Pattern Recognition Letters 1 (1982) 15–20
North-Holland Publishing Company

October 1982

The use of continuous variables for labeling objects

Robert P.W. DUIN
Department of Applied Physics, Delft University of Technology, Delft, The Netherlands

Received 7 July 1982

Abstract: A summary is given of the various pattern recognition situations in which continuous variables may be used for labeling objects. Specific problems may arise during the construction of classification functions, e.g. when discontinuities of the assigned labels have to be avoided. Solutions are discussed and an example is given.

Key words: Continuous labels, fuzzy labels, mixtures of classes, probabilistic labels, multiple nonlinear regression.

1. Introduction

In statistical pattern recognition one tries to assign a label λ to an object $\boldsymbol{x}$ that is represented by K featurevalues: $\boldsymbol{x}=(x_1,x_2,\dots,x_K)$. Such a label is usually a nominal variable: $\lambda\in\{1,2,\dots,L\}$ with L the number of classes. There is no order defined on the possible values of λ. They are just symbols referring to some class. In this paper the case will be investigated in which λ is a continuous variable, e.g. $\lambda\in[0,1]$ or $\lambda\in(-\infty,\infty)$. Below a number of situations is given where this kind of labeling can be used. Note that the number of classes may still be finite.

(1) Probabilistic labels, $\lambda\in[0,1]$ is the probability that the corresponding object, given a number of observations, belongs to a certain class. To a certain object $L-1$ probabilistic labels may be assigned independently.

(2) Fuzzy labels, $\lambda\in[0,1]$ is the membershipvalue of the corresponding object to a certain class. By this the labelvalue may represent some (subjectively estimated) distance to the class ideal. To a single object L fuzzy labels may be assigned independently.

(3) Mixtures. If each object is in fact a mixture of L different components (e.g. chemical compounds) $L-1$ continuous labels may be assigned to it independently, defining the mixture rates.

These three situations are really different but often confused. They are all closely connected to a situation with, in some way or another, a finite number of classes. In the two-class case a probabilistic label $\lambda=0.7$ implies that the corresponding object is a member of a family of objects of which 70% belongs to one of the classes. A fuzzy label of 0.7 implies that the corresponding object is a reasonable, but not very good example of the class of objects the label is referring to. A mixture label of 0.7 implies that the corresponding object consists for 70% of one of the components. Each of these three label types are in fact a refinement of the case of nominal labels. This does not hold for the next type.

(4) Class-continuum. In this case one of the continuous variables measured on the objects is treated as a label. The aim is to estimate the value of this 'label variable' from all other variables (features), instead of measuring it. For an example see Section 4.

The classification problem with continuous labels may look similar to the multiple nonlinear regression problem. However, a few differences exist. In regression one is interested in the relation

$\lambda = g(\boldsymbol{x}, \theta)$ between λ and the set of variables $\boldsymbol{x}$. Often the function $g(\text{-})$ is given without the exact values of the parameters θ. These have to be estimated from the learning set, consisting of the measured combinations $\{\lambda_i, \boldsymbol{x}_i, i = 1, \dots, m\}$.

In the classification problem there is usually no function $g(\text{-})$ given and one is primary even not interested in it at all, but just in the possibility of classification: find label estimates $\hat{\lambda}$ for new objects, based on some learning set of labeled objects. Another difference is that in regression each set of variable values generates some value of λ, including some noise, while in the classification problem each label λ generates a set of objects according to some distribution $f_\lambda(\boldsymbol{x})$. Therefore the relation between $\boldsymbol{x}$ and λ has here primary to be written as

$$\boldsymbol{x} = \boldsymbol{X}(\lambda) + \varepsilon \tag{1}$$

in which ε represents a K-dimensional additive noise vector. This difference in approach is caused by the fact that with each feature its own noise may be related and that for the noise-free case a label-value λ uniquely defines some featurevalues x_j $(j = 1, \dots, K)$, but not the other way around.

In Section 2 classification problems and strategies are discussed. The feature reduction problem is shortly treated in Section 3. An example in which some of the problems discussed before are illustrated is given in Section 4.

2. Classification strategies and problems

In the case of continuous labels classification errors cannot be measured in terms of probabilities of wrong classification as almost each estimated label will differ from the true one. In this case the difference between the label and its estimate is of interest. It seems natural to use the expected square error for measuring the performance

$$\delta = E_i(\hat{\lambda}_i - \lambda_i)^2 \tag{2}$$

is the true label of an arbitrary object i, and $\hat{\lambda}_i$ is its estimate. Other choices, however, may be possible. If the relation $\boldsymbol{X}(\lambda)$ is linear (1) can be written as

$$\boldsymbol{x} = \lambda \boldsymbol{a} + \boldsymbol{b} + \varepsilon.$$

The parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ may easily be estimated from a learning set $\{\boldsymbol{x}_i, i = 1, \dots, m\}$ by minimizing the mean square distance between $\boldsymbol{X}(\lambda_i)$ and $\boldsymbol{x}_i$ for the learning set. This problem is identical with the linear regression problem, except that $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{x}$ are vectors. The following estimators follow therefore immediately from the linear theory, e.g. see Draper and Smith (1966):

$$\hat{\boldsymbol{a}} = \sum_i \lambda_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) / (\overline{\lambda^2} - \bar{\lambda}^2), \tag{3}$$

$$\hat{\boldsymbol{b}} = \bar{\boldsymbol{x}} - \bar{\lambda} \hat{\boldsymbol{a}} \tag{4}$$

where $\bar{\boldsymbol{x}}$, $\bar{\lambda}$ and $\overline{\lambda^2}$ are the averages of respectively $\boldsymbol{x}$, λ and λ^2 over the learning set. An unknown label λ may now be estimated from a given $\boldsymbol{x}$ by minimizing the distance between $\boldsymbol{x}$ and $\hat{\boldsymbol{X}}(\lambda) = \lambda \hat{\boldsymbol{a}} + \hat{\boldsymbol{b}}$. If for this distance the Euclidean distance is used, differences in variances between features are not taken into account. If the variance–covariance structure may assumed to be constant over the feature space the covariance matrix Σ may be used for normalizing the feature space: rotate over the eigenvectors of Σ and divide by the square roots of the eigenvalues of Σ. The value of λ that minimizes the distance to $\hat{\boldsymbol{X}}(\lambda)$ is now given by

$$\hat{\lambda} = (\boldsymbol{x} - \hat{b}) \cdot \hat{\boldsymbol{a}} / (\hat{\boldsymbol{a}} \cdot \hat{\boldsymbol{a}}). \tag{5}$$

If $\boldsymbol{X}(\lambda)$ is an unknown nonlinear function, other strategies have to be followed such as:

(A) Piece-wise linear approximation. The range of values λ takes on for all learning objects is split into a number of nonoverlapping intervals, such that for each interval the number of corresponding learning objects is about equal. An unknown object $\boldsymbol{x}$ is first classified into one of the subsets by some classical multiclass classification technique. The resulting subset corresponds to a possible region for λ. An estimate $\hat{\lambda}_1$ may now be found by assuming that in this interval $\boldsymbol{X}(\lambda)$ is linear and applying the linear technique treated above. The nonlinear dependency between $\boldsymbol{X}$ and λ is thereby approximated by a piece-wise linear fucntion. How good this is depends upon the degree of nonlinearity, the number of subsets chosen and the number of available learning objects.

(B) The stochastic relation between $\boldsymbol{x}$ and λ may be estimated from the learning set by estimating the joint density distribution $f(\boldsymbol{x}, \lambda)$. An estimate $\hat{\lambda}$

for λ by a given value of $\boldsymbol{x}$ may be found by maximizing $\hat{f}(\boldsymbol{x},\lambda)$ for λ:

$$\hat{\lambda}_2=\arg\max_{\lambda}\{\hat{f}(x,\lambda)\} \tag{6}$$

or by the mean value of λ for the given value of $\boldsymbol{x}$:

$$\begin{aligned}\hat{\lambda}_3&=\int_{\lambda}\lambda\hat{f}(\lambda\mid\boldsymbol{x})\mathrm{d}\lambda\\&=\int_{\lambda}\lambda\hat{f}(\lambda,\boldsymbol{x})\mathrm{d}\lambda\Big/\int_{\lambda}\hat{f}(\lambda,\boldsymbol{x})\mathrm{d}\lambda.\end{aligned} \tag{7}$$

The joint distribution $f(\boldsymbol{x},\lambda)$ can have any form, because of the nonlinear dependency. For that reason a general, nonparametric estimator like the Parzen estimator may be used for estimating $f(\boldsymbol{x},\lambda)$. As the computation of a single Parzen estimate is already computational heavy, the computations of the estimates (6) or (7) for λ will become very unfeasible for any reasonable size of the learning set. An approximative method might be the following.

(C) Nearest neighbour method. For an object $\boldsymbol{x}$ to be classified its N nearest neighbours in the learning set are found: $\boldsymbol{x}^1,\boldsymbol{x}^2,\ldots,\boldsymbol{x}^K$. Let the corresponding labels be given by $\lambda^1,\lambda^2,\ldots,\lambda^N$. An estimate for λ is now:

$$\hat{\lambda}_4=\sum_{i=1}^{N}\lambda^i. \tag{8}$$

If the size of the learning set goes to infinity simultaneously with N, $\hat{\lambda}_4$ becomes identical with $\hat{\lambda}_3$. If the size of the learning set is finite, N should be small enough to obtain local linearity between λ and $\boldsymbol{X}$, otherwise a systematic error is introduced in the estimate $\hat{\lambda}_4$.

All the above methods linearize in some way or another the function $\boldsymbol{X}(\lambda)$. In (A) it is piece-wise linear, in (B) it is hidden in the density estimation procedure and in (C) it is caused by the local use of learning objects. The mean square error in the label estimate of an unknown $\boldsymbol{x}$ is therefore directly related to the mean square error in the linear case, which is given by

$$\delta=\boldsymbol{a}^{\mathrm{T}}\Sigma\boldsymbol{a}/(\boldsymbol{a}\cdot\boldsymbol{a})^2. \tag{9}$$

The effect on the error of using a finite learning set is primary dependent on the number of learning objects used for the local estimates: in (A) the number per subset, in (B) the number used for finding a local density estimate and in (C) the number of nearest neighbours. Second order effects are the additional error made by choosing the wrong subset in (A), or by having some systematic error due to a too heavy linearization of the nonlinear relation $\boldsymbol{X}(\lambda)$.

The classification methods (A) and (C) are discontinuous in the sense that an infinitesimal deviation of $\boldsymbol{x}$ may cause a step in the label estimate $\hat{\lambda}$. For a number of applications this may be very undesirable. For instance, if one studies the classification of a mixture of components with a continuous varying mixture rate, one does not expect a discontinuous mixture rate estimate (the label). As the use of method (B) may be unwanted for its computational complexity, some heuristic approach has to be used for avoiding this problem. A detailed example is given in Section 4.

3. Feature reduction

There may be two reasons for lowering the dimensionality of the feature space. One is to decrease the amount of computations and measurements to be done during classification. The second is to attempt to increase the classification accuracy by using less parameters to be estimated and by filling the feature space better by the available learning set.

The usual methods for feature reduction may be applied to the subset-classes as defined in the previous section, method (A). This method initially approximates the continuous labeling by a multiclass problem, thereby discretizing the labels. This will decrease the accuracy of the feature reduction. Therefore some method may be needed that treats the feature space as a whole, e.g. a Karhunen–Loeve expansion. If the noise is large for some features this method will focus on the noise structure instead of the discriminating power of the features. This again can be avoided by normalizing the feature space as indicated in the previous section, provided that the noise is constant over the space. It seems very hard to perform a reasonable feature reduction in the case of heavy, spatial dependent noise. A solution might be to select subsets of the learning set, like in method (A), such that for each subset $\boldsymbol{X}(\lambda)$ can be approximated by a linear function. For some applications, however, this may be impractical or impossible.

A problem that also exists in multiclass separation may arise here too: The selected feature set gives a very good performance for some intervals of λ but is very poor for other intervals. By this the classification accuracy may become strongly non-uniform over the range of λ. This effect may be restricted by using a max-min method, by which the minimum classification accuracy over the range of λ is maximized.

4. Example

In this section we will present an example where a number of problems described previously arose. The solutions developed so far will be worked out. The problem arose in a project in which the possibilities of controlling an fore-arm prosthesis are investigated. One of these possibilities is a principle originally formulated by Wirta and Taylor (1969) that states that with each distal effort of the arm of a normal subject an activity of the proximal shoulder musculature corresponds. The contraction of these muscles serve to provide reaction force and stabilize the shoulder joint in a natural way. A prosthesis controlled by these synergistic muscles may be operated in a way that it is natural and easy to learn.

In one of the experiments set up to investigate this principle for practical use the activities of 10 muscles in the shoulder girdle of a normal subject are measured, resulting in the features $x_1, x_2, \ldots, x_{10}$. Simultaneously the direction of the force exerted by the hand in the vertical plane is measured. This direction is treated as the label λ. For details of the experimental situation see Duin et al. (1977). The aim is to find out whether it is possible to estimate the direction λ $(0<\lambda\leq 2\pi)$ of a force from the muscles activities $(x_1, x_2, \ldots, x_{10})$. The classification accuracy has to be reasonable but not necessarily very good, as in practice the accuracy will be increased by visual feedback. More important are the speed and the complexity of the computations needed for classification as they have to be performed fast (less than 100 ms) by a microprocessor builtin in the prosthesis. Moreover, it is necessary that discontinuities as described in Section 2 are avoided in order to obtain stable classification results during small changes of the feature values. Finally it is desirable to obtain a uniform classification accuracy over the interval $0<\lambda\leq 2\pi$.

First a learning set was measured for 16 different values of λ: $\lambda_i = 2\pi i/16$, $i=1,\ldots,16$. For each value of λ, 150 measurements were made on the 10 features, constituting a learning set of $16 * 150$ objects. As each subset of 150 objects is measured for the same value of λ they may be used for obtaining local estimates of the noise structure. It appeared that there are considerable differences in the covariance matrices of the various subsets. Nevertheless, the mean covariance matrix was used for normalising the feature space. After this, feature reduction was obtained by using the first two eigenvalues of the covariance matrix of the complete learning set. An example of a plot of the subset means in this reduced feature space for a normal subject is given in Fig. 1. New objects $\boldsymbol{x}$ were classified with visual feedback for the subject. Method (A) (Section 2) was applied in the feature space of Fig. 1, pairing successive measurement directions. During the test stage the subject was asked to exert a force in a preset direction, such that the correct classification result was obtained. A typical result is given in Fig. 2, showing the difference $\hat{\lambda}-\lambda$ as a function of time. However, for a few directions it appeared sometimes impossible to obtain a stable result, see Fig. 3. The estimate $\hat{\lambda}$ kept wandering between two directions. The cause is the discontinuity effect described in Section 2.

In order to avoid these discontinuities we changed method (A) in the following way. After obtaining the two-dimensional reduced feature space, a point P was selected by an interactive or

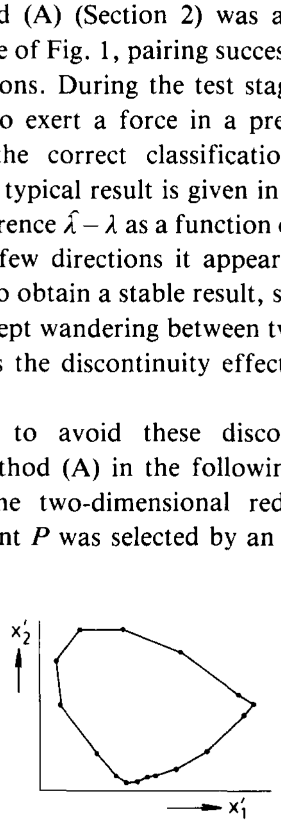

Fig. 1. The means of the 16 subsets of the learning set, projected on the first two eigenvectors (see text).

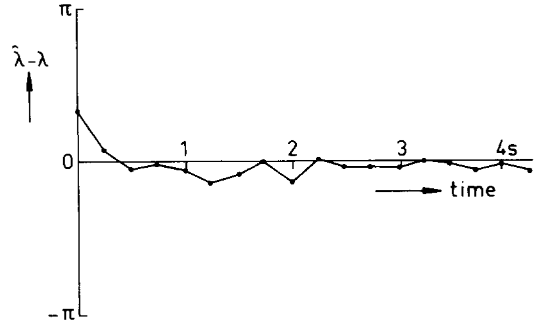

Fig. 2. The error $\hat{\lambda}-\lambda$ in the classification result as a function of time during online testing with visual feedback, stable situation.

automatic method such that from this point all intervals of the curve $\boldsymbol{X}(\lambda)$ could be 'seen' directly (this is not possible for every $\boldsymbol{X}(\lambda)$). An object $\boldsymbol{x}$ is now classified by projecting it from P on $\boldsymbol{X}(\lambda)$. All discontinuities are thereby concentrated in P. Objects which are close to P may be rejected. Moreover, the feature selection procedure was changed in such a way that the shortest interval between two subset means corresponding with two successive values of λ:

$$\min_{i=1,\ldots,16} \|\hat{\boldsymbol{x}}(\lambda_i)-\hat{\boldsymbol{x}}(\lambda_{i-1})\|$$

becomes as large as possible. By this stable classifications (1% reject) with a reasonable accuracy (0.3 rad) were obtained as described by Den Boer (1980).

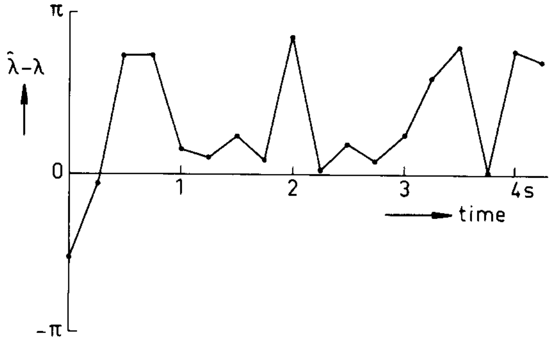

Fig. 3. The error $\hat{\lambda}-\lambda$ in the classification result as a function of time during online testing with visual feedback, instable situation.

5. Discussion

In this paper we have suggested some methods for the estimation of labels represented by continuous variables. The pattern recognition aspects of the problem are closely connected to the lack of knowledge on the functional relation between featurevalues and labels. We have suggested some methods for a linear approximation of nonlinear relations. As soon as knowledge becomes available on the functional form, parameter estimation methods can be used as well and will be probably more accurate. If this knowledge is not available classification techniques using continuous labels can be of use. They may be applied in statistical pattern recognition in several ways, e.g.:

(1) During the learning stage continuous labels enable the teacher to express his knowledge in a better, more subtle way. The classification accuracy, also for nominal classes, can be increased by this, e.g. see Beukema toe Water and Duin (1981).

(2) A classification result expressed in continuous labels preserves the doubt that may exist between two or more nominal classes. Such a label can smooth the cascading of pattern recognizers, automatic or human. For example, a human specialist can better integrate a probabilistic or a fuzzy label produced by an automatic diagnostic program with his own observations. Besides, he is more willing to. See Hermans and Habbema (1975).

(3) In a number of applications the classes really constitute a continuum, like in the example given in Section 4. It would be unnatural and decrease the accuracy if this is neglected.

The use of continuous variables for labeling objects makes problems like the need of a large learning set and the reduction of the dimensionality even more serious than they are in the case of nominal labels. A new problem, discussed in Section 2, is the possibility of an instable classification result. In Section 4 an heuristic solution is presented in the context of an example. However, in spite of these difficulties, continuous labeling can be applied succesfully, provided that the learning set can be made large enough. If this condition is fulfilled, techniques as described in this paper can be used for both, estimating unknown variable values (labels) and introducing continuous labels for classifying objects performing more flexibility during learning and testing.

Acknowledgment

The research project described in Section 4 is a corporation with E.H. Furnee of the Laboratory for Prosthesis Control, Delft University of Technology, The Netherlands. The measurements were set up and performed by J. den Boer and M. Loose.

References

Draper, N.R. and H. Smith (1966), *Applied Regression Analysis.* Wiley, New York.

Wirta, R.W. and D.R. Taylor (1970) Development of a multiple-axis myoelectrically controlled prosthetic arm. In: M.M. Gravilovic and A.B. Wilson eds., *Advances in External Control of Human Extremities.* Yugoslav committee for electronics and automation, Belgrado, pp. 245–253.

Duin, R.P.W., J.A. de Vos, J.G.M. Bakker, R.R. de Boer and E.H. Furnee (1977). The recognition of muscle activity patterns for prosthesis control. *Proc. SITEL-ULG Seminar on Pattern Recognition.* Liège, Belgium, November, pp. 811–819.

Den Boer, J. (1980), A new model for controlling an armprosthesis by EMG signals. Thesis, Delft University of Technology, August. [In Dutch.]

Beukema toe Water, F.T. and R.P.W. Duin (1981), Dealing with a priori knowledge by fuzzy labels. *Pattern Recognition* 14, 111–115.

Hermans, J. and J.D.F. Habbema (1975), Comparison of five methods to estimate a posteriori probabilities. *EDV Med. Biol.* 6, 14–19.