Research Challenges in Ubiquitous Knowledge Discovery

Michael May
Fraunhofer Institute Intelligent Analysis and Information Systems
michael.may@iais.fraunhofer de

Abstract

Ubiquitous Knowledge Discovery is a new research area at the intersection of machine learning and data mining with mobile and distributed systems. In this paper the main characteristics of the objects of study are defined and a high-level framework for analyzing ubiquitous knowledge discovery systems is introduced. Next, a number of examples from a broad range of application areas are reviewed and analyzed in terms of this framework. Based on this material, important characteristics of this field are identified and a number of research challenges are discussed.

Ubiquitous Knowledge Discovery

Knowledge Discovery in ubiquitous environments (KDUbiq) is an emerging area of research at the intersection of the two major challenges of highly distributed and mobile systems and advanced knowledge discovery systems.

Today, in many subfields of computer science and engineering, being intelligent and adaptive marks the difference between a system that works in a complex and changing environment and a system that does not work. Hence, projects across many areas, ranging from Web 2.0 to ubiquitous computing and robotics, aim to create systems which are "smart", "intelligent", "adaptive" etc., allowing to solve problems that could not be solved before. A central assumption of KDUbiq is that what seems to be a bewildering array of different methodologies and approaches for building smart, adaptive, intelligent systems, can be cast into a coherent, integrated set of key ideas centered on the notion of learning from experience.

Focusing on these key ideas, KDUbiq provides a unifying framework for systematically investigating the mutual dependencies of otherwise quite unrelated technologies employed in building next-generation intelligent systems: machine learning, data mining, sensor networks, grids, P2P, data stream mining, activity recognition, Web 2.0, privacy, user modeling and others. Machine learning and data mining emerge as basic methodologies and indispensable building blocks for some of the most difficult computer science and engineering challenges of the next decade.

From a high-level perspective, key characteristics of an ubiquitous knowledge discovery application are:

C1. Time and space. The objects of analysis exist in time and space. Often they are able to move.

C2. Dynamic environment. These objects might not be stable over the life-time of an application. Instead they might appear or disappear. They exist in a dynamic and unstable environment, evolving incrementally over time.

C3. Information processing capability. The objects are endowed with information processing capabilities.

C4. Locality. The objects never see the global picture, knowing only their local spatio-temporal environment.

C5. Real-Time. Because they typically have to take decisions or even act upon their environment, analysis and inference has to be done in real-time, and not only on historic data; the models have to evolve incrementally in correspondence with the evolving environment.

C6. Distributed. In many cases the object will be able to exchange information with other objects, thus forming a truly distributed environment.

Objects to which these characteristics apply are humans, animals, and, increasingly, various kinds of computing devices. It is the latter, that form the objects of study for KDUbiq.

For analyzing the different possible architectures of ubiquitous knowledge discovery systems within a high-level framework, we introduce six dimensions of KDUbiq:

1. Application Area.
2. Ubiquitous Technologies.
3. Resource Aware Algorithms.
4. Ubiquitous Data Collection.
5. Privacy and Security.
6. HCI and User-Modeling.

When designing a ubiquitous knowledge discovery system, major design decisions in each of these six dimensions have to be taken. These choices are mutually constraining each other and dependencies among them have to be carefully analyzed. KDUbiq thus adopts a *systems* view on how to build next generation knowledge discovery systems.

Two important aspects to be ubiquity have to be distinguished, namely

- *the ubiquity of data*, and
- *ubiquity of computing*.

In a prototypical application the ubiquity of the environment corresponds naturally to the ubiquity of the data – e.g. the spatio-temporally tagged data in case study 3 arise *because* the vehicles are moving, in case study 5 they arise because the collections are owned by different people. But there are borderline cases that are ubiquitous in one way but not in the other, e.g. clusters or grids for speeding up data analysis by distributing files and computations to various computers, or track mining from GPS data where the data a analyzed on a central server in an offline batch setting.

To stimulate research and further define the field, the KDUbiq research network (www.kdubiq.org), funded by the European Commission, was launched in 2006. Currently it has more than 40 members. It is organized around working groups for each of the dimensions of KDUbiq. It has launched workshops series at KDD, ICDM, and PKDD and ECML/PKDD, including mining data streams from sensor data, on privacy-preserving data mining, on spatial data mining, on user modeling, on ubiquitous web mining. The general points that emerge from these activities are discussed in a joint book, currently under preparation, the KDUbiq "Blueprint on Ubiquitous Knowledge Discovery". It aims for a comprehensive overview on the six design dimensions and the research agenda needed for implementing the KDUbiq vision.

To provide a more specific description of the content of KDUbiq, in this document we analyze a number of case studies (in this extended abstract the descriptions have to be shortened). The following selection criteria have been used: (1) each case study focuses on a different domain; (2) it presents a challenging real-life problem; (3) there is a body of prior technical work addressing at least some of the six dimensions of ubiquitous knowledge discovery.

Existing work is not necessarily done under the label of "Ubiquitous Knowledge Discovery". The subject is new and draws on work scattered around many communities. For a review of earlier work on distributed data mining see [7].

Case Study 1: Autonomous driving vehicle

The first case study provides an impressive example how machine learning can help to solve an important real world task: The DARPA grand challenge. The goal was to develop an autonomous robot capable of traversing unrehearsed road-terrain. The robot had to navigate a 228 km long course through the Mojave desert in no more than 10 hours. The challenge was won in 2005 by the robot Stanley, built by a Stanford-based team lead by Sebastian Thrun [17].

Modern vehicles fit the basic characteristics of ubiquitous knowledge discovery systems very well: they exist in a dynamic environment, moving in time and space, equipped with sensors, increasingly communicating with other devices, e.g. satellites, for navigation. What sets Stanley apart from traditional cars on the hardware side is the large number of additional sensors, computational power and actuators.

Stanley uses machine learning for a number of learning tasks, both offline and online. An offline classification task solved with machine learning is obstacle detection, where a first order Markov model is used. A second online task is road finding: classifying images into drivable and non-drivable areas. An adaptive Mixture of Gaussians algorithm is used to model a distribution that changes over time. It would be impossible to train the system offline for all possible situations.

Case Study 2: Activity recognition – inferring transportation routines from GPS-data

The widespread use of GPS devices led to an explosive interest in this type of data. One emerging area is assistive technologies: A personal guidance system helping cognitively impaired persons to find their way through a complex transportation system. This application has been proposed by the project *Opportunity Knocks* [14][11].

The basic infrastructure is a mobile device equipped with GPS and connected to a server. An inference module running on the server is able learn a person's transportation routines from the GPS data collected. It is able to give advice to persons, which route to take or where to get off a bus, and it can warn the user if he commits errors, e.g. takes the wrong bus line.

Machine learning algorithms are used to infer likely routes, activities, transportation destinations and deviations from a normal route. It is an unsupervised learning task. The basic knowledge representation mechanism is a Dynamic Bayesian Network. In further work, Conditional Random Fields are used.

Case Study 3: Intelligent Multi-Agent Systems - Smart Home

MavHome [3] is a project that aims to build an intelligent environment, *a smart home*, which is able to acquire and apply knowledge about its inhabitants and surroundings. A home is seen as a rational agent capable

of perceiving the state of the home through sensors an acting upon the environment through effectors.

MavHome uses a sensor network for perceiving light, humidity, smoke, gas (CO), motion, and door, window seat status sensors. Inhabitant localization is done using passive infrared sensors. The software architecture is based on CORBA for communication between agents.

The system is based on combining multiple heterogeneous machine learning algorithms in order to identify repeatable behaviors (patterns), to predict inhabitant activity and to learn a control strategy. The information is used for automation and optimization of the conditions in the house.

For detecting patterns a sequential pattern mining algorithm ED is used which minimizes description length, and processes data as they arrive, thus assuming a data stream setting. Behavior prediction is done via the ALZ algorithm, taking ideas from the well-known LZ78 text compression algorithm. The predictive performance on real-world data collected over a month, was 44% when ALZ was combined with ED.

Case Study 4: Real-Time Vehicle Monitoring

The Vehicle Data Stream Mining System VEDAS [6] is a mobile and distributed data stream mining application. It analyzes and monitors the continuous data stream generated by a vehicle. It is able to identify emerging patterns and reports them back to a remote control center over a low-bandwidth wireless network connection. Applications are real-time on-board health monitoring, drunk-driving detection, driver characterizations, and security related applications for commercial fleet management.

VEDAS uses a PDA or other light weight mobile device installed in a vehicle. It is connected to the On Board Diagnostic System (OBD-II); other sensory input comes from a GPS device. Significant mining tasks are carried out on board, monitoring the state of transmission, engine and fuel systems. Only aggregated information is transmitted to a central server via a wireless connection. The data-mining has to be performed onboard using a streaming approach, since the amount of data that would have to be transmitted to the central server is too huge.

The basic idea of the VEDAS data mining module is to provide distributed mining of multiple mobile data with little centralization. The data mining algorithms are designed around the following ideas: minimize data communication; minimize power-consumption; minimize onboard storage; minimize computing usage; respect privacy constraints.

VEDAS implements incremental PCA, incremental Fourier transform, online linear segmentation, incremental k-means clustering and several lightweight statistical techniques. The basic ideas of these algorithms are of course well-known; the innovation lies in adapting to a resource-constrained environment, resulting in new approximate solutions.

Case Study 5: Web2.0 – Music Mining

With the advent of Web 2.0, collaborative structuring of large collections of multi-media data based on meta-data and media features has become a significant task. Nemoz (NEtworked Media Organizer [12]) is a Web 2.0-inspired collaborative platform for playing music, browsing, searching and sharing music collections. It works in a distributed scenario, a loosely coupled P2P domain. Nemoz has facilities for Web 2.0-style tagging, but also allows users to automatically classify their audio-collection using machine learning.

Nemoz is motivated by the observation that a globally correct classification for audio files does not exist, since each user has its own way of structuring the files, reflecting his own preferences and needs. Still, a user can exploit labels provided by other peers as *features* for his own classification: the fact that Mary, who structures here collection along mood, classifies a song as “melancholic” might indicate to Bob, who classifies along genre, that it is not a Techno song. To support this, Nemoz nodes are able to exchange information about their individual classifications. These added labels are used in a predictive machine learning task. Thereby Nemoz introduces a new type of learning problem [18]: the collaborative representation problem.

This application is a representative of a innovative subclass of applications in a Web 2.0 environment. Whereas most Web 2.0 tagging applications use a central server where all media data and tags are consolidated, the current application is fully distributed.

Research Challenges

We reviewed examples from data mining, machine learning, probabilistic robotics, ubiquitous computing, ambient intelligence/smart homes, and Web 2.0. Collectively, these applications span a broad range of ubiquitous knowledge discovery applications from vehicle driving, assistive technologies, home automation, transportation, and leisure. In this section common lessons from the case studies are drawn and research challenges identified.

Thesis 1: *Across a large sector of challenging application domains, further progress depends on advances in the fields of machine learning and data mining; increasing the ubiquity sets the directions for further research and improved applications.*

In each case study providing adequate machine learning capabilities was a major cornerstone for success. For example, in case study 1 the authors sum up the role of machine learning: "The pervasive use of machine learning, both ahead and during the race, made Stanley robust and precise. We believe that those techniques, along with the extensive testing that took place, contributed significantly to Stanley's success in this race" [17].

Moreover, in many case studies additional benefit could be gained by addressing the dimensions of KDUbiq more fully. In case study 1, the desert driving scenarios is limited if compared to a normal traffic scenario. Directions for future research are set by the DARPA 2007 urban challenge, where the vehicles are required to navigate their way under normal traffic conditions, with other cars present, traffic light, turns etc. In this new challenge one of the key aspects of the ubiquitous knowledge discovery paradigm becomes highly important, namely the *presence of many distributed, interacting objects.* Furthermore, once cars similar to Stanley go into production, HCI and user modeling as well as privacy will play a central role, which have not been addressed so far: There are many questions starting from user acceptance (the autonomy of the car diminishes the autonomy of the user!) to liability and legal issues.

For the activity recognition prototype in case study 2, the architecture prototype will face a number of practical problems, firstly, when there is no reliable GPS or phone signal, secondly because constantly exchanging information with the server consumes a lot of battery power; and thirdly because this scenario inevitably creates privacy threats: a service provider offering these services could in principle track the mobile behavior of a person by storing an annotated diary of a person's activities. A "KDUbiq Upgrade" would result in a much more satisfactory design: *Once the machine learning moves to the mobile phone, many of the limitations vanish*: activity tracking could work also in the absence of a phone signal, and privacy is much better addressed, since information remains under the control of the user. Case study 2 might have been prevented from a more mobile solution because

of scalability implications: meeting the resource constraints put up by mobile devices.

A way to address such problems has been shown by case study 3: new implementations even for well-known algorithms are needed, which provide more efficient solutions by stepping back to approximations. This provides many opportunities both for theoretical and applied research. In this case study the impact of distributed and privacy-preserving data mining, as well as their interrelation, become very clear.

In case study 5 we saw that for learning in KDUbiq setting a new class of learning problems was introduced: the collaborative representation problem. *Ubiquitous knowledge discovery asks for the invention of new learning scenarios – not only new solutions to existing problems are found, but new classes of learning problems are invented.*

Thesis 2: *Ubiquitous Knowledge Discovery requires research beyond independent and identically distributed data.*

Despite the diversity of applications and communities, one universal feature that emerged is that of *concept drift*. It simply cannot safely be assumed that a data sample collected in the past sufficiently well describes the future situations the system will encounter. Thus each application at least partially applied algorithms that can adapt to distributions that change in an unforeseen manner. For case study 1, it was necessary to adapt to terrain that had not been explored before; in case study 2 and 3, unsupervised learning was used because behavioral patterns of persons change over time; in case study 4 the system has to be able to recognize deviating behavior (malfunctions) of some machine parts; in case study 4, the collections evolve incrementally, based on possible changing interests of the users. It is the combination of a dynamically changing complex environment combined with local information processing capabilities, which makes this feature so universal.

This is a significant departure from the current mainstream in data mining and machine learning. Most approaches in machine learning and data mining assume that the training data is randomly sampled from a fixed distribution. Thus leading paradigms in learning theory, PAC learning, Statistical Learning Theory, maximum likelihood estimation, and most practical algorithms are crucially based on this assumption.

In the last few years, some promising research in this direction is appearing. Krause and Guestrin [9] discuss active learning of non-stationary Gaussian processes in a ubiquitous setting. The task is to find a near-optimal placement of sensors for monitoring a river. There is also recent theoretical work by Ommen and Rueda [13] on so-called *weak estimators* for learning non-stationary distributions. The *1st KDD from Sensor Data* workshop [4] (supported by KDUbiq and partially organized by Working Group 3) contains further examples of learning from data streams and gives further pointers to current research activities on this topic.

Thesis 3: *Ubiquitous knowledge discovery requires new approaches in spatio-temporal data mining, privacy-preserving data mining, sensor data integration, collaborative data generation, distributed data mining,*

and user modeling. The most successful approaches will be those, that combine several aspects.

Case studies 1,2 and 4 highlighted the central role of spatio-temporal data mining, especially GPS track data. For an overview on recent developments in this area, see [15][8].

All case studies proved to be privacy sensitive, since ubiquitous devices reveal highly sensitive information about the persons that carry them. Privacy-preserving data mining in a distributed, spatio-temporal environment poses many challenges [16][1]. Privacy issues will become even more pressing once the application migrate from a research prototype status to real products.

On the data collection side two major issues arise. The first one is collection and integration of data collected from heterogeneous sensors as in case studies 1,3, and 4. Case study 5 highlights data collection issues in a collaborative Web 2.0 environment.

Finally, user modeling, and HCI are particularly challenging for case study 1, 2, and 5, since technically non-skilled end users will be confronted with those systems [2].

Acknowledgements. This work has been funded by the European Commission under IST-FP6-021321 KDUbiq Coordination Action. Contributions by all project partners are gratefully acknowledged.

References

[1] Bonchi, F. , Y. Saygin, V.S. Verykios, M. Atzori, A. Gkoulalas-Divanis, S. Volkan Kaya, and E. Savas, "Privacy in Spatio-temporal Data Mining", in Giannotti, F., Pedreschi, D., 2007, 253-276.

[2] Berendt, B., Kröner, A. , Menasalvas, E., Weibelzahl, S. (eds), *Proc. Knowledge Discovery for Ubiquitous User Modeling '07*, http://vasarely.wiwi.hu-berlin.de/K-DUUM07/.

[3] Cook, D., G. M. Youngblood, and S. K. Das, "A Multi-Agent Approach to Controlling a Smart Environment", *AI and Smart Homes*, pages 165-182, Springer, 2006.

[4] Ganguly, A., Gama, J., Omitaoumu, O., Gaber, M., Vatsavi, R. *Proceed. of the First International Workshop on Knowledge Discovery from Sensor Data, KDD'07*, 2007.

[5] Giannotti, F. and Pedreschi, D. (eds.), *Geography, mobility, and privacy:a knowledge discovery vision*, Springer, 2007, in print.

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

[7] Kargupta and K. Sivakumar, "Existential Pleasures of Distributed Data Mining" in H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (eds) *Data Mining: Next Generation Challenges and Future Directions*" AAAI/MIT Press, 2004.

[8] Kuijpers, B., Nanni, M., Körner, C., May, M., Pedreschi, D., "Spatio-Temporal Data Mining", in Giannotti, F., Pedreschi, D., 2007, p.277-306.

[9] Krause, A., C. Guestrin. "Nonmyopic Active Learning of Gaussian Processes - An Exploration-Exploitation Approach". *Proc. of 24th International Conference on Machine Learning (ICML) 2007.*

[10] Liao, L. , Patterson, D.J. , Fox, D., Kautz, H.: "Learning and Inferring transportation routines", Artificial Intelligence, 2007.

[11] Liao, L., Fox, D., Kautz, H.: "Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields", *International Journal of Robotics Research*, 2007.

[12] Mierswa, I., Morik, K., "Automatic Feature Extraction for Classifying Audio Data", *Machine Learning* 58, 2005, 127-149.

[13] Oommen, B. J. and Rueda, L., "Stochastic Learning-based Weak Estimation of Multinomial Random Variables and Its Applications to Pattern Recognition in Non-stationary Environments", *Pattern Recognition*, Vol. 39, 2006, pp. 328-341.

[14] Patterson, D.J, Liao, L, Gajos, K., Collier, M., Livic, N., Olson, K., Wang, S., Fox, D., Kautz, H., "Opportunity Knocks: a System to Provide Cognitive Assistance with Transportation Services", *Proc. Ubicomp* 2004, 433-450, Springer.

[15] Rinzivillo, S., Turini, S., Bogorny, V., Körner, C., Kuijpers, B., May, M, "Knowledge Discovery from Geographical Data," in Giannotti, F., Pedreschi, D., 2007, 253-276.

[16] Schuster, R. Wolff, and B. Gilburd, "Privacy-Preserving Association Rule Mining in Large-Scale Distributed Systems", *Proceedings of CCGRID`04*, Chicago, Illinois, April, 2004.

[17] Thrun, S. et al: Stanley, "The Robot that Won the DARPA Grand Challenge", *Journal of Field Robotics* 23(9), 661-622, 2006.

[18] Wurst, M., Morik, K., "Distributed feature extraction in a p2p setting – a case study", *Future Generation Computer Systems* 23 (2007), 69-75.