Mobile Data Stream Mining: From Algorithms to Applications

Shonali Krishnaswamy
Institute for Infocomm Research (I^{2}R), Singapore and Faculty of IT, Monash University, Australia
email: spkrishna@i2r.a-star.edu.sg, Shonali.Krishnaswamy@monash.edu

Joao Gama
Laboratory of Artificial Intelligence and Decision Support
University of Porto
Portugal
email: jgama@fep.up.pt

Mohamed Medhat Gaber
School of Computing
University of Portsmouth
United Kingdom
email:mohamed.gaber@port.ac.uk

***Abstract*—This paper presents an overview of the current state-of-the-art in mobile data stream mining. This area of mobile data stream mining is significant for a number of new application domains such as mobile crowd sensing and mobile activity recognition. The paper presents the strategies and techniques for adaptation that are essential in order to perform real-time, continuous data mining on mobile devices. We present an overview of the algorithms research in this area. Finally, we discuss the key toolkits, systems and applications of mobile data stream mining.**

Keywords-Mobile Data Mining;Data Stream Mining; Ubiquitous Data Mining

I. INTRODUCTION

Mobile devices are increasingly becoming the central computing and communication device in people's lives. Devices today are equipped with a growing number of sophisticated embedded sensors such as an accelerometer, digital compass, gyroscope, GPS, microphone, light intensity sensor, and camera. This creates the opportunity to develop applications that leverage on the sensing capability of these mobile devices, as well as data from other sensors such as bio/body sensors. Data from mobile users/devices is becoming increasingly important for numerous applications including urban modeling, transportation, and more recently for mobile crowd-sensing for citizen journalism and real-time traffic routing. While significant efforts are being focused towards the analysis of mobile user data, a key challenge that needs to be addressed in order to realize the full-potential of mobile user analytics is to address the scalability issues of real-time data collection. By scalability, we refer to both the challenges of data transmission from a large number of users, as well as the issues of energy consumed on individual devices as a result of that transmission.

Mobile data stream mining is a key technology for real-time analysis of data streams generated on-board the phone itself, for both data generated by sensors on the phone and/or in close proximity to the phone. The significant advantages that mobile data stream mining provides over traditional strategies for leveraging the phone as a "transmission device" for sensor data, are as follows: reduce the amount of data transmitted from the phone to servers/the cloud, as well as reduce the energy/battery usage on the phone due to transmission of sensor data. Mobile data stream mining is particularly significant for applications that need real-time analysis of continuous data streams such as such as mobile crowd sensing, mobile activity recognition, intelligent transportation systems, mobile healthcare, and so on.

Mobile data stream mining techniques typically focus on adapting data stream mining algorithms to be operational in the context of mobile devices. As such, the aim is to enable data stream mining to be performed in a manner such that it is congruent with the limited computational resources, screen real-estate, and energy considerations of the mobile device. In this context, strategies for adaptation of data stream mining algorithms to enable their effective and scalable operation on mobile devices has been a significant research focus in this area.

In addition to such adaptation techniques for mobile data stream mining, numerous "light-weight" mining algorithms for varies types of analysis such as clustering, classification, concept drift detection, change detection, and frequent items analysis have been developed. Finally, a number of applications/systems/toolkits for mobile data stream mining have also been presented in the literature. There is also an emerging focus on visualization techniques and strategies for mobile data mining. In summary, while the last few years have seen continuous evolution of research in this area, mobile data mining has now finally come of age with significant wider interest in this domain.

The goal of the Advanced Seminar is to present the state-of-the-art in mobile and ubiquitous data stream processing. and discuss open research problems, issues, and challenges in this area. We will present the fundamental techniques for data stream analysis such as change detection, clustering, classification, frequent patterns, and time series analysis from distributed data streams. We will present the critical factors that need to be considered in order to develop and deploy data stream mining in mobile/ubiquitous environments including the need for adaptation and context/situation-aware reasoning. We will then present state-of-the-art algorithms for mobile data stream mining. The seminar will also present the Open Mobile Miner (OMM) toolkit [26] for rapid deployment of mobile data stream mining and real-world application/case studies and demonstrations in the areas of Mobile Crowdsensing, Intelligent Transportation Systems, Patient Monitoring and Habitat Monitoring to stimulate the real need for this growing research field. The web demos of the Open Mobile Miner are accessible at: http://www.mobilemining.monash.edu.au. Finally the seminar will be concluded with open issues and future directions. The seminar also provides a substantial list of mobile data stream mining resources. The key learning objectives of the seminar are to enable understanding of the motivations, rationale and challenges of the emerging important area of data stream mining in mobile and ubiquitous environments, and in-depth knowledge of techniques for mobile/ubiquitous data stream mining and identification of the key research and application challenges in this domain. This key distinguishing feature of this seminar when compared to previous editions/versions of this seminar are as follows: the seminar will present the role, techniques, and benefits of mobile data stream mining for new and previously not explored applications such as *mobile crowd sensing,* and *mobile activity recognition,* and recent research in mobile data stream mining techniques.

This paper is organized as follows: Section 2 presents and overview of the research contributions in *adaptation strategies* for mobile data stream mining. Section 3 presents and overview of *analysis algorithms* and *visualization challenges*. Section 4 presents an overview of the key applications and systems for mobile data stream mining that have been developed till date. Section 5 concludes this paper and outlines directions for future research.

II. Adaptation Strategies For Enabling Mobile Data Stream Mining

There has been much work on developing adaptation strategies for data stream mining algorithms that vary the accuracy levels according to available computational resource levels and incoming data rates. Adaptation strategies have been shown to significantly enhance the longevity of continuous real-time processing of data mining in mobile environments. Adaptation can enable, if not guarantee, the continuity, cost-efficiency and consistency of a mobile/ubiquitous data stream mining application.

As discussed in [18], even efficient data stream mining algorithms such as Very Fast K-Means (VFKM) can cause device crashes on mobile devices when used without awareness to context such as variations in data rates and computational resource availability. Thus, there are primarily three different adaptation strategies that have been proposed for mobile data stream mining:

- *Resource-Aware Adaptation:* The key focus of the adaptation strategies has been to factor in varying levels of computational resources on the mobile device and use this as a continuous control parameter to adapt the behavior of the stream mining algorithms that are operational on the mobile device. Generic granularity-based adaptation [14, 16, 19] techniques that can be used with any data stream mining technique running on a resource-constrained device have been developed. This approach facilitates adaptation of data stream mining algorithms to varying data rates based on available computational resources in mobile devices by innovative strategies to perform knowledge integration, controlling the rate of learning and varying the accuracy levels of the discovered patterns. The granularity-based adaptation approach has three different variations. AOG (Algorithm Output Granularity) provides adaptability by adjusting the algorithm output rate (e.g. the number of clusters) according to the availability of memory, the remaining time to fill the available memory and the data stream rate. AOG uses a time threshold that is the time required to generate the results before any incremental integration according to some accuracy measure. AIG (Algorithm Input Granularity) is a process that adapts the data rates feeding into the algorithm according to the battery charge by using the methods of sampling, load shedding and synopsis. When the battery charge becomes critical, the adaptation process changes the sampling rate to reflect the current data rate according to the previous consumption pattern in the most recent time frame. APG (Algorithm Processing Granularity) performs adaptation of the processing settings of the algorithm with respect to the CPU usage. For example, in a a stream clustering analysis, it uses a novel approach termed Randomization Assignment. When the algorithm needs to make the micro-cluster assignment for a new data point, this method enables examining only a fraction of the current micro-clusters using the randomization factor. If the randomization factor is equal to 1, all micro-clusters need to be examined. However if the usage pattern of CPU increases this factor will be decreased; thereby reducing the number of micro-clusters examined and consequently the CPU consumption.
- *Situation-Aware Adaptation:* Resource-aware adaptation aims to vary the algorithm accuracy for efficient use of resources but it focuses only on resources such as memory and battery. With situation-aware adaptation, the focus of adaptation is with respect to the application's accuracy requirements changing based on the occurring situations [23]. When the current situation warrants for less frequent monitoring/analysis, the algorithm accuracy can be moderately decreased to preserve resources. On the other hand, in critical situations where there is a need for closer monitoring, it is important to increase the accuracy even if the resource availability is scarce. By situations we mean real-life situations such 'fire_threat', 'heat_stroke', 'driving' and many others. There are

certain situations in which applications do not need high accuracy (e.g. 'healthy' situation). When such situations occur, data mining algorithm settings can be adjusted to produce results with the low accuracy and thus consume less resource. On the other hand, when critical situations (e.g. 'fire_threat') occur, the application requires a higher level of accuracy and the settings of the mining algorithm can be adjusted to achieve this according to the need. Situation-aware adaptation is typically driven by leveraging context-aware and situation-aware engines customized for specific mobile data mining applications.

- *Hybrid Adaptation Strategies:* Hybrid adaptation strategies aim to integrate resource-aware strategies and situation-aware strategies to control the adaptation process [23].

Having briefly reviewed the state-of-the-art adaptation strategies for mobile data mining, we now present an overview of algorithms for analysis and visualization.

III. Algorithms For Mobile Data Stream Mining

There exists a large body of work in distributed data stream mining [6, 7, 9, 10, 11, 12, 13] to perform a range of stream analysis such *clustering, detecting concept drift, classification, frequent items, novelty detection, and change detection*. Distributed data stream mining has several applications ranging from intelligent transportation, sustainability, email classification, astronomy, and healthcare [1, 2, 3, 8]. Mobile data stream mining algorithms typically aim to perform the same data stream mining on-board a mobile device. In general, these algorithms are typically one-pass techniques that leverage sliding windows since they operate over streaming data. There are two strategies for development of mobile data stream mining algorithms. Basically, as explained a mobile data stream mining algorithm operates with variability of performance according to resource, situation and other constraints. This variable performance is typically effected by specifying an upper and lower bound for accuracy levels that are acceptable. Thus, there are two ways in which algorithms are developed according to this principal. Firstly, the algorithms themselves are developed as light-weight techniques that have in-built strategies for adaptation such as [14, 15]:

- Clusterers
    - Light-Weight Clustering
    - RA-Cluster and DRA-Cluster
- Change Detection
    - CHANGE-DETECT
- Classifiers
    - Light-Weight Class (LWC)
- Frequent Items and Associations
    - LWF (Light-Weight Frequent Items)
    - HiCoRE (Highly Correlated Energy-Efficient Rules)

Alternatively, the algorithms are general stream mining techniques that can be treated as a "black-box" (i.e. no modifications to the internal processing), but adaptation is leveraged by identifying an algorithm's potential control parameters. For example in [18], the Very Fast K-Means algorithm which is a variation of k-means for streaming data, was treated as a black box but the epsilon- and delta- error thresholds were identified and experimentally established as a good control parameters to specify the variable behavior of the algorithms and drive the drive the adaptation process. In [18], the well-known time series representation technique *Symbolic ApproXimation* (SAX) [21] was adapted to mobile devices for analyzing ECG data by leveraging the *word* and *segment* size parameters of the SAX representation. Thus, it was shown that varying word and segment sizes according to resource levels could achieve trade-offs of battery consumption and application longevity vis-à-vis accuracy levels. More recently in [25], the well-known naïve-bayes classifier has been adapted to learn from streaming data for mobile activity recognition.

While much research has focused on developing novel ways of analyzing data in real-time on mobile devices, there have not been specific techniques developed for visualization of the analysis. This can be primarily attributed to the fact that it is only now that even analysis is possible on mobile devices. Visualization of the results from analysis in real-time is therefore an emerging challenge - but one that is vital in order to effectively leverage the benefits of mobile data mining to enable real-time decision making by mobile users. The key challenges to visualization of mobile data mining are [22]: 1) The small screen real-estate of mobile phones/PDAs and therefore the need to effectively use this limited screen space to present useful and easy-to-understand information; 2) The need to dynamically perform computations relating to visualization; and 3) The need to rapidly change the visualization so that they capture and reflect accurately the current state of the underlying analysis process. In this context, we have also developed an adaptive approach for real-time cluster visualization for mobile data mining.

IV. Mobile Data Mining Applications, Systems and Toolkits

One of the earliest work in the area of mobile data stream mining was by Kargupta et al. [9] who developed a client/server data stream mining system: MobiMine which focuses on data stream mining applications for stock market data. Kargupta et al. also developed the Vehicle Data Stream Mining System (VEDAS) system and its commercial version [4] for fleet management. MOLEC [27] is a mobile cardiac monitoring system that aims to analyze ECG signals to identify a range of anomalies and arrhythmias using decision trees that were built off-line. More recently, there have applications for mobile crowdsensing [24] and mobile activity recognition [25] leveraging mobile data stream mining.

The Open Mobile Miner [26] is an integrated toolkit for performing data stream mining on mobile devices which has a range of algorithms to facilitate different types of applications. The primary features of the development of this toolkit are as follows:

1. Enable easy deployment of mobile data mining applications on a range of mobile devices;
2. Provide a platform for evaluation of new and existing mobile data stream mining techniques by the research community;
3. Encapsulate extensibility of the toolkit by being an open source resource;
4. Facilitation integration of new and existing data stream mining algorithms into the toolkit that may or may not have adaptation mechanisms incorporated;
5. Interface with a range of input sources for data streams including Bluetooth-enabled sensors, previously recorded data, distributed data, and synthetic data (i.e. data stream generation for evaluation purposes);
6. Allow flexible, application specific visualizations to be developed.

V. Conclusions And Future Directions

With the significant interest in mobile users and applications, driven by the ever increasing sophistication and capabilities of today's mobile device, mobile data mining is emerging as a key technology. We have presented an overview of the current state-of-the-art in terms of algorithms, adaptation strategies, and applications/systems/toolkits for mobile data stream mining. There are many key areas for future work including developing new application case studies that leverage mobile data mining such as in the network gaming area, as well as to develop activity recognition technologies that turn the phone to your "personal protection" device, as well as large scale gathering of data through mobile devices to sense urban phenomena and events. Furthermore, there is always need for more sophisticated analysis and visualization techniques. Finally, a key challenge is to take the next step from analysis to providing real-time decision making for mobile users.

References

[1] Kanishka Bhaduri, Kamalika Das, Kirk D. Borne, Chris Giannella, Tushar Mahule, Hillol Kargupta: Scalable, asynchronous, distributed eigen monitoring of astronomy data streams. Statistical Analysis and Data Mining 4(3): 336-352 (2011)

[2] Hillol Kargupta, Joao Gama, Wei Fan: The next generation of transportation systems, greenhouse emissions, and data mining. KDD 2010: 1209-1212

[3] Hillol Kargupta, Kakali Sarkar, Michael Gilligan: MineFleet®: an overview of a widely adopted distributed vehicle performance data mining system. KDD 2010: 37-46

[4] Kargupta, H., Bhargava, R., Liu, K., Powers, M., Blair, P., Bushra, S., Dull, J., Sarkar, K., Klein, M., Vasa, M. and Handy, D. 2004. VEDAS: A Mobile and Distributed Data Stream Mining System for Real-Time Vehicle Monitoring. Proc. of the SIAM DM Conference.

[5] H. Kargupta, B. Park, S. Pittie, L. Liu, D. Kushraj, K. Sarkar, "MobiMine: Monitoring the Stock Market from a PDA", SIGKDD Explorations, January, 2002, Vol. 3, No. 2.

[6] João Gama, Pedro Pereira Rodrigues, Luís M. B. Lopes: Clustering distributed sensor data streams using local processing and reduced communication. Intell. Data Anal. 15(1): 3-28 (2011)

[7] Pedro Pereira Rodrigues, João Gama, João Araújo, Luís M. B. Lopes: L2GClust: local-to-global clustering of stream sources. SAC 2011: 1006-1011

[8] José M. Carmona-Cejudo, Manuel Baena-García, Rafael Morales Bueno, João Gama, Albert Bifet: Using GNUsmail to Compare Data Stream Mining Methods for On-line Email Classification. Journal of Machine Learning Research - Proceedings Track 17: 12-18 (2011)

[9] João Gama, Raquel Sebastião, Pedro Pereira Rodrigues: Issues in evaluation of stream learning algorithms. KDD 2009: 329-338

[10] João Gama, Pedro Pereira Rodrigues: An Overview on Mining Data Streams. Foundations of Computational Intelligence (6) 2009: 29-45

[11] João Gama, Ricardo Rocha, Pedro Medas: Accurate decision trees for mining high-speed data streams. KDD 2003: 523-528

[12] Eduardo J. Spinosa, André Carlos Ponce Leon Ferreira de Carvalho, João Gama: Cluster-based novel concept detection in data streams applied to intrusion detection in computer networks. SAC 2008: 976-980

[13] Domingos, P. and Hulten, G. 2001. A General Method for Scaling Up Machine Learning Algorithms and Its Applications to Clustering. Proceedings of the 18th Int. Conf. on Machine Learning.

[14] Gaber, M, M., Krishnaswamy, S., and Zaslavsky, A. 2005. On-board Mining of Data Streams in Sensor Networks, A Book Chapter in Advanced Methods of Knowledge Discovery from Complex Data, (Eds.) S. Badhyopadhyay, U. Maulik, L. Holder and D. Cook, Springer.

[15]. M. M.Gaber, P. S. Yu, "Detection and Classification of Changes in Evolving Data Streams", International Journal of Information Technology & Decision Making, Vol. 5, No. 4, World Scientific Publishing Company, 2006.

[16] Gaber, M, M., Zaslavsky, A. and Krishnaswamy, S. 2004. A Cost-Efficient Model for Ubiquitous Data Stream Mining. Proceedings of the 10th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia Italy, July 4-9

[17] M. M. Gaber, A. Zaslavsky, S. Krishnaswamy, "Mining Data Streams: A Review", ACM SIGMOD Record, 34, 1 (June 2005).

[18] Shah R., Krishnaswamy S., and Gaber M. M. 2005. Resource-Aware Very Fast K-Means for Ubiquitous Data Stream Mining Proceedings of 2nd Int. Wshop on KD in Data Streams, ECML/PKDD 2005.

[19] Gaber M.., Yu P.S., A Holistic Approach for Resource-aware Adaptive Data Stream Mining, Journal of New Generation Computing, 25(1) pp. 95-115, 2006.

[20] Salim, F. D., Loke, S. W., Rakotonirainy, A., Srinivasan, B., Krishnaswamy, S., 2007, Collision pattern modeling and real-time collision detection at road intersections, Proc of the 2007 IEEE Intelligent Transportation Systems Conference, USA, pp. 16.

[21] Hossein Tayebi, Shonali Krishnaswamy, Agustinus Borgy Waluyo, Abhijat Sinha, Mohamed Medhat Gaber: RA-SAX: Resource-Aware Symbolic Aggregate Approximation for Mobile ECG Analysis. Mobile Data Management (1) 2011: 289-290

[22] Brett Gillick, Hasnain AlTaiar, Shonali Krishnaswamy, Jonathan Liono, Nicholas Nicoloudis, Abhijat Sinha, Arkady B. Zaslavsky, Mohamed Medhat Gaber: Clutter-Adaptive Visualization for Mobile Data Mining. ICDM Demo Paper 2010: 1381-1384

[23] Pari Delir Haghighi, Arkady B. Zaslavsky, Shonali Krishnaswamy, Mohamed Medhat Gaber, Seng Wai Loke: Context-aware adaptive data stream mining. Intell. Data Anal. 13(3): 423-434 (2009)

[24] Sherchan, W., Jayaraman, P.P., Krishnaswamy, S., Zaslavsky, A., Loke, S., Sinha, A.: Using on-the-move mining for mobile crowd-sensing. In: (to appear) Proceedings of MDM 2012 (July 2012).

[25] Gomes, J, P., Krishnaswamy, S., Gaber, M, M., Pedro, S., and Menasalvas, E.: MARS: A Personalised Mobile Activity Recognition Systems In: (to appear) Proceedings of MDM 2012 (July 2012).

[26] Krishnaswamy, S., Gaber, M, M., Harbach, M., Hugues, C., Sinha, A., Gillick, B., Delir Haghighi, P.,and Zaslavsky, A., (2009), Open Mobile Miner: A Toolkit for Mobile Data Stream Mining, ACM Knowledge Discovery in Databases (ACM KDD 2009), Demo and Short Paper, Paris, June-July 2009.

[27] Alfredo Goñi, Alfredo Burgos, Lacramioara Dranca, Jimena Rodríguez, Arantza Illarramendi, Jesús Bermúdez: Architecture, cost-model and customization of real-time monitoring systems based on mobile biological sensor data-streams. Computer Methods and Programs in Biomedicine 96(2): 141-157 (2009).