ZIP and data document visualization

Dora Alvarez-Medina, Hugo Hidalgo-Silva,
CICESE-Ciencias de la Computación
Km. 107 Carr. Tijuana-Eda.
Ensenada, 22800 México.
(dalvarez,hugo@cicese.mx)

Abstract

Text data modeling has been usually considered with Bernoulli or multinomial event models. Poisson distribution is considered inefficient for text information retrieval. In this work, we propose to incorporate the Zero Inflated Poisson model in the Generative Topographic Mapping algorithm. The modified algorithm is presented as a text document cluster extraction and visualization tool. Experimental results are presented for the Medlars, CISI and Cranfield collections, observing notable class separation.

1 Introduction

Data visualization is aimed at obtaining a graphic representation of high dimensional information. One particular technique used to represent data is the latent variable model. In this model, we would like to obtain a representation of the data distribution in a lower-dimensional space, usually two-dimensional. The latent variable model is also employed for the identification of clusters and outliers on data. Most of the information obtained with the latent class models can not be obtained using just a dimension reduction method as principal component analysis (PCA). From the several tools proposed for data visualization, the Generative Topographic Mapping (GTM) [1] has gained importance because of its topographic organization capacity. In the GTM model, the data are assumed a noisy version of a high-dimensional variable related via a non-linear function to the latent variables. Considering the Gaussian distribution for the noise, a continuous model was obtained. The GTM model has been also considered for binary data modeled with a multi-dimensional Bernoulli distribution [2]. Kaban and Girolami [3] adapted GTM to the exponential distribution family. Some application to cluster visualization with GTM has also been reported at [1], [4] and [5]. The performance of original GTM decreases when the data dimensionality is too high, such as in text documents. In the bag of words representation, the documents are considered as elements of a vector space. Each element of the vector represents the word frequency count in the document. When a word do no exist in the document the frequency term is zero. Usually in the matrix representing a document collection may exist many terms with a zero. Kaban and Girolami [3] considered this problem by modeling the expectation parameter as a nonlinear function that asymptotically reaches the value of zero.
In this work, we propose to incorporate the Zero Inflated Poisson (ZIP) scheme in GTM to visualize data documents in a plane. As Poisson process fails when there are too many zeros, in the ZIP regression model the data are considered as generated from a distribution with 0 with some probability p and $Poisson(\lambda)$ with probability $1-p$. A mixture model for the latent variables is assumed, and the ZIP scheme for the generative process.

2 ZIP and count data models.

In the bag of words document representation, a document is considered as an element of a vector space. The document set is represented as a term-document matrix T, where an element of the matrix can be represented as t_{nj}, indicating the occurrence times of the word j in the document n. In order to reduce the notation, from now on, we will use $\mathbf{t}_n$ when referring the n_{th} row of matrix T, and $\mathbf{t}_j$ for j_{th} column. We consider the complete vocabulary set of a document after preprocessing (stemming or stop words elimination) of D dimension. We will also assume that the term document matrix T has many elements with zero count, because not all the words are present in all the documents.

Count data processes have been studied in many statistics applications, Kaban and Girolami [3] considered it for document visualization. They obtained an accurate representational structure for text document data using the exponential family of distributions, except for Poisson distribution. Considering the special handling of zero counts in ZIP theory, we propose to use a probabilistic mixture

model of the latent classes along with ZIP.

For count data a Poisson model is commonly used, Li [9] estimates the distribution of the document vector using a Poisson mixture for document classification and word clustering. Lambert [8] proposes a careful modification of Poisson with the ZIP model in an attempt to consider excess of zeroes. She assumes that with probability p the observation is 0, and with probability $1-p$ a Poisson process. The ZIP regression model for the multivariate case with the data count variable t_{nj} can be represented as:

$$\begin{aligned} t_{nj} &= 0 \quad \text{with probability } p_{nj} + (1-p_{nj})e^{-\lambda_{kj}}\ , \\ &= t_{nj} \quad \text{with probability } (1-p_{nj})\frac{e^{-\lambda_{kj}}\lambda_{kj}^{t_{nj}}}{t_{nj}!}\ . \end{aligned}$$

where λ_{kj} is considered the mean of a subset of k elements that includes elements from the column $\mathbf{t}_j$, and p_{nj} is the probability that t_{nj} of having a zero count. The latent representation is assumed with a *log* link between λ and covariates w and the *l*ogit link of p_{nj} with the covariate γ:

$$log(\lambda_{kj}) = \sum_{m=1}^{M} \phi_{km} w_{mj} \tag{1}$$

$$logit(p_{nj}) = log\frac{p_{nj}}{1-p_{nj}} = G_{nj}\gamma_j \tag{2}$$

where w_{mj} is a weight applied to the basis function ϕ_{km}, evaluated on the latent variables. Wedel [10] considered the *log* link for a regression model of the explanatory variables, and assumed a mixture of Poisson distributed variables model for the observed frequencies. Assuming also a mixture of K Poissons model, the model for n_{th} row ($P_n(\mathbf{t}_n)$ can be considered as:

$$\begin{aligned} P_n(\mathbf{t}_n|w,\gamma) &= \sum_{k=1}^{K} \alpha_k p_{nk}(\mathbf{t}_n|w,\gamma) \\ &= \sum_{k=1}^{K} \alpha_k \Bigg(\prod_{\substack{j=1 \\ \forall t_{nj}=0}} p_{nj} + (1-p_{nj})e^{-\lambda_{kj}} \\ &\qquad \prod_{\substack{j=1 \\ \forall t_{nj}>0}}^{D} (1-p_{nj})\frac{e^{-\lambda_{kj}}\lambda_{kj}^{t_{nj}}}{t_{nj}!} \Bigg) \end{aligned} \tag{3}$$

Where α_k is the mixture coefficient parameter. Similar to GTM, the latent space is a grid X of size $K \times L$, with $L = 2$. After several tests with several radial basis vector functions we decided to use a *sinh* function of X as basis function with parameter μ_m:

$$\phi_{km} = sinh(x_k^t \mu_m) \tag{4}$$

where x_k is the k_{th} row.

3 General formulation of ZIP.

The log likelihood function of data considering (3) is given as

$$\ell = \sum_{n=1}^{N} log\left(\sum_{k=1}^{K} p(\mathbf{t}_n|w,\gamma)\theta_{nk} \right)$$

where θ_{nk} is the latent prior, i.e. the probability that $\mathbf{t}_n$ belongs to class k, and evaluated using the Bayes' rule

$$\theta_{nk} = \frac{\alpha_k p_n(\mathbf{t}_n|w,\gamma)}{\sum_{k=1}^{K} \alpha_k p_n(\mathbf{t}_n|w,\gamma)}. \tag{5}$$

For parameter estimation, we considered the EM algorithm, formulating the Expectation step from the relative likelihood as:

$$E(\log \mathcal{L}) = \sum_{n=1}^{N}\sum_{k}^{K} \theta_{nk} \log\left(\alpha_k P_n(\mathbf{t}_n|W,\gamma)\right)$$

$$\begin{aligned} &= \sum_{\substack{n,k,j \\ t_{nj}=0}}^{NKD} \theta_{nk} \log(e^{G_{nj}\gamma_j} + e^{-\lambda_{kj}}) \\ &\quad - \sum_{\substack{n,k,j \\ t_{nj}=0}}^{NKD} \theta_{nk} \log(1 + e^{G_{nj}\gamma_j}) \\ &\quad - \sum_{\substack{n,k,j \\ t_{nj}>0}}^{N,K,D} \theta_{nk} \log(1 + e^{G_{nj}\gamma_j}) \\ &\quad + \sum_{\substack{n,k,j \\ t_{nj}>0}}^{N,K,D} \theta_{nk} \left(t_{nj}\log(\lambda_{kj}) - \lambda_{kj}\right) \\ &\quad - \sum_{\substack{n,k,j \\ t_{nj}>0}}^{N,K,D} \theta_{nk} \log(t_{nj}!) \\ &\quad + \sum_{n,k,j}^{N,K,D} \theta_{nk} \log(\alpha_k). \end{aligned} \tag{6}$$

The separation of the sum of exponentials in the first term complicates the maximization of relative likelihood, but the evaluation can be realized by defining $Z_{nj} = (1 + e^{-G_{nj}\gamma_j - \lambda_{kj}})^{-1}$ as in [8], then:

$$\begin{aligned} E(\log \mathcal{L}_c) &= \sum_{n,k,j}^{N,K,D} \theta_{nk} Z_{nj} \mathrm{G}_{nj}\gamma_j \\ &\quad - \sum_{n,k,j}^{N,K,D} \theta_{nk} \log(1 + e^{G_{nj}\gamma_j}) \\ &\quad - \sum_{n,k,j}^{N,K,D} \theta_{nk}(1 - Z_{nj})\, \log(1 + e^{G_{nj}\gamma_j}) \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{n,k,j}^{N,K,D} \theta_{nk}(1 - Z_{nj})\Big(t_{nj}\log(\lambda_{kj}) - \theta_{nk}\lambda_{kj}\Big) \\
&- \sum_{n,k,j}^{N,K,D} \theta_{nk}(1 - Z_{nj})\ \log\ (t_{nj}!) \\
&+ \sum_{n,k,j}^{N,K,D} \theta_{nk}\log(\alpha_{kj}) \\
= \quad & \theta_{nk}\mathcal{L}_c(\gamma; t, Z) + \theta_{nk}\mathcal{L}_c(W; t, Z) \\
&- \sum_{n,k,j}^{N,K,D} \theta_{nk}(1 - Z_{nj})\ \log(t_{nj}!) \\
&+ \sum_{n,k,j}^{N,K,D} \theta_{nk}\log(\alpha_{kj}). \qquad (7)
\end{aligned}
$$

In the Maximization step, w_{mj}, γ_j and α_k are estimated as follows;

M step for w_{mj}: The maximization step is implemented with the Iterative Reweighed Least Square (IRLS) procedure [11]. Applying IRLS the estimation is obtained by iterating

$$
\begin{aligned}
w_{j'}^{new} &= w_{j'}^{old} \\
&\quad + H^{-1}\phi^T(\theta^T C t_{j'} - A\theta^T(1 - Z_{j'})) \qquad (8)
\end{aligned}
$$

with

$$
\begin{aligned}
H &= diag(-F^T W) \\
W &= B\theta^T(1 - Z_{j'}) \\
B &= diag\left(\frac{\lambda_{j'}(1 - e^{-\lambda_{j'}}(1 + \lambda_{j'}))}{(1 - e^{-\lambda_{j'}})^2}\right) \\
F_{j'} &= diag(\phi_{j'}\phi_j) \\
A &= diag\left(\frac{\lambda_{j'}}{1 - e^{-\lambda_{j'}}}\right) \\
C &= diag(1 - Z_{j'}).
\end{aligned}
$$

H, B, A y C are diagonal matrices and F is a $K \times D$ matrix. $\lambda_{j'}$, $z_{j'}$, $F_{j'}$ and $\phi_{j'}$ are the j'th column of λ, Z, F and ϕ respectively.

M step for $\gamma_{j'}$: this parameter is estimated using logistic regression [12], taking the first and second derivatives:

$$
\begin{aligned}
\gamma_{j'}^{new} &= \gamma_{j'}^{old} \\
&\quad + (-G_{j'}^T W G_{j'})^{-1}(G_{j'}^T(AB - CB)) \qquad (9)
\end{aligned}
$$

$$
\begin{aligned}
W &= FE \\
A_{nn} &= Z_{nj'} \\
F_{nn} &= p_{nj'}(1 - p_{nj'}) \\
B &= \theta I \\
E_{nn} &= \sum_{k=1}^{K} \theta_{nk} \\
C_{nn} &= p_{nj'}
\end{aligned}
$$

F, E, A, W and C are diagonal matrices, I is a vector of $K \times 1$ ones and B is a $N \times 1$ vector.

M step for α_k: as the sum of all α_k must be one, we maximize (7) with respect to α_k as an augmented function $\sum_{n,k,j}^{N,K,D} \theta_{nk} log\alpha_k - \mu\left(\sum_k^K \alpha_k - 1\right)$, where μ is a Lagrangian multiplier. The parameter actualization is given as

$$
\alpha_k^{new} = \sum_n^N \frac{\theta_{nk}}{N} \qquad (10)
$$

The algorithm begins initializing the parameters, then iterating until convergence.

- *Initialization*
  - X = random
  - μ = lines of sequence points
  - G = random
  - γ = random
  - w = random
  - $\alpha_k = 1/K$
  - Compute ϕ_{km} from (4)
- *Iterate until convergence*
  - E step:
    - Compute θ from (5)
  - M step: parameter update
    - Compute w^{new} from (8)
    - Compute γ^{new} from (9)
    - Compute α^{new} from (10)

4 Performance evaluation.

In this section, we present the experimental results for data visualization. The simulations were done using three data collections:

- 200 documents extracted from Medlars collections (1033 medical abstracts),
- 200 documents extracted from CISI collections (1460 science abstracts),
- 200 documents extracted from Cranfield collections (1398 aerodynamics abstracts).

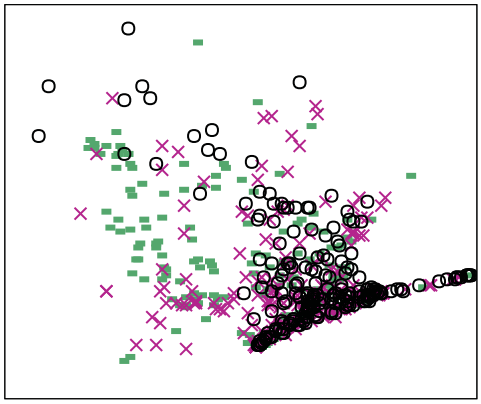

Figure 1: Data visualization of Medlars (-), CISI (o) and Cranfield (x) collection. With uniform grid latent variables and Gaussian RBF functions.

Before applying the algorithm, we have to perform a preprocessing task, where the first goal is to consider only the words with sense. It was necessary to remove the stop-words and use all words in their root form, by performing *stemming* [14]. Each document of the 600 was transformed in a vector by removing the stop words and applying the Porter's stemming algorithm[13]. The final data set contained 1,173 unique terms. Beside that, we selected some words in the dictionary in order to reduce the amount of computation. We use the resolving power of significant words [14] as a term selection method. The "resolving power" takes the words in the middle frequency range, which are supposed to be the relevant items. Experiments were done using a dictionary size of 329 terms, therefore the size of our term-document matrix T is 600×329.

4.1 Experimental results.

Experiments were done considering a uniform grid of 270 latent variables (X) and 25 basis function parameters (μ). Results are presented in Fig. 1. Separation efficiency is highly dependent on latent variables and basis function parameter initialization. After several tests with different basis functions, we noticed a better class separation with hyperbolic sin function. In Fig. 2 results are presented for sinh. A random sampling strategy was also considered for the latent variables distribution. Fig. 3 shows the results when a random sampling (Gaussian distribution for $\mu = 0$ and $\sigma = 1$) of the grid is used to select the latent variables. From Fig. 3 we observe the Cranfield class closer to Medlars, and at the right side some Cranfield elements appearing close to those from CISI.

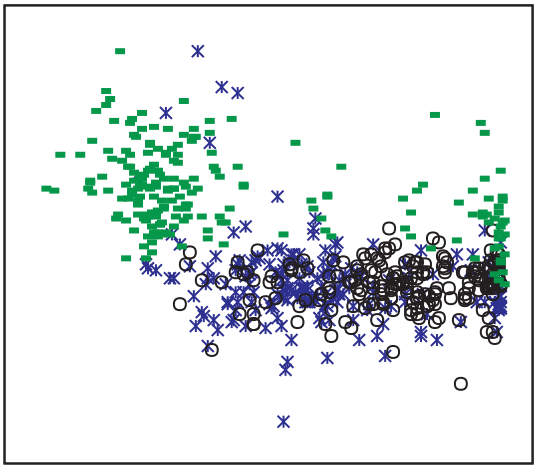

Figure 2: Data visualization of Medlars (-), CISI (o) and Cranfield (x) collection. With uniform grid latent variables and *sinh* basis function.

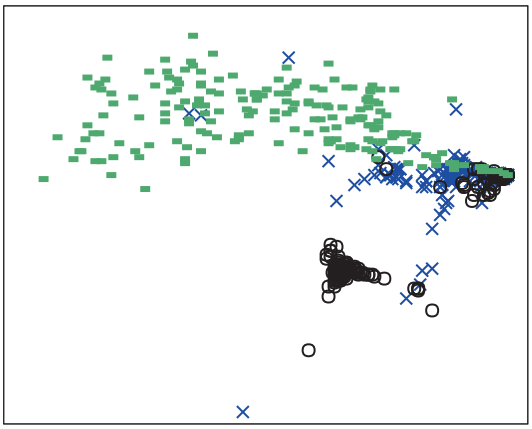

Figure 3: Data visualization of Medlars (-), CISI (o) and Cranfield (x) collection. With randomly selected latent variables and *sinh* basis function.

5 Conclusions.

A mixture model for data visualization is presented, the special zero accounting by ZIP is implemented in the vector space representation for text. The influence of latent space and basis functions is observed. Using the ZIP representation, a notable class separation is observed on the latent variables. The best results were obtained with the *sinh* basis function.

6 Acknowledgements

The first author was supported by CONACyT.

References

[1] C. M. Bishop, M. Svénsen, and C.K.I. Williams, GTM: The Generative Topographic Mapping, *Neural Computation*, 10:215-235, 1998.

[2] M. Girolami, "A generative model for sparse discrete Binary Data with Non-Uniform Categorical Priors", *P*roc. European Symp. Artificial Neural Networks (ESANN'00), pp. 1-6, 2000.

[3] A. Kabán, and M. Girolami, A combined latent class and trait model for the analysis and visualization of discrete Data, *IEEE transactions on pattern analysis and machine intelligence*, 23(8):859-871, 2001.

[4] P. Tino, and I. Nabney, Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):639-656, 2002.

[5] J. Yang, and B.T. Zhang, Customer Data Mining and Visualization by Generative Topographic Mapping Methods. In *Proceedings of the International Workshop on Visual Data Mining*, 55-66, 2001.

[6] K.W. Church and W. Gale, Poisson Mixtures, *Natural Language Eng.*, 1(2):163-190, 1995.

[7] DD. Lewis, Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *European Conf. Machine Learning*, 4-5, 1998.

[8] D. Lambert, Zero Inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, 34(1):1-13, 1992.

[9] J. Li, and H. Zha, Two-way Poisson mixture models for simultaneous document classification and word clustering, *Computational Statistics & Data Analysis*, 50(1):163-180, Elsevier, 2006.

[10] M. Wedel, W. S. Desarbo, J. R. Bult, and V. Ramaswamy, A latent class Poisson regression model for heterogeneous count data, *Journal of Applied Econometrics*, 8:397-411, 1993.

[11] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1983.

[12] T. Hastie and R. Tibshirani and J. Friedman, *The elements of statistical learning*, Springer, U.S.A., 2001.

[13] M. F. Porter, *An algorithm for suffix stripping*,Program, 14 1980, pp. 130–137.

[14] G. Salton and M. J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, U.S.A., 1983.