VIETNAM NATIONAL UNIVERSITY HOCHIMINH CITY
UNIVERSITY OF INFORMATION TECHNOLOGY
ADVANCED PROGRAM IN INFORMATION SYSTEMS

NGUYỄN DUY TIẾN

CLUSTERING STOCKS ACCORDING TO THE VARIABILITY AT A PRICES OF PERIOD

BACHELOR OF ENGINEERING IN INFORMATION SYSTEMS

HO CHI MINH CITY, 2014

NATIONAL UNIVERSITY HOCHIMINH CITY
UNIVERSITY OF INFORMATION TECHNOLOGY
ADVANCED PROGRAM IN INFORMATION SYSTEMS

NGUYỄN DUY TIẾN – 10520174

CLUSTERING STOCKS ACCORDING TO THE VARIABILITY AT A PRICES OF PERIOD

BACHELOR OF ENGINEERING IN INFORMATION SYSTEMS

THESIS ADVISOR

ASSOC. PROF. DR. ĐỖ PHÚC

HO CHI MINH CITY, 2014

ASSESSMENT (ADVISOR)

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

ASSESSMENT (COMMITTEE MEMBERS)

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor Associate Processor Dr. Đỗ Phúc who suggests, guides and inspires me in working with with thesis.

I would like to thank Mr. Nguyễn Thành Vị for support, their knowledge of Stock Market in Exchange VietNam.

I want to express my thankful to the staffs in faculty of Information Systems of University of Information Technology and my instructors who emit the passion to me.

Finally, I would like to thank my families for providing us the best condition to study in Advanced Education Program and finish this graduation thesis.

Fall 2014

Nguyễn Duy Tiến

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In this thesis, I use data mining and predictive technologies for analyzing amount of trades in the market. Data mining is well founded on the theory that I can use to analyze the historical for predicting the future direction of stock. This technology is designed to help investors discover hidden patterns from the historical data that have probable predictive capability in their investment decisions. The prediction of stock markets is regarded as a challenging task of finacial time series prediction. Data analysis is one way of predicting whether future stocks prices will increase or decrease of stock price. In this thesis, I use data mining approach for classification of stocks. After classification, the stocks could be selected from these groups for predicting.

Chapter 1
Introduction

1.1 Background

Forecasting stock price is an important financial subject that has attracted researchers' attention for many years. This research tries to help the investor decide the better timing for buying or selling stocks based on the knowledge extracted from the historical stock prices after processing.

Data mining can be described as "making better use of data". Every human being is increasingly faced with unmanageable amount of data; hence, data mining or knowledge discovery apparently affects all of us. It is therefore recognized as one of the key research areas. Ideally, I would like to develop techniques for "making better use any kind of data for any purpose".

Data clustering is the process of searching and discovering groups of similar data in large database. Most of the techniques have been used in solving data clustering problems in areas such as finace, geographic information, biology, image recognition, etc. Recently, clustering methods are proposed and used in stocks market: Dynamic Time Wrapping, Hierarchical Agglomerative clustering (HAC).

In this thesis, I focus on using the data daily stock price of the peroid and stock "group" to help inventors predict better trends prices of other days with clustering system .

1.2 Problem statement

Financial time series are very dynamic and sensitive to quick change. It is because in part of underlying nature of the financial domain and in part of the mix of know parameters (previous day's closing price, P/E ratio etc.) and unknown factors(like election results, rumors etc). Financial time series are difficult to forecast because the stock market is essentially a non-linear, non-parametric system that is extremely hard to model with any reasonable accuracy. Pattern discovery, dimensionality reduction,

clustering and classification, rule discovery and summarization are the main challenges of processing the time series data. Stock price is one kind of time series, and using data mining techniques to predict stock trend is one of the most important issues to be investigated. Stock trend prediction aimed on developing ap- proaches to predict the price in the future with high profits. Obviously, the prediction is difficult because the implemented models should capture the market volatility. Data mining uses techniques clustering rule. This techniques can forecast the future trend based on the itemsets.

1.3 Scope of the thesis

Stock trend prediction is an important finacial research. If a market can be predicted successfully, the investors can gain improved returns. There are many factors that affect the prediction results and thus no universal model that can predict everything well for all problems or even be a single best forecasting method for all situations. Therefore, many researchers have applied different analysis methods to do stock trend prediction.

In this thesis, the scope is set around two algorithm: Dynamic time warping and Hierarchical clustering algorithm with its respective Data Mining Technique. In this thesis I take closing price data of these 15 stocks for website to test my application.

At the beginning, the data collected contained 6 attributes (table 1.1). This number was reduced manually to only one attributes (Close) as the other attributes were found not important and not having a direct effect on the research.

Table 1.1 Attribute of Historical Stock Prices Data

Attribute	Description
Date	Date open sell or buy
Open	Current day open price of the stock
High	Current day maximum price of the stock
Low	Current day minimum price of the stock
Close	Current day close price of the stock
Volume	The action taken by the investor on this stock

1.4 Structure of this thesis

The thesis is organized as follows:

Chapter 1: Introduction

In this chapter, I give a short introduction about my problem and my methodology to solve the problem..

Chapter 2: Fundamental theories

In chapter 2, I introduce overview about predicting stock market trends based data mining , Time series with clustering using Dynamic Time Warping and Hierarchical Clustering algorithm.

Chapter 3: Problems identification and solving methods

In this chapter, I have two main parts: Using Dynamic Time Warping and Hierachical Clustering for clustering daily closing price stock time series

Chapter 4: System Implemention and Testing

In this chapter, I show my implementation and discuss experimental results of my problem. The testing on some stocks at stock exchange in Vietnam, should be mention when the implementation is finished.

Chapter 5: Conclusions and Future Works

I will be summarized and the final evaluation will be described. Also, possible topics of future research on my clustering for stock prices will be mentioned and briefly explained.

Chapter 2

Fundamental Theories

2.1 Stock market prediction

Financial market is a complex, noisy, nonstationary, nonlinear and always dynamic system but it does not follow random walk process. There are many factors that may cause the fluctuation of inancial market movement. The main factors include economic condition, political situation, traders' expectations, catastrophes and other unexpected events. Therefore, predictions of stock market price and its direction are quite difficult. In response to such difficulty, data mining techniques have been introduced and applied for this financial prediction. Most of the research have focused on the accurate forecasting of the value of stock price.

Stock market prediction is the act of trying to determine the future value of a company stock or other finacial instrument traded on an exchange. The successful prediction of a stock's future price could yied significant profit. The efficient- market hypothesis suggests that stock price movements are governed by the random walk hypothesis and thus are inherently unpredictable. Others disagree and those with this viewpoint possess a myriad of methods and technologies which purportedly allow them to gain future price information.

Stock market prediction includes the comprehension of two features: market factors and unknown random processes. When one is considering a trading strategy or even in any study of the markets, there are a considerable amount of factors that can be considered. But the factor *Time* should not be left on the side, especially when the some of the investors decisions are built on historical stock prices performances. Historical stock prices are an important piece of researching potential investment opportunities.

Such as, when an economic factors is released, based on its expectations, the investors will take a determined position, which will lead to movements in stocks

prices. At a certain point of the financial markets, an announcement such as the one previously referred to can cause an impact within a short time frame, such as a week.

The adjustment time is unable to refute. The main goal of this thesis is to substantiate the existence of certain patterns in the financial markets, which can or can't, detect from a historical. This is eager to perceive behaviours on financial stocks which can be detected after being adjusted through a time modulation. In this case of satisfactory results, the work developed may be used as a way of prediction or even designed to be an investment strategy, as it can be seen later in the present report. Furthermore, there might be space to final remarks regarding the time's complexity in this matter.

Data mining of financial data has been proven to be very effective and very profitable. The goal of this project is to apply data mining techniques to an interesting, albeit not very lucrative, area of finance. In this thesis, using techniques of data mining are discussed and applied to predict price movement of HOSE index of Ho Chi Minh stock market. Every stock sold on the Ho Chi Minh Stock Exchange is classified into industrial catergory, or just called an "industry". A company is identified with an industry based on its primary activities, which usually means the area from which it derives the largest share of its revenue. These categories include Chemicals, Biotechnology & Drugs, Retail, Healthcare, Utilities and so on. In addition, each industry is further divided into sub-categories. For example, the Media & Advertising industry consists of four sub-categories: Advertising, Movies & Music, Publishing & Printing, and TV & Radio. However, these sub-categories are not standardized. In this thesis, I will propose an effective clustering method (Dynamic time warping distance measure and Hierarchical clustering algorithm) to predict the short-term stock price movements of group in an industry.

A stock market is aggregation of buyers and sellers (a loose network of economic transactions, not a physical facility or discrete entity) of stocks (also called shares);

these many include securities listed on a stock exchange as well as those only traded privately.

2.2 Data-Mining Method

2.2.1 What is the Data-Mining

Data mining is the way to discover, extract the knowledge in a large data set. It uses many techniques such as artificial intelligence, statistic, or database system. Nowadays many problems relate to business have been using it for investigating situation and conducting business strategies for a better result.

Data mining's techniques include three tasks:

- **Clustering**: is a technique bases on some way to measure the attributes between item in a field and group those whose similar (close together measure in attributes).
- **Classification**: is a technique to "recognize" items and put them in the suitable place, usually called "clusters".
- **Prediction:** is a technique to predict the outcome if some conditions is satisfied.

The process of data mining may be variously described in some way, but in basic they are consist of three phases:

- **Pre-processing**: the first things when doing something would be prepare the ingredient. Indeed, data mining uses a giant amount of raw data, thus the incompleteness or integrity could be violated. To have the data mining works as it finest, the input (raw data) must be through a refine step in order to reach a standard in both quality and quantity.
- **Data mining**: this step uses some techniques for running. Depending on demand – which is clustering, classification, or prediction – one could use many models to serve the demand.

- **Validation:** the final step use some test set to check whether the result is trustable in how many percentage. Two scales are considered: minimum support and minimum confidence. The minimum support measure the association rules bases on its popularity, and the minimum confidence measure the result bases on its correction.

2.2.2 What is Time Series Clustering

In a data mining function, clustering can be used as a standalone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Clustering is one of the most fundamental issues in data recognition. It plays a significant role in searching for structures in data. It may serve as a preprocessing step for other algorithms, which will operate on the detected clusters. The time series data are grouped together based on similarity - similar waves cling together to form a cluster and dissimilar waves tend to stay far apart in separate clusters. Iteratively the time series clustering algorithm relocate the data points one step at a time to ensure that the data points inside the same cluster have the minimum intra-similarity and data points across different clusters have the maximum inter-similarity. The similarity is defined as the multidimensional distance between two data points whose multiple attributes are measured as how close they are in values. Two variables exist for time series clustering algorithm, one is for choosing the similarity function for measuring the distance between each pair of data points, and the other is the overall operation that converge from an initial assignment of data points to clusters to a converged or optimal assignment of data points to clusters.

Clustering is useful both in its own right as an exploratory technique. Conditional, Time series clustering can be divided into two large groups:

- Whole clustering: In this group the notion of clustering is similar to that of conventional clustering of discrete objects. Provided there is a set of individual time series data, the purpose is to classify similar time series into the same cluster.

- Subsequence clustering: Provided there is a single time series, individual time series (subsequences) are extracted with a sliding window. Clustering is then performed on the extracted time series.

Clustering algorithms have shown their best indifferent data mining tasks, consequently it would be efficient to evaluate their facilities in the analysis of time series.

A time series is a sequence of real numbers that representing the measurements of a real variable at equal time intervals. Time series analysis is a sufficiently well-known task. However, in recent years research has been carried out with the purpose to try to use clustering for the intentions oftime series analysis. A time series database is a large collection of time series. For example, measuring the daily stock price in the period would comprise a time series. This is because stock prices is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series. An observed time series can be decomposed into four components: the trend, the seasonal effects, cycles and residuals. (figure 2.1).

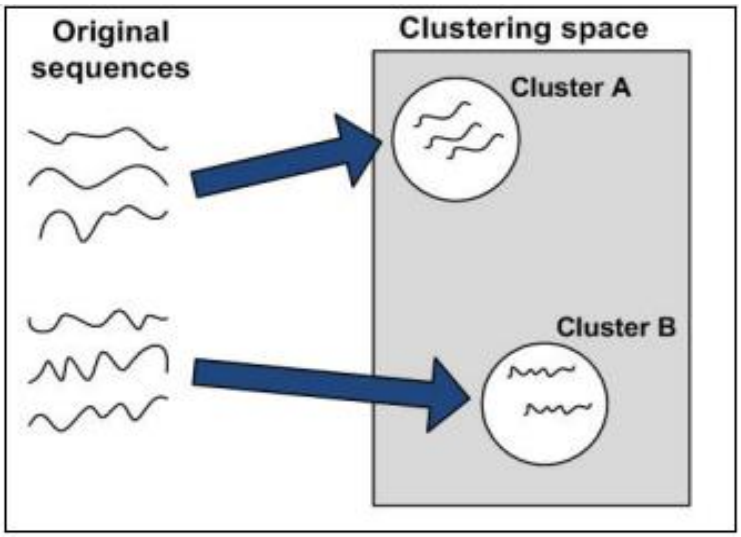

Figure 2.1 Clustering for time series

2.2.3 What is Dynamic Time Warping

The indexing of very large time series databases has attracted the attention of the database community in recent years. The vast majority of work in this area has focused on indexing under the Euclidean distance metric. But, there is a growing awareness that the Euclidean distance is a very brittle distance measure. Afterthat, Berndt and Clifford introduced the new technique, dynamic time warping (DTW) to the database community (Berndt and Clifford 1994). Although they demonstrate the utility ofthe approach, they acknowledge that its resistance to indexing is a problem andthat " ... performance on very large databases may be a limitation." Despite this short coming of DTW, it is still widely used in various fields. In bioinformatics, Aach and Church (2001) applied DTW to RNA expression data. In chemical engineering (Gollmer and Posten 1995), it has been used for the synchronization and monitoring of batch processes in polymerization. DTW has been used successfully to arrange biometric data, such as gait (Gavrila and Davis 1995), signatures (Munich and Perona 1999) and even fingerprints (Kovacs Vajna 2000). Many researchers, including Caiani et al. (1998) have demonstrated theutility of DTW for ECG pattern matching. Rath and Manmatha have applied DTW to the problem of indexing repositories of handwritten historical documents (Rath and Manmatha 2002). Finally, in robotics, Schmill et al. demonstrated a technique that utilizes DTW to cluster an agent's sensory outputs (Schmill et al. 1999).

Note that, while the two time series have an overall similar shape and different lengths, they are not aligned in the time axis. Euclidean distance, which assumes the i^{th} point in one sequence is aligned with the i^{th} point in the other, will produce a pessimistic dissimilarity measure. The nonlinear dynamic time warped alignment allows a more intuitive distance measure to be calculated (figure 2.3)

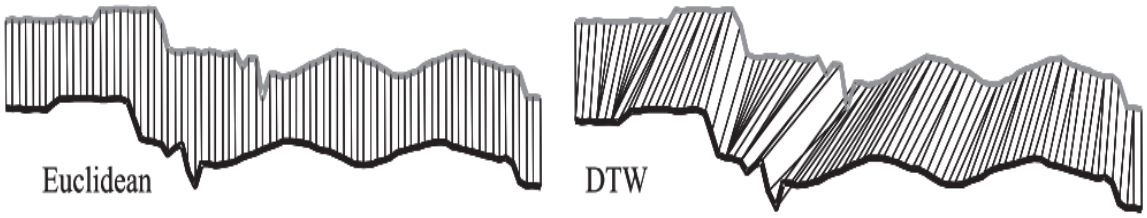

Figure 2.2 Euclidean distance and DTW distance

A distance measurement between time series is needed to determine similarity between time series and for time series classification. Euclidean distance is an efficient distance measurement that can be used. The Euclidian distance between two time series is simply the sum of the squared distances from each nth point in one time series to the nth point in the other. The main disadvantage of using Euclidean distance for time series data is that its results are very unintuitive. If two time series are identical, but one is shifted slightly along the time axis, then Euclidean distance may consider them to be very different from each other. Dynamic time warping (DTW) was introduced[11] to overcome this limitation and give intuitive distance measurements between time series by ignoring both global and local shifts in the time dimension.

Dynamic Time Warping (DTW) is also a mathematical method that allows the comparison of two arrays of data. As it has been previously mentioned, it is applied in several areas, mostly in time series. However, it is also possible to operate it on static data, such as fingerprints. For every two subsequences of data, the algorithm gives us, not only the information of how alike they are, but also the best correspondence among their data prints. Furthermore, the procedure is quite flexible, so it can be easily adjusted to the type of data to which one intends to operate in.

Moreover, the most important aspect of this procedure is that it comprises the fact that the sequences may endure different durations. Therefore, it tolerates compression and distension of time.

Consider two sets of points A and B, with lengths of respectively n and $m \in \mathrm{N}$. The DTW starts to measure the distance from all the points of A to all elements in B.

When one proceeds throughout this computation, there are some distance functions that can be used, for example the Manhattan Distance, the ρ-norm Distance or even the Discrete Metric. As it was suggested by Wong and Yeung (2008), in the present report one will apply the Euclidean Distance, which is a particular case of the ρ-norm Distance. One must highlight that that the DTW is more accurate than the regular Euclidean distance, since it returns a nonlinear alignment between sequences with different lengths.

The Euclidean Distance can be represented by the following formula:

$$d_{i,j} = \sqrt{(a_i - b_j)^2}\text{, where } i \in \{1,2,\dots,n\},\ j \in \{1,2,\dots,m\},\ a \in A, b \in B \qquad (1)$$

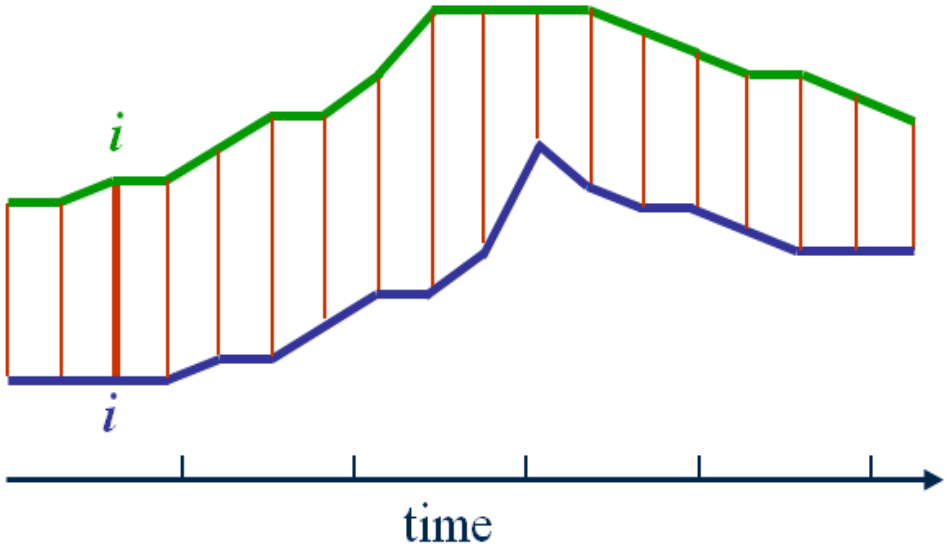

Figure 2.3 Euclidean measure distance

For instance, choose the second point of the set A and then calculate the distance to all the other points in B . Then the same procedure is repeated to the rest of the points in the first sequence. This results in a matrix that route can enclose the following aspect(figure 2.4) :

$$\begin{bmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nm} \end{bmatrix}$$

Figure 2.4 The results matrix

Remark that the entrance of the matrix (*i, j*) stands for the formula (1) applied to the point *i* in A and *j* in B.

Afterwards, the accumulated cost matrix is computed, following the rule to each cell:

$$c_{i,j} = \begin{cases} d_{i,j} & \text{if } i = 1 \wedge j = 1 \\ d_{i,j} + min\{d_{i,j-1}, d_{i-1,j}, d_{i-1,j-1}\} & otherwise \end{cases} \quad (2)$$

The element (*i, j*) of the previous matrix stands for the accumulated cost of inserting the correspondence between the points a_i with b_j in the final match. So, the value represents the cost of that correspondence, (*i, j*) plus the minimum backward correspondences in the considered path. This matrix is essential for the next step: the search for the minimum path in the accumulated cost matrix, which coincides with the best match between the sequences.

For the calculation of the optimal match, one can assume that the first elements of the sequences are connected, as well as the last ones, in order to obtain a reasonable correspondence (figure 2.5). Henceforth, the path must start at the entrance (1,1) and finish in (*n,m)*. Finally, the path is simply computed by continuously searching the smallest value in the neighbourhood of the matrix So, supposing that one is in the cell in the procedure, it means that, in order to have the best alignment, the point i of A must be connected to the element *j* of B, the next point to become part of the minimum path coincides with the lowest value in the set $G=\{d_{i, j+1}, d_{i+1,j}, d_{i+1,j+1}\}$. The cells that are not a part of the final match are

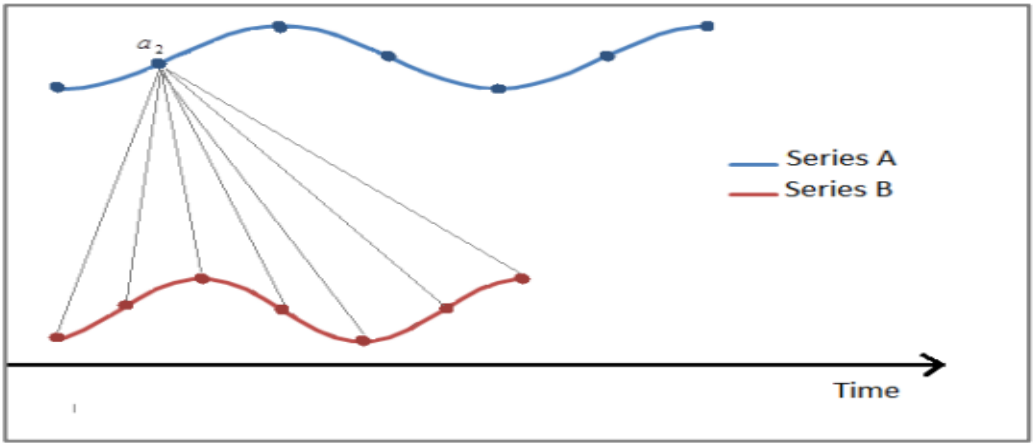

Figure 2.5 Find distance from a point in Series A to some point in Series B

The total cost of the path corresponds to the sum of the cells of the match found, which is never lower than 0. The smaller the value returned by the distance, the more alike the two sequences ought to be.

One should note that some restrictions are required, so that the result encompasses a logical output. All of them have been already referred. The Boundary Condition reassures that the point (1,1) and (*n*,*m*) are in the path (figure 2.6). The Monotonicity Condition and the Step Size Condition were implied in the set G, by guaranteeing that each cell that is added to the path, results in a movement in the sequence and that each point has a correspondence on the other series

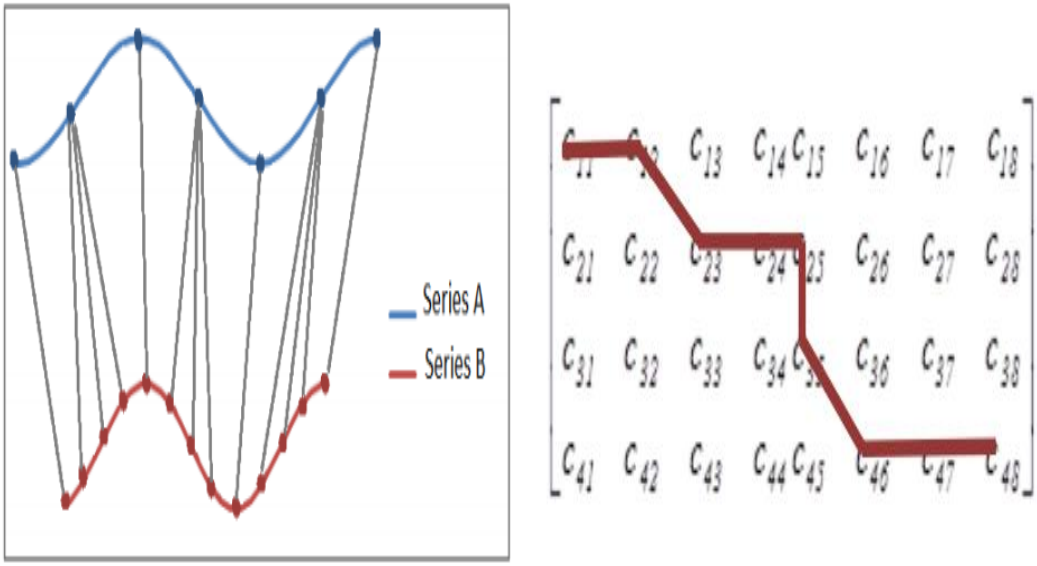

Figure 2.6 Optimal Path in an accumulated cost matrix

The DTW algorithm is shown in figure **2.7**:

- Start the calculation of g(1,1) = d(1,1).
- Calculate the first row g(***i***, 1) = g(***i***–1, 1) + d(***i***, 1).
- Calculate the first column g(1, ***j***) = g(1, ***j***) + d(1, ***j***).
- Move to the second row g(***i***, 2) = min[g(***i***, 1), g(***i***–1, 1), g(***i*** – 1, 2)] + d(***i***, 2). Book keep for each cell the index of this neighboring cell, which contributes the minimum score (red arrows).
- Carry on from left to right and from bottom to top with the rest of the grid g(***i***, ***j***) = min(g(***i***, ***j***–1), g(***i***–1, ***j***–1), g(***i*** – 1, ***j***)) + d(***i***, ***j***).
- Trace back the best path through the grid starting from g(***n***, ***m***) and moving towards g(1,1) by following the red arrows.

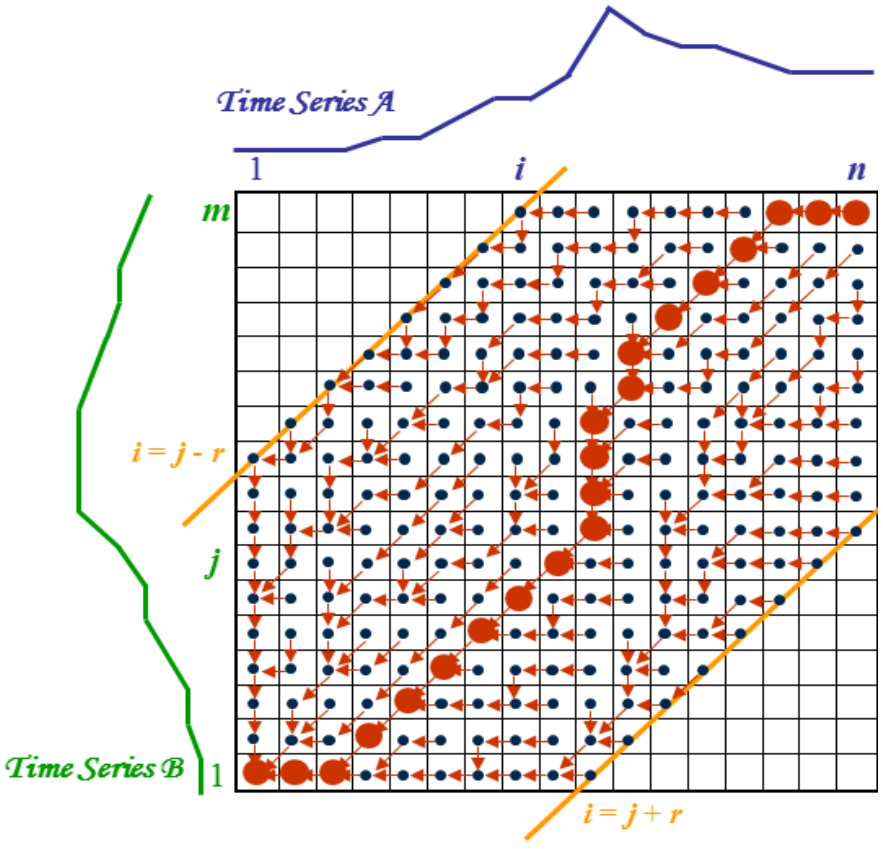

Figure 2.7 Implement Dynamic Time Warping between two time series

2.2.4 What is Hierarchical clustering algorithm

Clustering methods are usually classified with respect to their underlying algorithmic approaches. Hierarchical algorithms find successive clusters using previously established ones, whereas partitional algorithms determine all clusters at one.

Hierarchical clustering (also called hierarchical cluster analysis or HAC) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- **Agglomerative**: This is a "bottom up" approach: begin with each element as a separate cluster and merge them into successively larger clusters (figure 2.8)
- **Divisive:** This is a "top down" approach: begin with the whole set and proceed to divide it into sucessively smaller clusters.(figure 2.8)

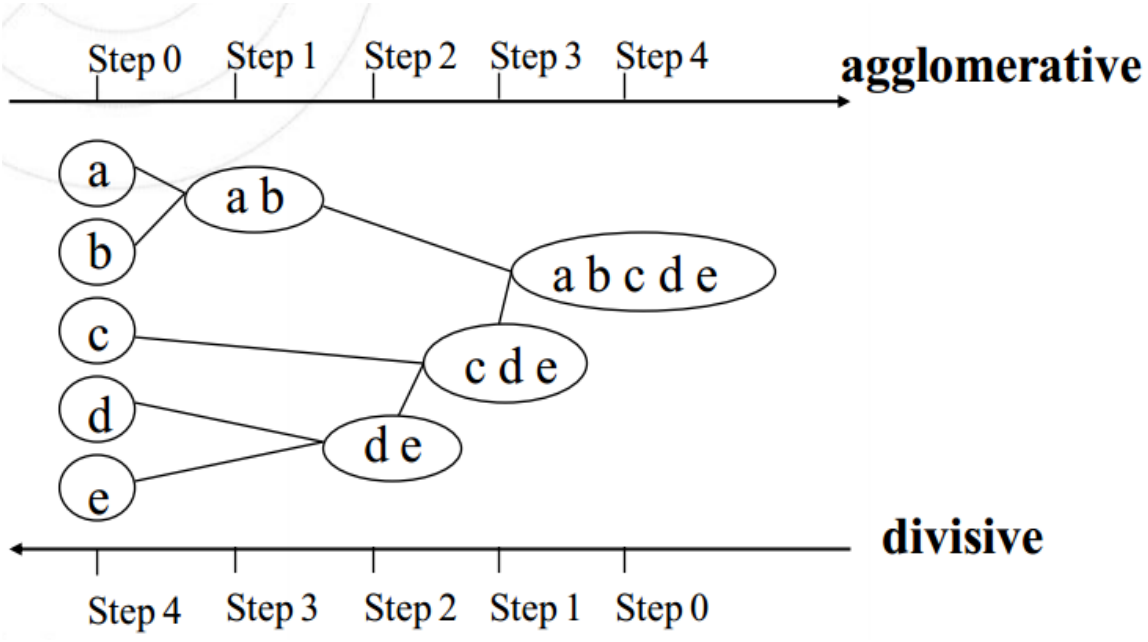

Figure 2.8 The hierarchical clustering dendrogram[1]

[1] A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes or samples.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. the complexity of agglomerative clustering is $\mathbf{O}(n^3)$, which makes them too slow for large data sets. Divisive clustering with an exhaustive search is $\mathbf{O}(2^n)$, which is even worse. However, for some special cases, optimal efficient agglomerative methods (of complexity $\mathbf{O}(n^2)$ are known: SLINK for single-linkage and CLINK for complete-linkage clustering.

This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed? For this there are many available methods. Some of them are:

Single-nearest distance or single linkage: Distance between two clusters is the distance between the closest points. Also called "neighbor joining." In the single linkage method , ***D(n,m)*** is computed as:

D(n,m) = Min {*d(i,j)* :Where object *i* is in cluster ***m*** and object *j* is in cluster ***n*** }

Here the distance between every possible object pair *(i,j)* is computed, where object *i* is in cluster ***n*** and object *j* is in cluster ***m***. The minimum value of these distances is said to be the distance between clusters ***n*** and ***m.*** .This measure of inter-group distance is illustrated in the figure below(figure 2.9)

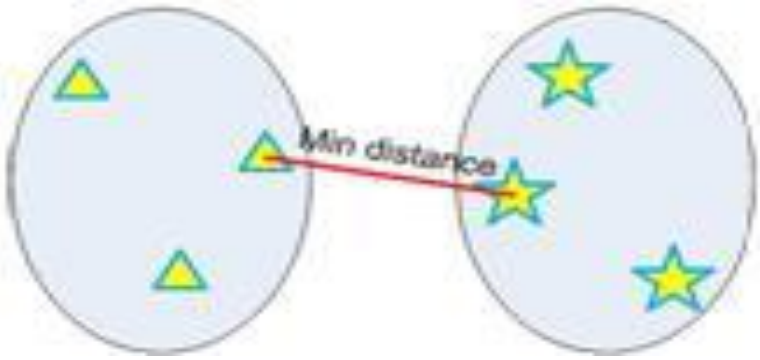

Figure 2.9 the min distance

Complete-farthest distance or complete linkage: Distance between clusters is distance between farthest pair of points. In the complete linkage method, ***D(m,n)*** is computed as:

D(m,n) = Max {*d(i,j)* : Where object *i* is in cluster ***m*** and object *j* is cluster ***n*** }

Here the distance between every possible object pair *(i,j)* is computed, where object *i* is in cluster ***m*** and object *j* is in cluster n and the maximum value of these distances is said to be the distance between clusters ***m*** and ***n.*** The measure is illustrated in the figure below (Figure 2.10):

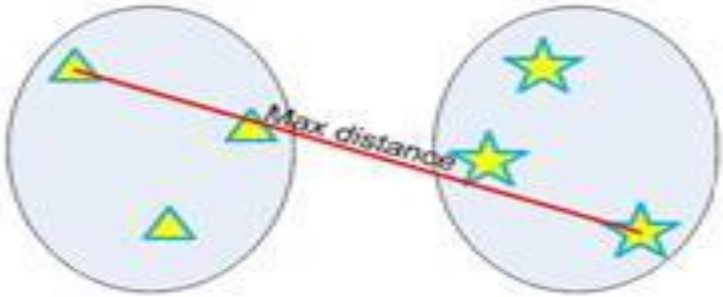

Figure 2.10 The max distance

Average-average distance or average linkage: Distance between clusters is average distance between the cluster points. In the average linkage method, ***D(r,s)*** is computed as

$$D(n,m) = T_{n,m} / (S_n * S_m)$$

Where T_{mn} is the sum of all pairwise distances between cluster ***m*** and cluster ***n.*** $S_n$and S_m are the sizes of the clusters ***m*** and ***n*** respectively.

The figure below illustrates average linkage clustering (figure 2.11):

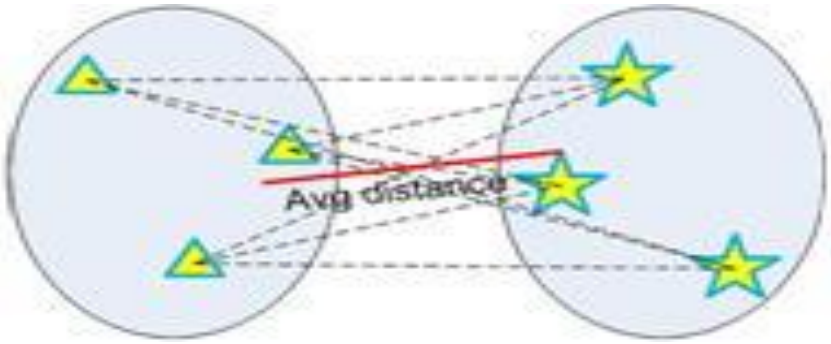

Figure 2.11 Illustrates average linkage clustering

Centroid distance: Distance between clusters is distance between centroids. In the Centroid method, the two clusters **m** and **n** are merged such that, after merger, the average pairwise distance within the newly formed cluster, is minimum. Suppose we label the new cluster formed by merging clusters **m** and **n**, as **t**. Then **D(m,)** , the distance between clusters **m** and **n** is computed as:

D(m,n) = Average { d(i,j) : Where observations i and j are in cluster **t**, the cluster formed by merging clusters **m** and **n** }

The figure below illustrates Centroid clustering (figure 2.12)

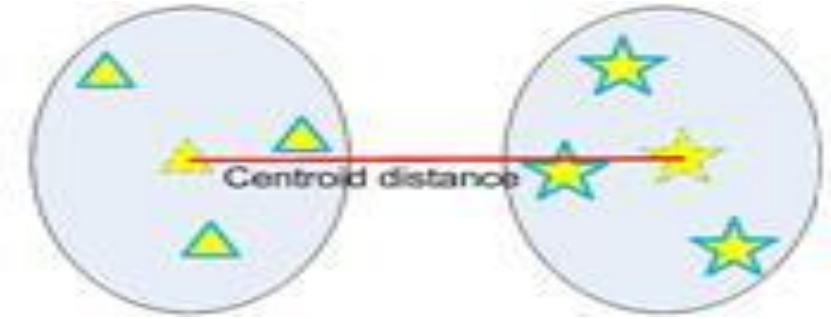

Figure 2.12 Centroid distance clustering

In the summary I have overview about hirerchical algorithm :

1. Hierarchical cluster analysis of nobjects is defined by a stepwise algorithm which merges two objects at each step, the two which havethe least dissimilarity.
2. Dissimilarities between clusters of objects can be defined in several ways; for example, the maximum dissimilarity (complete linkage), minimum dissimilarity (single linkage) or average dissimilarity (average linkage).
3. Either rows or columns of a matrix can be clustered– in each case we choose the appropriate dissimilarity measure that we prefer.
4. The results of a cluster analysis is a binary tree,or dendrogram, with n – 1 nodes. The branches of this tree are cut at a level where there is a lot of ‘space’ to cut them, that is where the jump in levels of two consecutive nodes is large.
5. A permutation test is possible to validate the chosen number of clusters, that is to see if there really is a non-random tendency for the objects to group together

Chapter 3

Problems Identification and Solving Methods

3.1 Data description

For the clustering analysis, I choose the interday stock price (Close Price) time series instead of the intraday stock price data. The choice is made considering the fact that the intraday data is very much "dynamic" and more of random nature. Price changes for different stock counters over short period at intraday level of resolution present few distinct features, making it difficult to differentiate for clustering. For the proposed implementation, I have used a dataset of 15 selected interday stock price time series. The stocks chosen span at Banking and Insurance industries at Stock exchange in Vietnam. It is believed that this diversity will make my data more suitable for the testing of my proposed clustering framework.

3.2 Clustering stock price time series data

3.2.1 Initializing Characterizations of Stock Price Clustering

Stock price clustering is pervasive in markets of many kinds. Clustering analysis of stocks is necessary when investigating in stocks.

After the data has been prepared and transformed, the next step was to build the clustering model using clustering algorithms based on distance measurement between two stocks (Time series). For clustering time series, likewise many variants of algorithm are applicable. They range from simple ones like K-means and K-medoids, to sophisticated algorithms like DBSCAN, density-based clustering for clustering structures. In my case, hierarchical clustering is desirable because it allows the time series which are stock prices data to be grouped in different levels automatically that helps a user to explore the structure of the groupings from coarse to refined. The hierarchical clustering technique was selected because the construction of hierarchical clustering does not require any domain knowledge about financial report, thus it is

appropriate for exploratory knowledge discovery. Another benefit is that the steps of hierarchical clustering induction are simple and fast. Next issue, I must choose a good solution to find distance time series. In the fact, many similarity measures are available such as Manhattan, Euclidean and Minkowski just to name a few and a range of popular similarity functions are compared in performance in order to observe which one performs the best (table 3.1). The Euclidean distance metric has been widely used, although its known weakness of sensitivity to distortion in time axism. In recent years, the Dynamic Time Warping distance measure was introduced to the data mining community as a solution to this particular weakness of Euclidean distance metric. Because the dynamic of the stock prices data that I am working with is time series, I choose to use Dynamic Time Warping function (DTW) as a distance measure that finds optimal alignment between two sequences of time series data points. DTW a pairwise comparison of the feature (or attribute) vectors in each time series. It finds an optimal. In other words, it is flexibility allows two time series that are similar but locally out of phase to align in a non-linear manner and best solution known for time series problems in a variety of domains. The sequences are "warped" nonlinearly in the time dimension to determine a measure of their similarity independent of certain nonlinear variations in the time dimension.

Table 3.1 Percentage of correctly clustered groups of some similarity functions

Canberra	DTW	Euclidean	Manhattan	Minkowski	Minkow.2	Minkow.3
86.67	91.67	63.33	63.33	63.33	63.33	66.67
Minkow.4	Minkow.5	Minkow.6	Minkow.7	Minkow.8	Minkow.9	Minkow.10
78.33	83.33	86.67	76.67	66.67	66.67	66.67

DTW optimal alignments between some points in the two stock prices time series. The alignment is optimal in the sense that it minimizes a cumulative distance measure consisting of "local" distances between aligned samples. This procedure is called time warping due to the fact that it warps the time axes of the two time series in such a way that corresponding samples appear at the same location on a common time axis.

3.2.2 Learning Procedure of Stock Price Clustering

This section examines individual stock price clustering distributions to more completely characterize the clustering .

According to the previous chapters and also my knowledge of hierarchical clustering and dynamic time series. So what is the Dynamic Time Warping for stock prices data which calculated distance measurement?

The solution of this situation I transfer the each stock data prices into the vector contains corresponding daily closing stock prices for finding distance them. Then I can apply the DTW to find distance. Of course, the distance measure affect hierarchical clustering and the process time, if the distance measure is closed, the number group stocks is exactly and also the time needed for clustering are reduced as well.

However, in this case of distance between stock prices data, they are difficult and ambitious to find the list distance at a time . My solution assigns all stocks data in the matrix.Then I apply DTW to calculate distance between pair elements and recursive with the other pair of elements in this matrix.

Back to the Dynamic Time Warping section , suppose I have two time series Q and C as first stock and second stock, of length n, m (but in this thesis I choose m=n) respectively, where

$$Q=q1,q2,...,qi,...,q_m$$

$$C=c1,c2,...,cj,...,c_n$$

For example, Q such as stockA= {0, 2, 1, 1, 1} and C such as stockB={ 0, 0, 2, 1.5, 1, 1}. In this case I align two sequences stock prices data using DTW, I construct a square matrix where the (i^{th}, j^{th}) element of the matrix contains the distanced (q_i,c_j) between the two points q_i and c_j. Each matrix element (i,j) corresponds to the alignment between the points q_i and c_j. A warping path W is a contiguous (in the sense stated below) set of matrix elements that defines a mapping between Q and C.The k^{th} element of W is defined as $w_k=(i,j)_k$. So we have

$$W=w_1,w_2,...,w_k,...,w_K \quad \max(m, n) \leq K < 2m-1$$

The warping path is typically subject to several constraints. There are exponentially many warping paths. However, we are only interested in the path that minimizes the warping cost:

$$DTW(Q, C) = \min \left\{ \sqrt{\sum_{k=1}^{K} w_k} \right.$$

This path can be found using dynamic programming to evaluate the following Recurrence, which defines the cumulative distance γ(i,j)as the distanced(i,j) found in the current cell and the minimum of the cumulative distances of the adjacent elements:

$$\gamma(i,j)=d(qi,cj)+\min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\}$$

Let me go back with the example, finding the distance between stockA, stockB suppose that the each value prices of each stock assigned by time series or vector

DTW($\overrightarrow{stockA}$, $\overrightarrow{stockB}$)= distance measure($\begin{Bmatrix} stockA1 & stockA2 & \dots stockA5 \\ stockB1 & stockB2 & \dots stockB6 \end{Bmatrix}$

With:

stockA1, stockA2, stockA 5, stockB1, stockB2, stockB6 are daily stock price values

γ[1,1] = |a[1] - b[1]| + min(γ [0,1], γ [1,0], γ[0,0]) = 0
γ [2,1] = |a[2] - b[1]| + min(γ [1,1], γ [2,0], γ [1,0]) = 2
γ [3,1] = |a[3] - b[1]| + min(γ [2,1], γ [3,0], γ [2,0]) = 3
γ [4,1] = |a[4] - b[1]| + min(γ [3,1], γ [4,0], γ [3,0]) = 4
γ [5,1] = |a[5] - b[1]| + min(γ [4,1], γ [5,0], γ [4,0]) = 5
γ [1,2] = |a[1] - b[2]| + min(γ [0,2], γ [1,1], γ [0,1]) = 0
γ [2,2] = |a[2] - b[2]| + min(γ [1,2], γ [2,1], γ [1,1]) = 2
γ [3,2] = |a[3] - b[2]| + min(γ [2,2], γ [3,1], γ [2,1]) = 3
γ [4,2] = |a[4] - b[2]| + min(γ [3,2], γ [4,1], γ [3,1]) = 4
γ [5,2] = |a[5] - b[2]| + min(γ [4,2], γ [5,1], γ [4,1]) = 5
γ [1,3] = |a[1] - b[3]| + min(γ [0,3], γ [1,2], γ [0,2]) = 2
γ [2,3] = |a[2] - b[3]| + min(γ [1,3], γ [2,2], γ [1,2]) = 0
γ [3,3] = |a[3] - b[3]| + min(γ [2,3], γ [3,2], γ [2,2]) = 1
γ [4,3] = |a[4] - b[3]| + min(γ [3,3], γ [4,2], γ [3,2]) = 2
γ [5,3] = |a[5] - b[3]| + min(γ [4,3], γ [5,2], γ [4,2]) = 3
γ [1,4] = |a[1] - b[4]| + min(γ [0,4], γ [1,3], γ [0,3]) = 3.5
γ [2,4] = |a[2] - b[4]| + min(γ [1,4], γ [2,3], γ [1,3]) = 0.5
γ [3,4] = |a[3] - b[4]| + min(γ [2,4], γ [3,3], γ [2,3]) = 0.5
γ [4,4] = |a[4] - b[4]| + min(γ [3,4], γ [4,3], γ [3,3]) = 1
γ [5,4] = |a[5] - b[4]| + min(γ [4,4], γ [5,3], γ [4,3]) = 1.5

γ [1,5] = |a[1] - b[5]| + min(γ [0,5], γ [1,4], γ [0,4]) = 4.5
γ [2,5] = |a[2] - b[5]| + min(γ [1,5], γ [2,4], γ [1,4]) = 1.5
γ [3,5] = |a[3] - b[5]| + min(γ [2,5], γ [3,4], γ [2,4]) = 0.5
γ [4,5] = |a[4] - b[5]| + min(γ [3,5], γ [4,4], γ [3,4]) = 0.5
γ [5,5] = |a[5] - b[5]| + min(γ [4,5], γ [5,4], γ [4,4]) = 0.5
γ [1,6] = |a[1] - b[6]| + min(γ [0,6], γ [1,5], γ [0,5]) = 5.5
γ [2,6] = |a[2] - b[6]| + min(γ [1,6], γ [2,5], γ [1,5]) = 2.5
γ [3,6] = |a[3] - b[6]| + min(γ [2,6], γ [3,5], γ [2,5]) = 0.5
γ [4,6] = |a[4] - b[6]| + min(γ [3,6], γ [4,5], γ [3,5]) = 0.5
γ [5,6] = |a[5] - b[6]| + min(γ [4,6], γ [5,5], γ [4,5]) = 0.5

So warped distance of stockA and stockB is γ [m,n] = γ [5,6]

In order to make more clear, these figure below is the pseudocodes of the Dynamic Time Warping distance function (Figure 3.2):

```
DTW(v1, v2) {
//where the vectors v1=(a1,…,an), v2=(b1,…,bm) are the time series with n and m
time points
  Let a two dimensional data matrix S be the store of similarity measures
such that S|0,…,n, 0,…,m|, and i, j, are loop index, cost is an integer.
  // initialize the data matrix
  S|0, 0| := 0
  FOR i := 1 to m DO LOOP
    S|0, i| := ∞
  END
  FOR i := 1 to n DO LOOP
      S|i, 0| := ∞
  END
  // Using pairwise method, incrementally fill in the similarity matrix
  with the differences of the two time series
  FOR i := 1 to n DO LOOP
    FOR j := 1 to m DO LOOP
    // function to measure the distance between the two points
      cost := d(v1|i|, v2|j|)
      S[i, j] := cost + MIN(S|i-1, j|,      // increment
                      S|i, j-1|,    // decrement
                      S|i-1, j-1|)  // match
    END
  END
  Return S|n, m|
}
```

Figure 3.1 Pseudo code of dynamic time wrap algorithm

The next step finding the distance measure stock prices time series is clustering based on distance. Because hierarchical clustering allows the data to decide the suitable number of groups by themselves. I chose agglomerative mode which is also

known as the "bottom up" with average linkage approach is used. Initially each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. For deciding which clusters should be merged, a similarity function is used between sets of observations. The clustering algorithm constructs the hierarchy from the individual time series by progressively merging clusters up. The basic process of hierarchical clustering comprises of the following steps, given n time series as number stocks in the input data, and a two dimensional $n \times n$ similarity matrix S:

Step 1. Each stock prices time series is assigned to a cluster of its own, with a total of n clusters for n number stocks. Initialize S with similarity measures between the clusters which are the same as the similarity measure between the stock prices time series that they contain.

Step 2. The most similar pair of clusters are merged into a single cluster. Retain the current level of clusters and move up a level in the hierarchy.

Step 3. Calculate the new similarity measures in S between the new clusters and each of the old clusters.

Step 4. Finally repeat Step 2 and Step 3 until all the stock prices time series are clustered into a single cluster of size n. When this happens, the highest level of the hierarchy is attained.

I can write this as an algorithm (figure 3.3)

Each $x_i \in X$ is a separate cluster S_i.
while Two clusters are *close enough* **do**
  Find the *closest* two clusters S_i, S_j
  Merge S_i and S_2 into a single cluster

Figure 3.2 Hierarchical Clustering

Let me find some group stock after I have distance matrix of them. For example, I have 7 stocks with stock A, B, C, D, E, F, G (figure 3.4)

NameStock	A	B	C	D	E	F	G
A							
B	0.5						
C	0.4286	0.7143					
D	1	0.8333	1				
E	0.25	0.6667	0.4286	1			
F	0.625	0.2	0.6667	0.8	0.7778		
G	0.375	0.7778	0.3333	0.8571	0.375	0.75	

Figure 3.3 Distance matrix of some stock

The first step in the hierarchical clustering process is to look for the pair of stocks that are the most similar, that is are the closest in the sense of having the lowest dissimilarity – this is the pair B and F, with dissimilarity equal to 0.2000. These two stocks are then joined at a level of 0.2000 in the first step of clustering tree. The point at which they are joined is called a *node*

I am going to keep repeating this step, but the only problem is how to calculated the dissimilarity between the merged pair (B,F) and the other stock. This decision determines what type of hierarchical clustering I intend to perform and there are several choices. I choose one of the most popular ones, called the average method: the dissimilarity between the merged pair and the others will be the the average distance between each point. For example, the dissimilarity between B and A is 0.5, F and A is 0.625. Hence I choose the average of the two is 0.5625 to quantify the dissimilarity between (B,F) and A. Continuing in this way we obtain a new dissimilarity matrix (figure 3.4)

NameStock	A	B,F	C	D	E	G
A						
B,F	0.5625					
C	0.4286	0.6905				
D	1	0.81665	1			
E	0.25	0.72225	0.4286	1		
G	0.375	0.7639	0.3333	0.8571	0.375	

Figure 3.4 Distance similarities calculated after B and F are merged

The process is now repeated: find the smallest dissimilarity in figure 3.5 which is 0.25 for samples A and E, and then cluster these at a level of 0.25. Then recomputed the dissimilarities between the merged pair (A,E) and the rest to obtain figure 3.5. For example, the dissimilarity between (A,E) and (B,F) is the average of 0.5625 (A to (B,F)) and 0.72225 (E to (B,F)).

NameStock	A,E	B,F	C	D	G
A,E					
B,F	0.642375				
C	0.4286	0.6905			
D	1	0.81665	1		
G	0.375	0.7639	0.3333	0.8571	

Figure 3.5 Distance similarities calculated after A and E are merged

After first two steps of hierarchical clustering, using the ‘average’ method, I have clustering tree in below figure

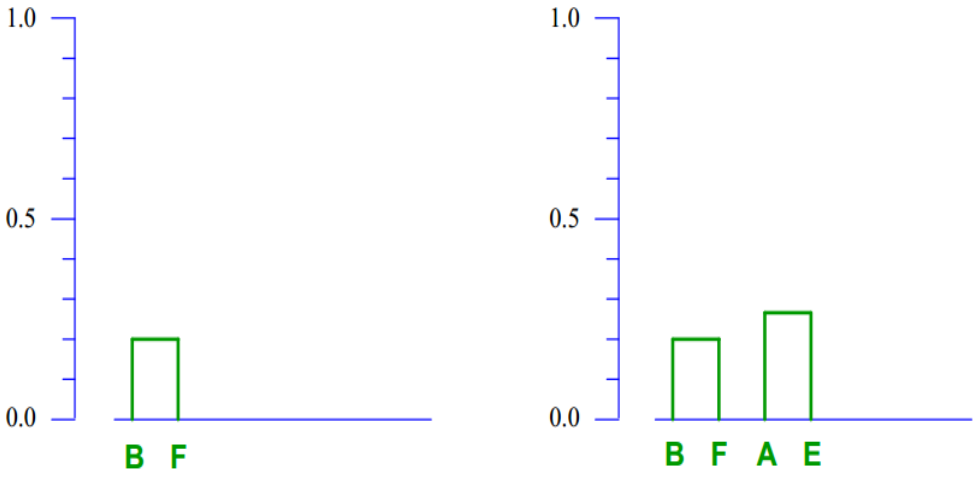

Figure 3.6 Two cluster of two step clustering

Similar, I have distances tables after merging in the below figures (firgure 3.7, 3.8, 3.9)

NameStock	A,E	B,F	C,G	D
A,E				
B,F	0.642375			
C,G	0.4018	0.7272		
D	1	0.81665	1	

Figure 3.7 Distance similarities calculated after C and G merged

NameStock	A,E,C,G	B,F	D
A,E,C,G			
B,F	0.684788		
D	1	0.81665	

Figure 3.8 Distance similarities calculated after A,E and C,G merged

NameStock	A,E,C,G,B,F	D
A,E,C,G, B, F		
D	0.750719	

Figure 3.9 Distance similarities calculated after A,E,C, G and B,F merged

Because there are 7 stocks to be clustered, there are 6 steps in the sequential process to arrive at the final tree where all stock are in a single cluster (figure 3.7). For botanists that may be reading this: this is an upside-down tree. The dendrogram on the right is the final result of the cluster analysis. In the clustering of nobjects, there are n – 1 nodes (i.e. 6 nodes in this case).

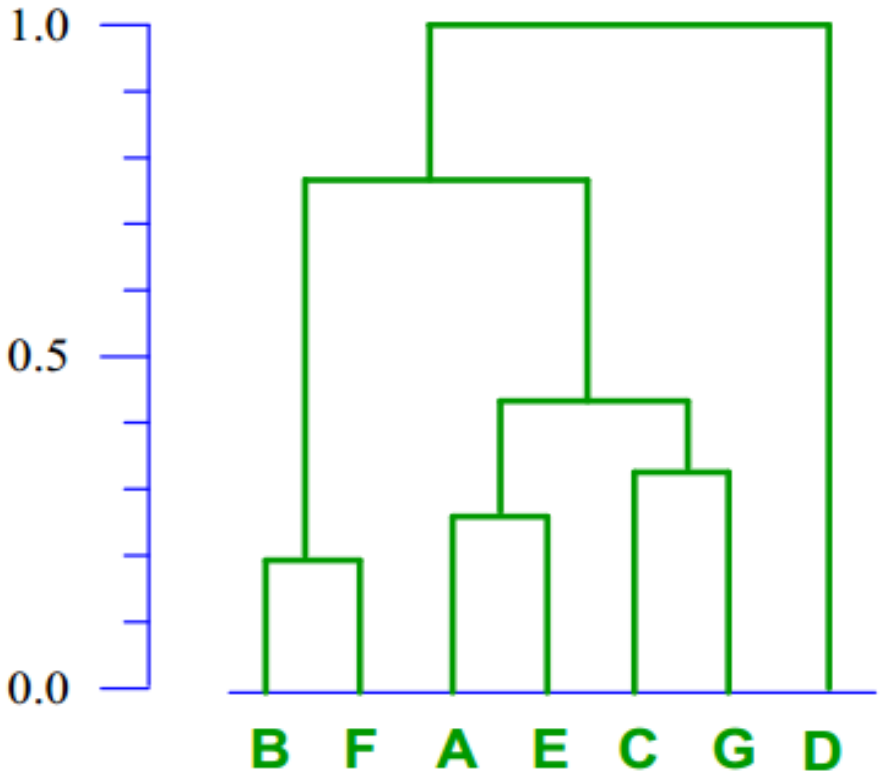

Figure 3.10 Sixth steps of hierarchical clustering

In summary, I need find distance measure between stock prices data by Dynamic DTW to assigned matrix of distances. Then I perform HAC to do initial clustering and construct a dendrogram. Finally I will implement them in next chapter.

Chapter 4

Sytem Implementation and Testing

This chapter is an answer for the question how to develop the application to test and demonstrate these theories in the real computing system.

This chapter mainly indicates the two parts to implement clustering actors by HAC and DTW:

- System Implementation
- System Test

4.1 System Implementation

4.1.1 Programming Framework

In order to clustering stock prices data from historical data, I decide to implement my approach as desktop application, which can be done easily with the investors. I have two iusses need to be considered:

- Why Microsoft .NET Framework and Windows Form C# programming language is chosen?
- Why DevExpress tool is chosen to design Graphic User Interface?

To begin with, let get an overview of Microsoft .NET Framework. As I know,

Microsoft .NET Framework was introduced in 2002 by Microsoft. this is a software framework that supports many programming languages such as C++, C#, VB, JScript and provides languages interoperability, which means that each language can use code in other languages. Included with .NET is a substantially number of library, which is available to all languages supported by .NET. It means like Java that code of a program is not complied as native code for computer such as C programming language anymore, the code is interpreted to bytecode in order to run in

any computing system platform with support this build-inframework.

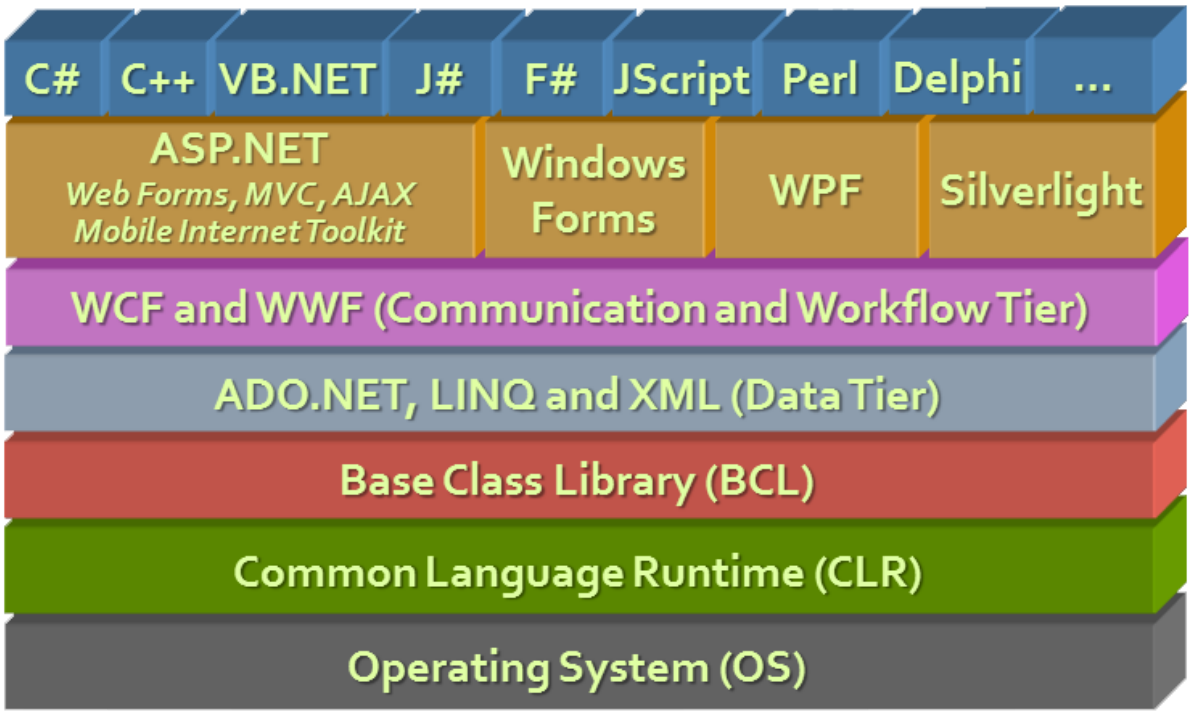

Figure 4.1.NET Architecture

The execution process of .NET framework is as follow:

- Code written is compiled using its appropriate compiler, this code is then called IL (Intermediate Language) or managed code
- The CLR (Common Language Runtime) compile the managed code again and convert it to native code by JIT (Just In-time Compiler)
- The native code is passed to operating system or hardware

So the problem is why we choose .NET Framework and Windows Forms C# ? The main reason is .NET Framework has more ability to interact with Windows Operating System than other language such as representing Graphic User Interface (GUI) to make the demo become clearly and easily. Secondly, .NET Framework has several powerful tools in editing programming code or designing GUI such as Microsoft Visual Studio so the time to building the demo application is reduced. The last reason is the .NET Framework's programming language. C# is approximately similar to Java so if we want to reuse the code in another computing platform which

Java has more advantages to implement such as web platform, it is not so hard to convert and reuse the code.

Due I need to view chart data stock, I need strong tool support to display detail stock data. Furthermore, Some control and button need to make great feature. Therefore I choose DevExpress which is 3rd party tool.

In this thesis, I code my application using Windows Form C# combine DevExpress tool to design interface. This version of DevExpress is v14.1

4.1.2 Programming Implement

. In this sector I will explain my implementation of main methods to perform clustering. This will be presented according to the following process (figure 4.2):

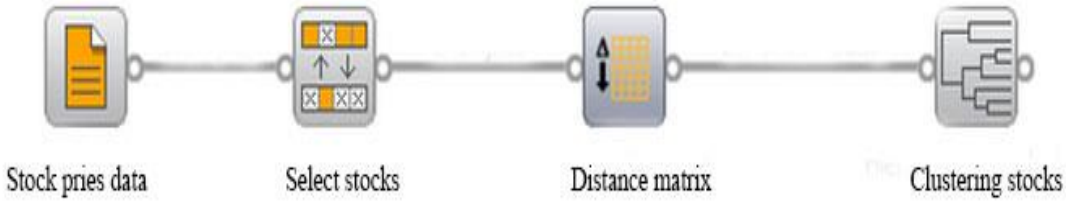

Figure 4.2 Clustering process in stock prices data

The above process can be described as follow:

1. **Stock prices data**: The only thing user need to input file excel includes stocks prices data which that they want. User can input a number from 1 to 8, more than 8 clusters bring no advantage or better meaning to the clustering result.
2. ***Select stocks:*** User must choose stock name which want to cluster them.
3. ***Distance matrix:*** the initial seeds and the set of data taken from the excel are passed to Dynamic Time Warping function to calculator distance measure between stocks to their distance matrix.
4. ***Clustering stocks*** :The clustering process is run some times and after each time, the clustering result was saved and display all groups.

Thus, this is the main interface of the application:.

1. Import Data Tab: By clicking “Browse”button, users can be able to open stock prices data file by Open File Dialog, then user click “Load” button to assign all context at file to data of application
2. List Stock Tab: It will show a check list box which includes name of stock of file, users can choose stock name in check list to start distance measure time seris and the clustering process by clicking “Clustering” button.

   3. Clustering Stock Tab: It will show a check list box again which is automatically generated stock name after finishing the clustering process. If users want to view the chart of a group stock, user click the “Chart” button after checked the stocks.

The following figure is the main interface of the demo application (figure 4.3) with the number denoting each function which I have already discussed:

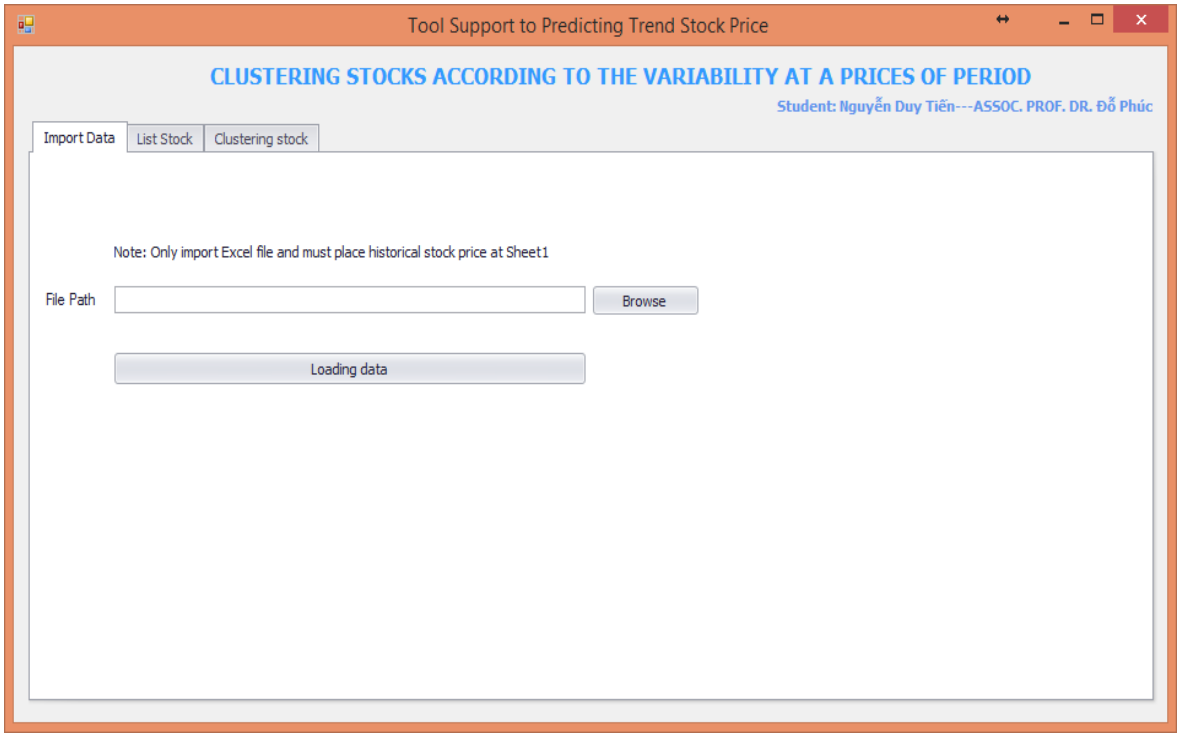

Figure 4.3 The main interface of the demo application

Assume that users click the open button to load the input file, they should see Open File Dialog shown in figure 4.4:

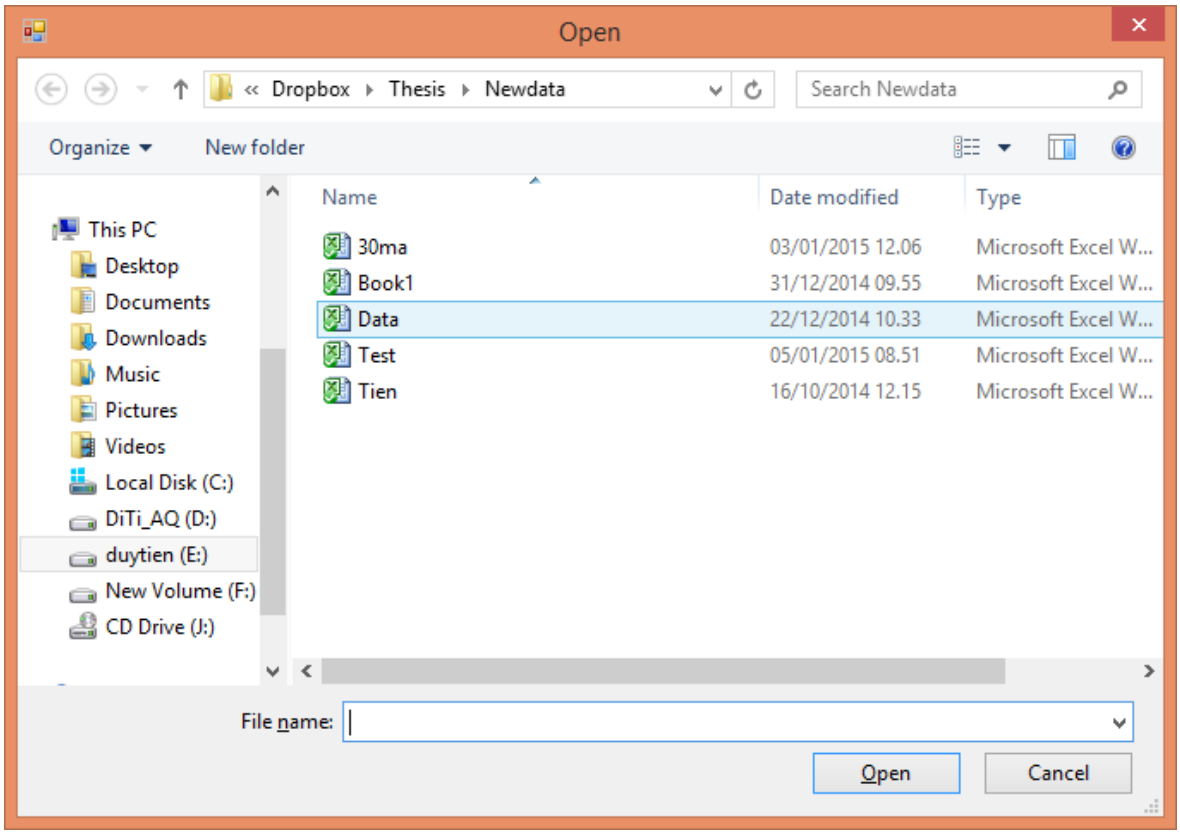

Figure 4.4 Open File Dialog for input data file

Then, Path file assign to File Path textbox and user must click Loading data to import to application.

Then, List Stock tab appears:

1. By checking stock name, user can choose some stock to view trend of cluster stock after clustering for predicting
2. By choosing format from date and to date, user can be filter number of days in process clustering
3. By clicking “Clustering stock” to start clustering process with selected stocks

To be more understandable, figure 4.5 describes all of these functions are denoted as number as follow:

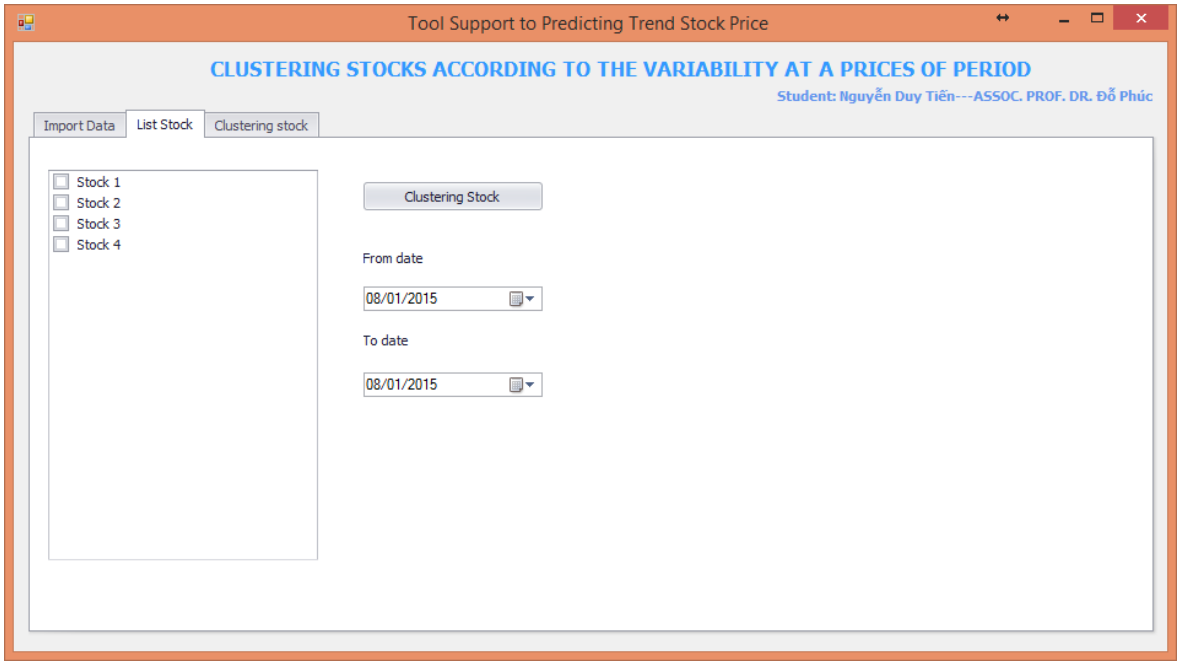

Figure 4.5 List Stock form

Finally, Clustering Stock tab appear, after all the processes have been succeeded, they show some cluster stock , users can click to checked cluster stock and can be view chart each cluster by clicking button in this tab (figure 4.6)

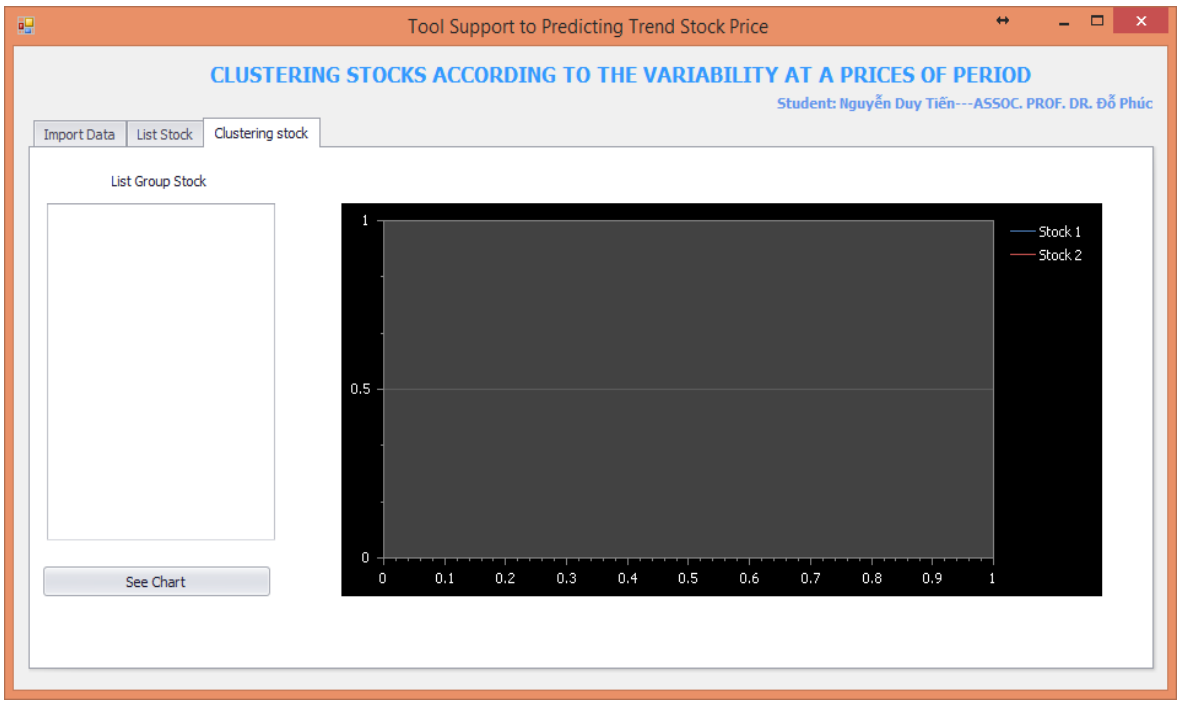

Figure 4.6 Clustering Stock form appear after clustering process success.

Note: I use the chart control of DevExpress to show time series prices data of each stock. That's why chart type of DevExpress have support for Stock. So I want to use it to show clearly.

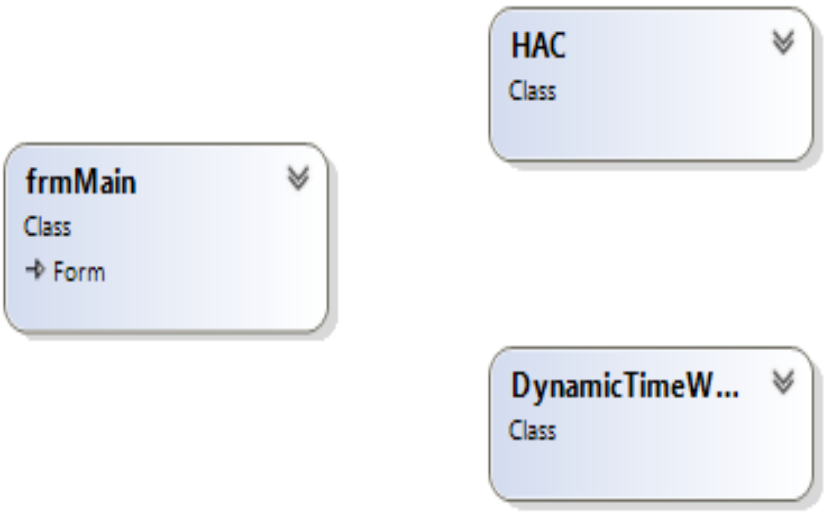

Figure 4.7 Main Form, Dynamic Time Warping and HAC class

4.2 System Test

To begin with, I go back to two algorithms of clustering stock which has been discussed in chapter 3. The prerequisite of my processes is the encoding process which is based on DTW measuring similarity and HAC method average linkage in this case. Thus, the testing and assessment procedure is these steps.

Firstly, I start with an initial of 15 stocks of Banking and Insurance industries. I get historical stock datas from website http://www.cophieu68.vn which is free for everyone to use . First I download each stock prices data. Next I merge and format all data to get datetime stock name and close prices . For each of the 15 stocks, I use a stock prices of 661 different times from 3th January 2012 to 29th August 2014. Then I export a file which include 15 stock prices data as below figure (figure 3.1)

NameStock	20140829	20140828	20140827	20140826	20140825	20140822	20140821	20140820	20140819	20140818	20140815	20140814	20140813	20140812
ACB	15.2	15.2	15.4	15.3	15.2	15.2	15.3	15.3	15.3	15.4	15.4	15.4	15.4	15.5
BIC	12.3	12.3	12.1	12.1	12	12.2	12.3	11.7	11.7	12.1	11.7	11.6	11.7	11.6
BMI	17.4	17.4	17.7	17.7	17.7	17.7	17.4	17.4	17	17.4	17	16.8	16.8	16.3
BVH	40	42.7	43.9	42.7	45.1	46.4	46	46.4	45.5	46.4	45.9	45.8	46.3	44.3
CTG	14.5	14.5	14.6	14.6	14.6	14.6	14.6	14.6	14.6	14.8	14.9	14.9	15.2	14.8
CTS	10.5	10.6	11.5	10.5	10.2	10.6	11.1	10.7	10.9	11	10.8	10.7	11.1	10.9
EIB	12.3	12.3	12.4	12.6	12.6	12.6	12.5	12.5	12.5	12.6	12.6	12.8	12.7	12.8
MBB	13.5	13.6	13.7	13.6	13.6	13.7	13.7	13.7	13.9	14.1	14.2	14.2	14.3	14
VCB	26.3	27.3	27.2	27.4	27.8	28.5	28.5	28.6	28.6	29.3	29.3	28.8	28.3	28
NVB	6.6	6.7	6.7	6.7	6.3	6.5	6.8	6.8	6.5	6.3	6.7	6.7	6.8	6.3
PGI	10	10	10	10	10	10	10	9.9	9.7	10.3	9.7	9.4	9.5	9
PTI	11.2	10.7	10.9	10.8	10.9	11	11.1	10.6	10.2	10.2	10.4	10.7	10.7	10.3
PVI	18.7	18.5	18.8	18.6	19	18.5	18.5	18.6	18	18.9	18.9	18.8	18.8	18.9
SHB	9.1	9.3	9.3	9.3	9.2	9.4	9.4	9.5	9.4	9.7	9.8	9.8	9.7	9.5
STB	18.5	19.1	19	19.2	19.5	19.7	19.8	20	19.6	19.7	19.6	19.5	19.5	19.4

Figure 4.8 Data includes 15 Stock Prices Time Series

After modifying the data file, I can import them into my application using Import data tab to implement (figure 4.9). Having the data now in place, I can now start testing my application

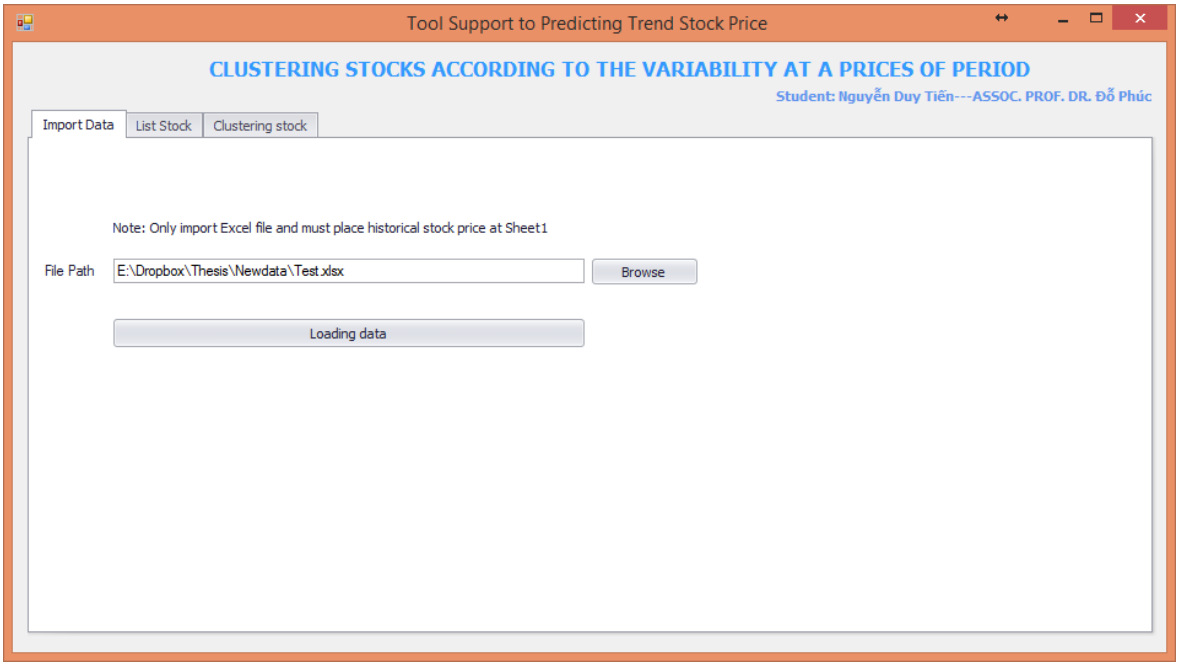

Figure 4.9 Import data file to application

Next tab I can choose stocks name to performing clustering analysis on the set of previously inputted data. In this case I choose 6 stocks to analysis, from data is 03/01/2012 to data is 29/08/2014 (figure 4.10). After that I must click "Clustering Stock" button to processing.

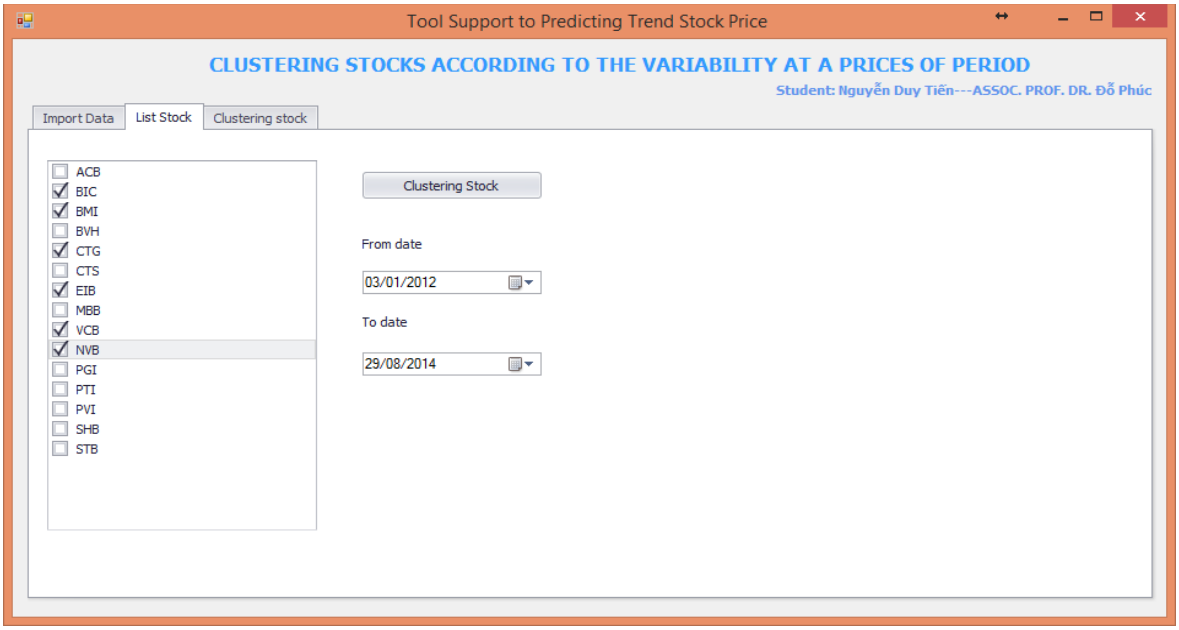

Figure 4.10 Selected stocks name to analysis

Based on DTW distance measure and HAC to clustering, I can see 3 stocks cluster in "List Group Stock" list check box after processing by using (figure 4.11)

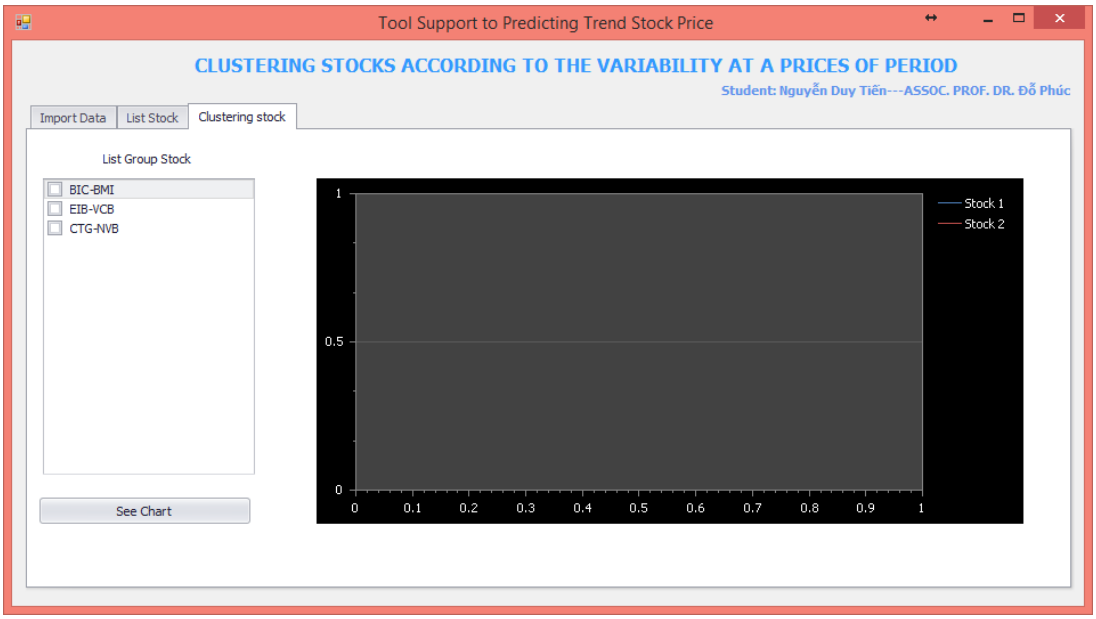

Figure 4.11 Stock Cluster after processing

Finally, based on the clustering result, I can select cluster to show the chart time series of each cluster by using checked and clicking “See chart” button. Such as (figure 4.12)

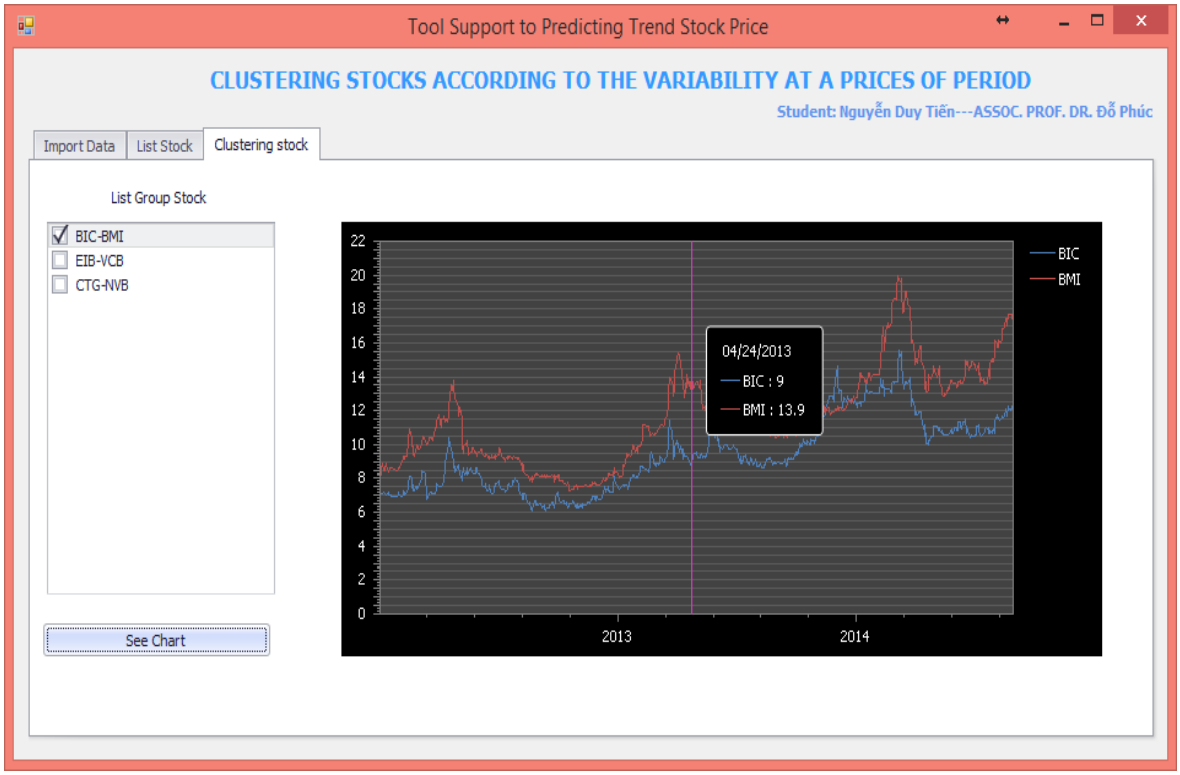

Figure 4.12 The chart of stock cluster which includes: PGI and CTS

Let me began exploring knowledge on the historical stock prices, I first make a plot of the raw data, i.e. the daily close prices of BIC and BMI stock. The plot is shown below (figure 4.13)

Figure 4.13 Daily close prices of two stocks: BMI and BIC

This plot shows that the close price of two stocks seems to increases lightly with time. This shows me that the two companies developed well during this times. Another case, this helps me choose right time to buy or sell for BIC and BMI stocks, based on the cycle of prices trend in this chart.

Chapter 5

Conclusion and Future Development

5.1 Conclusion

In summary, stock market analysis is recognized as a highly complex domain of research. I have proposed a novel and effective framework for classifying stock time series based on similarity in the price trends. This article characterizes stock price clustering and its relation to several observable characteristics of stock prices. Clustering increases with price level and volatility. Clustering decreases with market value and transaction frequency: Clustering is greater in dealer markets than in public auction exchange markets.

The proposed clustering framework for stock time series, benefits from the multi-resolution capability to analyze the nonlinear similarities between stock price time-series at different time. In this thesis I present the measuring similarities using the algorithm Dynamic Time Warping then based on them to cluster stocks by hierarchical cluster analysis. As the results for the proposed model were not perfect because many factors including but not limited to political events, general economic conditions, and investors' expectations influence stock market.

After that, the thesis states in details of the structure of the demo application which is the combination of the implementation of the DTW and the HAC and developed in C# programming language in .NET Framework 4.5. In the same vein, the testing and assessing procedure and the testing historical stock prices data also has been cited

5.2 Future Development

My demo application consists of basic analysis and functionalities of historical prices data(the close prices). Therefore, required more works to make it a real-life working system.

Firstly, a module for financial report should be integrated right into the my application so that investors can view and compare prices between stocks in each quarter which consistent with the financial situation of each company.

Next, it is necessary to do is to develop automatically update the stock prices data of each companies. More attributes also means more cells and require more computations, so it will present all attributes in chart which help more inventors to consider the stock market so much.

Then, improving code to implement two algorithm processing and language support must be desirable. For improving algorithm processing, the system should be able to have good collections to record temp data procedure. For improving language support, encoding process should be support several languages such as French, Chinese and so on. The solution for this issue is that we enlarge the encoding word dictionary for these languages

Finally, the current approach is only 2 algorithm for clustering close prices data and thus the system should be able to have more option method suitable for clustering such as K-Mean, Decision Tree, so on.

REFERENCES

[1]"Dynamic Time Warping" [Online]. Available http://en.wikipedia.org/wiki/Dynamic_time_warping

[2]“Hierarchical clustering” [Online]. Available http://en.wikipedia.org/wiki/Hierarchical_clustering

[3]Banavas, Denham, S. and Denham, “Fast nonlinear deterministic forecasting of segmented stock indices using pattern matching and embedding techniques”, Computing in Economics and Finance, 2000.

[4]Berndt, D. and Clifford, “Using Dynamic Time Warping to Find Patterns in Time Series”, Workshop on Knowledge Discovery in Databases, Seattle, Washington, 1994.

[5]C. Giles, S. Lawrence and A. Tsoi “Noisy Time Series Prediction using a Recurrent Neural Network and Grammatical Inference”, Machine Learning, Vol. 44, p. 161-183, 2001.

[6]Ehsan Hajizadeh, Hamed Davari Ardakaniand Jamal Shahrabi, “Appilication of data mining techniques in stock market”, Journal of Economics and International Finance Vol. 2, pp. 109-118, 2010

[7]Enke, D., Thawornwong, “The use of data mining and neural networks for forecasting stock market returns”, Expert Systems with Applications, pp. 927-940, 2005

[8]Fama, E.F., French, K.R., “Common risk factors in the returns on stocks and bonds”, The Journal of Finance, 33, pp. 3-56, 1993

[9]G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth,“Rule Discovery from Time Series.”, Proc. of the KDD, p. 16-22, 1998.

[10]Giorgino,“Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package” In Journal of Statistical Software 3,pp.1-24,2009

[11]Keogh, E. Pazzani, M. “Derivative Dynamic Time Warping”, First SIAM International Conference on Data Mining, Chicago, USA, 2001

[12]L.-A.Yu and S.-Y. Wang,“Kernel Principal Component Clustering Methodology for Stock Categorization,” Sys-tem Engineering-Theory & Practice, Vol. 29,pp.1-8, 2009

[13]M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani,“Mining the Stock Market: Which Measure is Best?” , Proc. of the KDD, p. 487-496, 2000

[14]Niederhoffer, V., “ A new look at clustering of stock prices”, Journal of Business 39, pp.309-313, 1966

[15]Wang, Y.F., “Predicting stock price using fuzzy grey prediction system”, Expert Systems with Applications, pp. 33-39, 2002

[16]Wang, Y.F. “Mining stock price using fuzzy rough set system”, Expert Systems with Applications, pp. 13-23, 2003

[17]Wang, J.L., Chan, S.H, “Stock market trading rule discovery using two-layer bias decision tree”, Expert Systems with Applications, pp. 605-611, 2006

[18]Xiaozhe Wang,Kate Smith and Rob Hyndman“Characteristic-Based Clustering for Time Series Data”, Data Mining and Knowledge Discovery, Springer Science + Business Media, pp. 335–364, 2006

APPENDICES

Stock data

These are the 15 stocks used by the clustering methods described in Sec 4.. They are Banking and Insurance industrial groups in Viet Nam (Table 5.1) . Data for all stocks was daily change from 3th January 2012 to 29th August 2014

Table 5.1 Stocks of Banking and Insurance Industrial

Stock Name	Company Name
ACB	Asia Commercial Bank
BIC	BIDV Insurance Corporation
BMI	Bao Minh Insurance Corporation
BVH	Bao Viet Holdings
CTG	Vietnam Joint Stock Commercial Bank for Industry and Trade
CTS	Viet Nam Bank For Industry & Trade Securities JSC
EIB	Vietnam Commercial Joint Stock Export Import Bank
MBB	Military Commercial Joint Stock Bank
NVB	National Citizen Commercial Joint Stock Bank
PGI	Petrolimex Insurance Corporation
PTI	Post - Telecommunication Joint - Stock Insurance Corporation
PVI	PVI Holdings
SHB	Saigon Hanoi Commercial Joint Stock Bank
STB	Sai Gon Thuong Tin Commercial Joint Stock Bank
VCB	Bank for Foreign Trade of Vietnam