Linear Regression for Dimensionality Reduction and Classification of Multi Dimensional Data*

Lalitha Rangarajan and P. Nagabhushan

Department of Studies in Computer Science,
University of Mysore, Mysore, India
lali85arun@yahoo.co.in, pnagabhushan@hotmail.com

Abstract. A new pattern recognition method for classification of multi dimensional samples is proposed. In pattern recognition problems samples (pixels in remote sensing) are described using a number of features (dimensions/bands in remote sensing). While a number of features of the samples are useful for a better description of the image, they pose a threat in terms of unwieldy mass of data. In this paper we propose a method to achieve dimensionality reduction using regression. The method proposed transforms the feature values into representative patterns, termed as symbolic objects, which are obtained through regression lines. The so defined symbolic object accomplishes dimensionality reduction of the data. A new distance measure is devised to measure the distances between the symbolic objects (fitted regression lines) and clustering is preformed. The efficacy of the method is corroborated experimentally.

Keywords: Pattern Classification, Dimensionality Reduction, Feature Sequence, Regression, Clustering, Data Assimilation, Multi Dimensional Data, Symbolic Data.

1 Introduction

Measurements made in pattern recognition applications are inherently multi dimensional in nature. Larger the numbers of features, more severe are the problems of storage and analysis time requirements. It is a tacit assumption that all these features are useful in obtaining a good classification [2]. In fact, it is more rewarding to use a few significant features or their functions in suitable combinations [1]. Further, reducing the feature set to two or three facilitates visual inspection of the data. Aim of dimensionality reduction is to describe the samples by means of minimum number of features that are necessary for discrimination of objects (patterns) in the image. A variety of methods for dimensionality reduction are proposed in the literature [many described in 2]. Most of these belong to subsetting methods or feature space transformation methods.

In pattern recognition applications, each sample is described using a set of values (features), the size of the set being the dimension of the data set. Each sample can be viewed as quantitative multi valued symbolic variable, as termed in an advanced

* This research is supported by ISRO project RESPOND 10/4/317, Oct 1999.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 193 – 199, 2005.

exploratory data analysis technique called Symbolic Data Analysis [3]. The gist of symbolic approach in pattern classification is to extend the problems and methods defined on classical data to more complex data called 'symbolic objects', which are well adapted to represent knowledge [3: Michalski, 1981; Diday, 1988, 1989b; Tu Bao Ho et al., 1988]. A symbolic object is a description of the properties of a set of elementary objects. The description of symbolic objects may depend on the relations existing between elementary objects, frequency of occurrence of values [3: Bertrand and Goupil, 2000] and so on. Each variable of a symbolic object may take a set of values [6, 1994; 3: De Carlvaho, 1998] or an interval of values [3: Gowda and Diday, 1991] or even a sub object [6, 1988; 3: Gowda and Diday, 1991]. Symbolic representation of multi dimensional temporal observations can be found in [5, 10].

Each sample in our data set is described by a set of feature values. Ichino, Yaguchi and De Carlvaho have suggested interesting metric on set type symbolic data. These distance measures are inadequate to determine the distance between two samples. We have introduced a new symbolic data type of the set of feature values and a suitable distance measure. We have made an attempt to summarize the feature values of a sample into a new symbolic object namely "regression curves". This results in data assimilation of feature values of a sample. [9] is concerned with the dimensionality reduction of temporal data through regression lines. Here lines of best fit are identified for temporal observations of a specific feature. However in the proposed method we have achieved dimensionality reduction by assimilating the features of a sample into a regression curve. Perhaps feature sequence may affect the nature of best fit.

Section 2 describes the method proposed. In Section 3 we have described computation of distance between two regression lines, which is the foundation for clustering samples, and clustering procedure. Experiments and results are discussed in Section 4. We have concluded the paper with suggestions for further improvements in Section 5.

2 Description of the Method

In an n dimensional image the collection of all n feature values $\{f_1, f_2, \ldots, f_n\}$ decide the class the sample belongs to. This collection can be regarded as quantitative multi valued symbolic variable (set) [3]. The distance between such symbolic variables (sets) can be measured using (i) the number of total elements in both the sets (Cartesian join $\oplus/\cup$) and (ii) the number of common elements in both the sets (Cartesian meet $\otimes/\cap$) [6]. The distance between features can also be measured using agreement ($\cap$ of both sets) and disagreement ($\cap$ of a set and complement of the other set) as suggested by [3: De Carvalho, 1998] The component distances can be aggregated with generalized Minkowski's metric. Both these distance measures are not suitable for our symbolic object although our data is also multi valued quantitative. Our 'set' is sensitive to arrangement of elements within the set, unlike the 'set' in [3]. That is the following two samples x and y described by the features sets namely, {1.6, 2.4, 4.7, 10.8} and {10.8, 4.7, 2.4, 1.6} are not identical and also unlikely to be from the same class. Our features are not only sets but also vectors. Hence the symbolic representation should take care of not only the set contents but also the arrangement of the set elements.

We have used 'regression curves' for describing each feature of a sample. This symbolic transformation can take care of both the contents of the set and the arrangement of the set elements. For the above example of feature values of samples x and y, our symbolic objects are regression curves fitted for points (1,1.6), (2, 2.4), (3, 4.7), (4, 10.8) and for points (1, 10.8), (2, 4.7), (3, 2.4), (4, 1.6). These regression curves are entirely different. We have used 'least square error' regression curves to symbolize features of a sample. The transformation results in the assimilation of f features into a regression curve.

A good regression curve provides an apt description of the samples. But it is difficult to determine the nature of regression curves. Even if we do, the feature values of different samples could yield different types of regression curves. For instance the best regression curves of two samples may become a quadratic curve and a line. This amounts to keeping track of type of regression curves and also the parameters of regression curves. This results in too much of book keeping, making the problem of classification too tedious. Probably this may even contradict the possibilities of the very theme of the research, that is, dimensionality reduction and data assimilation. Therefore all samples are represented by symbolic data object namely "least square error regression lines" fitted for points $(1, f_1)$, $(2, f_2)$, …, (n, f_n). Our goal is not a better description of pixels (samples), but a description that is good enough that retains the vital information hidden in the feature values so that the classification is truthful. Our regression lines for points $(1,f_1)$, $(2,f_2)$, …, (n, f_n) possesses this property. A new distance measure has been proposed to measure the distance between the set of regression lines similar to the one in [9], where dimensionality reduction of multi temporal data using regression is discussed.

Samples belonging to different classes differ significantly in atleast one dimension. The corresponding regression lines reflect this significant difference in feature values. This is illustrated in the figure 1. Samples belonging to same class have the respective feature values close to each other and hence the regression lines are close too. An example of this case is illustrated in the figure 2. Observe that the regression lines of samples belonging to same class are close to each other only in the interval [1,n]. The lines may differ significantly beyond this interval. Yet the samples are similar.

Suppose that the data items are represented by d[i,j] where i is the sample and j is the feature and $1 \leq i \leq s$, $1 \leq j \leq f$. Regression lines are fitted for the points (1, d[i,1]), (2, d[i,2]), …, (x,d[i,x]), …, (f, d[i,f]). The above process of fitting regression lines is repeated for all samples. In the end we have s regression lines. The number of features of a sample is reduced from f to 2.

3 Computation of Distance Measure for Clustering and Sequencing the Features

The distance between the samples m,n is defined to be maximum of the lengths LmLn, RmRn (refer figure 3). Here Lm, Ln are the points of intersections of the regression lines of samples m, n with the (left) ordinate at x=1 and Rm, Rn are the points of intersections of the regression lines of samples m, n with the (right) ordinate at x=f. The new reduced set of feature values of the sample m are Lm, Rm. Our

distance measure is a supremum distance measuring dissimilarity between samples. It is obvious from figure 4 that the distance measure is a metric.

Sequencing of Features

Suppose that there exists a sequence of features such that they are approximately linear (true for all classes) and is according to the order of variances. Then we have an ideal and most preferred scenario for our transformation into symbolic data, namely regression lines (refer figure 5). But in reality this may not be possible. That is sequencing features according to variances may not preserve linearity for all classes. As we measure distances between samples by looking at ordinates at 1 and f, we require an ordering that would discriminate samples from distinct classes at either the ordinate at 1 or at f. The feature with maximum variance is likely to discriminate many classes. The exhaustive experimentation suggests that the best sequence of features is according to variance. The way the first few PCs are employed for classification, support the argument on sequencing the features based on magnitude of variances. However if classes are well separable any sequence is good enough.

For figures 1 to 6 feature numbers are along X axis and feature values along Y

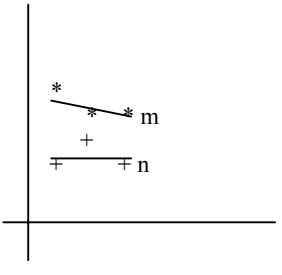

Fig. 1. m,n from different class

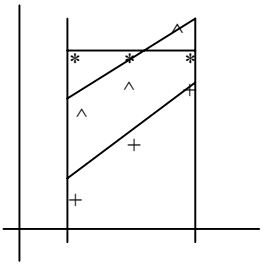

Fig. 2. m,n from same class

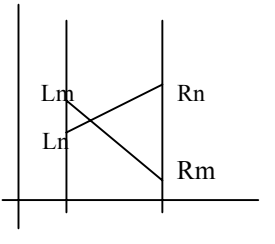

Fig. 3. Distance between m,n dis(m,n)=max{LmLn, RmRn} =RmRn

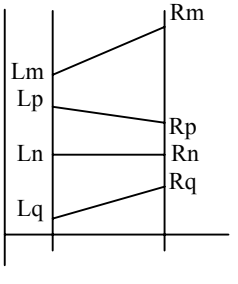

Fig. 4. Dist measure is metric dis(m,n)=dis(m,p)+dis(p,n) dis(m,n)<dis(m,q)+dis(q,n)

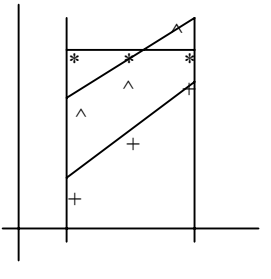

Fig. 5. Favorable scenario Features according to order variances and linear. Classes well separable at f=1

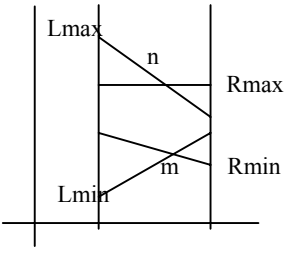

Fig. 6. First two centers m,n LminLmax > RminRmax Initial two centers: m,n

4 Experiments

To illustrate the feasibility of the proposed method we conducted experiments on supervised data sets (Iris and synthetic remote sensed image).

The first experiment is on a well separable subset (size 30, 10 from each class) of Iris data with 3 classes. Our method and PCA yield the same result (misclassification is nil). Also classification is found to be independent of feature sequence.

The second experiment is on a synthesized remotely sensed image. The synthesized image (figure 7) is of size 50 x 50 pixels, consisting of 7 classes found in the Coorg forest area in the state of Karnataka, India. The classes are arranged as found in nature and spectral responses generated randomly to be within the predetermined ranges [11]. After transformation and reduction of bands, we performed clustering in two rounds. Initially we made a partition of the data into twice the number of expected clusters. The reason for making too many clusters is that some classes in the data set are hardly separable. The regression lines of such classes tend to average out the small differences that exist in feature values. With more clusters such classes also become distinct (with the new transformed reduced set of features). Initial two centers are selected to be the two most distant regression lines

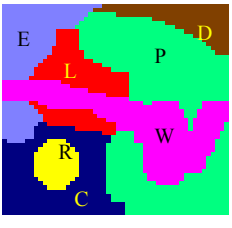

Fig. 7. Synthetic image E-Evergreen forest W-Water R-Rubber P-Paddy C-Coffee L-Cropland D-Deciduous forest

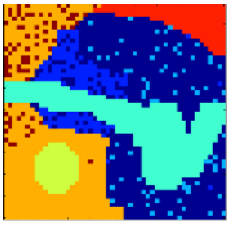

Fig. 8. Classification map of image in fig 7 using first 2 PCs

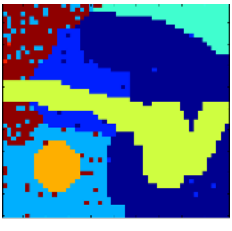

Fig. 9. Classification map of image in fig 7 using proposed method

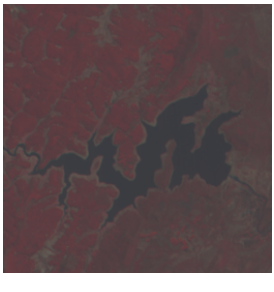

Fig. 10. Registered IRS 1A LISS II B1 – FCC of Coorg forest cover. Size: 260x260 Date: Feb 3rd 1991

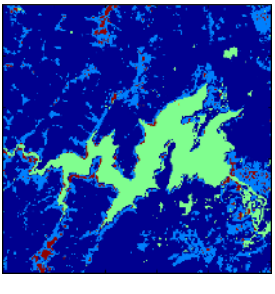

Fig. 11. Classification map of image in fig 9 using first two PCs

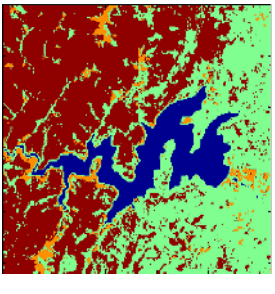

Fig. 12. Classification map of image in fig 9 using proposed method

(figure 6). Subsequent centers are selected to be the regression lines of the samples that are beyond a threshold distance from all previous centers. The pre specified distance is chosen to give twice as many centers as there are number of classes in the data set. Some amount of misclassification observed only in classes that are overlapping namely Paddy - Cropland and Evergreen - Coffee both with our method and using PCA. These overlapping classes are better separated with our method, than with PCA. The misclassification is 4.4% with our method. Clustering with first two PCs resulted in misclassification of 15.1% (misclassification is more when clustering is done with more or less than 2 PCs). Out of all sequence of features the order of features according to variances is found to outperform any other sequence. Figures 8 and 9 are the classification maps of the synthetic image using first two PCs and the proposed method.

The proposed method is tested on an image from which the synthetic image was derived. Figure 10 is registered IRS 1A image of Coorg forest cover. Classification maps using PCs and the proposed methods are given in figures 11 and 12.

5 Conclusion

The method suggested in this paper performs dimensionality reduction of multi dimensional data. Feature values of a sample are assimilated in the form of a regression line, thus reducing the dimensions from f (the number of features) to 2. A new distance measure is devised to measure distances between the regression lines. Experiments demonstrated the ability of the method to produce fairly good classifications. The method is very versatile and can be used with any multi spectral data. The method is simple and can perform well on classes that are separable. A disadvantage of the method is that the regression lines tend to average out the existing small differences between the classes. Perhaps the method suggested can be used as pre clustering procedure on data sets with large number of classes to isolate well separable few classes and then coarse clusters obtained may be clustered again using original features. Also if the number of features is too many a single regression line for a sample may bring two well-separated classes together. A higher order regression curve may be a better fit. Such as arbitrary curve can be approximated by piecewise linear segments [4,8], the higher order regression curve can be approximated by piece wise regression line segments. Probable solution is to go for slicing number of features and multiple regression lines for each sample as done in [9] for temporal data.

References

1. Dallas E. Johnson, “Applied Multivariate Methods for Data Analysis”, Duxbury Press (1998)
2. P.Nagabhushan, “An effient method for classifying Remotely Sensed Data (incorporating Dimensionality Reduction)”, Ph.D thesis, University of Mysore, Mysore, India (1988)
3. H.H.Bock, E.Diday (editors), “Analysis of Symbolic Data”, Springer Verlag, (2000)
4. Dunham J.G, “Piecewise linear approximation of planar curves”, IEEE Trans. PAMI, vol 8 (1986)

5. Getter-Summa, M., MGS in SODAS, Cahiers du CEREMADE no. 9935, Universite' Paris IX Dauphine, France (1994)
6. Ichino, M., Yaguchi, H., Generalized Minkowski metrics for mixed feature type data analysis, IEEE Trans. Systems Man Cybernet, 24(4) (1994)
7. Jolliffee, I.T., Principal Component Analysis, Springer Verlag, NY (1986)
8. Leung, M.K., Yang, Y.H., Dynamic strip algorithm in curve fitting, Computer Vision Graphics and Image Processing, 51 (1990)
9. Lalitha Rangarajan, Nagabhushan, P., Dimensionality reduction of multi dimensional temporal data through regression, J of PRL, Elsevier, vol 25/8 (2004), 899 - 910
10. Srikantaprakash, H.N., Nagabhushan, P., Gowda, K.C., Symbolic data analysis of multi spectral temporal data, IEEE International Geosci and Remote Sensing symposium, Singapore (1997)
11. Lalitha Rangarajan, Nagabhushan, P., Content driven dimensionality reduction at block level in the design of an efficient classifier for multi spectral images, J of PRL, Elsevier, vol 25 (2004), 1833 - 1844