AN ABSTRACT OF THE THESIS OF

Michael M. Anderson for the degree of Master of Science in Computer Science presented on January 1, 2013.

Title: Activity Detection on Free-Living Data Using Change Point Detection

Abstract approved: ____

Weng-Keen Wong

(Abstract text)

Activity Detection on Free-Living Data Using Change Point Detection

by

Michael M. Anderson

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented January 1, 2013
Commencement June 2013

Master of Science thesis of Michael M. Anderson presented on January 1, 2013.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Michael M. Anderson, Author

ACKNOWLEDGEMENTS

(Acknowledgement text)

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: Introduction

1.1 Motivation

1.2 Previous Research

Chapter 2: Experimental Setup

2.1 Featurization

2.2 Base Classifiers

2.3 Datasets

2.4 Performance Metrics

Chapter 3: Top-Down Approach

3.1 Change Point Detection

For this approach, the data was split into non-overlapping segments for featurization using techniques from the statistical field of change point detection. Change point detection has found application in many problem domains that require analysis of time series data from dynamic systems, including failure detection [1], quick detection of attacks on computer networks [7], and monitoring of heartbeat fluctuations during sleep [5]. Change point detection assumes that each tick of a time series is a draw from some process, but that the process may suddenly change as time passes. The goal is to predict when these changes have occurred. A *score* is generated for each time tick, and if the score is above a given threshold a change is predicted to have occurred between that tick and its immediate predecessor. To generate a score at a time tick, a window of data that immediately preceeds it (the *reference data*) is compared to it along with a window of data that immediately follows it (the *test data*).

Model-based approaches to change point detection assume that each tick in a time series is a draw from some underlying probability distribution. Scores are generated by estimating the distribution of the reference data and the test data, and then by calculating the likelihood that the two distributions are different. Where it is reasonable

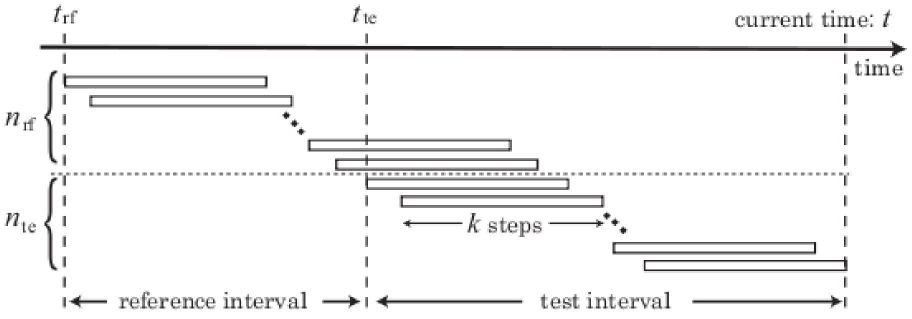

Figure 3.1: Reference and Test Data [2]

to assume that the data belongs to a particular family of distributions then parametric estimation methods have been employed [8]. If no such modeling assumptions are reasonable then non-parametric methods have also been found to be viable [3]. Finally, distance-based approaches such as Singular Spectrum Analysis generate scores through other metrics of dissimilarity or difference between the reference data and the test data [4]. Notationally, we say that for each tick i in a time series:

$$s_i = D(x_{r,i}, x_{t,i})$$

Where s_i is the score of the ith tick, $x_{r,i}$ is the reference data associated with the ith tick, $x_{t,i}$ is the test data associated with the ith tick, and $D(A, B)$ is a function that computes the dissimiliarity between a matrix of data A and matrix of data B, and is particular to the given change point algorithm. Note that for a given algorithm it may not be possible to generate scores right at the very beginning of the time series (insufficient reference data) or at the very end of a time series (insufficient test data).

3.2 Methodology

There are many different modeling assumptions and associated algorithms for generating change point detection scores, but one simple baseline approach that we tested was the Control Chart. This approach assumes that the reference data is drawn from a multivariate normal distribution, and scores are calculated by the Mahalanobis distance of the target time tick from the estimated multivariate normal:

$$s_i = \sqrt{(\bar{x}_{r,i} - x_i)^T \, S_{r,i}^{-1} \, (\bar{x}_{r,i} - x_i)}$$

where $\bar{x}_{r,i}$ is the sample mean of the reference data, $S_{r,i}$ is the sample covariance matrix of the reference data, and $x_i = x_{t,i}$ is the ith data point.

We were also interested in testing the performance of a newer and more sophisticated change point detection algorithm: the Kullback-Leibler Importance Estimation Procedure (KLIEP), introduced by Kawahara and Sugiyama [2] [6]. This approach generates scores using the Kullback-Leibler (KL) divergence between the reference data and the test data. One method of doing this is to estimate the density of the reference distribution and test distribution separately, and then compare them using a likelihood ratio (known in the change point detection literature as *importance*). Instead, KLIEP estimates the importance directly using a non-parametric model.

Let the estimate of the importance $\hat{R}$ be represented by this model:

$$\hat{R} = \frac{p_t}{\hat{p}_r} = \sum_{j=1}^{n_t} \alpha_j K_G(x, x_{t,j})$$

Where p_r and p_t are the probability densities of the reference data and the test data, n_t is the number of ticks in the test window, α is a vector of model parameters to solve for, x is the concatenation of the reference and the test data, $x_{t,j}$ is the jth element of the test data, and $K_G(A, B)$ is the Gaussian kernel with width σ:

$$K_G(A, B) = \exp\left(-\frac{||A - B||^2}{2\sigma^2}\right)$$

Now solve for α so that the empirical KL divergence between $\hat{p}_t$ and $p_t = p_r\hat{R}$ is minimized, which is equivalent to the following convex optimization problem:

$$\begin{cases} \max\limits_{\alpha} & \displaystyle\sum_{j=1}^{n_t} \log\left(\sum_{k=1}^{n_t} \alpha_k K_G(x_{t,j}, x_{t,k})\right) \\ \text{s.t.} & \displaystyle\frac{1}{n_r}\sum_{j=1}^{n_r}\sum_{k=1}^{n_t} \alpha_k K_G(x_{r,j}, x_{t,k}) = 1 \\ & \text{and } \alpha_1 \ldots \alpha_{n_t} \geq 1 \end{cases}$$

Finally, the scores that we wish to generate are just the estimate of the importance given by the solution to the complex optimization problem, i.e. $s_i = \hat{R}_i$.

Since this approach uses a Gaussian kernel, it requires the selection of a kernel width σ for each time tick. We used an implementation of KLIEP that is available at Sugiyama's website, which included a cross-validation procedure for the value of σ. The CV procedure chooses a number of disjoint splits of the test data along with a number of different candidate σ's, and runs KLIEP with each combination of split and candidate σ. Then it chooses the candidate σ that, on the average across all of the splits, maximizes the KL divergence (the $\max_\alpha$ equation above) the most.

For the OSU Hip dataset, we used this CV to choose the the kernel width at each individual time tick. This computationally- intensive approach was impractical for the

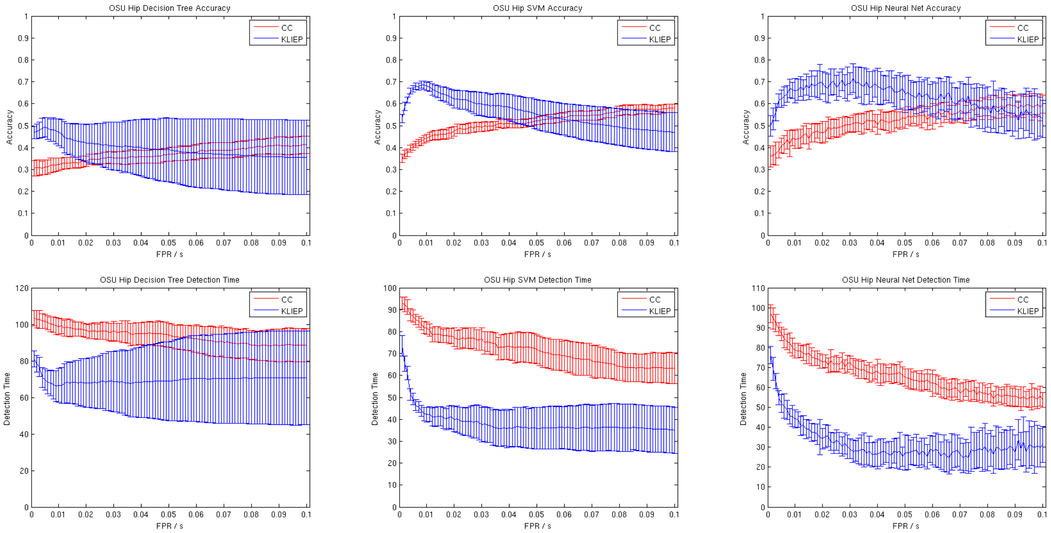

Figure 3.2: CPD-Based Classification Performance

UQ dataset because it is orders of magnitude larger, so instead of running it on every tick of that data, we ran the CV procedure on a number of random ticks drawn from the data. From this we were able to empirically identify 0.01 as a plausible σ, and so fixed σ at that value for our experiments on this dataset. Our reference windows were fixed at a length of 10 seconds, and our test windows were fixed at a length of 1 second.

Once scores were generated, we chose a number of threshold values to determine which scores were considered high enough to predict a changepoint. Threshold values were chosen by considering the false positive rates of change prediction for the change point detection algorithms. A smaller false positive rate corresponded to a higher and more conservative threshold, which split the time series into fewer segments for featurization. A larger false positive rate corresponded to a lower threshold, which split the time series into more segments.

3.3 Results

Figure 3.2, shows accuracy and detection time as a function of the false positive rate per second, for the OSU Hip dataset, for each of the SVM, Decision Tree, and Neural Net base classifiers.

3.4 Discussion

Chapter 4: Bottom-Up Approach

4.1 HMMs

4.2 Methodology

4.3 Results

4.4 Discussion

Chapter 5: Conclusion

5.1 Discussion

5.2 Directions for Future Research

Bibliography

[1] Suk Joo Bae, Byeong Min Mun, and Kyung Yong Kim. Change-point detection in failure intensity: A case study with repairable artillery systems. *Computers and Industrial Engineering*, 64:11–18, January 2013.

[2] Yoshinobu Kawahara and Masashi Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. *Proceedings of the SIAM International Conference on Data Mining*, pages 389–300, 2009.

[3] David S. Matteson and Nicholas A. James. A nonparametric approach for multiple change point analysis of multivariate data. 2012.

[4] V. Moskvina and A. A. Zhigjavsky. An algorithm based on singular-spectrum analysis for change-point detection. *Communication in Statistics. Statistics and Simulations*, 32:319–352, 2003.

[5] M. Staudacher and all. A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep. *Statistical Mechanics and its Applications*, 349:582–596, April 2005.

[6] Masashi Sugiyama and all. Direct importance estimation with model selection and its application to covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

[7] A.G Tartakovsky and all. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54:3372–3382, September 2006.

[8] G. Thatte and all. Parametric methods for anomaly detection in aggregate traffic. *IEEE/ACM Transactions on Networking*, 19:512–519, April 2011.