importance):

- 7 - Fruit Shape Idx Ext1
- 13 - Circular
- 12 - Ellipsoid
- 11 - Fruit Shape Triangle
- 14 - Rectangular
- 10 - Distal Fruit Blockiness
- 8 - Fruit Shape Idx Ext2
- 1 - Perimeter

We also investigate the k-nearest neighbors classifier as we will use this classification method in combination with our attribute selection approach. Figure 2 shows the k - nearest neighbors classification error for k = 2,...,15. The top (blue) line corresponds to runs that include all 34 attributes and the bottom (red) line shows the error when only the best five attributes (identified by CART classification) are used.

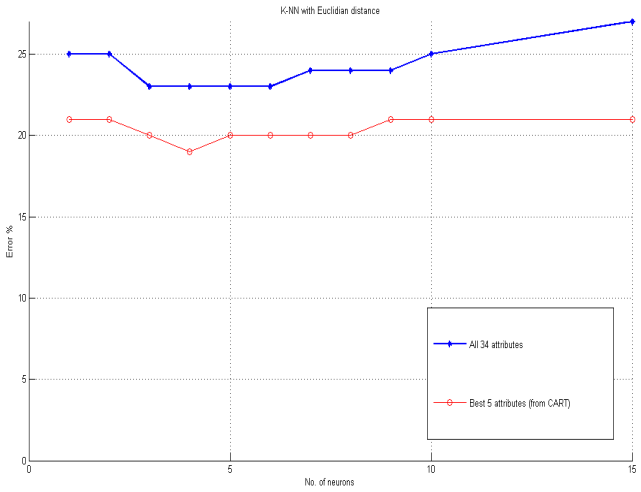

Figure 2: K - nearest neighbors classification error for k = 2,...,15. The top (blue) line corresponds to runs that include all 34 attributes and the bottom (red) line shows the error when only the best five attributes are used (these attributes were identified through CART classification).

As shown in Figure 2, the k-nearest neighbors classification technique consistently scores lower error when using only 8 attributes (the one selected by CART), rather then all 34. Thus, a natural question arising here is: is there a better combination of attributes then the one selected by CART for classification? For k = 4 the k-nearest neighbors technique yields the lowest error, which justifies our choice of using k = 4 in the next experiments.

The Confusion Matrix for the Tomato Data Set

When using all 34 attributes and all 8 classes, the confusion matrix obtained from the k-nearest neighbors clustering with k=4 is shown in Figure 3, where along the x-axis are listed the true class labels and along the y-axis are the k-nearest neighbors class predictions. Along the first diagonal are the correct classifications, whereas all the other entries show misclassifications. The bottom right cell shows the overall accuracy.

In this confusion matrix it can be seen that 8 examples of class 7 are wrongly predicted as class 1. Additionally, from the CART classification, the classes 1 and 7 are identified as the two classes overlapping the most. Thus the experiment presented next uses the confusion matrix attribute selection to better separate these two classes. Namely, we search for a subset of the 34 attributes such that the attributes are complementary in the sense described above and quantified in equation (3).

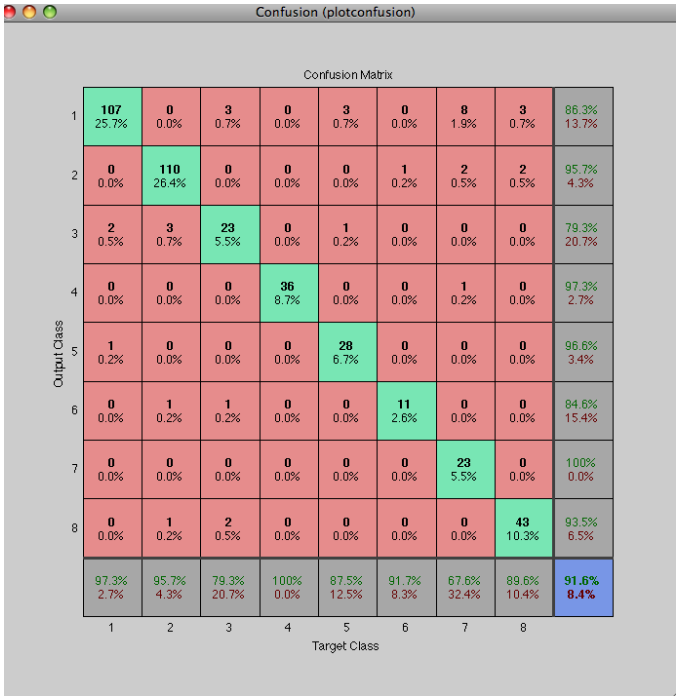

Figure 3: Confusion matrix for all classes and all attributes. Class 7 has 8 examples wrongly predicted as class 1 (see top row).

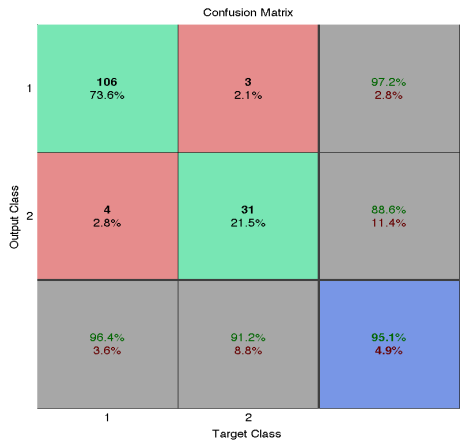

Figure 4: Confusion matrix for class 1 and 7 along attribute 14. Four examples of class 1 are misclassified as class 7, and 3 examples of class 7 belong to class 1.

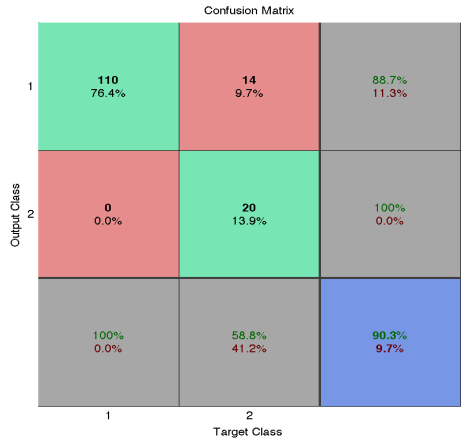

Figure 5: Confusion matrix for class 1 and 7 along attribute 20. Fourteen examples of class 7 are misclassified as class 1.

Confusion Matrix-based Attribute Selection for Classes 1 and 7

When using the data from classes 1 (Ellipse) and 7 (Rectangular), a data set of size 145 is obtained (110 in class 1 and 34 from class 7).

As illustrated in Algorithm 1, for each of the 34 attributes, the k-nearest neighbor algorithm (with k = 4) is used for classification and the corresponding classification accuracy and confusion matrix are obtained (for each attribute). Further, the 34 attributes are ranked in the order of their individual performance in distinguishing between class 1 and 7, leading to the ranking set R = 14, 7, 8, 17, 1, 3, 6, 12, 30, 4, 9, 20, 29, 18, 26, 2, 10, 21, 34, 32, 11, 33, 5, 13, 19, 16, 15, 31, 25, 28, 24, 27, 22, 23.

We first create growing nested subsets of attributes in the order specified by their individual classification abilities. Note that this particular choice of subsets is not part of Algorithm 1 and makes no use of the complementarity. We simply introduce it as a comparative model for our selection approach, which, besides the accuracy ranking, incorporates complementarity information as well.

Figure 6 shows the classification accuracy for subsets of attributes consisting of the top 1, top 2, top 3, etc. attributes from R (the subsets are shown on x-axis, while the y-axis shows classification accuracy). From Figure 6 it can be seen that the highest accuracy is achieved when the top 3 attributes are used together (i.e. attributes 14, 7, and 8),

Table 3: Attribute ranking based on disagreement score. The best classification attributes found by CART are shown in bold (they are also underlined). The attributes marked by (*) are the ones identified by our selection algorithm.

Attr. number	Disagreement score	Class of largest error
20	1	7
22	1	7
24	1	7
25	1	7
26	1	7
27	1	7
31	1	7
23	0.9655	7
28	0.9565	7
15	0.9524	7
2	0.9375	7
21	0.9375	7
29	0.9231	7
12	**0.9167**	**7**
30	0.9167	7
1	**0.9091**	**7**
3	0.9091	7
7	**0.9000**	**7**
17	0.9000	7
5	0.8889	7
11	**0.8824**	**7**
32	0.8750	7
34	0.8667	7
18	0.8462	7
13	**0.8235**	**7**
19	0.8235	7
6	0.8182	7
33	0.8125	7
4	0.7273	7
9*	**0.7273**	**7**
16*	**0.6875**	**7**
8*	**0.6250**	**7**
10*	**0.3000**	**7**
14*	**0.2500**	**1**

yielding a 97.2% correct classification.

The above approach is a (greedy) forward selection method, i.e. the attributes are progressively incorporated in larger and larger subsets. However, here we incorporate the attributes in the order dictated by their individual performance, yielding nested subsets of attributes. For example, we do not evaluate the performance of the subset consisting of first and third attribute. Indeed, it may be that this subset can perform better then considering all top three attributes. However, as indicated earlier in this paper, evaluating all possible combination is not a feasible approach. Thus, we propose **to combine the attributes that are complementary, i.e. two (or more) attributes that may achieve individually similar classification accuracy but they have the largest disagreement (this information is extracted from the confusion matrix).**

The disagreement scores for all 34 attributes when classify-

ing data from classes 1 and 7 are listed in Table 3, column 2. As column 3 of the same table shows, only attribute 14 misclassifies more examples of class 1 then of class 7 (see bottom row). All other 33 attributes have the largest number of misclassifications attributed to class 7. Thus, for this particular data set, we will consider subsets that combine attribute 14 with one or more other attributes having the largest disagreement.

For example, Figures 4 and 5 show the confusion matrix for classes 1 and 7 when using attribute 14 and 20, respectively. These two attributes disagree the most as the above figures and Table 3 show.

Algorithm 1 is illustrated here for k = 2,3,4, and 5 only (note for k =1 the classification results are the same as in Figure 6). The classification accuracy for these experiments is shown in Figures 7, 8, 9, and 10, respectively. In addition to selecting the top k attributes in Table 3 (this combination yields the first star in the above plots), we also plot the classification accuracy of the sliding (moving from top to bottom in Table 3) window of k attributes.

Figures 7, 8, 9, and 10 show the classification accuracy of the k-nearest neighbor algorithm when attribute 14 is combined with all the other attributes in decreasing order of their disagreement scores (see Table 3). Figure 7 shows results for 2-member subsets and the results from Figure 8 are obtained for 3-member subsets: attribute 14 and two consecutive attributes from Table 3.

As Figure 10 illustrates, simply selecting the top attributes from Table 3 (having the largest disagreement) does not ensure a better classification, nor is an increasing or decreasing trend observed when sliding down the table. This is because classification ability is not additive and opens up the question of whether a better k-subset of attributes can be obtained by mixing attributes across the table, not only the k-neighbors selection used here.

Among the k-member subsets investigated here (note, there are more sliding window subsets for $k > 5$), the largest classification accuracy (98%) is achieved for a 5-member subset, namely for the attribute-set 14, 9, 16, 8, and 10. The CART classifier recognizes these two classes with 93% accuracy (using attributes 7, 13, 12, 14, 11, and 10), and the accuracy-ranking only (no complementarity information incorporated) selection achieves 97.3% (using top 3 attributes: 14, 7 and 8). The attribute subset with the largest discriminating power (when using a k-nearest neighbors clustering, k = 4) is obtained with the confusion matrix-based attribute selection; however, it is not a large improvement as the classes are pretty well separated to begin with.

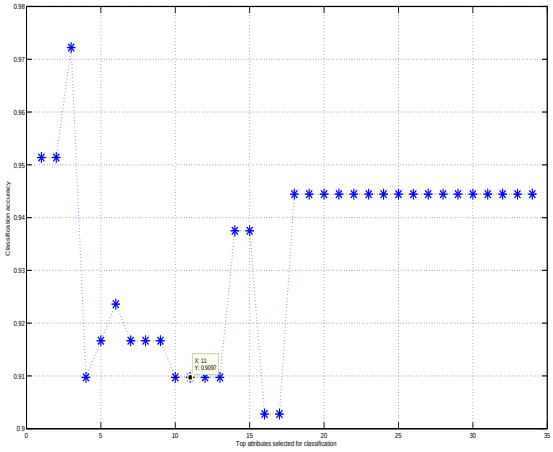

Figure 6: K - nearest neighbors classification accuracy for k = 4 when using data from classes 1 and 7 only. On x-axis are listed the nested subsets of attributes having top 1,2,3,...,34 attributes. The highest accuracy (97.2%) is obtained for the subset having the top 3 attributes: 14,7, and 8.

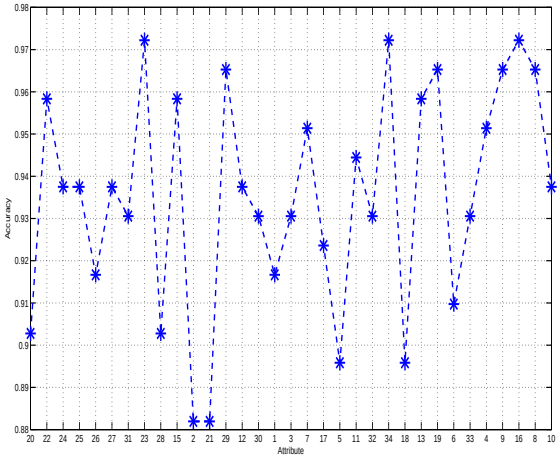

Figure 7: K - nearest neighbors classification accuracy for k = 4 when using 2-member subsets of attributes. Each subset contains attribute 14 and one of the remaining attributes; x-axis shows these subsets listed in the order of their complementarity - see Table 3. Largest accuracy is 97.3%.

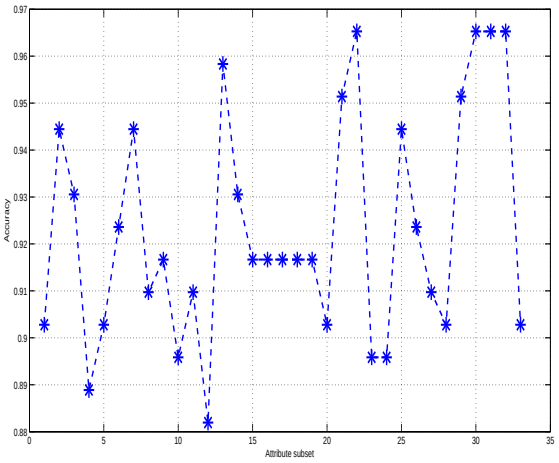

Figure 8: K - nearest neighbors classification accuracy for k = 4 when using 3-member subsets of attributes. Each subset contains attribute 14 and two consecutive attributes (in the order of their complementarity - see Table 3). Largest accuracy is 97.3%.

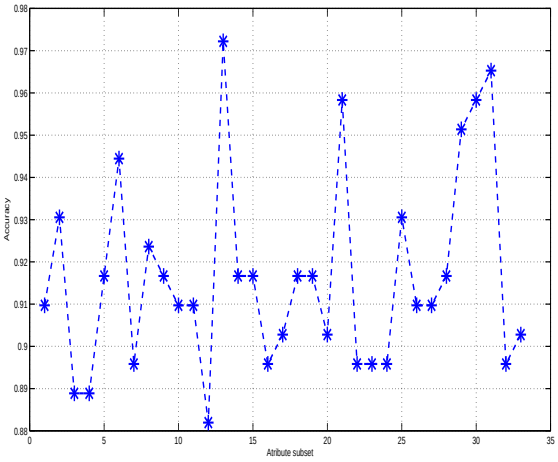

Figure 9: K - nearest neighbors classification accuracy for k = 4 when using 4-member subsets of attributes. Each subset contains attribute 14 and three consecutive attributes (in the order of their complementarity - see Table 3). Largest accuracy is 96.7%.

Conclusions and Future Work

A new technique for attribute selection is proposed here. The method selects attributes that are complementary to each other, in the sense that they misclassify different classes, and favors attributes that have good classification abilities by themselves. This new approach is illustrated on a real data set. For two classes of interest within this data set, this technique found a better (i.e. yielding higher classification accuracy) subset of attributes, than using all attributes or even using the 8 attributes identified by CART. However, we must investigate this new approach in more data sets and in combination with other classification techniques (here only the k-nearest neighbor classifier was investigated). Another future direction is to investigate the use of subsets that combine complementary attributes, even if these attributes are weak classifiers by themselves. The challenging factor for this approach is the large number of subsets that must be investigated. Depending on the data set, if this search space is very large, then genetic algorithms can be used to explore the version space. We must also extrapolate this method to multi-class data sets and investigate its scalability factor.

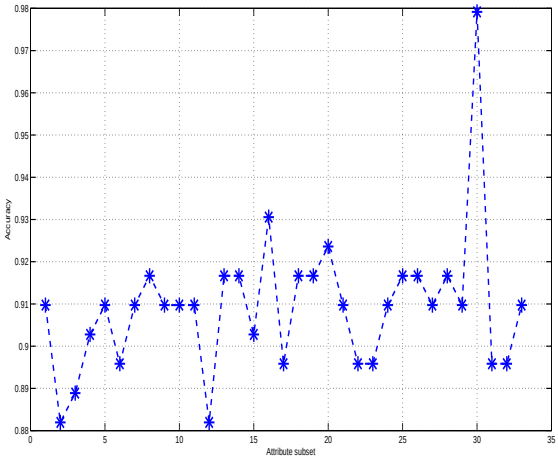

Figure 10: K - nearest neighbors classification accuracy for k = 4 when using 5-member subsets of attributes. Each subset contains attribute 14 and four consecutive attributes (in the order of their complementarity - see Table 3). Largest accuracy is 98%.

Acknowledgments

Esther van der Knaap acknowledges support from the NSF grant NSF DBI-0922661. Sofia Visa was partially supported by the NSF grant DBI-0922661(60020128) and by the College of Wooster Faculty Start-up Fund.

References

Breiman, L.; Friedman, J.; Olshen, R.; and Stone, C., eds. 1984. *Classification and Regression Trees*. CRC Press, Boca Raton, FL.

Gonzalo, M.; Brewer, M.; Anderson, C.; Sullivan, D.; Gray, S.; and van der Knaap, E. 2009. Tomato Fruit Shape Analysis Using Morphometric and Morphology Attributes Implemented in Tomato Analyzer Software Program. *Journal of American Society of Horticulture* 134:77–87.

Guyon, I., and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3:1157–1182.

Jain, A., and Zongker, D. 1997. Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(2):153–158.

Kira, K., and Rendell, L. 1992. A practical approach to feature selection. In *International Conference on Machine Learning*, 368–377.

Kohavi, R., and John, G. 1997. Wrappers for features subset selection. *Artificial Intelligence* 97:273–324.

Kohavi, R., and Provost, F. 1998. On Applied Research in Machine Learning. In *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, Columbia University, New York*, volume 30.

Pudil, P.; Novovicova, J.; and Kittler, J. 1994. Floating search methods in feature selection. *Pattern Recognition Letters* 15(11):1119–1125.

Rodriguez, S.; Moyseenko, J.; Robbins, M.; Huarachi Morejn, N.; Francis, D.; and van der Knaap, E. 2010. Tomato Analyzer: A Useful Software Application to Collect Accurate and Detailed Morphological and Colorimetric Data from Two-dimensional Objects. *Journal of Visualized Experiments* 37.

Sugeno, M., and Yasukawa, T. 1993. A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on fuzzy systems* 1(1):7–31.

Xin, E.; Jordan, M.; and Karp, R. 2001. Feature Selection for High-Dimensional Genomic Microarray Data. In *Proceedings of the 18 International Conference in Machine Learning ICML-2001*, 601–608.