Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content

Article · January 1982

CITATIONS
93

READS
155

2 authors, including:

Michele Forina
Università degli Studi di Genova
176 PUBLICATIONS **3,571** CITATIONS

SEE PROFILE

Annali di Chimica 72 (1982), by Società Chimica Italiana

PATTERN RECOGNITION METHODS IN THE PREDICTION OF ITALIAN OLIVE OIL ORIGIN BY THEIR FATTY ACID CONTENT (*)

Michele FORINA (°) and Enrico TISCORNIA

Institute of Pharmaceutical Science, University of Genoa.

Summary - Some methods of parametric (linear discriminant analysis, bayesian analysis) and non-parametric (piecewise linear discriminant analysis = KNN, linear learning machines) multivariate statistical analysis have been used for recognizing the origin region of Italian olive oils by their fatty acid content. High predictive ability proves the effectivness of these methods. Bayesian analysis and KNN show prediction ability about 94 %, and, whenever prediction is wrong, the sample is ascribed to a bordering region.

Riassunto - Alcuni metodi dell'analisi multivariata parametrica (analisi discriminante lineare, analisi bayesiana) e non parametrica (piecewise linear discriminant analysis = KNN, macchine intelligenti lineari) sono stati utilizzati per il riconoscimento della regione di provenienza di olii di oliva italiani dal loro contenuto in acidi grassi. L'alta abilità predittiva dimostra l'efficacia di questi metodi. Nell'analisi bayesiana e nel KNN l'abilità predittiva risulta intorno al 94 %, e, quando la predizione è errata, il campione viene attribuito ad una regione confinante.

- - - - -

(*) Work partly supported by the National Research Council of Italy (CNR).

Great importance is given to the geographical origin of olive oils because their origin considerably affects flavour and nutritional characteristics and consequently the marketing value of oil.

In this work we apply to the fatty acid content of selected samples of Italian virgin olive oils the parametric and non-parametric multivariate methods of supervised pattern recognition, in order to measure the ability of these analytical data and methods for predicting the site of origin of oils.

Multivariate analysis extracts the maximum information from experimental data; the pattern of the oils from a country is obtained objectively. This pattern contains the range of variability of each acid allowed for that country, and, not less important, the cross-correlation between each pair of acids. This pattern is built without assigning any a priori importance to the chemical variables, based either on their quantity or on their nutritional value. So, the multivariate chemical "picture" of the oil from a region is a powerful basis of characterization both for prediction and against adulteration, since man cannot economically reproduce the complex pattern made by nature.

DATA

Data were selected from literature and unpublished data of one of the Authors, as previously described[1].

On the basis of graphical unsupervised pattern recognition[1], it was possible to divide the 572 datavectors (ordered column of eight fatty acid percentages in a sample) into nine oil producing regions, as shown in Table 1.

TABLE 1: Some characteristics of the data used in multivariate analysis.

Region code	Region name	Datavector number	Range of datavector number in the training set
1	North Apulia	25	13 - 19
2	Calabria	56	34 - 43
3	South Apulia	206	126 -160
4	Sicily	36	20 - 28
5	Inland Sardinia	65	35 - 48
6	Coast Sardinia	33	20 - 26
7	East Liguria	50	29 - 38
8	West Liguria	50	25 - 38
9	Umbria	51	31 - 42
	TOTAL	572	364 -405

PREPROCESSING AND TECHNIQUES

Data reported as "trace" in the original literature were processed as a random number in the range 0 - 0.1.

The rough data were generally autoscaled. We refer to non-scaled data in the other cases. In a few cases, the percentage of eicosenoic acid was not used. In some cases, only one or two fatty acid percentages were used for measuring the effectiveness of multivariate analysis compared with univariate and bivariate analysis.

The techniques used were:

non-parametric	K nearest neighbour	KNN
	Linear learning machines	LLM
parametric	Linear discriminant analysis	LDA
	Weighted LDA	WLDA
	Bayesian statistical analysis	BSA

All the techniques require a subdivision of the datavectors into two sets: training or recognition set, and test or prediction set. During the training, the datavectors in the training set are used to generate classification or prediction rules. The performance of the so obtained rules is evaluated, first on the same data from which they were obtained (RECOGNITION) and then on the data of the test set (PREDICTION).

The subdivision between training and test set was made ten times, each time with a random attribution, with the preselected probability p that a datavector may be assigned to the recognition set.

p was selected between 65 and 75 % in the ten subdivisions. The minimum and maximum number of datavectors obtained in the recognition set is given in Table 1. In the following the techniques we have used are briefly described.

KNN - This technique does not require training, and recognition ability is, by definition, 100 %. Indeed, the pattern of each region is given by the collection of the datavectors of that region in the training set. The classification into a region of a datavector **x** of the prediction set is made by measuring the euclidean distances between **x** and all the other datavectors in the training set (we have used other distances too[2], without significant change of the results).

The K nearest datavectors of the training set are retained (we generally used K = 5). A vote is given to each nearest datavector, the vote K to the first nearest, K - 1 to the second one, and

so on up to the vote 1 to the K-th nearest one. A category vote is obtained by the sum of the datavector votes of the same category, and the category with the highest vote is assigned to the datavector **x**.

LLM - The pattern of each category is given by a volume in the hyper space of the variables; this volume is delimited by a number of hyperplanes. The equations of the hyperplanes which best classify the datavectors in the recognition set are obtained by a stepwise training procedure[3]. In the first step LLM locates, in the J-dimensional space of the variables, a hyperplane that divides the space into two half-spaces. Each half-space includes a group of one or more categories. The hyperplane equation has the form:

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_J x_J + w_{J+1} = 0$$

The J + 1 coefficients of **w** are obtained by an iterative inspection of the datavectors in the training set. When a datavector of this set is not correctly classified by the actual hyperplane, the weight vector **w** given by the J+1 weights w in the J+1-dimensional space of the weights is appropriately rotated. Iteration stops when recognition is complete. When the datavectors in the training set are not linearly separable, i.e. when no hyperplane can perfectly separate the datavectors into two groups, we use a dampened learning machine, in which the maximum allowed rotation of the weight vector is gradually reduced, in order to avoid erratic rotation of the vector due to misclassified datavectors. Otherwise we use a quadratic learning machine, operating in the J(J+3)/2 dimensional hyperspace of the original variables plus their squares and cross-products.
However the results obtained by this quadratic learning machine are not significantly better than those obtained by LLM.

The performance of a LLM is evaluated by the effectiveness of the classification of the datavectors in the training set (degree of linear separability, i.e. recognition ability) and by the predictive ability on the datavectors in the prediction set.

By repeated use of the LLM, we can divide the datavectors into groups of decreasing number of categories. The series of binary decisions constitutes a logical decision tree (overall learning machine).

LDA - Parametric methods draw out some statistical parameters from the datavectors in the training set, and by these parameters the pattern of each category is built according to some statistical hypothesis. In Linear Discriminant Analysis the pattern of each region is given by a confidence hyperellipsoid in the hyperspace of variables[4]. At the same confidence coefficient, the hyperellipsoids are equal, but they have a different position in the hyperspace. All the points on the hyperellipsoid surface have the same Mahalanobis distance

from the center of the hyperellipsoid (baricenter of the category).

In order to compute the Mahalanobis distances in the training set the pooled dispersion matrix **V** of the datavectors is obtained. Both for recognition and prediction, from each datavector **x** the L datavectors $\mathbf{x} - \mathbf{m}_l$ are obtained, where L is the number of categories, and each datavector is given by the J axial distances between the datavector **x** and the baricenter $\mathbf{m}_l$ of each category.

The product

$$(\mathbf{x} - \mathbf{m}_l)' \, \mathbf{V}^{-1} \, (\mathbf{x} - \mathbf{m}_l)$$

is the Mahalanobis distance of the datavector **x** from the baricenter of the category l. The datavector is classified into the category that has its baricenter at the minimum Mahalanobis distance.

Weighted Linear Discriminant Analysis - In this method the pattern of each region is given by hypersphere in the space of the eigenvectors of the matrix $\mathbf{V}^{-1}\mathbf{L}$ where **L** is the covariance matrix of the baricenters of the categories[5]. The datavectors are projected from the original J-dimensional space of the variables into the space of the first J' eigenvectors of the matrix $\mathbf{V}^{-1}\mathbf{L}$ (J' is selected so as to optimize the recognition); these eigenvectors are generally not orthogonal, so that we have not only a simple rotation of axes but a deformation of the original space of variables. This is equivalent to weighting the variables to optimize the separation between categories.

The euclidean distances between each datavector and the category baricenters in the space of the eigenvectors are used as a classification criterion.

Bayesian Statistical Analysis - The pattern of each region is given by a confidence hyperellipsoid, as in the case of LDA; however the confidence hyperellipsoids differ both in the center position and in their form[4].

For each datavector the L Mahalanobis distances from the category baricenters are computed by using the intracategory covariance matrix $\mathbf{V}_l$ instead of the pooled dispersion matrix **V**. From the Mahalanobis distance the conditional probability p_l is computed, and the datavector is assigned to the category l showing the highest probability. As a result, the distribution of each variable is no longer seen as category-independent, and a more accurate statistical picture of training data ensues. Moreover, the a priori probability that a datavector may lie in a category and the risk of misclassification must be taken into account.

COMPUTING

All statistical and mathematical treatments were made with an Olivetti P-6060 16 kbyte desk computer. The BASIC programs used are part of the package PARVUS, developed in this Institute and available upon request. The running time of each technique is closely related to the low memory capacity. A very approximate picture of the relative time requirements of the techniques used is the following (where time for LDA is considered 1):

Technique	training	recognition	prediction
KNN	-	-	12
LLM	24	0.1	0.1
LDA	1	1	1
WLDA	2	1	1
BSA	10	16	8

RESULTS AND DISCUSSION

Tables 2-7 and Figure 1 summarize the results obtained with the pattern recognition techniques.

The high recognitive and predictive abilities shown by all the techniques (Table 2) point out that the fatty acid content is very useful to locate the geographical origin of an olive oil. Moreover, in the case of incorrect prediction (or recognition) the errors are slight, i.e. the oil is classified in a region very close to the true one (see Table 5). The advantages of multivariate analysis are evident from data in Tables 3 and 5. The use of eight parameters greatly lowers the number of incorrect attributions, and furthermore removes serious classification errors, as the attribution of a Ligurian oil to Apulia or Calabria, or of an Umbrian oil to Sicily.

The data in Table 3 show that the relative importance of fatty acids in Italian olive oil classification is: 4 (oleic) - 5 (linoleic) - 1 (palmitic). This order results from bivariate analysis, and it is slightly different from that given by univariate analysis, where palmitoleic acid is the third in predictive power, the second in the recognitive power.

However, both recognition and prediction with the oleic-palmitic acid pair are better than that with the oleic-palmitoleic acid pair; thus showing that a large part of the information contained in the palmitoleic percentage is the same as in the oleic acid percent.

TABLE 2: Results of the classification analysis. The results are reported as the range of percent recognitive (R) and predictive (P) abilities obtained with ten random subdivisions between training and prediction set.

Region	KNN		LLM		LDA		WLDA		BSA		Mean
	P	R	P	R	P	R	P	R	P	P
North Apulia	83-100	100	67-100	94-100	87-100	93-100	87-100	100	38-88	86
Calabria	89-100	89-92	92-95	83-88	75-95	52-90	45-89	95-100	75-95	85
South Apulia	96-100	97-99	89-97	95-98	93-100	91-94	86-96	96-99	93-100	96
Sicily	31-67	67-77	20-35	76-91	54-80	73-91	23-77	90-100	70-87	55
Inland Sardinia	100	100	92-95	100	95-100	97-100	100	98-100	100	98
Coast Sardinia	92-100	100	90-100	100	100	91-100	100	100	90-100	98
East Liguria	93-100	94-100	90-94	90-97	75-100	80-94	65-81	100	93-100	92
West Liguria	100	94-100	100	100	100	97-100	92-100	100	100	99
Umbria	100	97-100	65-100	94-98	90-100	87-100	94-100	100	85-100	93
Total	91-96	95-97	85-92	94-97	90-96	89-92	81-93	98-99.5	91-96	92
Total (mean)	94.5	96	89	96	93	90	88	99	93.5	

TABLE 3: Results of univariate and bivariate Bayesian statistical analysis.(*)

UNIVARIATE variable index	(acid) name	% R	% P	BIVARIATE variables indexes	% R	% P
1	Palmitic	43.4	40.9	4,5	83.8	85.8
2	Palmitoleic	51.3	41.5	1,4	80.3	82.4
3	Stearic	23.8	26.7	2,5	78.8	80.1
4	Oleic	57.1	60.8	2,4	75.8	69.3
5	Linoleic	50.5	59.7	3,4	72.2	71.0
6	Eicosanoic	39.4	39.8	1,2	71.7	66.5
7	Linolenic	33.8	29.6	6,8	62.9	56.8
8	Eicosenoic	40.2	36.4	3,7	50.0	46.0

(*) With the same random subdivision between training and prediction set, multivariate Bayesian analysis with eight acids gives 99 % recognitive ability (R) and 96 % predictive ability (P).

TABLE 4: Details of a prediction run with 5-NN

				All the nearest in ONE category			The nearest in TWO categories				The nearest in THREE categories			
Region index	Number of datavectors	Correct predictions	Incorrect predictions	Total	Correct	Incorrect	Total	First correct	2-nd correct	Both incorrect	Total	First correct	2-nd correct	All incorrect
1	9	8	1	7	6	1	2	2	0	0	0	0	0	0
2	19	19	0	16	16	0	2	2	0	0	1	1	0	0
3	64	60	4	48	47	1	10	8	1	1	2	1	0	1
4	12	6	6	4	1	3	8	5	3	0	0	0	0	0
5	22	22	0	22	22	0	0	0	0	0	0	0	0	0
6	10	9	1	9	9	0	1	0	1	0	0	0	0	0
7	17	15	2	10	10	0	7	5	2	0	0	0	0	0
8	16	15	1	11	11	0	4	3	1	0	0	0	0	0
9	17	17	0	16	16	0	1	1	0	0	0	0	0	0
ALL	186	171	15	148	143	5	35	26	8	1	3	2	0	1

TABLE 5: Misclassification matrix - The number of cases in which an oil of the row category is attributed to the column category is reported both for multivariate and (in brackets) bivariate Bayesian analysis (variables 4,5 (oleic acid, linoleic acid). Zeroes and correct classifications are not reported.

	1	2	3	4	5	6	7	8	9
1							(1)		(1)
2							(1)		
3		(2)		1(1)					
4	1	2(4)	1(1)				(1)	(1)	
5		(1)							
6					1				
7	(4)	(1)						3	(2)
8			(1)				(2)		
9				(1)					

TABLE 6: Some results with KNN method. The results refer to the same subdivision between training and test sets.

Conditions	Percent predictive ability
8 variables, $\underline{K} = 5$	96.0
8 variables, $\underline{K} = 1$	95.8
7 variables, $\underline{K} = 5$ (*)	93.8
8 variables, Fisher weights, $\underline{K} = 5$	94.3
8 variables, non scaled, $\underline{K} = 5$	90.0

(*) Eicosenoic acid omitted.

TABLE 7: Results of weighted linear discriminant analysis (the results refer to two subdivisions between training and test sets).

J'	% variance of j'-th eigenvector		% total variance of the J' eigenvectors		% Rec.	% Pred.
1	45.3	43.8	45.3	43.8	-	-
2	29.1	30.7	74.4	74.5	78.3-80.5	78.4-78.8
3	12.9	12.0	87.3	86.5	89.4-89.0	88.6-91.4
4	7.3	7.6	94.6	94.1	89.9-89.6	89.8-90.9
5	3.1	3.8	97.7	97.9	91.4-91.7	89.8-92.9
6	2.1	2.0	99.8	99.9	91.4-92.3	89.8-92.9

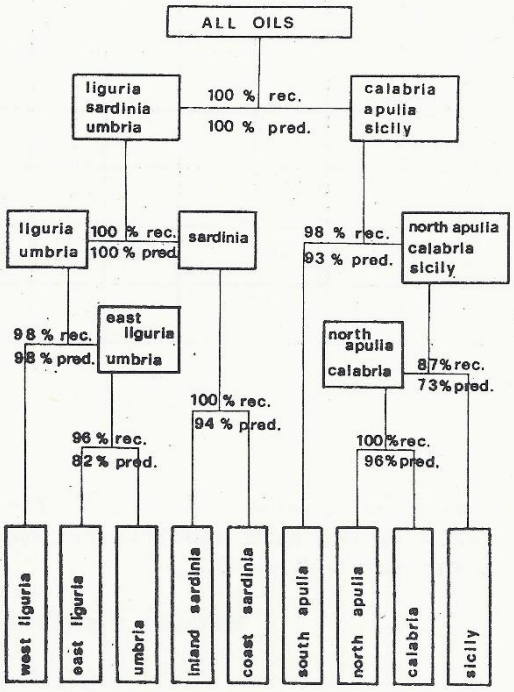

FIG.1 - Binary decision tree of the overall learning machine. Recognitive and predictive abilities refer to the same random subdivision between training and test sets.

The relative importance oleic-linoleic-palmitic acid ascribes the maximum classification power to the oil components with the greatest percentages. Very likely, the analytical percent error of these acids is lower than that of the minor component. On the one hand, this supports the results of multivariate analysis, because they are founded on low-error parameters; on the other hand the small contribution of minor components can be explained with greater intraclass variance due to a higher experimental error; perhaps, a more accurate analytical procedure can increase their importance, and therefore the overall predictive efficiency.

However, the importance of the above results is lowered by the small number of Italian oil-producing regions we have used in the multivariate analysis. Moreover, besides fatty acids, many other chemical variables (as sterols, headspace components) and physico-chemical ones (spectra, refractive index, viscosity,...) could be used in multivariate analysis to discriminate between a larger number of areas, or to increase the predictive ability.

The following aspects can also be deduced from the results. Bayesian analysis shows very high recognitive and predictive ability; in spite of the non-normality of univariate marginal distribution, multivariate normal distribution behaves very well in the description of the statistical data. At present, BSA requires a higher computing time than KNN. However, in view of the use of a greater number of datavectors, or with a slightly larger computer, BSA can be faster than KNN, with a negligeable loss of efficiency.

The oversimplification introduced by linear discriminant analysis, which considers the variables to have category-dependent means but category-independent variances and correlation coefficients, has a 3 percentage-point effect on the recognitive ability and almost no effect on the predictive ability. So, the high computing speed of LDA can justify its use instead of KNN or BSA.

Weighted linear discriminant analysis is a technique which tends to substitute LDA in computer program packages (e.g. in BMCP)[4]. However, here it undoubtely gives both lower recognitive and predictive abilities. Indeed, the use of the eigenvectors of the ratio matrix causes an average improvement of the separation between categories; as a consequence, however, we can obtain a better separation between distant categories at the expense of a worse separation between adjoining categories. Because of that, the overall discrimination is worse.

The number of eigenvectors used in weighted linear discriminant analysis can be reduced to three without any performance degradation. The elimination of the third eigenvector is critical as regards the separation between West Liguria and both East Liguria and Inland Sardinia, and between North Apulia and Sicily. The data in Table 7 show the high repeatibility of the variance content of each eigenvector. About 15 % of the total information is unimportant as regards discrimination.

When we consider the predictive ability for each region, we can see that it is very low in the case of Sicilian oils. From a general overview, the mean regional results of all classification techniques show a very good agreement with the results of the graphical analysis obtained with the simplified non-linear mapping[1].

The results obtained with KNN show that the weighting slowly reduces the predictive ability. The same discussion applies as in the case of WLDA. Fisher weights give great importance to eicosenoic acid: in effect, since this acid is at trace level in North Italian oils, we have a very low intraclass variance, and a great interclass variance when two regions are compared, one in the North, the other in southern Italy. However, this great Fisher weight does not justify the small decrease in predictive ability, when the eicosenoic acid is not taken into account in the multivariate analysis.
The value of K seems almost of no importance; however, the computing time is quite independent of K. Indeed, in most cases all five near neighbours are in the same class; when two of them are in two categories it is very probable that the first one is in the category with the highest vote.

The preprocessing by autoscaling we have used makes the initial importance of all variables equal. However, the analytical error per cent is higher in the case of minor components. So the autoscaling gives excessive importance to these variables. However, when autoscaling is not used, the predictive ability remains high. Perhaps, without autoscaling, the a priori importance of minor components is too low. The difficulties in selecting appropriate weighting, between equal weights, weights proportional to mean percentage and Fisher weights, could be surmounted by optimizating the weights by means of a "weighting set", having maximum prediction ability when the weights are optimum. This procedure, suggested by Byers and Perone[6], is very time consuming.

At present, we think that the increase of prediction ability we were able to obtain is not such as to justify its utilization.

In conclusion, LLM shows a relatively low predictive ability (higher, however, than that shown by WLDA). This was the only multivariate technique showing some serious classification errors (only twice in the ten random subdivisions, and the errors consisted in the classification of a Sicilian oil as Ligurian). In spite of these imperfections, this technique could be of some interest because of its very low computing time, and its immediate improvement when more datavectors become available.

CONCLUSIONS

The information given by fatty acid content is sufficient to give a very high predictive ability both with parametric and non parametric methods of pattern recognition.

The differences between oils from various Italian regions can be attributed to many factors: olive variety, climate, soil geochemistry, and so on, which do not appear directly from the data analysis.

So, an attribution to the "region X" could, perhaps, be correctly read as attribution to "variety Y" typical of the region X.

More and more samples and treatments will be necessary to answer the questions about olive oil origins. The results reported here can be a beginning for more extensive work, with more samples, more countries, more chemical and physical data.

Received September 24th, 1981

REFERENCES

1) M.FORINA, C.ARMANINO, Annali Chim.(Rome), 71, (1982).
2) D.L.MASSART, A.DIJKSTRA, L.KAUFMAN, "Evaluation and Optimization of Laboratory Methods and Analytical Procedures", Elsevier Sci. Publ. Co., Amsterdam, The Netherlands, 1978, p. 378.
3) P.C.JURS, B.R.KOWALSKI, T.L.ISENHOUR, Anal.Chem., 41, 21 (1969).
4) W.J.DIXON, ed., "Biomedical Computer Programs", Univ.California Press, Berkeley, Calif., 1974, pp. 221-254.
5) F.K.KAWAHARA, Y.Y.YANG, Anal. Chem., 48, 651 (1976).
6) W.A.BYERS, S.P.PERONE, Anal.Chem., 52, 2173 (1980).