Anomalous Pattern Detection Using Context Aware Ubiquitous Data Mining

Zahoor ur Rehman[1], Muhammad Shahbaz[1], Muhammad Shaheen[1], Sajid Mehmood[1], Syed Athar Masood[2]

[1]Department of Computer Science and Engineering, University of Engineering and Technology, Lahore-54890, Pakistan
[2]Department of Engineering Management, NUST, College of E & ME, Rawalpindi, Pakistan
M.shahbaz@uet.edu.pk, Muhammad.Shahbaz@gmail.com

Abstract— Due to the developments in technology number of applications emerged that produce huge amount of data in the form of streams. Dealing with this and extracting useful information from that data is a real challenge. In this paper, we have developed an architecture that can be used to manage data streaming applications and can extract useful information from that data in online fashion. To achieve mining results online, different phases in our model are parallelized. In this model we have also introduced the concept of context-awareness to improve performance of the proposed architectural model. In this model information from heterogeneous sources is gathered, fuse that information, and generate real-time results. These real-time results can be beneficial in different application area like web usage mining, online monitoring, fraud detection, network security, telecommunication calls monitoring, network monitoring and security, etc. To fulfill the objectives of this research, we incorporate lightweight online mining algorithms to extract useful but hidden information from the data gathered. Contextual information is exploited to detect anomalous behaviors. In this paper we have designed an architectural model to extract frequent patterns in the streaming data.
[Zahoor ur Rehman, Muhammad Shahbaz, Muhammad Shaheen, Sajid Mehmood, Syed Athar Masood. **Anomalous Pattern Detection Using Context Aware Ubiquitous Data Mining.** Life Science Journal. 2012;9(3):6-12] (ISSN:1097-8135). http://www.lifesciencesite.com. 2

Keywords- stream mining; context-aware; anomalous pattern mining; ubiquitous data mining

I. INTRODUCTION

Developments in technology led us to the scenario where we have to deal with very voluminous data that is continuously arriving without limits. For example data being received from satellites, web-clicks, sensor network, stock exchange, shopping at big malls, telecommunication calls, etc. is very huge and continuous without limits. The data which is continuous, unbounded and without limits is called data stream. This data is fast changing, massive in nature and potentially infinite. To deal with this data, there is need to devise some architectural model and efficient techniques to extract useful information in timely fashion. Traditional data mining models and algorithms mainly focused on modeling [1], regression [1], clustering[5,6], classification [[2][3][4] and querying [5][6]. Unlike traditional data mining, stream mining need to extract information from continuously varying, huge and dynamic data therefore some complex challenges are need to tackle in streaming environment. Some inherent characteristics of stream mining are:

- data streams are in order and that cannot be controlled
- Data is coming continuously at some fixed or variable rate
- Concept drift might occur after some interval
- Storage of data being received is not possible or simply not required
- Only one scan of data is possible because of scarce computational resource
- Analysis of data should be instantly available when requested
- Error rate in outputs should be kept as small as possible

Due to these constraints, motivation for stream mining model and architectural design emerged. In other words, development of efficient algorithms and model for stream mining is direly needed. Many efforts have been made to develop efficient algorithms for stream mining [5] [7][8][9][10][11][12].

According to the stream mining model [13], there is a need to collect only sample data from the sensing devices to use their resources more effectively. Once the data is received, perform data mining operation on that with efficient algorithms and then store those results in the knowledge-base to improve learning ability of the system. Similarly, stream mining research can be categorized into *landmark window, sliding window and damped window* [14]. In the landmark window, transactions from some specific time marker to the present are considered to generate mining results. While in sliding window model, two types are time-sensitive and transaction-sensitive sliding windows. In time-sensitive sliding window, time slot expires while in transaction-sensitive window transactions are expired. The third type of sliding window model is

damped-window, deals with the concept-drift, meaning that recent-most transactions are more important than the older ones.

A large number of models and algorithms have been proposed for mining streaming data, many of those do not give importance to the recent transactions on the older ones. Similarly, most of those algorithms produce approximate results on streaming data [15][16][10][17]. To overcome this limitation of approximate results, a new CP-Tree based sliding window algorithm for streaming data has been proposed [18]. Although, sliding window based frequent pattern mining algorithm is efficient enough but it cannot be used in ubiquitous environment.

To overcome the problems in the domain for streaming data, we propose architecture for sensor stream mining. We have also incorporated the context-awareness to improve effectiveness and reliability of results. This framework will analyze the real-time contextual information of a particular environment although such information will of course impede the computational and communicative powers of ubiquitous devices. In the event of anomalous behavior, alerts and alarms caution security personnel to take necessary actions in a timely manner. These alerts help minimize the chances of disrupting the smooth flow as desired.

Rest of the paper is organized as:

Section 2 is dedicated to provide an overview of Ubiquitous Data Mining (UDM) and its potential for implementation in real-world scenarios. This emergent technology can play a vital role in predicting behaviors and intentions of an agent in an environment under observation. Applications and implementations of sensor streams and ubiquitous data mining are described in section 3 of this paper. Section 4 describes an example scenario for integration of ubiquitous data mining and contextual information for the prediction of human intentions. In Section 5, we have presented the design of the system, its sub-modules and schematic flow of data. Finally, Section 6 concludes and describes implications for future research.

II. PRELIMINARIES

The process of extracting interesting, hidden and useful patterns is called data mining. With the help of data mining algorithms and techniques, we can identify clusters, classification of data and extraction of association rule can be performed. Statistical analysis and data mining tasks utilize considerable computational, memory and communicational resources. Therefore it is required and important to pay attention to the computational and communicative powers of systems. In conventional data mining systems, data is normally gathered at some central location in the form of a data warehouse to perform data analysis by incorporating statistical techniques and machine learning algorithms [13]. The emergence of wireless and mobile devices has introduced a new dimension and enabled access to a large amount of data located at distributed and remote locations in the form of continuous streams. Ubiquitous Data Mining (UDM) is the process of analyzing data and information being received directly from the environment or retrieved from remote systems on mobile devices like cell phones, PDA or touch-pad [19]. Ubiquitous computing and data mining enables users to monitor, retrieve and analyze data from distributed and heterogeneous devices like sensors and mobiles [20][21][22][23].

As computational power of wireless and portable devices is continuously increasing so we are able to perform tasks that need high resources in terms of memory and computing power. Currently available portable devices can perform data mining operation on the bases of spatial and temporal constraints [24][22] . The basic techniques for analyzing data and extracting hidden patterns are usually derived from traditional data mining, statistical techniques and machine learning methodologies. However, the existing techniques cannot be used straightway due to resource limitations of these portable devices. There is a need to cater traditional data mining algorithms to fit in the ubiquitous environment.

Ubiquitous data mining normally need to perform pattern extraction task in online fashion. Similarly, the data in continuously arriving at high rate therefor very fast algorithms are required to be developed. These fast techniques need to compromise a bit on accuracy as compared with traditional data mining algorithms [25][26]. There is need of ubiquitous data mining (UDM) software to extract hidden but useful patterns from streaming data. Objective of this UDM module is to analyze data in real-time and then transfer that information to central location for further processing. Due to this functionality, usage of bandwidth can be improved with decreased traffic towards the central server. Personalization and aggregation tasks will also be performed locally.

Data-intensive applications are starting to appear on PDAs and cell phones such as cell-phone-based patient monitoring systems [27][28], vehicles and driving monitoring systems [29], and wireless security systems. In the near future, some of the applications to be exercised include monitoring and analyzing data in embedded devices for smart applications, and the use of Nano-scale devices for on-board monitoring. Thus, it is necessary to provide support for such applications in terms of advanced data analysis and prediction. Such applications pose various challenges and problems in order to analyze data and apply data mining techniques, which, in this domain, include:

- Efficient single-pass algorithms need to be

developed to analyze and extract useful information from streaming data in ubiquitous environments;

- How to visualize results on mini screens of smart phones and PDA's; This is the major area need to be worked so that extracted results and patterns can be presentable more effectively on these mini devices;
- Communication bandwidth in sensor network and cell phones is normally low as compared with the other computing devices like desktop or laptop computers. Cannels used by normal computing devices for communication is fast enough though there is need to optimize utilization of bandwidth in that case too [22][30].

A lot of research is has been carried out to optimize battery usage in mobile devices and similarly researcher are also focusing to develop battery modules that can store more energy to provide longer backup time. In spite of all these efforts and research, limitation in terms of battery is still there. It also direly needed to use battery resources in optimal or sub-optimal way to increase network overall lifetime. Battery resource is still a barrier to extract useful information in ubiquitous data mining [31][32].

III. LITERATURE REVIEW

Advances in technology and software enable us to gather huge amounts of data from various sources. To exploit the full potential of this data, we are in need to perform data analysis to extract useful patterns from data streams. A dramatic decrease in the cost of data storage technology has enabled us to store huge volumes of data streams. Extracting data from those data generating and storage devices for analysis and fulfilling users' queries is not efficient enough and has become an interesting area of research. To overcome this problem, researchers from artificial intelligence, machine learning and data mining community focused their attention to this specific problem. As a result technology of intelligent data analysis emerged to extract hidden and useful information by using automated or semi-automated means from data streams. Initially intelligent data analysis is incorporated in already existing statistical techniques to extract interesting patterns from historical and static data. Gradually with advances in computational devices and increase in database size, efficient and scalable machine learning and artificial intelligence algorithms were developed. This improved data analysis tasks both in terms of accuracy and reliability. In order to address the problem of very large databases, statistical analysis and machine learning techniques have been incorporated [33].

Developments in data gathering and wirelessly sending devices in last few years are exponentially high [34]. Voluminous data is being transmitted from satellites, sensors, web clicks, stock market, etc. and it is a sheer challenge to store, manipulate and analyze such huge quantities of data. Similarly, in time critical applications, analysis of data is important only in some specific time intervals. Storing all data might be of no use in future [35]. Data coming continuously in the form of streams needs to be analyzed as soon as it arrives at the processing unit. This online analysis is a challenge with the constraints of available storage, computational powers and communicational bandwidths. To overcome these problems some efforts have been made in last few years. To address the challenges of streaming data mining some systems, models and algorithms have been developed [35].

Incorporating contextual information in ubiquitous data mining is very useful. It is used to predict car accidents before it occur. Similarly the same information is used to warn drivers in smart cars. Sources of data for these alarms are both on-board sensing devices and sensors from the environment [36] [29]. This is a good breakthrough to develop such cars that use context-awareness to reduce number of accidents in smart cars. Analysis of test data reflects that most of the accidents are due to tiredness of drivers or the environmental conditions being logged as contextual information.

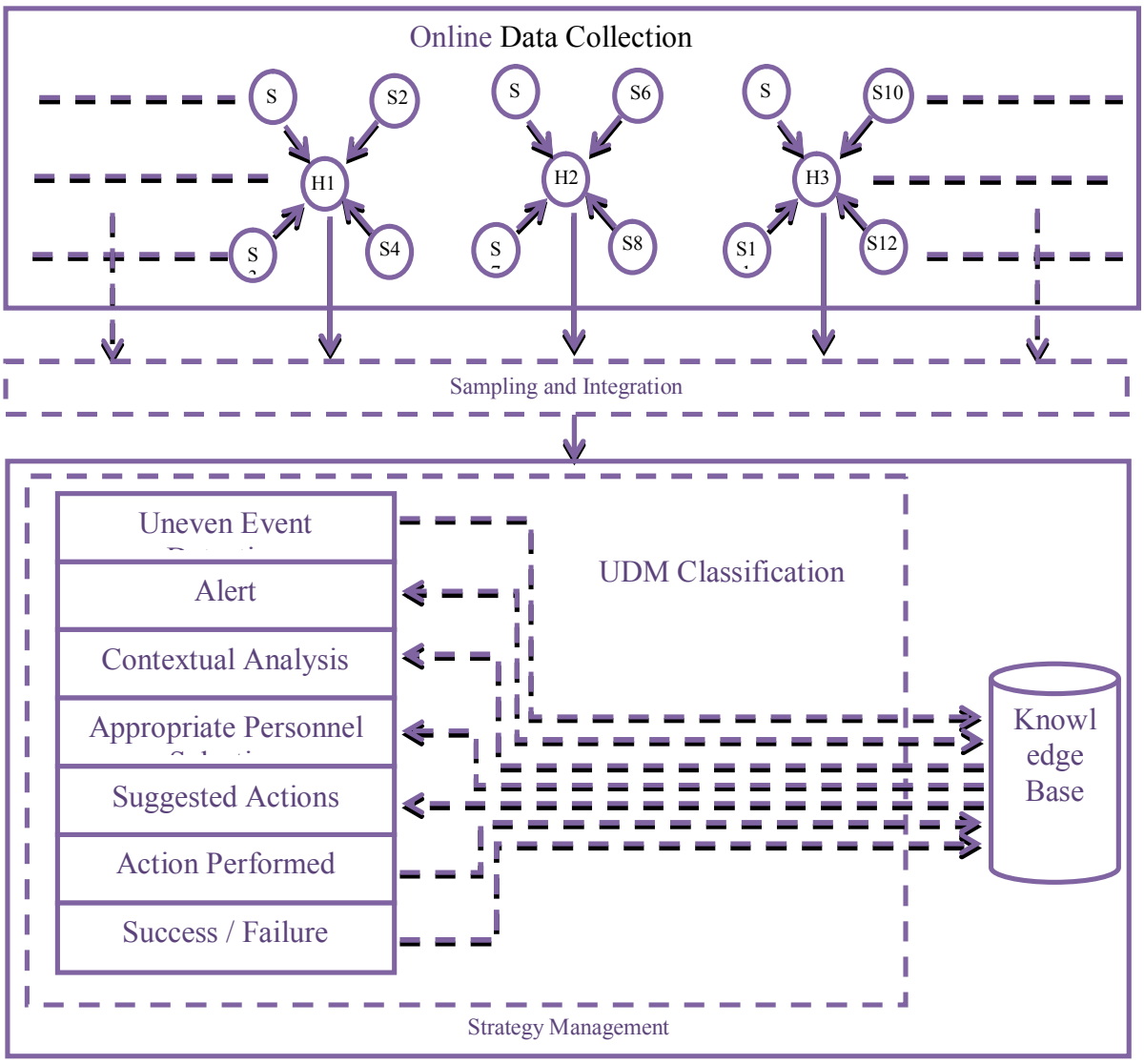
Monitoring patients who are suffering critical diseases like cardiac problem or asthma is one of the most important application area of UDM [27][28]. Using patient cell phone, home-based sensor network and then connecting with the hospital information system provide good opportunity for the better treatment of critical patients. Connecting patient information with the contextual circumstances, early treatment can be started if needed.

Link analysis techniques are used to detect human behavior in real-time. To achieve this feature extraction using unsupervised method is used and discrete human activities are extracted [37].

In this research we have developed an architectural framework to extraction pattern using portable devices. To improve performance, we have also incorporated context-awareness. Combining technological developments and data gathering techniques, our proposed architecture is expected to perform much better.

IV. EXAMPLE SCENARIO

In this section, we explain our proposed architecture with example application of security and sustainability of society. There is need to define some basic things before explaining the example application.



Situation: The state of affair of an entity is called its situation i.e. what is happening inside or outside of an entity and how is the entity at the moment.

Awareness: Knowing exactly what is happening inside or outside of an entity under observation is called awareness. Awareness gives us some values about the entity and its surrounding.

Context-awareness: It means that getting information about the entity and its surroundings and then utilizing that information on the bases of context. It includes information about the location, person and the environment in which the person is currently located, among other things.

Security sensitive areas like bus stops, railway stations, airports, some events, etc. can be monitored using this UDM architecture. Environmental sensing devices can collect necessary information about the persons in the area under observation. This live information is fused with the existing knowledgebase to extract useful patterns. Security personnel are provided with the UDM enabled smart-phones to

analyze situation in online fashion. Unusual behavior of a person or vehicles or other moving things can be predicted in timely fashion to minimize chances of any devastation. Once the uneven behavior is detected, an alert will be generated on the smart-phone of the security personnel. These alerts will also consider context-awareness and will be generated or forwarded to the closer 'k' security personnel. These alerts will not only alarm security personnel but also suggest some preventive measures to be taken to avoid any miss happening. Along with the contextual information, history data can also be very useful for predicting intention of a person. Similarly, importance

of place and size of crowd also reflect useful information whether that place can be at risk or not. Uneven movements in one place cannot be similar to the other at some other place. So context-awareness will be playing vital role in these security applications.

IV. PROPOSED METHODOLOGY

The proposed architecture is novel as it integrates contextual information and performs ubiquitous data mining in online fashion. In this architecture, contextual information can be different depending upon the application area. In human security case, context may refer to historical data about criminal activities, personal profiles, current circumstances, place under observation, high profile official presence, etc. while in buying trend case context might be weather conditions, day of week, seasonal information, etc. These factors with ubiquitous data mining have not been used collectively to extract anomalous patterns from the streaming data as per our knowledge.

Fusing person's information like his/her bio data, area of residence, previous living records, etc. will be very useful to predict behavior of each individual with great accuracy. Current information in vicinity under observation is gathered using electronic devices like wireless sensors and cameras.

We also define usual patterns in each area so that unusual patterns can be distinguished easily. Here we apply Naïve Bayes technique to predict the occurrences of uneven events.

In real-time, we analyze our data on PDAs or handheld devices available from security personnel in order to assess the risk of criminal activities. Using past patterns of criminal activity, historic data, personal profiles, and current contextual information, the model detects odd events. For example, if a person had been involved in some criminal activity, has been identified as suspicious, or his or her current contextual information clusters him in a criminal category, the model alerts security personnel to take necessary countermeasures.

Initial training of the model can be performed using historical data about criminal activities. In case of unavailability of such data, synthetic data can be used to achieve this training objective. Once the training has been completed and system becomes live with the actual environment then the database will be automatically growing and subsequently knowledgebase will be building for future use. One of the important research goals is to find uneven events and crime patterns in this specific case therefore our proposed model will be mining that information. Knowledgebase built from these real world scenarios will be used for prediction of uneven events well before time and suggest necessary actions required to be taken.

We use probability theory to find out the chances of an unusual event so that the severity can be calculated. If severity level is some predefined threshold, then necessary alert will be generated on contextual bases and forwarded to the staff at duty for appropriate actions. Similarly, if calculated severity is below the threshold but there are some detected uneven patterns then those specific persons will be kept under keen observation to avoid any loss. The success and failure rates of the alerts generated, as well as the resulting predictions will be analyzed and the main repository database will be updated. This process of continuous learning improves system performance and future decisions become more accurate and reliable.

To save energy and communication cost, data sampling is performed. Only that data is recorded which is predicted as malicious and propagated towards main repository database for futuristic use. It will decrease network traffic and utilization of communication channel is expected to optimize. This mechanism increase overall lifetime as well as performance of network.

To convert this proposed model into practical shape, several issues need consideration and catering those factors is necessary to obtain expected results as well as accuracy. Few factors are:

- *Data:* multidimensional data from heterogeneous sensors arrive in the form on continuous streams at a high rate. To deal with such a substantial amount of data in real-time is a complex challenge. Moreover, classifying this continuous stream of data by incorporating a predictive model requires access to historic data about criminal events. Normally, data about a location where criminal action has occurred is available, as is information about the person who committed the crime. However, personal information about history of the criminal is often unavailable. Similarly, the movements and actions performed before committing a crime are not available as there is no existing system to record such information. To obtain such data, a simulator is ideal at initial stages; later, continuous learning processes enable models to become more realistic.
- *Analysis:* As lightweight algorithms have already been developed which perform well in resource constraint environment but there is a need to optimize those algorithms and cater those so that data obtained from integration of heterogeneous streams can be analyzed and process in effective manner.
- *Human Rights and Legal Issues:* Unfortunately, public organizations, constitutional rights, and basic rules of independence are core hindrances in the implementation of this model. To deploy

this model, constitutional shelter as well as acceptance by citizens is required, given that such a system will be perceived an invasion of their privacy due to modern ethics, which do not allow viewing an individual's personal details without sufficient and appealing proof.

V. CONCLUSION AND FUTURE WORK

We have proposed and architecture based on information and communication technology to extract useful and uneven patterns from the streaming data. Our model uses state-of-the-art technological devices like smart phone, wireless sensors and imaging devices. This paper presented a novel approach to detect uneven patterns and suggestion for the actions to be taken to avoid from devastating results of those unusual patterns. It is expected that this proposed architecture will bring considerable improvement in finding uneven events in the streaming data environment. It will also enable researcher to perform ubiquitous data mining in streaming data to get results in real-time for better performance.

As a next step, we have developed a prototype of this system and it can detect anomalous patterns in the streaming data. It also generate set of suggested actions in case of unusual event has been detected. Once we are fully successful in prototype, development for real environment will be initiated on the bases of application areas.

REFERENCES

[1] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang, "Multi-dimensional regression analysis of time-series data streams," in *Proceedings of the 28th international conference on Very Large Data Bases*, 2002, pp. 323–334.

[2] H. Wang, W. Fan, P. S. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers," 2003, p. 226.

[3] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "On demand classification of data streams," 2004, p. 503.

[4] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," 2001, pp. 97-106.

[5] S. Babu and J. Widom, "Continuous queries over data streams," *ACM SIGMOD Record*, vol. 30, no. 3, p. 109, Sep. 2001.

[6] M. Greenwald and S. Khanna, "Space-efficient online computation of quantile summaries," 2001, pp. 58-66.

[7] Ruoming Jin and G. Agrawal, "An Algorithm for In-Core Frequent Itemset Mining on Streaming Data," pp. 210-217.

[8] H.-fu Li, S.-yin Lee, and M.-kwan Shan, "An Efficient Algorithm for Mining Frequent Itemsets over the entire History of Data Streams," in *In Proc. of First International Workshop on Knowledge Discovery in Data Streams*, 2004.

[9] Hua-Fu Li, Suh-Yin Lee, and Man-Kwan Shan, "Online Mining (Recently) Maximal Frequent Itemsets over Data Streams," pp. 11-18.

[10] C.-hsiang Lin, D.-ying Chiu, and Y.-hung Wu, "Mining frequent itemsets from data streams with a time-sensitive sliding window," in *In SDM*, 2005.

[11] G. S. Manku and R. Motwani, "Approximate Frequency Counts over Data Streams," in *VLDB*, 2002, pp. 346–357.

[12] W.-guang Teng, M.-syan Chen, and P. S. Yu, "A regression-based temporal pattern mining scheme for data streams," in *In VLDB*, 2003, pp. 93–104.

[13] Z. ur Rehman, M. Shahbaz, M. Shaheen, and A. Guergachi, "Situation-Awareness and Sensor Stream Mining for Sustainable Human Life," 2009, pp. 616-620.

[14] Y. Zhu and D. Shasha, "StatStream: statistical monitoring of thousands of data streams in real time," in *Proceedings of the 28th international conference on Very Large Data Bases*, 2002, pp. 358–369.

[15] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu, *Mining Frequent Patterns in Data Streams at Multiple Time Granularities*. 2002.

[16] K.-F. Jea, C.-W. Li, and T.-P. Chang, "An Efficient Approximate Approach to Mining Frequent Itemsets over High Speed Transactional Data Streams," 2008, pp. 275-280.

[17] J. Yu, Z. Chong, H. Lu, Z. Zhang, and A. Zhou, "A false negative approach to mining frequent itemsets from high speed transactional data streams," *Information Sciences*, vol. 176, no. 14, pp. 1986-2015, Jul. 2006.

[18] S. K. Tanbeer, C. F. Ahmed, B.-S. Jeong, and Y.-K. Lee, "Sliding window-based frequent pattern mining over data streams," *Information Sciences*, vol. 179, no. 22, pp. 3843-3865, Nov. 2009.

[19] R. Shah, S. Krishnaswamy, and M. M. Gaber, "Resource-aware very fast K-Means for ubiquitous data stream mining," in *In Proceedings of 2nd International Workshop on Knowledge Discovery in Data Streams, to be held in conjunction with the 16th European Conference on Machine Learning (ECML'05) and the 9th European Conference on the Principals and Practice of Knowledge Disc*, 2005.

[20] R. Chen, K. Sivakumar, and H. Kargupta, "An approach to online Bayesian learning from multiple data streams," in *In Proceedings of Workshop on Mobile and Distributed Data Mining, PKDD '01*, 2001, pp. 31–45.

[21] International Conference on Enterprise Information Systems, M. Piattini, J. Filipe, and J.

Braz, "Enterprise information systems IV," Dordrecht; Boston; London, 2003.
[22] H. Kargupta, B.-H. Park, S. Pittie, L. Liu, D. Kushraj, and K. Sarkar, "MobiMine," *ACM SIGKDD Explorations Newsletter*, vol. 3, no. 2, p. 37, Jan. 2002.
[23] V. Ganti, J. Gehrke, and R. Ramakrishnan, "Mining data streams under block evolution," *ACM SIGKDD Explorations Newsletter*, vol. 3, no. 2, p. 1, Jan. 2002.
[24] S. Krishnaswamy, *Delivering Distributed Data Mining E-Services*. .
[25] M. Krogmann, M. Heidrich, D. Bichler, D. Barisic, and G. Stromberg, "Reliable, Real-Time Routing in Wireless Sensor and Actuator Networks," *ISRN Communications and Networking*, vol. 2011, pp. 1-8, 2011.
[26] N. Jiang and L. Gruenwald, "Research issues in data stream association rule mining," *ACM SIGMOD Record*, vol. 35, no. 1, pp. 14-19, Mar. 2006.
[27] S. Kumar, K. Kambhatla, F. Hu, M. Lifson, and Y. Xiao, "Ubiquitous Computing for Remote Cardiac Patient Monitoring: A Survey," *International Journal of Telemedicine and Applications*, vol. 2008, pp. 1-19, 2008.
[28] C. Orwat, A. Graefe, and T. Faulwasser, "Towards pervasive computing in health care – A literature review," *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, p. 26, 2008.
[29] S. Krishnaswamy, S. W. Loke, A. Rakotonirainy, O. Horovitz, and M. M. Gaber, "Towards Situation-awareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application," in *Proceedings of Conference on Intelligent Vehicles and Road Infrastructure*, 2005, pp. 16–17.
[30] H. Kargupta and Byung-Hoon Park, "A fourier spectrum-based approach to represent decision trees for mining data streams in mobile environments," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 216-229, Feb. 2004.
[31] D. Puccinelli and M. Haenggi, "Wireless sensor networks: applications and challenges of ubiquitous sensing," *IEEE Circuits and Systems Magazine*, vol. 5, no. 3, pp. 19-31, 2005.
[32] L. Q. Zhuang, K. M. Goh, and J. B. Zhang, "The wireless sensor networks for factory automation: Issues and challenges," 2007, pp. 141-148.
[33] D. H. Dejong, *Medical informatics : knowledge management and data mining in biomedicine*. [S.l.]: Springer, 2005.
[34] M. M. Gaber, "Ubiquitous data stream mining." .
[35] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams," *ACM SIGMOD Record*, vol. 34, no. 2, p. 18, Jun. 2005.
[36] F. D. Salim, S. Krishnaswamy, S. W. Loke, and A. Rakotonirainy, "Context-Aware Ubiquitous Data Mining Based Agent Model for Intersection Safety," in *Proc. of EUC 2005 Workshops*, 2005.
[37] J. Hunter and M. Colley, "Feature Extraction from Sensor Data Streams for Real-Time Human Behaviour Recognition," in *Knowledge Discovery in Databases: PKDD 2007*, vol. 4702, J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, and A. Skowron, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 115-126.

2/20/12