Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Henriques dos Santos Teixeira

Data Stream Anomaly Detection through Principal Subspace Tracking

Dissertação de Mestrado

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática do Centro Técnico Científico PUC–Rio as partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Prof. Ruy Luiz Milidiú

Rio de Janeiro
September 2009

Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Henriques dos Santos Teixeira

Data Stream Anomaly Detection through Principal Subspace Tracking

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática do Centro Técnico Científico PUC–Rio as partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the following commission:

Prof. Ruy Luiz Milidiú
Advisor
Departamento de Informática — PUC–Rio

Prof. Marco Antonio Casanova
Departamento de Informática — PUC-Rio

Prof. Raúl Pierre Rentería
Departamento de Informática – PUC-Rio

Prof. Prof. José Eugenio Leal – PUC-Rio
Coordinator of the Centro Técnico Científico — PUC–Rio

Rio de Janeiro — September 04, 2009

Pedro Henriques dos Santos Teixeira

Graduated from the University of Warwick (England) in Computer Science in 2004, and worked as a senior software developer for the last five years building complex scalable distributed systems. His academic interests are in the area of Optimization and Automatic Reasoning and throughout the program he has participated in the Spock and Netflix challenges, developed an AI video game for the XNA framework and implemented many heuristics for optimizing NP-hard problems such as TOP and TSP. He is currently working on machine learning solutions for monitoring data streams at his own start-up.

Bibliographic data

Teixeira, Pedro Henriques dos Santos

Data Stream Anomaly Detection through Principal Subspace Tracking / Pedro Henriques dos Santos Teixeira; advisor: Ruy Luiz Milidiú. — 2009.

95 f. : il. ; 29,7 cm

Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui bibliografia.

1. Informática – Teses. 2. Deteção de Anomalia. 3. Rastreamento do Subespaço Principal. 4. Aprendizado não-supervisionado. 5. Fluxo de dados. 6. Séries temporais. 7. Redução de dimensão. I. Milidiú, Ruy. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

To my parents for their unconditional support and for being role models of moral integrity and kindness throughout the years. I am forever grateful to them for giving me the freedom to pursue my own goals and for having invested in my education.

To Gabriela for the caring love and for having all the patience in the world throughout the past two years.

To my adviser Professor Ruy Milidiú, for sharing his wisdom. The incisive comments were essential to my journey, and I am especially thankful for the early guidance on conciliating my academic work with my professional life.

To the Professors in the comission, Raúl Rentería and Marco Casanova, whose reviews helped improve this dissertation.

To the colleagues at LEARN and to all the teachers and friends at the DI for providing an excellent master's degree.

To my fellow associates who missed me at the company during the last month.

To Globo.com, especially to Marco Lucio and Denis Vieira, who supported our project.

I thank Professor Peter Strobach for replying my e-mails and enlightening me with some history on subspace trackers. I also thank Professor Roland Badeau for sharing his Matlab implementation of many subspace trackers and Professor Mark Coates who answered my inquiries about KOAD's code.

Abstract

Teixeira, Pedro Henriques dos Santos; Milidiú, Ruy. **Data Stream Anomaly Detection through Principal Subspace Tracking**. Rio de Janeiro, 2009. 95p. MScThesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The complexity of data centers and the high volume of generated monitoring data poses many challenges to system administrators. Current tools rely on experts to configure fixed thresholds for each data stream, which is not efficient nor appropriated for dynamic systems. We study an unsupervised learning technique based on spectral analysis proposed for anomaly detection. In this technique, just one pass over the data is allowed. It tracks the principal subspace of r dimensions from N numerical data streams. An anomaly is considered to be a sudden change in system behaviour and is indicated by a change in the number of latent variables. A previous approach [Papadimitriou et al., 2005] relies on a PAST-type subspace tracker, which is based on an inverse matrix update and is known to be unstable. It requires an extra normalization step in order to guarantee the expected reconstruction error, which really adds to $\mathcal{O}(Nr^2)$ in time complexity per update instead of the $\mathcal{O}(Nr)$ advertised. In this work, we present FRAHST, the first rank-adaptive principal subspace tracker based on the new recursive row-Householder algorithm, which is the state-of-the-art for an algorithm of this kind. It is stable, provides orthonormal basis and has a true dominant complexity of only $\mathcal{O}(Nr)$ flops. Our technique reduces in 75% the number of false positives when compared against the previous subspace technique in the public Abilene Packets dataset while still doubling the number of detections. By embedding lagged values to the input vector to explore temporal correlations, FRAHST successfully detects subtle anomalous patterns in the data and when compared against four other anomaly detection techniques, it is the only one with a consistent F_1 score above 80% in all four datasets. As part of this work, a real-time system is successfully implemented to monitor the infrastructure of a ISP, and is a good use case for unsupervised anomaly detection in the industry.

Keywords

Anomaly Detection. Principal Subspace Tracking. Unsupervised Learning. Data Streams. Time Series. Dimensionality reduction.

Resumo

Teixeira, Pedro Henriques dos Santos; Milidiú, Ruy. **Deteção de Anomalia em Fluxo de Dados através de Rastreamento do Subespaço Principal**. Rio de Janeiro, 2009. 95p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A complexidade de um data center combinada com o alto volume de dados gerados para monitoração traz diversos desafios para administradores de sistemas. As ferramentas atuais requerem especialistas que configurem limiares constantes para cada fluxo de dado, o que é ineficiente e inapropriado para estes sistemas dinâmicos. Estudamos uma técnica não supervisionada baseada em teoria espectral que foi proposta para deteção de anomalia, na qual é permitida uma única passada sobre os dados e monitora o subspaço principal de dimensão r apartir de N fluxos de dados númericos. Uma anomalia é considerada uma mudança brusca no comportamento do sistema e pode ser indicada por uma mudança no número de variáveis latentes nos dados. A abordagem anterior depende de um algoritmo do tipo PAST, que é baseado em atualizações sobre uma matrix invertida e que é coqnhecido por ser instável. Além disso, é necessário esforço adicional de ortonormalização para garantir o erro de reconstrução condizente com os paramêtros, o que adiciona $\mathcal{O}(Nr^2)$ em complexidade de tempo por atualização. Neste trabalho, apresentamos FRAHST, o primeiro algoritmo de rank adaptativo baseado no método de row-Householder recursivo para aproximar o subespaço principal, que representa o estado-da-arte para um algoritmo deste tipo. FRAHST é estável, fornece estimativas orthonormais e tem complexidade dominante de $\mathcal{O}(Nr)$ flops. Nossa ténica reduz em 75% o número de falsos positivos quando comparado com a ténica anterior no dataset público Abilene Packets e ao mesmo tempo dobra o número de deteções. Ao incorporar valores passados no vector de entrada para explorar correlações temporais, FRAHST deteta sutis padrões anômalos nos dados e quando comparado com outras quatro técnicas para deteção de anomalia, é o único que obtem de forma consistente um valor de F_1 acima de 80% em todos datasets. Como parte deste trabalho, um sistema de tempo-real foi implementado para monitorar a infraestrutura de um provedor de internet e é um caso de sucesso para deteção de anomalia não-supervisionada na indústria.

Palavras–chave

Deteção de Anomalia. Rastreamento do Subespaço Principal. Aprendizado não-supervisionado. Fluxo de dados. Séries temporais. Redução de dimensão.

Contents

List of Figures

List of Tables