

Multi-View Learning from Disparate Sources for Poverty Mapping

Neeti Pokhriyal

Department of Computer Science and Engineering
University at Buffalo
State University of New York, US
neetipok@buffalo.edu

Many data analytics problems involve data coming from multiple sources, sensors, modalities or feature spaces, that describe the object of interest in a unique way, and typically exhibit heterogeneous properties. The varied data sources are termed as *views*, and the task of learning from such multi-view data is known as *multi-view learning*.

In my thesis, I target the problem of poverty prediction and mapping from disparate multi-source data, namely mobile phone data, satellite imagery, ground weather measurements and open street maps. Additionally, census data was used to get the regression targets for poverty. Currently, poverty is estimated through intensive socioeconomic household surveys. However, this approach is costly and time consuming and can only be realistically carried out for a small sample of households, thereby making timely updates of poverty challenging. Thus, the need is to timely and accurately predict poverty and map it to spatially fine-grained baseline data (Devarajan 2013).

The primary aim of my thesis is to develop novel multi-view algorithms which are focused on the unique constraints and challenges of mapping poverty. Several challenges lie in establishing relationships between auxiliary data sources (that are not collected to directly measure socioeconomic deprivations, example mobile phones, satellite imagery, weather measurements and open street maps) and poverty, namely varying spatial, temporal and conceptual granularity at which the different data are available. This necessitates an aggregation mechanism to *link* them to individuals whose poverty is to be estimated. Other challenges include quantifying and mitigating the uncertainty associated with each data source, and need for a validation mechanism due to unavailability of ground-truth data.

Another aim is to relax the core assumptions faced by existing multi-view learning algorithms. Most existing approaches either assume that the views are completely independent or fully dependent. Methods operating under the former assumption typically involve multi-kernel learning, while those following the latter assumption aim at learning a shared latent subspace or manifold. However, in real scenarios, these assumptions are almost never truly satisfied. My

aim is to develop learning algorithms that learn factorized representations, where the subspace is divided into a shared (capturing common variance across different views), and private subspaces (capturing remaining variance). Besides dimensionality reduction, the aim of this factorized representation is to provide understanding of underlying structure of data, and, thus, giving better accuracy for prediction, and handling domain issues of missing data. I pose the problem in both regression and classification settings.

Preliminary research has been done towards the goals of the thesis. A naive way to work with multiple data sources is concatenating the different feature spaces, and employing a single classifier. I have explored the concatenation approach while conceptualizing the idea of cognitive biometrics using large scale social media data (Pokhriyal, Nwogu, and Govindaraju 2014; Pokhriyal et al. 2017).

Using Disparate Data Sources via GPs

I constructed two *semantically distinct* views: 1). *mobile phone view*, using call data records and, 2). *environmental view*, using satellite imagery, ground weather measurements and open street maps. Here, poverty mapping is studied as a regression problem, with continuous poverty values associated with each micro-region. The regression targets are poverty values that are calculated from census data.

Gaussian process regression (GPR) (Rasmussen and Williams 2006) models are independently trained on each view, to infer poverty values. Gaussian Processes (GP) fall under the class of kernel methods, where the choice of different kernel functions enables one to learn different non-linear relationships between the independent and target variables. Each GP-based model provides a probabilistic estimate of poverty for a given location, including the mean and variance of the estimates. The variance provides a measure of uncertainty, which allows us to combine the predictions from the multiple views.

To predict poverty for a location from a single data view, the following model is assumed:

$$y_i = \beta^T \mathbf{x}_i + f(\mathbf{x}_i) + \epsilon \quad (1)$$

where y_i is the target poverty value and \mathbf{x}_i is a vector of independent variables derived from the particular view for

the i^{th} location. The first term is a linear combination of the independent variables. The function $f()$ models the non-linear relationship between y_i and \mathbf{x}_i . The residual term, ϵ , models the remaining unexplained noise, and is modeled as a zero-mean Gaussian random variable, i.e., $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. We assume a Gaussian Process (GP) prior on $f()$, with following kernel function:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right) \exp\left(-\frac{\|\mathbf{x}_s - \mathbf{x}'_s\|^2}{2\ell_s^2}\right) \quad (2)$$

where \mathbf{x}_s and \mathbf{x}'_s are the spatial coordinates (latitude, longitude) of the commune centers corresponding to \mathbf{x} and \mathbf{x}' .

To predict poverty for a location, we use independently trained GPR models on the two views, and their outputs i.e. the posterior Gaussian distributions are combined in weighted manner, where the view that provides small predictive variance is assigned a higher weight.

The model parameters are estimated by maximizing the marginalized log-likelihood of the training data. We couple the GP model with elastic net regularization during the model learning phase to remove the effect of spurious features. This allows for automatic relevant feature selection and learning a parsimonious and interpretable model.

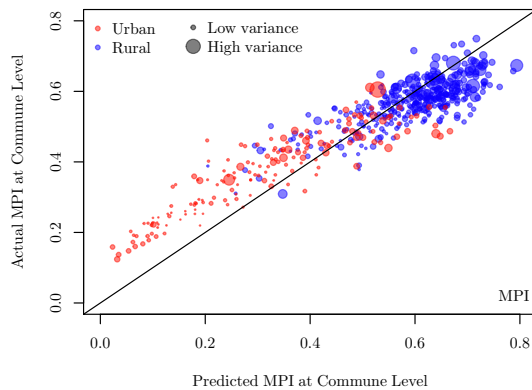


Figure 1: Comparison of actual and predicted poverty values for all communes and urban areas of Senegal.

Figure 1 shows the relationship between poverty values predicted by our model and those estimated from census (this is the validation set). We observe a linear relationship with lower values for urban areas (shown in red) and higher values for rural areas (shown in blue), along with uncertainty values depicted (Pokhriyal and Jacques 2017). We get a Pearson correlation of 0.91 using disparate data, compared to single source data (Pearson correlation of 0.89 using only mobile data, and 0.84 using only environment data).

Discriminative Factorized Representations

Motivated by the additional challenges like missing views, I focus on learning factorized subspace representations in a classification setting, which consists of shared and private subspaces corresponding to each view.

Existing works on factorized subspaces is focused on unsupervised settings (Salzmann et al. 2010), (Ek et al. 2008),

mainly with linearity assumption (Xue et al. 2017). My focus is learn representations, which can give insight into the underlying relationships among the disparate views, beneficial for prediction and in case of missing views. I propose a linear method (called Discriminative Factorized Subspace (DFS)), and its non-linear version (kernel DFS (KDFS)). The optimization formulation is efficiently solved using generalized eigenvalue decomposition, which provides a globally optimal solution, via Constrained Generalized Rayleigh Quotient. Our algorithms are rigorously tested using spatial cross-validation strategy, and with missing views; and can classify a location as poor or not, with 83% accuracy.

I propose a latent variable model, assuming that the observed data is generated by smooth mappings from a factorized latent space to the observed views. Smooth reverse mappings are assumed too. Training step involves learning these mappings, which are used during inference to output the representations of an unseen multi-view instance. While any existing subspace learning algorithm can be used to learn the shared space, the view-specific representations are learnt by extracting projections, that are orthogonal to the shared subspace, while maximizing the between-class and minimizing the within-class variances.

Significance to the Domain and Next Steps

The significance of this work, lies in its capability to generate spatially detailed poverty maps in a timely and cost-effective manner, which can provide policy makers assistance between intercensal periods, or in areas of conflicts.

As future work, I plan to extend the concept of discriminative factorized representations to regression, and multiple outputs setting, where the different outputs correspond to variety of socio-economic deprivations.

References

- Devarajan, S. 2013. Africa’s statistical tragedy. *Review of Income and Wealth*.
- Ek, C. H.; Rihan, J.; Torr, P. H. S.; Rogez, G.; and Lawrence, N. D. 2008. Ambiguity modeling in latent spaces. In *Machine Learning for Multimodal Interaction*.
- Pokhriyal, N., and Jacques, D. C. 2017. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*.
- Pokhriyal, N.; Tayal, K.; Nwogu, I.; and Govindaraju, V. 2017. Cognitive-biometric recognition from language usage: A feasibility study. *IEEE Trans. Information Forensics and Security*.
- Pokhriyal, N.; Nwogu, I.; and Govindaraju, V. 2014. Use of language as a cognitive biometric trait. In *IEEE Intl Joint Conf on Biometrics, Clearwater*.
- Rasmussen, C. E., and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. The MIT Press.
- Salzmann, M.; Ek, C. H.; Urtasun, R.; and Darrell, T. 2010. Factorized orthogonal latent spaces. In *AISTATS*.
- Xue, X.; Nie, F.; Wang, S.; Chang, X.; Stantic, B.; and Yao, M. 2017. Multi-view correlated feature learning by uncovering shared component. In *AAAI*, 2810–2816. AAAI Press.