A Research Proposal Submitted by

S.Syedsafi
Assistant Professor,
Department of Computer Applications,
Meenakshi Chandrasekaran College of Arts and Science,
Karambayam, Pattukkottai.

Research Area **:** Data mining

Research Topic **:** Real Time Adaptive of Massive Complex Data Streams with single scan for Drift

Introduction:

Over past decade there has been a significant increase in the volume of online data. Extracting meaningful knowledge from this high volume data is considered as important aspect of research. It is very difficult to completely store full data, because of its perpetual nature. Therefore, analysis is needed while the "data is moving". This moving data is known as data stream and analyzing it without storing it completely is termed as data stream mining. Data streams are continuous flows of data. Examples of data streams include network traffic, sensor data; call center records and so on. Their sheer volume and speed pose a great challenge for the data mining community to mine them.

Problem Description:

In this research one of the biggest challenge is how to extract valuable information from the massive continuous data streams during single scanning and discuss the following challenges in data stream mining - making models simpler, protecting privacy and confidentiality, dealing with legacy systems, stream preprocessing, timing and availability of information, Evaluation of stream mining algorithms. Data streams also suffer from scarcity of labeled data since it is not possible to manually label all the data points in the stream. Each of these properties adds a challenge to data stream mining. We are interested in drift detection over data streams. Data streams are unbounded sequence of examples received at so high a rate that each one can be read at most once.

Proposed Methodology:

In this work we focus the concept drift and data drift, we use the sliding window technique for single scan that has been widely used during many researches on it and we use the basic data mining techniques like classification, regression tress with the support of output granularity technique. We used the VFDT (very fast decision tree learner) to find the decisions and where the drift occur and overcome the drift.

Research Tools:

(i) Apache spark
(ii) Apache Storm
(iii) Apache flink

- RapidMiner (formerly YALE (Yet another Learning Environment)): free open-source software for knowledge discovery, data mining, and machine learning also featuring data stream mining, learning time-varying concepts, and tracking drifting concept.
- EDDM (Early Drift Detection Method): free open-source implementation of drift detection methods in Weka (machine learning).
- MOA (Massive Online Analysis): free open-source software specific for mining data streams with concept drift. It contains a prequential evaluation method, the EDDM concept drift methods.

Conclusion:

In this research, we discussed research challenges for data streams, originating from real-world applications and we have found the drift where they occur and how to overcome the drift using our VFDT and machine learning algorithm.

References:

[1].https://www.researchgate.net/publication/227081327_Mining_ConceptDrifting_DataStreams.

[2]. Masud, M.M., Gao, J., Khan, L., Han, J., Thuraisingham, B.: A Multi-partition Multi-chunk Ensemble Technique to Classify Concept-Drifting Data Streams. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 363–375. Springer, Heidelberg (2009), doi:10.1107/9-78-3-642-01307-2_34.

[3].https://www.researchgate.net/publication/321750028_Concept_drift_in_Streaming_Data_Classification_Algorithms_Platforms_and_Issues.

[4]. http://bbrc.in/bbrc/streaming-data-classification-with-concept-drift/

[5]. BASSEVILLE, M., & NIKIFOROV, I. V. Detection of Abrupt Changes. Upper Saddle River, NJ, USA: Prentice-Hall. Retrieved marzo 11, 2016, from ftp://ftp.irisa.fr/local/as/mb/k11.pdf, 2012.