Real Time Mining

Prof. Pratiksha Rodge

***Abstract*—Data collection has recently become a wide spread phenomenon for web users. A key challenge is that how analysis the data in real time to know what kind of knowledgeable information we should be looking for. Evolving data streams are contributing to the growth of data created over the last few years. Data stream mining is the process of extracting knowledge structures from continuous, rapid data records in real time. Evolving data stream methods is becoming a key area, low-cost methodology for the goal of the prediction in real time. We discuss the current & future trends of mining evolving data streams.**

I. INTRODUCTION

The volumes of automatically generated data are constantly increasing. According to the digital universe study[1] , over 2.8 ZB of data were created & processed in 2012 , with a projected increase of 15 times by 2020. People carrying smart phones produce data, database transactions are being counted & stored, streams of data are extracted from virtual environments in the form of logs or users generated content. A significant part of such data is volatile, which means it needs to be analyzed in real time as it arrives.

Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from such applications as: twitter posts, email, blogging, sensor network, measurements in network monitoring & traffic management, call details records and others [2]

In the data stream model, data arrive at high speed & algorithms that process them must do so under very strict constraints of space and time. D data streams pose several challenges that milestones on the road to better algorithm designs for real world data stream mining systems. To verify if this challenges newly proposed solutions. First, algorithms must make use of limited resources (time & memory). Second, they must deal with data whose nature or distribution changes over the time.

We need to deal with resources in an efficient way. Green computing is major approach to improving energy usage efficiency & reducing energy usages. With optimized task scheduling, computers can complete tasks using less energy. It also reduces energy consumptions from supporting devices.

From green computing perspective, efficient task scheduling can be defined as either minimizing energy consumption with schedule length constraint or minimizing schedule length with energy consumption constraint. The objective of first problem is to use the least amount of energy to complete all task within given time frame. It is used mainly in real time processing environments. The second problem is to complete tasks as fast as possible under given energy limitations. Its objective is to use to energy efficiently and has usage in mobile computing etc. [3]

A main approach to green computing is based on algorithmic efficiency. In data stream mining, we are interested in three main dimensions:

- accuracy
- amount of space (computer memory)

Necessary

- the time required to learn from training

Examples and to predict

These dimensions are typically interred Pendent: adjusting the time and space used by an algorithm can innocence accuracy. By storing more pre-computed information, such as look up tables, an algorithm can run faster at the expense of space. An algorithm can also run faster by processing less information, either by stopping early or storing less, thus having less data to process. The more time an algorithm has, the more likely it is that accuracy can be increased.

II. STRUCTURE CLASSIFICATION

A new important and challenging task may be the structured pattern classification problem. Patterns are elements of (possibly infinite) sets endowed with a partial order relation. Examples of patterns are item sets, sequences, trees and graphs.

The structured pattern classification problem is defined as follows. A set of examples of the form (t; y) is given, where y is a discrete class label and t is a pattern. The goal is to produce from these examples a model ^y = f(t) that will predict the
Classes y of future pattern examples

Most standard classification methods can only deal with vector data, which is but one of many possible pattern structures. To apply them to other types of patterns, such as graphs, we can use the following approach: we convert the pattern classification problem into a vector classification learning task, transforming patterns into vectors of attributes. Each attribute denotes the presence or absence of particular sub patterns, and we create attributes for all frequent sub patterns, or for a subset of these.

As the number of frequent sub patterns may be very large, we may perform a feature selection process, selecting a subset of these frequent subpatterns, maintaining exactly or approximately the same information. The structured output classification problem is even more challenging and is defined as follows. A set of examples of the form
(t; y) is given, where t and y are patterns. The goal is to produce from these examples a pattern model ^y = f(t) that will predict the patterns y of future pattern examples. A way to deal with a structured output classification problem is to convert it to a multilabel classification problem, where the out-put pattern y is converted into a set of labels representing a subset of its frequents sub patterns. Therefore, data stream multi-label classification methods may over a solution to the structured output classification problem.

III. SOCIAL NETWORKS APPLICATIONS

Social network have become extremely popular in recent years because of numerous online social network such as Facebook, LinkedIn and MySpace.

In addition, many chat applications can also be modeled as social networks. Social networks provide a rich & flexible platform for performing the mining process with different kinds of data such as text, images, audio and video. Therefore, a tremendous amount of research has been performed in recent years on mining. In particular, it has been observed that the use of a combination of linkage structure and different kinds of data can be a very powerful tool for mining purposes. The one can combine the text in social networks with the linkage structures in order to implement more effective classification models. Other recent work uses the linkage structure in image data in order to perform more effective mining and search in information networks. Therefore, it is natural to explore whether sensor data processing can be tightly integrated with social network construction and analysis. Most of the afore-mentioned data types on a social network are static and changed slowly overtime. On the other hand, sensor collect vast amount of data which need to be stored and process in real time. There are a couple of important drivers for integrating sensor and social networks :

- One driver for integrating sensors and social networks is to allow the actors in the social networks to both publish their data subscribe to each other's data either directly, or indirectly after discovery of useful information from such data. The idea is such that collaborative sharing on social networks can increase real-time awareness of different users about each other, and provide unprecedented information and understanding about global behavior of different actors in the social network.
- A second driver for integrating sensors and social networks is to better understand or measure the aggregate behavior of self-selected communities or the external environment in which these communities function. Examples may include understanding environmental pollutions levels, or measuring obesity trends. Sensors in the possession of large numbers of individuals enable exploiting the crowed for massively distributed data collections and processing.

In recent years, sensors data collection techniques and services have been integrated into many kinds of social networks.

IV. NEW TECHNIQUES: HADOOP OR APACHE S4

Hadoop MapReduce is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster.
The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.
These reduce tasks use the output of the maps to obtain the final result of the job.
S4 (Simple Scalable Streaming System) is a general-purpose, distributed, scalable, partially fault-tolerant, pluggable platform that allows programmers to easily develop applications for processing continuous, unbounded streams of data.

Developed by **Yahoo** (which released the Yahoo's S4 paper) and then open sourced to Apache. Inspired by MapReduce and Actor model for computation. Basic components are:

- **Processing Element (PE)**: Basic computational unit which can send and receive messages called Events.
- **Processing Node (PN)**: The logical hosts to PEs
- **Adapter:** injects events into the S4 cluster and receives from it via the Communication Layer.

Ensemble learning classifiers are easier to scale and parallelize than single classifier methods. They are the first, most obvious, candidate methods to implement using parallel techniques.

V. CONCLUSIONS

We have discussed the challenges that in our opinion, mining evolving data streams will have to deal during the next years. We have outlined new areas for research. These include structured classification and associated application areas as

social networks. Our ability to handle many Exabyte of data across many application areas in the future will be crucially dependent on the existence of a rich variety of datasets, techniques and software frameworks. There is no doubt that data stream mining offers many challenges and equally many opportunities as the quantity of data generated in real time is going to continue growing.

REFERENCES

[1] J.Gantz & D.Reinsel. The digital universe in 2020 : Big data , bigger digital shadows & biggest growth in the fareast , December 2012.

[2] J.Gama knowledge discovery from data streams. Chapman & Hall/ CRC,2010.

[3] Evolutionary green computing solutions, for Distributed Cyber physical systems(chapter1). (Zahra Abbsi, Michael Jonas, Ayan Benerjee, Sandeep Gupta and Georgious Varsamopoulos.

[4] A. Bifet and E. Frank. Sentiment knowledge discovery in Twitter streaming data. In Proc 13th International Conference on Discovery Science, Canberra, Australia, pages 1{15. Springer, 2010.

[5] Bifet, G. Holmes, and B. Pfahringer. Moa-tweetreader: Real-time analysis in twitter streaming data. In Discovery Science, pages 46{60, 2011.

[6] A. Bifet, G. Holmes, B. Pfahringer, and E. Frank. Fast perceptron decision tree learning from evolving data streams. In PAKDD, 2010.

[7] J. Gama. Knowledge discovery from data streams. Chapman & Hall/CRC, 2010.

[8] B. Liu. Web data mining; Exploring hyperlinks, contents,and usage data. Springer, 2006.

[9] http://blog.andreamostosi.name/tag/apache-s4/