OPSEARCH **38**(2001), 443-455

AN ESTIMATION ALGORITHM USING DISTANCE CLUSTERING OF DATA

ADI BEN-ISRAEL AND YURI LEVIN

ABSTRACT. The problem is to predict a value $y \in Y$ (output, class) from an observed value of a vector $\mathbf{x} \in X$ (predictors, inputs, attributes), the relations between y and $\mathbf{x}$ given in (empirical) data $\mathcal{D} = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$, listing N observed pairs. We propose an estimation algorithm using a classification of $\mathcal{D}$ in clusters $\{\Omega_1, \ldots, \Omega_m\}$, based on a distance function in $X \times Y$. For each cluster Ω_i compute the centroid $\overline{y_i}$ of y, and denote the X–projection of Ω_i by Ω_i^X. Prediction of y given $\mathbf{x} \in X$ is done by assigning the point $\mathbf{x}$ to a nearest projected cluster, say Ω_i^X, and using $\overline{y_i}$ as estimate for y. Numerical tests show the method, in its basic general form, to give accurate predictions for well–known data sets.

1. INTRODUCTION

A variable y (*dependent variable, class membership*, or *output*) is assumed to depend in some fashion on n variables $\mathbf{x} = (x_1, \ldots, x_n)$ (*independent variables*, *predictors*, *attributes*, or *inputs*). The variables y and $\mathbf{x}$ take values in sets Y and $X = X_1 \times X_2 \times \cdots \times X_n$, respectively, where Y and the X_i are real intervals or finite sets, in particular $\{0, 1\}$.

The relation between y and $\mathbf{x}$ is known only through an empirical *data set*

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$$

consisting of N previously observed points $(\mathbf{x}, y) \in X \times Y$.

The problem is to predict the value of y corresponding to an observed value of $\mathbf{x} \in X$. This problem appears in many areas and contexts, including statistical estimation, regression, learning theory, and artificial intelligence. In typical applications, the values of $\mathbf{x}$ can be observed or measured with low cost, but the exact determination of y is complicated and costly, hence the need to predict y given $\mathbf{x}$.

For example, in typical medical applications y takes two values (e.g. 0 or 1), denoting respectively the absence or presence of disease. The values $\mathbf{x} = (x_1, \ldots, x_n)$ come from diagnostic tests. The determination of y dictates the course of treatment, in particular, $y = 1$ may result in additional tests or even surgery. In general, the two possible errors:

Key words and phrases. Cluster analysis, estimation, prediction, data analysis, diagnostics.

type 1 (false positive): declaring $y = 1$ when it is $= 0$, and
type 2 (false negative): declaring $y = 0$ when it is $= 1$,

differ in their consequences, with type 2 more serious.

Many data sets are available in the public domain. A good repository of machine learning databases is available from the University of California–Irvine (UCI), see [17].

We assume that suitable *distance functions*, denoted by d_X and d_Y, are defined on X and Y, respectively[1]. For example, if $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}$ we can use

$$d_X(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\| \,, \tag{1}$$

$$d_Y(y_1, y_2) = |y_1 - y_2| \,, \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$, restricted to X.

The distances d_X and d_Y can be combined to form a distance function d on $X \times Y$ in several ways. We use the distance

$$d\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha\, n\, d_Y^2(y_1, y_2)} \,, \tag{3}$$

where:

- the product $\alpha\, n$ is a scaling factor, with
- n, the dimension of the vector $\mathbf{x}$, used to equalize the effects of $\mathbf{x}$ and the scalar y, and the
- parameter $\alpha \geq 0$ controls the relative importance of the y–component for the distance d, see § 4 below.

We propose a method for predicting y given $\mathbf{x} \in X$, using a classification of the data $\mathcal{D}$ into clusters $\{\Omega_1, \ldots, \Omega_m\}$. The i th–cluster Ω_i has a centroid $\boldsymbol{\mu}_i = (\overline{\mathbf{x}_i}, \overline{y_i})$ of its $\mathbf{x}$ and y components, respectively. Each cluster Ω_i is computed, recursively, as all points $(\mathbf{x}, y)$ in $\mathcal{D}$ that are closer to $\boldsymbol{\mu}_i$ than to any other mean $\boldsymbol{\mu}_j$, see § 2 for details. The X–projection of a cluster Ω_i is the set $\Omega_i^X \subset X$ consisting of all vectors $\mathbf{x}$ with $(\mathbf{x}, y) \in \Omega_i$.

Having thus classified the data $\mathcal{D}$, any point $\mathbf{x} \in X$ can be assigned to some projected cluster Ω_i^X with closest X–mean $\overline{\mathbf{x}_i}$. The corresponding Y–mean, $\overline{y_i}$, is then used as prediction of y. If Y is a discrete set, the values of $\overline{y_i}$ need discretization. In particular, if $Y = \{0, 1\}$, a cut–off value p is used to infer

$$y = \begin{cases} 1 & \text{if } \overline{y_i} > p \\ 0 & \text{if } \overline{y_i} \leq p. \end{cases} \tag{4}$$

2. The nearest mean reclassification algorithm

A well known clustering technique is the *Nearest Mean Reclassification Algorithm* (NMRA) [9] or *Iterative Self–Organizing Data Analysis* (ISODATA), [19]. The number of clusters m is specified. If the class variable is

[1] In the absence of linear structure on X and Y, the distance functions d_X and d_Y are not associated with norms.

discrete, $y \in \{1, \ldots, l\}$ then m must be $\geq l$. The k th–iteration begins with a *partition* (or *clustering*)

$$\Omega^k = \{\Omega_1^k, \ldots, \Omega_m^k\}$$

of the given data $\mathcal{D} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$. The initial partition Ω^0 is selected randomly.

A *center* (mean, centroid) $\boldsymbol{\mu}_i^k$ of each Ω_i^k is computed (Algorithm 1, step 2), and points $\mathbf{v}_j \in \Omega_i^k$ are reassigned to other clusters if closer to their centers than to $\boldsymbol{\mu}_i^k$, (Algorithm 1, step 4). The algorithm stops (if no reassignments are possible) or proceeds with the new partition $\Omega^{k+1} = \{\Omega_1^{k+1}, \ldots, \Omega_m^{k+1}\}$ reflecting the reassignments.

The general iteration is described as follows:

Algorithm 1 (Nearest Mean Reclassification Algorithm. Iteration k)**.**

Given a partition $\Omega^k = \{\Omega_1^k, \ldots, \Omega_m^k\}$ of the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$.

1	**set**	$r := 0$	(the number of reassignments)
2	**compute**	**for** $i = 1, \ldots, m$	the center $\boldsymbol{\mu}_i^k$ of Ω_i^k
3	**compute**	**for** $j = 1, \ldots, N$	distances $d(\mathbf{v}_j, \boldsymbol{\mu}_i^k),\ i = 1, \ldots, m$
4		**for** $j = 1, \ldots, n$	**if** $d(\mathbf{v}_j, \boldsymbol{\mu}_\ell^k) = \min\{d(\mathbf{v}_j, \boldsymbol{\mu}_i^k) : i = 1, \ldots, m\}$ **and** $\mathbf{v}_j \in \Omega_p^k\ ,\ p \neq \ell$, **then** $\Omega_\ell^k := \Omega_\ell^k \cup \{\mathbf{v}_j\}\ ,\ \Omega_p^k := \Omega_p^k \setminus \{\mathbf{v}_j\}$ (reassign $\mathbf{v}_j$) $r := r + 1$ **endif**
5	**if**	$r = 0$	**stop**
	endif	$\Omega^{k+1} := \Omega^k$	
		$k := k + 1$	**go to 1**

Remark. Step 4 (reassignment) in Algorithm 1 may leave a cluster Ω_p^k empty, having "lost" all its customers to nearer facilities. The next partition Ω^{k+1} may therefore have fewer than m nonempty clusters.

3. Numerical experience

The proposed method was tested on many datasets. We report the results for six representative datasets from UCI [17], described briefly in the Appendix. A summary of our procedure:

(1) Each dataset was partitioned at random into a *training set* (consisting of 80% of the observations) and a *testing set* (the remaining 20%)
(2) The training set was clustered using Algorithm 1

Name of Data Set	% Correct Predictions			% Errors		Results of [14]	
	Mean	Max	Min	Type 1	Type 2	Max	Min
Breast Cancer	96.5	100	93.1	2.5	1.0	97	91
Liver	63.2	79.3	49.7	19.7	17.1	72	57
Diabetes	74.7	79.9	65.7	10.2	15.1	78	69
Voting	92.0	98.78	82.3	3.8	4.2	96	94
Wine	93.7	100	82.35	2.6	3.7	100	NA
Hepatitis	86.03	96.42	71.43	8.12	5.85	83	NA

TABLE 1. Summary of results for 6 datasets, see Appendix

(3) Each point $(\mathbf{x}, y)$ in the testing set was then matched with a projected cluster Ω_i^X, and the correct value y was compared with the prediction $\overline{y_i}$.
(4) Steps 1–3 were repeated 50 times, each time with a different initial random partition, giving 50 values of the percentage of correct predictions. The mean, maximum and minimum of these 50 values are listed in columns 2–4 of Table 1.
(5) The mean errors (of type 1 and type 2) are listed in columns 5–6. Note that the percentages of means of correct predictions, type 1 errors and type 2 errors add to 100%.

The last two columns of Table 1 give the best (Max) and the worst (Min) performances, in percentages of correct predictions, from among the 33 algorithms (22 decision tree, 9 statistical and 2 neural network algorithms) compared in [14]. The procedure was ten–fold cross validation, with 90% of the dataset in the training set, and 10% used for testing. There was no over–all champion; the winning algorithm in one dataset, may be an also–ran in another dataset.

We are confident that our results, based on 50 random replications, represent the average performance of our algorithm. Not knowing the statistical procedure used in [14], it is not possible to relate our results to the "Max" and "Min" of the last two columns.

In spite of its simplicity, the proposed algorithm, even in its elementary form given above, performed very well on some datasets, notably *Breast Cancer* and *Hepatitis*, and performed credibly on others. An explanation would require statistical analysis, deferred for future research. For now it suffices to note that for the method to work, the relation between $\mathbf{x}$ and y needs some kind of monotonicity, e.g.,

$$d_X(\mathbf{x}, \overline{\mathbf{x}_1}) < d_X(\mathbf{x}, \overline{\mathbf{x}_2}) \implies d_Y(y, \overline{y_1}) < d_Y(y, \overline{y_2}) \,, \tag{5}$$

for any two clusters Ω_1, Ω_2 with means $(\overline{\mathbf{x}_1}, \overline{y_1}), (\overline{\mathbf{x}_2}, \overline{y_2})$. In particular (for clusters consisting of single points),

$$d_X(\mathbf{x}, \mathbf{x}_1) < d_X(\mathbf{x}, \mathbf{x}_2) \implies d_Y(y, y_1) < d_Y(y, y_2) \,, \tag{6}$$

for any three points $(\mathbf{x}, y), (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)$. The monotonicity (6) holds in the case of affine relation between $\mathbf{x} = (x_1, \ldots, x_n)$ and y, say,

$$y = \xi_0 + \sum_{i=1}^{n} \xi_i \, x_i \tag{7}$$

for some constants $\{\xi_i : i \in \overline{0, n}\}$. Therefore, the method proposed here is expected to work well if linear regression works.

In general, the best one can expect is a probabilistic version of (6), with the metric inequality in the right side replaced by a probabilistic inequality.

4. The role of the parameter α

We recall that the role of the parameter α in the distance function

$$d\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha \, n \, d_Y^2(y_1, y_2)} \, , \tag{3}$$

is to control the relative importance on the y–component. For $\alpha = 0$ the distance function $d((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2))$ reduces to $d_X(\mathbf{x}_1, \mathbf{x}_2)$. As α increases, the distance d depends more on y, and less on the $\mathbf{x}$–component. If the class variable y is binary, and if α is sufficiently large, then the clustering algorithm will partition the dataset into two clusters, one with $y = 0$, the other with $y = 1$, regardless of how many clusters were in the initial partition[2].

We call a partition where all points $(\mathbf{x}, y)$ in a cluster have the same value of y a *pure class partition*. As expected, increasing the parameter α eventually results in a pure class partition. If the prediction algorithm still gives good predictions, there is indication that the dataset in question has a special structure, that we call *class separability*. A data set is *binary class separable* if it is class separable, and the class variable is binary.

Our algorithm is greatly simplified for binary class separable datasets, requiring no computation in the clustering stage, and a trivial computation in the other two stages:

Clustering: assign all points $(\mathbf{x}, 0)$ to the cluster Ω_0, and all points $(\mathbf{x}, 1)$ to the cluster Ω_1.

Centers: compute the $\mathbf{x}$–centers $\overline{\mathbf{x}_0}$ and $\overline{\mathbf{x}_1}$ of Ω_0^X and Ω_1^X, respectively.

Prediction: given a point $(\mathbf{x}, y)$, determine the class value:

$$y = \begin{cases} 0 & \text{if } d_X(\mathbf{x}, \overline{\mathbf{x}_0}) < d_X(\mathbf{x}, \overline{\mathbf{x}_1}) \\ 1 & \text{otherwise.} \end{cases}$$

We see below that the *Breast Cancer* dataset is binary class separable: our algorithm gives good predictions even for large values of α that resulted in pure class partition. Treating the *Breast Cancer* dataset with sophisticated tools (see, e.g., [15]–[16]) does not give significantly better results than the above elementary procedure.

[2]Similarly, if the class variable y is discrete with m values, and if α is sufficiently large, the algorithm will result in m clusters, corresponding to the values of y.

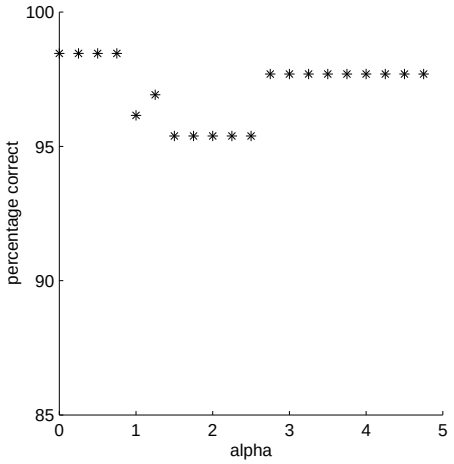

FIGURE 1. Breast cancer: % correct as function of α

The *Diabetes* data set, on the other hand, is not class separable, and may even be missing some important explaining variables.

We illustrate here the role of α for these two extreme datasets: *Breast Cancer* and *Diabetes*.

Summary of procedure: The algorithm was tested for 20 different values of α, starting at $\alpha = 10^{-6}$, and having step 0.25. In all runs, we used the same initial partitioning, the same initial number of clusters = 6, and the same cutoff $p = 0.5$.

Figure 1 displays the dependence of correct predictions percentage on α for the *Breast Cancer* dataset. Figure 2 shows that the number of nonempty clusters decreases as a function of α, giving a pure class partition for $\alpha \geq 2.75$. Figure 1 then shows, for all $\alpha \geq 2.75$, a stable correct predictions percentage of about 97%. The *Breast Cancer* dataset is thus class separable, as claimed.

Figures 3–4 are the analogous illustrations for the *Diabetes* dataset. Figure 4 shows that the number of nonempty clusters decreases to 2 for $\alpha \geq 8.5$, and the percentage of correct predictions (in Figure 3) approaches 76%.

5. Sensitivity to the number of clusters

We next study the importance of the initial number of clusters. For the *Breast Cancer* dataset the results do not depend on the number of initial clusters (as is the case in general for class separable datasets). The results for the *Diabetes* dataset are shown in Figure 5. The algorithm was tested for initial numbers of clusters from 3 to 10, using the same initial partition, the same $\alpha = .4$ and the same cutoff $p = 0.5$. We see that the percentage of correct predictions is almost insensitive to the initial number of clusters.

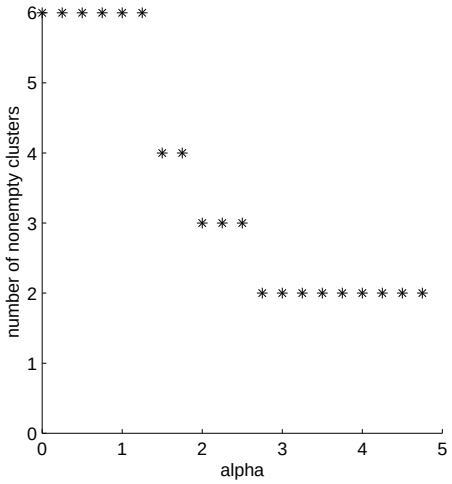

FIGURE 2. Breast cancer: Number of nonempty clusters as function of α

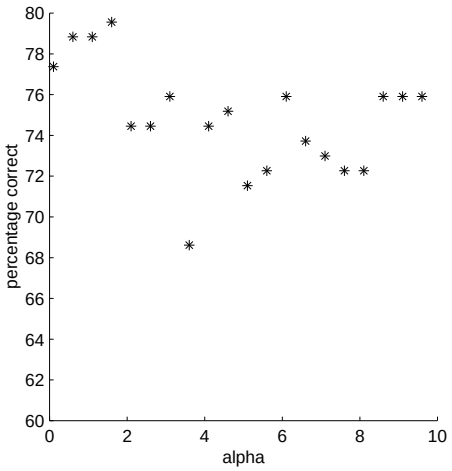

FIGURE 3. Diabetes: % correct as function of α

6. Sensitivity to the cutoff p

Finally we tested the dependence of the predictions accuracy on the cutoff value p used in (4). The results are shown in Figure 6 for *Breast Cancer*, and in Figure 7 for *Diabetes*.

Summary of procedure: The algorithm was tested for values of p from 0.2 to 1, with step 0.1. All tests used the same initial partition, the same initial number of clusters = 6, and the same $\alpha = 0.4$.

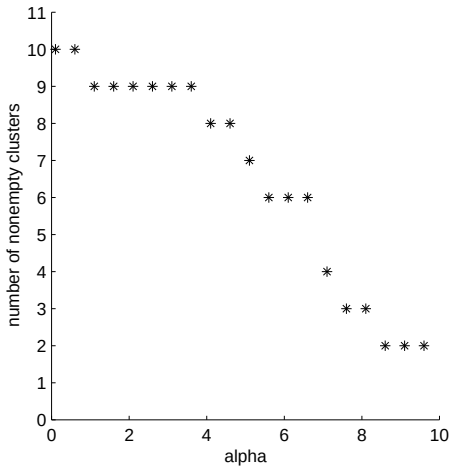

FIGURE 4. Diabetes: Number of nonempty clusters as function of α

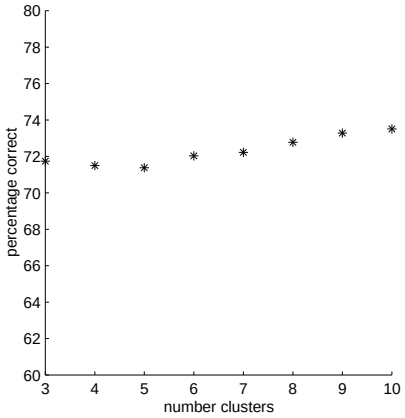

FIGURE 5. Diabetes: % correct as function of the initial number of clusters

Since the *Breast Cancer* dataset is class separable, it is remarkably insensitive to the cutoff value p. Indeed, any value $0 < p < 1$ can serve as cutoff. This is illustrated in Figure 6.

The *Diabetes* dataset, on the other hand, shows a dependence of the accuracy of predictions on the cutoff p, with the optimal cutoff at about 0.5.

7. Future research

The importance of class separable datasets, such as *Breast Cancer*, suggests studying statistical properties of datasets that may shed light on this phenomenon.

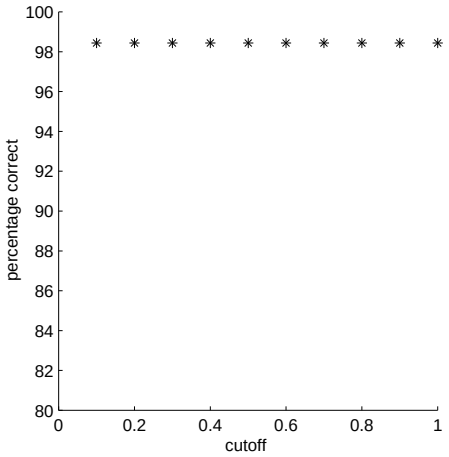

FIGURE 6. Breast Cancer: % correct as function of p

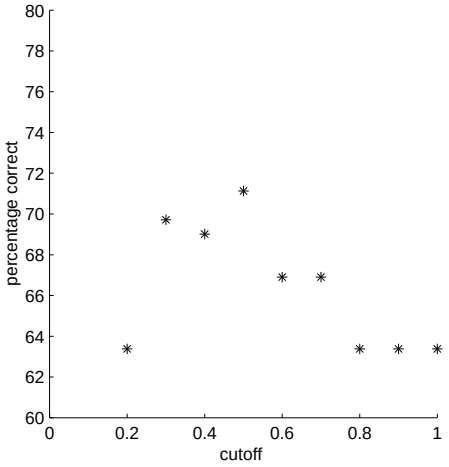

FIGURE 7. Diabetes: % correct as function of the cutoff p

Monotonicity properties of datasets, such as (5), need to be clarified.

For improved performance on given datasets, the method will need modifications, specific for the data set in question. The objects of modification include:

- the distance d_X on X (including the Mahalanobis distance [10], variance and entropic distances),
- the distance d on $X \times Y$,
- the parameter α,
- the initial number of clusters, and
- the binarization cut–off p in (4).

These topics are left for future research.

References

[1] S. Aeberhard, D. Coomans and O. de Vel, Comparison of classifiers in high dimensional settings, *Technical Report 02-1992*, Dept. of Computer Science, James Cook University, Australia

[2] S. Aeberhard, D. Coomans and O. de Vel, The classification performance of RDA, *Technical Report 01-1992*, Dept. of Computer Science, James Cook University, Australia

[3] K. Bennett and O. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, *Optimization Methods and Software* **1**(1992), 23–34

[4] E. Boros, P. Hammer, T. Ibaraki, and A. Kogan, Logical analysis of numerical data, *Mathematical Programming* **79**(1997), 163–190

[5] G. Cestnik, I. Konenenko, and I. Bratko, Assistant-86: A knowledge-elicitation tool for sophisticated users, *In Progress in Machine Learning*, Sigma Press, 31–45

[6] B. Duran and P. Odell, Cluster Analysis, *Springer-Verlag*, Berlin, 1974

[7] P. Diaconis and B. Efron, Computer-intensive methods in statistics, *Scientific American* **48**, 1983

[8] B. Everitt, Cluster Analysis, 3nd edition, *Edward Arnold*, London, 1993

[9] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, *Academic Press Inc.*, Boston, MA, 1990

[10] R. Gnanadesikan, J. Harvey and J. Kettenring, Mahalanobis metrics for cluster analysis, *The Indian Journal of Statistics. Series A* **55**(1993), 494–505

[11] P. Hansen, B. Jaumard, Cluster analysis and mathematical programming, *Math. Programming* **79**(1997), 191–215

[12] M. Jambu and M. Lebeaux, Cluster Analysis and Data Analysis, *North-Holland Publishing Co.*, Amsterdam, 1983

[13] Y. Levin and A. Ben-Israel, A heuristic method for large–scale multifacility location problems, to appear

[14] T. Lim, W. Loh, and Y. Shih, A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms, *Machine Learning* **40**, 203–228

[15] O. Mangasarian and W. Wolberg, Cancer diagnosis via linear programming, *SIAM News*, **23**(1990), 1–18

[16] O. Mangasarian, R. Setiono and W. Wolberg, Pattern recognition via linear programming: theory and application to medical diagnosis, In *Large–Scale Numerical Optimization*, T. Coleman and Y. Li, editors, SIAM Publications, Philadelphia 1990, 22–30

[17] C. Merz and P. Murphy, UCI Repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA, 1996. (http://www.ics.uci.edu/ mlearn/MLRepository.html)

[18] J. Smith, J. Everhart, W. Dickson, W. Knowler and R. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, *In Proceedings of the Symposium on Computer Applications and Medical Care*, 261–265, IEEE Computer Society Press

[19] J. T. Tou and R. C. Gonzales, Pattern Recognition Principles, Addison–Wesley, Reading, Mass. 1974

[20] W. Wolberg and O. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *Proceedings of the National Academy of Science, USA*, **87**–1990, 9193–9196

[21] J. Zhang, Selecting typical instances in instance-based learning, *In Proceedings of the Ninth International Machine Learning Conference*, Aberdeen, Scotland, 1992, 470–479

Appendix A: Brief description of the datasets

We give a short description of the 6 datasets reported in § 3. For further details see [17], [14].

Breast Cancer (Wisconsin). The data (collected at the University of Wisconsin by W. Wolberg) consist of 699 observations pertaining to Breast Cancer. The problem is to predict whether a tissue sample from a breast is malignant. There are 9 numerical attributes (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses), each taking values in $\{0, 1, \ldots, 10\}$. Further references: [3], [4], [15], [16], [20], [21].

Liver (BUPA Liver Disorders). This dataset lists 345 observations on liver disoders. There are 6 numerical attributes (mean corpuscular volume, alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, gamma-glutamyl transpeptidase, number of half-pint equivalents of alcoholic beverages drunk per day). The class variable is binary. The problem is to predict a liver disorder based on the observed attributes.

Diabetes (PIMA Indian Diabetes). The patients in this dataset are adult females of Pima Indian heritage, living near Phoenix, Arizona. The problem is to predict a positive Diabetes test result given a number of physiological measurements and medical test results. There are 768 observations and 8 numerical attributes (number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, age). In [14] one attribute (serum insulin) and 236 observations were removed due to unrealistic values. This change drastically improved the classification accuracy. Further references: [4], [18].

Voting (Congressional Voting Records). This dataset concerns votes on 16 key issues by the 435 US Representatives, considered as observations. There are 16 categorical attributes ("yes" or "no"). The different vote schemes (9 at all) are reduced to a binary variable ("for" or "against"). The problem is to predict a a US Representative's vote given his/her attributes.

Wine (Wine Recognition Data). Wines from the same region in Italy are derived from three different cultivars. The problem is to identify the cultivar given the values of 13 attributes from chemical analysis of the wine. There are 178 observations. Further references: [1], [2].

Hepatitis. This dataset lists 155 Hepatitis patients, with 19 attributes (e.g., age, sex, steroid, bilirubin, albumin, liver firm, etc.) and a binary class variable of survival from the disease. The problem is to predict whether a patient will survive. Further references: [5], [7].

Adi Ben-Israel and Yuri Levin **{bisrael,ylevin}@rutcor.rutgers.edu**
RUTCOR–Rutgers Center for Operations Research
Rutgers University
640 Bartholomew Rd
Piscataway, NJ 08854-8003, USA