Classifier Optimization Via Graph Complexity Measures

J. L. Solka
Code B10
NSWCDD
Dahlgren, VA 22448

D. A. Johannsen
Code B10
NSWCDD
Dahlgren, VA 22448

Abstract

This paper examines the use of minimal spanning trees as an alternative measure of classifier performance. The ability of this measure to capture classifier complexity is studied through the use of a gene expression dataset. The effect of distance metric on classifier performance is also detailed within.

Keywords

minimal spanning trees, gene expression data, classifier complexity, distance metric

Introduction

Given a set of observations X along with a set of associated class labels C one is often interested in constructing a classifier which is essentially a mapping from X to C. The problem with constructing such a mapping comes when the inherent dimensionality of the observations is high. In fact many modern datasets may consist of less than one hundred observations in several thousand dimensions. Despite the requisite considerations for the curse of dimensionality the community that produced such data still wishes to be able to determine the classification mapping. In fact they are often interested in identifying a small, less than 100, subset of the collected data dimensions wherein the various classes are well separated. Given the fact that in the multiclass case the identified dimensions might be different for each of the particular classes or even different when we compare each class against every other class, one by one, the selection of these particularly fruitful dimensions might be particularly daunting.

Selection of the features is of course predicated on some sort of measure of their ability to contribute to the discriminant analysis of the particular class. One can formulate various figures of merit to use including cross-validated nearest neighbor classifier performance. This figure of merit while simple to implement is quite computationally intensive. Potential benefits may be obtained from the use of other figures of merit that might serve as an alternative to the cross-validated nearest neighbor performance, while still capturing the ability of the the cross-validated nearest neighbor performance to quantify the difficulty of the classification problem.

Another factor that can contribute to the classifier performance is the manner in which one measures the distance between observations. Our previous efforts (ref

CEP JSM 2000) indicated the benefit of varying the choice of Minkowski p parameter when one measures the distance between observations. In fact the creation of any sort of classification scheme is predicated on the ability to measure distances between observations. We are of course interested not only in how classifier performance might change as a function of the Minkowski p parameter but also how any sort of proposed complexity measure might change as one varies the p parameter.

Given these preliminary discussions the paper unfolds as follows. First we provide some background discussions as to our approach to the feature identification, classification complexity quantification, metric space adaptation methodology. Second we illustrate the application of the discussed approaches to an artificial nose and gene expression dataset. Finally we close our discussions with some indications of our future plans for continuing work on these issues.

Background

Given X a n x d matrix of observations and a set C of n associated class labels chosen from the set $1, 2, \cdots, nc$ then one is interested in constructing a mapping $Y : X \rightarrow C$ that associates with each $x \in X$ a class label. The simplest way to measure the efficiency of said classifier is by the cross-validated single nearest neighbor error rate $\hat{L}$ or alternatively by the cross-validated single nearest neighbor probability of correct classification $1 - \hat{L}$.

The performance of the classifier is in part determined by the way that we choose to measure distances between the observations. We choose to write a generalized distance function $d(x, y)$ between two observations in our original space as

$$d(x, y) = (\sum_{i=1}^{d} w_i |s(x_i) - s(y_i)|^p)^{\frac{1}{p}} \quad (1)$$

where w is a weighting vector that indicates how much each of the d dimensions contributes to the distance, s is a smoothing function, and p is a parameter that indicates the form of the Minkowski metric. We have previously presented results that examine the effects of varying these parameters on single nearest neighbor classifier performance as applied to an artificial nose dataset [Priebe et al., 2000].

The purpose of the current article is to continue these discussions and also to examine the use of a minimal spanning tree as a classifier complexity measure. Previously Friedman and Rafsky proposed a test based on the minimal spanning tree to determine whether two samples of observations were drawn from the same distribution [Friedman and Rafsky, 1979]. Recall that given a graph G a minimal spanning tree is a connected subgraph that has no cycles which covers each point in the graph and whose sum of edge weights is minimal [Gross and Yellen, 1999].

One can use this method to characterize classifier complexity as follows. First compute the minimal spanning tree based on the full set of observations. Second, count those edges in the minimal spanning tree that travel between disparate class types. The number of such edges is then used to characterize the complexity of the problem.

Figure 1 shows a representative spanning tree obtained based on a sample of 100 points from two bivariate normals, $\mu_1 = [1, 1]$,$\mu_2 = [-1, -1]$, $\sigma^1_{xx} = 1.5$, $\sigma^1_{yy} = 1.$, $\sigma^2_{xx} = 1.$, $\sigma^2_{yy} = 1.5$. The red edges indicate edges between disparate classes.

At times in the results section we will be interested with which subset of d-dimensional features contributes in a positive manner to the discernibility of the two

classes. One classic way to characterize the manner in which a particular feature contributes is the scatter. There are numerous ways that this term has been defined in the literature but we will follow Duda, Hart, and Stork, 2000 [Duda et al., 2000]. We begin by considering two classes with $\mathcal{C}_1$ and $\mathcal{C}_2$, with n_1 and n_2 members (respectively).The class means are given by

$$m_i = \frac{1}{n_i} \sum_{x \in \mathcal{C}_i} x.$$

The class scatter matrices are given by

$$S_i = \sum_{x \in \mathcal{C}_i} (x - m_i)(x - m_i)^t.$$

The within class scatter is defined by

$$S_W = S_1 + S_2.$$

The between class scatter is defined by

$$S_B = (m_1 - m_2)(m_1 - m_2)^t.$$

In the univariate case S_B and S_W are scalers and we choose those features with large values of S_B/S_W. In the multivariate case S_B and S_W are no longer scalers and we choose those features with large values of $\mathrm{tr}(S_B)/\mathrm{tr}(S_W)$.

Results

We have chosen to illustrate the performance of our algorithms using two datasets. The first dataset is an artificial nose dataset. The response of a nonlinear fiber optic artificial nose was measured when it was exposed to a target compound, trichloroethylene, in conjunction with various confusers, Coleman fuel, chloroform, and others, against just the confusers mixed with air. The dataset used for our analysis consisted of 80 observations with target compound and confuser along with 80 observations or just the confuser compound. Each observation consists of 19 fibers sampled at 2 wavelengths 60 times during the fiber exposure phase. Hence the response of the system consists of a point in a 2280 dimensional space.

The second dataset that we will be discussing is a gene expression dataset that was previously analyzed by Golub et al 1999, [Golub and Slonin, 1999]. This Affeymetix dataset consists of measurements of over 7000 genes on a group of 72 patients all of whom were currently ill with leukemia. The patient population was divided between those patients with the acute lymphoblastic variant (ALL), 47, and the acute myeloblastic variant (AML), 25. The discriminant analysis problem in this case was distinguishing between the ALL and AML varieties.

Figure 2 presents average cross-validated single nearest neighbor classifier performance, green curve, along with average cross-validated minimal spanning tree based complexity, blue curve, as a function of Minkowski p parameter obtained using the raw artificial nose data. We notice that the complexity curve is low when the performance curve is high, and that the complexity curve is high when the performance curve is low. We also note that the optimal performance is obtained at a Minkowski p parameter of 5. This type of deviation from the standard Euclidean parameter value of $p = 2$ is in keeping with our previous work on this dataset [Priebe et al., 2000].

Given an optimal p parameter value of 5 we would like to know how the performance varies as a function of the included scatter selected fibers. Figure 3 shows average cross-validated single nearest neighbor performance as a function of scatter selected fibers at $p = 5$ obtained using the raw artificial nose data. We notice that this plot indicates that one can improve upon the optimal performance of .75 obtained using all 38 fibers at a $p = 5$ by using the top 21 scatter selected fibers at a $p = 5$, measured performance level of .78125. We need to clarify that the ranking of the scatter selected fibers was obtained by computing the scatter value for each of the fibers individually and then ranking them based on the obtained scatter values. One again we point out the fact that the performance and complexity curves seem to be mirror images of one another. This is the type of behavior that we would expect if the minimal spanning tree complexity measure was truly capturing the difficulties associated with the classification problem.

The previous results detail the benefits of first choosing an optimal Minkowski p parameter and then choosing an optimal weighting or w parameter. Next we consider first choosing a smoothing function s, followed by a Minkowski p parameter, and finally an optimal w parameter. It is the hope that one can improve classifier performance by first choosing s, then p and finally w. First we smoothed the nose data using a spline-based smoother. Figure 4 presents average single nearest neighbor cross-validated performance and average complexity as a function of Minkowski p parameter for the spline smoothed nose data. The optimal p value of 29 with an associated performance of .85625 is in keeping with our previous studies of this data. The behavior of the complexity curve is as expected.

Proceeding as we did when we analyzed the raw nose data we next present a plot of performance as a function of scatter selected fibers at the associated optimal p parameter value. Figure 5 presents average single nearest neighbor cross-validated performance as a function of scatter selected fibers at $p = 29$ obtained using the spline smoothed nose data. In this case we were not able to improve upon the performance obtained using all 38 of the fibers

We next turn our attention to an analysis of the Golub gene expression dataset. Figure 6 presents average cross-validated single nearest neighbor performance, green curve, and and average complexity, blue curve, as a function of Minkowski p parameter for the full Golub gene expression dataset. We note that the minimal spanning tree complexity measure seems still to perform quite well and that the optimal performance level of .8194 is obtained at a p parameter value of 4.

Figure 7 shows average cross-validated single nearest neighbor performance a function of scatter selected genes for the full Golub data at $p = 4$. We notice that we are able to improve upon the performance associated with using all 7000 of the genes by using the top 372 scatter selected genes, performance level = .84722.

The last analysis that we will present uses a reduced set of Golub genes obtained as follows. First we use only those genes that have an expression level of 20 or greater across all patients. Now consider a n_g genes by n_s patients data matrix. We first divide each column by its mean. Next we subject each row to a standard normalizing transformation. Our analysis then proceeds forward with this reduced set of 1753 genes. Figure 8 presents performance and complexity as a function of Minkowski p parameter for the reduced Golub gene expression dataset.

We notice that the optimal performance is obtained at a Minkowski p parameter value of 4. We also notice that once again the MST based complexity seems to capture the associated discriminate difficulty.

Finally we wish to examine the performance as a function of scatter selected genes for the reduced Golub dataset. Figure 9 shows performance as a function of

Table 1: Results Summary.

Data	Number of Features	p	$1 - \hat{L}$
Nose	38	2	.7
Nose	38	5	.75
Nose	21	5	.78125
Smoothed Nose	38	2	.70
Smoothed Nose	38	29	.85625
Smoothed Nose	13	29	.7825
Golub Gene	7129	2	.805
Golub Gene	7129	4	.8194
Golub Gene	372	4	.84722
Reduced Golub Gene	1853	2	.8
Reduced Golub Gene	1853	4	.84722
Reduced Golub Gene	1698	4	.8611

scatter selected genes for the reduced Golub data at $p = 4$.

In this particular case the full performance level is obtained when one uses 1698 of the original 1753 genes that appear in the reduced gene dataset. We do note that we can obtain a performance level of on the order of .82 utilizing fewer than 200 of the original reduced genes along with an associated Minkowski p parameter value of 4.

Conclusions

Our results are summarized in Table 1. We can see that one can make great improvements in classifier performance by merely adjusting the Minkowski p value. This is in keeping with our previous paper that just focused on the artificial nose data. We also note that in many cases one can improve performance not only by optimizing the p parameter value but by also following this optimization with a down select of the features utilizing a measure such as scatter for a figure of merit to use in our forward selection process. Finally we note that in the case of the artificial nose data, we demonstrated the advantage of using a smoother prior to the p value calculation.

We have not discussed how the magnitude/value of the Minkowski p parameter might be related in some sense to the structure of the data that we are dealing with. We have also not examined the simultaneous optimization of p parameter selection, smoother and feature down selection process. This research is relegated to future endeavors.

On the complexity front, we have shown that for the two applications at hand, the artificial nose data and the gene expression dataset that the MST based complexity measure does a good job of capturing the difficulty of a classification problem. Currently we merely provide this information as an observation. We have not to date examined the incorporation of the MST based complexity characterization into the general feature selection process. This too must be relegated to future work.

Acknowledgments

The authors would like to thank Wendy Martinez of the Office of Naval Research for funding this effort.

References

[Duda et al., 2000] Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. Wiley-Interscience, ssecond edition.

[Friedman and Rafsky, 1979] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717.

[Golub and Slonin, 1999] Golub, T. R. and Slonin, D. K. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

[Gross and Yellen, 1999] Gross, J. and Yellen, J. (1999). *Graph Theory and Its Applications*. CRC.

[Priebe et al., 2000] Priebe, C. E., J., M. D., and Solka, J. L. (2000). On the selection of distance for a high dimensional cla ssification problem. *ASA Proceedings of the Sections on Statistica and Statistical Graphics*, (58–63).

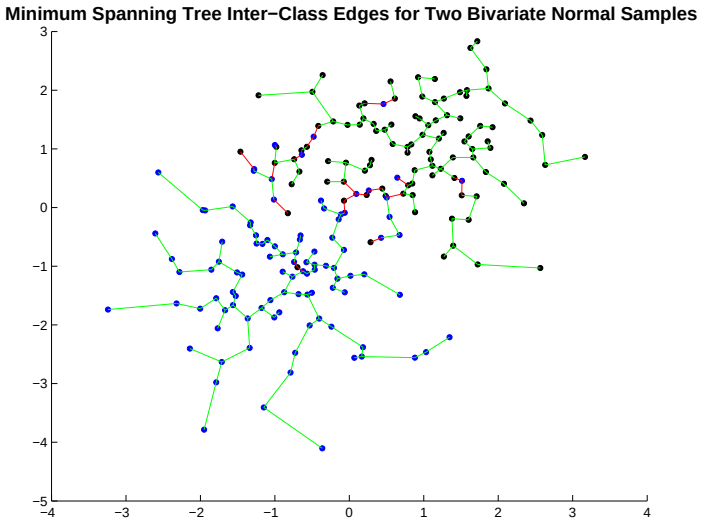

Figure 1: Example minimal spanning tree computed based on two samples from fairly well separated bivariate normals. The red edges are between disparate classes.

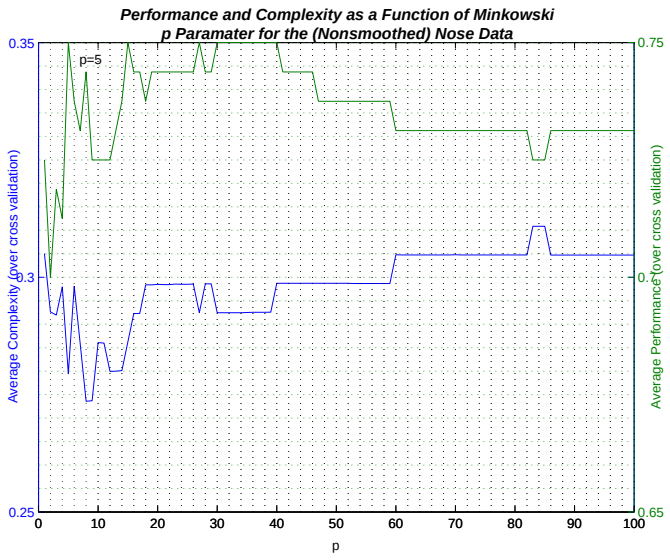

Figure 2: Average single nearest neighbor cross-validated performance, green curve, and average minimal spanning tree based complexity, blue curve, as a function of Minkowski p parameter based on the raw artificial nose data.

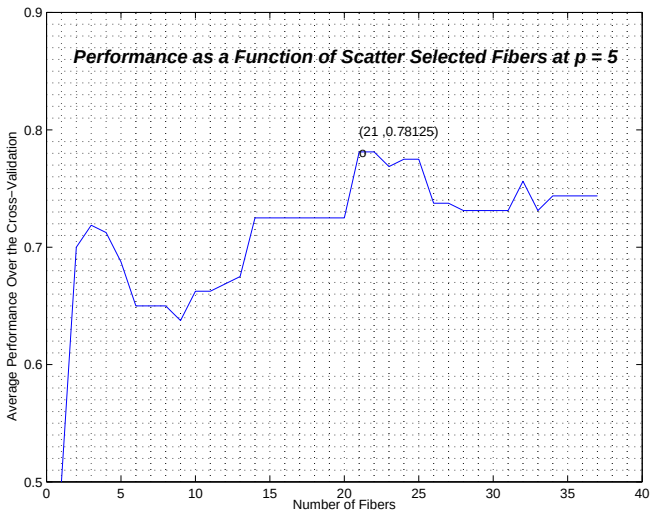

Figure 3: Average single nearest neighbor cross-validated performance as a function of number of scatter selected fibers at $p = 5$ for the raw artificial nose data.

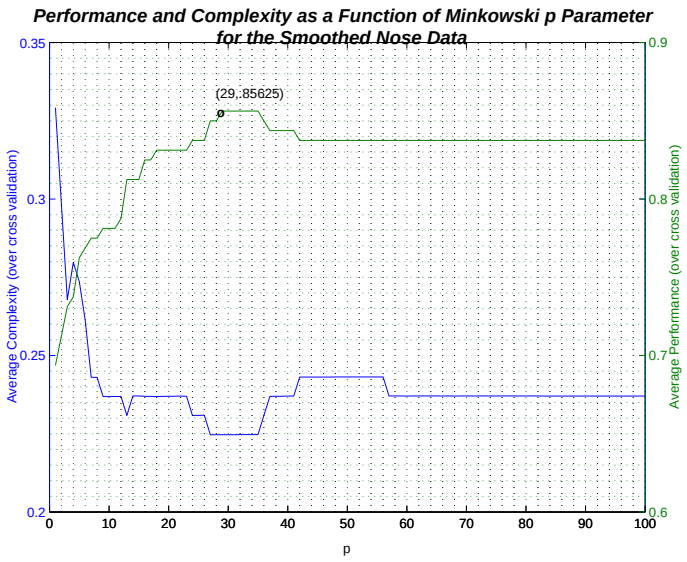

Figure 4: Average cross-validated single nearest neighbor performance, green curve, and average complexity,blue curve, as a function of Minkowski p parameter for the spline smoothed nose data.

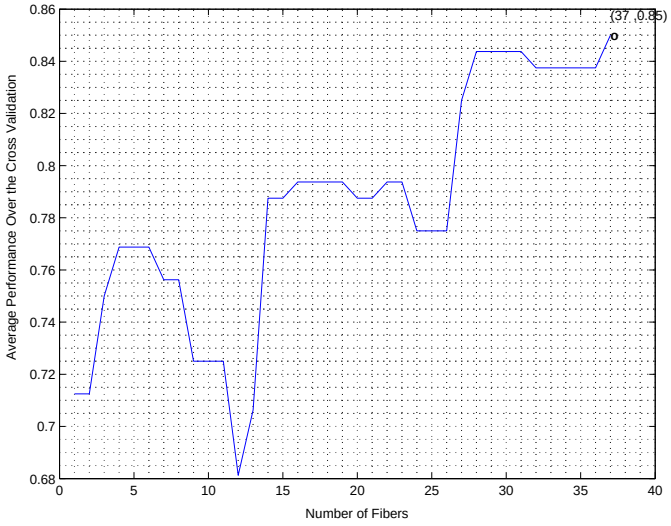

Figure 5: Average single nearest neighbor cross-validated performance as a function of number of scatter selected fibers at $p = 29$ obtained using the spline smoothed nose data.

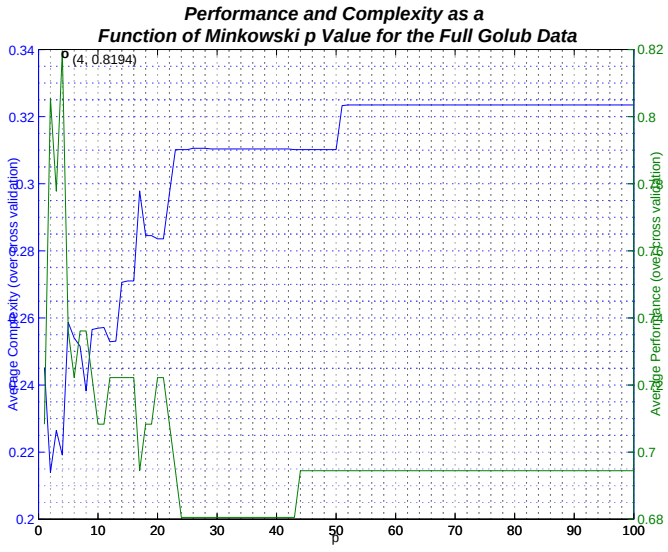

Figure 6: Average cross-validated single nearest neighbor performance, green curve, and average complexity, blue curve, as a function of the Minkowski p parameter for the full Golub gene expression dataset.

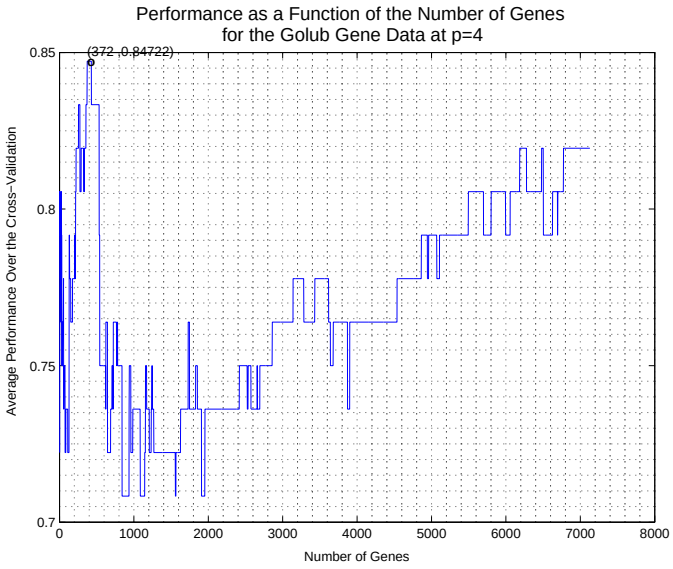

Figure 7: Average cross-validated single nearest neighbor performance as a function of scatter selected genes for the full Golub dataset at an optimal $p = 4$.

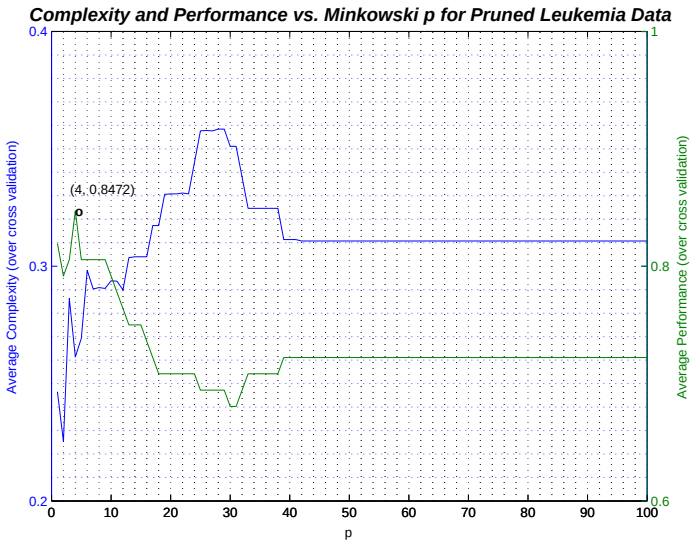

Figure 8: Performance and complexity as a function of the Minkowski p parameter for the reduced Golub gene expression dataset.

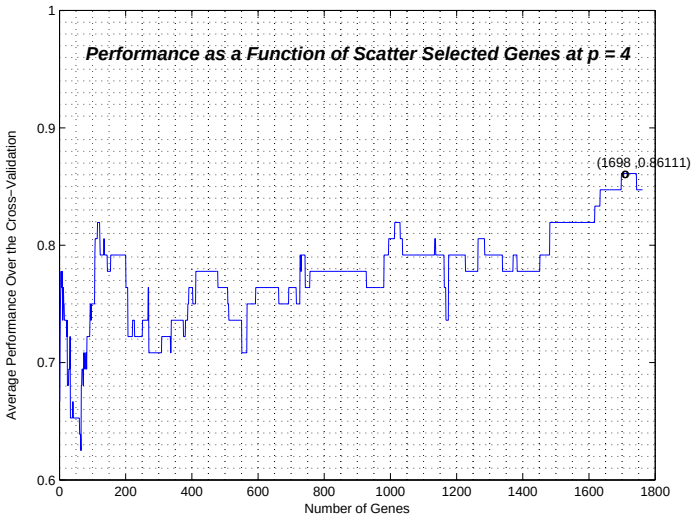

Figure 9: Performance as a function of scatter selected genes for the reduced Golub dataset at an optimal $p = 4$.