A Robust Approach for Multivariate Binary Vectors Clustering and Feature Selection

Mohamed Al Mashrgy, Nizar Bouguila, Khalid Daoudi

A Robust Approach for Multivariate Binary Vectors Clustering and Feature Selection

Mohamed Al Mashrgy[1], Nizar Bouguila[1], and Khalid Daoudi[2]

[1] Concordia University, QC, Cannada
m_almash@encs.concordia.ca, bouguila@ciise.concordia.ca
[2] INRIA Bordeaux Sud Ouest, France
khalid.daoudi@inria.fr

Abstract. Given a set of binary vectors drawn from a finite multiple Bernoulli mixture model, an important problem is to determine which vectors are outliers and which features are relevant. The goal of this paper is to propose a model for binary vectors clustering that accommodates outliers and allows simultaneously the incorporation of a feature selection methodology into the clustering process. We derive an EM algorithm to fit the proposed model. Through simulation studies and a set of experiments involving handwritten digit recognition and visual scenes categorization, we demonstrate the usefulness and effectiveness of our method.

1 Introduction

The problem of clustering, broadly stated, is to group a set of objects into homogenous categories. This problem has attracted much attention from different disciplines as an important step in many applications [1]. Finite mixture models have been widely used in pattern recognition and elsewhere as a convenient formal approach to clustering and as a first choice off the shelf for the practitioner. The main driving force behind this interest in finite mixture models is their flexibility and strong theoretical foundation. The majority of mixture-based approaches have been based on the Gaussian distribution. Recent researches have shown, however, that this choice is not appropriate in general especially when we deal with discrete data and in particular binary vectors. The modeling of binary data is interesting at the experimental level and also at a deeper theoretical level. Indeed, this kind of data is naturally and widely generated by various pattern recognition and data mining applications. For instance, several image processing and pattern recognition applications involve the conversion of grey level or color images into binary images using filtering techniques. A given document (or image) can be represented by a binary vector where each binary entry describes the absence or presence of a given keyword (or visual word) in the document (or image) [2]. An important problem is then the development of statistical approaches to model and cluster such binary data.
Several previous researches have addressed the problem of binary vectors classification and clustering. For example, a likelihood ratio classification method

based on Markov chain and Markov mesh assumption has been proposed in [3]. A kernel-based method for multivariate binary vectors discrimination has been proposed in [4]. A fuzzy sets-based clustering approach has been proposed in [5] and applied for medical diagnosis. An evaluation of five discriminations approaches for binary data has been proposed in [6]. A multiple cause model for the unsupervised learning of binary data has been proposed in [7]. Recently, we have tackled the problem of unsupervised binary feature selection by proposing a statistical framework based on finite multivariate Bernoulli mixture models which has been applied successfully to several data mining and multimedia processing tasks [8, 2]. In this paper, we go a step further by tackling simultaneously , with clustering and feature selection, the challenging problem of outlier detection. We are mainly motivated by the fact that learning algorithms should provide accurate, efficient and robust approaches for prediction and classification which can be compromised by the presence of outliers as shown in several research works (see, for instance, [1, 9]). To the best of our knowledge the well-known binary data clustering algorithms offer no solution to the combination of feature selection and outlier rejection in the case of binary data.
The rest of this paper is organized as follows. First, we present our model and an approach to learn it in the next section. This is followed by some experimental results in Section 3 where we give results on a benchmark problem in pattern recognition namely the classification of handwritten digits and in a second problem which concerns visual scenes categorization. Finally, we end the article with some conclusions as well as future issues for research.

2 A Model for Simultaneous Clustering, Feature Selection and Outliers Rejection

In this section we first describe our statistical framework for simultaneous clustering, feature selection and outliers rejection using finite multivariate Bernoulli mixture models. An approach to learn the proposed statistical model is then introduced and a complete EM-based learning algorithm is proposed.

2.1 The Model

Let $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\} \in \{0,1\}^D$ be a set of D-dimensional binary vectors. In a typical model-based cluster analysis, the goal is to find a value $M < N$ such that the vectors are well modeled by a multivariate Bernoulli mixture with M components:

$$p(\boldsymbol{X}_n|\Theta_M) = \sum_{j=1}^{M} p_j p(\boldsymbol{X}_n|\boldsymbol{\pi}_j) = \sum_{j=1}^{M} p_j \prod_{d=1}^{D} \pi_{jd}^{X_{nd}} (1-\pi_{jd})^{1-X_{nd}} \qquad (1)$$

where $\Theta_M = \{\{\boldsymbol{\pi}_j\}, \boldsymbol{P}\}$ is the set of parameters defining the mixture model, $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jD})$ and $\boldsymbol{P} = (p_1, \ldots, p_M)$ is the mixing parameters vector, $0 \leq p_j \leq 1$, $\sum_{j=1}^{M} p_j = 1$. It is noteworthy that the previous model assumes

actually that all the binary features have the same importance. It is well-known, however, that in general only a small part of features may allow the differentiation of the different present clusters. This is especially true when the dimensionality increases and in this case the so-called curse of dimensionality becomes problematic in part because of the sparseness of data in higher dimensions. In this context many of the features may be irrelevant and will just introduce noise and then compromise the uncovering of the clustering structure. A major advance in feature selection was made in [10] where the problem was defined within finite Gaussian mixtures. In [8, 2], we adopted the approach in [10] to tackle the problem of unsupervised feature selection in the case of binary vectors by proposing the following model

$$p(\boldsymbol{X}_n|\Theta) = \sum_{j=1}^{M} p_j \prod_{d=1}^{D} \left[\rho_d \pi_{jd}^{X_{nd}} (1-\pi_{jd})^{1-X_{nd}} + (1-\rho_d) \lambda_d^{X_{nd}} (1-\lambda_d)^{1-X_{nd}} \right] \quad (2)$$

where $\Theta = \{\Theta_M, \{\rho_d\}, \boldsymbol{\Lambda}\}$, $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_D)$ are the parameters of a multivariate Bernoulli distribution considered as a common background model to explain irrelevant features, and $\rho_d = p(\phi_d = 1)$ is the probability that feature d is relevant such that ϕ_d is a missing value equal to 1 is feature d is irrelevant and equal to 0, otherwise. Feature selection is important not only because it allows the determination of relevant modeling features but also because it provides understandable, scalable and more accurate models that prevent data under- or over-fitting. Unfortunately, the modeling capabilities in general and the feature selection process in particular can be negatively affected by the presence of outliers. Indeed, a common problem in machine learning and data mining is to determine which vectors are outliers when the data statistical model is known. Removing these outliers will normally enhance generalization performance and interpretability of the results. Moreover, it is well-known that the success of many applications usually depends on the detection of potential outliers which can be viewed as unusual data that are not consistent with most observations. Classic works on outlier rejection have considered being an outlier as a binary property (i.e. either the vector in the data set is an outlier or not). In this paper, however, we argue that it is more appropriate to affect to each vector a degree (i.e. a probability) of being an outlier or not as it has been shown also in some previous works [9]. In particular, we define a cluster independent outlier vector to be one that can not be represent by any of the mixture's components and then associated with a uniform distribution having a weight equal to p_{M+1} indicating the degree of outlier-ness. This can be formalized as follow

$$p(\boldsymbol{X}_n|\Theta) = \sum_{j=1}^{M} p_j \prod_{d=1}^{D} \left[\rho_d \pi_{jd}^{X_{nd}} (1-\pi_{jd})^{1-X_{nd}} + (1-\rho_d) \lambda_d^{X_{nd}} (1-\lambda_d)^{1-X_{nd}} \right] + p_{M+1} U(\boldsymbol{X}_n) \quad (3)$$

where $p_{M+1} = 1 - \sum_{j=1}^{M} p_j$ is the probability that $\boldsymbol{X}_n$ was not generated by the central mixture model and $U(\boldsymbol{X}_n)$ is a uniform distribution common for all data to model isolated vectors which are not in any of the M clusters and which show

significantly less differentiation among clusters. Notice that when $p_{M+1} = 0$ the outlier component is removed and the previous equation is reduced to Eq. 2.

2.2 Model Learning

The EM algorithm, that we use for our model learning, has been shown to be a reliable framework to achieve accurate estimation of mixture models. Two main approaches may be considered within the EM framework namely maximum likelihood (ML) estimation and maximum a posteriori (MAP) estimation. Here, we use MAP estimation since it has been shown to provide accurate estimates in the case of binary vectors [8, 2]:

$$\hat{\Theta} = \arg\max_{\Theta} \{\log p(\mathcal{X}|\Theta) + \log p(\Theta)\} \tag{4}$$

where $\log p(\mathcal{X}|\Theta) = \log \prod_{i=1}^{N} p(\boldsymbol{X}_n|\Theta)$ is our model's loglikelihood function and $p(\Theta)$ is the prior distribution and is taken as the product of the priors of the different model's parameters. Following [8, 2], we use a Dirichlet prior with parameters $(\eta_1, \ldots, \eta_{M+1})$ for the mixing parameters $\{p_j\}$ and Beta priors for the multivariate Bernoulli distribution parameters $\{\pi_{jd}\}$. Having these priors in hand, the maximization in Eq. 4 gives us the following

$$p_j = \frac{\sum_{n=1}^{N} p(j|\boldsymbol{X}_n) + (\eta_j - 1)}{N + M(\eta_j - 1)} \quad j = 1, \ldots, M+1 \tag{5}$$

where

$$p(j|\boldsymbol{X}_n) = \begin{cases} \dfrac{p_j \prod_{d=1}^{D} \left(\rho_d p_{jd}(X_{nd}) + (1-\rho_d) p(X_{nd})\right)}{\sum_{j=1}^{M} \left(p_j \prod_{d=1}^{D} \left(\rho_d p_{jd}(X_{nd}) + (1-\rho_d) p(X_{nd})\right)\right) + p_{M+1} U(\boldsymbol{X}_n)} & \text{if } j = 1, \ldots, M \\ \dfrac{p_{M+1} U(\boldsymbol{X}_n)}{\sum_{j=1}^{M} \left(p_j \prod_{d=1}^{D} \left(\rho_d p_{jd}(X_{nd}) + (1-\rho_d) p(X_{nd})\right)\right) + p_{M+1} U(\boldsymbol{X}_n)} & \text{if } j = M+1 \end{cases} \tag{6}$$

where $p_{jd}(X_{nd}) = \pi_{jd}^{X_{nd}}(1-\pi_{jd})^{1-X_{nd}}$ and $p(X_{nd}) = \lambda_d^{X_{nd}}(1-\lambda_d)^{1-X_{nd}}$. $p(j|\boldsymbol{X}_n)$ is the posterior probability that a vector $\boldsymbol{X}_n$ will be considered as an inlier and then assigned to a cluster $j, j = 1, \ldots, M$ or as an outlier and then affected to cluster $M+1$. Details about the estimation of the other model parameters namely π_{jd}, λ_d, and ρ_d can be found in [8, 2]. The determination of the optimal number of clusters is based on the Bayesian information criterion (BIC) [11]. Finally, our complete algorithm can be summarized as follows

Algorithm

For each candidate value of M:

1. Set $\rho_d \leftarrow 0.5$, $d = 1, \ldots, D$, $j = 1, \ldots, M$ and initialization of the rest of parameters [3].

[3] The initialization is based on the K-Means algorithm by considering that $M+1$ clusters are present in the data.

2. Iterate the two following steps until convergence:
   (a) E-Step: Update $p(j|\boldsymbol{X}_n)$ using Eq. 6.
   (b) M-Step: Update the p_j using Eq. 5, and π_{jd}, λ_d and ρ_d as done in [8].
3. Calculate the associated BIC.
4. Select the optimal model that yields the highest BIC.

3 Experimental Results

In this section, we validate our approach via two applications. The first one concerns handwritten digit recognition and the second one tackles visual scenes categorization.

3.1 Handwritten Digit Recognition

Fig. 1. Example of normalized bitmaps.

In this first application which concerns the challenging problem of handwritten digit recognition (see, for instance, [12, 13]), we use a well-known handwritten digit recognition database namely USPS [4]. The UCI database contains 5620 objects. The repartition of the different classes is given in table 1. The original images are processed to extract normalized bitmaps of handwritten digits. Each normalized bitmap includes a 32×32 matrix (each image is represented then by 1024-dimensional binary vector) in which each element indicates one pixel with value of white or black. Figure 1 shows an example of the normalized bitmaps. For our experiments we added also 50 additional binary images (see Fig. 2), which are taken from the MPEG-7 shape silhouette database [14] and do not contain real digits, to the UCI data set. These additional images are considered as the outliers.

Evaluation results by considering different scenarios namely recognition without feature selection and outliers rejection (Rec), recognition with feature selection and without outlier rejection (RecFs), recognition without feature selection and with outliers rejection (RecOr), and recognition with feature selection and outlier rejection (RecFsOr) are summarized in table 2. It is noteworthy that we were able to find the exact number of clusters only when we have rejected the outliers. According to the results in table 2 it is clear that feature selection improves the recognition performance especially when combined with outliers rejection.

[4] ftp://ftp.mpik-tueb.mpg.de/pub/bs/data/

Table 1. Repartition of the different classes.

class	0	1	2	3	4	5	6	7	8	9
Number of objects	554	571	557	572	568	558	558	566	554	562

Fig. 2. Examples of the 50 images taken from the MPEG-7 shape silhouette database and added as outliers.

Table 2. Error rates for the UCI data set by considering different scenarios.

Rec	RecFs	RecOr	RecFsOr
14.37%	10.21%	9.30%	5.10%

3.2 Visual Scenes Categorization

In this second application, we consider the problem of visual scenes categorization by considering the challenging PASCAL 2005 corpus which has 1578 labeled images from multiple sources grouped into 4 categories (motorbikes, bicycles, people and cars) as shown in Fig. 3 [15]. In particular, we use the approach that we have previously proposed in [2] which consists on representing visual scenes as binary vectors and which can be summarized as follows. First interest points are detected on images using the difference-of-Gaussians point detector [16]. Then, we use PCA-SIFT descriptor [17] which allows the description of each interest point as a 36-dimensional vector. From the considered database, images were taken, randomly, to construct the visual vocabulary. Moreover, extracted SIFT vectors were clustered using the K-Means algorithm providing 5000 visual-words. Each image was then represented by a 5000-dimensional binary vector describing the presence or the absence of a set of visual words, provided from the constructed visual vocabulary. We add 60 outlier images from different sources to the PASCAL data set. In order to investigate the performance of our learning

(a) (b) (c) (d)

Fig. 3. Example of images from the PASCAL 2005 corpus. (a) motorbikes (b) bicycles (c) people (d) cars.

approach, we ran the clustering experiment 20 times. Over these 20 runs, the clustering algorithm successfully selected the exact number of clusters, which is equal to 4, 11 times and 5 times with and without feature weighting, respectively, when outliers were taken into account. Without outliers rejection, we were unable to find the exact number of clusters. Table 3 summarizes the results and it is clear again that the consideration of both feature selection and outliers rejection improves the results.

Table 3. Error rates for the visual scenes categorization problem by considering different scenarios.

Cat	CatFs	CatOr	CatFsOr
34.02%	32.43%	29.10%	27.80%

4 Conclusion

In this paper we have presented a well motivated approach for simultaneous binary vectors clustering and feature selection in the presence of outliers. Our model can be viewed as a way to robustify the unsupervised feature selection approach previously proposed in [8, 2], to learn the right meaning from the right observations (i.e inliers). Experimental results that address issues arising from two applications namely handwritten digit recognition and visual scenes categorization have been presented. The main goal in this paper was actually the rejection of the outliers. Some works, however, have shown that these outliers may provide useful information and an expected knowledge, such as in electronic commerce and credit card fraud, as argued in [18] (i.e. "One person's noise is another person's signal" [18]). Thus a possible future application of our work could be of the extraction of useful knowledge from the detected outliers for applications like intrusion detection [19].

Acknowledgment

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. M. Ester, H.-P. Kriegel, J. Sander and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Poc. of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996.

2. N. Bouguila and K. Daoudi. Learning Concepts from Visual Scenes Using a Binary Probabilistic Model. In *Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, oct. 2009.
3. K. Abend, T. J. Harley and L. N. Kanal. Classification of Binary Random Patterns. *IEEE Transactions on Information Theory*, 11(4):538–544, 1965.
4. J. Aitchison and C. G. G. Aitken. Multivariate Binary Discrimination by the Kernel Method. *Biometrika*, 63(3):413–420, 1976.
5. J. C. Bezdek. Feature Selection for Binary Data: Medical Diagnosis with Fuzzy Sets. In *Proc. of the National Computer Conference and Exposition*, pages 1057–1068, New York, NY, USA, 1976. ACM.
6. D. H. Moore II. Evaluation of Five Discrimination Procedures for Binary Variables. *Journal of the American Statistical Association*, 68(342):399–404, 1973.
7. E. Saund. Unsupervised Learning of Mixtures of Multiple Causes in Binary Data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 27–34, 1993.
8. N. Bouguila and K. Daoudi. A Statistical Approach for Binary Vectors Modeling and Clustering. In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T. B. Ho, editors, *PAKDD*, volume 5476 of *Lecture Notes in Computer Science*, pages 184–195. Springer, 2009.
9. M. M. Breunig, H-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying Density-Based Local Outliers. In *Proc. of the ACM SIGMOD International Conference on Management of Data (MOD)*, pages 93–104, 2000.
10. M. H. C. Law, M. A. T. Figueiredo and A. K. Jain. Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004.
11. G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 16:461–464, 1978.
12. Y. Freund and R. E. Schapire. Experiments with a New Boosting Algorithm. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 148–156, 1996.
13. J. Dahmen, R. Schlüter and H. Ney. Discriminative Training of Gaussian Mixtures for Image Object Recognition. In W. Förstner, J. M. Buhmann, A. Faber, and P. Faber, editors, *DAGM-Symposium*, pages 205–212. Springer, 1999.
14. S. Jeannin and M. Bober. Description of core experiments for MPEG-7 motion/shape. Technical Report ISO/IEC JTC 1/SC 29/WG 11 MPEG99/N2690, MPEG-7 Visual Group, Seoul, March 1999.
15. M. Everingham, A. Zisserman, C. K. I. Williams, L. Van Gool, et al. The 2005 PASCAL Visual Object Classes Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 117–176. Springer-Verlag, LNAI 3944, 2006.
16. D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
17. Y. Ke and R. Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 506–513, 2004.
18. E. M. Knorr and R. T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proc. of 24rd International Conference on Very Large Data Bases (VLDB)*, pages 392–403, 1998.
19. R. Durst, T. Champion, B. Witten, E. Miller, and L. Spagnuolo. Testing and Evaluating Computer Intrusion Detection Systems. *Commun. ACM*, 42:53–61, July 1999.