Drug Screening
by Nonparametric Posterior Estimation

Alexander Gray

February 2004
CMU-CS-04-109

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

To be presented at ENAR 04.

Abstract

Automated high-throughput drug screening constitutes a critical emerging approach in modern pharamaceutical research. The statistical task of interest is that of discriminating active versus inactive molecules given a target molecule, in order to rank potential drug candidates for further testing. Because the core problem is one of ranking, our approach concentrates on accurate estimation of unknown class probabilities, in contrast to popular non-probabilistic methods which simply estimate decision boundaries. While this motivates nonparametric density estimation, we are faced with the fact that the molecular descriptors used in practice typically contain thousands of binary features. In this paper we attempt to improve the extent to which kernel density estimation can work well in high-dimensional discrimination settings. We present a synthesis of techniques (SLAMDUNK: Sphere, Learn A Metric, Discriminate Using Nonisotropic Kernels) which yields favorable performance in comparison to previous published approaches to drug screening, as tested on a large proprietary pharmaceutical dataset.

The author was supported by the NASA Graduate Research Fellowship.

Keywords: virtual screening, high-throughput screening, classification, nonparametric, metric learning, nonisotropic, kernel density estimation

1 Introduction: Discrimination for Drug Screening

1.1 Automated Drug Screening

Virtual screening refers to the use of statistical and computational methods for prioritizing candidate molecules for biological testing for their possible use as drugs. Because these assays are time-consuming and expensive, accurate "virtual" assays, or prioritization of molecules by computer, has direct impact in cost savings and more rapid drug development. Virtual screening, which is part of the more general enterprise of *high-throughput screening*, has thus become an increasingly pressing new component of modern drug development research.

The discrimination problem. Several scenarios exist for the specific setup of the virtual screening problem, and these demand slightly different emphases. In this paper we are concerned with a scenario that is representative of that of a large pharmaceutical research and development laboratory, which is as follows: We assume there is a single *target* molecule. There are multiple molecules which are known to interact in the desired fashion with the target molecule, or are *active* with respect to the target. There are also a number of molecules which are known to be *inactive* with respect to the target. The number of inactive molecules available is generally much larger than the number of active molecules. A very active research area within computational chemistry continues to explore the use of statistical *discrimination* (also called classification) using a set of *features* (or measurements) describing molecular properties to predict whether a previously unseen molecule will be active with respect to the target or not.

The labels. Building such a virtual screening system begins by collecting the *training set*, or the set of labelled molecules (labelled as "active" or "inactive"), often by a mixture of human-intensive but fairly certain biological testing and automated or semi-automated testing whose outcome may retain some uncertainty. Labels are often obtained by thresholding a continuous "activity" level. Datasets are also sometimes formed combining molecules obtained from outside sources, such as the purchase of datasets from other research groups.

The features. To date, a succinct characterization of the properties of molecules which are relevant to their activity with respect to a target is not known. The structure of a molecule determines its interaction with a target molecule – whether and how it will interlock, or "dock" with the target – but the interaction is itself a complex dynamic process guided by the pairwise potential energies between the atoms of the two molecules (and atoms within the same molecule), whose complete characterization remains an oustanding problem of science. Thus, molecular descriptions used in virtual screening typically contain hundreds or thousands of binary (0/1) features, collecting all manner of both generic and target-specific properties which might be relevant to the discrimination task. Typical binary features record the absence or presence of a certain kinds of atom or substructures, proximity relationships, and so on. Exploration of different ways to characterize molecules in terms of fixed-length vectors of features is itself an active research topic.

Our dataset and goal. In this work our goal is to design a classifier with the best possible prediction performance based on a proprietary commercial training set of 26,733 molecules, 6,348 binary features, and one output variable ("active" or "inactive"). While further details about this dataset cannot be disclosed, it is similar in nature to the kind of data that can be found in public archives such as the NCI AIDS file, which contains compounds which are known to be active or inactive with respect to the HIV virus.

In general, we are primarily empirically motivated rather than philosophically motivated, and take an approach of synthesizing insights and techniques across both statistics and pattern recognition (an outgrowth of electrical engineering) and machine learning (an outgrowth of computer science). We believe that this unified cultural viewpoint is both fruitful and inevitable.

1.2 Previous Discrimination Methods

Recent work in virtual screening. Virtual screening is a rapidly developing area of research. Our work is strongly motivated by two of the most recently published comparisons of discrimination methods for virtual screening ([16],[10]).

The primary methods that have been proposed are more or less those which have enjoyed recent popularity in general, drawn mainly from the fields of pattern recognition and machine learning; they include decision trees, neural networks, and naive Bayes classifiers. However, among them, support vector machines (SVM)

([15]), the subject of perhaps the most recent attention in the study of discrimination, have been distinguished as one of the most successful empirically.

Among the lesser-known methods proposals is the 'binary kernel discriminator' (BKD) of [9], a fairly standard kernel estimator for discrimination using a kernel based on the Hamming distance, which was demonstrated to yield good performance in the virtual screening problem. (We note that the BKD is not formulated directly in terms of decision theory.) In [16], a fairly extensive comparison (by a different group of researchers) between support vector machines and the binary kernel discriminator was performed, demonstrating surprisingly clear superiority in the performance of BKD's over SVM's. In that work, molecule descriptions containing up to about 1,000 features were used.

In [10], which performed experiments using the *same dataset* used in this paper, a conjugate gradient-based logistic regression (LR) method was demonstrated to have consistently favorable performance compared with several popular methods including SVM's with both linear and nonlinear (radial basis function) kernels, decision trees, naive Bayes classifiers, and k-nearest-neighbor classifiers. Interestingly, though logistic regression finds only linear decision boundaries, it outperformed both linear SVM's, which finds linear decision boundaries via a different procedure, and RBF-based SVM's, which have the capacity to represent more complex decision boundaries.

Our work helps to understand and relate the success of both of these methods, and ultimately combines aspects of both to achieve a method with performance superior to either one.

1.3 Ranking Versus Binary Decision-making

To score the ranking performance of a discriminator, we use the standard device of *receiver operating characteristic* (ROC) curves ([2]), which captures more information than simply the percentage of correctly-classified data. An ROC curve is constructed by by sorting the data according to the predicted probability for the "active" class, *i.e.* $P(C_1|x)$. Starting at the origin and stepping through the data in order of decreasing "active" probability, a point on the curve is plotted by moving up one unit if the true label was actually "active" and moving right one unit if the prediction was incorrect. The ROC curve for any discriminator begins at the origin of the graph and ends at the top right corner. One curve reflects ranking performance superior to another to the extent that it sits above it. A summary of an ROC curve is the *area under the curve* (AUC), which is 0.5 for a discriminator which guesses randomly and 1.0 for one which ranks perfectly.

The starting point for the approach of this paper is that the ranking problem is different from the "pure" discrimination problem, and in fact is more difficult, because the quantity of interest is the posterior class probability rather than simply the error rate of making binary decisions. A discriminator may estimate class probabilities with very large bias, but still perform well when scored in terms of accuracy in binary decision-making as long as the order relation between the class probabilities is maintained. As noted by [3], this simple fact may explain the historically puzzling observation that discriminators that are very different in concept (*i.e.* modeling assumptions) tend to yield similar error rates.

In this work we pursue the extent to which direct estimation of posterior class probabilities, as opposed to pure discrimination designed to minimize the binary error rate, might yield superior ranking performance.

There are additional practical advantages to obtaining accurate class-conditional densities. Among them: imputation of missing data is naturally treated, outliers are more naturally identified, and ambiguous data which are difficult to classify are easy to isolate.

2 General Approach: Decision Theory with Nonparametric Density Estimation

Decision theory. Based on the motivation above, we are led naturally to the general framework of statistical decision theory. The posterior class probability $P(C_1|x)$ is expressed in terms of the class-conditional density $p(x|C_1)$:

$$P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)} \tag{1}$$

If the class-conditional distributions on the right-hand side are known, the so-called Bayes error rate is achieved, meaning that no better performance can be achieved.

Typically in pattern recognition applications an empirical Bayes stance is implicitly taken, in which $P(C_1)$ and $P(C_2)$ are estimated from the data. This has significance in our setting, in which the class proportions are significantly different, and extra accuracy is in fact obtained in practice by incorporating this information.

Nonparametric density estimation. When the class-conditional distributions are normal (with diagonal covariance), the resulting estimator is called the (naive) Bayes classifier. We consider the classifier obtained by estimating $p(x|C_1)$ and $p(x|C_2)$ with minimal assumptions, using the nonparametric *kernel density estimator*:

$$\hat{p}(x) = \frac{1}{N} \sum_{i}^{N} K_h(x, x_i) \tag{2}$$

where N is the number of data, $K()$ is called the kernel function and satisfies $\int_{-\infty}^{\infty} K_h(z)dz$, and h is a scaling factor called the bandwidth. We refer to the resulting discriminator as a *nonparametric Bayes classifier* (NBC), for lack of a standard name.

The standard form of kernel which is most often used is the *product kernel*, in which

$$K_h(x, x_i) = \prod_{d}^{D} K_d \left(\frac{\|x - x_i\|}{h} \right), \tag{3}$$

where D is the number of dimensions, *i.e.* the kernel function is a product of D univariate kernel functions, and all share the same bandwidth h. Though we could consider a setup in which separate bandwidths can be adjusted for each dimension, this creates a combinatorial problem which is intractable in our high-dimensional setting. If we ensure that the scales of the respective features are roughly the same, we need only adjust a single parameter h.

Note that a particular advantage of the decision-theoretic framework which is relevant for this problem is that unequal misclassification costs are easily handled.

2.1 Pros and Cons

Since we cannot assume any parametric plausible model for the class-conditional densities, nonparametric estimation is required. Kernel density estimation is the most widely-used and well-studied method for nonparametric density estimation, owing to both its simplicity and flexibility, and the many theorems establishing its consistency for near-arbitrary unknown densities and rates of convergence for its many variants ([14], [13]). However, two main factors have traditionally kept it (and nonparametric density estimation in general) from more widespread applicability, particularly in contexts like the present one:

- *Computational intractability.* Estimation of the density at each of the N points, when performed in the straightforward manner, has $O(N^2)$ computational cost. This quickly becomes prohibitive even for moderate sizes of N.

- *Statistical inefficiency in high dimensions.* Theoretical bounds establish that in the worst case, the number of samples required for accurate kernel density estimation rises exponentially with the dimension. Even for relatively small D, these worst-case numbers are discouraging.

- *Ignores simpler decision problem.* One of Vapnik's central arguments for the non-probabilistic approach underlying the support vector machine is that if the error rate is the desired quantity to be minimized, estimation of entire densities rather than simpler decision boundaries is unnecessary and wasteful of modeling capacity ([15]). Stated differently, the straightforward decision-theoretic approach does not make use of information which can be obtained from the decision boundary, which is possibly more easily characterized than the entire class-conditional densities.

3 SLAMDUNK: Sphere, Learn A Metric, Discriminate Using Nonisotropic Kernels

The SLAMDUNK methodology consists of a set of procedures designed to mitigate the traditional limitations of nonparametric density estimation in the setting of high-dimensional discrimination, so that its distinct advantages may be exploited. We now treat in turn each of the three roadblocks mentioned in the last Section.

3.1 Fast Algorithm for Kernel Density Estimation

Computational intractability is the first major roadblock hit in practice. Computational efficiency impacts statistical inference directly – for example in [16] only 200 data were subsampled for each class to form the training set, due to the computational cost of BKD. In our experiments we use the entire set of 26,733 data. Any high-dimensional context demands the use of as much data as possible, exacerbating the computational issue.

Fortunately, this problem has been largely mitigated in very recent work presenting a fast algorithm yielding simultaneously fast and accurate computation of kernel density estimates ([8]). The method casts the kernel density estimation computational problem within a larger class called 'generalized N-body problems' and is a special case of a more general algorithmic approach called 'higher-order divide-and-conquer' which achieves the best known time complexity (asymptotic order of growth in runtime for a given problem size N) for this class of problems. It is proven in [8] that the algorithm reduces the $O(N^2)$ cost of density estimation at each of the N points to $O(N)$, or *constant* time per point.

It is shown empirically in [8] that the algorithm's time complexity is not exponential in the dimension D, as indicated by well-known worst-case theoretical results ([4]). It is instead conjectured that such algorithms are sensitive to the *intrinsic dimensionality*, the local dimensionality of the manifold upon which the data lies ([5]) (see below).

The algorithm employs techniques of computational geometry, in particular space-partitioning data structures called *ball-trees*, also called metric trees, which are constructed as a preprocessing step in $O(N \log N)$ time. The major constraint imposed by this approach which is most relevant in this context is the fact that ball trees require that the underlying distance be a true metric, heavily relying for example on the triangle inequality. This will become relevant in constraining other parts of our methodology.

3.2 Nonstationary and Nonisotropic Estimators

We now consider extensions to the standard kernel density estimator as described earlier, regarding the parametrization of the kernel function.

Nonstationary estimators. It has long been noted that the assumption of spatial *stationarity*, or a single scale h holding across the entire space is deficient. Visually it is clear that smoothing with a fixed bandwidth is unappealing when the dataset contains regions of differing density, which is inevitable in practice. This problem is clearly seen, for example, in the sparser tails of a typical univariate dataset.

Adaptive (or variable-kernel) kernel density estimators have been studied and shown to be more effective than fixed-width kernel density estimators in experimental studies, *e.g.* [1]. In these estimators, the variable bandwidth h_i for each point x_i is obtained by scaling the single global bandwidth h by a factor

$$\lambda_i \propto \{\tilde{p}(x_i)\}^{-1/2} \tag{4}$$

where $\tilde{p}()$ is a pilot estimate of the density, to which the overall estimator is largely insensitive. Many simple choices can be used for this pilot estimate, including adaptive Gaussian mixture models or piecewise-constant estimates based on multivariate binning, for example by *kd*-trees ([4]).

Nonisotropic estimators. It has been noted by many authors (particularly in the field of machine learning, in which high-dimensional data discrimination and clustering is routinely performed) that in practice it is virtually never the case that a dataset's intrinsic dimensionality is equal to its explicit dimensionality D, *e.g.* [3]. With the assumption that the data lie on a linear manifold, the dimension of the subspace can be estimated using the eigenspectrum from a principal components analysis ([5]). However in general the data may lie on a nonlinear manifold ([12]). A common way estimator of the intrinsic dimension with minimal assumptions has been called, among other things, the correlation dimension ([7]), but amounts to the 2-point correlation function used in spatial statistics. Very often in practice the intrinsic dimension $D' << D$, regardless of which variant of its definition is used.

With this in mind, the standard product kernel, which is *isotropic*, *i.e.* has equal extent in all directions, is a poor match to realistic high-dimensional data. Further, as noted earlier, the behavior of volumes in high dimensionalities, rising exponentially in D, is disastrous when D is large.

Instead we use an estimator in which the univariate bandwidths h_i are replaced by matrices H_i, resulting in a multivariate kernel such as the multivariate Gaussian

$$K_{H_i}(x, x_i) = \frac{1}{(2\pi)^{D/2}|H|^{1/2}} \exp\left\{-\frac{1}{2}(x - x_i)^T H^{-1}(x - x_i)\right\} \tag{5}$$

where $H_i = h\lambda_i\widehat{\Sigma_k}$, with $\widehat{\Sigma_k}$ the covariance matrix estimated from the k nearest neighbors of x_i and x_i. Such estimators have received relatively little study, though one example showing their consistency is [6].

By allowing increased sensitivity to the local manifold of the data, we deflate the extent of the curse of dimensionality in kernel density estimation, relative to the naive product kernel estimator.

4 Coordinate Transformation and Metric Learning

Metric learning. An implicit part of the kernel estimator is the underlying metric used to obtain the distances. The standard Euclidean distance is used by default. It can be seen as a special case of a more general weighted Euclidean distance

$$d(x, y) = \|x - y\| = \sqrt{(x - y)^T W(x - y)} \tag{6}$$

in which the matrix W is diagonal containing all 1's. Rather than assume this special case, we take the stance that the metric weight matrix W should be considered a free parameter to adjust to maximize the performance of our estimator. We refer to this as "learning the metric".

The question of the optimal metric has been heavily studied in the context of discrimination by the nearest-neighbor rule (which can be regarded as a special case of the kernel estimator for discrimination). Although asymptotic results imply that the choice of metric does not affect performance, finite-sample experiments show that marked improvements can be made by adjusting the metric to the task at hand. A general theory has been developed ([11]) which formulates the optimal distance in terms of the Bayes optimal posterior class probabilities. However, this form of "distance" does not in general yield a formal metric. We will ensure that metric properties are retained, for the purpose of using the fast algorithm described earlier, by staying within the confines of weighted Euclidean distances.

The linear discriminant metric. We propose a form of W which relates the metric to the decision boundary corresponding to a linear discriminant.

We consider only forms of W which are diagonal. First we obtain the vector w which is the result of a linear classifier such as logistic regression or a linear support vector machine (we use logistic regression based on the favorable experimental results described earlier). The weight vector w describes a discriminator where the class prediction for x is obtained by computing wx and comparing it to a threshold w_0. Thus if two points x and y lie on the decision boundary of the discriminator, we have that

$$w^T(x - y) = 0, \tag{7}$$

i.e. the vector w is orthogonal to the decision boundary.

By taking the metric formed by the norm

$$d(x, y) = \|w^T(x - y)\| \tag{8}$$

we obtain a metric which measures distance along w, or between the class means (with the appropriate Gaussian assumptions). This can be interpreted as measuring the extent to which the linear discriminant prefers class 1 or class 2.

This can be regarded as an implicit form of dimensionality reduction, by realizing that values of w tending to zero will cause the metric to assign negligible weight in those directions, which in the limit is akin to removing the corresponding features.

In this manner, we use discriminant information to weight our metric rather than assume isotropy in the metric.

Sphering. Our diagonal restriction on W motivates the removal of correlation between the features in advance. Normalizing each feature so that they all have roughly the same scale is also important for kernel density estimation as noted earlier. For this reason we perform these operations (*sphering* the data) as the first step of our methodology using principal component analysis (PCA). We also take the opportunity at this stage to examine the resulting eigenspectrum and remove low-eigenvalue features.

Our overall dimension reduction scheme thus includes two kinds of steps: this PCA-based explicit feature removal, which aims to 'denoise' the data, and the implicit direction weighting performed by our metric learning procedure.

5 Experimental Results

Our dataset contains 26733 rows and 6348 attributes, and is sparse, containing 3732607 non-zero input values. It has 804 positive output values ("active" class).

A pre-analysis of the data, however, reveals that 2290 columns are empty. Furthermore, 388 out of 8235711 pairs of columns are identical. These are also removed. Among the remaining columns, a column reduction scheme also reveals linear dependencies. Removal of 406 columns from the remaining 3871 columns is performed. This leaves about half of the original dimensions. We then perform PCA, keeping only 100 of these dimensions.

All experiments were performed using 10-fold cross-validation, in which the data is broken into 10 equally-sized disjoint subsets, and testing (evaluation) is performed on one of them while training is performed on the other 9 put together.

The following table lists the results of the experimental evaluation of [10] performed on the same data. It shows the best performance yielded by each method, with and without the use of PCA projecting to 100 dimensions.

Method	AUC
k-nearest neighbors	0.862 ± 0.017
Bayes classifier	0.891 ± 0.012
Decision tree	0.893 ± 0.011
linear support vector machine	0.918 ± 0.010
RBF support vector machine	0.927 ± 0.013
Logistic regression	0.931 ± 0.012

The next table shows the results of the SLAMDUNK methods on this data.

Method	AUC
SLAMDUNK fixed isotropic kernel	0.933 ± 0.017
SLAMDUNK fixed isotropic kernel with metric learning	0.937 ± 0.012
SLAMDUNK variable nonisotropic kernel with metric learning	0.940 ± 0.012

6 Conclusion

We have presented a methodology called SLAMDUNK which we have designed to have favorable properties for the problem of virtual screening. We have demonstrated its favorable performance on a real pharmaceutical dataset as evidence that this line of thinking may hold promise for this important contemporary problem.

Additionally, this work represents a foray into the more general problem of high-dimensional discrimination, in particular exploring the extent to which probabilistic methods can be successful in high-dimensional problems. We plan to continue developing the seeds of the ideas which have been presented here.

Acknowledgements

The author would like to thank Andrew Moore for his enthuasiastic support of this work and for providing a first-class research environment, Paul Komarek and Ting Liu for the excellent software framework which facilitated this work, and our pharmaceutical collaborators for providing the data and their insights regarding this problem.

References

[1] L. Breiman, W. Meisel, and E. Purcell. Variable Kernel Estimates of Multivariate Densities. *Technometrics*, 19:135–144, 1977.

[2] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1973.

[3] J. H. Friedman. Flexible Metric Nearest Neighbor Classification. Technical report, Stanford University, 1994.

[4] J. H. Friedman, J. L. Bentley, and R. A. Finkel. An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Transactions on Mathematical Software*, 3(3):209–226, September 1977.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2nd ed.* Academic Press, 1990.

[6] G. H. Givens. Consistency of the Local Kernel Density Estimator. Technical report, Colorado State University, 1994.

[7] P. Grassberger and I. Procaccia. Measuring the Strangeness of Strange Attractors. *Physica D*, pages 189–208, 1983.

[8] A. G. Gray and A. W. Moore. Very Fast Multivariate Kernel Density Estimation via Computational Geometry. In *Joint Statistical Meeting 2003*, 2003. to be submitted to JASA.

[9] G. Harper, J. Bradshaw, J. C. Gittins, and D. V. S. Green. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.*, 41:1295–1300, 2001.

[10] P. Komarek and A. W. Moore. Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs. In *Workshop on AI and Statistics*, 2003.

[11] T. P. Minka. Distance Measures as Prior Probabilities. Technical report, Massachusetts Institute of Technology, 2000.

[12] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500), December 2000.

[13] D. W. Scott. *Multivariate Density Estimation.* Wiley, 1992.

[14] B. W. Silverman. *Density Estimation.* Chapman and Hall, New York, 1986.

[15] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[16] D. Wilton and P. Willett. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. *J. Chem. Inf. Comput. Sci.*, 43:469–474, 2003.