Gaining Insights into Support Vector Machine Pattern Classifiers Using Projection-Based Tour Methods

Doina Caragea
Artificial Intelligence Research Laboratory
Department of Computer Science
Iowa State University
Ames, Iowa 50011
dcaragea@cs.iastate.edu

Dianne Cook
Virtual Reality Applications Center
Department of Statistics
Iowa State University
Ames, Iowa 50011
dicook@iastate.edu

Vasant Honavar
Artificial Intelligence Research Laboratory
Department of Computer Science
Iowa State University
Ames, Iowa 50011
honavar@cs.iastate.edu

ABSTRACT

This paper discusses visual methods that can be used to understand and interpret the results of classification using support vector machines (SVM) on data with continuous real-valued variables. SVM induction algorithms build pattern classifiers by identifying a maximal margin separating hyperplane from training examples in high dimensional pattern spaces or spaces induced by suitable nonlinear kernel transformations over pattern spaces. SVM have been demonstrated to be quite effective in a number of practical pattern classification tasks. Since the separating hyperplane is defined in terms of more than two variables it is necessary to use visual techniques that can navigate the viewer through high-dimensional spaces. We demonstrate the use of projection-based tour methods to gain useful insights into SVM classifiers with linear kernels on 8-dimensional data.

Keywords

Dynamic graphics, visualization, machine leaning, classification, support vector machines, tours, classification, multivariate data

1. INTRODUCTION

Support vector machines [9, 15] offer a theoretically well-founded approach to automated learning of pattern classifiers for mining labeled data sets. They have also been shown to build accurate rules in complex classification problems, for example, gene expression analysis using microarray data [2], and text classification [13]. However the algorithms are quite complex and the solutions sometimes difficult to understand. For many data mining tasks understanding a classification rule is as important as the accuracy of the rule itself.

Graphical methods, especially dynamic graphics, provide an important complement to automated classification algorithms in data mining. Pictures can provide easy to digest summaries of complex information. In classification problems, graphics can help us understand the nature of the boundaries between classes and the relative importance of variables for differentiating classes. We explore the use of dynamic graphics methods called tours [1, 8, 7, 4] to examine results from SVM.

Tours provide mechanisms for displaying continuous sequences of low-dimensional linear projections of data in high-dimensional Euclidean spaces. They are generated by constructing an orthonormal basis that represents a linear subspace. Tour methods are most appropriate for data that contain continuous real-valued variables. They are useful for understanding patterns, both linear and non-linear, in multi-dimensional data. However, because tours are defined as projections (analogous to an object shadow) rather than slices some non-linear structures may be difficult to detect. Tours are also limited to applications where the number of variables is less than 20 because otherwise the space is too large to randomly explore within a reasonable amount of time. Hence when we have more than 20 variables, it is important to perform some dimensionality reduction prior to applying tour methods.

This paper describes the use of tour methods for exploring the results of SVM for an application on classifying olive oils according to production region. The data set is one that the authors have some experience in working with. It is interesting because it is a multi-class data set and the boundaries between classes are both simple and complex.

2. METHODS

2.1 Support Vector Machines

Let $\mathcal{E} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l)\}$, where $\mathbf{x}_i \in \mathcal{R}^N$ and $y_i \in \{-1, 1\}$ be a set of training examples for a 2-category classifier. Suppose the training data is *linearly separable*. Then it is possible to find a hyperplane that partitions the N-dimensional pattern space into two half-spaces R^+ and R^-. The set of such hyperplanes (the solution space) is given by $f_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$. SVM selects among the hyperplanes that correctly classify the training set, the one that minimizes $\|\mathbf{w}\|^2$, which is the same as the

KDD '01 San Francisco, CA, USA

hyperplane for which the *margin* of separation between the two classes, measured along a line perpendicular to the hyperplane, is maximized.

If the goal of the classification problem is to find a linear classifier for a non-separable training set, a new set of weights, called slack weights (measuring the extent to which the constraints are violated) can be introduced. In this case the margin is maximized, paying a penalty proportional to the cost of constraint violation. The decision function is similar to the one for the linearly separable problem.

If the training examples are not linearly separable, the SVM works by mapping the training set into a higher dimensional *feature* space using an appropriate kernel function ψ. Therefore, the problem can be solved using linear decision surfaces in the higher dimensional space. Any consistent training set (i.e., one in which no instance is assigned more than one class label) can be made separable with an appropriate choice of a feature space of a sufficiently high dimensionality. However, in general, this can cause the learning algorithm to overfit the training data resulting in poor generalization. In this paper, though, we concentrate on linear classifiers, but discuss extensions to non-linear cases in the final section.

For the experiments in this paper, we used SVM^{light}3.50 [14] implementation of SVM algorithm, that can handle large data sets, as opposed to the traditional quadratic optimization programs that have stringent limitations regarding memory and time. SVM^{light} is currently one of the most widely used implementations of SVM algorithm.

2.2 Tours

Tours are dynamic views of data provided by the manipulation of low-dimensional (D) projections of high-dimensional (N) spaces, where N is the number of variables. The *grand tour*, originally proposed by Asimov [1], consists of a visual presentation of randomly determined low-dimensional projections. A grand tour path is dense in the set of all unoriented D-dimensional planes in $\mathcal{R}^N$, meaning that, if the viewer could watch until the end of time she would see every possible D-dimensional projection of the data. Technically, we define a D-dimensional projection matrix, P, to be a matrix of size $N \times D$, where the columns are orthonormal. P actually defines a plane in $\mathcal{R}^N$. Then a projection of the data would be written as $\mathbf{x}_i P, \quad i = 1, \ldots, l$.

There have been several approaches to implementing grand tours. The method proscribed in Buja et al. [4] is as follows: (1) Generate a series of anchor planes, $P_j, \quad j = 1, \ldots, t$, from the space of all such planes, (2) Interpolate over a geodesic (shortest) path from P_j to P_{j+1}. Generating an interpolation path between two planes requires numerous calculations. The principle is that there is a set of canonical angles between the planes which can be found using singular value decomposition of the space spanned by the two anchor planes. Some examples of tours can be found in [16].

There have been several recent developments in tour methods. Guided, correlation and manually controlled tours are the adaptations that are applied in this paper. In a guided tour more interesting projections can be given a higher probability of a visit than less interesting views during the tour path [8]. A correlation tour [3] is defined as 2×1-dimensional tours, one displayed horizontally and the other vertically in a 2-dimensional display space. This is useful in situations where there are two disjoint sets of variables, in this paper the two sets are explanatory variables ($\mathbf{x}_i$) and class identity (y_i) (or even predicted values). In a manual tour, the user can adjust the projection coefficient of a single variable by rotating it into or out of the current projection [7].

2.3 SVM and Tours

There are several approaches to exploring SVM results using tours: the location of the support vectors in the data space, the SVM predicted values in relation to the explanatory variables, and the weight vectors, $\mathbf{w}$, for examining importance of the explanatory variables to the classification. The grand (random) tour is used for generally exploring support vectors and classification boundaries in the data space. Manually controlled tours are used for studying variable importance. A correlation tour (and manually controlled correlation tour) is used to examine predicted values in relation to explanatory variables.

We examine the distribution of support vectors relative to the other instances in the data set to explore whether tour methods can provide some insight into the behaviour of the SVM algorithm. If the two classes are linearly separable, we expect to see support vectors from each group roughly indicating a boundary between the groups in some projection. The variables contributing to the projection provides an indication of relative importance to separating the groups. The coefficients of the projection (elements of P) are used to examine the variable contribution.

We examine the predicted value $\mathbf{w}\cdot\mathbf{x}+b$ for each instance $\mathbf{x}$ in the space of explanatory variables. By using tour methods and focusing on the predicted values that are close to 0, we can explore the nature of the decision boundary in the space of the explanatory variables. Predictions in the neighborhood of 0 represent instances on the edge of the two groups in linearly separable problems.

Examining the weights $(\mathbf{w}^*, b^*)$ of decision boundary which maximizes the margin of separation between the two classes is a way to explore the importance of variables. If the variables are standardized (zero mean, unit variance) prior to fitting the SVM, the magnitude of the components of $\mathbf{w}^*$ provide some indication of their relative importance in defining the separating hyperplane between the two classes. (If the variables are not standardized then correlations between the predicted values and the variables can be used similarly. The correlation measures the strength of the linear relationship between predicted values and explanatory variables.) Elimination of variables that have negligible correlation with predicted values should result in simpler decision boundaries with little loss in accuracy.

This paper illustrates these three procedures for one example data set but they generalize to any linearly separable problems. Non-linearly separable problems pose further challenges not addressed here.

3. APPLICATION

The *olive oil* data consists of the percentage composition ($\times100$) of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) found in the lipid fraction of 572 Italian olive oils. (An analysis of this data is given in [11]). There are 9 collection areas, 4 from southern Italy (region 1), 2 from Sardinia (region 2), and 3 from northern Italy (region 3). The samples from the

southern region are North Apulia (area 1), Calabria (area 2), South Apulia (area 3) and Sicily (area 4). From Sardinia there are two areas: Inland (5) and Coastal (6) and from northern Italy there are East (7) and West Liguria (8), and Umbria (9). This data was chosen because the authors have a good understanding of the class structure in high-dimensional space from extensive visual exploration, and because it poses some interesting challenges to building classifiers. It is also an interesting practical problem. Oil producers in Italy are very protective of the quality of their product, and the competition seems to arise along regional boundaries. For quality control it is important to be able to verify the oil is from the region it is purported to be from. Classifiers can help here if the growing region can be identified by the composition of fatty acids in the oil samples.

Here is a summary from our visual exploration. The samples from the southern region can be separated from the other two regions by 1 variable alone: eicosenoic acid. The samples from the north can be separated from the Sardinian samples in numerous ways with just 2 variables using a quadratic boundary, or linear using 3 variables. The samples from areas within the southern region are very difficult to separate cleanly. All variables appear to be necessary. The two areas in Sardinia have a clean linear boundary. The areas from within the northern region separate reasonably well with linear boundaries allowing for a small amount of confusion, using all of the variables. In general, the classes also have heterogeneous shape and variance in the multivariate space. Methods such as linear discriminant analysis and tree classifiers don't perform well here. The visual exploration is important for gaining some initial insight into the class structure. It assists in assessing the sensibility of the results from SVM (or indeed any other type of classifier).

3.1 Application of SVM^{light} 3.50

The data was divided by 100, to put it back in the range 0-100 of raw percentage values. Since the SVM^{light} algorithm works with only two classes, we need to work sequentially through pairs of classes (see Table 1). There are many ways to do this. The order used here was in part suggested by the hierarchical class structure, but somewhat arbitrarily chosen for classifying sub-regions. Only one problem with the ordering was encountered, related to the sub-regions of the southern samples.

The original data set is randomly divided into two subsets, training set and test set. The training set contains roughly three fourths of the data, while the test set contains the remaining one fourth. The kernel used for the experiments with SVM^{light} is a polynomial with degree 1 (this is the same as a linear kernel), and the bound C was chosen to be 1000. All the other parameters are the default parameters. After the model was constructed, it was used to classify both the training set and the test set. The accuracy results are reported in Table 1. Although there are more sophisticated ways to estimate accuracy in prediction the results reported here are reasonable and reflect the expected accuracies based on the preliminary visual inspection.

3.2 Visual Examination of SVM Results

In Figure 1 the support vectors for separating region 1 from regions 2,3 are marked as large solid circles. In the leftmost plot Region is plotted against eicosenoic. Eicosenoic acid was identified by the preliminary visual analysis as the most important variable for separating regions. We would expect that in this variable the support vectors would all be in the boundary of the regions for eicosenoic acid, that is, take values close to 10 for region 1, and values close to 3 for regions 2,3. Interestingly, the support vectors do not lie on the edge of the two groups. Using a manually controlled correlation tour we explore the influence of other variables. Rotating palmitic acid with a positive component in a linear combination with eicosenoic brings the support vectors from region 1 closer to the boundary with regions 2 and 3 (middle plot). Also, subtracting a small component of stearic acid from the linear combination of eicosenoic and palmitic brings the support vectors from regions 2,3 closer to the boundary with region 1. It appears that these three variables were detected to be important by the SVM for separating region 1 from regions 2,3.

If we examine the correlations with predicted values (Table 2) with each variable we find that eicosenoic acid is the most important, followed by palmitic, palmitoleic, and negative oleic acids. Starting with eicosenoic acid and manually touring a small number of each of these variables into the view (i.e. altering the projection coefficients to include more of these variables in the projection) gives similar results. It is clear that the SVM has used a combination of variables to generate the space of best prediction. Figure 2 shows a histogram of the predicted values from SVM, and also a histogram of eicosenoic acid alone. Both show good separations between the two groups. On re-fitting the SVM using only one variable, eicosenoic acid, we obtain the same accuracy as that obtained using the more complex model.

Figure 3 shows the support vectors as large solid circles in plots of oleic vs linoleic acid, and a linear combination of oleic, linoleic and arachidic acids from a manually controlled correlation tour. The location of the support vectors are roughly where we would expect them: on the boundary between the two groups in the variables oleic and linoleic acids. From the visual examination of the data these two variables emerged as the major variables for separating the two regions. The correlation tour is used to examine the location of the support vectors in the combination of variables oleic, linoleic and arachidic acids. The separation is stronger in these 3 variables, and the support vectors are still roughly in the boundary between the two groups. The correlations of predicted values and the individual variables (Table 2) also indicates that these 3 variables are the most important in the SVM prediction. Refitting the model with only these variables gives the same 100% accuracy.

Understanding the results of classifying areas within southern Italy are much more difficult. These are not linearly separable groups. To separate areas 1 (North Apulia) and 2 (Calabria) from areas 3 (South Apulia) and 4 (Sicily) 7 support vectors are used, and 29 vectors are between the two groups (Table 1). The correlation between predicted values and variables (Table 2) suggest that variables oleic, linoleic, palmitic and palmitoleic are important. Figure 4 shows the support vectors (solid circles) and the slack vectors (open circles) in relation to all the instances for Region 1 in a projection of these variables. The support vectors and slack vectors are spread throughout the view. If we ignore the area Sicily (area 4) then this view provides a good separation of all three other areas. In the initial visual examination of the data we did find that Sicily was the most difficult area to classify, it overlaps with all three other areas. The other

Table 1: Accuracy results

Groups	Subgroups	Tr.Ex.	Ts.Ex.	SV	Slack	Tr. acc.	Test acc.
All	1:2,3	436	136	9	0	100	100
	2:3	190	59	4	0	100	100
Group 1	11,12:13,14	246	77	7	29	89.02	89.47
	11:12	60	20	4	0	100	100
	13:14	186	57	8	0	100	98.21
Group 2	25:26	74	24	3	0	100	100
Group 3	37,38:39	116	35	7	0	100	97
	37:38	76	24	6	0	100	100

Table 2: Correlations of predictions with individual variables

Group	palm	p'oleic	stearic	oleic	lino	l'enic	arach	eico
1:2,3	0.688	0.613	0.036	-0.616	0.267	0.506	0.288	0.940
2:3	0.052	0.198	-0.131	-0.798	0.851	0.153	0.477	-0.001
11,12:13,14	-0.604	-0.766	0.398	0.782	-0.808	0.588	0.199	0.290

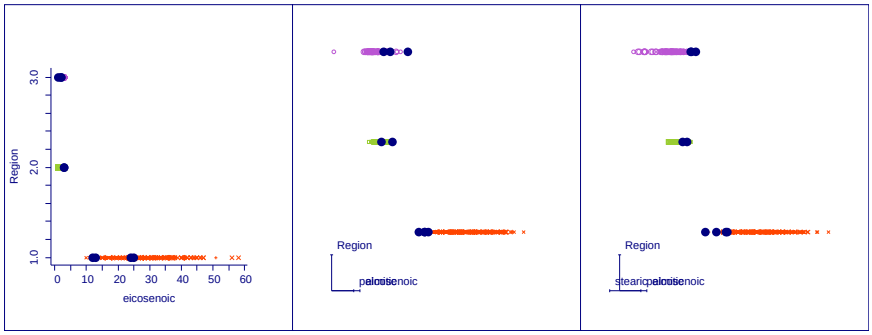

Figure 1: Olive oil data: Support vectors for separating region 1 from regions 2,3 are marked as large solid circles. (Left) Region vs eicosenoic. Interestingly, the support vectors do not lie on the edge of the two groups in the plot of eicosenoic acid alone. (Middle, Right) The manual correlation tour is used to introducing palmitic in an additive linear combination with eicosenoic, and subtract a small amount of stearic acid. The support vectors are now clearly on the boundary between the two groups.

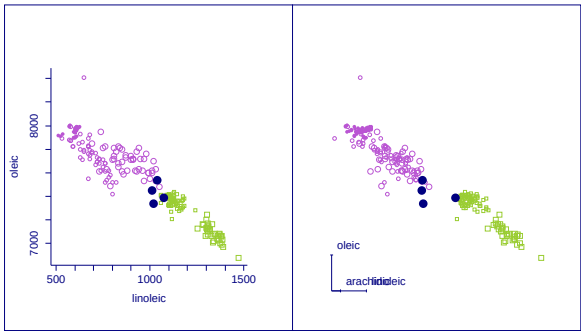

Figure 3: Olive oil data: Support vectors for separating region 2 from regions 3 are marked as solid circles. (Left) Oleic vs linoleic. The support vectors are where we would roughly expect them to be relative to the boundary between the two groups. (Right) Correlation tour used to explore the influence of arachidic acid. Arachidic acid is added to linoleic acid, increasing the separation between the two groups.

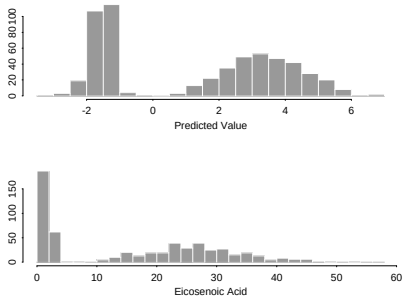

Figure 2: Olive oil data: (Top) Histogram of predicted values from SVM classification of region 1 vs regions 2,3. (Bottom) Histogram of eicosenoic acid shows as good a separation.

three are basically linearly separable. The misclassification table (Table 3) confirms this: roughly half this group (14) are misclassified. One might wonder if there is something suspect about the sample from Sicily: could it be that they are not truly from the same region, or are the growing conditions more variable in Sicily than in other areas? It may be that this high classification error truly reflects the variability in the sample, and that using a non-linear classifier to build a more accurate rule may give misleading information. In light of these observations, though, a slightly simpler strategy for building a better classification rule is to use a different pairing of groups. For example, using a different kernel to classify oils from Sicily (4) from the other three groups, and then a linear separator for areas North/South Apulia and Calabria (1,2,3) works better (Table 4).

Table 3: Error distribution in the case 11,12:13,14

Group(class)	Pos prediction (+1)	Neg prediction
11(+)	16	3
12(+)	31	10
13(-)	156	3
14(-)	16	11

Figure 5 examines the explanatory variable space of instances with predictions close to zero. Instances with predictions close to zero are highlighted as solid circles in these plots. For the SVM classifying region 2 from 3 (Figure 5) it is clear that instances with predictions close to zero are on a (3 variable) linear boundary between the two groups. They are not on the (2 variable) non-linear boundary (right plot) which shows clearly that the SVM, as expected, detected the linear separation not the non-linear separation.

4. SUMMARY AND DISCUSSION

The research reported in this paper is part of a larger project on visualization methods for large data sets [17]. The project is exploring the use of SVM as a preprocessor, both in terms of reducing the number of variables to enter into the tour, and to reduce the number of instances to the set of support vectors (which is much smaller than the data set). In related work, we are exploring distributed learning algorithms [5], [6] for visualization of large distributed data

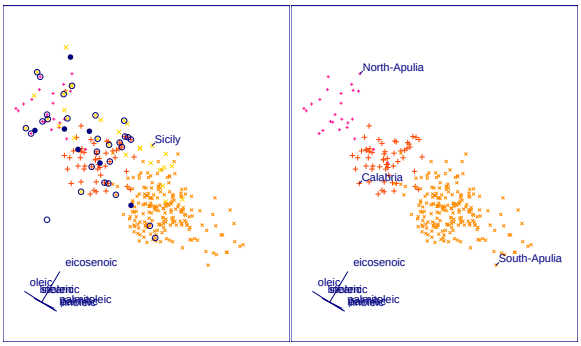

Figure 4: Olive oil data: (Left) Support vectors for the areas 1,2 vs 3,4 in the southern region are highlighted as large solid circle. View is the combination suggested by correlations between variables and predicted values. (Right) Same view with instances corresponding to Sicily are removed, revealing the neat separation between the other three areas.

sets.

In general, the tour methods can be applied for a small number of variables, not thousands of variables. Tour methods can provide us better insight into the nature of class structure in data from 2 to 20 dimensions than almost any other graphics method. Hence there is a need for ways to select variables or subspaces to explore with the tour. Methods such as principal component analysis are not well-suited for dimension reduction in classification problems (see [10]). The tour methods discussed here can be applied to data containing up to 1 million instances. The main problem with a large number of instances is that the rendering method, scatterplots, produces too much overplotting.

Tours are defined for linear projections. Linear projections are simple to understand, and there is a real-world analogy: shadows that objects make from a light source. They have the same strengths and weaknesses as shadows; we can see object curvature with a shadow but concavities and hollow, empty centers may be difficult to detect. Furnas & Buja [12] discuss methods for detecting such structure using sections or slices rather than projections. However this introduces considerable complexity and greatly increases the number of parameters needed to run an appropriate tour. The methods described in previous sections will work with non-linear SVM to some extent. It is possible to explore the pattern of the support vectors from a non-linear kernel, and it may be possible to detect a non-linear boundary in $\mathcal{R}^N$. However, using correlation between predicted values and explanatory variables may not accurately describe the relative importance of variables in non-linear SVM. Similarly exploring the instances with predicted values near zero may not be helpful because the relationship is non-linear. Rather a correlation tour could be used with predicted values plotted against combinations of explanatory variables to explore the non-linear dependencies.

Lastly, the approach we have described is very labor intensive for the analyst. It cannot be automated because it relies heavily on the analyst's visual skills and patience for watching rotations and manually adjusting projection coefficients. However, it provides insights which we may not otherwise be able to make. However, in the quest for automating classifi-

Table 4: Accuracy results for group 1

Group	Subgroups	Tr.Ex.	Ts.Ex.	SV	Slack	Tr. acc.	Test acc.
Group 1	11,12,13:14	246	77	30	1	100	95.59
	11:12,13	219	67	6	0	100	100
	12:13	200	59	7	0	100	96.72

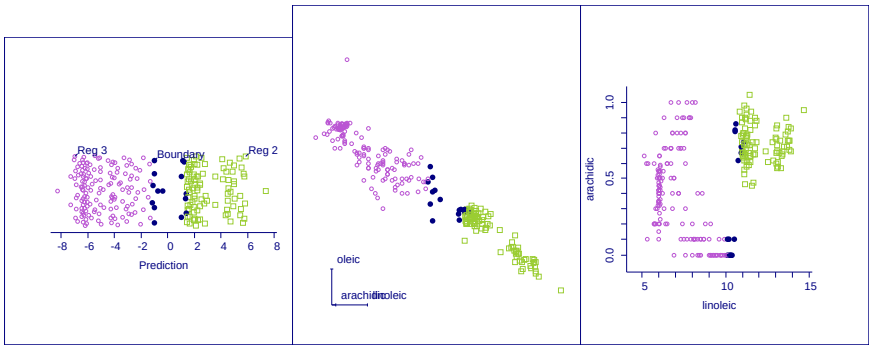

Figure 5: Olive oil data: (Left) Instances in the neighborhood of predicted value zero are highlighted as solid circles. (Middle) These instances are clearly in the decision boundary for the two groups in the explanatory variables oleic, linoleic and arachidic acids. (Right) They are clearly not in the non-linear boundary between the two groups when only arachidic and linoleic acids are considered.

cation tasks these methods can only be used in preliminary stages, to refine and understand algorithms.

Acknowledgements

This work has been supported in part by grants from the National Science Foundation (#9982341, #9972653), the Carver Foundation, Pioneer Hi-Bred, Inc., John Deere Foundation, and the Iowa State University Graduate College.

5. REFERENCES

[1] D. Asimov. The Grand Tour: A Tool for Viewing Multidimensional Data. *SIAM Journal of Scientific and Statistical Computing*, 6(1):128–143, 1985.

[2] M. Brown, W. Grundy, D. Lin, N. Christianini, C. Sugnet, T. Furey, M. Ares Jr., and D. Haussler. Knowledge based analysis of microarray gene expression data using support vector machines. TR UCSC CRL-99-09, CRL, Santa Cruz, CA., 1999.

[3] A. Buja, D. Asimov, C. Hurley, and J. A. McDonald. Elements of a Viewing Pipeline for Data Analysis. In W. S. Cleveland and M. E. McGill editors, *Dynamic Graphics for Statistics*, pages 277–308. Wadsworth, Monterey, CA, 1988.

[4] A. Buja, D. Cook, D. Asimov, and C. Hurley. Dynamic Projections in High-Dimensional Visualization: Theory and Computational Methods. TR, AT&T Labs, Florham Park, NJ, 1997.

[5] D. Caragea, A. Silvescu, and V. Honavar. Agents that learn from distributed dynamic data sources. In *Proc. of the Workshop on Learning Agents, Agents 2000/ECML 2000, Barcelona, Spain*, 53–61, 2000.

[6] D. Caragea, A. Silvescu, and V. Honavar. Analysis and synthesis of agents that learn from distributed dynamic data sources. In S. E. A. Wermter, Ed., *Neural Network Architectures Inspired by Neuroscience*. Springer-Verlag, 2001.

[7] D. Cook and A. Buja. Manual Controls For High-Dimensional Data Projections. *Journal of Computational and Graphical Statistics*, 6(4):464–480, 1997. Also see `www.public.iastate.edu/~dicook/research/papers/manip.html`.

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

[9] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[10] D. J. Donnell, A. Buja, and W. Stuetzle. Analysis of Additive Dependencies using Smallest Additive Principle Components. *Annals of Statistics*, 22, 1994.

[11] M. Forina, C. Armanino, S. Lanteri, and E. Tiscornia. Classification of olive oils from their fatty acid composition. In H. Martens and H. Russwurm Jr., editors, *Food Research and Data Analysis*, pages 189–214. Applied Science Publishers, London, 1983.

[12] G. Furnas and A. Buja. Prosection Views: Dimensional Inference Through Sections and Projections. *Journal of Computational and Graphical Statistics*, 3(4):323–385, 1994.

[13] M. Hearst, B. Scholkopf, S. Dumais, E. Osuna, and J. Platt. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, 1998.

[14] T. Joachims. *Making Large-Scale SVM Learning Practical*. MIT Press, 1999.

[15] V. Vapnik. *The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science)*. Springer-Verlag, New York, NY, 1999.

[16] `www.public.iastate.edu/~dicook/JSS/paper`

[17] Limn: `www.public.iastate.edu/ dicook/Limn`