Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Haodong Zhao Date: Mar 31st. 2019

Please note that assignments in this class may be submitted to

www.turnitin.com, a web-based anti-plagiarism system, for an evaluation of their originality.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Reading review**

**Stream Data Mining: A Survey**

Data streams are large, continuous, and fast data elements of data.

Data mining is the process of discovering useful patterns in data and is often used in machine learning methods. Traditional data mining techniques cannot be used for data streams, because most of them need to scan data multiple times to extract information, so for streaming data, it is not realistic. Stream data has the following three characteristics:

* Data arrives continuously from the data stream
* Cannot make any assumptions about the ordering of data streams
* There is no limit to the length of the data stream

The basic method of data flow: From the perfect statistics and calculation methods, you can solve the problem of mining data flow by using the following methods:

1. Check a subset of the entire data set or transform the data to reduce the size: a summary of the entire data set or a subset of the input stream to be analyzed. The techniques used are sampling, cropping, sketching, summary data structures and aggregation. Specifically, the following methods are included:

* Sampling
* Load shedding
* Sketching
* Synopsis data structure: Sampling methods; Histograms; Wavelets; Sketches; Micro cluster-based summarization; Aggregation

1. Algorithms that make efficient use of time and space. Since the existing algorithms for data mining are modified from generating efficient algorithms for data streams, new algorithms are needed to solve the computational challenges of data streams:

* Approximation algorithm
* Sliding window
* Algorithm output granularity (AOG)

The need to understand the large amount of data generated in real time has led to the birth of a new data processing model - data stream processing. In this model, data arrives in the form of continuous, high-capacity, fast and time-varying streams, and these streams. Processing requires near real-time constraints. The following are several algorithms for extracting useful information from the data stream:

* Clustering
* Classification
* Association
* Frequency counting
* Time series analysis

The conclusion of this paper is that most current mining methods use a mining algorithm, and a small part of them even solve the drift problem. Due to the limited storage, the current technology produces approximate results.

The research of data flow is still in the early stage. The paper proposes several challenges and problems in the research of data flow mining. If it is solved, it will play an important role in the business world.