Strategies for Winnowing Microarray Data

D.B. Skillicorn
School of Computing, Queen's University
Kingston Canada
skill@cs.queensu.ca

Simeon Simoff
Paul Kennedy
Faculty of Information Technology, University of Technology
Sydney Australia
{simeon,paulk}@it.uts.edu.au

Daniel Catchpoole
Westmead Childrens' Hospital
Sydney Australia
DanielC@chw.edu.au

Abstract

The analysis of microarray datasets is complicated by the magnitude of the available information. Most data mining techniques are significantly hampered by irrelevant or redundant information. Hence it is useful to reduce datasets to manageable size by discarding such useless information. We present techniques for winnowing microarray datasets using singular value decomposition and semidiscrete decomposition, and show how they can be tuned to extract some information about the internal correlative structure of large datasets.

1 Introduction

Microarrays have the potential to explicate the relationships between biochemical pathways and higher-level biological processes such as disease and drug activity. At present, knowledge extraction from microarrays is the bottleneck to fully exploiting their potential, since a substantial amount of human interaction is still required. This is slowly changing as data mining is applied and customized for the microarray setting, and more experience of how much can be automated is gained.

1.1 Technology Background. A microarray captures information about the levels of expression of a large number of genes (and other proteins) in a single sample. Collecting the results from identical microarrays processed with different samples produces datasets with a large number of rows (corresponding to the number of genes on the microarray and typically in the tens of thousands) and relatively few columns (corresponding to samples from patients and typically in the tens to hundreds).

There are two main kinds of microarrays: those in which each location measures the absolute expression level of a particular gene, and so which are used with a single sample; and those in which each location measures the relative expression level of a particular gene, and so are used with a mixture of two samples, one representing the background ('normal') expression and one representing the foreground ('disease'), each labelled with a different dye. These two different kinds of microarrays produce datasets with different properties that must be taken into account in subsequent analysis.

1.2 Gene Data Issues. The main application of microarrays is to solve an inverse problem: which changes in gene expression account for observed

differences among arrays (representing an organism at different times or different organisms with different conditions, e.g. patients). Such information can reveal previously unknown subclasses of conditions (perhaps corresponding to new subtypes of diseases), can be used to build predictors to predict conditions, and may reveal the biochemical pathways that result in different conditions. For example, many cancers contain subclasses that are difficult to distinguish and for which different treatments have dramatically different success rates. Correct identification of subclass is therefore critical to successful treatment.

In some cases, it is possible to classify patients, and so samples, based on clinical knowledge (e.g. diagnosis) and so the problem becomes to connect this classification to its gene expression origins. In other cases, the appropriate classification may not be known in advance (e.g. prognosis for disease course); even the appropriate number of classes may not be known. Here the problem is to discover both appropriate clusters of patients and genes, and also the connections between them. In many clinical settings, the problem is somewhere between these two extremes – pragmatic classifications are known but assigning patients to classes is not always completely accurate, and the classes may contain as yet unrealized subclasses.

Consider two genes that are both expressed differently for different target classes. The magnitudes of their relative expression may be different: for example one may have a moderate expression level for one class and a high expression level for the other class; while the second has a low expression level for the first class and a moderate expression level for the second. If each row is considered as a point in a metric space, then the points corresponding to these genes may be far apart. Of course, this may be mitigated by suitable normalization of the expression levels, but such normalization is necessarily affected by all of the values for a particular gene and hence susceptible to noise. It is also possible that the two genes are strongly negatively correlated, in which case they will necessarily be far apart (on opposite sides of the origin) in a metric space representation of the dataset. In any case, normalization will change the shape of a localized set of genes which may mislead clustering algorithms that assume, for example, spherical clusters.

From this we conclude that both proximity-based methods and density-based methods should be treated with caution as techniques for clustering gene expression data. Such clusters are not localized, and so are not easily visible to techniques that treat the dataset as a representation of a metric space. (This may partially explain why some attempts at clustering have found large numbers of clusters in settings where quite small numbers might have been expected.)

On the other hand, gene expressions that are unrelated to target classes tend to be uncorrelated with each other, while gene expressions that are related to target classes are correlated (positively or negatively), even if they are not localized. In other words, even though clusters of genes related to classes are not localized, those genes that are unrelated are unlocalized, and the remaining correlation among the interesting genes may be visible against that background. Hence data-mining techniques that are based on correlation are likely to detect gene expression clusters.

One obvious initial strategy is to improve datasets by discarding genes and/or samples that are (probably) irrelevant to any interesting biochemistry. When target classes are known, the expression level of a relevant gene should differ between the columns corresponding to each class. Significance tests can be applied to try to detect such genes: for example, Significance Analysis of Microarrays (SAM) [13], thresholding [11], or neighborhood analysis [4]. In a setting where higher-order correlation is commonplace, discarding genes on the basis of their individual properties runs the risk of discarding a significant set of genes whose expression, although small in magnitude, is tightly correlated.

PCA and SVD are standard techniques for dimensionality reduction and factor analysis. They have been used to analyze microarrays in three related ways:

- To reduce the dimensionality of the data and/or remove 'noise' from the data;

- To generate 'eigengenes' and 'eigensamples' that capture concerted behavior of a number of genes or samples [1];
- To cluster genes or samples using spectral clustering methods [7].

1.3 Proposed Approach. In this paper we show how two matrix decompositions, singular value decomposition (SVD), and semidiscrete decomposition (SDD) can be used to winnow a set of genes to a much smaller set. Unlike the standard application of SVD, the goal is not to produce some set of eigengenes, but to remove from consideration genes whose correlative relationship to the other genes suggests that they are not of interest in the context of a given set of samples. The reduced dataset can then be analyzed using any of a number of other data mining techniques. The techniques described here can also be used when the target classes are unknown, so they can be applied even when thresholding cannot.

Both SVD and SDD map the original data into a structure with the following two important properties:

1. Genes that are similar and/or correlated are mapped to close locations within the structure; and
2. The structure contains a neutral point to which genes that are either correlated with everything or correlated with nothing (and hence are probably 'uninteresting') are mapped.

The second property is the key to winnowing, since these uninteresting genes can be discarded with little risk of losing useful information. Notice that this is sharply different to most attribute selection techniques which try to find the genes with the *most* predictive power, a strategy that does not work well when many genes have a small amount of predictive power. The first property can help with prediction since, for example, those genes that predict a given target gene well will be close to it within the structure.

A further advantage of these matrix decompositions is that they are symmetric with respect to genes and samples, so that identical techniques can be used to discover relationships among samples (including outlier samples that may correspond either to unexpected subclasses of disease or to process flaws).

We will illustrate these techniques using a dataset comparing pediatric patients diagnosed with acute lymphocytic leukemia (ALL) and healthy subjects. This dataset was gathered using cDNA technology and postprocessed using an Axon GenePix scanner. Per sample analysis of the quality of the measured expression levels showed high reliability, so only four attributes per patient sample were used – the median contrasts at each laser frequency. We used a total of 24 samples from 9 different patients. The coordinates of each spot were also included so that we could check for location effects.

The paper is organized as follows: Section 2 outlines some of the related work. Section 3 defines the singular value decomposition and shows how it can be used for winnowing using correlation. Section 4 defines the semidiscrete decomposition and shows how it is used for winnowing using difference. Section 5 illustrates how weighting can be used to focus attention on particular structures. Finally we draw some conclusions.

2 Related Work

There is a vast amount of work related to data mining of microarray datasets and we sketch only a few techniques that have been used for leukemia-based data. A useful survey is Yang and Speed [14]. The unusual difficulties of mining such datasets are (a) the large number of genes compared to the number of samples, which makes it likely that spurious correlations will be observed (see [12] for a summary of these issues), (b) the wide error bounds for microarray readings arising from the complex processes of preparation, hybridization, and microarray preparation and reading [2], and (c) the fact that most biochemical processes are the result of multiple weak influences rather than one strong one, whereas most data-mining techniques, for example decision trees, look for the smallest set of influences.

Most techniques are supervised, that is it is assumed that the classification of samples (with

respect to subclasses of disease or prognosis) is known. Golub *et al.* [4] used a measure based on correlation with the target class to determine the predictive power of each individual gene, and provided lists of genes predictive of AML (Acute Myeloid Leukemia) and ALL (Acute Lymphoblastic Leukemia). Yeoh *et al.* [15] use two-dimensional hierarchical classification trees to learn the genes most predictive of six subclasses of ALL.

3 SVD Winnowing of Microarray Data

3.1 Singular Value Decomposition. Singular Value Decompositions (SVD) [3] is a well-known matrix transformation that is often used to reduce the dimensionality of data. Suppose that a microarray dataset is a matrix A with n rows (corresponding to genes) and m columns (corresponding to measurements). Then the SVD expresses A in the form

$$A = USV'$$

where U is an $n \times m$ orthogonal matrix, S is an $m \times m$ diagonal matrix whose r non-negative entries (where A has rank r) are in decreasing order, and V is an $m \times m$ orthogonal matrix. The superscript dash indicates matrix transpose. The diagonal entries of S are called the *singular values* of the matrix A.

One way to understand SVD is as an axis transformation to new orthogonal axes (represented by V), with stretching in each dimension specified by the values on the diagonal of S. The rows of U give the coordinates of each original row in the coordinate system of the new axes. Hence U is a new representation for the genes. SVD measures variation with respect to the origin, so it is usual to transform the matrix A so that the attributes (i.e. columns) have zero mean. If this is not done, the first singular vector (the first axis of the transformed space) represents the vector from the origin to the center of the data, and this information is not usually particularly useful.

The most powerful property of SVD is that the maximal variation among genes is captured in the first dimension, as much of the remaining variation as possible in the second dimension, and so on. Hence, truncating the matrices so that U_k is $n \times k$, S_k is $k \times k$ and V_k is $m \times k$ gives a representation for the dataset in a lower-dimensional space. Moreover, such a representation is the best possible with respect to both the Frobenius and L_2 norms. (Note that SVD is a decomposition into linearly independent components, not statistically independent components, a distinction that is sometimes important.)

SVD has often been used for dimensionality reduction in data mining. When m is large, Euclidean distance between objects, represented as points in m-dimensional space is badly behaved in the sense that the expected distance between the farthest and nearest neighbors of a given point is very small. Choosing some smaller value for k allows a faithful representation in which Euclidean distance is practical as a similarity metric. When $k = 2$ or 3, visualization is possible.

3.2 SVD-based Information Extraction from Microarray Data. The property of most interest for winnowing is the following: view each row of U as a vector in the transformed m-dimensional space. Two vectors that are close together are positively correlated, and the angle between them is small (their dot product is a large positive number). Similarly, two vectors that are negatively correlated are on opposite sides of the origin, and their dot product is a large negative number. Two vectors that are at right angles to each other are completely uncorrelated, so their dot product is (close to) zero. However, the space only has m dimensions and there are many more vectors than this. Vectors that are uncorrelated to all of the others must have (almost) zero dot products with all of them; the only way this can happen is if all of their coordinates are close to zero. In other words, genes that are uncorrelated with most of the others will tend to be positioned close to the origin in the transformed space, while genes with significant correlation to other genes will tend to be positioned further from the origin.

The structure to which SVD maps genes in the dataset is a low-dimensional space in which proximity corresponds to similarity or correlation, and whose neutral point in the origin.

SVD transforms the given dataset into a struc-

ture in which distance from the origin is a surrogate for interestingness, in the sense of possessing significant correlative structure. We can draw the following conclusions about genes from their position in a transformed SVD space:

- Genes that are uncorrelated with any other genes will be close to the origin. Intuitively, the point corresponding to such a gene is being pushed in towards the origin by the other genes. It is unlikely that processes of biological interest will involve only a single gene, so such genes may be removed from the dataset with low risk of losing information.

- Genes that are correlated with many other genes will also be close to the origin. Intuitively, the point corresponding to such a gene is being pulled towards many of the other genes and reaches equilibrium near the center. Again, such genes are unlikely to be of great interest since they have little predictive power.

- Genes that are correlated with a moderate number of other genes will tend to be far from the origin. The distance from the origin provides some information about the tightness of the correlation with other 'interesting' genes. Since proximity corresponds to correlation, the direction from the origin is also significant.

  It is worth distinguishing two structures that commonly occur. When a group of genes are correlated among themselves, but also correlated with many other genes, they may appear as a bulge in a central spheroid. If the group are less correlated with other genes, such bulges may actually be separated from the central spheroid, and may appear as distinct clusters. This is, of course, the most interesting case since it corresponds to a clear subset of genes with mutually correlated behavior.

Note that we are not using SVD as a dimensionality reduction technique in the usual way – we are in fact reducing the dimensionality of the *patient/sample* space.

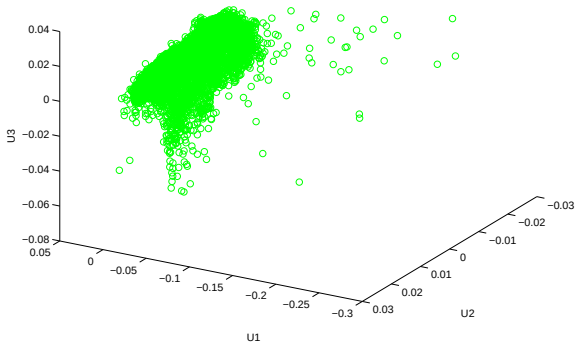

Figure 1: 3-dimensional plot of genes after SVD

3.3 Example of the application of SVD. We now show how this strategy works using a dataset of samples from pediatric leukemia patients. Data related to housekeeping spots were removed, leaving a total of 10752 genes, only a few of which were duplicates. The dataset was normalized by column to zero mean and unit standard deviation. Figure 1 shows a plot in 3 dimensions of the entire set of genes.

The main cluster is elliptical, oriented along the U1 axis, the axis of maximal variation in the transformed space. This axis would typically represent the magnitude of normalized gene expression, with relatively highly expressed genes at one end and relatively weakly expressed genes at the other. The points far from the origin along the U2 axis represent a process that is uncorrelated with the process expressed along the U1 axis. The top twenty known extremal genes in the plot of the entire microarray dataset are (in decreasing order): B2M, RPS27, RPL13A, SERPINB6, HLA-C, RPL30, RPS6, KIAA0404, HLA-A, RPS8, LYZ, RPL9, and HBE1. Note that the main cluster is hard to analyze in any further depth because of the sheer number of points within it.

Figure 2 shows the singular values for this decomposition. The magnitude of these values indicates how much variation is being captured by each dimension. In particular, much variation is captured by the first 10 dimensions.

We now remove genes that are unlikely to be interesting, by removing those points closer to the origin than the median Euclidean distance. We

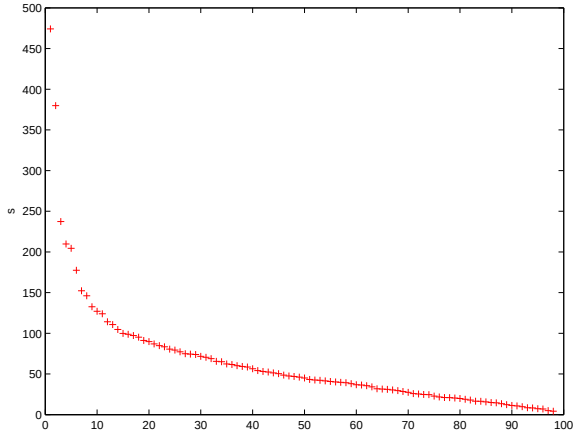

Figure 2: Magnitude of the singular values

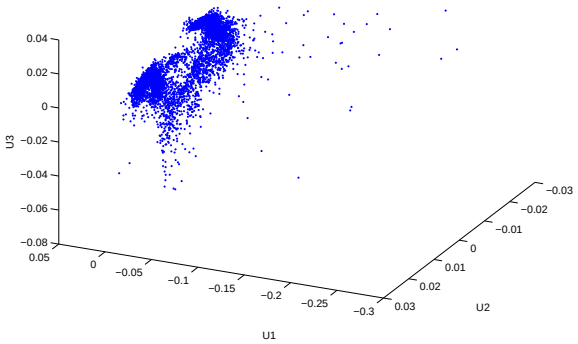

Figure 3: 3-d plot of genes further than the median distance from the origin

could compute the distances in all 98 dimensions, but having transformed the dataset using SVD and knowing the magnitudes of the singular values, we can instead compute the median using only 10 dimensions with confidence that those points discarded will be essentially the same.

Figure 3 shows that the points removed are, as expected, not extremal in any direction – indeed it is difficult to see that half of the points have been removed.

Removing points closer than 1.5 times the median distance leaves only 1660 genes, but as Figure 4 shows, the extremal points remain intact, and it is possible to see some of the structure of the main cluster. Note, for example, the small tight cluster just to the left of the top of the main cluster in this view.

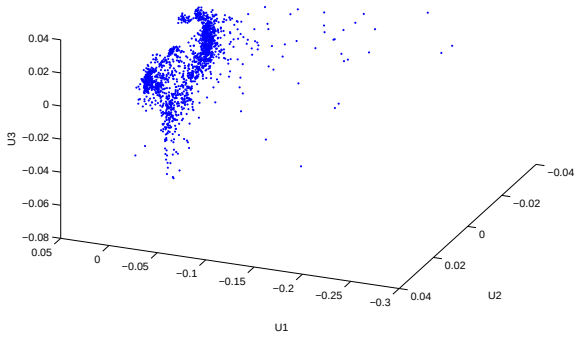

Figure 4: 3-d plot of genes further than 1.5 times the median distance from the origin

The list that remains after winnowing has been examined using CONGO, which uses the GO gene ontology to determine whether sets of genes share common function. The list does show substantial functional coherence [6].

Winnowing based on distance from the origin could have been used directly on the original matrix A to discard genes whose differential expression is very small. It is instructive to consider the different effect of applying this technique to the U matrix instead. Winnowing based on the U matrix will keep as interesting groups of genes with small differential expression *provided that they occur as part of a group with similar, though small, expression.* Conversely, single genes with large differential expression will be preserved by winnowing on A directly, but probably not by winnowing on U. Hence, winnowing on U is more discriminating. This argument shows some of the dangers of blindly thresholding during early analysis.

4 SDD Winnowing of Microarray Data

SVD views the data in a geometrical way, in which the rows and columns of matrices are regarded as coordinates in vector spaces. In contrast, semidiscrete decomposition works with the matrix A itself, looking for 'bumps', rectilinearly aligned regions of similar value.

4.1 SemiDiscrete Decomposition. The semidiscrete decomposition [8–10] expresses the

matrix A as a sum

$$A = \sum A_i$$

where each A_i is the product of a column vector, x_i, of length n, a row vector, y_j, of length m, and a scalar, d_i like this:

$$A_i = d_i x_i y_i$$

where the entries of x_i and y_i are constrained to be only -1, 0, or $+1$. Note that the product $x_i y_i$ is $n \times m$, that is the same shape as A itself. This product forms a stencil or footprint that describes the locations in A that are accounted for by this 'bump' while d_i describes the magnitude of the value that is accounted for at each of these locations. In other words, the SDD describes how to recreate A as the sum of a set of submatrices, each of which is (negatively or positively) constant at the locations described by the footprint and zero elsewhere.

Note that SDD regards values in the matrix as correlated if they are of about the same magnitude, whether positive or negative (corresponding to locations where the entries of x_i and y_j are $+1$ or -1).

SDD can be expressed as a matrix equation similar in form to that of SVD as follows:

$$A = XDY$$

where X is the horizontal concatenation of the X_is, Y the vertical concatenation of the Y_is, and D a diagonal matrix whose entries are the d_is. Each row of X corresponds to a row of A, and hence to a gene. Collectively X defines a ternary hierarchical decomposition of the genes: they are divided into 3 groups according to the value in the first column of X; each of these groups is further subdivided into three subgroups according to the value in the second column of X, and so on.

Because SDD uses the *volume* of bumps to decide which to remove next, it has one tunable parameter, the relative magnitudes of the matrix entries. If the entries are increased, say by signed squaring, the 'height' of each location in the matrix changes while the area each bump covers remains the same – hence small, high bumps are likely to be considered more significant. On the other hand, if the entries are decreased, then the effect is to make low bumps covering many locations seem more important. This can be exploited to look for either outlier structure or mainstream structure in a dataset.

Genes whose expression values have similar magnitudes will tend to fall into the same bumps. The rows of X corresponding to genes of little interest will contain zeroes – hence the zero branch of the hierarchical classification represents the neutral point.

4.2 SDD-based Winnowing of Microarray Data. In an SDD, the number of columns of the X matrix may become quite large, even larger than m the number of columns of the original matrix. In other words, A can be expressed as the sum of more than m matrices. However, just as in SVD, the earlier terms of this sum tend to be more important; the magnitudes of the d_is decrease. This has two implications for the hierarchical classification induced by the columns of X: first, the higher levels of the tree represent more important distinctions; second, the 0 branches of the tree represent genes that have not participated in any 'bumps' and so are perhaps less interesting.

Some care is needed. Because the decision about the next bump to include in the sum is based on the *volume* of the bump, bumps of low height but which occur in a large number of locations can be considered more important than localized high bumps. Hence, bumps at level 1 may not be more 'important' than those at level 2, although they are certainly more important than those at level 10, say.

It is useful to visualize the classification induced by SDD by overlaying it on plots of points from SVD. This combined plot shows how SVD and SDD agree and disagree about the classification of genes.

Figure 5 shows the SDD classification superimposed on the positions from SVD. Here color is used to indicate the first column of X (green = 0, red = $+1$, the -1 branch is empty); and shape is used to indicate the second column of X (dot = -1, circle = 0, cross = $+1$).

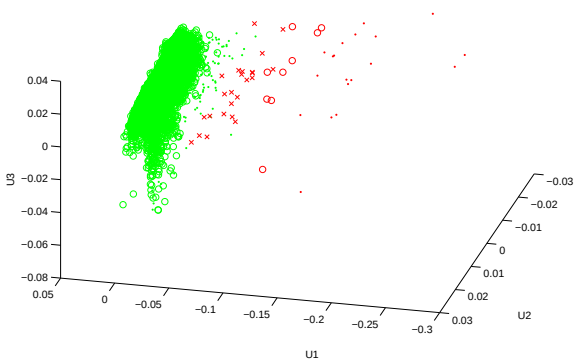

Figure 5: SDD classification imposed on SVD positions

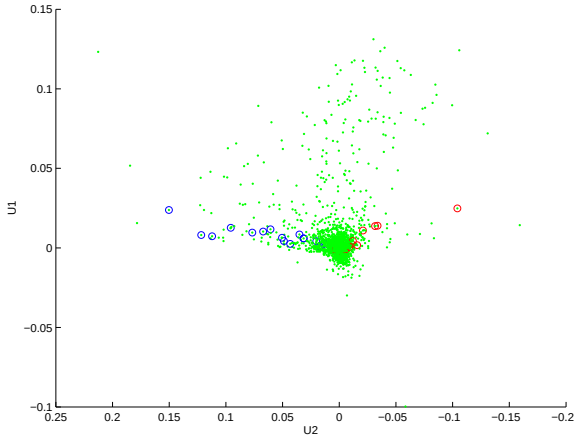

Figure 6: SVD plot of Golub dataset, blue circles: genes predictive of AML, red circles: genes predictive of ALL

This figure shows two interesting things about the SDD classification. First, it is clear where the boundary between ordinary genes and those differentially expressed in leukemia might be (those that are colored green and red respectively). The genes related to leukemia form a bump that is strongly visible in the data. Second, further information is provided about how unusual each of the genes related to leukemia is. For example, those genes labelled with red dots are clearly more unusual than those labelled with red crosses, according to both SVD and SDD.

Golub *et al.* [4] published an early paper analyzing a dataset for genes predictive of the difference between AML and ALL, and data from this paper has been extensively analyzed by others [5, 15]. This dataset is in some ways easy, since the two classes can be correctly predicted based on a single gene (zyxin). However, Golub *et al.* provide a list of the top fifty genes predictive of each case. Figure 6 shows the SVD plot from this dataset with these 100 genes labelled. It is clear that SVD produces the same kind of results for this dataset as for the dataset above: there is a dimension that discriminates the classes well (in both cases the U2 dimension), and the extremal points appear to be most interesting. The original paper makes it clear that discrimination between the classes could be done well using other genes, and this is clear from the plot – there are two well-defined opposite 'arms' of points, each of which is predictive of one leukemia type. Indeed, others [15] have defined smaller sets of discriminating genes.

5 Focus as Fine Tuning of Winnowing Techniques

When certain genes (or patients) are known to be of interest, both matrix decompositions allow this information to be used to focus the analysis.

For SVD, adding scalar weights to rows or columns of the (previously normalized) data has the effect of moving them further from the origin, and hence making them seem more interesting. However, as a side-effect, the points corresponding to other genes that are correlated with the weighted genes also move further from the origin because of their correlation. This often reveals a cluster of correlated genes that was previously hidden because it was inside another larger cluster and hence hard to see.

The amount of weight to be applied to the 20 most extremal genes in the U2 direction in the original plot requires some experimentation. If a weight of 10 is applied to these genes, the resulting plot has these genes as outliers, and the remaining genes in one undifferentiated cluster. However, it is clear from this weighting that these genes divide themselves into four subclusters: HLA-A and HLA-C; LYZ; HBE1; and the rest. Such a clustering is at least superficially plausible. When a weight of 2 is applied to these genes, there is some movement of related genes outward from the main cluster. Although this improves the visualization

of these genes, no new information results because they were detected by sorting the data positions. Some examples of such genes are: RPS7, VRP, and TPT1.

Weighting of the columns (corresponding to patients) can also reveal properties of genes. For this dataset, we have physician ratings of patient risk (normal versus high risk) for the patients in this dataset. Increasing the weight on high-risk patients induces different positions for genes in the SVD plot. In effect, we are now able to examine those genes that are associated with high-risk ALL leukemia, rather than just with ALL leukemia.

Figure 7 shows the SVD plot when the weight on the high-risk patients is increased to 8. There is little visible change from the unweighted case. However, when the SDD classification is overlaid, as shown in Figure 8, a new set of genes, labelled by green dots, begin to separate from the main cluster. There are also an increasing number of genes separating from the main cluster along its long axis.

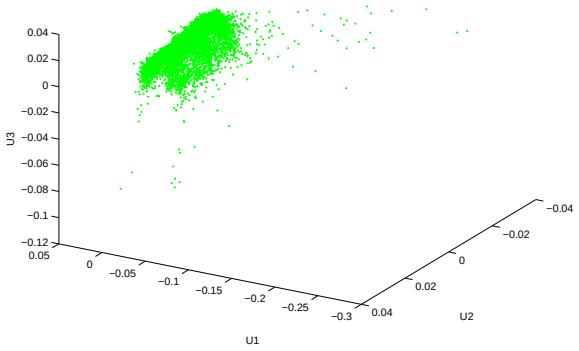

Figure 7: SVD plot of genes with high-risk patients weighted

6 Conclusions

Microarray datasets plausibly contain a great deal of information that is not helpful for determining the biochemical structures underlying particular conditions. Most data mining techniques, even those with built-in abilities to perform attribute selection, do not perform well in the presence of irrelevant or redundant information. We have presented two techniques, based on singular value decomposition and semidiscrete decomposition, that are able to generate a kind of ranking of the presumptive interestingness of genes. This ranking can be used to discard substantial fractions of the genes, allowing more sophisticated techniques to be applied robustly to what remains. The two matrix decompositions interact in ways that are more revealing than using them separately; each also allows for some tuning to allow particular structures to be searched for. We have illustrated these techniques on a dataset of patients with pediatric ALL, where more than 80% of the genes appear not to be significant.

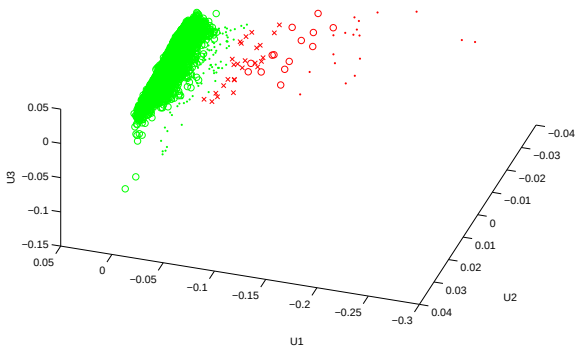

Figure 8: SDD labelling of SVD plot of genes with high-risk patients weighted

Matrix decompositions have the potential to extract clustering information from datasets, but background knowledge and fairly sophisticated use of the techniques is required. Here we only consider their use as a preprocessing step to reduce dataset size and complexity for other, downstream model building techniques.

References

[1] O. Alter, P.O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Science*, 97(18):10101–10106, 2000.

[2] M.E. Futschika, A. Reeveb, and N. Kasabov. Evolving connectionist systems for knowledge discovery from gene expression data of cancer tis-

sue. *Artificial Intelligence in Medicine*, 28:165–189, 2003.

[3] G.H. Golub and C.F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[4] T.R. Golub, D.K. Slonim, P. Tamayo, C.Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 15 October 1999.

[5] J. Hardin, D.M. Rocke, and D.L. Woodruff. Robust model-based clustering of genes in microarray data: Are there gene clusters. In *Proceedings of CAMDA 2000*, 2000.

[6] P.J. Kennedy, S.J. Simoff, D.B. Skillicorn, and D. Catchpoole. Extracting and explaining biological knowledge in microarray data. In *Pacific Asia Knowledge Discovery and Data Mining Conference (PAKDD2004)*, Sydney, May 2004.

[7] Y. Kluger, R. Basri, J.T. Chang, and M. Gerstein. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 13:703–716, 2003.

[8] G. Kolda and D.P. O'Leary. A semi-discrete matrix decomposition for latent semantic indexing in information retrieval. *ACM Transactions on Information Systems*, 16:322–346, 1998.

[9] T.G. Kolda and D.P. O'Leary. Computation and uses of the semidiscrete matrix decomposition. *ACM Transactions on Information Processing*, 1999.

[10] S. McConnell and D.B. Skillicorn. Semidiscrete decomposition: A bump hunting technique. In *Australasian Data Mining Workshop*, pages 75–82, December 2002.

[11] D. Nguyen and D. Rocke. Classification of acute leukemia based on DNA microarray gene expressions using partial least squares. In *Methods of Microarray Data Analysis*, pages 109–124. 2002.

[12] G. Piatetsky-Shapiro, T. Khabaza, and S. Ramaswamy. Capturing best practise for microarray gene expression data analysis. In P. Domingos, C. Faloutsos, T. Senator, H. Kargupta, and L. Getoor, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD2003*, pages 407–415. ACM Press, 2003.

[13] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Science*, 98(9):5116–5121, 2001.

[14] Y.H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nature Reviews*, 3:579–588, 2002.

[15] E.J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.H. Pui, W.E. Evans, C. Naeve, L. Wong, and J.R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1, March 2002.