

Microarray Analyzes Towards Cancer Classification

Leonardo Araujo

Abstract—We present here the analysis and classification of cancer data of two types: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The data is highly dimensional and there are only a few samples available. The simple k-nearest neighbors (KNN) was used and achieved a good result. The method here proposed consists in taking the dissimilarity between samples to create a dissimilarity matrix. This one is used by a multidimensional scaling (MDS) method to represent the data in a principal component space. The representation in 2 dimensions shows also a good performance in the classification using KNN. This simpler representation seems promising for the application of other classification techniques that would have a high cost for high dimensional data.

Index Terms—cancer, acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), multidimensional scaling (MDS), k-nearest neighbors (KNN).

I. INTRODUCTION

The classification of cancer is important to determine the therapy to used because pathogenetically distinct tumor types requires specific therapies, to achieve a better efficacy and lower the toxicity levels the patient is subject to. Cancer classification is based nowadays mainly in morphological analysis of the tumor, what is known to be a limited analysis that may leads to false conclusions Golub et al. (1999).

In the present work, we chose acute leukemias as a test case, and we aim at classification between two types: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). In 1970s, the classification between those two types was solidified by the development of antibodies recognizing either lymphoid or myeloid cell surface molecules. “Although the distinction between AML and ALL has been well established, no single test is currently sufficient to establish the diagnosis”, it is always subjected to the interpretation of a specialist in a laboratory Golub et al. (1999).

The analysis of gene expression using DNA microarrays is important to establish a cancer classification procedure in order to get a more accurate diagnosis and then apply the right treatment, increasing the chances of the patient recovery.

The data set used here is the same provided by Golub et al. (1999). It consists of 38 bone marrow samples (27 ALL, 11 AML) obtained from acute leukemia patients. A microarray was used with 7129 probes for 6817 human genes. For each gene, a quantitative expression level is obtained.

II. MICROARRAY

Microarray technology is a technology used by many biologists to monitor genome expression levels of genes in a given organism. It consists of an array with thousands of spots filled with a few million copies of DNA oligonucleotides¹, called features, or probes.

The most popular way of measuring gene expression with microarrays is to compare expression of a set of genes in a query cell with the expression of the same set of genes in a reference cell. RNA is extracted from the cells and reverse transcribed into cDNA. Fluorescent dyes are used to label the cDNA samples, usually Cy3 and Cy5, what will correspond to a green or red fluorescence respectively. The samples from a query cell and a reference cell

¹DNA oligonucleotides are short nucleic acid polymer that may be synthesized into a sequence of fewer bases (typically up to 200 bases). DNA oligonucleotides are used as probes because they readily bind to their respective complementary nucleotide.

are labeled with different colors, and at this point, the samples may be used into the microarray, so they will hybridize to specific spots containing their complementary sequences. The amount of cDNA bound to a spot is proportional to the initial number of RNA molecules present in the cells, and so related to the gene expression.

The microarray is scanned with a laser to detect the red and green dyes. The intensity of the emitted fluorescence, upon excitation by the laser, corresponds to the amount of nucleic acid bound in that spot. If the spot is found to be red, that means the samples labeled with red are majority; if it is found to be green, then the green labeled samples are majority; if it is found to be yellow, both samples are equally found bound to the spot; and if it is found black, that means neither of them is present in the spot. In the end of the experiment, we have an image, in which each spot is associated to a gene and has a fluorescence value representing the relative expression level of that gene.

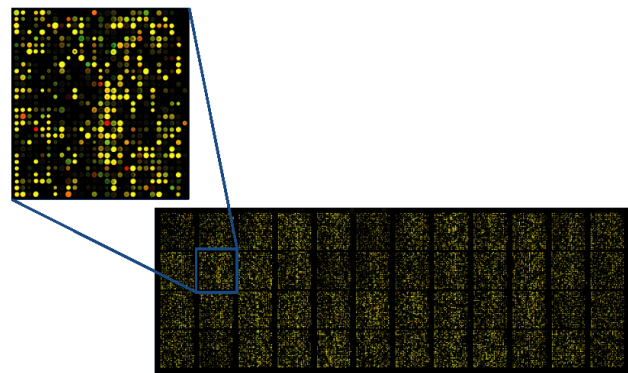


Fig. 1: Example of a microarray with approximately 40,000 probes. (Wikipedia)

The data used in the present work was collect using Affymetrix microarray containing probes for 6817 human genes. The Affymetrix technology does not used the two-colors base schema as explained before. To get the relative expression of the query case against the control case it is necessary to perform the microarray analysis of both samples separately and then compare the results to get a relative expression measure.

Regardless of the approach used (use of single label and independent arrays, or a single array with distinguishable fluorescent dye labels), after hybridization the arrays are scanned into distinct grayscale images, typically 16-bit TIFF images with no compression. These images must then be analyzed to identify the arrayed spots and to measure the relative fluorescence intensities for each element.

“The hypothesis underlying microarray analysis is that the measured intensities for each arrayed gene represent its relative expression level. Biologically relevant patterns of expression are typically identified by comparing measured expression levels between different states on a gene-by-gene basis. But before the levels can be compared appropriately, a number of transformations must be carried out on the data to eliminate questionable or low-quality measurements, to adjust the measured intensities to facilitate comparisons, and to select genes that are significantly differentially expressed between classes

of samples” Quackenbush (2002).

According to Golub et al. (1999), a primary control criteria was used on the microarray analysis consisting of eliminating the samples with weak hybridization or any visible defects in the array (such as scratches). From the initial samples, 10% were excluded on this a priori quality control criteria.

A. Data Expression Rate

The data available for the problem were expressions rates achieved by a primary comparison between the query and reference samples. We are going to call the query case by red (R) and the control case by green (G) (colors commonly used to represent the data), and the ratio (T) is the ration between the luminescence output of red and green. For each gene, this ratio will be given by

$$T_i = \frac{R_i}{G_i} . \quad (1)$$

B. Normalization

In order to verify if the hybridization process occurred with no problems, it is necessary to inset into the microarray chip some control probes that will be used as a reference to validate the performance of the whole process. It is necessary to use two different types of control probes: positive and negative control. As a positive control it is usually assigned gens that are know to be always expressed, as the constitutive gens. As a negative control it is usually used gens that know known not to be expressed, as gens from other species. The results presented in the control probes are used to evaluate if the process occurred with errors and not. They are also used as references to normalize the observed expression values.

The hybridization process is not uniform over the whole array. It is necessary then to place control spots over each region of the array. The array is usually subdivided into several subarrays, each containing its own control group. The expressions levels from each control group should be taken in account as the upper and lower limits of expressions in the subarray.

Although ratios provide an intuitive measure of expression changes, they have the disadvantage of treating up- and down-regulated genes² differently. When a gene is up-regulated by a factor of 2, the expression level found will be 2. When a gene is down-regulated also by a factor of 2, the expression level found will be 1/2. If we wish to treat up- and down-regulated genes in a similar fashion, we should use a logarithm function, because $\log 1/x = -\log x$, and now we may deal with $\log x$ and $-\log x$. The same factors for up- and down-regulated genes expressions will experience the same amplitude in a logarithm scale.

Normalization is an important step to adjusts the individual hybridization intensities to balance them appropriately so that meaningful biological comparison can be made. There are many other reasons why the intensities have to be normalized: the quantities of initial RNA may differ, the fluorescent dyes used may not have the same efficiency, and the measured expression levels may be subject to some sort of systematic bias.

Let us assume that the initial quantity of RNA in each sample is the same. There are millions of RNA molecules in each sample, assuming an equal average molecular mass across samples, and since we use the same mass in each sample, we may assume an equal average number of RNA molecules in each sample.

²Up-regulated genes are those genes highly expressed. Down-regulated genes are those with lower expression level.

III. DATA

The dataset used in this study comes from Golub et al. (1999), where there is gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays (HU6800 chip) containing probes for approximately 6,800 human genes and ESTs. The chip actually contains 7,129 different probe sets; some of these map to the same genes and others are there for quality control purposes. The data comprise 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML. Samples are divided into a learning set with 38 observations and a test set of 34 observations. We will make the assumption that the data have been suitably pre-processed, that means: image analysis, normalization and computation of expression measures. The analysis of the control probes, seen in figure 2, points in that direction. We will then trust that the data were suitably pre-processed.

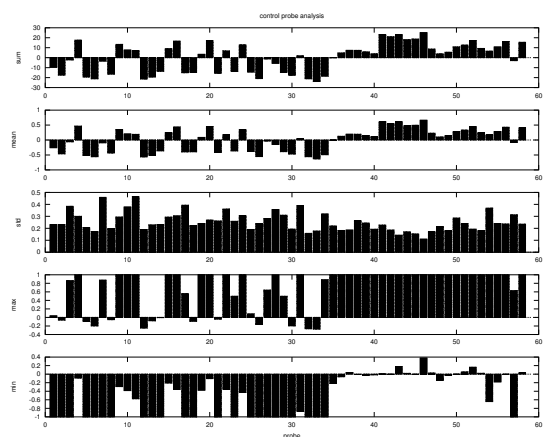


Fig. 2: Control Probes Analysis.

IV. APPROACH

The approach here used to analyze the data and classify is based on Multidimensional Scaling (MDS) (Young and Householder, 1938; Torgerson, 1952; Kruskal and Wish, 1978; Shepard, 1963; Rothkopf, 1957) and K-Nearest Neighbors (KNN) (Cover and Hart, 1967). The microarray data is highly dimensional, and the database provided contains a very small set of samples. The process of acquiring new samples is expensive and we may not expect to see in a near future a large database of such data. To overcome the ‘curse of dimensionality’ it would be necessary to have a set of samples many times greater than the dimensionality of them, and that it completely unpractical for our problem at hand.

The method here proposed consists of computing the dissimilarity between every two samples, creating a dissimilarity matrix. We tried different types of distance measures to assign our dissimilarities measures: *euclidean*, *chebyshev*, *manhattan* and *pearson*. The figure 3 shows one example of such dissimilarity matrix computed using the manhattan distance as a dissimilarity measure between microarray data samples. The figure 4 is the same matrix with elements reordered so that the smallest dissimilarities are placed on the top left of the matrix.

This dissimilarity matrix is used as input to a MDS that represents the samples in a principal components (PC) space, regarding the dissimilarity among them. One example of such MDS representation is plotted in figure 5, where only the first two principal components

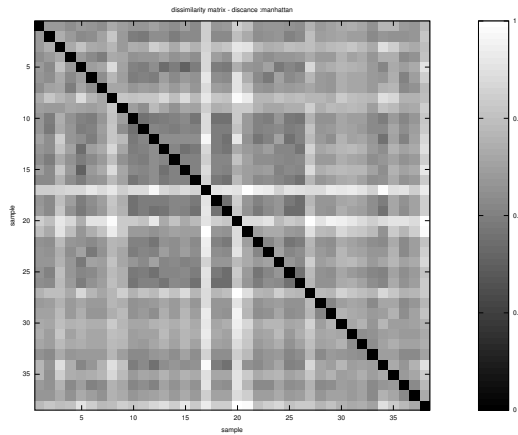


Fig. 3: Dissimilarity Matrix.

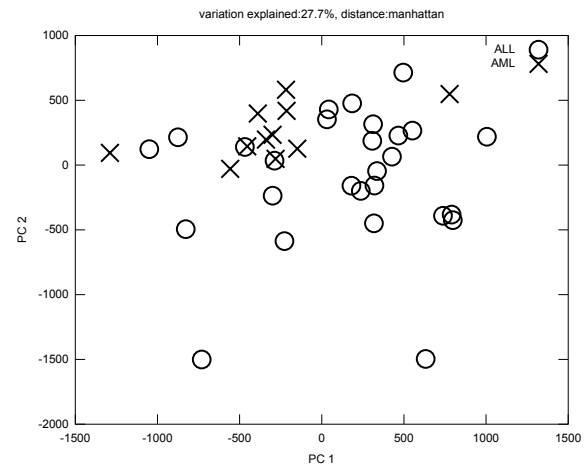


Fig. 5: MDS Result Using Manhattan Distance Measure.

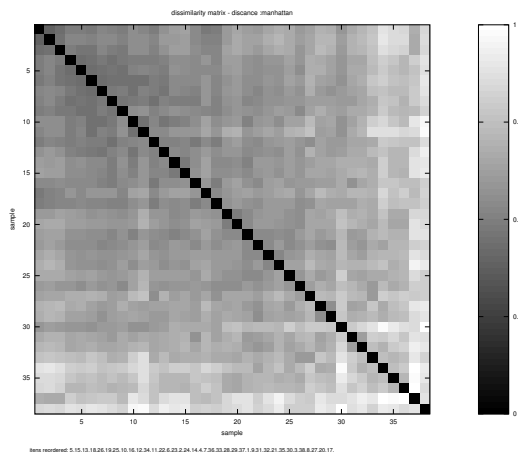


Fig. 4: Dissimilarity Matrix Reordered.

are used. These two components are chosen because they are the components with greater variance. The two of them together are responsible for almost 28% of the data variance.

The MDS output data, as plotted in figure 5 (only 2 PCs are represented), were used by a KNN method to classify between ALL and AML. Every time KNN was performed it was run 100 times, selecting 75% of the samples as known labeled samples and the other 25% as test samples, which we want to label using the KNN method. Every iteration a new set of samples was randomly selected to create the training set and test set.

For the approach here adopted we may see that some variables play important roles: (1) the number of neighbors considered by the KNN method, (2) the number of features used to create the dissimilarities, (3) the number of principal components used in the classification stage. The section Results presents some analysis of the role each of these variables plays and the final results of the KNN classification method.

V. RESULTS

The figure 6 shows that the KNN performance for $k = 3$ along with different number of features used to create the dissimilarity measures. As the performance varies so much for small changes of the number of features used, a moving average, with windows length equals 100, is shown in red in the same figure. It may be observed

that the best values (on average) are around 1500 features and 5000 features. It is important to note that the features are always selected at change, that means that the results on figure 6 is just one among other results, but which represents the typical profile observed on the others. Considering the results, we chose to work with the fewest number of features as possible that gives us the best result possible, so it is better to work with a number of features around 1500 instead of a number of features around 5000.

Another analysis was made to check the performance of the KNN for various k values. The results are on figure 7. It was considered only a number of features between 1400 and 2000 features, as pointed out before, to avoid a computational cost with no reason. We observe that when working with a greater number of PCs (6 PCs, as shown in the figure 7), the performance of the KNN decreases as the number for k increases. With a smaller number of PCs (in our example, the results for 2 PCs are shown in black), the pattern is almost the same, but the performance for small values of k is almost the same, and starts decreasing when k is greater than 8.

Using the KNN for classification of the data, it reached a 81% of mean hit rate. The KNN method was performed 100 times, and the mean hit rate was calculated.

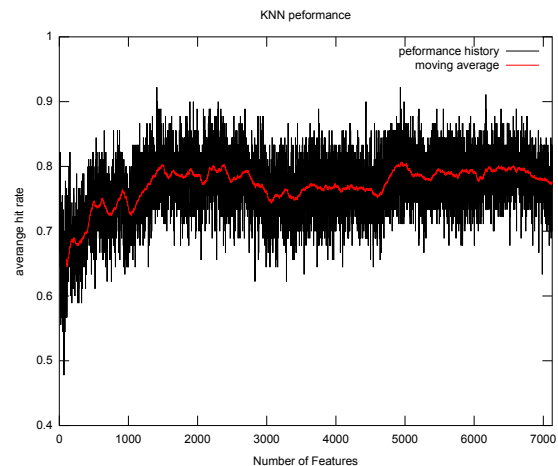


Fig. 6: KNN performance for different number of features.

If we use the KNN method directly on the data, we achieve a very good result. The table I shows the best, worst and average results for different distance measures to obtain a dissimilarity between samples.

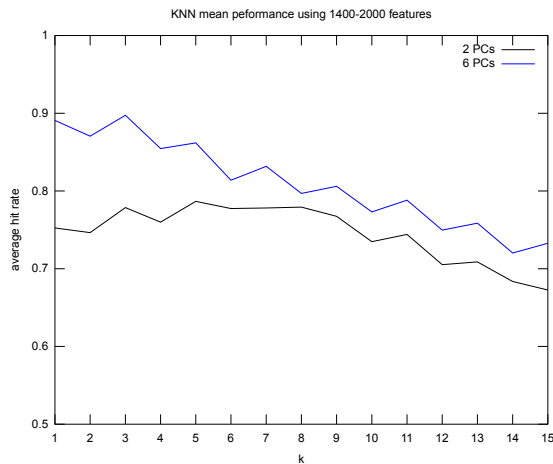


Fig. 7: KNN performance for different k values and for number of features between 1400 and 2000 features.

In every case, it was achieved a very good result. The best one was using pearson distance, which lead us to a 89% of hit rate on average.

distance	max	min	avg
euclidean	100	44	86
chebyshev	100	33	72
manhattan	100	44	87
pearson	100	56	89

TABLE I: KNN results. Hit rates (% value)

Comparing the results from the simple KNN method and our here proposed method, using only 2 PCs and all features, we realize that the performance is not increased. Although the performance is not better, we have to emphasize that this approach achieves a result almost as good as the simple KNN, but it may bring more insight, since the data is represented is represented in only 2 dimensions, what brings a easy visual analysis.

distance	max	min	avg
euclidean	100	44	77
chebyshev	100	44	72
manhattan	100	56	83
pearson	100	44	82

TABLE II: KNN using MDS results. Hit rates (% value)

From the results shown in this paper, it is difficult to tell that there are some features that are more prominent for the characterization of the samples between ALL and AML types. It was observed that most of them are not well correlated to the others. As the number of features is incredible great, it would not be possible to compute all correlations between features in a reasonable time. We computed instead the correlations among small sets of 100 features selected by chance. The results, as the one presented in figure 8, shows that there might be a small number of features well correlated one to another, but that would not help so much to decrease the dimensionality of the problem. The Principal Component Analysis (PCA) also show that there is no preferred dimension along which the data is distributed, as may be seen in the figure 9.

The figure 11 shows the effect on the variance of the first two PCs when more features are added to the calculation. When only one feature is used, the first PC has 100% of the variance. This value

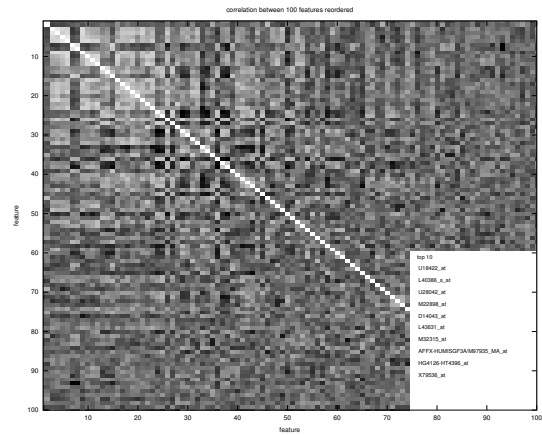


Fig. 8: Correlation Analysis.

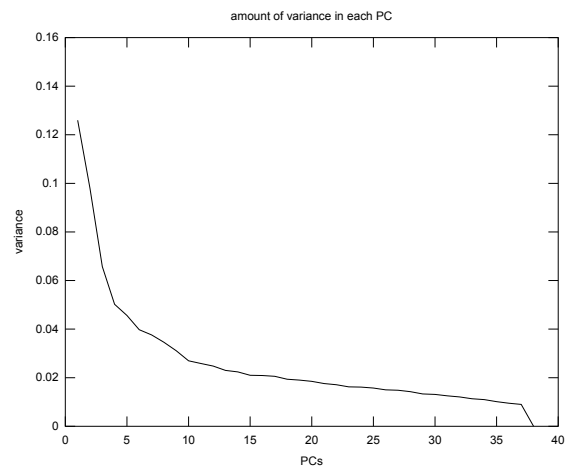


Fig. 9: PCA - Variance in each PC.

drops rapidly as other features are added to the calculations. The second PC start at null and rises rapidly, without overtaking the first PC. As more features are added the variance of the seconds PC also starts to decay. It might be seem, in a long term, that the variances of both PCs converge to a final value around 0.15 and 0.1 for the first and second PCs, respectively. The plot bellow, figure 11, shows the derivative of the variance of each PC. Greater derivative values are related to features that cause more variation on the information explained by the others considered until the moment they were added. The zoom plot in figure 11 (b) shows the labels of the features added. This procedure was performed a few times, and the typical result illustrated in figure 11 is always the same.

For each PC, as the number of features used increase, the tendency is to have a decrease in the variance associated with it. The variance of each PC was analyzed, as the number of features were increased, and the mean variance observed for each PC is plotted in figure 10. Using a threshold value of 0.1, we realize that most PCs have a mean variance bellow that threshold, only the first and second PCs have a mean variance greater. As seen before, as the number of features increase, the variance of the first and second PCs converge to a value greater that 0.1. Taking both information in account, we conclude that a good representation with the minimum number of PCs associated, would be a representation using only the first and second PC. One more analysis of figure 10 shows that the variance is almost equally

distributed among the other PCs, which have a variance around 0.02.

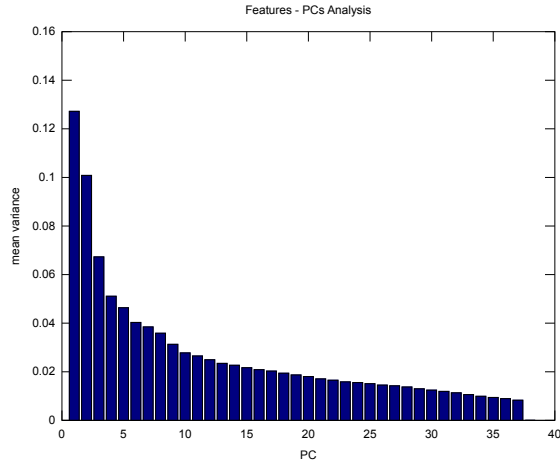


Fig. 10: Mean variance (taken through different number of features used) of the Principal Components.

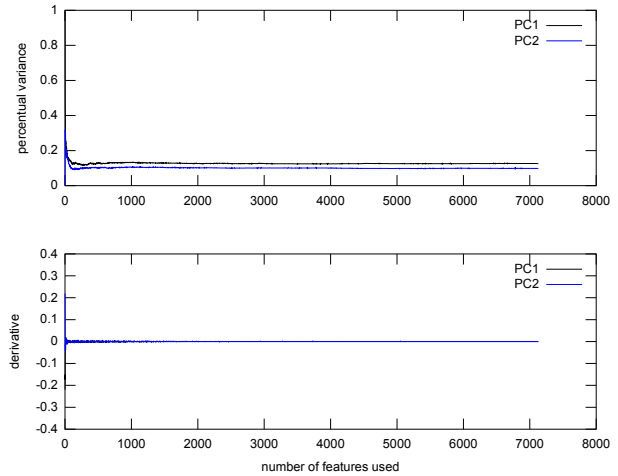
It seems convenient to apply a classification method on the data represented by the first and second PCs. Some results of the KNN method might be seen in figure 12.

VI. CONCLUSIONS

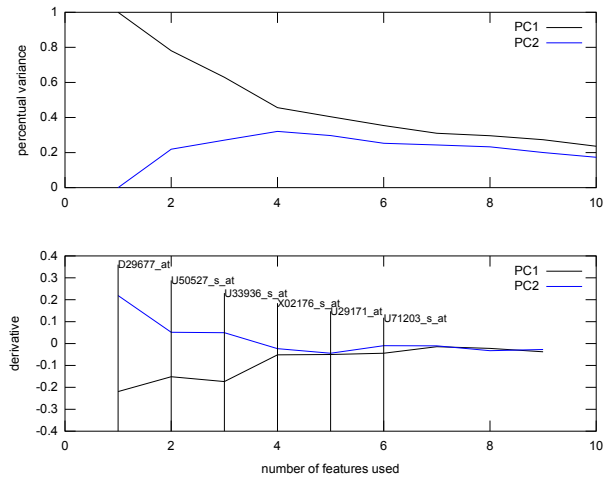
The KNN itself shows a good performance in the classification of the cancer data between AML and ALL. The proposed method showed no increase in the performance, what could be expected, since the MDS representation is created from the dissimilarity measure, and the KNN classification is also made from the dissimilarity measure. On the other hand, the performance was not considerably decreased, what means that it was possible to draw a simpler representation that elicits the class separation with almost the same discriminability. A simpler representation might be useful to acquire an insight over the data. Some analysis of the effect of using different features were taken, and it was not possible to draw deep conclusions on what are the best features or how are they correlated one to another. Some other classification techniques should be test in the future. The simpler representation might be very useful in classification methods that usually have a better performance but have a high cost when the data has such a high dimensionality.

REFERENCES

- Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, 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.
- Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage.
- Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics Supplement*, 32.
- Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, 53:94–101.
- Shepard, R. N. (1963). Analysis of proximities as a technique for the study of information processing in man. *Human Factors*, 5:33–48.
- Torgerson, W. (1952). Multidimensional scaling: 1. theory and method. *Psychometrika*, 17:401–419.



(a) Percentage variance explained by the first two Principal Components (PCs) along the number of features used (selected randomly). The plot below shows how this percentage changes.



(b) Zoom of the previous image, showing the name of the 6 first features added.

Fig. 11: Analysis of the effect of different number of features to perform the MDS.

Young, G. and Householder, A. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22.

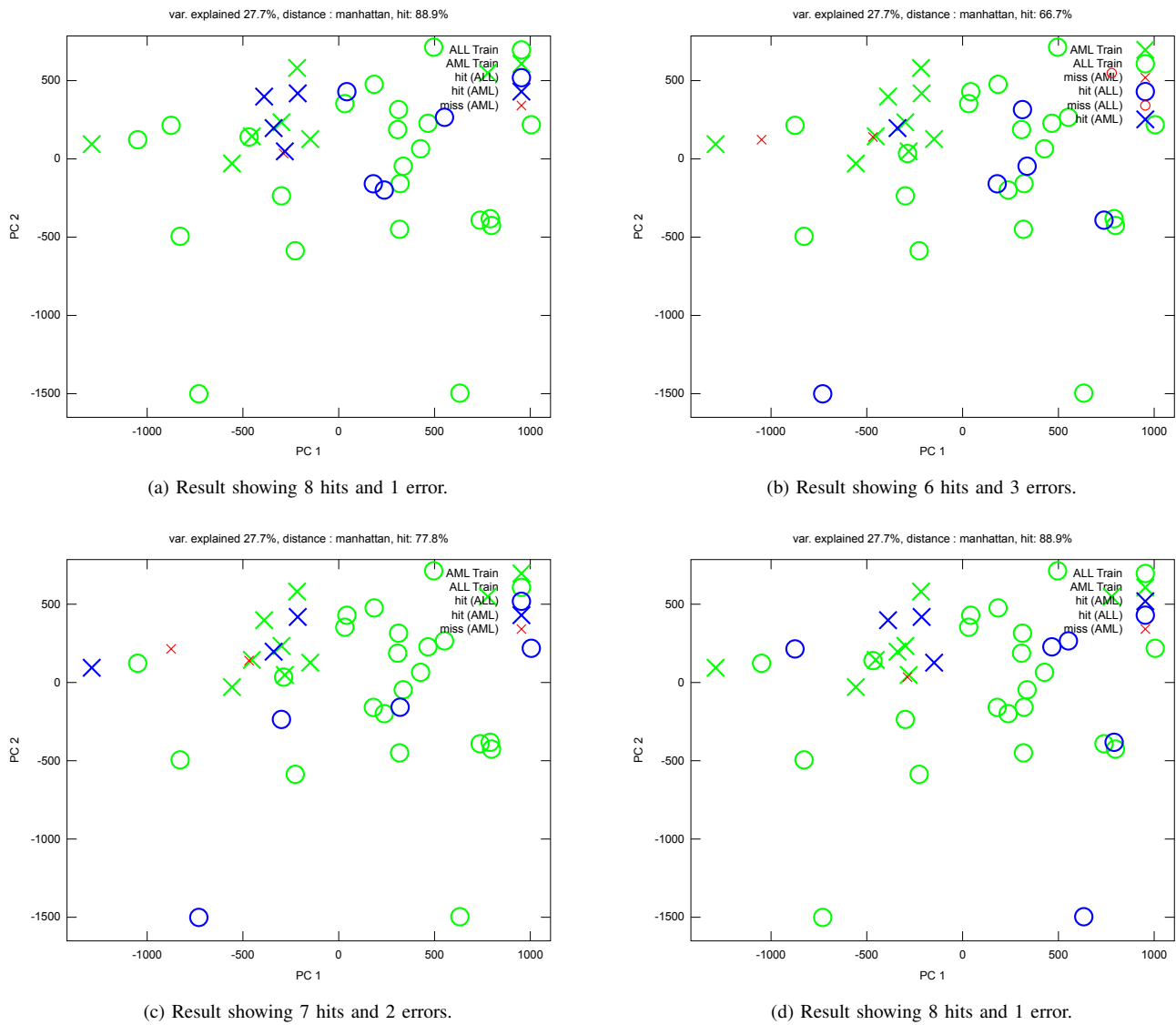


Fig. 12: Some results of KNN method applied to the MDS of the data. For all results show above it was used the Manhattan Distance to create the Dissimilarity Matrix and to perform the KNN method with $k=3$.