Learning Morphological Data of Tomato Fruits

Joshua Thomas, Matthew Lambert, Bennjamin Snyder,
Michael Janning, Jacob Haning, Yanglong Hu, Mohammad Ahmad, Sofia Visa
Computer Science Department
College of Wooster
jet4416@gmail.com, mlambert13@wooster.edu, benn.snyder@gmail.com,
mjanning13@wooster.edu,jacob.haning1@gmail.com,
yhu12@wooster.edu, mahmad12@wooster.edu, svisa@wooster.edu

Abstract

Three methods for attribute reduction in conjunction with Neural Networks, Naive Bayes, and k-Nearest Neighbor classifiers are investigated here when classifying a particularly challenging data set. The difficulty encountered with this data set is mainly due to the high dimensionality and to some inbalance between classes. As a result of this research, a subset of only 8 attributes (out of 34) is identified leading to a 92.7% classification accuracy. The confusion matrix analysis identifies class 7 as the one poorly learned across all combinations of attributes and classifiers. This information can be further used to upsample this underrepresented class or to investigate a classifier less sensitive to imbalance.

Keywords: classification, attribute selection, confusion matrix;

Introduction

Knowing (or choosing) the best machine learning algorithm for classifying a particular real world data set is still an ongoing research topic. Researchers have tackled this problem more as experimental studies, such as the ones shown in (Michie, Spiegelhalter, and Taylor 1999) and (Visa and Ralescu 2004), than as theoretical ones. Currently, it is difficult to study the problem of the best classification method given a particular data set (or the reverse problem for that matter), because data classification depends on many variables, e.g. number of attributes, number of examples and their distribution across classes, underlying distribution along each attribute, etc. Additionally, it is difficult to study classifier induction in general, because different classifiers learn in different ways, or stated differently, different classifiers may have different learning biases. Thus, this research focuses on finding the best classification method for a particular data set of interest through experimental means.

We investigate several machine learning techniques for learning a particular 8-class domain having 34 attributes and only 416 examples. We also combine these methods with various subsets of attributes selected based on their discriminating power. The main goal of this research is to find the best classification algorithm (or ensemble of algorithms) for this particular data set.

The research presented here is part of a bigger project, with the classification of morphological tomato data (i.e. data describing the shape and size of tomato fruits such as the data set used here) being the first step. Namely, having the morphological and gene expression data, the dependencies between these two sets are to be investigated. The goal of such computational study is to reveal genes that affect particular shapes (e.g. elongated or round tomato) or sizes (e.g. cherry versus beef tomato) in the tomato fruit. However, as mentioned above, a high classification accuracy of tomatoes based on their morphological attributes is required first.

Table 1: Data distribution across classes. In total, there are 416 examples each having 34 attributes. (Table from (Visa et al. 2011))

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 8 classes are illustrated in Figure 1 and the distribution of the 416 examples is shown in Table 1 (Visa et al. 2011).

The Tomato Fruit Morphological Data

The experimental data set was obtained from the Ohio Agricultural Research and Development Center (OARDC) research group led by E. Van Der Knaap (Rodriguez et al. 2010).

This morphological data of tomato fruits consists of 416 examples having 34 attributes and distributed in 8 classes. The

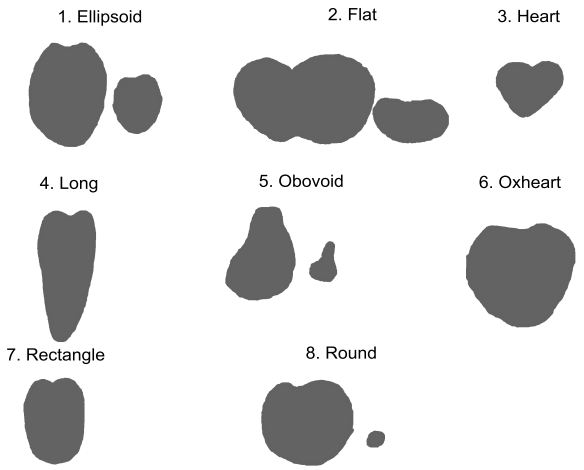

Figure 1: Sketch of the 8 morphological classes of the tomato fruits.

34 attributes numerically quantify morphological properties of the tomato fruits such as perimeter, width, length, circularity (i.e. how well a transversal cut of a tomato fits a circle), rectangle (similarly, how well it fits a rectangle), angle at the tip of the tomato, etc. A more detailed description of the 34 attributes and how they are calculated can be found in (Gonzalo et al. 2009).

Problem Description and Methodology

The focus of this research is to find the best (defined here as high classification accuracy, e.g. 90%) classification technique (or combination of classifiers) for the morphological tomato data set. In addition to tomato fruit classification, it concentrates on finding which attributes have more discriminative power and finding a ranking of these attributes.

As seen in Figure 1, the round class and several others may have smaller or much larger instances of tomato fruits. Thus, attributes 1 (perimeter) and 2 (area), for example, might have no positive influence in classifying these classes; at worst, it may hinder classification. The tomato data set of interest here has 34 attributes and only 416 examples available for learning and testing. One can argue that many more examples are needed to have effective learning in such high-dimensional space. Furthermore, the class-distribution is imbalanced with the largest and the smallest classes having 115 (class 2, Flat tomatoes) and 12 examples (class 6, Oxheart tomatoes), respectively (see Table 1).

For these reasons, our strategy is to investigate several machine learning classifiers on subsets of top-ranked attributes in an effort to reduce the data-dimensionality and to achieve better data classification. Finding if different classification algorithms make identical errors (for example, they all misclassify class 7 with class 1) is also of interest in this experimental study. Our hypothesis is that (some) different classifiers misclassify different data-examples and thus, by combining different classifiers, one can achieve better accuracy merely through their complementarity. The misclassification error for each individual class is tracked through the use of confusion matrices.

Attribute Selection Techniques

Two filter-methods (analysis of variance ANOVA (Hogg and Ledolter 1987) and the RELIEF method (Kira and Rendell 1992b), (Kira and Rendell 1992a)) and one wrapper-method are used in our experiments for attribute-ranking. The first two algorithms are independent of the choice of classifier (Guyon and Elisseeff 2003), whereas the third one is "wrapped" around a classifier - here the attributes selected by the CART decision tree are used (Breiman et al. 1984).

The first attribute ranking method considered here is based on the analysis of variance which estimates the mean value of each attribute by comparing the variation within the data (Hogg and Ledolter 1987).

The second ranking method we use is the RELIEF algorithm, introduced by (Kira and Rendell 1992b) and described and expanded upon by (Sun and Wu 2009). In short, the algorithm examines all instances of every attribute and calculates the distance to each instance's nearest hit (nearest instance that has the same classification) and nearest miss (nearest instance that has a different classification). It then calculates the differences of the nearest misses and the nearest hits over all instances of each attribute. as shown in equation (1).

$$d_n = \|x_n - NearestMiss(x_n)\| - \|x_n - NearestHit(x_n)\| \quad (1)$$

where d_n is an instance of an attribute, $NearestMiss(x_n))$ is the nearest miss of the instance, and $NearestHit(x_n))$ is the nearest hit of the instance. Then, the d-values are summed over all instances and the attributes are ranked from largest value to smallest value. Zero may provide an appropriate cut-off point when selecting attributes.

The third ranking is obtained as a result of decision trees classification (CART), which through a greedy approach places the most important attributes (based on information gain) closer to the root of the tree.

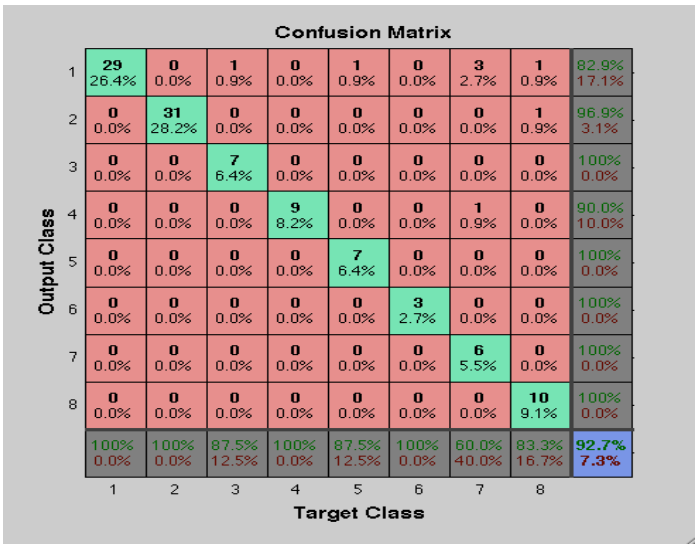

Figure 2: Confusion matrix of NN for top 8 CART attributes. This case achieved the highest classification accuracy when using NN (92.7%).

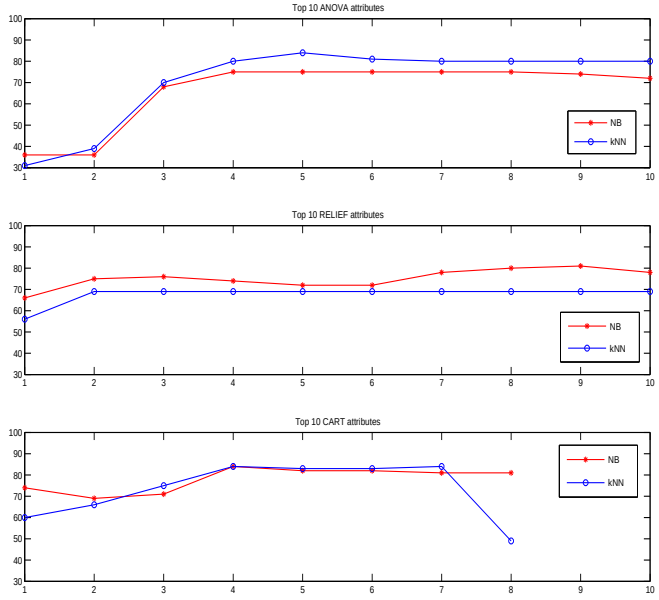

Figure 3: Accuracy of NB and kNN for top k (k=1,10) ANOVA attributes (top figure), top k (k=1,10) RELIEF attributes, and top k (k=1,8) CART attributes (k is shown on x-axis).

Table 2: Top 10 ANOVA and RELIEF attribute rankings. Column 3 shows the top 8 ranked attributes resulted from classification and regression trees (CART) (Visa et al. 2011)

ANOVA	RELIEF	CART
17	21	7
20	18	13
18	7	12
21	8	11
2	33	14
1	13	10
28	11	8
26	9	1
6	19	-
5	22	-

Classification Techniques

We use Matlab to conduct these experiments. For each experiment 75% of data is randomly selected for training, and the remaining 25% of data is used for testing.

Matlab implementations of the Naive Bayes (NB), k-nearest Neighbors (kNN) for k=4, and various Artificial Neural Network (NN) configurations are tested in conjunction with the three reduced-attribute tomato data sets, as well as with the whole data sets (i.e. having all 34 attributes). For the latter case, the classifiers are ordered by their accuracies as follows: NN (89.1%), NB (80%), kNN (79.1%). Here, kNN is investigated for k=4 only because (Visa et al. 2011) shows that it achieves the lowest error over a larger range of k.

Results

The top 10 ANOVA and RELIEF attribute rankings are shown in the first two columns of Table 2. Column 3 shows the top 8 ranked attributes resulted from classification and regression trees.

NN Results

Many NN configurations (in terms of number of layers, number of neurons in each layer, training method, and activation function) for each of the three data sets obtained from selecting the subsets of attributes shown in Table 2 were tried. However, only the ones leading to the best results are reported in Table 3. Among the subsets of attributes studied here, the 8 attributes resulting from the decision tree classification lead to the best classification in the NN case (92.7%). The confusion matrix associated with this case is shown in Figure 2. From this matrix, it can be seen that the largest error comes from misclassifying 3 test data points of class 7 (Rectangle) as class 1 (Ellipsoid). Indeed, Figure 1 shows that these two clases are the most similar in terms of shape.

Table 3: Best NN configurations and their corresponding classification accuracies.

No. of attributes	No. of layers	No. of neurons	Accuracy
Top 10 ANOVA	1	10	84.5%
Top 10 RELIEF	2	25+15	88.2%
Top 8 CART	1	10	92.7%
All 34	2	25 +15	89.1%

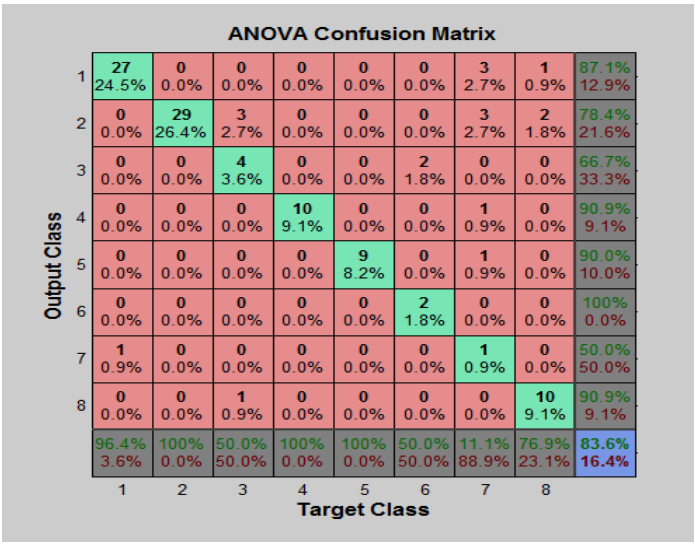

Figure 4: Confusion matrix of kNN for top 5 ANOVA attributes. This case achieved the highest classification accuracy when using kNN (83.6%).

NB and kNN Results

Figure 3 shows the accuracy of NB and kNN for top k (k=1,10) ANOVA attributes (top figure), top k (k=1,10) RELIEF attributes, and top k (k=1,8) CART attributes (k is shown on x-axis). The two largest accuracy values are obtained for kNN (83.6%) for the top 5 ANOVA attributes, and for NB (81.1%) in the case of top 9 RELIEF attributes. For these two cases, the confusion matrices showing the misclasifications across the 8 classes are shown in Figures 4 and 5, respectively. Similar to NN classifier, NB and kNN both misclasify class 7 as class 1 (by 4 and 3 examples, respectively). However, contrary to NN, NB and kNN carry some additional class confusions:

- NB misclassifies class 3 as class 1 (3 instances) and as class 8 (3 instances);
- Additional error for kNN comes from misclassifying class 3 as class 2 (3 examples) and class 7 as class 2 (3 examples).

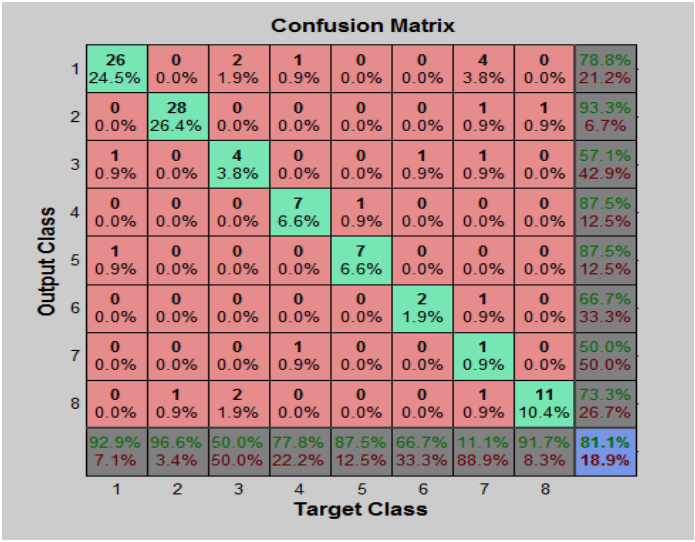

Figure 5: Confusion matrix of NB for top 9 RELIEF attributes. This case achieved best classification accuracy when using NB (81.1%).

Conclusions and Future Work

Several machine learning algorithms for classifying the 8-class tomato data are investigated here. In addition, 3 attribute selection strategies are combined with these learning algorithms to reduce the data set dimensionality. The best combination of attribute selection and classification method among the ones investigated here leads to a 92.7% classification accuracy (for the NN classifier on the 8 CART attributes).

The confusion matrix analysis points out that class 7 (Rectangle) is the one most frequently miscclassified (or very poorly learned) across all three classifiers. It is more often missclasified as class 1. This is consistent with the observation that (1) based on Figure 1, these two classes are very similar, and (2) since class 1 is larger in terms of available examples (110 versus only 34 in class 7, see Table 1), we can conclude that the classifiers are biased toward the larger class. This situation is known in literature as learning with imbalanced data (Visa and Ralescu 2004). As a future direction, we point out that, for imbalanced data sets, classifiers less sensitive to the imbalance can be used such as the one proposed in (Visa and Ralescu 2004). Also, the imbalance can be corrected by intentional upsampling (if possible) of the underrepresented class.

A similar study that considers some additional classification techniques applied to a larger overall data set (the 416 examples in the current data sets poorly covers the 34-dimensional space) in which the classes are less imbalanced will provide more insight as to what attributes should be selected for better classification accuracy. Also, a more thorough analysis of the confusion matrices will identify com-

plementary classification techniques which can be subsequently combined to obtain a larger classification accuracy for the data set of interest.

Acknowledgments

This research was partially supported by the NSF grant DBI-0922661(60020128) (E.Van Der Knaap, PI and S.Visa, Co-PI) and by the College of Wooster Faculty Start-up Fund awarded to Sofia Visa in 2008.

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.

Hogg, R., and Ledolter, J., eds. 1987. *Engineering Statistics*. New York:MacMillan.

Kira, K., and Rendell, L. 1992a. The Feature Selection Problem: Traditional Methods and a New Algorithm. In *Proceedings of AAAI*, 129–134.

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

Michie, D.; Spiegelhalter, D.; and Taylor, C. e., eds. 1999. *Machine Learning, Neural and Statistical Classification*. http://www.amsta.leeds.ac.uk/ charles/statlog/.

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.

Sun, I., and Wu, D. 2009. Feature extraction through local learning. In *Statistical Analysis and Data Mining*, 34–47.

Visa, S., and Ralescu, A. 2004. Fuzzy Classifiers for Imbalanced, Complex Classes of Varying Size. In *Proceedings of the Information Processing and Management of Uncertainty in Knowledge-Based Systems Conference, Perugia, Italy*, 393–400.

Visa, S.; Ramsay, B.; Ralescu, A.; and Van der Knaap, E. 2011. Confusion Matrix-based Feature Selection. In *Proceedings of the 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati*.