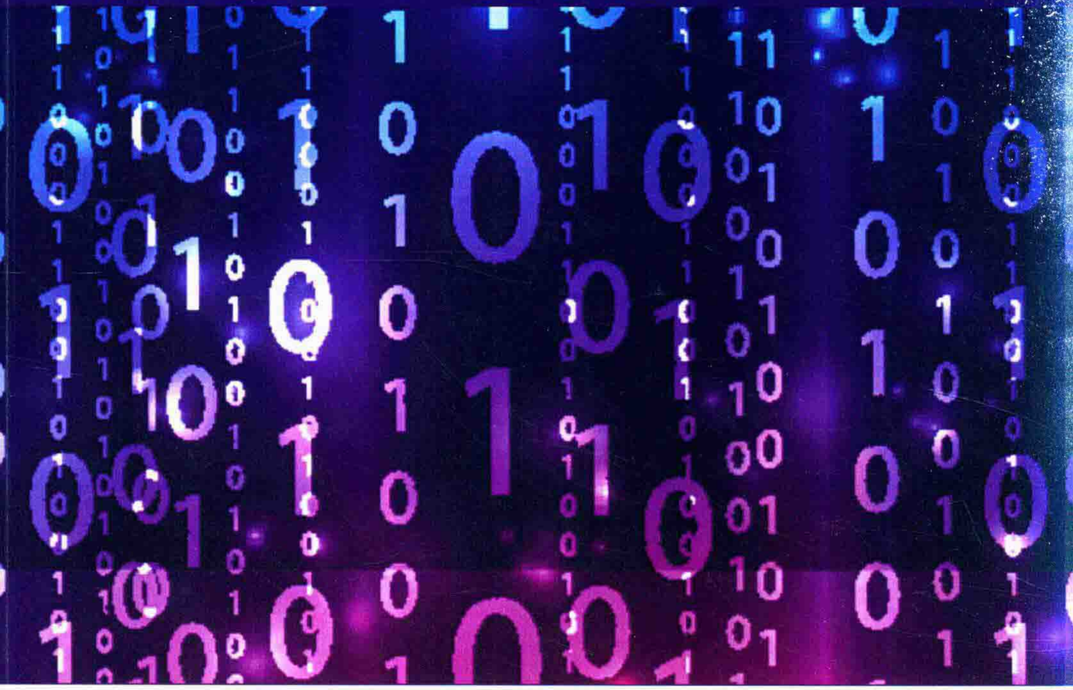

Tegjyot Singh Sethi

Mining Drifting Data

Automated learning in changing environments with limited feedback

LAP LAMBERT Academic Publishing

Tegjyot Singh Sethi

Mining Drifting Data

Automated learning in changing environments with limited feedback

LAP LAMBERT Academic Publishing

Impressum / Imprint

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
Alle in diesem Buch genannten Marken und Produktnamen unterliegen warenzeichen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.
Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Coverbild / Cover image: www.ingimage.com

Verlag / Publisher:
LAP LAMBERT Academic Publishing
ist ein Imprint der / is a trademark of
OmniScriptum GmbH & Co. KG
Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany
Email: info@lap-publishing.com

Herstellung: siehe letzte Seite /
Printed at: see last page
ISBN: 978-3-659-69276-5

For this book is dedicated to the

memories of Daduji and Dadiji

ACKNOWLEDGEMENTS

The author would like to thank his advisor Dr. Mehmed Kantardzic for all his support, and guidance. This thesis would not have been possible without his continued encouragement and advice. The author would also like to convey his sincere appreciation to the members of his thesis committee: Dr. Adel S. Elmaghraby and Dr. Gail W. DePuy, for their valuable suggestions and patience. The author would like to thank Dr. Olfa Nasraoui for her advice and insight in this field. The author would like to thank his family in India, Man Mohan Singh and Gurinder Sethi, for their support and encouragement to pursue a graduate education. The author also thanks the members of his family in Chicago, Illinois, Devinder and Rachita Singh, for their continued motivation. Finally, special thanks are extended to Angad, Ankita and Jia for their unconditional love and trust.

ABSTRACT

THE GC3 FRAMEWORK

GRID DENSITY BASED CLUSTERING FOR CLASSIFICATION OF STREAMING DATA WITH CONCEPT DRIFT

Tegjyot Singh Sethi

July 24, 2013

Data mining is the process of discovering patterns in large sets of data. In recent years there has been a paradigm shift in how the data is viewed. Instead of considering the data as static and available in databases, data is now regarded as a stream as it continuously flows into the system. One of the challenges posed by the stream is its dynamic nature, which leads to a phenomenon known as Concept Drift. This causes a need for stream mining algorithms which are adaptive incremental learners capable of evolving and adjusting to the changes in the stream.

Several models have been developed to deal with Concept Drift. These systems are discussed in this thesis and a new system, the GC3 framework is proposed. The GC3 framework leverages the advantages of the Grid Density based Clustering and the Ensemble based classifiers for streaming data, to be able to detect the cause of the drift and deal with it accordingly. In order to demonstrate the functionality and performance of the framework a synthetic data stream called the TJSS stream is developed, which embodies a variety of drift scenarios, and the model's behavior is analyzed over time.

Experimental evaluation with the synthetic stream and two real world datasets demonstrated high prediction capability of the proposed system with a small ensemble size and labeling ratio. Comparison of the methodology with a traditional static model with no drifts detection capability and with existing ensemble techniques for stream classification, showed promising results. Also, the analysis of data structures maintained by the framework provided interpretability into the dynamics of the drift over time. The experimentation analysis of the GC3 framework shows it to be promising for use in dynamic drifting environments where concepts can be incrementally learned in the presence of only partially labeled data.

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

PAGE

CHAPTER 1

INTRODUCTION

1.1 Stream data mining and its challenges

A data stream refers to a continuous flow of ordered data records in and out of a system [1]. With the growth in sensor technology and the big data revolution, large quantities of data are continuously being generated at a rapid rate. Whether it is from sensors installed for traffic control or systems to control industrial processes, data from credit card transactions to network intrusion data, streaming data is ubiquitous. Today almost all forms of data being collected is streaming as we do not stop collecting data to make analysis on it, but instead the analysis and the data collection happens simultaneously. This poses a major challenge with the timeliness of the prediction results. The analysis results from historical data would fail to account for the current state of the system and as such will not be totally reliable. Also the rate at which this data is being generated (real time in many cases) is much higher than the rate at which it can be analyzed by traditional data mining techniques.

In such a dynamic environment, the basic tasks of Data mining such as Clustering, Classification, Summarization, etc. are no longer trivial. There is a paradigm shift from the traditional techniques where the system is presented with all the historical data and a model is built on it, using if needed a validation set, and this model once built is used without change for all future predictions. Such a static model does not fit well in the real world scenario as the data encountered in most cases is itself dynamic and embodies constant changes in the environment. Also, the huge amount of data flowing into the system poses practical restrictions on the memory and the amount of data that can be stored and processed at each time interval. Thus the algorithms for stream mining need to be more selective as to what data they store and what they disregard.

All these concerns have led to a lot of research in the recent years to overcome these challenges posed by streaming data. The main characteristics of streaming data that need to be addressed by any model developed was described in [1,2] and is summarized below.

- *Scalability and Response Time:* As the data stream may, in principle, be an infinite source of data, it is not possible to store all the data for performing the analysis. Thus the model needs to analyze data in chunks and store only a very small portion of this data in the main memory. Also, since the data is continuously pouring in, the response needs to be in near real time in most cases, for it to be of any practical use.
- *Robustness:* Any real world process is bound to exhibit noise and distortions in the data being generated. A system needs to be robust to these factors and work even in the presence of such changes. This problem is even more challenging in case of streaming data, as the data is dynamic and it is necessary to distinguish the noise from the changes in the environment. The model needs to balance between being overly sensitive to noise and at the same time being able to detect changes and learning from them.
- *Concept Drift:* The major challenge with streaming data is that of adaptability. The distribution generating the data might change over time and the model

generated needs to detect and adjust to such changes automatically. This is what makes mining of streaming data different from traditional mining techniques. This change in the generating model with the passage of time is known as Concept Drift.

In this thesis the challenge posed by Concept Drift is considered. The GC3 incremental learning framework is proposed to detect and adapt to changes in an evolving data stream.

1.2 Formal Problem Statement: Classification of Streaming Data

The data mining task considered here is: *Classification*. A stream of data is evaluated one at a time and the class label associated with each sample is estimated. A good model would provide high estimation capability within the limitations of time and memory resources.

Consider a stream of samples represented by X_1, X_2,...X_{Inf} ; where X_i is a vector representing an input sample. Each X_i has an associated Y_k which is its class label. Furthermore consider the value initial_train_stream. For all X_j, $j<$ initial_train_stream, the corresponding Y_k are available. For X_j , $j>$ initial_train_stream, only a few of the Y_k's are available. The task is to predict these Y_k's given only the prior information about the samples. Probabilistically the task of classification is the probability that the class label is Y_k given the sample is X_i denoted by the conditional probability $p(Y_k|\, X_i)$.

When dealing with static data, the common assumption is that the probability distribution of the data does not change with time. i.e. $p_t(Y|\, X_i) = p_{t+1}(Y|\, X_i)$. However, in case of streaming data with concept drift, this assumption is not valid as the distribution