CITY UNIVERSITY OF HONG KONG
香港城市大學

Data Stream Mining in Financial Securities Databases

金融證券數據庫的數據流挖掘

Submitted to
Department of Information Systems
資訊系統學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

Liu Xiaoyan
劉曉艷

November 2007
二零零七年十一月

ABSTRACT

In recent years, advances in hardware technology have allowed us to automatically record everyday transactions in stock trading market at a rapid rate. Such processes lead to large amounts of data which grow at an unlimited rate. These data processes are referred to as data streams. One of the key issues for data stream mining is online mining of changes. By understanding the nature of changes in stock data stream, a user may be able to convert this information into valuable trading decisions. Therefore, it is useful to develop tools and techniques which track changes in securities database in real time.

In this research, we focus on the analysis of "T&Q" (Trade and Quote) data, which contains detailed transaction information of securities tick by tick. One data preprocessing technique and several online algorithms to monitor the flow changes in the stock data are presented and their implications in the real securities data are discussed in the thesis.

Since data streams are often in high volume, to reduce the data dimension, we propose a new measure of class heterogeneity and develop a heuristic method to find the approximate optimal discretization scheme in Chapter 3. The numeric evaluation shows that the proposed method can be a good alternative to entropy-based discretization methods.

Segmentation of data streams is useful to find the change in the trend of stock price. In Chapter 4, we propose two novel online segmentation algorithms: the Feasible Window Space method and its extension the Stepwise Feasible Window Space method. They are piecewise-linear-model-based algorithms and always generate fewer segments with acceptable representation error and less computation time. Extensive experiments on a variety of real-world time series are conducted to evaluate the proposed methods.

Monitoring the change in the stock order flow is a meaningful topic in the financial intelligence field since the change in order follow precedes stock price change. In Chapter 5 and Chapter 6, we propose two online change-point detection methods for the stock order flow. One is the multilayer change-point detection algorithm which makes use of the multiresolution property of wavelet transform. It is a non-parametric method. The change-points obtained by this method are more reliable than those detected only from the original time series. The other one is based on the Poisson distribution assumption of the sequence which is a result of empirical finance study. This method identifies the change-points incrementally.

The contributions of this research are two-fold: 1) From the viewpoint of technique, we propose a data reduction method and a set of effective online change detection methods for data streams and develop a series of theoretical results. Detection of changes supports building data mining models in data streams more effectively and accurately. And 2) From the viewpoint of application, the mining of changes in T&Q databases provides a tool to reduce the information asymmetry in the securities market and support the market short-term players making short-term trading decisions.

TABLE OF CONTENTS