Neural Sequence Embedding generalization with Density-Based Clustering

Milan van der Meer

***Abstract*—The abstract goes here.**

***Index Terms*—IEEEtran, journal, LaTeX, paper, template.**

I. Motivation

General motivation

THIS demo file is intended to serve as a "starter file" for IEEE journal papers produced under LaTeX using IEEEtran.cls version 1.8a and later. I wish you the best of success.

mds
September 17, 2014

A. Subsection Heading Here

Subsection text here.

Abstract; Introduction; Previous Research; Problem Formulation; Model or Methods and Results; Conclusion; References; Acknowledgements.

II. Previous Research

Literatuurstudie een beetje uitleggen met de verschillende methodes.

III. Problem Definition

In this section we describe our problem and each of the subproblems.

The medical history of a patient is a time series with each medical status a data point in time. The main goal is to predict a label for each time series. For example: patient will be cured.
A high dimensional numerical vector represents a medical status where a value can express for example the bloodpressure of a patient. Between data points, there can be long time periods and also irregular intervals. The numerical values of the vectors need to be standarized. Typical for large datasets, are missing values which have to be taken into account.

1) Time Series: Each patient is an independent time series. But in the time series, several indepedent disease periods can occur. In medical data, there are a large amount of unique events because of the high dimensional data points. Each possible combination of the vectorspace represents an unique event. Machine learning techniques are harder to apply when there are a large amount of different events, especially when rare events are possibly important which is the case in medical data.

2) Long Time Periods: Between the events are a long range of dependencies possible. When machine learning techniques try to model those dependencies, adecay or blow up of events can happen, this is called the vanishing gradient problem.

3) Irregular Intervals: Irregular intervals are a form of missing data. Our method has to handle the irregular intervals or transform them accordingly to regular intervals

4) High Dimensionality: A well known problem is the Curse of Dimensionality. It causes the data to be sparse and implies the need for more data. Also the effects and importance of attributes is unclear because of the large amount of attributes.

5) Sequence Labeling: Our prediction method should handle all above problems and be able to build a model to label a given sequence. The label is an indicator on what the outcome will be of the patients disease trajectory. The labeling proces gives the prediction of the disease trajectory. Sequence labeling can be done with Recurrent Neural Networks (RNN). A RNN which handles problems like III-2, III-3, and III-4 is Long Short Term Memory (LSTM).

6) Embedding: DESCRIBE EMBEDDING PROBLEM

The method described in IV tries to handle the preprocessing needed for the LSTM network.

IV. Methods

We first describe our approach on the problems as irregular intervals, standarization, and the high dimensionality. Secondly, we describe our method to make the embedding matrix. Lastly, we introduce a method to generalize our events to get a more general embedding matrix but still keep outliers.

A. Preprocessing

V. Future Work

Untill now we described and implemented a method to generalize our dataset and find connections between events. Those results can be used to feed into the LSTM network to label the sequences. With the learned model and the embedding matrix, we can also handle unseen events by looking for the k-nearest neighbours of this unseen event and base our results on the findings of those neighbours.

VI. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.