Imbalanced Clustering of Microarray Time-Series

Ronald K. Pearson PEARSON@MAIL.DBI.TJU.EDU
Gregory E. Gonye GGONYE@MAIL.DBI.TJU.EDU
James S. Schwaber JAMES.SCHWABER@MAIL.TJU.EDU
Daniel Baugh Institute for Functional Genomics and Computational Biology, Department of Pathology, Anatomy, and Cell Biology, Thomas Jefferson University, 1020 Locust St., Philadelphia, PA 19107-6799, USA

Abstract

In their survey on learning from imbalanced datasets, Japkowitz and Stephen (2002) consider the influence of four key variables: degree of class imbalance, complexity of the learning task, size of the training set, and the learning algorithm used. This paper specializes these variables to the problem of clustering short, irregularly sampled time-series, motivated by the practical problem of analyzing cDNA microarray time-series. These time-series usually arise from experiments that attempt to characterize changes in gene expression following the application of an external stimulus. If, as is often the case, this stimulus only causes significant expression changes in a small subset of genes, the resulting clustering problem is highly imbalanced.

1. Introduction

This paper analyzes the influence of class imbalance on the cluster analysis of short, irregularly-sampled time-series like those currently being generated in our laboratory. In particular, we are using cDNA microarrays to study the dynamics of cellular gene expression changes in response to the neuropeptide angiotensin II activating its cognate AT_1 receptor. Cluster analysis of these time-series is expected to yield useful insights into the *functional* associations between genes, but the resulting cluster size distribution is expected to be highly imbalanced, with small groups of genes exhibiting certain characteristic responses and the majority of genes exhibiting either no response or certain generic responses. As a preliminary step toward the design of improved experimental protocols and data analysis procedures, we specialize the analysis of the four key issues considered by Japkowitz and Stephen (2002) in their recent survey on learning from imbalanced datasets: degree of class imbalance, complexity of the learning task, size of the training set, and the learning algorithm used. Our ultimate objective is to identify which of these factors are most important in obtaining reliable clustering results to provide a basis for enhanced biological understanding. The specific aims of this paper are more modest: first, to specialize the factors considered by Japkowitz and Stephen to the clustering problems we are attempting to solve and second, to compare the influence of these factors on the basis of a simple simulation-based example that captures the essential characteristics of these microarray time-series clustering problems. Later publications will apply the results of this paper to the analysis of real microarray time-series datasets.

1.1. Microarray data analysis

One of the most significant changes in biology in the last century has been the development of high-throughput analytical methods like cDNA microarrays (Schena, 1995), which permit the simultaneous measurement of expression levels essentially across an entire genome (i.e., in $\sim 10^4$ genes). A typical microarray dataset contains results for multiple microarrays, obtained with varying degrees of biological and methodological replication (e.g., duplicate slides prepared from the same tissue sample, slides prepared from different tissue samples of the same type, etc.). A typical example is that of Pritchard *et al.* (2001), who consider a collection of 72 microarrays, obtained from three different tissue types for each of six different mice. Microarrays were prepared from each of these 18 tissue samples using two-fold replication with two method variations (specifically, two different dye labellings). Analysis of the resulting datasets involves a wide variety of issues, including data prefiltering to detect and eliminate various types of data anomalies,

Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003.

normalization to correct for slide-to-slide differences, and various types of summary analysis that attempt to either classify genes into groups on the basis of behavior differences across the set of microarrays (e.g., normally variable vs. highly variable genes in the analysis of Pritchard *et al.* (2001)), or classify tissue types on the basis of differences in the expression patterns of a subgroup of the genes (Golub *et al.*, 1999).

Our own work is closely related to that of Tavazoie *et al.* (1999), who obtained significant biological insights by applying the popular k-means clustering procedure to the yeast microarray time-series discussed by Cho *et al.* (1998). There, microarrays were used to monitor the expression changes in approximately $6,000$ yeast genes, generating a time-series of length $L = 15$ that was sampled approximately uniformly in time. Two important practical differences between the yeast time-series considered by Tavazoie *et al.* and our time series are first, that our time-series are shorter (6 samples vs. 15) and second, that our sampling times are highly non-uniform (i.e., $t = 0$, 5, 15, 30, 60, and 240 minutes). Unlike typical engineering (Kay and Marple, 1981) and statistical (Brockwell and Davis, 1991) time-series analysis problems where uniform sampling is the norm, the approximately exponential sampling scheme considered here is quite common in biology, arising from a dose-response view: a key objective is to determine the time required for the biological system to exhibit a significant response. Further, because microarray generation involves significant manual effort, time-series of length $L \sim 10^2$ or 10^3, typical in many engineering applications, are not currently feasible.

1.2. Cluster analysis

Many different clustering methods have been proposed and examined to varying degrees, both theoretically and experimentally, and space does not permit a detailed survey here. In this paper, we are concerned exclusively with *unsupervised* clustering procedures, which take a dataset and generate one of two results: *partitioning methods* partition a dataset summarizing N objects into k mutually exclusive subsets, while *hierarchical methods* generate a hierarchy of such partitionings, ranging from a finest partitioning where each object represents its own cluster, to a coarsest partitioning that contains all objects. Because it is best suited to the biological questions of primary interest to us, we restrict consideration here to partitioning methods. This class includes both the extremely popular k-means procedure considered by Tavazoie *et al.* (1999) and others in the analysis of gene expression data, and the *Partitioning Around Medoids (PAM)* algorithm described in detail by Kaufman and Rousseeuw (1990).

In practice, cluster analysis involves more than the choice of a general method (hierarchical vs. partitioning) and a specific computational algorithm. This point is nicely illustrated for the PAM algorithm, which generates a partitioning of a dataset based on a computed matrix $\mathbf{D}$ of *dissimilarities* D_{ij} between every pair of objects, i and j in the dataset. Like all partitioning methods, this algorithm requres the number of clusters k to be specified as an input parameter. Given k, the PAM algorithm finds a set $\{x_i^*\}$ of k *medoids* or *representative objects* from the dataset and a collection of k sets S_i such that $x_i^* \in S_i$ and the following aggregate dissimilarity measure is minimized:

$$J = \sum_{i=1}^{k} \sum_{j \in S_i} D_{ij}^*, \tag{1}$$

where D_{ij}^* is the dissimilarity between the representative object x_i^* and the object $x_j \in S_i$. A detailed description of this algorithm is given in Chapter 2 of the book by Kaufman and Rousseeuw (1990).

Since the choice of clustering method represents one of the four key variables considered by Japkowitz and Stephen, it is important to consider the influence of this choice. However, because so many partitioning methods have been proposed (see, for example, the discussions in the books by Kaufman and Rousseeuw (1990) and Gordon (1999)), it is important to restrict the range of algorithms considered, especially in light of the ranges of the other three factors to be considered here.

Initially, we restrict consideration to the PAM algorithm, but with different choices of dissimilarity measures D_{ij} between objects. We chose this algorithm over the better known k-means procedure in part because it addresses some of the known limitations of k-means (e.g., the dependence of the results on the order in which the objects appear in the dataset), discussed further by Kaufman and Rousseeuw (1990, p. 114), and in part because it is inherently more flexible, permitting the use of arbitrary dissimilarity measures.

2. Case Study Structure

To compare the influence of the four factors just described, we use the Generalized Sensitivity Analysis (GSA) framework proposed recently for comparing data analysis results (Pearson, 2003). This framework consists of the following five steps:

1. Define a collection $\{\Sigma_\ell\}$ of *scenarios* to be compared (here, different degrees of class imbalance, task complexity, and dataset size).

2. For all scenarios, specify a common *sampling scheme* that generates a collection of datasets $\{S_j\}$, each of which are expected to yield "equivalent" results.

3. Select a common, real-valued *descriptor* that will be used to compare the results obtained across the different scenarios.

4. Select a collection of methods to be compared, all of which are compatible with the descriptor chosen in Step 3.

5. To compare the results generated in Steps 1 through 4, construct boxplots summarizing the variation in the descriptor values defined in Step 3 that are seen across the datasets generated in Step 2, for each scenario defined in Step 1 and each method defined in Step 4.

The GSA framework provides a systematic basis for managing comparisons between different problem characteristics, and it is particularly useful in cases like the one considered here where the number of interesting comparisons is too large to explore exhaustively, a point discussed further in Sec. 3.

The methods compared here correspond to the PAM algorithm described in Sec. 1.2 with the two most popular dissimilarity measures: Euclidean distance between feature vectors, and the product-moment correlation coefficient between feature vectors (Gordon, 1999; Kaufman and Rousseeuw 1990). The other three of the four factors considered by Japkowitz and Stephen correspond to the three scenarios considered here, described in detail in Secs. 3 through 5. In all cases, a dataset is generated that consists of time-series from three simulation-based classes: N_A series from class A, N_B from class B, and N_C from class C. As in the survey of Japkowicz and Stephan (2002), the total dataset size $N = N_A + N_B + N_C$ is fixed, allowing us to separate imbalance effects from size effects. All time-series $\{y_k\}$ in each class are of fixed length L and are of the form:

$$y_k = a x_k + \epsilon_k, \quad k = 1, 2, \ldots, L, \tag{2}$$

where a is a positive random amplitude variable, $\{\epsilon_k\}$ is an independent, identically distributed (i.i.d.) random noise sequence, statistically independent from a, and $\{x_k\}$ is a class-specific deterministic sequence that defines the noise-free response. The random amplitude a is uniformly distributed on the interval $[1-\lambda, 1+\lambda]$ for a fixed value of λ, and each noise sample ϵ_k is normally distributed with mean zero and standard deviation σ. The parameters λ and σ define the variation between individual members of each class, since taking $\lambda = 0$ and $\sigma = 0$ reduces all objects in the class to the deterministic sequence $\{x_k\}$. The deterministic class templates considered here correspond to the solutions of three simple dynamic models:

A: exponential decay, $x_k = \alpha^{t_k}$,

B: hyperbolic decay, $x_k = 1/(1 + \beta t_k)$,

C: decaying oscillation, $x_k = \alpha^{t_k} \cos \omega t_k$.

Specifically, Sequence A comes from a simple first-order, linear ordinary differential equation model, Sequence B comes from an inherently more complex first-order *partial* differential equation model, and Sequence C comes from a second-order linear ordinary differential equation model. Here α, β and ω are positive constants and the sequence $\{t_k\}$ corresponds to one of the following three sampling patterns:

U: uniform sampling, $t_k = (k-1)T/(L-1)$,

E: exponential sampling, $t_k = [T+1]^{(k-1)/(L-1)} - 1$,

R: random sampling: t_k is an increasing sequence of random times, uniformly distributed on $[0, T]$.

Note that the deterministic sequences U and E both satisfy the condition $0 = t_1 < t_2 < \cdots < t_L = T$.

3. Learning task complexity

The learning task complexity is essentially determined by the difficulty of distinguishing objects from different classes. For the simulation case studies considered here, this difficulty is determined by:

a. the length of the individual time-series: L,

b. the noise-free response characteristics: α, β, ω,

c. the observation noise characteristics: λ, σ,

d. the sampling pattern: class U/E/R, duration T.

Even if we only consider three values for each of the eight variables listed here, the total number of combinations is $3^8 = 6,561$, illustrating the point noted earlier that exhaustive comparison of interesting scenarios is not feasible here. To overcome this difficulty, we adopt the following strategy:

a. To keep the series length L appropriate to the biological problems motivating our investigation, we consider $L = 6$, corresponding to the length

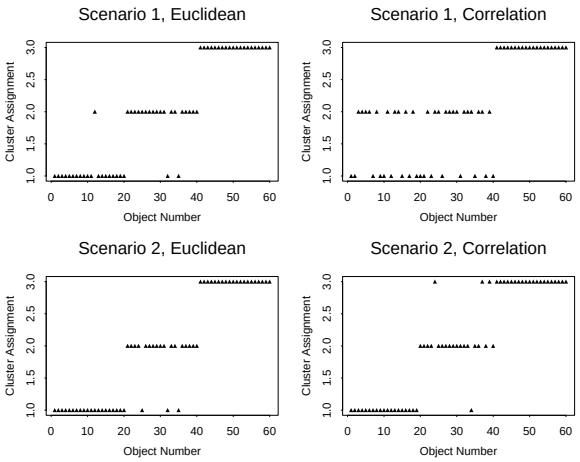

Figure 1. PAM clustering results for Scenarios 1 and 2, Euclidean vs. correlation dissimilarities

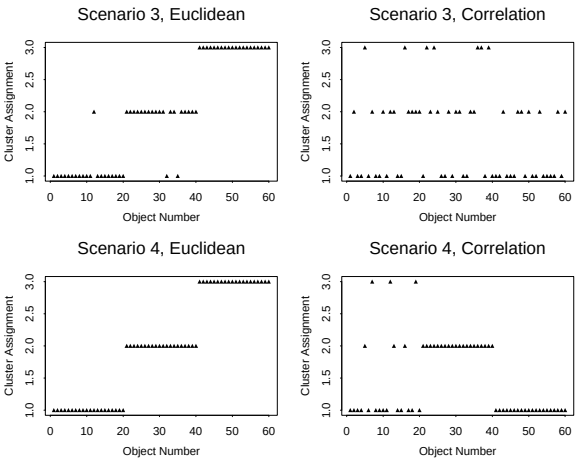

Figure 2. PAM clustering results for Scenarios 3 and 4, Euclidean vs. correlation dissimilarities

of the time-series generated in our laboratory and $L = 15$, corresponding to the length of the yeast time-series considered by Cho *et al.* (1998).

b. We consider four noise-free responses:

   b1: $\alpha_1 = 0.8$, $\beta_1 = 0.5$, $\omega_1 = 30$,
   b2: $\alpha_2 = 0.2$, $\beta_2 = 15.0$, $\omega_2 = 30$,
   b3: $\alpha_3 = 0.8$, $\beta_3 = 0.5$, $\omega_3 = 1$,
   b4: $\alpha_4 = 0.8$, $\beta_4 = 5.0$, $\omega_4 = 1$.

c. To compare both scaling and noise effects, we consider the following combinations of λ and σ:

   c1: low noise, no scaling: $\lambda = 0$, $\sigma = 0.1$,
   c2: moderate noise, no scaling: $\lambda = 0$, $\sigma = 0.2$,
   c3: high noise, no scaling: $\lambda = 0$, $\sigma = 0.4$,
   c4: low noise, random scaling: $\lambda = 0.2$, $\sigma = 0.1$,
   c5: moderate noise, random scaling: $\lambda = 0.2$, $\sigma = 0.2$,
   c6: high noise, random scaling: $\lambda = 0.2$, $\sigma = 0.4$.

d. The fixed duration $T = 2.5$ was chosen so that the noise-free sequences $\{x_k\}$ are neither essentially constant, as would be the case if T were chosen too small, nor essentially zero for all $k > 1$, as would be the case if T were chosen too large.

Even under these restrictions, the number of combined scenarios to consider would be 144, far too many to attempt to summarize here. Hence, we further simplify the problem by first comparing the 24 possible combinations of the four noise-free sequences b1 through b4 with the six scaling and noise scenarios c1 through c6.

Fig. 1 compares the results obtained by the two clustering methods considered here (i.e., Euclidean vs. correlation dissimilarities) for Scenarios 1 and 2, corresponding to the noise-free responses b1 and b2 listed above. Here, we consider the case of balanced clusters ($N_A = N_B = N_C = 20$) and uniform sampling (U), so the differences between the objects in each cluster are determined both by random scaling (with $\lambda = 0.1$) and additive noise (with $\sigma = 0.1$). The results shown in Fig. 1 are the cluster assignments made for each time-series, plotted against its index. The correct classification in this case corresponds to the assignment of the first 20 objects to cluster number 1 (Cluster A), the next 20 objects to cluster number 2 (Cluster B), and the last 20 objects to cluster number 3 (Cluster C). The results obtained for both Scenarios 1 and 2 using Euclidean dissimilarities are shown in the left-hand two plots in Fig. 1 and it is clear from these plots that three objects of the 60 are misclassified. The corresponding results obtained for these scenarios using correlation-based dissimilarities are shown in the right-hand plots. For Scenario 2, this result is only slightly worse than the corresponding Euclidean result, but for Scenario 1, the correlation-based results are dramatically worse: 17 misclassifications *vs.* 3.

The corresponding results for Scenarios 3 and 4, representing the noise-free responses b3 and b4 listed above with the same noise and random scaling as before, are shown in Fig. 2. As in Fig. 1, it is clear that the Euclidean results are substantially better than those obtained using correlation-based dissimilarities. This result is particularly pronounced for Scenario 4, where the Euclidean dissimilarity gives perfect classification, whereas the correlation-based dissimilarity results exhibit 17 misclassifications. This last result also illustrates an important practical issue that arises in assessing misclassification rates (Pearson *et al.*, 2003):

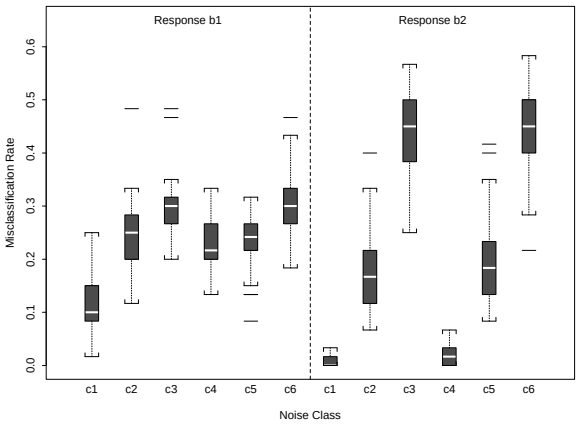

Figure 3. Influence of noise-free responses, random scaling, and noise amplitude on misclassification rates

the clustering procedure has no knowledge of the cluster identities, but can only assign objects to arbitrarily ordered clusters. That is, the mapping of the empirically identified clusters 1, 2, and 3 into the data generating clusters A, B, and C requires knowledge of these cluster characteristics that is not available to the clustering algorithm. Hence, this assignment must be made after the fact, and it can be made in different ways that generally give different results. Since we are ultimately interested in comparing misclassification results for different scenarios, the results presented here are based on the pairing of the triples (A, B, C) and $(1, 2, 3)$ that give the lowest misclassification rate. For example, note that if we choose the pairing $A \to 1$, $B \to 2$, and $C \to 3$ for Scenario 4 under the Euclidean dissimilarity measure, we obtain perfect classification, but if we adopt this choice for the correlation-based results, we obtain 26 misclassifications (6 for Cluster A and all 20 for Cluster C). In contrast, the assignment $A \to 3$, $B \to 2$, and $C \to 1$ yields only 17 misclassifications for this case.

These results demonstrate that the choice of clustering method is profoundly influential, even for perfectly balanced clusters. Because the results obtained for this preliminary test case are so much better for Euclidean dissimilarities than for correlation-based dissimilarities, we restrict consideration to the Euclidean case for the remainder of this paper.

Fig. 3 summarizes the misclassification rates for 12 scenarios, corresponding to two different choices of noise-free response (b1 vs. b2) and the six noise effect classes, c1 through c6. More specifically, each boxplot in this figure summarizes the misclassification rates for 50 statistically independent repetitions of the clustering problem considered here. The white line at the center of each boxplot corresponds to the median misclassification rate for the 50 clusterings, the top and bottom of the black box represent the upper and lower quartiles, respectively, and the whiskers correspond to the most extreme non-outlying data values. Outliers are defined as points lying further than 1.5 times the interquartile range from the median, and are marked with separate horizontal lines in the plot.

Several conclusions are clear from a comparison of these boxplots. First, as expected, the learning task becomes more difficult with increasing object variability: misclassification rates increase strongly with increasing σ (noise classes c1 vs. c2 vs. c3 and c4 vs. c5 vs. c6) and somewhat less strongly with increasing λ (c1/c2/c3 vs. c4/c5/c6). These differences also depend on the noise-free response characteristics, b1 vs. b2 shown here, and qualitatively similar dependences are obtained for noise-free responses b3 and b4, not shown. For example, the effect of increasing λ is negligible relative to the effect of increasing σ for response class b2, as may be seen by comparing the c1/c2/c3 results with the c4/c5/c6 results. In contrast, for response class b1, the misclassification rates are significantly higher for noise class c4 than for noise class c1 ($\lambda = 0.2$ vs. $\lambda = 0$, both with $\sigma = 0.1$). Also, the rate at which the number of misclassifications increases with σ is a very strong function of the response class. This point may be seen most dramatically when comparing noise scenarios c1, c2, and c3 in Fig. 3: response class b2 is consistently better than response class b1 when $\sigma = 0.1$ (noise class c1), but consistently worse when $\sigma = 0.4$ (noise class c3).

To compare the influence of sequence length and sampling pattern, Fig. 4 summarizes the results obtained for two of the 12 scenarios compared in Fig. 3 as these factors are varied. Specifically, Scenario A corresponds to the response and noise class combination b1/c1, and Scenario B corresponds to the combination b1/c3. As noted earlier, the sequence lengths considered here are $L = 6$ and $L = 15$, and the sampling patterns considered are uniform (U), exponential (E), and random (R). Note that in the boxplots shown in Fig. 4, the *same* random sampling time sequence $\{t_k\}$ is used in all simulations to facillitate comparisons: the variability seen in each boxplot is entirely due to differences between the 50 statistically independent realizations defined by the noise class, c1 or c3. The results presented in Fig. 4 give some evidence that the clustering problem is more difficult for shorter sequences, but these differences are negligible at the higher noise level (Scenario B). These results also suggest that the sampling pattern—uniform vs. exponential vs. random—

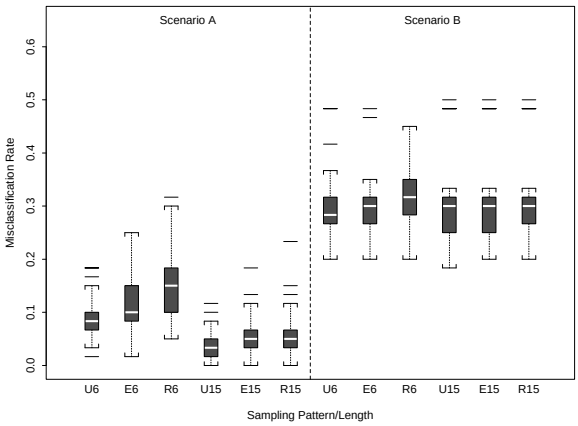

Figure 4. Influence of sequence length and sampling pattern on misclassification rates

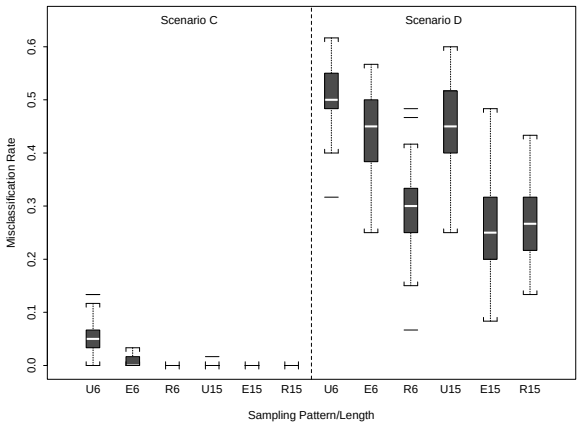

Figure 5. Influence of sequence length and sampling pattern on misclassification rates

has a significant influence at low noise levels, particularly for shorter sequences. In particular, note that for Scenario A, uniform sampling is clearly the best strategy, followed by exponential and then random sampling, both for $L = 6$ and $L = 15$, but the sampling pattern has essentially no effect on the results for Scenario B.

Fig. 5 presents analogous results for two additional scenarios: Scenario C corresponds to the combination b2/c1 and Scenario D corresponds to the combination b2/c3. As before, it is clear that noise effects are dominant: results for the high-noise Scenario D are uniformly worse than those for low-noise Scenario C. For the low-noise case, the effects of sequence length L are much more pronounced than they were for Scenario A; in particular, increasing the sequence length from $L = 6$ to $L = 15$ in Scenario C yields a much more dramatic improvement than in Scenario A, reflecting an inherent difference in our ability to separate the b1 patterns vs. the b2 patterns. Finally, note that in marked contrast to Scenario A, the different sampling schemes have precisely the opposite ordering for both Scenarios C and D: random sampling appears best, followed by exponential with uniform sampling poorest. Again, note that sampling pattern effects appear more significant for shorter sequences than for longer ones.

4. Class imbalance effects

For fixed sample size N, the membership of each individual cluster can vary between 0 and N, with $N_A = N_B = N_C = N/3$ for the perfectly balanced case. The number of possible partitionings into three classes is $(N^2 + 3N + 2)/2$, meaning that there are $1,891$ possible cases for the case $N = 60$ considered here. Again, exhaustive search is impractical so we initially examine the 10 special cases listed in Table 1. As before, since we are principally interested in the case of very short sequences ($L = 6$) and exponential sampling patterns (E), we restrict consideration to these cases and consider the consequences of the ten different class imbalance patterns listed in Table 1 on the clustering of time-series generated under Scenarios C and D from the previous example.

Class	N_A	N_B	N_C	Class	N_A	N_B	N_C
1	20	20	20	6	10	10	40
2	20	10	30	7	5	35	20
3	20	5	35	8	5	20	35
4	10	20	30	9	5	10	45
5	10	30	20	10	5	5	50

Table 1. Imbalance classes considered here

Fig. 6 gives a boxplot summary of the misclassification results obtained under each of these different imbalance patterns, numbered as in Table 1, for Scenarios C and D. It is clear from this plot that class imbalance effects are extremely significant, comparable to the noise effects that distinguish these two scenarios ($\sigma = 0.1$ for Scenario C and $\sigma = 0.4$ for Scenario D). For Scenario C, generally good results are obtained for most imbalance classes. In particular, only Classes 3, 9, and 10 give consistently poor results, while Classes 6, 7, and 8 occasionally give poor results, as indicated by the outlying cases with high misclassification rates. Classes 1, 2, 4, and 5 exhibit the best results, consistent with the

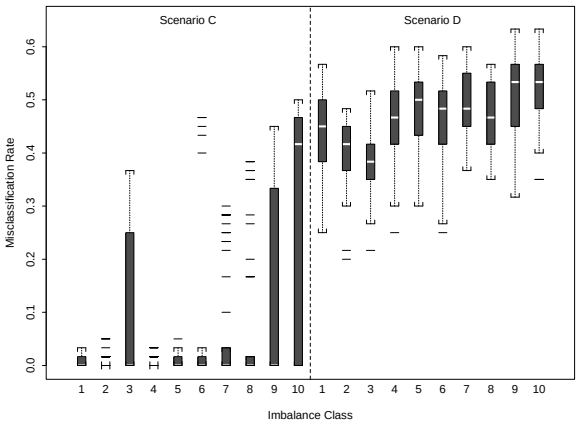

Figure 6. Effects of class imbalance on Scenarios C and D

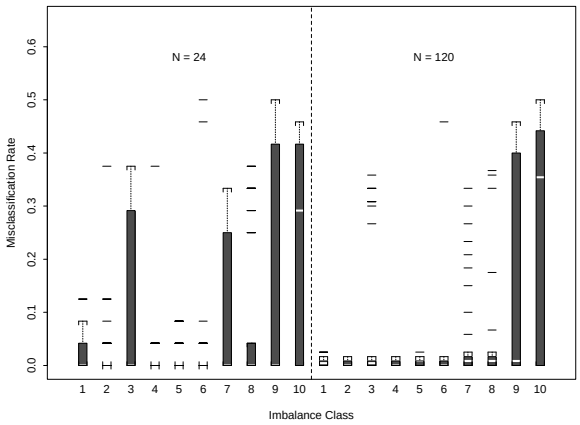

Figure 7. Effects of dataset size on Scenario C

fact that these cases also exhibit the lowest degree of class imbalance of the ten cases considered here. Further, note that the five cases with very small classes (Cases 3 and 7–10) all exhibit either consistently poor misclassification rates or outlying examples with high misclassification rates. As before, differences are much less pronounced for high-noise Scenario D, where results are uniformly poorer.

5. Size effects

The last problem characteristic considered here is the size of the dataset, N. Since it gives the most consistent results, we examine this effect for Scenario C for the three sizes $N = 24$, $N = 60$, and $N = 120$, corresponding to scalings of the previously considered problem by factors of 2/5, 1, and 2. These scalings were chosen because the larger and smaller datasets are scaled by roughly the same factor, while keeping integer values for cluster sizes N_A, N_B, and N_C in all of the ten imbalance classes defined in Table 1. These effects are shown in Fig. 7, which should be compared with the left-hand side of Fig. 6, which has the same format and scaling. Careful comparison of these plots generally reveals little change in the median misclassification rates with total sample size, but the variability of these results does decrease with increasing N, as indicated both by the width of the solid portion of the boxplots and the indicated outliers. Exceptions are Classes 3 and 7, where the misclassification rates seem to improve dramatically as N increases from 24 to 60 to 120. In contrast, note that imbalance classes 9 and 10 are consistently the hardest to cluster correctly in this example, giving results that are almost entirely independent of sample size, at least over the range considered here.

6. Summary

Motivation for this study was to provide the necessary background information to develop a systematic procedure for the analysis of microarray time-series that are being generated in our laboratory. Consequently, we have attempted to incorporate characteristic features of this problem to determine their influence on the results we obtain. To accomplish this, we have specialized the four characteristics of the imbalanced learning problem considered by Japkowicz and Stephan (2002) as follows. First, the degree of class imbalance is determined by the biological aspects of the problem; second, the nature of the time-series to be clustered is dictated by the length L, the noise-free response, the magnitude of the noise or other sources of variability, and the sampling pattern; third, the overall problem size is the number of gene responses to be clustered; and fourth, the learning method, is the clustering algorithm used in analyzing the time-series data.

A key practical question is whether it is possible to extract useful information from time-series as short as $L = 6$, given that classical results in engineering and statistical time-series analysis (e.g., spectrum estimation) are almost entirely useless in such cases. Both our previous results (Pearson *et al.*, 2003) and the results presented here suggest that analysis of these time-series can yield useful results. The comparisons presented here with time-series of length $L = 15$, a choice motivated by the results of Tavazoie *et al.* (1999) and Cho *et al.* (1998), demonstrate that increasing the length of the time-series can make the analysis problem easier, but that this effect is not dominant. Similar conclusions apply to the total sample size, N, which appears to be the least influential of all of the factors considered here.

The most significant factors appear to be the noise characteristics, the class imbalance pattern, and the clustering algorithm used. Beyond showing that it was a highly influential factor, we did not investigate the influence of the clustering method in detail here, primarily because the choice of clustering method is entirely under our control in analyzing the data and is not directly influenced by experimental considerations, in contrast to factors like the time-series length L. The results presented here show, not surprisingly, that noise effects ultimately become dominant, essentially eliminating our ability to cluster the time-series. A less obvious insight is the degree to which this noise-induced degradation depends on both the character of the noise-free response and the class imbalance pattern. It is also worth emphasizing that the class imbalance pattern appears to be a more significant influence than the fact that highly imbalanced problems exhibit some very small classes. In the microarray time-series analysis problems of interest to us, these results suggest the exploration of preprocessing procedures that improve the imbalance pattern by removing candidate genes from consideration that are not expected to be biologically interesting.

Finally, a particularly interesting observation was the influence of sampling pattern on the time-series clustering results considered here. As noted, engineering and statistical time-series tend to be uniformly sampled in time, whereas biological sampling patterns tend to be highly irregular, often approximately exponential, motivated by a dose-response view of response times. The results presented here show that sampling pattern differences can have a significant influence on clustering results, but that these differences are strongly case-specific and thus impossible to predict *a priori*. Also, it was seen that, at least in the examples considered here, these differences are more important for shorter time-series like those of primary interest to us than for longer time-series like those typical of engineering and statistical applications.

Acknowledgements

The authors wish to acknowledge first, their colleagues at the Daniel Baugh Institute for their interest and support in this activity and second, informative discussions with Prof. Javier Garcia-Frias and Ms. Yujing Zeng with the Department of Electrical and Computer Engineering at the University of Delaware. Funding support for this work was provided by NIH/NIGMS project MH64459-01, NIH/NIAAA project AA13203-01, and DARPA project F30602-01-0578.

References

Brockwell, P. and Davis, R., (1991). *Time Series: Theory and Methods.*. Springer-Verlag.

Cho, R., Campbell, M., Winzeler, E. Steinmetz, L., Conway, A., Wodica, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., and Davis, R. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, *2*, 65–73.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeck, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S., (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*, 531–537.

Gordon, A.D., (1999). *Classification*, 2nd ed., Chapman and Hall.

Japkowicz, N. and Stephen, S., (2002) The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis Journal*, *6*.

Kaufman, L. and Rousseeuw, P., (1990). *Finding Groups in Data*. Wiley.

Kay, S.M. and Marple, S.L., (1981) Spectrum analysis—a modern perspective. *Proc. IEEE*, *69*, 1380–1319.

Pearson, R.K. (2003). Generalized Sensitivity Analysis: A Framework for Evaluating Data Analysis Results. *Proc. 3rd SIAM Internat. Conf. Data Mining*. San Fransisco, May, 2003, pp. 212–223.

Pearson, R.K., Liu, H., Miller, D., and Gonye, GE, (2003) Clustering Short, Irregularly-sampled Time-series. Submitted for publication.

Pritchard, C.C., Hsu, L., Delrow, J., and Nelson, P.S., (2001), Project normal: Defining normal variance in mouse gene expression. *Proc. Natl. Acad. Sci.*, *98*, 13266–13271.

Schena, M., Shalon D., Davis, R.W. and Brown, P.O., (1995), Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*, 467–470.

Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G., (1999) Systematic determination of genetic network architecture. *Nature Genetics*, *22*, 281–285.