Intrinsic Dimensionality in Gene Expression Analysis

Harri Lähdesmäki[1,2], Olli Yli-Harja[2], Wei Zhang[1], Ilya Shmulevich[1]
[1]Cancer Genomics Laboratory, University of Texas M.D. Anderson Cancer Center,
[2]Institute of Signal Processing, Tampere University of Technology

***Abstract*— Gene expression based methods have become popular for classification and unsupervised class separation of cancer and other diseases. Due to the large number of genes relative to typical sample sizes in current experiments, the classification results depend crucially on the number of genes used in the classification. We propose to use a well-studied, unsupervised concept called intrinsic dimensionality for finding the size of the optimal gene subset for classification. In particular, the intrinsic dimensionality serves as a lower bound for the number of genes to be used in the classification. Furthermore, because the proposed method is completely unsupervised, it can be applied to cases where the correct class labels, say cancer subtypes, are not known. By applying our method to several publicly available gene expression data sets we show that the intrinsic dimensionality can be reliably used to predict the optimal number of genes for classification.**

I. BACKGROUND

Many data mining and pattern recognition tasks face the following problem with dimensionality: the number of features measured greatly exceeds the number of observations, thereby reducing the effectiveness of statistical methods. This problem is particularly evident in classification tasks arising from high-throughput gene expression studies. The large number of features (genes) hinders the classifier learning task by provoking classifier generalization (overfitting) problems, difficulties in error estimation of the designed classifier, and feature subset selection (FSS) problems [1]. In addition, many features may be completely uninformative for the classification task and should not be used. The most evident and often used method to alleviate these problems is to reduce the number of features in the classification. For instance, several different approaches to cancer classification based on high-dimensional feature (gene) vectors have been introduced [2], [3], [4], all of which use some dimensionality reduction method.

An alternative approach to the above mentioned problems is discussed here. Even though the true distribution of measurements is typically unknown, we may pose a question about its fundamental properties. Of particular interest is the intrinsic dimensionality (ID) of the observed random data. Fukunaga [5] defines the ID as "the minimum number of parameters required to account for the observed properties of the data". This is often referred to as the number of free parameters of the data generating process. Thus, the number of such free parameters would serve as a guide for determining how small the number of features should be [5]. Thus, if the intrinsic dimensionality of a data generating process is d then the n-dimensional measurement vector x can be modeled as $x = f(u)$, where $x \in \mathbb{R}^n = \mathcal{X}$, $u \in \mathbb{R}^d = \mathcal{U}$, and f is a possible stochastic mapping from $\mathcal{U}$ to $\mathcal{X}$ (the possible stochasticity can be considered as measurement noise or other source of variation). That is, the n random variables in x are intrinsically functions of d random variables. The usual case is that $n \gg d$.

From the point of view of classification, the essential (optimal) features for class prediction are, of course, the ones in vector u since they describe the intrinsic behavior of observations (disregarding the possible stochasticity in f). It is obvious that the ID corresponds to the upper bound for the optimal number of features in the $\mathcal{U}$ space. Because the mapping f is generally unknown, we are primarily concerned with classification in the $\mathcal{X}$ space. Further, consider a case where the optimal decision boundary between two classes forms a "complex" manifold in the feature space. In other words, most of the elements of u are needed to form a good classifier. Assuming f preserves the complexity of the optimal decision boundary, the intrinsic dimensionality d should serve as a lower bound for the optimal number of features in the $\mathcal{X}$ space. This constitutes a worst case scenario since not all intrinsic features are necessarily needed for classification. Thus, the ID has the potential to be of great help in classification, unsupervised class separation and FSS: no matter what type of classifier is used, the number of features should be no less than the ID.

Let us illustrate the ID with an example. Consider random points on the surface of a unit sphere. Since the 3-dimensional Cartesian coordinates $x_1 = \cos\phi\sin\theta$, $x_2 = \sin\phi\sin\theta$, $x_3 = \cos\theta$, can be expressed in terms of only two parameters the intrinsic dimensionality of such a data set should be equal to 2. The three Cartesian coordinates are further embedded in 3000 dimensional feature space where the added irrelevant features contain Gaussian noise with a small standard deviation. Using the ID estimation method discussed in Methods section on these very high-dimensional vectors, we obtain an ID estimate of 2.08. It is interesting to note, however, that standard methods, such as principal components analysis (PCA), will fail to capture the intrinsic dimensionality and will produce an estimate of the embedding dimensionality, which is equal to 3. Additionally, PCA can only capture linear behavior, whereas ID copes with non-linearities, too.

II. METHODS

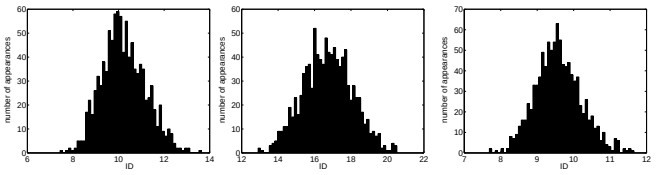
Here, we introduce computational methods for the estimation of intrinsic dimensionality. As discussed above, we are primarily interested in the size of the optimal feature subset. In this paper, we concentrate only on local ID estimation methods, the global ones being related to multidimensional scaling and its variants.

One of the most effective ID estimation methods is due to Pettis *et al*., see e.g. [5], [8]. This local method, called the *near neighbor* algorithm, estimates the ID directly from the local configuration of the measurement points instead of generating any lower dimensional projections. Let $d_k(X)$ denote the distance from a point X to its kth nearest neighbor among the measurements $S = (x^{(1)}, \ldots, x^{(N)})$. The intrinsic dimensionality d can then be expressed as [5], [8]

$$d \cong \frac{\mathbb{E}_S[d_k(X)]}{(\mathbb{E}_S[d_{k+1}(X)] - \mathbb{E}_S[d_k(X)])\,k} \quad (1)$$

$$\cong \frac{\mathbb{E}_X\mathbb{E}_S[d_k(X)]}{(\mathbb{E}_X\mathbb{E}_S[d_{k+1}(X)] - \mathbb{E}_X\mathbb{E}_S[d_k(X)])\,k} \quad (2)$$

where $\mathbb{E}_S$ (resp. $\mathbb{E}_X$) is the expectation relative to random sample (resp. (random) point X). Note that Eq. (1) provides an estimator of local property, around X, but Eq. (2) takes the average of local properties over $\mathcal{X}$. However, Eq. (2) still provides an estimate of local property. Since the distributions are usually unknown and



(a) (b) (c)

Fig. 1. Intrinsic dimensionality estimation for all of the three data sets: (a) colon, (b) ALL/AML, and (c) SRBCT. Each graph shows a histogram of the ID estimates over 1000 bootstrap iterations. ID estimates for different number of neighbors are combined by computing the median.

only one sample is available, Eq. (2) is applied such that the expectations $\mathbb{E}_X\mathbb{E}_S[d_k(X)]$ are replaced by sample averages $\overline{r}_k = \sum_{i=1}^{N} d_k(x^{(i)})$, where $d(\cdot)$ is the Euclidean distance. This results in [9]

$$d = \frac{\overline{r}_k}{(\overline{r}_{k+1} - \overline{r}_k)\, k}. \tag{3}$$

Eq. (3) implicitly assumes the ID to be constant in the whole measurement space. This assumption is often necessary in practise and will be adopted here. Although the estimator in Eq. (3) is fairly insensitive to the selection of k, the problem of selecting a value for the parameter k can be overcome by computing the median of the ID estimates for $k = 1, \ldots, 30$. In order to assess the statistical variability of the ID estimator, we used the standard bootstrap method to get confidence bounds for the estimates.When using the bootstrap we computed the median of the ID estimates for $k = 5, \ldots, 30$. [1]

III. RESULTS & CONCLUSIONS

In order to test the above mentioned concepts, the ID estimation method introduced in Eq. (3), with estimates for different values of k being combined by the median, was applied to three popular data sets [2], [3], [4] which are abbreviated in the following as colon, ALL/AML and SRBCT, respectively. The variability of the estimates was assessed by the standard bootstrap with 1000 iterations. The histogram of the ID estimates, as obtained by the bootstrap, for colon, ALL/AML and SRBCT data sets are shown in Figs. 1 (a), (b) and (c), respectively.

Recently, colon and ALL/AML data sets have been studied extensively in [6]. Using a novel pair-wise gene selection method, the cardinality of the optimal feature subset was found to be as small as about 15–20 (colon) and 20–30 (ALL/AML). The ID estimates are slightly smaller than the optimal number of features for classification: 10.8 for colon data set and 18.2 for AML/ALL data set (see also Figs. 1 (a) and (b)). These findings are in good agreement with the above discussion about the ID being a lower bound for the optimal number of features in the $\mathcal{X}$ space.

The SRBCT data set from [4] has also been studied in several papers. The original work by Khan *et al*. used the 10 most significant PCA components of the 96 highest scoring genes. A recent paper introduced a novel method for finding the most promising feature subset for classification [7]. The optimal feature subset for the SRBCT data set was found to have cardinality "less than 15" [7]. The ID estimate for the SRBCT data set is 10.1, with the bootstrap results shown in Fig. 1 (c). The ID estimate is again in agreement with the previously published results.

[1]We do not use $k = 1, \ldots, 4$ to estimate the confidence bounds due to the fundamental principle of the bootstrap method, which resamples the data with replacement. When a sample appears twice among the bootstrap samples, the distance between them becomes zero and this introduces a notable low-bias for the bootstrap estimates. However, for the sample-based ID estimator, we use all values of $k = 1, \ldots, 30$.

The feature subset selection is particularly challenging for microarray data. Most often the problem is attributed to the large number of features, small number of samples, and a considerable amount of noise present in the measurements. Also, a major part of the genes are usually uninformative for the class separation and some genes are redundant. Therefore, several different small feature subsets can be found to provide good classification performance. A preferred feature subset contains no uninformative or redundant genes. We propose the use of intrinsic dimensionality to guide the selection of the gene subset. The redundant features should have no effect on the ID since the ID measures the dimensionality of the (local) data generating process. A simple illustration can be given in the framework of the sphere example discussed earlier. For example, the ID estimate remains virtually the same even if any ten features are repeated ten times.

In biological classification problems, it is not uncommon to have uncertainty in the class labels. For instance, because of tissue heterogeneity, samples may represent mixed or even unknown tumor types, thus precluding confident assignment of class labels to these samples. Since the proposed method is completely unsupervised, its use is not limited to standard classification tasks and can also be used in cases where the class information is not available. The method is also computationally efficient to implement.

From another point of view, one could also say that the ID can support/contradict the number of features obtained by another FSS algorithm. For example, if the estimated ID is considerably lower than the number of features deemed optimal by some FSS algorithm, this may be cause for concern and may indicate that the FSS algorithm gratuitously selected an excessive number of features. Therefore, we also propose that the ID be used as a confirmation of other, supervised and unsupervised, methods. Topics of future studies will include a development of a methodology where the actual features are selected by the ID and a supervised ID estimation method, i.e., an ID estimation method which utilizes the class information.

IV. ACKNOWLEDGEMENTS

This study was partially supported by Tampere Graduate School in Information Science and Engineering (TISE) (HL), Academy of Finland (HL,OY-H), the Tobacco Settlement fund (IS,WZ), and the Kadoorie Foundation (IS,WZ).

REFERENCES

[1] E.R. Dougherty, "Small sample issues for microarray-based classification," *Comp. Funct. Genom.*, Vol. 2, pp. 28–34, 2001.

[2] U. Alon, *et al*., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA*, Vol. 96, pp. 6745–6750, 1999.

[3] T.R. Golub, *et al*., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, Vol. 286, pp. 531–537, 1999.

[4] J. Khan, *et al*., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat. Med.*, Vol. 7, pp. 673–679, 2001.

[5] K. Fukunaga, *Statistical pattern recognition.* San Diego: Academic Press, 1990.

[6] T. Bø, *et al*., "New feature subset selection procedures for classification of expression profiles," *Genome Biol.*, Vol. 3, pp. 1–11, 2002.

[7] J.M. Deutsch, "Evolutionary algorithms for finding optimal gene sets in microarray prediction," *Bioinformatics*, Vol. 19, pp. 45–52, 2003.

[8] K. Pettis, *et al*., "An intrinsic dimensionality estimator from near-neighbor information," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 1, pp. 25–37, 1979.

[9] P.J. Verveer, and R.P.W. Duin, "An evaluation of intrinsic dimensionality estimators," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 17, pp. 81–86, 1995.