DTU Library

Analysis of Pap-smear Image Data

Jantzen, Jan; Dounias, George

[Link back to DTU Orbit](#)

ANALYSIS OF PAP-SMEAR IMAGE DATA

Jan Jantzen[1] and George Dounias[2]

[1]Technical University of Denmark
Oersted-DTU, Automation
Building 326, 2800 Kongens Lyngby, Denmark
Phone: +45 4525 3561, fax: +45 4588 1295, email: jj@oersted.dtu.dk

[2]University of the Aegean
Dept. of Financial & Management Engineering
31 Fostini Str., 82100 Chios, Greece

ABSTRACT: The pap-smear benchmark database provides data for comparing classification methods. The data consists of 917 images of pap-smear cells, classified carefully by cyto-technicians and doctors. The classes are difficult to separate, since class membership is not clearly defined. A basic data analysis provides numerical measures indicating how well the classes are separated, based on the Mahalanobis distance norm. The paper compares the results of three advanced classifiers against a simple minimum distance classifier. The results show that while the simple classifier provides an error rate just over 6%, error rates down to 1-2% can be achieved with a combination of feature selection together with an advanced classsifier such as ant colony optimization. Students and researchers can access the database via the Internet, and use it to test and compare their own classification methods.

INTRODUCTION

The term *Pap-smear* refers to samples of human cells (Fig. 1) *smeared* onto a glass slide and coloured by means of the *Papanicolau* method. The colouring enables examination under a microscope for any abnormality indicating a precancerous stage.

Figure 1: Superficial squamous cell stained in order to enhance contrast.

A database of single cells has been collected at the Herlev University Hospital, Denmark, by means of a digital camera and a microscope. Skilled cyto-technicians and doctors manually classified each cell into one of 7 classes. Each cell was examined by two cyto-technicians, and difficult samples also by a doctor. In case of disagreement the sample was discarded. The database thus holds diagnoses that are as certain as possible, given the practical and economical constraints at the hospital.

The staff used a commercial software package CHAMP (Dimac) to segment the images. The extracted features are combined features of the segmented and non-segmented cell images.

The data have been shaped into a benchmark database for classifiers, specified in an earlier NiSIS paper (Jantzen *et al*. 2005). Three Master's projects have analysed the data, or an older version of the database, and they are all available from the World Wide Web (Norup 2005, Martin 2003, Byriel 1999). The data have been used to test an improved second order learning algorithm for neural networks (Ampazis, Dounias and Jantzen 2004). Furthermore, the data have been used for

benchmarking various classification methods (Dounias *et al.* 2006, Panagi *et al.* 2004, Tsakonas *et al.* 2004, Dounias *et al.* 2002, Tsakonas *et al.* 2001).

The objective of the present paper is to provide additional basic data analysis, and to compare some new classification results.

PRELIMINARY ANALYSIS OF CLASS SEPARATION

The database consists of 917 samples distributed unevenly in 7 classes (Table 1). The set of classes $\{1, 2, 3\}$ contain normal cells, while the set of classes $\{4, 5, 6, 7\}$ contain abnormal cells. A minimal requirement is to separate normal from abnormal, which is a 2-class problem.

Class	Category	Cell type	Cell count	Subtotals
1	Normal	Superficial squamous epithelial	74	
2	Normal	Intermediate squamous epithelial	70	
3	Normal	Columnar epithelial	98	242 normal
4	Abnormal	Mild squamous non-keratinizing dysplasia	182	
5	Abnormal	Moderate squamous non-keratinizing dysplasia	146	
6	Abnormal	Severe squamous non-keratinizing dysplasia	197	
7	Abnormal	Squamous cell carcinoma in situ intermediate	150	675 abnormal

Table 1: The distribution of the 917 cells in the database. Classes 1-3 are normal cells, and 4-7 abnormal.

Separating each class from the rest is a 7-class problem, which is harder. Each sample is described by 20 features (Table 2) extracted from images of single cells.

Column	Feature	Name
B	Nucleus area	Narea
C	Cytoplasm area	Carea
D	N/C ratio	N/C
E	Nucleus brightness	Ncol
F	Cytoplasm brightness	Ccol
G	Nucleus shortest diameter	Nshort
H	Nucleus longest diameter	Nlong
I	Nucleus elongation	Nelong
J	Nucleus roundness	Nround
K	Cytoplasm shortest diameter	Cshort
L	Cytoplasm longest diameter	Clong
M	Cytoplasm elongation	Celong
N	Cytoplasm roundness	Cround
O	Nucleus perimeter	Nperim
P	Cytoplasm perimeter	Cperim
Q	Nucleus position	Npos
R	Maxima in nucleus	Nmax
S	Minima in nucleus	Nmin
T	Maxima in cytoplasm	Cmax
U	Minima in cytoplasm	Cmin

Table 2: Summary of the 20 features in the database.

We can immediately give up the hope that the classes be linearly separable. For example, classes $4, 5$, and 6 denote mild, moderate, and severe *dysplasia* (abnormal form), indicating that the boundaries are unclear, even to professionals. To achieve an indication of the degree of overlap, we measure the distance between each class centre compared with the variation. For instance, if two class centres are far from each other in the feature space, and the standard deviation within each class is small, then the separation is good.

There is a vast difference in magnitude among the 20 features, however. The largest measurement is in the order of 10^5 while the smallest is in the order of 10^{-3}, a span of 8 orders of magnitude. The *Euclidian distance* in the 20-dimensional feature space will certainly be dominated by the features of largest magnitude, which may obscure gaps between classes.

The *Mahalanobis distance*, on the other hand, takes the local variation into account by using the *standard deviation* as a yardstick (see appendix). Figure 2 shows classes 1 and 2 with respect to just two of the features (nucleus area and the nucleus/cytoplasm ratio of areas) in order to plot them in a 2-dimensional plot. Ellipses are drawn at Mahalanobis distance 1, 2, and 3 respectively. Even though the two features differ by roughly 4 orders of magnitude, the Mahalanobis distance is only affected by the shape of the class. The Mahalanobis distance measure is relative to each class, because it depends on the standard deviation within the class.
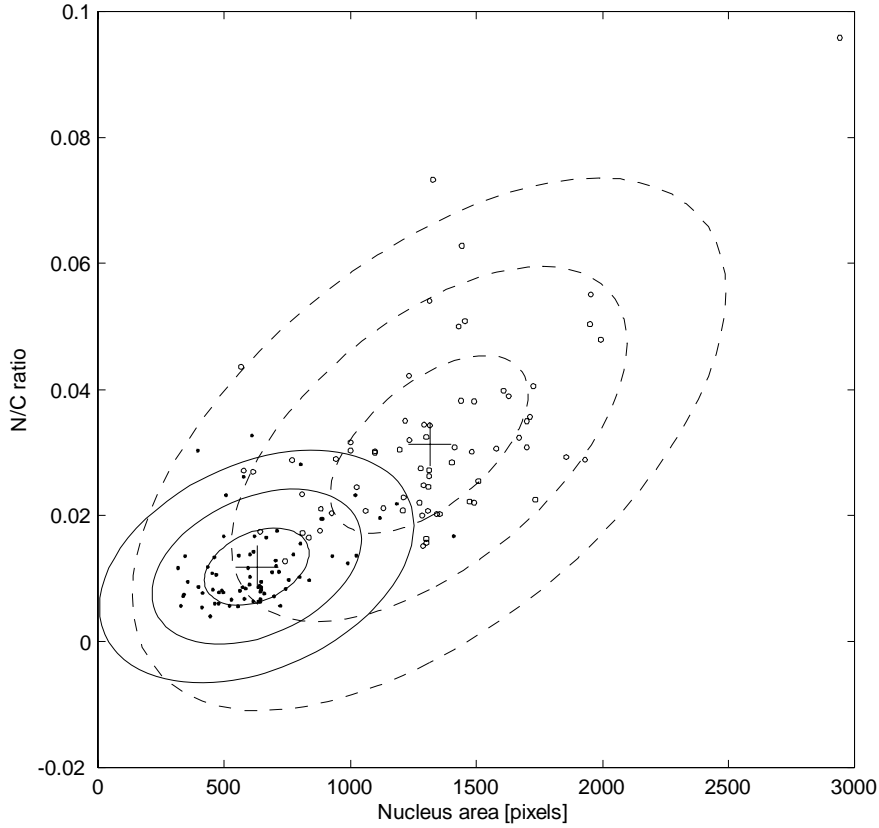

Figure 2: Scatter plot of class 1 (dots) and class 2 (circles) using just two features. Ellipses show Mahalanobis distances $d = 1, 2, 3$; solid ellipses for class 1, and dashed ellipses for class 2. The distance seen from class 1 to the centre of class 2 is larger than 3, while the distance seen from class 2 to the centre of class 1 is less than 2.

The Mahalanobis distance is appealing, because it allows us to work directly on the raw data, avoiding scaling. Thus the distance will be unbiased by our arbitrary selection of subsets of features, and we can in fact keep all 20 features and measure the Mahalanobis distance between each class centre. Table 3 defines what we shall call the *distance matrix*, that is, a 7-by-7 table **D** of distances from each class to all the other class centres. The distance is relative to the standard deviation of the current class, therefore the table is not symmetric. The table shows that distances vary over a large range. The distance from class 6 to class 7 is thus 1.1, while the distance from class 1 to class 7 is 412.4.

Centre	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
μ_1	0.0	6.0*	115.4	19.2	31.3	72.6	108.4
μ_2	9.0*	0.0	76.5	12.1	19.6	47.6	73.1
μ_3	222.8	76.0	0.0	5.3	3.5	2.6	4.5
μ_4	197.9	45.9	18.8	0.0	2.6	9.7	17.8
μ_5	259.6	73.9	7.8	1.7*	0.0	2.5	5.0
μ_6	328.7	107.7	4.3*	4.5	2.1*	0.0	1.4*
μ_7	412.4	140.0	5.2	6.7	3.4	1.1*	0.0

Table 3: Distance matrix **D**. An element $\mathbf{D}_{ij}$ is the Mahalanobis distance seen from class j to the centre μ_i of class i. For a given class, the shortest distance to a centre is marked by an asterisk (*).

Each class has a nearest neighbour, indicated by the shortest distance in each column. Figure 3 is a digraph showing the neighbour relationship for all classes. Clearly, classes 1 and 2 are each other's neighbour, and the distance to class 4 is relatively large, indicating that classes 1 and 2 are lying separately from the other classes. Furthermore, there is a forward path from class 4 to class 5 to classes 6 and 7, the last two being mutual neighbours.
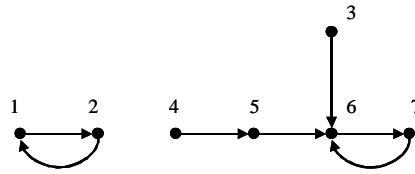

Figure 3: Digraph of class neighbours. The numbered nodes are classes, and the arrows point from a class to its nearest class neighbour, identified by the shortest distance in the columns of matrix **D**.

It is helpful that the normal classes 1 and 2 are well separated from the rest with regard to a classification of normal versus abnormal cells (the 2-class problem). But the normal class 3 represents a problem, since its nearest neighbour is the abnormal class 6, and it is quite far from classes 1 and 2. Figure 4 shows the distribution of distances within each class. For class 3 the distribution is centred more or less about 4 (mean 4.2). The distance from class 4 to class 6 is 4.3 (Table 3), thus we expect overlap between classes 3 and 6, and it will be difficult to separate them. Similarly, there is overlap between classes 3 and 7, but it is less. Furthermore, classes 4, 5, 6, and 7 are pairwise close, and all are rather dispersed, indicating overlap.

An outlier would show up as a point lying at a distance much greater than the rest, but it seems there are no outliers.
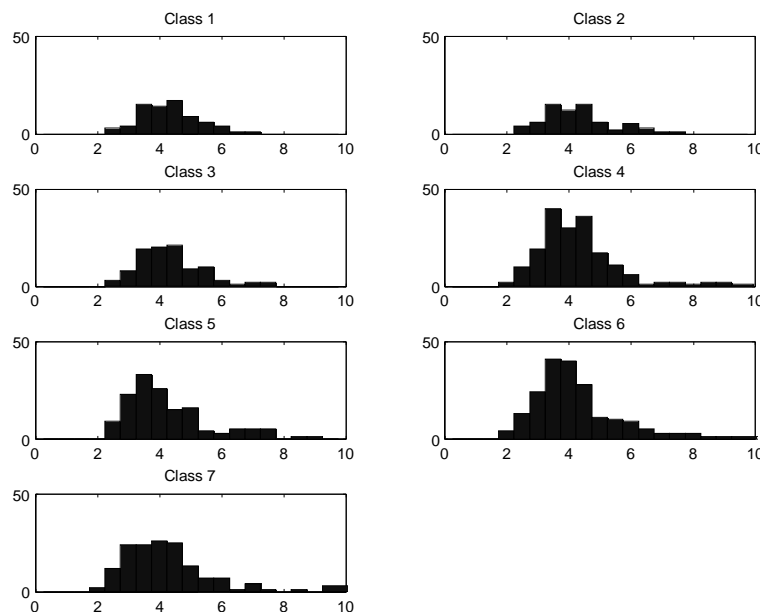
Figure 4: Histograms of point distances (Mahalanobis) within each class. Few points are closer than distance 2 to their own centre. All 20 features are used.

RESULTS

In the following we briefly compare three advanced classifiers with a simple minimum distance classifier.

MINIMUM DISTANCE CLASSIFIER

If we partition the data into a set of training data and a set of test data, we can apply a simple minimum distance classifier: To classify a feature vector $\mathbf{x}$, measure the distance $d(\mathbf{x}, \boldsymbol{\mu}_i)$ from each $\mathbf{x}$ to each of the class centres $\boldsymbol{\mu}_i$ in the training

set, and assign x to the class of the nearest centre.

The result will depend on the arbitrary partitioning of the data into training data and test data. We therefore apply k-fold validation (m-fold in Duda, Hart and Stork 2001): divide randomly the data into k equally large sets, and run the classifier k times, each time using a different test set. The estimated performance is the mean of the k error measures. The error measure is the overall percentage of misclassified cells.

We apply $k = 10$, since experiments by Norup (2005) indicated that 10 was sufficient. Since there are 917 objects in total, each set contains 91 objects with 7 objects in total left out. There is still some variation in the result, and the classifier is therefore rerun 500 times each with a different random selection of k-fold sets.

Table 4 shows the resulting confusion matrix. It shows a little confusion between classes 1 and 2, which is to be expected, and also between classes 6 and 7. Furthermore, class 3 is confused with classes 4, 5, 6, and 7, and mostly with class 6. This is as predicted by the class neighbour relationship (Fig. 3).

Estimate	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
$\widehat{C_1}$	6.6	0.6	0.0	0.0	0.0	0.0	0.0
$\widehat{C_2}$	0.7	6.1	0.0	0.0	0.0	0.0	0.0
$\widehat{C_3}$	0.0	0.0	6.2	0.1	0.4	1.1	0.3
$\widehat{C_4}$	0.0	0.1	0.3	14.7	5.1	2.3	0.4
$\widehat{C_5}$	0.0	0.1	0.9	2.4	6.7	2.8	1.9
$\widehat{C_6}$	0.0	0.0	2.0	0.9	1.8	10.8	6.6
$\widehat{C_7}$	0.0	0.0	0.2	0.0	0.5	2.6	5.7

Table 4: Average confusion matrix **C**. An element $\mathbf{C}_{ij}$ is the number of objects of true class j estimated as class i. The diagonal is the number of correctly classified samples of the test data (91 objects, 10-fold, 500 reruns).

The overall error in percent is 100 minus the ratio of correct classifications, which is the sum of the diagonal elements, to the sum of all objects, which is 91 (it is 92 in the table because of rounding). If we partition the confusion matrix into two classes, normal $\{1, 2, 3\}$ and abnormal $\{4, 5, 6, 7\}$, by adding the numbers in the four submatrices that appear, we can calculate the overall error for the two-class problem in a similar manner. Table 5 summarizes the results.

We can thus deduce

- that Mahalanobis distance provides better results than Euclidian distance

- that the minimum Mahalanobis distance classifier is slightly better than least squares

- that the best is the nearest class centre (NCC) by Norup

The table also shows that error rates for the 7-class problem are much larger than for the 2-class problem.

Classifier	2-class	7-class	Comment
LS (Norup 2005)	6.4 ± 1.9	42.9 ± 4.7	Least squares, Matlab 'mldivide'
NCC (Norup 2005)	5.1		21 neighbours, nearest centre (Euclidian)
Our min. dist.	6.2 ± 2.4	37.7 ± 4.8	Mahalanobis distance
Our min. dist.	15.5 ± 3.6	50.7 ± 5.0	Euclidian distance

Table 5: Performance. Comparison of percent misclassifications in the test set. The format of the numbers is: mean ś standard deviation.

NEAREST NEIGHBOUR WITH GA FEATURE SELECTION

Marinakis and Dounias (2006a) initially use a genetic algorithm for feature subset selection. They then use a number of variants of the nearest neighbour classification method (1-Nearest Neighbor, k-Nearest Neighbor, wk-Nearest Neighbor) in the classification phase of the proposed approach. Various experiments took place, for different feature sets selected. Features 3,4,5 and 7 were the ones most often selected for the two class problem. The overall classification error in the 2-class problem is usually smaller than 2% (some cases even smaller than 1%) in the 10- and 20-fold cross validation experiments. Features 3,4,5,7 and 14 were the ones most often selected for the 7-class problem, with the overall error found to be in the area of 3%.

NEAREST NEIGHBOUR WITH TABU SEARCH FEATURE SELECTION

In another paper, Marinakis and Dounias (2006b) propose a tabu search algorithm for the solution of the feature selection problem. The algorithm is then combined with a number of nearest neighbor based classifiers. Various experiments took place, for different feature sets selected. Features 1, 3, 4, 5 and 7 were the ones most often selected for the two class problem. The overall classification error in the 2-class problem similarly as above, is found to be usually smaller than 3-4% in the 10- and 20-fold cross validation experiments. Features 3, 4, 5, 7, 13, and 16 were the ones most often selected for the 7-class problem, with the overall error found to be in the area of 3-4%.

ANT COLONY OPTIMIZATION

In a third paper, Marinakis and Dounias (2006c) use an ant colony optimization (ACO) methodology, an approach derived from the foraging behaviour of real ants in nature. The method in fact models the problem as the search for a minimum cost path in a graph. Artificial ants walk through this graph searching for good paths. Single ants have relatively poor performance in finding good paths, but better paths are found as the emergent result of the global cooperation among ants in the colony. The ACO algorithm is then combined with a number of nearest neighbour based classifiers. Features 1,3,4,5,7 are again the ones most often selected for both, the 2-class and the 7-class problem. The overall error for the 2-class problem is around 1-2% in the 10- and 20-fold cross validation experiments (in one case is smaller than 1%). For the 7-class problem the overall error ranges between 2-4%.

SUMMARY AND CONCLUSION

The seven pap-smear classes are inherently unsharp, and even the cyto-technicians and doctors that perform a manual classification agree that there is an unavoidable overlap between classes, especially mild, moderate, and severe dysplasia.
Our main thoughts and conclusions arising from observing and comparing results, are:

- The number of examples belonging to each class in the given data set, is a matter for further discussion. What if the images were selected in a better balanced way among the 7 classes? In addition, would it be better to use a leave-one-out method in order to test the accuracy of each method against unknown data? In a real world setting, the normal cells (classes 1,2,3) are much more numerous than the cancerous ones (classes 4, 5, 6, and 7). Furthermore, there are many cells in one image, some of them overlapping, which obscures the measurements, making an automatic classification less reliable.

- It is interesting to observe that, apart from comparing the accuracy performance itself, there are approaches which use a very small subset of the initial feature set, thus reducing cosiderably the problem complexity.

- Feature selection indicates that features 3, 4, and 5 (N/C ratio, Nucleus Brightness, Cytoplasm Brightness) are generally the most capable for discriminating classes in both the 2-class and the 7-class problems for most of the hybrid intelligent schemes used for classifying the cells. Features 7 and 1 are the next two most important features (nucleus longest diameter, nucleus area). It is surprising that the nucleus area is not among the top 2 features. In practice, the dependency on the colouring method (nucleus brightness, cytoplasm brightness) is somewhat unfortunate, since the colouring method, or the dye, is not standardized; the colours may turn out differently in different hospitals.

- The nature inspired methods reached an excellent performance, fully competitive to other well-known intelligent approaches, when combined with other algorithmic data analysis approaches, in hybrid intelligent schemes.

One idea for comparing the effectiveness of the methods in the really hard classification tasks, should be to focus on the classification performance obtained in 4, 5, and 6 class separation. Furthermore, we suggest a hierarchical classification, that filters out classes 1 and 2 first, since the preliminary data analysis shows they are easy to identify.

References

[1] Dimac: Digital imaging company. URL http://www.dimac-imaging.com/.

[2] Nikolaos Ampazis, George Dounias, and Jan Jantzen. Pap-smear classification using efficient second order neural network training algorithms. In Vouros and Panayiotopoulos, editors, *Methods and Applications of Artificial Intelligence*, volume 3025 of *Lecture Notes in Artificial Intelligence*, pages 230–245. Springer, 2004. Third Helenic Conference on AI, SETN 2004, Samos, Greece, May 5-8, 2004, Proceedings.

[3] Jens Byriel. Neuro-fuzzy classification of cells in cervical smears. Master's thesis, Technical University of Denmark: Oersted-DTU, Automation, February 1999.

[4] G. Dounias, B. Bjerregaard, J. Jantzen, A. Tsakonas, N. Ampazis, G. Panagi, and E. Panourgias. Automated identification of cancerous smears using various competitive intelligent techniques. *Oncology Reports*, 15:1001–1006, 2006.

[5] George Dounias, Athanasios Tsakonas, Jan Jantzen, Hubertus Axer, Beth Bjerregaard, and Diedrich Graf von Keyserlingk. Genetic programming for the generation of crisp and fuzzy rule bases in classification and diagnosis of medical data. In *Proc. First International NAISO Congress on Neuro Fuzzy Technologies, Havana, Cuba*, page 7 pp. NAISO, ICSC Academic Press, Canada / The Netherlands, Jan 2002. cd rom paper 100027-01-GD-025.

[6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, 2. edition, 2001.

[7] J. Jantzen N. Ampazis A. Tsakonas E. Panourgias G. Dounias G. Panagi, B. Bjerregaard. A comparison among different intelligent techniques for the automated identification of cancerous smears. Presentation at 9th World Congress on Advances in Oncology, Crete, Greece, 14-16 Oct 2004,, 2004.

[8] Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. In *Proc. NiSIS 2005, Albufeira, Portugal*, pages 1 – 9, http://www.nisis.de, 2005. Nature inspired Smart Information Systems (NiSIS), EU co-ordination action, NiSIS.

[9] Yannis Marinakis and George Dounias. Nature inspired intelligent techniques for pap smear diagnosis: Ant colony optimization for cell classification. In *Proc. NiSIS 2006, Tenerife, Spain*, http://www.nisis.de, 2006. Nature inspired Smart Information Systems (NiSIS), EU co-ordination action, NiSIS.

[10] Yannis Marinakis and George Dounias. Nearest neighbor based pap-smear cell classification using tabu search for feature selection. In *Proc. NiSIS 2006, Tenerife, Spain*, http://www.nisis.de, 2006. Nature inspired Smart Information Systems (NiSIS), EU co-ordination action, NiSIS.

[11] Yannis Marinakis and George Dounias. Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. In *Proc. NiSIS 2006, Tenerife, Spain*, http://www.nisis.de, 2006. Nature inspired Smart Information Systems (NiSIS), EU co-ordination action, NiSIS.

[12] Erik Martin. Pap-smear classification. Master's thesis, Technical University of Denmark: Oersted-DTU, Automation, 2003.

[13] Jonas Norup. Classification of pap-smear data by transductive neuro-fuzzy methods. Master's thesis, Technical University of Denmark: Oersted-DTU, Automation, 2005.

[14] A. Tsakonas, G. Dounias, J. Jantzen, H. Axer, B. Bjerregaard, and D. G. v. Keyserlingk. Evolving rule-based systems in two medical domains using genetic programming. *Artificial Intelligence in Medicine*, 32(3):195–216, Nov 2004.

[15] Athanasios Tsakonas, Georgios Dounias, Jan Jantzen, and Beth Bjerregaard. A hybrid computational intelligence approach combining genetic programming and heuristic classification for pap-smear diagnosis. In *Proc. European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, EUNITE 2001*, pages 10 pp, CD–ROM. EUNITE, http://www.eunite.org, ELITE Foundation, Pascalstrasse 69, D-52076 Aachen, Germany, December 2001. ISBN 3-89653-916-7.

APPENDIX Mahalanobis Distance

If x is a random observation based on the components $x_1, x_2, \ldots, x_p$, and the components have vastly different magnitudes, then the *Euclidian distance* from the origin of the coordinate system, or the length of x, will be dominated by the larger components. Similarly, a change of unit from, say, kilometres to metres increases a component by the factor 10^3, which may dominate the distance measure.

Thus the *Mahalanobis distance* between an observation of a random variable x and its arithmetic mean value μ is measured in units of standard deviation σ (Duda, Hart, and Stork 2001),

$$d = \frac{|x - \mu|}{\sigma} \tag{1}$$

For a normal distribution, the probability is 0.68 that d will be less than 1, the probability is 0.95 that d will be less than 2, and the probability is 0.997 that d will be less than 3. Our objective is to understand how Equation (1) generalizes to p dimensions.

Figure 5 shows a cluster of 2-dimensional observations. We shall approximate the distribution by a bivariate normal distribution, the shape of which is an ellipse (Duda, Hart and Stork 2001). The ellipse is not parallel to either axis, indicating that the two variables x_1 and x_2 are correlated. The *covariance matrix* is one measure of the degree of correlation.
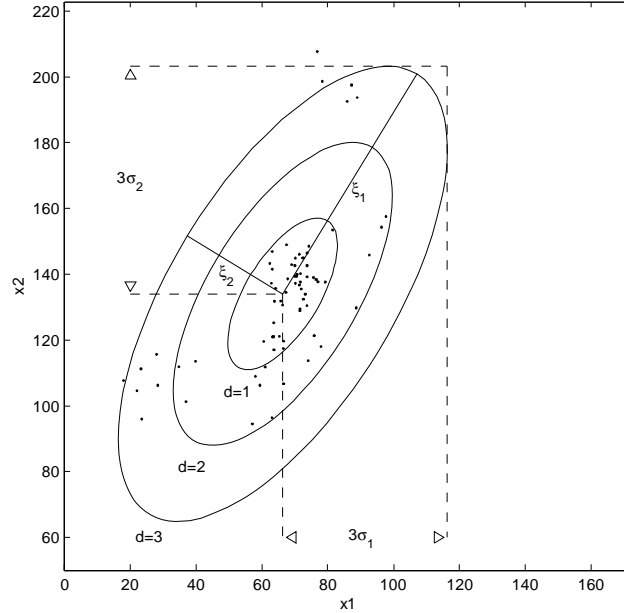

Figure 5: A cluster of observation points (dots). Each ellipsis corresponds to Mahalanobis distance d. The projection of an ellipse on an axis is a measure of the standard deviation on that axis, a property of the multivariate normal distribution.

The covariance of two random variables x and y is

$$\sigma_{xy} = \varepsilon \left[\left(x - \mu_x \right) \left(y - \mu_y \right) \right]$$

where ε is the expectation operator (the arithmetic mean) and μ_x, μ_y are the arithmetic mean values of x and y respectively. When $y = x$ the result is the variance σ_x^2 of x, the mean of the squared deviations from the mean. For a 2-dimensional observation vector $\mathbf{x}$, with the mean vector $\boldsymbol{\mu} = \varepsilon \left[\mathbf{x} \right]$, the covariance matrix is the outer product,

$$\boldsymbol{\Sigma} = \varepsilon \left[\left(\mathbf{x} - \boldsymbol{\mu} \right) \left(\mathbf{x} - \boldsymbol{\mu} \right)^t \right]$$

The diagonal holds the variances of the components σ_1^2, σ_2^2 (with $\sigma_{ii} = \sigma_i^2$). The elements outside the diagonal are a measure of the degree of dependence between components x_i and x_j ($i \neq j$). The matrix is symmetrical, that is $\sigma_{ij} = \sigma_{ji}$. If $\sigma_{ij} = 0$ the components are uncorrelated. If α is a constant and $x_j = \alpha x_i$ then $\sigma_{ij} = \alpha \sigma_i^2$. Thus the covariance is positive if x_i and x_j both increase or decrease together, and is negative if x_j decreases when x_i increases.

Figure 5 defines two orthogonal coordinate systems. One with its basis along the two axes ξ_1, ξ_2 of the ellipse; this we shall refer to as the *local* coordinate system. The coordinate axes coincide with the principal axes, such that the first axis is along the direction with the greatest variation and the second axis, perpendicular to the first, in the direction of the least variation. The other coordinate system is defined by a basis along the two axes x_1, x_2; this we shall refer to as the *global* coordinate system. The derivations in the following are in two dimensions, but all vector-matrix equations generalise to vectors in the p-dimensional space $\mathbb{R}^p$.

In the local coordinate system the cluster centre is in the origin, and we can ignore μ in Equation (1). If we scale the axes by their corresponding standard deviations, the ellipse becomes a circle, and the Euclidian distance will depend equally on the components. A point $\mathbf{x}^\xi = \left(\xi_1, \xi_2 \right)^t$ in local coordinates is then scaled to

$$\left(\mathbf{x}^\xi \right)' = \begin{pmatrix} \xi_1' \\ \xi_2' \end{pmatrix} = \begin{pmatrix} \frac{\xi_1}{\sigma_1^\xi} \\ \frac{\xi_2}{\sigma_2^\xi} \end{pmatrix}$$

where σ_1^ξ is the standard deviation in the direction of the first principal axis, and σ_2^ξ is the standard deviation in the direction of the second principal axis. The scaling operation is in matrix notation,

$$\left(\mathbf{x}^\xi\right)' = \begin{pmatrix} \frac{1}{\sigma_1^\xi} & 0 \\ 0 & \frac{1}{\sigma_2^\xi} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \mathbf{A}\mathbf{x}^\xi$$

The scaling matrix $\mathbf{A}$ defines a linear mapping of the vector $\mathbf{x}^\xi$ to the image $\left(\mathbf{x}^\xi\right)'$. The components of the scaled observations have zero mean and standard deviation 1. The Euclidian length of the scaled vector is

$$
\begin{aligned}
\left\|\mathbf{x}^\xi\right\| &= \sqrt{\left(\frac{\xi_1}{\sigma_1^\xi}\right)^2 + \left(\frac{\xi_2}{\sigma_2^\xi}\right)^2} \\
&= \sqrt{\left(\mathbf{x}^\xi\right)^t \mathbf{A}^2 \mathbf{x}^\xi}
\end{aligned}
\tag{2}
$$

This is the distance measure, but it is in local coordinates, and we wish to express it in global coordinates.

Let $(\mathbf{e}_1, \mathbf{e}_2)$ be an orthonormal basis of the local coordinate system, then any vector $\mathbf{x}$ can be defined as a linear combination in the local coordinate system,

$$\mathbf{x} = \xi_1 \mathbf{e}_1 + \xi_2 \mathbf{e}_2$$

The coordinates of $\mathbf{e}_i$ $(i = 1, 2)$ in the global coordinate system are (e_{1i}, e_{2i}) such that

$$
\begin{aligned}
x_1 - \mu_1 &= e_{11}\xi_1 + e_{12}\xi_2 \\
x_2 - \mu_2 &= e_{21}\xi_1 + e_{22}\xi_2
\end{aligned}
$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2)$ are the coordinates of the centre. In matrix notation

$$\mathbf{x} - \boldsymbol{\mu} = \mathbf{E}\mathbf{x}^\xi$$

where $\mathbf{E}$ is the matrix of column vectors $[\mathbf{e}_1, \mathbf{e}_2]$. The matrix $\mathbf{E}$ is invertible since the column vectors are linearly independent, thus

$$\mathbf{x}^\xi = \mathbf{E}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right) \tag{3}$$

Insertion into Equation (2) yields

$$
\begin{aligned}
\left\|\mathbf{x}^\xi\right\| &= \sqrt{\left(\mathbf{E}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)\right)^t \mathbf{A}^2 \mathbf{E}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)} \\
&= \sqrt{\left(\mathbf{x} - \boldsymbol{\mu}\right)^t \left(\mathbf{E}^{-1}\right)^t \mathbf{A}^2 \mathbf{E}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)}
\end{aligned}
$$

Since $\mathbf{E}$ is an orthonormal basis, its inverse equals its transpose, and we can write

$$\left\|\mathbf{x}^\xi\right\| = \sqrt{\left(\mathbf{x} - \boldsymbol{\mu}\right)^t \mathbf{E}\mathbf{A}^2\mathbf{E}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)} \tag{4}$$

The transformation $\mathbf{E}\mathbf{A}^2\mathbf{E}^{-1}$ is a similarity transformation,

$$\mathbf{C} = \mathbf{E}\mathbf{A}^2\mathbf{E}^{-1}$$

or conversely

$$\mathbf{E}^{-1}\mathbf{C}\mathbf{E} = \mathbf{A}^2 \tag{5}$$

Eigenvalues are invariant under a similarity transformation, thus $\mathbf{C}$ and $\mathbf{A}^2$ have the same eigenvalues. A solution is to choose $\mathbf{C}$ as the inverse of the covariance matrix $\boldsymbol{\Sigma}$ in the global coordinate system,

$$\mathbf{C} = \boldsymbol{\Sigma}^{-1}$$

and $\mathbf{E}$ the matrix of its column eigenvectors. By that choice Equation (5) results in a diagonal matrix $\mathbf{A}^2$ with the eigenvalues of $\boldsymbol{\Sigma}^{-1}$ in the diagonal. Insertion into Equation (4) yields the *Mahalanobis norm*

$$\left\|\mathbf{x}^\xi\right\| = \sqrt{\left(\mathbf{x} - \boldsymbol{\mu}\right)^t \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)} \tag{6}$$

In general, the *Mahalanobis distance* between two points $\mathbf{x}$ and $\mathbf{y}$ is

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^t \, \Sigma^{-1} (\mathbf{x} - \mathbf{y})}$$

By Equation (6), the set of points lying in the same Mahalanobis distance r from the centre $\boldsymbol{\mu}$ satisfies the equation

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^t \, \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{7}$$

which is a quadratic form with Σ^{-1} symmetric; in other words, the equation for an ellipse (hyperellipsoid in general) with principal axes along the eigenvectors of Σ^{-1}. By definition, an eigenvector $\mathbf{e}$ of the covariance matrix Σ satisfies the equation

$$\Sigma \mathbf{e} = \lambda \mathbf{e}$$

where λ is its corresponding eigenvalue. Premultiplying by Σ^{-1},

$$\mathbf{e} = \Sigma^{-1} \lambda \mathbf{e}$$

or

$$\frac{1}{\lambda} \mathbf{e} = \Sigma^{-1} \mathbf{e}$$

which shows that $1/\lambda$ is an eigenvalue of the inverse covariance matrix, and Σ and Σ^{-1} share the same eigenvectors. Thus the axes of the ellipse in Equation (7) are defined by the eigenvectors of the covariance matrix Σ, with the extent $\sqrt{\lambda_1}$ corresponding to the local standard deviation σ_1^ξ, and $\sqrt{\lambda_2}$ corresponding to the local standard deviation σ_2^ξ.

In order to draw an ellipse in two dimensions, given a fixed Mahalanobis distance r, we can start from a unit circle defined by the parametric form

$$\left(\begin{array}{c} x_1^0 \\ x_2^0 \end{array} \right) = \left(\begin{array}{c} \cos \theta (k) \\ \sin \theta (k) \end{array} \right)$$

where we let the angle $\theta (k)$ traverse a period of length 2π at suitable sampling points k, for example

$$\theta (k) = \{0, 5, 10, \dots, 360\} \, (2\pi / 360)$$

The scaling matrix

$$r \mathbf{A} = \left(\begin{array}{cc} r \sqrt{\lambda_1} & 0 \\ 0 & r \sqrt{\lambda_2} \end{array} \right)$$

scales the axes of the circle to the extent of the principal axes of the ellipse. The rotation matrix

$$\mathbf{E} = \left(\begin{array}{cc} e_{11} & e_{12} \\ e_{21} & e_{22} \end{array} \right)$$

performs a change of basis. The first column is the unit length eigenvector $\mathbf{e}_1$ corresponding to λ_1 and the second column is the unit length eigenvector $\mathbf{e}_2$ corresponding to λ_2. Finally a translation from the origin to $\boldsymbol{\mu}$ will place the ellipse at the centre of the cluster. Thus, in global coordinates

$$\left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right) + r \mathbf{E} \mathbf{A} \left(\begin{array}{c} x_1^0 \\ x_2^0 \end{array} \right)$$

This is an ellipse with centre $\boldsymbol{\mu}$ and axes determined by the eigenvectors in $\mathbf{E}$. The length of the axes are determined by the diagonal matrix $\mathbf{A}$, holding the square roots of the eigenvalues of the covariance matrix. The axes are multiplied by the desired distance r.