**Time Series Data: Stock Data Prediction**

Undergraduate # 12

Team member 1 (Michael Capriles, student ID, *mcaprile@kent.edu*)

Team member 2 (Adam Duryea, 810599988, *aduryea@kent.edu*)

Team member 3 (Shane Favours, student ID, *sfavours@kent.edu*)

Team member 4 (Samuel Pittman, 810666476, *spittma4@kent.edu*)

**1. Introduction**

* Motivation examples of this project(remove this later)
* Real applications

With the Stock Market always going up and down wouldn’t it be wonderful if there was a way to predict whether a company’s stock would go up and down in price. With time-series data, we can use past data to predict what the future of a company’s stock will do tomorrow. By studying the past data of a company, you can have an educated prediction to know whether you should buy or sell tomorrow.

This can be used in real applications for tracking and predicting what a company stock will do tomorrow. Using a company’s stock past data, we can look for patterns to predict whether or not a company stock will either rise or fall tomorrow. It could also be used with other data for predicting other things in life. In addition to this, any application that uses a time series data could be used to predict future data e.g. Network Monitoring, Monitoring Weather Temperature.

**2. Project Description**

* Brief descriptions of your project
* Challenges and technical contributions (new problems or new solutions?) in your project
* The workload distribution for each member in your team

Our project is to examine a company stock data and look for patterns that could be used to predict how the stock would perform tomorrow. Our project includes research to compare the different Efficiency Similarity Search prediction approaches outlined in the paper “Efficient Similarity Search over Future Stream Time Series” by Xiang Lang, and Lei Chen. The three prediction approach methods outlined in the paper (and to be compared in our project) are the polynomial prediction method, Discrete Fourier Transform prediction, and finally, probabilistic prediction methods.

Once successfully implemented and analyzed, we will choose the most effective method and further implement functionality to support not only our goal project of stock prediction, but develop an effective and efficient platform to be used with a wide array of time series based prediction applications, such as weather forecast predictions.

Each team member worked on one of the methods provided in the research paper, *Efficient Similarity Search over Future Stream Time Series* we read by Xiang Lian. Team members are helped to contribute and peer review and collaborate on each method of prediction. Once the a prediction method is chosen the team will divide the implementation up further for things such as further efficiency tuning, GUI, documentation to implement the solution into other industries/projects, testing, and so forth.

**3. Background**

* Related papers (or surveys for graduate teams)
* Software tools (DBMS, GUI, IDE, existing library, …)
* Required hardware
* Related programming skills (functions, Internet programming, object-oriented programming, distributed environment, etc.)

The paper that we read for this project was *Efficient Similarity Search over Future Stream Time Series* by Xiang Lian. This paper showed us different techniques on how to predict future data using past data. The type of hardware required would be a computer running an OS within the past ten years. Our recommendations are that you use a Windows computer that runs Windows 7 or later.

Programming skills are Object Oriented Programming, using Databases, et al.

Programming Language we are using is C++.

Software tools we are using Microsoft Visual Studios, CLion IDE, GitHub for a team project repository, etc.

**4. Problem Definition**

* Formal (mathematical) definitions of problems
* Challenges of tackling the problems
* A brief summary of general solutions in your project

The biggest difficulty we had was trying to understand how to implement the solutions given to us in the paper *Efficient Similarity Search over Future Stream Time Series* by Xiang Lian. It was hard converting the techniques into code that could be used to process the data and predict the next value. We found that translating mathematical and theoretical problems, equations, and solutions into tangible, function code to be the most challenging issue. Coupled with understanding how big data works, how to process and normalize the data, and just working with it in general, the project has been challenging.

The primary basis for our project has been the three different prediction methods outlined in the research paper of focus. Below we will outline general solutions to each prediction approach:

Polynomial Prediction: As the program processes information, it begins to learn and create a polynomial curve that serves as a historical data representation so that we can accurately predict data in the future. By using commonly used curves we can better approximate our historical data as well as future data. Once a curve is selected, we try to minimize the approximate error by measuring the squared Euclidean distance between the actual series and our prediction to get a more accurate prediction. A major benefit of this method is that this solution is easily adaptable for accuracy by changing the number of entries that are used to build our approximation, important for good performance for things such as online streaming.

Discrete Fourier Transform (DFT) Prediction: This method is approximates historical by using the frequency domain. Semantically, this approach is similar to the polynomial prediction model and follows a similar train of thought, however, it is based off of the frequency domain. If the historical and future data remains within the frequency domain, then the prediction will be accurate.

Probabilistic Prediction: The data stream is tokenized based on value range bands. These bands are used to create a trie where each node branches into one child node for each band. This trie keeps track of the occurrences of the patterns represented by the paths down its nodes. To predict the next value, the trie is traversed using the tree height - 2 latest tokens which leaves only one level of the trie. The predicted value is found by selecting the next node from this final level based on previous outcomes stored in the trie.

**5. The Proposed Techniques**

* Framework (problem settings)
* Details of major techniques (e.g., pruning methods in lemmas/theorems; illustrated with toy examples)
* Encoding or indexing of data
* Query processing algorithms (pseudo code) and query optimizations
* This section can be split into multiple sections if you have many contents to present

**6. Visual Applications**

* GUI design
* Design modules (with descriptions, figures, and/or flowcharts)

**7. Experimental Evaluation**

* Experimental settings
  + Descriptions of real/synthetic data sets
  + Competitors (baseline method, or existing techniques to compare with)
  + Parameter settings
  + Evaluation measures
* The performance report (pruning power, recall/precision/f-measure, CPU time, I/O cost, communication cost, index construction time/space, etc.)
* Screen captures

**8. Future Work**

* Possible project extensions

**9. References**

[1] Xiang Lian, Lei Chen. Efficient Similarity Search over Future Stream Time Series. In *XXX*, pages XXX-XXX, 2008.

[2] FirstName1 LastName1, FirstName2 LastName2, and FirstName3 LastName3. Journal paper title. In *XXX*, Vol. X, No. X, pages XXX-XXX, 19XX.