From SVM to LSTM: Classfication of Time Series Gene Expression Data

Aiyappa Parvangada, Shefali Umrania

Carnegie Mellon University

Computational Biology Department

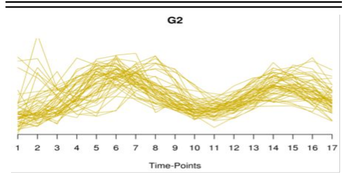
Introduction

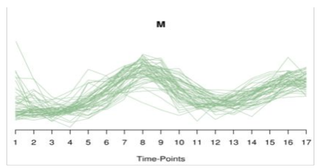
- Time series microarray experiments generate labelled temporal profiles which are suitable for the following tasks:
  - Classification of unlabelled time course expression data.
  - Identification of dynamic biological networks.
- Classification finds motivation most importantly in clinical applications where prediction of a patient's response to a drug treatment is crucial for timely alteration of therapeutic strategies.
- We implemented a novel CNN and LSTM, and compared the performance against a temporal SVM, HMM and kNN_{DTW}.

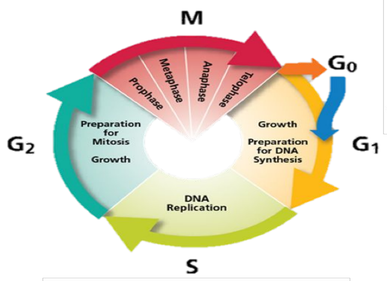
Objective

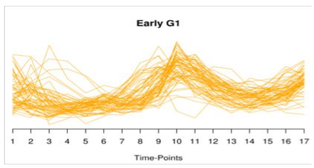
To classify *Saccharomyces cerevisiae* time series gene expression profiles into the 5 mitotic cell cycle phases - Early G1, Late G1, S, G2, M.

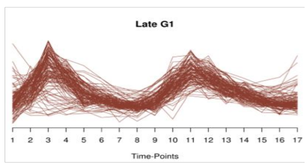
Dataset	
Summary	*Yeast*
Examples	388
Classes	Early G1 (67), Late G1 (136), S (77) G2 (54), M (54)
Time series length	17

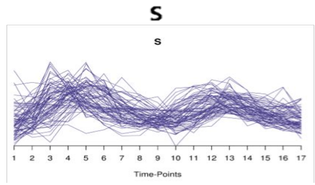










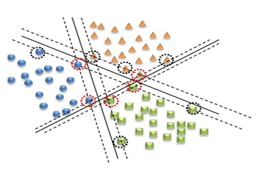


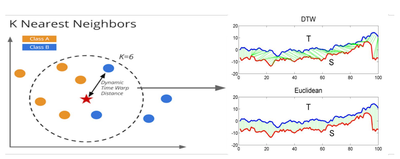
Dataset	
Summary	*Yeast*
Examples	6000
Classes	No class labels
Time series length	17

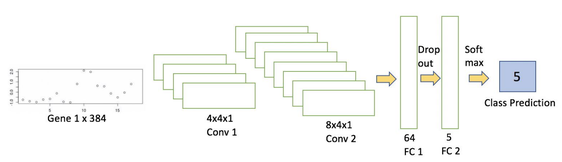
Methods

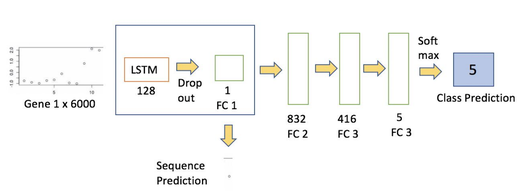
We compare our classification accuracy on the yeast cell cycle dataset with 3 other established machine learning models.

1. L_1-norm Temporal SVM: developed by Orsenigo et al.
2. HMM - TRAM: developed by Z. Bar-Joseph et al.
3. K-Nearest Neighbors with Dynamic Warping Distance
4. 1-D Convolutional Neural Net
5. Long Short-Term Memory Network with Convolutional Networks









Results

Method					
CNN	LSTM	$k\text{-}NN_{DTW}$	Gen-HMM	Disc-HMM	L_1-TSVM
87.5%	86.28%	65.26%	68%	54%	73.9%

- We used the Python deep learning library 'keras' to construct and train the neural network models. The prediction accuracy of the convolutional neural network was 87.50% and that of the LSTM was 86.28%.
- An implementation of kNN based on dynamic time warping as a distance metric (k=11 and warping window of 2) with 20-fold cross validation resulted in an accuracy of 65.26%.
- A freely available matlab implementation of an HMM was trained on the yeast dataset. An accuracy of 68% using the generative model and 54% using the discriminative model was achieved on a hold-out subset of 50 genes (10 from each class).
- The accuracy of the L1-norm temporal SVM was obtained from the publication that served as the inspiration for this project.
- Surprisingly, the 1D convolutional neural net had the highest accuracy compared to all other model classification strategies in spite of 'inadequate' training instances.

Discussion

- One major cause for the low accuracy of kNN_{DTW} and the discriminative HMM could be the high level of noise in the data.
- We conclude that our semi-supervised approach of pre-training using an LSTM and subsequent fine-tuning of the neural network performs better at classification than other established methods.
- The reason for this could be that because the data is small, both neural net methods learn similar set of features and hit an upper limit with regards to increasing accuracy. We anticipate varying performances with larger training data.

References:

1. Z. Bar-joseph et al.. "Access : Studying and modelling dynamic biological processes using time-series gene expression data : Nature Reviews Genetics." Nature.com. 18 Jul. 2012.
2. Cho et al. "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle: Molecular Cell." Cell.com. 1 Jul. 1998.
3. Orsenigo et al. "Time Series Gene Expression Data Classification via L1-norm Temporal S." SpringerLink. 2010.
4. Tien-ho et al. "Alignment and classification of time series gene expression in clinical studies | Bioinformatics | Oxford Academic." Academic.oup.com. 1 Jul. 2008.
5. Lipton et al.. "[1511.03677] Learning to Diagnose with LSTM Recurrent Neural Networks." Arxiv.org.
6. Mark Regan : Timeseries Classification: KNN & DTW