SHIV SHAKTI

International Journal of in Multidisciplinary and Academic Research (SSIJMAR)

Vol. 4, No. 3, June 2015 (ISSN 2278 – 5973)

The Study of Active Learning with Evolving Streaming Data

Shruti Jain

(Somany (PG) Institute of Technology and Management, Rewari)

E-mail:shrutijain.info@gmail.com, Mobile: 9899359979

Dr.Pushpender Sarao

Professor , SITM, Rewari

Impact Factor = 3.133 (Scientific Journal Impact Factor Value for 2012 by Inno Space Scientific Journal Impact Factor)

Global Impact Factor (2013)= 0.326 (By GIF)

Indexing:

Abstract

Active learning (Supervised Learning) is a branch of Artificial Intelligence. It can be defined as a setting where instead of using all the available data for training a computational model we actively select which data points to include into the training set. The main of this branch is to develop predictive models from the labeled data. These techniques can be used in real world scenarios like automatic organization, extraction of information from large pools of data.

The project area is "Computational Intelligence". There are various active learning strategies for streaming data. Streaming data is likely to change over time and can be defined as and rapid flow of data from a source like data coming from a sensor network or telephone records, web logs, multimedia data, retail transaction, railway ticket counter, surveillance systems etc. For example, let us consider a scenario of spam filters in Google mail. These filters scan the incoming data (Electronic mails- email) and classify them into a spam mail and non- spam mail on the basis of certain features like size of the email, number of addresses, etc. We use these features to train our models. Since the data changes over time, the models or classifiers need to be updated to classify accurately. To train or update a classifier we need features (length, number of addresses, words, etc.) and the true labels (is this message actually 'spam' or 'not'). When changes happen (e.g. spammers make their messages shorter) and we want to update our classifier/model accordingly we can extract the features automatically from the new incoming messages (e.g. count words), but we need to ask the user to tell the true labels (spam or not).

Key Words — Data mining, Active Learning, Data Stream, MOA, WEKA
Mouse Observers.

Introduction

Data stream can be defined as a continuous and rapid flow of data from a source like data coming from a sensor network or telephone records, web logs, multimedia data, retail transaction, railway ticket counter, surveillance systems, communication networks, Internet traffic, scientific and engineering experiments, remote sensors etc. (Khan 2010). These data sources are considered as continuous data generator at a very rapid rate. Data stream also show some other unique features like infinite length, concept drift, concept evolution and feature evolution. Because of these properties, handling of data mining of data streams has become more challenging (Khan 2010). There is an urge to develop new techniques which can handle this large amount of data in a reasonable time and extract the most beneficial information from it.

There are different challenges in data stream mining. As data streams have some unique features therefore some unique featured algorithms have to be developed which can answer these challenges. The challenges faced by data stream mining are classified into 5 categories by Kholghi and Keyvanpour in 2011 and can be summarized in the Table 1 shown below:

Issues	Challenges	Approaches
Memory Management	Fluctuated and irregular data arrival rate and variant data arrival rate over time	Summarizing techniques
Data pre processing	Quality of mining results and automation of pre processing techniques	Light-weight pre processing Techniques
Compact data structure	Limited memory size and large volume of data streams	Incremental maintaining of data structure, novel indexing, storage and querying techniques
Resource aware	Limited resources like storage and computation capabilities	AOG
Visualization of results	Problems in data analysis and quick decision making by user	One of the approach is Intelligent Monitoring

Table 1: Classification of Data Stream Mining Challenges (Kholghi and Keyvanpour 2011)

To develop a successful algorithm which could extract the information from data stream it is very important that the system should analyze the data in multidimensional, multi-level, single pass and online manner. Some solutions which have been proposed to develop such algorithms can be categorized into two broad categories (Gaber et.al. 2005) which are shown in Figure 1:

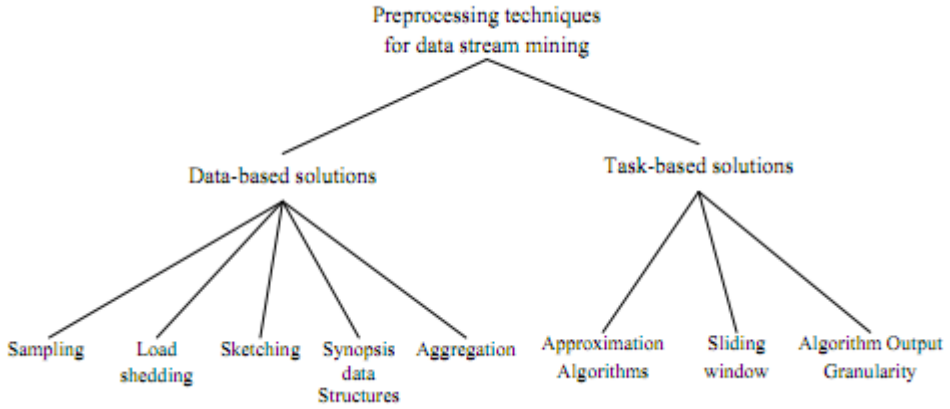

Figure 1: Classification of data stream pre-processing methods (Kholghi and Keyvanpour 2011)

The general process of Data stream mining is shown in the figure 2 below:

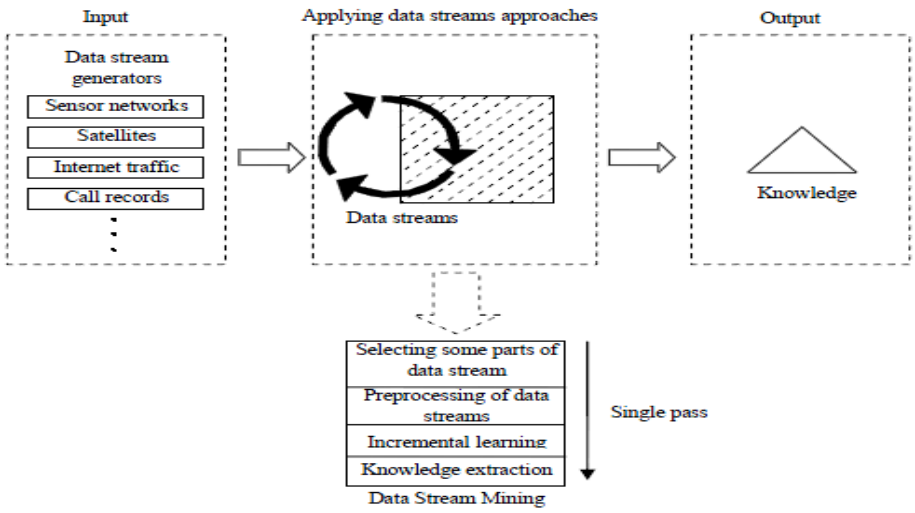

Figure 2: General Process of data streams mining (Kholghi and Keyvanpour 2011)

Active Learning

The learning can be accomplished in two ways: Active learning & Passive learning. The major difference between the active learning and passive learning is that in passive learning there is no query for seeking training examples to learn.

The active learning can be said that it is similar to semi-supervised learning as they will be requiring human experts while development of classifiers. The difference between them is that the queries are not done in semi supervised learning as labelling is done previously by human experts. In case of active learning the learner has to query the labels from the human experts. After obtaining the labelled instances these instances are added to the training instances.

Related Work

Indre Zliobaite published a research paper on "Active Learning with evolving streaming data" (Bifet et.al. 2011). This paper' main aim was to analyze the different existing active learning strategies and demonstrate the advantages and disadvantages of each strategy. The paper focused on learning with streaming data and it was said that to obtain true label may require excessive cost. For example if in a chemical plant a production is in process and in the end output quality is need to be predicted. The output quality depends on many constants like manual efficiency, chemical quality; machinery

working etc. and these conditions can change at any time. Therefore to predict the actual quality of output new experiments need to be conducted whenever these constants change, which can be very costly.

Zhu presented a research paper in 2007 on "Active learning from Data Streams". This paper proposed a new research topic i.e. active learning from data streams where volume of data continuously keeps on increasing. The main aim is to keep the labelling budget low by queering a small part of streaming data to develop a classifier which will be able to predict efficiently. As the volume of data increases to investigate it manually a manual labor need to be increased which can be quite costly.

In 2005, Frank and Witten published a book: "Data Mining- Practical Machine Learning Tool and Techniques" (Frank and Witten 2005). The main aim of this book was to provide the researchers an idea how the available data can be analyzed and understand and the WEKA toolkit is provided to understand it better by practically implementing it. This book also provides a step by step procedure and approach to analyze the data starting with a simple technique and if required to move to a sophisticated one.

Tools & Methodology

To start my research on Active learning with evolving streaming data I have tested the existing strategies using different data sets from Forest cover, Electricity and Airlines. For my research I have used Massive Online Analysis (MOA) software. This helped me in implementing active learning algorithm on evolving data streams. I thought of developing an open source framework, from where the required data sets can be obtained and these data sets can be used for testing and analyzing the existing active learning strategies. For this open source framework I developed Mouse observer software. This software is developed with an aim to obtain the datasets from the movement of mouse on the screen while performing certain activities. I was not able to record the mouse clicks on the screen as JAVA do not have the functionality to record the mouse clicks out of a form. I recorded the time stamped mouse movements and saved those movements in a CSV file format which gave me the coordinates along with the time stamp with a time interval of 1 millisecond while performing a task. After obtaining the CSV file I labelled the dataset obtained manually. I also tried to generate instances manually but as the data obtained was quite huge and was not feasible to generate it manually. Therefore I developed a code which was able to produce a file with instances. To test the strategies I decided to use Massive Online Analysis (MOA) software, but the problem was it only accepts ARFF file format as input. To convert the CSV file obtained from mouse observer software to ARFF file format I used WEKA software. It generated an ARFF file which I used to test the existing active learning strategies.

Software used

a) Massive Online Analysis (MOA)

MOA is related to Waikato Environment for Knowledge Analysis (WEKA) project, which is a framework used for data stream mining. It is a set of both online and offline tolls and algorithm required for evaluation of data sets. It is more related to classification problems, which will help in developing a model which can predict the class of unlabeled data (Bifet et.al. 2010).

b) WEKA (Waikato Environment for Knowledge Analysis)

It is a popular machine learning open source workbench, developed by University of Waikato and first implementation was done in 1997 (Bouckaert et.al. 2010). It is a Java based software which has a User Interface which helps user to feed the data in the software and has an ability to represent results in form of tables or curves. There are three major types of implemented schemes: Implemented schemes for classification, Implemented schemes for numeric prediction and Implemented "meta-schemes" (Bouckaert et.al. 2010). This provides a platform to the users which provide them with the variety of tools which allow them to focus on algorithm and take care of other tasks like reading the data from the files provided or filtering algorithms. I have used WEKA to convert the CSV files to ARFF.

Mouse Observer Framework

A new framework is developed for obtaining time stamped data sets. The framework is a JAVA based Mouse observer Software, which will be able to track the time stamped mouse movements while performing certain tasks. This software will provide a CSV file in which the coordinates of the mouse on the screen will be produced along with the time stamp. While designing this mouse observer software I have kept these tasks in mind:

- Data Collection Method
- Features & Instances Development Method
- Data Collection for framework testing

Application of Mouse Observer framework

1) **Personal strategy performance-** This mouse observer software can be used for obtaining data sets and for testing your own personalized strategy. For instance you would like to develop any strategy which requires some data sets for training purpose and testing the strategy performance, the required data sets can be obtained performing certain tasks and can make the data sets as complex as you require

2) **Existing strategy performance-** This mouse observer software can also be used for testing already existing strategies in any field (Ambady and Freeman 2010). For example you would like to test the psychology of a human being which can be judged by the human hand movement, and then this software could be used for recording the user hand movement while performing certain tasks(Dale et.a. 2011)

3) **Research purpose-** This software could be used in various fields for research purposes. Some of the research areas where this software is currently used are: Visual perception, auditory perception, social perception and cognition, neurophysiology, psycholinguistics, high-level cognition, embodied cognition, cognitive neuroscience, language acquisition, physiology, audiovisual and multimodal integration, synaesthesia, motor control, memory encoding and retrieval, tone deafness, moral cognition, hearing, visual attention, group processes, artificial intelligence.

4) **Personal evaluation-** This software could be used for personal judgment as well. Some people believe they work throughout the day on their system but are not able to distinguish what actual work they have done. For this we can record the mouse movement of the user throughout the day and then collect the data sets for it. These data sets can be used for extracting the information about the different works done by the user in the work time.

5) **Office evaluation** – This software could also be used in offices as well for recording the mouse movements of their employee and analyzing how employee spend most of their time in the office. Which work required most of the time and which work requires less time? On the basis of this workforce can be decided for the work. This work should be done in an ethical manner by taking consent from the employee.

6) **Market Analysis**- This software could be used in web applications for market analysis. The mouse movement of the users can be recorded and could be processed in order to judge how user is using the web service. What is most visited section of our website or which product is most liked by the user as maximum hits are on a certain product. By collecting this useful information market analysis can be done and in future this will become an essential part because experts cannot provide all the information from this huge leap of data.

Future work

In my Thesis I have proposed a Mouse observer framework which was used for analyzing active learning strategies. During the research I encountered various challenges and interested results which I have inspired me to provide some future work that can be done. This section introduces some of these ideas which I feel are fruitful directions for future work.

- Framework extension- Firstly due to scope of my project I was restricted to develop featured which fulfilled my requirements. In future this framework can be extended for several other purposes. One of the possible extensions could be that as I was not able to record the clicks of the mouse which can be an important feature that can be developed. One other feature could be that an interactive user interface can be developed as initially it is developed for developers only not for end users. The user interface could have features like running of the program, stopping the program, saving the .CSV file at desired location. Other feature could be that this framework can be integrated with active learning algorithms which can continuously process the data stream without saving the data which will help in saving memory. This will also help in testing in real world scenario.

- New features Development- In my project I have developed certain features which were required like Min X, Max X, Min Y, Max Y etc. New features could be developed which could result in development of better instances and better testing of strategies. New features that can be developed as per the user requirement like Mode, Median etc.

- Individual level or group level testing- The user interface of this framework can be developed as per the user requirement. If this framework is to be used for individual designed according to the requirement. If an individual level performance is to be tested then the user interface should be according to this requirement or else used in an office or an institution like place where group of people performance needs to be tested then it should be modified so that multiple user data can be recorded at the same time. The system should be capable of recording the data of multiple user as well as processing the same data with help of active learning algorithm in order to get the results in real time.

- Testing of other Strategies- My work scope was to test the classification strategies using Massive Online Analysis Software. This scope can be extended by testing clustering or classification by using different setting in the system. Like we can use Holdout technique instead of prequential or we can use other classifier instead of Naïve bayes. This will help in judging which type of strategy will be best suited for different tasks.

- Commercial purpose- This framework can be developed in order to develop an adaptive data analysis or market analysis or business model predicting software which are currently used for successful enterprises. This framework can help to gather all the information from internet by help of recording the mouse movements and clicks of the user while surfing the content on particular website. By collecting a large data it can be visualized what is the best for the enterprise success.

Conclusion

In today Information Intense world, supervised machine learning approach are taken in order to extract the exact information, updating the system from a large pool of continuous flowing data stream of infinite length. There are many existing active learning algorithms which are used for data mining and have their own limitations and advantages depends on certain situation like data sets, scenarios etc.

The main aim of developing this software was to collect datasets and analyze the strategies. I have analyzed the strategies. For analyzing I have used the datasets ARFF file and fed into MOA software and used the output to generate graphs. The graphs are plotted on which X-axis is the budget of labeling and Y- axis is the classification correct (%) i.e. accuracy. For Barclays SHL test datasets we can see that variable uncertainty and random variable uncertainty have performed the best. In the graph it was seen that they were overlapping but if we look at the datasets we can realize that the difference between the predicted accuracy was in decimals therefore they were overlapping. In case of Facebook random strategy performance was outstanding. In case of general datasets Random variable uncertainty is the best due to the changes occurred far from the decision boundary. In case of game played data sets we can see that random strategy is the lowest performer whereas all other strategies performed equally well when 15% of labeling budget is provided. However random variable uncertainty performed the best when provided with a lower budget value.

References

[1] Avila, J.C., Bifet, A., Bueno, R.M., Garcia, M.B. and Gavalda, R. Early Drift Detection Method. Available from: http://www.lsi.upc.edu/~abifet/EDDM.pdf. Accessed on [2nd October 2011].

[2] Ambady, N. and Freeman, J. 2010.MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. Behavior Research Methods, 42(1), 226-241. Available from: http://www.jbfreeman.net/pubs/Freeman_BRM.pdf. Accessed on [15th November 2011].

[3] Barto, A. and Jonsson, A. Active Learning of Dynamic Bayesian Networks in Markov decision Processes. 1-11. Available from: http://www-anw.cs.umass.edu/pubs/2007/jonsson_b_SARA07.pdf. Accessed on [20th September 2011].

[4] Bifet, A. 2009. Adaptive Learning and Mining of Data Streams and Frequent Patterns. Doctoral Thesis to the Department de Llenguatges i Sistemes Informatics Universitat Politecnica de Catalunya. Available from: http://www.lsi.upc.edu/~abifet/Thesis.pdf. Accessed on [15th August 2011].

[5] Bifet, A. and Kirkby, R. 2009. Data Stream Mining A Practical Approach. Centre for Open Software Innovation. Available from: http://www.cs.waikato.ac.nz/~abifet/MOA/StreamMining.pdf?&lang=en_us&output=json&session-id=058ab3def69d098a1a22632bb6ef037e. Accessed on [20th September 2011]

[4] Castillo, G., Gama, J., Medas, P. and Rodrigues, P. Learning with Drift Detection. Available from: http://www.liaad.up.pt/~jgama/Papers/sbia04.pdf. Accessed on [3rd January 2012].

[5] Dale, R., Farmer, T.A. and Freeman, J.B. 2011. Hand in motion reveals mind in motion. Frontiers in Psychology, 2, 1-6. Available from: http://www.jbfreeman.net/pubs/2011_FrontiersInCognition.pdf. Accessed on [15th November 2011].

[6] Kholghi, M. and Keyvanpour, M.2011. An Analytical Framework For Data Stream Mining Techniques Based On Challenges and Requirements. International Journal of Engineering Science and Technology, 3(3), 2507- 2513. Available from: http://arxiv.org/ftp/arxiv/papers/1105/1105.1950.pdf. Accessed on [20th August 2011].

[7] Lin, X., Shi, Y., Zhang, P. and Zhu, X. 2007. Active Learning from Data Streams. Seventh IEEE International Conference on Data Mining, 757-762. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4470323. Accessed on [16th September 2011].

[8] Mitchell, T.M. 1997. Does Machine Learning Really Work?. AI Magazine, 18(3), 11-20. Available from: http://www.aaai.org/ojs/index.php/aimagazine/article/view/1303/1204. Accessed on [12th August 2011].

[9] Settles, B. 2008. Curious Machines: Active Learning with Structured Instances. Doctor of Philosophy (Computer Sciences) at the UNIVERSITY OF WISCONSIN–MADISON. Available from: http://www.cs.cmu.edu/~bsettles/pub/settles.thesis.pdf. Accessed on [30th August 2011]

[10]Tatbul, N. 2009. Load Shedding. 45-49. Available from: http://www.inf.ethz.ch/personal/tatbul/publications/load_shedding_EDBS.pdf. Accessed on [23rd August 2011].

[11] Vinzamuri, B. 2011. Embedding Robust Data Mining Models in Active Learning. A Thesis Submitted to International Institute of Information Technology.1-76. Available from: http://www.web2py.iiit.ac.in. Accessed on [23rd September 2011]