ELSEVIER

CrossMark

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 122 – 130

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

An Iterative Projective Clustering Method

Renata Avros[a], Zakharia Frenkel[a]*, Dvora Toledano-Kitai[a] and Zeev Volkovich[a]

[a]Ort Braude College of Engineering, Karmiel 21982, Israel

Abstract

In this article we offer an algorithm recurrently divides a dataset by search of partitions via one dimensional subspace discovered by means of optimizing of a projected pursuit function. Aiming to assess the model order a resampling technique is employed. For each number of clusters, bounded by a predefined limit, samples from the projected data are drawn and clustered through the EM algorithm. Further, the basis cumulative histogram of the projected data is approximated by means of the GMM histograms constructed using the samples' partitions. The saturation order of this approximation process, at what time the components' amount increases, is recognized as the "true" components' number. Afterward the whole data is clustered and the densest cluster is omitted. The process is repeated while waiting for the true number of clusters equals one. Numerical experiments demonstrate the high ability of the proposed method.

Keywords: Projective Clustering; Projected Pursuit; Image Segmentation

1. Introduction

General clustering procedures applied for high-dimensional data classification are frequently based on the Gaussian Mixture Model (GMM). Such model can expose an unsteady performance when the size of the considered dataset is overly small compared to the number of parameters to estimate or when an expected partition is composed

* Corresponding author. Tel.: +972-4-990-1764; fax: +972-4-990-1852.
E-mail address: zfrenkel@braude.ac.il

doi:10.1016/j.procs.2015.08.111

from clusters with significantly diverse sizes. To avoid these difficulties a projection clustering approach can be applied aiming to find a balance between the parameters' quantity and the generality of the replica. Although the unrelated characteristics of the projected set may actually "hide" the clusters by imaging two items belonging to the same cluster observe as dissimilar as an arbitrary couple of items. Likewise, items could cluster inversely in varying subspaces.

The projected clustering concept presumes that consequential partition can be discovered by projecting onto subspaces of lower dimensionality. Practically the most of existing projected clustering algorithms (see e. g. [1-5]) are definitely based on an assumption that underlying clusters are depicted by areas of the data of high density separated by sparse areas. This fact is expressed by separated "picks" or "islands" in a subspace corresponding to the overall density of the full space. Hence, the seeking of attractive cluster structure in the high dimensional space can be altered by a corresponding procedure in lower dimensional subspaces. Formally speaking, interestingness is measured by a distance between the distribution of the projected data and a distribution of recognized as uninteresting, which is typically suggested to be normal. So, (see [6-7]) any test statistic for testing non-normality (or departure from normality) might be applied as a projection index, quantifying the "interestingness".

In this article we offer a one-dimensional projection pursuit algorithm in the framework of the general Gaussian Mixture Model (GMM) supposing that each cluster is represented by a Gaussian probability density. Note that each projection of GMM distributed data is also GMM distributed. The parameters of GMMs are mostly estimated by the well-known Expectation Maximization Algorithm (EM) finding a maximum likelihood solution. A weakness of this fitting method consists of poor functioning, once high-dimensional data are operated as a large sample size is required in order to attain the required precision.

We propose here a hierarchical projective clustering approach in the spirit of the mentioned earlier projection pursuit perspective based on searching of appropriate one-dimensional subspaces (directions). Note that such an approach is naturally connected to the color space optimizations (see, e.g. [8]) where an image transformation is constructed in a way that saves as much of the information as possible from the source space though remaining as authentic as possible to the natural mapping. Actually, such a transformation is appearing in our approach as a weighted sum of the three linear-intensity values with the weights evaluated via the clustering projections goodness. The algorithm recurrently divides a dataset by search of partitions via one dimensional subspace discovered by means of optimizing of a projected pursuit function. Aiming to assess the model order (the components' quantity) a resampling technique is employed. For each number of clusters, bounded by a predefined limit, samples from the projected data are drawn and clustered through the EM algorithm. Further, the basis cumulative histogram of the projected data is approximated by means of the GMM histograms constructed using the samples' partitions. The saturation order of this approximation process, at what time the components' amount increases, is recognized as the "true" components' number. Afterward the whole data is clustered, by the EM algorithm with this found number of clusters, and the densest cluster is omitted. The process is repeated while waiting for the true number of clusters equals one.

The paper is organized as follows: In section 2 we present proposed method and discuss its ingredients: the Gaussian Mixture model and the closely connected EM algorithm, Criteria for Projections Selecting and the model selection method. The remaining sections are devoted to the numerical experiments consisting of an application of the presented method to image segmentation and to conclusion.

Nomenclature

GMM	Gaussian Mixture Model
EM	Expectation Maximization
KS	Kolmogorov-Smirnov

2. Proposed method

The following pseudo-code describes the meta-algorithm of the suggested method. Now assume that the observed multivariate data $X = \left(X_1,...,X_N\right)$, have been generated from a mixture Gaussian distributions $\left(N_i,\ i=1,...,k\right)$ on the Euclidian space R^m

Meta-Algorithm:

```
1. For given data $X$ do:
2. Discover an appropriate direction $d^*$ by optimizing of a projection pursuit function $I(d)$;
3. Project the data on the direction $d^*$;
4. Estimate the number of clusters $k^*$ of the projected data in the framework  the GMM (model order selection step);
5. Partition the projected data with the EM algorithm into $k^*$ clusters;
6. Uncover the most concentrated cluster of the found partition and exclude it from the data;
7. If the stopping criterion is met then stop, else return to step 1.
```

Since, the partitions are being created in the GMM framework, the most concentrated cluster is expressed by the minimal cluster inner standard deviation. Let us consider in details an implementation of the supplementary method ingredients.

2.1. GMM and EM algorithm

The Gaussian Mixture Model (GMM) is a probabilistic replica that supposes all the data are produced from a mixture of a finite number of Gaussian distributions with undiscovered parameters (see, e.g. [9]). In the one-dimensional case GMM assumes that the probability density function of X is follows:

$$f(x) = \sum_{i=1}^{k} w_i N(x \mid \mu_i, \sigma_i),$$

where k is the components number, $N(\bullet \mid \mu_i, \sigma_i)$ is a normal probability density function of the i – th component, having the mean value μ_i, the standard deviation σ_i, $i=1,...,k.$; w_i are components' weights, $i=1,...,k$. The fit of the GMM model to the data can be evaluated by the log likelihood function of the data. Maximizing this function can be done using standard, iterative, numeric optimization methods or by EM algorithm to maximize log-likelihood:

$$ln(f(X)) = \sum_{n=1}^{N} ln\left\{\sum_{i=1}^{k} w_i N(x_t \mid \mu_i, \sigma_i)\right\}.$$

The consequences of the EM algorithm are very sensitive to the initial values of the parameters due to local maxima of the likelihood function. At the algorithm initialization an auxiliary set $Z=\{z_{it},\ i=1,...,k,\ t=1,..,N\}$ is introduced as a vector of k binary indicator variables that are mutually exclusive and exhaustive (i.e., one and only one of the z_{in}'s equals to 1, and all the others are 0, for a given n). $\{z_{in},\ i=1,...,k\ \}$ is an array representing the identity of the mixture component including x_n. Then, in every EM step, the succeeding calculations are performed which assure a monotonic growth in the likelihood value:

- Mixture Weights Calculation (E-Step):

$$\overline{w_i} = \frac{1}{N}\sum_{n=1}^{N} z_{in},\ i=1,...,k.$$

- M-Step:
  - Means Calculation:

$$\overline{\mu_i} = \frac{\sum_{n=1}^{N} z_{in} x_n}{\sum_{n=1}^{N} z_{in}},\ i=1,...,k.$$

  - Variances Calculation:

$$\overline{\sigma_i^2} = \frac{\sum_{n=1}^{N} z_{in} x_i^2}{\sum_{n=1}^{N} z_{in}} - \overline{\mu_i^2},\ i=1,...,k.$$

  - *A posteriori* Probabilities Calculation:

$$z_{in} = \frac{\overline{w_i} \boldsymbol{N}(x_n \mid \overline{\mu_i}, \overline{\sigma_i})}{\sum_{m=1}^{k} \overline{w_m} \boldsymbol{N}(x_n \mid \overline{\mu_m}, \overline{\sigma_m})},$$

$$i = 1,...,k,\ \ n = 1,..,N.$$

After the new parameters values have been calculated, the M-step is finished, and the process returns to the E-step to recalculate the membership weights, and so on. The steps are executed until the parameters discontinue changing.

2.2. Criteria for Projections Selecting

We would like to operate with a criterion for choosing an appropriate projecting direction which is comfortable to compute, even if the data dimension and the sample size are sufficiently large. From the clustering standpoint, an interesting direction is someone producing projected clusters placed nearby well separated midpoints. In this case, (see [10]), a multimodal projected distribution predictable occurs. Hence, a purposeful criterion appears to be like seeking for directions exploiting the bimodality distribution property. Such an attitude was actually adopted in [11] where the directions are selected to minimize the kurtosis of the projected data:

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{argmin}\left(I_1(d) = kurtosis\left(<\boldsymbol{X}, \boldsymbol{d}>\right)\right) \ . \quad (1)$$

On the other hand, under some weak assumptions [12] distributions of linear projections can be reflected as approximately normal in the high dimension case; i.e., practically speaking, most projections are approximately Gaussian distributed. According to the well-known Cramer-Wold principle, if all one-dimensional projections are normal then the underlying distribution is normal. So, a statistic measuring "departure from normality" can be used as a pursuit index. Resting upon this concern we employ also the following index:

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{argmax}\left(I_2(d) = KS\left(<\boldsymbol{X}, \boldsymbol{d}>\right)\right), \quad (2)$$

where KS stands the Kolmogorov-Smirnov quantifies a distance from normality (see e.g. [13]). To avoid degenerative solutions we constrain these optimization tasks as follows:

$$\text{s.t.}: \|\boldsymbol{d}\| = 1. \quad (3)$$

2.3. Model selection

This “ill posed” cluster analysis problem can bear more than one answer [14-15]. A review of clustering model selection methods is presented, for example, in [16]. The offered approach employs the stability attitude; more explicitly: a stable approximation of the underlining distribution by means of the GMM learned from samples. For cluster amounts, bounded by a predefined upper limit, we extract samples from the projected data and divide them using the EM algorithm. The underlying cumulative data function is approximated with the GMM ones resulted from the samples’ partitions obtained. The goodness of fit is evaluated by the Kolmogorov-Smirnov sample test distance (KS) from normality averaged over all drawn samples. Evidently, the goodness of fit is being improved if the number of components grows to the correct one, and it is getting worse after. So, the process saturation appears when the number of components corresponds to the most stable GMM state.

Model Selection Algorithm
Input

- X- projected data;
- K_{max}-maximal number of components to be tested;
- $NSAMP$-number of samples to be drawn from the data;
- $NSIZE$-fraction of the drawn samples in the current data;
- ε-saturation threshold.

Algorithm

```
Construct the cumulative distribution $\boldsymbol{F}$ of $\boldsymbol{X}$.
1.  For $k=1$ to $K_{max}$
2.      For $n=1$ to $NSAMP$
3.        $S_n=SAMPLE(X,NSIZE)$
4.        Construct the GMM cumulative distribution $G_n$ partition of $S_n$ obtained by the EM
          algorithm.
5.        Calculate a distance statistic $Dis_n = KS\left(F,G_n\right)$.
6.      End
7.  Calculate $C(k)$=mean$(Dis_n,\ n=1,...,NSAMP)$
```

A suitable number of components k^* is given by the first saturation point of $C(k)$:

$$k^* = \underset{k}{agrmin}\left(\left(C\left(k\right)-C\left(k+1\right)\right)<\varepsilon\right).$$

3. Applications to image segmentation

We provide several experiments in order to demonstrate the capability of the proposed method on images. The following parameters’ values are used:

- $K_{max} = 10$;
- $NSAMP = 50$;
- $NSIZE$ = 10% of the remaining data;
- $\varepsilon = 0.001$.

In the pre-processing step the one dimensional Haar discrete wavelet transform (DWT) is applied to gain high level details in the projected data. After using DWT the approximation coefficients matrix is nulled, and the result of the inverse DWT is added to the projected data. Such a technique exploiting the advantage of the noise-robust nature of wavelets is actually applied in order to contrast images enhancement (see, e.g. [17]). However, we apply it here in a slightly different manner aiming to stress separation of the high density regions. The segmentation procedure is performed twice: with the two introduced projection pursuit functions defined in (1) and (2).

Example 1

The first image considered is presented in Fig. 1 and Fig. 2 with its final segmentations obtained by means of the proposed method.

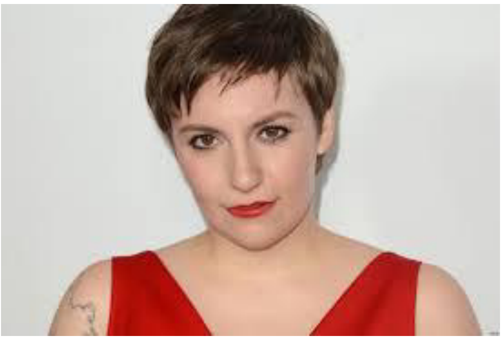

Fig. 1 The first segmented image

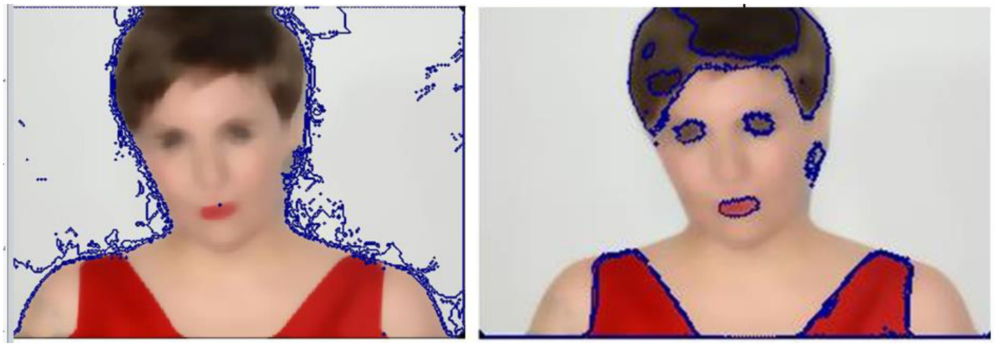

Fig. 2 The final segmentations of the first segmented image performed by means of two pursuit functions

As have been seen, the second index provides more accurate segmentation of the image. Let us consider the evolution of this process in the second case. Actually, just three segmentations were performed. The directions and the numbers of the components chosen are presented in the following table (Table 1).

Table 1 Characteristics of the segmentation process

Iteration numbers	d(1)	d(2)	d(3)	Number of the components
1	0.0506	0.7754	-0.6294	2
2	0.0357	0.9412	-0.3359	2
3	0.6092	0.5077	-0.6092	3

The following figure (Fig. 3) demonstrates the approximation of the underlying cumulative projection function (marked in red) via the 50 sample cumulative ones in the case where the suggested components number equals to 1.

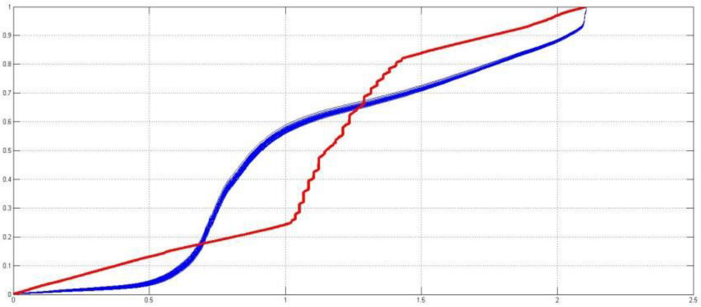

Fig. 3 Approximation of the underlying cumulative projection function (marked in red) via the 50 sample cumulative ones for *k=1*

The two first iterations of the segmentation process are also exhibited in Fig. 4 to 5 where the omitted clusters are marked in red.

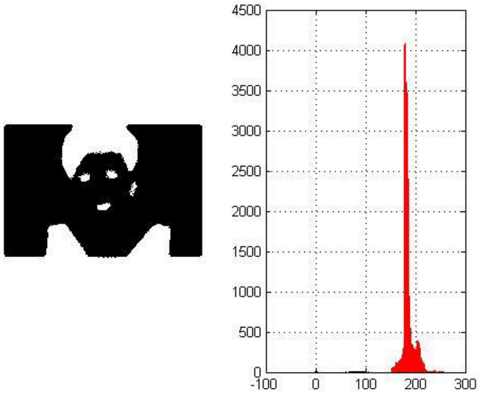

Fig. 4 The first segmentation iteration

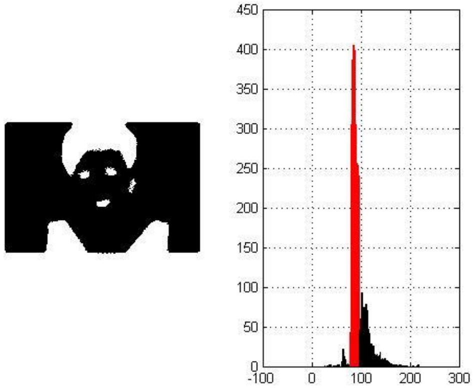

Fig. 5 The second segmentation iteration

Example 2

Here a brain MRI is considered (Fig. 6)

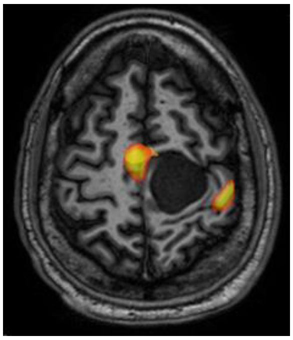

Fig.6 The first segmented image

Note that this image appears to be more complicated in comparison with first one. Particularly, two relatively small objects, supposedly potential tumours, are presented. However our algorithm is capable to recognize them (Fig. 7).

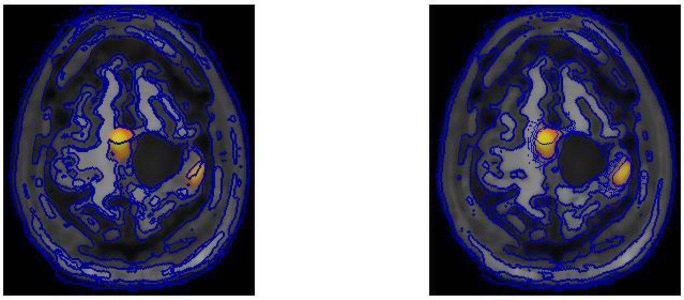

Fig. 7 The final segmentations of the second segmented image performed by means of two pursuit functions.

The segmentation process with the second pursuit function passed 8 iterations (Table 2).

Table 2. Characteristics of the segmentation process

Iteration numbers	d(1)	d(2)	d(3)	Number of the components
1	0.8130	0.5641	0.1442	4
2	0.8632	0.4773	-0.1645	4
3	0.8610	0.4942	-0.1205	3
4	0.8628	0.4810	-0.1552	2
5	0.6380	0.1160	-0.7612	2
6	0.6980	0	-0.7161	5
7	0.6078	0.1421	-0.7812	2
8	0.6103	0.1501	-0.7778	3

Please note that the process has adapted at each step in order to get the best partition and eventually found the smallest details in the picture. The number of clusters at 8-th iteration is not 1, but the process has been stopped due to the small amount of items intended to be clustered at the next step.

4. Conclusion

We proposed a new a one-dimensional projection pursuit algorithm in the framework of the general Gaussian Mixture Model. The algorithm progressively removes data elements which divert the emphasis of attention. The algorithm is capable to reveal data groups owning different sizes. The numerical experiments demonstrate a high ability of the proposed model and we plan to intensify its evaluation and to compare its performance to those of other models

References

1. Aggarwal CC, Procopiuc C, Wolf JL, Yu PS, and Park JS. Fast algorithms for projected clustering. *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*,1999. ACM New York, p. 61-72.
2. Procopiuc CM, Jones M, Agarwal PK., Murali TM. A Monte Carlo algorithm for fast projective clustering. *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, ACM New York, 2002. p. 418-427.
3. Yip KY, Cheung DW, Ng MK. On discovery of extremely low-dimensional clusters using semi-supervised projected clustering, *Proceedings of Data Engineering ICDE 2005,* IEEE, 2005. p. 329-340.
4. Ng E., Fu A., Wong R. Projective Clustering by Histograms. *IEEE Transactions on Knowledge and Data Engineering*, 2005; **17**(3):369-383.
5. Moise G., Sander J., Ester M.: "P3C: A Robust Projected Clustering Algorithm", *Proceedings of. 6th IEEE Int. Conf. on Data Mining (ICDM 2006)*, Hong Kong, 2006. p. 414-425.
6. Friedman JH., Tukey JW. A Projection Pursuit Algorithm for Exploratory Data Analysis, *Ann. Statist.* 1985;**13**:435–475.
7. Huber PJ. Projection Pursuit, *Ann. Statist*, 1985;**13**: 435–475.
8. Lau C, Heidrich W, Mantiuk R. Cluster-Based Color Space Optimizations, *Proceedings of IEEE International Conference on Computer Vision, 2011*, p. 1172-1179.
9. Hartley H. Maximum likelihood estimation from incomplete data. *Biometrics*, 1958:**14**:174–194.
10. Switzer P. Comments on "Projection Pursuit," by P. J. Huber, *Ann. Statist.*, 1985:**13**:515–517.
11. Peña D and Prieto FJ, Cluster Identification Using Projections. *Journal of the American Statistical Association,* 2001: 96: 1433 – 1445.
12. Diaconis P, Freedman D. Asymptotics of Graphical Projection Pursuit. *Ann. Statist*. 1984:12:3: 793-815.
13. Stephens MA EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*, 1974;**69**:730–737.
14. Gordon AD., *Classification*, Chapman and Hall, CRC, Boca Raton, FL, 1999.
15. Jain A, Dubes R, *Algorithms for Clustering Data*, Englewood Cliffs, Prentice-Hall, New Jersey; 1988.
16. Volkovich Z, Barzily Z, Weber GW, Toledano-Kitai D, Avros R, Resampling Approach for Cluster Model Selection, *Machine Learning*, 2011;**85**(1-2):37-43.
17. Valliammal N, Geethalakshmi SN. Leaf Image Segmentation Based On the Combination of Wavelet Transform and K Means Clustering, *International Journal of Advanced Research in Artificial Intelligence*, 2012;**1**(3):97-194.