Pap-smear Benchmark Data For Pattern Classification

Jan Jantzen[1], Jonas Norup[1], George Dounias[2], Beth Bjerregaard[3]

[1]Technical University of Denmark
Oersted-DTU, Automation
Building 326, DK-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

[3]Herlev University Hospital
Dept. of Pathology
Herlev Ringvej 75, DK-2730 Herlev, Denmark

ABSTRACT: This case study provides data and a baseline for comparing classification methods. The data consists of 917 images of Pap-smear cells, classified carefully by cyto-technicians and doctors. Each cell is described by 20 numerical features, and the cells fall into 7 classes. A basic data analysis includes scatter plots and linear classification results, in order to provide domain knowledge and lower bounds on the acceptable performance of future classifiers. Students and researchers can access the database on the Internet, and use it to test and compare their own classification methods.

KEYWORDS: Linear regression, clustering, false negative, cancer.

INTRODUCTION

The term *Pap-smear* refers to samples of human cells stained by the so-called *Papanicolau* method. A specimen of cells is *smeared* onto a glass slide and coloured, making it easier to examine the cells under a microscope for any abnormalities indicating a pre-cancerous stage.

We wish to publish a database of Pap-smear cell images with image measurements, and we propose to use the database as a benchmark case for classification methods. Since the data origin is human we find it relevant to host the database under NiSIS, Nature-inspired Smart Information Systems (EU co-ordination action, contract 13569), with special relevance to the focus group Nature-Inspired Data Technology. It will thus be accessible on the Internet for anyone (http://fuzzy.iau.dtu.dk/download/smear2005). This paper is to be included, acting as a first common sense analysis of the data, from which students and researchers can quickly gain some domain knowledge. It is hoped that future classification results will be added to the database, so that the number of benchmark results will increase for the benefit of solid comparison.

The database consists of 917 samples (Fig. 1) distributed unevenly in 7 different classes. Each sample is described by 20 features extracted from images of single cells. The data were collected at the Herlev University Hospital by means

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

of a digital camera and microscope. Skilled cyto-technicians and doctors manually classified each cell into one of the 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 purpose of the Papanicolau method is to diagnose pre-cancerous cell changes before they progress to *invasive carcinoma* (Meisels & Morin, 1997). It does take a skilled cyto-technician to differentiate between the different kinds of cells. Every glas slide can contain up to 300,000 cells and it is a time consuming job, so semi-automatic systems have been developed to aid the cyto-technicians (Papnet in for instance Koss, 2000; Aphrodite, Dimac).

In order to provide a baseline for future benchmark studies, we present below a common sense analysis. For example, executing a linear, simple classifier yields a worst case result, which provides a lower bound for future benchmark studies. The reasoning is that any classifier must be superior to the linear case to be acceptable. In the same spirit, a random classification also yields a lower bound on the performance measure. These lower bounds may later turn out to be conservative, but they do provide numbers for clear and convenient comparison.

METHOD AND DATA

Ideally specimens are taken from several areas of the cervix (Fig. 2). The specimens most often contain cells from the *columnar epithelium* and the *squamous epithelium*. The columnar epithelium is located in the upper part of the cervix, and the squamous epithelium in the lower part. Between these two is the *metaplastic epithelium*, also called the transformation zone or the *squamo-columnar junction*.

In the *squamous* epithelium there are 4 layers of cells. The cells form at the *basal* layer and while maturing they move up through the *parabasal* layer, the *intermediate* layer, and finally the *superficial* layer. The cells in the basal layer divide and deliver cells to the layers above it. While the cells mature and move through the layers, they change shape, color and other characteristics. When the cells reach the superficial layer they are rejected and replaced by the cells coming from below. The basal layer has small round cells with a relatively big nucleus and small cytoplasma. When maturing, the nucleus becomes smaller and the cytoplasma becomes larger. The shape of the cells becomes less round the more mature they are.

The *columnar* epithelium only contains a single layer of cells containing columnar cells and reserve cells. The reserve cells divide into new reserve cells and new columnar cells. In normal columnar epithelial cells, the nucleus is located at the bottom of the cytoplasm. When viewed from the top, the nucleus seems larger. When viewed from the side, the cytoplasm seems larger.

The *metaplastic* epithelium consists of reserve cells from the columnar epithelium. When the cells have matured fully in the metaplastic epithelium, they look like the cells found in the squamous epithelium.

In *dysplastic* cells, the genetic information is somehow changed, and the cell will not divide as it should. This is a pre-cancerous cell. Depending on which kind of cell that divides incorrectly, it is given diagnoses like *dysplasia* and *carcinoma in situ*. The term 'plasia' means growth, and 'dysplasia' means disordered growth. The dysplastic cells are divided into *mild*, *moderate* and *severe* dysplastic. The grading is determined from the likelihood of the cells later on turning into malignant cancer cells. A high amount of the mild dysplastic cells will disappear without becoming malignant, whereas

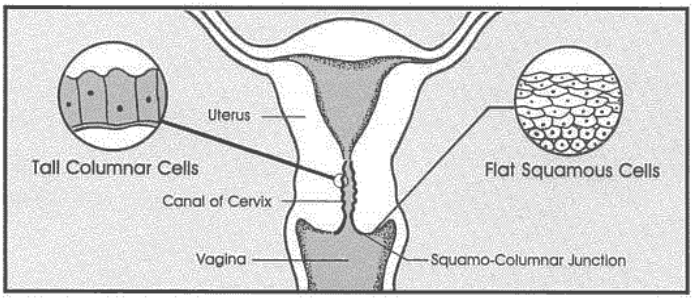

Figure 2: The location of columnar cells and squamous cells. Squamous cells develop bottom-up from the basal, to the parabasal, the intermediate, and the superficial layer. Dysplasia develops in the squamo-columnar junction. (Indman, http://www.gynalternatives.com/cervix_structure.htm)

severe dysplastic cells likely will turn into malignant cells. Squamous dysplastic cells generally have larger and darker nuclei and tend to cling together in clusters. In severe dysplasia, nuclei are large, with dark granules and usually deformed. The cytoplasm is small and dark compared to nuclei.

Cervical carcinoma is the most common malignancy of the female genital system. Carcinogenesis is a long-lasting process, which begins from normal epithelium, that becomes dysplastic, evolves to carcinoma in situ, and then to cancer. The long time interval between the stages, allows the possibility of an early diagnosis, with complete cure.

Data description

The Pap-smear database is the latest of two versions built by the Herlev University Hospital. The images were prepared and analyzed by the staff at the hospital using a commercial software package CHAMP (Dimac) for segmenting the images. The cells were selected, not to collect a natural distribution, but to make a good collection of the important classes. There are at least twice as many abnormal cells as normal cells; the precise composition of the classes is shown in TABLE 1.

The features, defined in the Appendix, were extracted from the images using Matlab programs written by Martin (2003). They are extracted from a combination of the segmented and non-segmented cell images. The data have been analysed in two Master's projects (Martin, 2003; Norup, 2005).

Performance measures

From a medical point of view, it is worse to misclassify an abnormal cell as normal, than oppositely. Since we are looking for *abnormal* cells this is called a *positive* finding, while a *normal* cell is called a *negative* finding. An abnormal cell misclassified as normal is called a *false negative* finding, and it is important that a classifier, whether human or automatic, minimises the number of false negatives. By analogy, a normal cell misclassified as abnormal is called *false positive*,

Estimated class	Actual class N	P
$\widehat{N}$	TN	FN
$\widehat{P}$	FP	TP

The above are definitions of true negative (TN), false negative (FN), true positive (TP), and false positive (FP). For instance, if a cell estimated as negative $(\widehat{N})$ is actually positive (P), the estimate is a false negative (FN, upper right corner).

The only possibility of correcting a false negative diagnosis is for the 'patient' to turn up at the next screening, which is voluntary. The consequence of a false positive diagnosis is, that the 'patient' is called back for a repeat of the Pap-test, where the mistake is likely discovered. Denoting the number of false negative cells by FN, the number of true negative cells by TN, the number of false positive cells by FP, and the number of true positive cells by TP, we define two relative performance measures in percentages

$$FN\% = \frac{FN}{TP+FN} * 100 \quad (1)$$

$$FP\% = \frac{FP}{TN+FP} * 100 \quad (2)$$

The denominator $TP + FN = P$, where P is the number of truly positive cells. Thus the false negative percentage $FN\%$ ranges from 0, when all positive cells have been classified correctly $(FN = 0)$, to 100 percent, when all positive

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.

cells have been classified incorrectly ($TP = 0$). Analogously with the $FP\%$. Furthermore we use a third performance measurement, the overall error,

$$OE\% = \frac{FN + FP}{TP + FN + TN + FP} * 100 \tag{3}$$

This is the percentage of misclassified cells $FN + FP$ relative to the total number of cells. For a 2-class problem with $P \neq N$, this number cannot be 0 and it cannot be 100. The following relationships apply,

$$\begin{aligned} FN + TP &= P \\ TN + FP &= N \\ N + P &= U \end{aligned}$$

Here U (for universe) is the total number of cells. The three performance measures $FN\%, FP\%$, and $OE\%$ are discrete functions, since a cell is either classified correctly or misclassified. For instance, if a cell is reclassified from false negative to true positive, the FN will decrease by 1 and TP will increase by 1, causing a jump in the values of the performance measures. The three measures only make sense for 2-class problems.

A fourth alternative measure, which can be applied to n-class problems, is the *root-mean-squared error*,

$$RMSE = \sqrt{\sum_{i=1}^{K} \frac{(y_i - \widehat{y}_i)^2}{K}}$$

The $RMSE$ can be interpreted as the normalised distance between the classifier model output $\hat{y}$ and the true class y for K cells in a batch.

The Pap-smear data fall into 7 classes, but a minimal requirement is to separate normal from abnormal, which is a 2-class problem. To recall, the normal classes are $\{1, 2, 3\}$, while the abnormal classes are $\{4, 5, 6, 7\}$.

To make the best use of the amount of data, and to neutralise the problem of which data to select as test data and training data, we use *k-fold cross-validation* (see for example Duda, Hart & Stork, 2001). The whole set of U cells is divided into k segments, randomly distributed. Temporarily one segment is hidden and the remaining $k - 1$ segments used as training data to build a model. The test is then performed on the hidden segment. This process is repeated k times, hiding a different segment each time, and the total performance measure is calculated as the average of the k tested models.

The fraction $k - 1/k$ is the proportion of training data to the total amount of data. When k is large, the amount of training data is large. When $k = 2$ the training data and test data have equal size, and when $k = U$ each segment consists of one cell. The computational load grows with the magnitude of k, but an experiment with different values of k indicated that $k = 10$ is sufficient in our case (Norup, 2005).

Since the k segments are randomly selected, the results may vary from run to run. Therefore we apply *reruns*, which means running the same tests R times. The final performance measure, the average over R reruns, smooths out the statistical variation, thus making the performance measures more consistent.

Classifier

The simplest classifier is a linear model that minimises the squared error, a *least squares* approach. It corresponds to a neural network with linear activation functions. It is easy to implement in Matlab using the backslash operator (*mldivide*), and it executes fast. Mathematically the model is a linear combination of the features $x_1, x_2, \ldots, x_p$,

$$\begin{aligned} \hat{y} &= w_1 x_1 + w_2 x_2 + \ldots + w_p x_p + w_0 \\ &= \sum_{j=1}^{p} w_j x_j + w_0 \end{aligned}$$

With an eye to the Matlab implementation we rearrange, generalise, and write it in matrix notation,

$$\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} & 1 \\ x_{21} & x_{22} & \ldots & x_{2p} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \\ w_0 \end{bmatrix} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \vdots \\ \vdots \\ \widehat{y}_n \end{bmatrix}$$

Or,

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \tag{4}$$

Or objective is to find the weight vector $\mathbf{w}$ that satisfies (4), where $\mathbf{X}$ is the augmented matrix of all the training points (dimension n-by-$(p+1)$, $n = U * (k-1)/k$) and for $\hat{\mathbf{y}}$ we insert the known classes of the training set. If $\mathbf{X}$ were nonsingular, we could write $\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$ and obtain a solution at once. However, $\mathbf{X}$ is usually rectangular with more rows than columns, so $\mathbf{w}$ is overdetermined and no exact solution exists. However, we can seek a weight vector $\mathbf{w}$ that minimizes the sum-of-squared-errors, and this is exactly what Matlab's *mldivide* (matrix left divide) function does. It should be noted, however, that the least squares approach does *not* guarantee a separating hyperplane, even if one exists (Duda, Hart & Stork, 2001).

RESULTS AND DISCUSSION

To begin with, we note that the medical experts emphasize the importance of the ratio between nucleus area and cytoplasm area (N/C ratio). As a first step, one can list the mean value and standard deviation (Std) of the raw N/C ratios (data column D),

Class	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Mean	0.01	0.03	0.35	0.27	0.38	0.49	0.60
Std	0.01	0.01	0.10	0.10	0.12	0.14	0.13

Classes 1 and 2 have a small N/C ratio; classes 3, 4, 5, and 6 medium N/C ratio; and class 7 has a large N/C ratio. This corresponds more or less with the statements from the medical experts. The table also suggests that classes 1 and 2 can be separated from the rest using N/C ratio, and that class 3 may be easily confused with class 5.

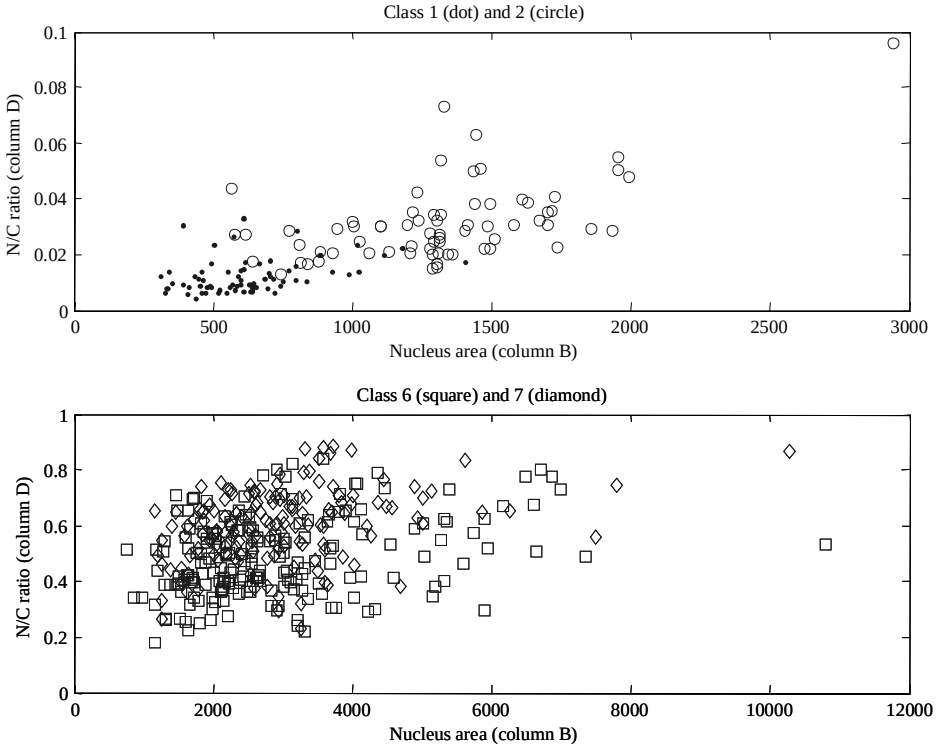

Figure 3: Scatter plots of nucleus area versus N/C ratio. Classes 1 and 2 have little overlap, while 6 and 7 have a large overlap.

If we include the nucleus area as well as a next step, scatter plots (Fig. 3) confirm that classes 1 and 2 indeed are separated from the rest, and also from each other more or less. Oppositely, classes 6 and 7 have large overlap according

to the figure. We notice in passing that there is at least one outlier in the upper plot (top right corner), and possibly two in the lower plot.

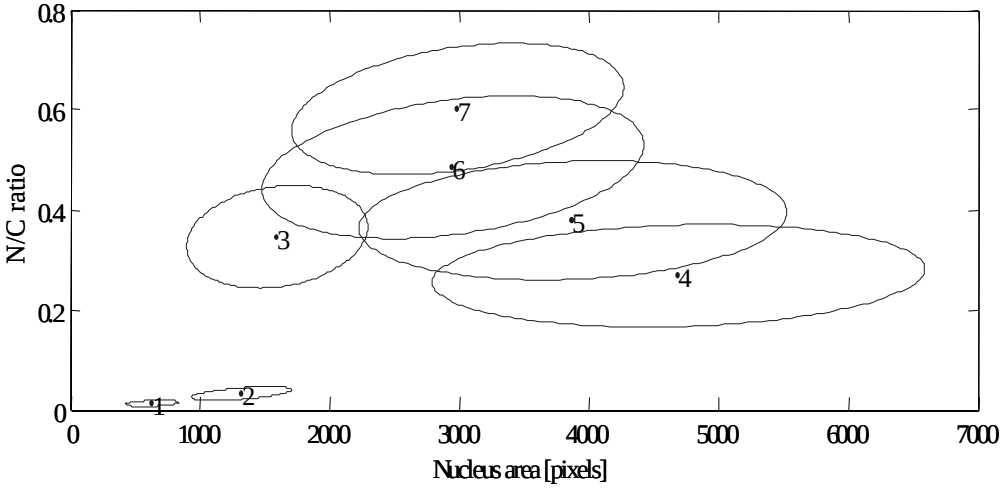

Figure 4: Class centers and their standard deviations along the principal axes.

By means of the covariance matrices and principal components of each class (e.g., Duda, Hart & Stork, 2001), we can get an idea of the shape and overlap of each class. Figure 4 confirms that classes 1 and 2 lie separated from the rest. Classes 4, 5, 6, and 7 successively overlap, they have decreasing nucleus area, and the N/C ratio increases. The transition from one to the next class is gradual. Class 3 can clearly be separated from 1 and 2, but it has some overlap with class 6. The ellipses extend only one standard deviation in the direction of the major and minor axes, so the overlap is worse than the plot indicates.

The 2-class problem

There are three ways to make a, shall we call it *naive* classifier. All three provide performance measures, that certainly are lower bounds on the acceptable performance. Assume firstly that all cells are estimated as positive. Since we know the composition of the database, we have $TP = P = 675$, $FN = TN = 0$, $FP = N = 242$. Insertion into equations (1, 2, 3) yields $FN\% = 0$, $FP\% = 100$, and $OE\% = 26$. Assume secondly that *all* cells are estimated as negative. We have then $TN = N = 242$, $FP = TP = 0$, $FN = P = 675$. As a result $FN\% = 100$, $FP\% = 0$, and $OE\% = 74$. However naive, the results do show at this point that the overall error should never exceed 26 %, and that we automatically achieve better results by leaning towards positive, due to the distribution of the data.

Assume thirdly that the classifier is a random selection from a uniform distribution. In other words, each cell has a 50 percent chance of being estimated as positive and a 50 percent chance negative. As a result, half of the negative cells are expected classified correctly, and the other half wrongly. We thus have $TN = FP = 0.5 * 242 = 121$. Likewise with the positive cells, and we have $TP = FN = 0.5 * 675 = 337.5$. In practice the decimal 0.5 is not feasible, but we will ignore that for the sake of clarity. Insertion yields $FN\% = 337.5 * 100/675 = 50$, $FP\% = 121 * 100/242 = 50$, $OE\% = (337.5 + 121)/917 = 50$. This classifier is worse in terms of overall error, but it does show that all three performance measures can be equal. In fact this happens whenever $FN = \alpha * P$ and at the same time $FP = \alpha * N$, where α is some constant of proportionality; in the case of the random classifier $\alpha = 0.5$. The results indicate there is a tradeoff between FN and FP: if one is high, then the other is low.

We now progress to the linear least-squares classifier. To begin with, our objective is to separate the cells into normal and abnormal, the 2-class problem. In accordance with (4) the matrix **X** holds the training data, **w** is the weight vector that

Measure	mean	std	min	max
$FN\%$	1.2	1.3	0.0	7.3
$FP\%$	20.7	6.3	4.0	36.0
$OE\%$	6.4	1.9	1.1	11.7
$RMSE$	0.25	0.02	0.21	0.30

Table 2: Results from least squares model, 2 class problem, 10-fold validation, 50 reruns.

we wish to find, and the right hand side is the class, a vector of ones and twos, where 1 indicates normal and 2 indicates abnormal, during training. With 10-fold cross validation and 50 reruns, the results by Norup (2005) are displayed in TABLE 2. It is remarkable that the overall error $OE\%$ is only 6.4%. The false positive percentage is high, while the false negative percentage is low, the latter being the more important from a medical viewpoint.

It is possible to shift the false negative and false positive percentages. The training result is a weight vector **w**, and the estimated class is found by inserting test data $\mathbf{X}^*$ into (4). The estimates $\widehat{y}$ are real numbers, not just integers, and one has to decide where to split, in order to assign the class membership. When the model is trained using arbitrary class numbers 1 and 2, it is natural to choose a threshold in the middle, $\tau = 1.5$, such that any $\widehat{y} \geq \tau$ is assigned to class 2, otherwise class 1. Shifting τ upwards, however, implies more negative estimates and less positive. Consequently the $FN\%$ will increase, and the $FP\%$ will decrease, and vice versa.

The 7-class problem

Stepping up to the problem of classifying all 7 classes, we notice that the concepts of false negative and false positive are not applicable, so we shall just calculate the overall error and the root-mean-squared error.

It is straight forward to generalise the linear least-squares classifier to 7 classes. One possibility is to number the classes $\{1, 2, \ldots, 7\}$ and fill the vector on the right hand side of (4) accordingly during training. That way we require that the trained weight vector **w** applies to all classes. It is more accurate, however, to train 7 classifiers, i.e. 7 weight vectors concatenated to form a weight matrix $\mathbf{W}$. During training we insert a matrix of known classes on the right hand side of (4). Each row (1-by-7), corresponding to each row of $\mathbf{X}^*$, is a boolean vector with a 1 in the position of the known class, zeros elsewhere. The weight matrix is now found in the usual way using *mldivide*.

With 10-fold cross validation and 50 reruns, the results by Norup (2005) are displayed in TABLE 3. When comparing with the 2-class model, we see that the overall error has increased, but not that much. The $RMSE$ has increased by almost 5 times. An increase is to be expected, since the 7-class problem is much harder than the 2-class problem.

A *confusion* matrix is a generalisation of the false negative / false positive concept. It shows (TABLE 4) that classes 1 and 2 are classified more or less correctly, but with a little overlap. Class 3 is more or less correctly classified, but there is some confusion with class 6. Classes 4, 5, 6, and 7 successively overlap, especially class 5 is mostly misclassified as class 4. These results confirm the earlier observations from plots (Figures 3 and 4).

Summary of results

From this basic analysis we have gained qualitative and quantitative insight. We have established

- that classes overlap, but 1 and 2 are separated more or less from the rest;
- that there is a gradual, successive transition through classes 4, 5, 6, and 7;
- that class 3 is somewhat 'odd', and mixes with especially classes 5 and 6;
- that the separation into normal $\{1, 2, 3\}$ and abnormal $\{4, 5, 6, 7\}$ therefore is unfortunate;
- that an acceptable overall error $OE\%$ must be less than 6.4% for the 2-class problem, and less than 7.9% for the 7-class problem; and
- that a false negative rate $FN\%$ down to 1.2% for the 2-class problem is feasible, it could even go as low as 0%, but there is a tradeoff with the false positive rate $FP\%$.

These results are achieved using simple methods. The sensitivity of the results has not been investigated in detail, that is, whether the results change a lot if there is only a small change in the underlying data. On the other hand, the results can be regarded as quite reliable, since we have avoided scaling of data and feature selection. It is usual practice to scale or standardise data before modelling, but by virtue of the linear model, this was not necessary. With 20 features, feature selection is usually necessary. We have, implicitly in our plots selected only two features, but the linear classifier operates on the full set of 20 features, and thus avoids another degree of complexity.

Measure	**mean**	**std**	**min**	**max**
$OE\%$	7.9	2.7	1.1	18.5
$RMSE$	1.13	0.13	0.76	1.57

Table 3: Results from least squares model, 7 class problem, 7 models, 10-fold validation, 50 reruns.

Estimated	Actual class						
class	1	2	3	4	5	6	7
1	76.1	23.9	0	0	0	0	0
2	21.4	69.0	0	0	0.6	0.4	0
3	2.5	1.4	74.7	0.6	10.3	9.3	4.9
4	0	5.6	0.9	88.0	51.8	14.1	5.0
5	0	0	0	3.0	13.4	4.2	2.3
6	0	0.1	19.0	7.2	16.3	43.4	24.7
7	0	0	5.4	1.2	7.6	28.6	63.1
Sum	100	100	100	100	100	100	100

Table 4: Confusion matrix. Least squares model, 7 class problem, 10-fold validation, 50 reruns. Each column shows the percentage correctly classified (on diagonal) and misclassified (off diagonal).

CONCLUSION

The objective of this case study is to provide a good database for benchmarking of classification methods. We have provided data, which have been selected and examined as carefully as possible in the hospital. The data are now on the Internet for public use, and the present paper provides a first, basic analysis so that students and researchers can get a head start, and avoid making any fundamental mistakes. The next step is to apply nonlinear classification methods. We are interested in hearing about new results, and we would like to include references to other studies in the database, in the hopes that the quality and accuracy of future classification results will increase.

References

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

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

[3] L. Koss. *Artificial neural networks in biomedicine*, pages 51–68. Perspectives in Neural Computing. Springer, 2000.

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

[5] A. Meisels and C. Morin. *Cytopathology of the uterus*. ASCP Press, 2. edition, 1997.

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

APPENDIX: Data definitions

Each cell in the Pap-smear database is associated with 20 features, summarised in TABLE 5. The heading 'Column' refers to the column in the Excel spreadsheet (http://fuzzy.iau.dtu.dk/download/smear2005). Each feature is defined below. The shorthand notation *N* stands for nucleus and *C* stands for cytoplasm. The letter in square brackets [·] refers to the name of the column in the data spreadsheet (Excel), where column *A* is the cell identifier, which again refers to the name of the file with the image of the cell. Calculations were executed in Matlab (Martin, 2003); the programs are available for public use, http://fuzzy.iau.dtu.dk/download/martin2003.

- *N area [B], C area [C]*. Count of N and C pixels respectively, 1 pixel = (0.201 micrometer) squared.
- *N/C ratio [D]*. Size of nucleus relative to cell size, $NC_{ratio} = N_{area}/(N_{area} + C_{area})$
- *N brightness [E], C brightness [F]*. The average perceived brightness, a function of colour wavelengths: $Y = 0.299 * red_{\mu} + 0.587 * green_{\mu} + 0.114 * blue_{\mu}$ Here $(red_{\mu}, green_{\mu}, blue_{\mu})$ are average measured intensities of the colours, and these are weighted by the perceived brightness of the human eye.
- *N longest diameter [H], C longest diameter [L]*. The diameters N_{long} and C_{long} of the circles circumscribing the nucleus and the cytoplasm respectively.

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 5: Summary of the 20 features in the database (Excel spreadsheet)

- *N shortest diameter [G], C shortest diameter [K]*. The diameters N_{short} and C_{short} of the inscribed circles of the nucleus and cytoplasm respectively. The diameter is approximated by the sum of two distances S_1 and S_2 perpendicular to the line L defined in Fig. 5.
- *N elongation [I], C elongation [M]*. The ratio between the shortest diameter and the longest, $N_{elong} = N_{short}/N_{long}$, and $C_{elong} = C_{short}/C_{long}$
- *N roundness [J], C roundness [N]*. The ratio between the actual area and the area of the circumscribed circle, $N_{circle} = \pi/4 * N_{long}^2 \implies N_{roundness} = N_{area}/N_{circle}$ and $C_{circle} = \pi/4 * C_{long}^2 \implies C_{roundness} = C_{area}/C_{circle}$
- *N perimeter [O], C perimeter [P]*. Length of perimeter of nucleus and cytoplasm respectively.
- *N relative position [Q]*. A measure of how well the nucleus is centered in the cytoplasm. The distance between nucleus centre (x_n, y_n) and cytoplasm centre (x_c, y_c), $Npos = 2a\sqrt{(x_n - x_c)^2 + (y_n - y_c)^2}/C_{long}$. The scaling factor $a = 0.201\ \mu\mathrm{m}/pixel$.
- *N maxima [R], N minima [S], C maxima [T], C minima [U]*. Counts of the number of pixels with the maximum / minimum value within a 3 pixel radius. Intended for measuring the degree of keratinisation.

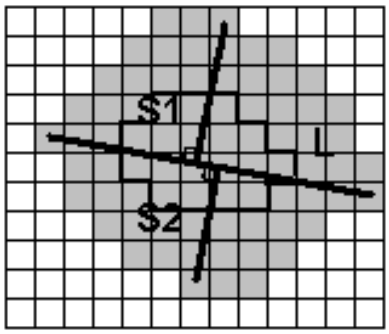

Figure 5: Definitions of cytoplasm longest diameter L, and auxiliary lines $S1$ and $S2$ for approximating the shortest diameter (adopted from Martin, 2003).