

VISUALIZATION IN CLASSIFICATION PROBLEMS

Cook, D., Caragea, D. and Honavar, H.

Key words: Tour methods, classification problems, support vector machines.
COMPSTAT 2004 section: Data Visualisation.

Abstract: In the simplest form support vector machines (SVM) define a separating hyperplane between classes generated from a subset of cases, called support vectors. The support vectors "mark" the boundary between two classes. The result is an interpretable classifier, where the importance of the variables to the classification, is identified by the coefficients of the variables defining the hyperplane. This paper describes visual methods that can be used with classifiers to understand cluster structure in data.

1 Introduction

The classification community has been overly focused on predictive accuracy. For many data mining tasks understanding a classification rule is as important as the accuracy of the rule itself. Going beyond the predictive accuracy to gain an understanding of the role different variables play in building the classifier provides an analyst with a deeper understanding of the processes leading to cluster structure. Ultimately this is our scientific goal, to solve a problem and understand the solution. With a deeper level of understanding researchers can more effectively pursue screening, preventive treatments and solutions to problems.

In the machine learning community, which is driving much of the current research into classification, the goal is that the computer operates independently to obtain the best solution. In data analysis, we're still a long way off this goal. Most algorithms will require a human user twiddle with many parameters in order to arrive at a satisfactory solution. The human analyst is invaluable at the training phase of building a classifier.

This paper describes the use of graphics to build a better classifier based on support vector machines (SVM). We will plot classification boundaries in high-dimensional space and other key aspects of the SVM solution. The visual tools are based on manipulating projections of the data, and are generally described as tour methods.

Our analysis is conducted on a particular data problem, the Italian olive oils data where the task is to classify oils into their geographic area of production based on the fatty acid composition.

We focus on SVM because they operate by finding a hyperplane which maximizes the margin of separation between the two classes. This is similar to how we believe the eye perceives class boundaries. As a result of visualizing

class structure in relation to SVM we have suggestions about how to find simpler but accurate classifiers.

2 Support Vector Machines

SVM is a binary classification method that takes as input labeled data from two classes and outputs a model for classifying new unlabeled data into one of those two classes. SVM can generate linear and non-linear models. In the linear case, the algorithm finds a separating hyperplane that maximizes the margin of separation between the two classes. The algorithm assigns a weight to each input point, but most of these weights are equal to zero. The points having non-zero weight are called *support vectors*. The separating hyperplane is defined as a weighted sum of support vectors. It can be written as, $\{\mathbf{x} : \mathbf{x}'\mathbf{w} + b = 0\}$, where $\mathbf{x}$ is the p-dimensional data vector, and $\mathbf{w} = \sum_{i=1}^{s}(\alpha_i \cdot y_i)\mathbf{x_i}$, where s is the number of support vectors, y_i is the known class for case $\mathbf{x}_i$, and α_i are the support vector coefficients that maximize the margin of separation between the two classes. SVM selects among the hyperplanes that correctly classify the training set, the one that minimizes $\|\mathbf{w}\|^2$, which is the same as the hyperplane for which the *margin* of separation between the two classes, measured along a line perpendicular to the hyperplane, is maximized. The classification for a new unlabeled point can be obtained from $f_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$.

We will be using the software SVM Light 3.50 [5] for the analysis. It is currently one of the most widely used implementations of SVM algorithm.

3 Tours Methods for Visualization

The classifier resulting from SVM, using linear kernels, is the normal to the separating hyperplane which is itself a 1-dimensional projection of the data space. Thus the natural way to examine the result is to look at the data projected into this direction. It may also be interesting to explore the neighborhood of this projection by changing the coefficients to the projection. This is available in a visualization technique called a manually-controlled tour.

Generally, tours display linear combinations (projections) of variables, $\mathbf{x}'\mathbf{A}$ where $\mathbf{A}$ is a $p \times d(< p)$-dimensional projection matrix. The columns of A are orthonormal. Often $d = 2$ because the display space on a computer screen is 2, but it can be 1 or 3, or any value between 1 and p. The earliest form of the tour presented the data in a continuous movie-like manner [1], but recent developments have provided guided tours [3] and manually controlled tours [2]. Here we are going to use a $d = 2$-dimensional manually-controlled tour to recreate the separating boundary between two groups in the data space.

We will be using the tour methods available in the data visualization package GGobi (`www.ggobi.org`).

4 The Analysis of the Italian Olive Oils

To explain the visualization techniques we will use a data set on Italian olive oils. The olive oil data [4] contains 572 instances (cases) that have been chemically assayed for their fatty acid composition. There are 8 attributes (variables), corresponding to the percentage composition of 8 different fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic), 3 major classes corresponding to 3 major growing regions (North, South, Sardinia), and 9 sub-classes corresponding to areas in the major regions.

The data was collected for studying quality control of olive oil production. It is claimed that olive oils from different geographic regions can be distinguished by their fatty acid signature, so that forgeries could be recognized. The multiple classes create a challenge for classification, and the class structure is interesting: the classes have dramatically different variance structure, there are both linear and nonlinear boundaries between classes and some classes are difficult to separate. There are some surprises in the data.

4.1 Approach

Because there are 9 classes and SVM works only with 2 classes, we need to define an approach to building a classifier working in pairwise fashion. The most obvious start is to develop a classifier for separating the 3 regions, and then hierarchically work within region to build classifiers for separating areas.

In each classification we will run SVM. In the visualization we will highlight the cases that are chosen as support vectors and use the weights of the support vectors, and correlations between predicted values and variables to determine the best separating projection.

4.2 Analysis

South vs Sardinia/North: This separation is too easy so it warrants only a short treatment. If the analyst blindly runs SVM with the oils from southern Italy in one class and the remaining oils in another class, then the result is a perfect classification, as shown in the plot of the predicted values on the horizontal axis of Figure 1. However if the analyst plots eicosenoic acid alone (vertical axis in Figure 1) she would notice that this also gives a good separation between the two classes. Eicosenoic acid alone is sufficient to separate these regions. When we examine the correlations between the predicted values and the variables we can see that although eicosenoic acid is the variable that is most correlated with predicted values that several other variables, palmitic, palmitoleic and oleic also contribute strongly to the prediction. Clearly, this classifier is too complicated. The simplest, most accurate rule would be to use only eicosenoic acid, and split the two classes by:

```
Assign to south if eicosenoic acid composition is more than 5%.
```

The minimum eicosenoic acid value for southern oils is 10%, and the maxi-

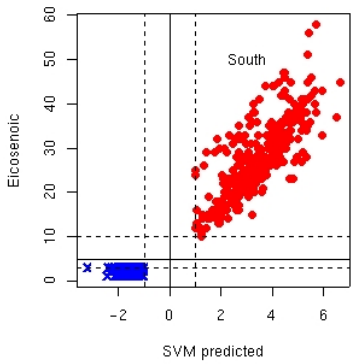

Figure 1: Scatterplot of eicosenoic acid vs predicted values from fitting SVM. The bimodality due to the two classes is clear in eicosenoic acid alone but the SVM fit uses a combination of more variables.

mum eicosenoic acid value for Sardinian and northern oils is 3%. Thus given the variance difference between the two groups a boundary at 5% makes sense. To obtain a solution numerically, the analyst could use eicosenoic acid alone in the SVM, quadratic discriminant analysis or logistic regression. Linear discriminant analysis does not give a good boundary because the two groups have very different variance.

Sardinia vs North: This is an interesting classification problem, so it warrants an in-depth discussion. Plotting the data (Figure 2) two variables at a time reveals both a fuzzy linear classification (left) and a clear non-linear separation (middle) which would be difficult to model. A clean linear separation can be found by a linear combination of linoleic and arachidic acids (right). The circle in the plot is an axis that represents the combination of variables that are shown, and the numbers at right are the numerical values of the projection. Variable 7, linoleic, has a projection coefficient equal to 0.969, and variable 9, arachidic, has a projection coefficient equal to 0.247. This class boundary warrants some investigation. Very often explanations of SVM are accompanied by an idealized cartoon of two well-separated classes, support vectors highlighted, and with the separating hyperplane drawn as a line along with the margin hyperplanes. We would like to reconstruct this picture in high-dimensional data space using tours to examine the classification results. To do this we need to turn the SVM output into visual elements. First, we generate a grid of points over the data space, select the grid points within a tolerance of the separating hyperplane. Then we will use the manually controlled tour to rotate the view until we find the projection of the hyperplane through the normal vector, where the hyperplane reduces to a

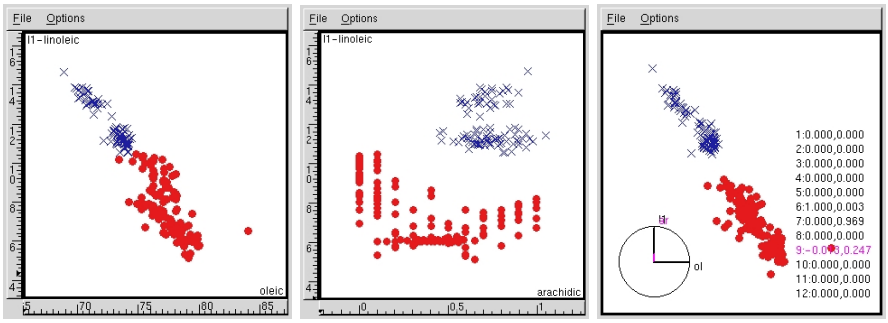

Figure 2: (Left) Close to linear separation in two variables, linoleic and oleic, (Middle) Clear nonlinear separation usng linoleic and arachidic acid, (Right) Good linear separation in combination of linoleic and arachidic acid in the vertical direction.

straight line. Figure 3 shows the sequence of manual rotations made to find this view. Grid points on the hyperplane are small green squares. The large open circles are the support vectors. Samples from Sardinia are represented as blue crosses, and from northern Italy as solid red circles. The large circle at lower left is an axis where the radii represent the projection coefficient of the variables in the current plot. The numbers at the right are the numeric values of the projection coefficients for each variable. Purple text indicates the values for the variable just manipulated, rotated, into the projection. The rotation sequence follows the values provided by the weights, $\mathbf{w}$ and the

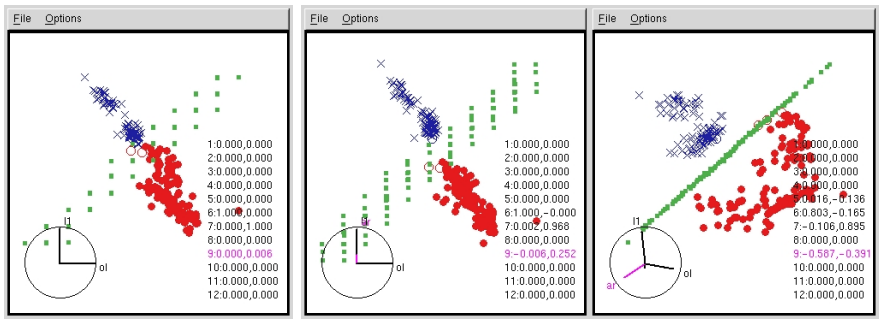

Figure 3: Sequence of rotations that reveals the bounding hyperplane between oils from Sardinia and northern Italy: (Left) Starting from oleic vs linoleic, (Middle) Arachidic acid is rotated into the plot vertically, giving a combination of linoleic and arachidic acid, (Right) The separating hyperplane in the direction of the nonlinear separation.

correlations between predicted values and variables. According the the correlations linoleic is the most important variable followed by oleic, arachidic,

palmitoleic, but stearic acid has a large weight value so we are curious about the contribution of this variable. We begin with the projection of the data into linoleic vs oleic acids because these two variables provided a simple linear boundary. Then other variables will be rotated into the plot in combination with linoleic to match the order of importance as provided by the weights. Since oleic acid has a low weight value we believed that its contribution to the separating hyperplane is mostly due to its strong correlation with linoleic acid. We were wrong. The separating hyperplane runs orthogonally to the linear relationship between oleic and linoleic, lower left to upper right rather than horizontally. Thus, it is clear that oleic acid is used even though its weight is small relative to linoleic acid. Next, arachidic acid is rotated into the plot, in combination with linoleic acid. Here the clear linear separation between the two classes can be seen, but it is not the projection corresponding to the SVM separating hyperplane. So, stearic acid is rotated into the plot in combination with linoleic and arachidic acid. Finally, the separating hyperplane is clearly visible, and the support vectors defining the hyperplane lie on opposing edges of the groups. Its clear that stearic acid is used in building the classifier. The boundary is too close to the northern oils. This surprised us, and we re-checked our planar computations several times, but this is the boundary. A quick fix could be obtained by adjusting the shift parameter value to shift the plane to closer to the middle of the separation between the two classes.

In general, the simplest but as accurate solution would be obtained by entering only two variables, linoleic and arachidic acids, into a classifier which should roughly give a rule as follows:

```
Assign to Sardinia if 0.97 x linoleic+2.5 x arachidic > 11%
```

This was obtained by using the projection coefficients provided by the tour, but it could just have easily been obtained by fitting the SVM model or some other classification model using only linoleic and arachidic acids.

North: The oils from the 3 areas in the north (Umbria, East/West Liguria) are difficult to separate. Working purely from graphical displays we might conclude the areas are not separable. But SVM tells us otherwise. The results are that the 3 areas can be perfectly separated using a linear kernel. So we attempt to find the projection of the data which separates the 3 areas. Working from the correlations between the variables and the predicted values, and from the weights for each variable in the separating hyperplane we rotate variables into view. Figure 4 displays the results. Sure enough, with a combination of most of the variables, a projection of the data where the 3 classes are separated can be found. The combination uses stearic, linoleic, linolenic and arachidic horizontally, and palmitoleic, stearic, linoleic and linolenic vertically. What we learn is that the SVM solution is about as good as possible, and although not easy to simplify using 5 of the 8 variables may provide an adequate classification.

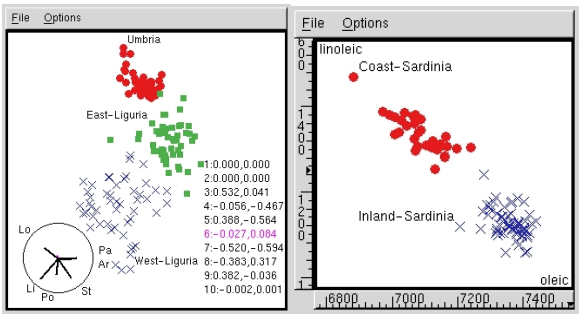

Figure 4: (Left) Projection of northern Italy oils. (Right) Projection of Sardinian oils

Sardinia: The oils from coastal Sardinia are easily distinguishable from the inland oils using oleic and linoleic acids (Figure 4). A simple rule would be provided by: Assign to Inland Sardinia if $0.5oleic - 0.75linoleic > 27$.
South: Now this is the interesting region! There are 4 areas. There is no natural way to break down the 4 areas into a hierarchical pairwise grouping to build a classification scheme using SVM. The best but still poor results are obtained if the analyst first classifies Sicily against the rest. The remaining 3 areas can be almost perfectly separated using linear kernels. Why is this?

Looking at the data, using tours, with the oils from Sicily excluded, it is clear that the oils from the 3 other areas are readily separable (Figure 5, left). But when Sicily (orange open squares) is included it can be seen that the variation on the Sicilian oils is quite large (Figure 5, right). These points almost always overlap with the other oils in the tour projections. These pictures raised suspicions about Sicilian oils. It looks like they are a mix of the oils from the other 3 areas. Indeed, from informal enquiries, it seems that this is the case, that Sicilian olive oils are made from olives that are imported from neighboring areas. The solution is to exclude Sicilian oils from any analysis.

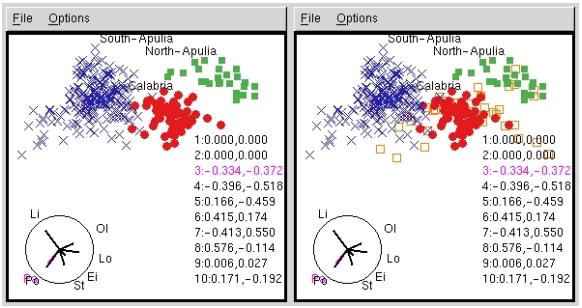

Figure 5: Projection of southern Italy oils

Summary of analysis: In summary the concluding remarks about this data set are to beware of the oils designated from Sicily, as they do not appear to be pure oils. The remaining geographic production areas do seem to produce distinct fatty acid signatures, and its possible to quantify the differences with linear classifiers.

5 Summary and Conclusion

SVMs provide a good complement to visual methods. The results from SVM are visually intuitive, unlike results from methods such as linear discriminant analysis where variance differences can produce a boundary that is too close to the group with the larger variance. Methods such as trees provide boundaries that are restricted to separations in individual variables. Logistic discriminant analysis, though, should compete with SVM and provide similarly accurate and interpretable linear boundaries.

As we raised in the introduction it is reasonable to be laborious in a training phase of building a classifier. Human input to the machine learning process can provide valuable insight into the scientific problem. With the visual tools described in this paper it is possible to visualize class structure in high-dimensional space and use this information to tailor better classifiers for a particular problem. We used just one example data set and one classification technique, but the approach works generally on other real-valued multivariate data, and for understanding other classification techniques.

References

[1] Asimov, D. (1985). *The Grand Tour: A Tool for Viewing Multidimensional Data*. SIAM Journal of Scientific and Statistical Computing **6(1)**, 128 – 11.

[2] Cook, D. and Buja, A. (1997). *Manual Controls For High-Dimensional Data Projections*. Journal of Computational and Graphical Statistics **6(4)**, 464 – 480.

[3] Cook, D. and Buja, A. and Cabrera, J. and Hurley, C. (1995). *Grand Tour and Projection Pursuit* Journal of Computational and Graphical Statistics **4(3)**, 155 – 172.

[4] Forina, M. and Armanino, C. and Lanteri, S. and Tiscornia, E. (1983). *Classification of olive oils from their fatty acid composition*, 189 – 214.

[5] Joachims, T. (1999). *Making Large-Scale SVM Learning Practical*.

Acknowledgement: This work has been supported in part by grants from the National Science Foundation (#9982341, #9972653), and the Iowa State University Graduate College...

Address: Iowa State University, Ames, IA 50011

E-mail: {dicook,dcaragea,honavar}@iastate.edu