Unsupervised Learning with Imbalanced Data via Structure Consolidation Latent Variable Model

Fariba Yousefi, Zhenwen Dai, Carl H. Ek, Neil Lawrence

- University of Sheffield
- F.Yousefi@sheffield.ac.uk

Motivation

- In medical applications, negative labelled data is easy to obtain, while positive labelled data is much harder to acquire.
- Imbalanced data are very challenging for both generative and discriminative methods.
- We build a latent variable model that can accommodate a very large number of negative examples, while sharing their characteristics appropriately with the positive class.
- Our model can perform classification as well as generating new samples.

Dataset Definitions

- Data is publicly available dataset (AMIDA13).
- The pre-processed images of the training set from Snell (2013).
- $146,562$ grey-scale image patches, of which 550 are positive.
- Randomly taking 80% of positive images and 5000 negative images as training set.

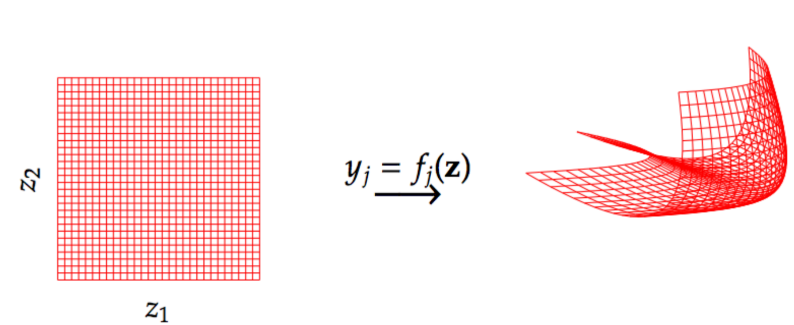
[1]

[1] The background image is taken from http://www.princeton.edu/ samory/

Gaussian Process Latent Variable Model

- The dataset is represented as a set of fixed length vectors $\mathbf{Y} \in \Re^{N \times D}$.
- A label of category is associated with each data point, $\mathbf{c} = (c_1, \ldots, c_N)$, $c_i \in \{1, \ldots, C\}$.
- The data associated with a set of latent representations $\mathbf{X} \in \Re^{N \times Q}$.
- The latent representations are related to the observed data through an unknown mapping function f and f follows a Gaussian process prior distribution.

- Gaussian Process Latent Variable Model



Structure Consolidation Latent Variable Model (SCLVM)

$x = [x_s^\top, x_p^\top]^\top, x_s \in \Re^{Q_s}, x_p \in \Re^{Q_p}$

$$k((x, c_x), (x', c_{x'})) = k_s(x_s, x_s') + k_p((x_p, c_x), (x_p', c_{x'})),$$

The private kernel is defined to take the following form:

$$k_p((x_p, c_x), (x_p', c_{x'})) = \begin{cases} k'(x_p, x_p'), & c_x = c_{x'}, \\ 0, & c_x \neq c_{x'}, \end{cases}$$

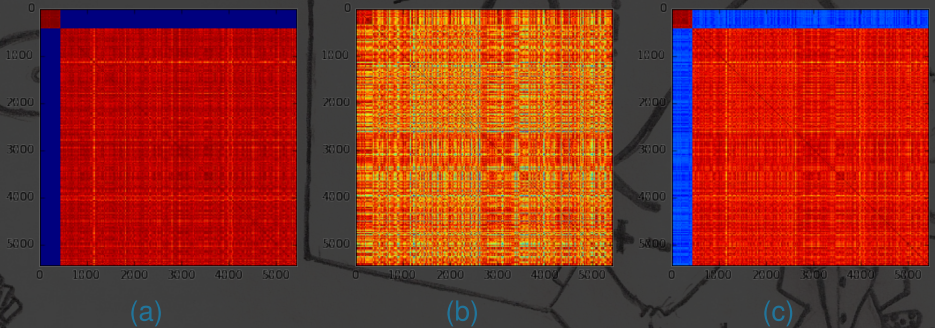


Figure: (a) Covariance matrix for Private space, (b) Covariance matrix for public space, (c) Covariance matrix for the whole kernel.

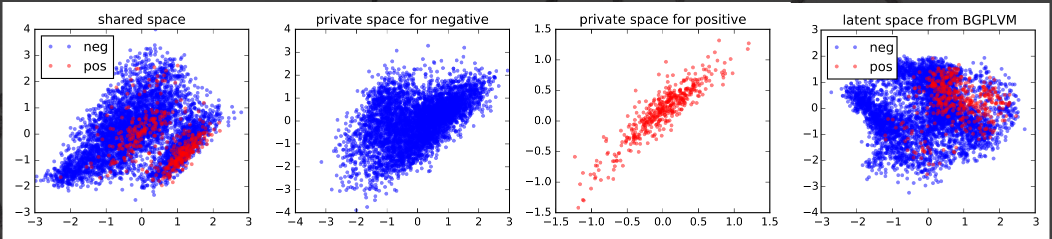
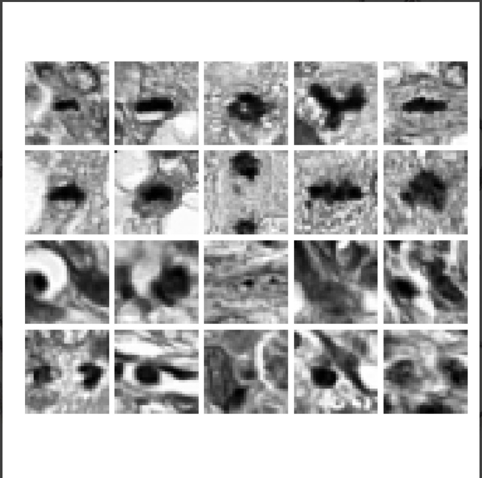
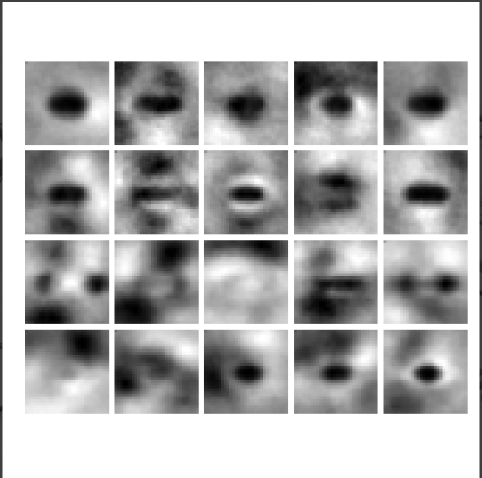


Figure: The visualization of the training data in the learned latent spaces.



(a) Some examples in the data sets



(b) Samples generated from the trained SCLVM

Results

Table: Classification performance. The mean and standard deviation from ten test sets are as below:

	SCLVM	BGPLVM (SVM)	DGPLVM (SVM)
precision	0.426 ± 0.024	0.306 ± 0.013	0.242 ± 0.008
recall	0.555 ± 0.007	0.827 ± 0.000	1.000 ± 0.000
F1 score	0.482 ± 0.015	0.447 ± 0.014	0.390 ± 0.011

Conclusion

- We presented a probabilistic latent variable model that can cope with imbalanced data.
- We developed a kernel that separates the latent space into a shared spare and a private space.
- An efficient variational inference method is proposed by deriving a closed form lower bound of marginal likelihood.
- Beyond the shown example, the ability of jointly modelling multiple data categories and handling imbalanced datasets can be linked to many other areas such as transfer learning.


THANK YOU