A Fast and Scalable Tool towards Collaborative Data Mining in Mobile Computing Environments

G.Jagatheeshkumar [#1], N.Anbazhagan [#2], Dr.S.Selva brunda [#3]
[1,2] *Asst Professor, Department of Information Technology.*
KSG College of Arts and Science,
Bharathiar University, Coimbatore. Tamilnadu, India.
[3]*. Professor & Head, Department of CSE*
Cheran College of Engineering,Karur,
Anna University,Chennai.Tamilnadu, India.
[1] jagatheeshkumaar@gmail.com,[2] anbuksg@gmail.com
[3] *brundhaselva@yahoo.com*

Abstract: **Systems that construct classifiers are one of the commonly used tools in data mining. new term describing collaborative mining of streaming data in mobile and distributed computing environments. Proposed using mobile software agents. The proposes a new architecture that prototyped for realizing the significant applications in this area. With the continuous advances in handheld mobile devices including smart phones, PDAs (Personal Digital Assistants) and smart sensors, there is an unpreceded opportunity to perform significantly useful data analysis.**

INTRODUCTION

The systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. Pocket Data Mining PDM is our new term describing collaborative mining of streaming data in mobile and distributed computing environments. Graphs are an increasingly important data source, with such important graphs as the Internet and the Web. The networking community has used different measures of node "importance" to build a hierarchy of the Internet.

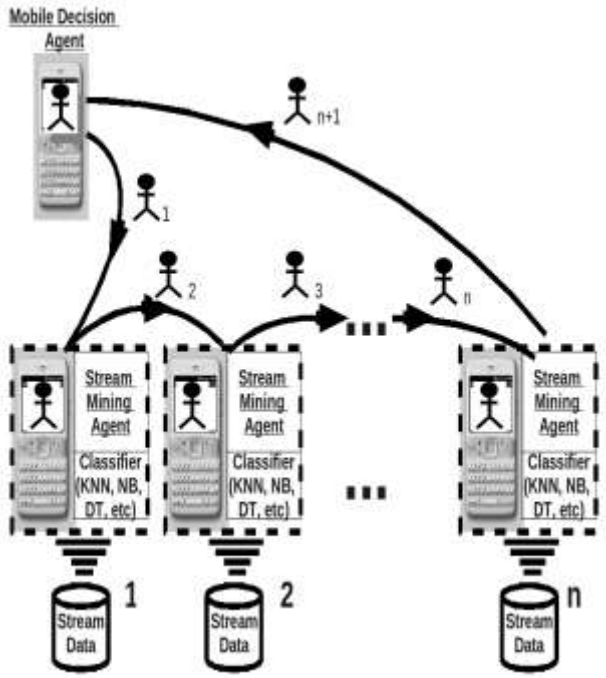
Another source of graph data that has been studied are citation graphs are an increasingly important data source, with such important graphs as the Internet and the Web. There are many more examples of graphs which contain interesting information for dates raining purposes. For example, the telephone calling records from a long distance carrier can be viewed as a graph, and by mining the graph we may help identify fraudulent behaviour or marketing opportunities.

This is same method to mining the data from the data center using the Networks. Wireless communication among these devices using Bluetooth and WiFi technologies has opened the door wide for collaborative mining among the mobile devices within the same range that are running data mining techniques targeting the same application. in this application for several reasons. Most importantly the autonomic intelligent behaviour of the agent technology has been the driving force for using it in this application.

Tasks in an ad hoc computing environment. This can be realized with the help of several established areas of study including:

- Data stream mining
- Mobile software ag
- ent ents
- Embedded programming.

It is possible to Computing three graph properties pertaining to the connectivity of neighborhood structure of the graph.



- Graph Similarity
- Sub graph Similarity
- Vertex Importance

This generic scenario, when applied to collaborative data mining, would include mobile software agents of different types. These types could be identified as follows:

(Mobile) agent miners (AM): these agents are either distributed over the network when the mining task is initiated or are already located on the mobile device.

Mobile agent resource discoverers (MRD): these agents are used to explore the available computational resources, processing techniques, and data sources.

Mobile agent decision makers (MADM): these agents roam the network consulting the mobile agent miners to collaborate in reaching the final decision.

II. RELATED WORK

Related work in this area includes systems developed by Kargupta el al in for mobile data mining for mobile brokering and road safety, and by Pirttikangas et al for context-aware health club. Brief descriptions of these systems will follow.Kargupta et al have developed the firstubiquitous data stream mining system termed MobiMine. It is a client/server PDA-based distributed data mining application for financial data streams. The system prototype has been developed using a single data source and multiple mobile clients; however the system is designed to handle multiple data sources. The server functionalities in the proposed system are data collection from different financial web sites and storage, selection of active stocks using common statistics methods,and applying online data mining techniques to the stock data.

The client functionalities are portfolio management using a mobile micro-database to store portfolio data and information about user's preferences, and construction of the WatchList and this is the first point of interaction between the client and the server. The server computes the most active stocks in the market, and the client in turn selects a subset of this list to construct the personalized WatchList according to an optimisation module. The second point of interaction between the client and the server is that the server performs online data mining and then transforms the results using Fourier transformation and finally sends this to the client. The client in turn visualises the results on the PDA screen. It is worth pointing out that the data mining process in MobiMine has been performed at the server side given the resource constraints of a mobile device. With the increase need for onboard data mining in resourceconstrained computing environments, Kargupta et al have developed Vehicle Data Stream Mining System (VEDAS). It is a ubiquitous data stream mining system that allows continuous monitoring and pattern extraction from data streams generated on-board a moving vehicle. The mining component is located on the PDA. VEDAS uses online incremental clustering for modeling of driving behaviour. A commercial version of VEDAS termed as MineFleet has been successfully deployed. Pirttikangas et al have implemented a mobile agentbased ubiquitous data mining for a context-aware health club for cyclists. The system is called Genie of the Net. The process starts by collecting information from sensors and databases in order to recognize the needed information for the specific application. This information includes user's context and other needed information collected by mobile agents. The main scenario for the health club system is that the user has a plan for an exercise. All the needed information about the health such as heart rate is recorded during the exercise. This information is analysed using data mining techniques to advise the user after each exercise.Other related work includes the large body of data stream mining algorithms. Key techniques and approaches in the area are discussed in and more recently in the tutorial presented by Gama et al in. Addressing the resource consints of small computational devices like smart phones and Personal Digital Assistants PDAs has been reported in work conducted by Gaber el al in
[4], [6], [5], [7]. The approach taken in this body of work has been termed as Granularity-based approach. It adapts the data mining algorithm to adjust the resource consumption pattern according to availability of resources. Notably, successful applications of the approach in road safety and healthcare have been reported in.

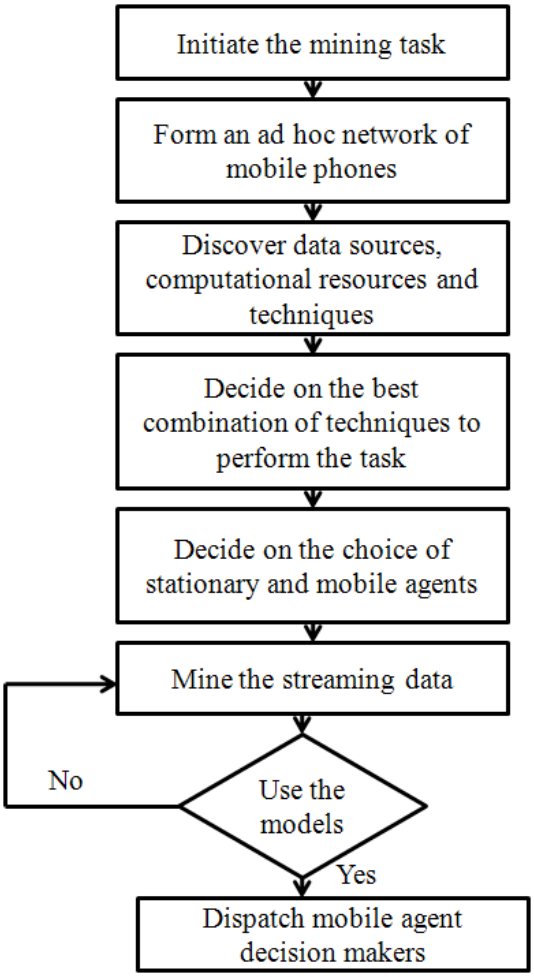
III. PDM ARCHITECTURE

The architecture of our PDM framework is illustrated in Figure 1. From this point onwards in the paper, we shall use the terms PDM architecture and PDM framework interchangeably.

As the figure shows, the data stream mining process runs onboard the users' smart mobile phones. As the data streams in, the model is continuously updated to cope with the possible concept drift of the streaming environments. The process of stream mining is carried out using an Agent Miner, denoted as AM. AMs are distributed at the beginning of initiating the mining task. Some of these miners could be stationary and some others could be mobile. Stationary agents are instructed by the task initiator to mine the streaming data to the mobile device without making any hops. However, the mobile agents could travel to one or more nodes in order to perform the mining task. The choice of using stationary or mobile agent relies on the nature of the task and the number of nodes involved in the processing. Typically, AMs are data stream classification techniques. But the use of other techniques is also possible according to the required task.

If at any point in time, a user decides to use the models built using the different AMs on all the mobile phones to collaborate in inding the class label of a set of unlabeled

instances, a Mobile Agent Decision Maker MADM is fired to visit the nodes consulting the models about the local class label. While these agents are visiting the different nodes, it may decide to terminate its itinerary given that clearly there is a dominant class. This clearly makes the agent framework the suitable technology for this task.A simple flow chart of the process of collaborative data mining using our PDM architecture is given in Our PDM framework raises a number of research issues that are important to be addressed to optimize the task. These issues are currently being investigated by our research team.



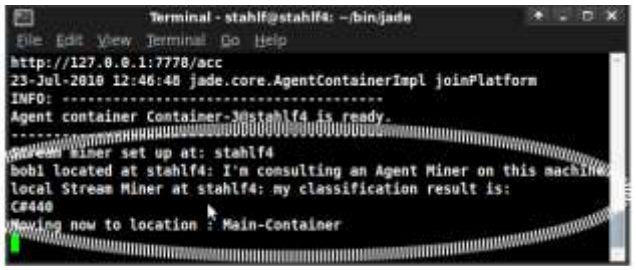
The following is a list of these issues.

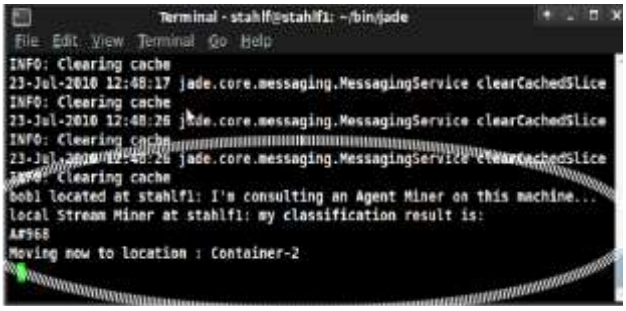
_ The number of AMs that are decided by the task initiator.In an ad hoc environment, the number of participants may vary. Thus, it is important to involve the number of AMs that cover the largest number of attributes and data instances. For example, if the number of mobile phones in a setting is 5, and 2 of these share the same instances and attributes and run the same data mining technique.

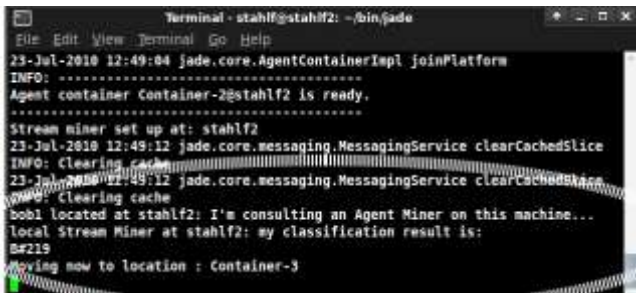
IV. EVALUATION OF THE PDM FRAMEWORK

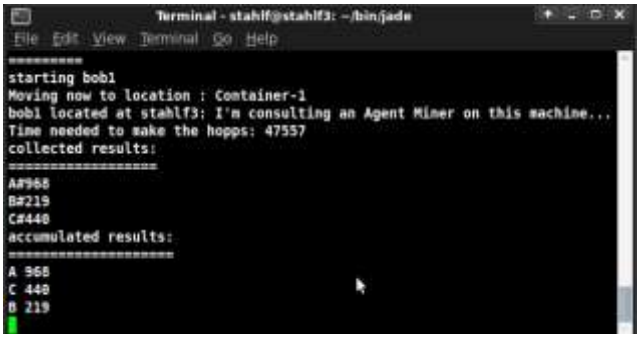
A prototype of the PDM framework as described in Section III has been implemented and empirically evaluated in a LAN. For the implementation the well known JADE framework has been used with the reasoning that there exist a version of JADE, JADE-LEAP (Java Agent Development Environment- Lightweight Extensible Agent Platform), that is designed for the implementation of agents on mobile devices and can be retrieved from the JADE project website as an 'add on' [23]. As JADE works on standard PCs as well as on mobile devices it was possible to develop and test the first prototype of the PDM framework on the LAN described below. The LAN consisted of four computers interconnected with a network switch. Three of the computers (computers A, B and C) have a CUP with a 2.8 GHz clock-speed and 1 GB of memory, the fourth computer (computers D) has a CPU with 2.20 GHz clock-speed and 500 MB of memory. The switch used is a standard CISCO Systems switch of the catalyst 2950 series. Computer D was used as the base from which all MADMs were started from. Computers A, B and C were hosting AMs that actually implement the data mining algorithms.

Computers A, B and C where hosting more than just one AM in order to simulate more nodes in the network than actual computers were available. In order to have a realistic scenario the MADMs were not permitted to visit two consecutively AMs located on the same machine. One constraint in this setup is that it may happen that two or more MADMs visit different AMs hosted on the same computer at the same time. That makes it more likely that there are collisions in the network, which would not be the case if all AMs were hosted on separate computers. So the performance of the PDM framework in the current setup will be worse compared with a setup with one physical computer per AM, like it would be in a real application of the PDM framework.Data instance to be test for a classification, the values are entered separated by a comma. Further information that the MADM needs in order to calculate the order in which AMs are visited, is the total number of AMs or "stream mining agents available" in the LAN, the number of AMs per machine, the number or ID if this particular MADM and the number of AMs this MADM shall visit. As mentioned before the order of the agents to visit is relevant for this evaluation as several AMs may be hosted on the same computer and an MADM should not visit consecutively AMs that are hosted on the same computer. In a real application only the test instance would be required as a input parameter. The information that the GUI requests would normally be collected from the MRD agent, which has not been implemented yet.GUI of an MADM.





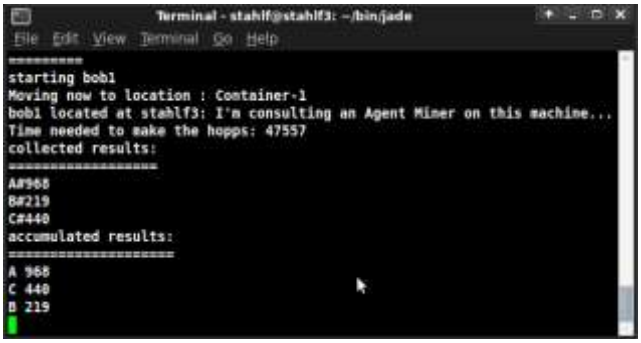




Figures show screenshots of the JADE agents running on computers A, B, and C. While computers A, B, and C are running the data stream classification process, another computer fires an MADM to consult the three aforementioned computers.

The purpose of this section is to evaluate the feasibility of the PDM framework in computational terms, in particular the communication performance and its parallel processing performance. With respect to the communication performance, the property of interest is the time needed for the MADMs to visit all the AMs and return to computer D. With respect to the parallel performance the property of interest is how much faster the PMD framework becomes the more MADMs are used.

A. Evaluation of the Communication Performance In order to evaluate the communication performance the actual data mining algorithms of the AMs were replaced by a random result generator. Assuming a classification task, the result produced by each AM would consist of the class label and a weight to indicate how reliable or important the classification produced by this particular AM is.



The random Fig.. A Screenshot of Computer D. result generator simply generates a random class label and a random weight. The reason for doing this is that generating a random result consumes only a very little amount of CPU time compared with an actual classification algorithm. In order to measure how quickly an MADM visits all the nodes it is important to bring the execution time of the AM visited to a minimum.

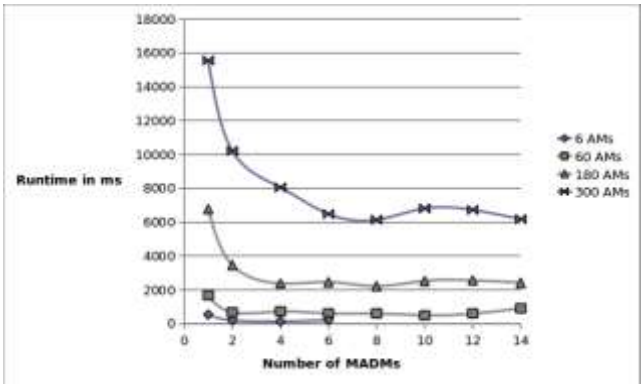


Figure depicts the time needed by one or more MADMs to visit all AMs in the network on several network configurations(different numbers of AMs). What can be seen is that the time needed for communication is decreasing at a high rate at the beginning for using more MADMs. The rate of decrease levels off very quickly and the time needed for communication seems to increase slightly for larger numbers of MADMs. The decrease of communication time with only a few MADMs can be explained by the fact that several MADMs perform their hops in parallel and the more MADMs are used the less ”hops” each MADM has to perform. However the benefit

of using more MADMs is contradicted by the fact that the more MADMS are used the higher the traffic on the network and thus the risk of collisions. Regarding that the risk of collisions is increased by the experimental setup itself, as mentioned earlier in this Section, in particular by hosting several AMs on the same computers, it is likely that the communication time would still decrease for a larger number of MADMs if there would be more computers available for the experimental setup. Nevertheless it can be seen that the communication overhead using one ore more MADMs is very low in general. B. Evaluation of the Parallel Performance using Speedup Factors One may say that regarding low communication overhead it not necessary to use more than one or two MADMs. However taking into consideration that the actual data mining algorithms embedded in the AMs will consume CPU time draws a different picture. If an MADM visits a AM it waits In this Fig. Time consumed by the MADMs to visit all AMs until the AM has derived a result and then the MADM will visit the next AM. So the more MADMs are used the more AMs can be visited by different MADMs and are executed in parallel (at the same time) and thus reduce the overall runtime. In order to evaluate this parallel behaviour all AMs have been forced to wait a period amount of time in order to simulate the execution time of their local data mining algorithm.

V. CONCLUSION

The paper introduced our Pocket Data Mining framework to enable collaborative mining of streaming data in mobile environments. The framework uses the mobile software agent's technology benefiting from its autonomous behaviour and computational efficiency. Experimental results using JADE toolkit have proved the applicability of the system. Future directions in our research would explore the numerous alternatives of collaborative mining techniques and strategies. These include varying the mining techniques, the sharing of attributes and instances of the data among nodes, and the distribution of the roles among agents that would yield the highest accuracy.

REFERENCES

[1] Agnik, MineFleet Description, http://www.agnik.com/minefleet.html

[2] Fabio Bellifemine, Agostino Poggi, and Giovanni Rimassa, Developing multi-agent systems with JADE. In Cristiano Castelfranchi and Yves Lesperance, editors, Intelligent Agents VII. Agent Theories Architectures and Lan- guages, 7th International Workshop, ATAL 2012, Boston, MA,USA, July 7-9, 2012, Proceedings, volume 2012 of Lecture Notes in Computer Science, pages 89 - 103. Springer Verlag, 2012.

[3] Gaber, M, M., Zaslavsky, A., and Krishnaswamy, S., Mining Data Streams: A Review, ACM SIGMOD Record, Vol. 34, No. 1, pp. 18-26, June 2012, ISSN: 0163-5808.

[4] Gaber M. M., and Yu P. S., A Holistic Approach for Resource-aware Adaptive Data Stream Mining, Journal of New Generation Computing, ISSN 0288-3635 (Print) 1882-7055 (Online), Volume 25, Number 1, November, 2012, pp. 95-115, Ohmsha, Ltd., and Springer Verlag.

[5] Phung N. D., Gaber M. M., and Rhm U, Resource-aware Online Data Mining in Wireless Sensor Networks, Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2007, pp. 139-146, part of the IEEE Symposium Series on Computational Intelligence 2007, Honolulu, Hawaii, USA, 1-5 April 2007. IEEE 2007, ISBN: 1-4244-0705-2.

[6] Gaber M. M., Data Stream Mining Using Granularity-based Approach, a book chapter in Foundations of Computational Intelligence Volume 6, Abraham A., Hassanien A., Carvalho A., and Snase V. (Eds), Volume 206/2009, pp. 47-66, ISSN 1860-949X (Print) 1860-9503 (Online), ISBN 978-3-642-01090-3, Springer Berlin/Heidelberg, Germany, 2009.

[7] Gaber, M, M., Zaslavsky, A., and Krishnaswamy, S., A Cost-Efficient Model for Ubiquitous Data Stream Mining, Proceedings of the tenth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2004), pp. 747-754, Perugia Italy, July 4-9.

[8] Gama J., and Gaber M. M. (Eds), Learning from Data Streams: Processing Techniques in Sensor Networks, a book published by Springer Verlag, ISBN 3540736786, 9783540736783, 2007.

[9] Gama J., Gaber M. M., and Krishnaswamy S., Data Stream Mining: From Theory to Applications and From Stationary to Mobile, presented in the ACM 25th Symposium On Applied Computing, available online at: http://www.csse.monash.edu.au/ shonali/ACM-SAC10-DSTutorial/Tutorial-SAC10-Final.pdf

[10] Haghighi P. D., Zaslavsky A., Krishnaswamy S., Gaber M. M., Mobile Data Mining for Intelligent Healthcare Support, Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS08), pp. 1-10, Hawaii, USA, January 5-8, 2009, IEEE 2009.

[11] Horovitz O., Gaber M. M., and Krishnaswamy S., Making Sense of Ubiquitous Data Streams: A Fuzzy Logic Approach, Rajiv Khosla, Robert J. Howlett, Lakhmi C. Jain (Eds.): Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, pp. 922-928, Melbourne, Australia, September 14-16, 2005, Proceedings,
Part II. Lecture Notes in Computer Science 3682 Springer 2005, ISBN 3-540-28895-3.

[12] Horovitz, O., Krishnaswamy, S., and Gaber, M, M., A Fuzzy Approach for Interpretation of Ubiquitous Data Stream Clustering and Its Application in Road Safety, Intelligent Data Analysis, Special Issue on Knowledge Discovery from Data Streams JooGama and Jesus Aguilar-Ruiz (Eds.), Vol. 25, No. 1, pp 89-108, 2007, 1088-467X (Print) 1571-4128 (Online), IOS Press.

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

