MLnet Technological Roadmap

30 November 1999

MLNET-II

Project 29288

Universiteit van Amsterdam

MLNET Technological Roadmap

30 November 1999

Preface

Acquiring knowledge is fundamental for the development of intelligent systems. Three main approaches have been followed: to elicit knowledge from experts, through interviews and analysis of protocols, the approach taken by Knowledge Acquisition (KA); to automatically discover general rules to be used for future decisions, as done in Machine Learning (ML); finally, to exploit past experience gained in concrete cases to be recalled and adapted in new situations, as suggested by Case-Based reasoning and learning (CBR).

This technical roadmap document embodies a vision on where the research areas in the three related fields are heading to in the near future. It describes the current research topics, possible application areas, and the emerging new topics with a focus on the future. This document is based on an inventory of opinions of MLNET members on the main future issues and on a number of discussions with members of MLNET. Mark Keane and Nick Kushmerick provided major contributions on Case-Based Reasoning, Richard Benjamins and Bob Wielinga on Knowledge Acquisition and Lorenza Saitta and Maarten van Someren on Machine Learning. It is a version that will be a guideline for MLNET actions in the year 2000 and it will also be the input for a process in which the main directions will be articulated and revised on the basis of developments in 2000.

The main focus of this document is on future developments in Europe. Sections 1 and 2 summarise the main current research topics in Europe and the main achievements. Section 3 outlines the main future developments. Sections 4 briefly discusses external factors that affect application of results and section 5

1. Current Research Focus in Europe

Above the diversity of learning tasks approached inside the European ML, KA and CBR communities, some common themes emerge, driven by the rapidly changing landscape of the fields and by new application needs.

One of the theme concerns **Data Mining** and **Knowledge Discovery in Databases**, which currently shows a large potential of applicability is several industrial, economical, and commercial domains. In particular, hot issues are those related to *scalability* to really large databases (through sampling or partitioning), deep *integration* with database

management systems, creation of *Data Mining Environments,* providing the end-user with help and guide to use the available methodologies, and designing tools to enable to user to effectively exploit his/her background knowledge of the domain. In fact, the effective management of knowledge discovery processes requires flexible and extensible supporting software and systems.

Another emerging theme is **Relational learning**, oriented to the development of algorithms to acquire knowledge expressed in richer languages than a vector of attribute values. This form of learning is studied in depth in the context of *Inductive Logic Programming*. Particularly important is the issue of *complexity,* which can be related to the definition of suitable language and procedural biases, and is fundamental for making relational learning practical in Data Mining. Neuronal nets are becoming an important tool for relational learning and the generalisation of knowledge structures leading to new possibilities in Data Mining, image processing and retrieval for example in databases of chemical reactions and protein structures.

A third theme, connected to a flourishing activity in the AI community at large, is **Multi-Agent learning**, which has several ramifications; *Multistrategy learning* and *Multirelational Instance-based learning* can be considered some of them.

A further theme, that makes Machine Learning closer to Cognitive Science, is **Modelling human learning**, or, at least some aspects of it. This attention to human learning is also driven by the essential role assumed by the **human expert/user** in the life cycle of a learning application development, both through the design of intelligent and friendly interfaces, and by providing the user with means for capitalising his/her previous experience in machine learning applications and for **reusing** the acquired knowledge and expertise.

Finally, an important trend in today ML research in Europe is an increasing awareness of the relations existing between Machine Learning and **Statistics**, especially for what concerns model selection and result interpretation. Recently, in fact, it is being realised that results from statistics greatly contribute to a better understanding of machine learning. Also, methods with roots in statistical approaches, such as the **Support Vector Machines**, are experiencing a large diffusion.

In KA, the main current topic emerges as the study of **ontologies**. In particular, semi-automatic *construction* of ontologies (high-level ontologies, linguistic methods/machine learning techniques for building ontologies from text), and systematic *integration* of ontologies (characterise "competent"/modelling assumptions, similarity measures). Other important topics are **Reuse/Configuration** of components, knowledge acquisition from the **Web** with heterogeneous and multiple sources, **Text mining** (integration of methods from computational linguistics), **specification languages** methodologies, **Knowledge management** for large scale enterprises, and intelligent integration and retrieval.

The idea of **re-using** previous experience is also at the basis of the CBR approach, which tries to make a virtue of re-using prior experience in classification, problem solving and

reasoning tasks. CBR systems deploy case bases of previously-solved problems to tackle new problems. Typically, when a new problem is encountered, it is used as a probe to retrieve a similar problem from the case base which is then modified or adapted to fit the target problem. CBR systems can also learn by storing the results of a successful problem-solving episode in its case base to be used in the future.

2. Achievements

In the last decade, ML has made substantial advances both on the theoretical side, and on the application front. Among the most fundamental **methodological** results, three deserve special mention: the design of *boosting* and bagging algorithms; with the associated deep understanding of the bias/variance trade-off; the *No-Free-Lunch* theorems, which has roots in Bayesian statistics and has enhanced our understanding of how good learning algorithms can possibly be. The *feasibility* of learning algorithms with understandable, justified and theoretically analysable results can also be included in the recent successes of ML.

CBR allowed a strong reduction of the knowledge engineering load in the construction of intelligent systems. Where possible, existing data-base representation can be reused without significant re-representation, even though, in some more complex systems, there may be a need to hand-craft a small set of cases from which to seed the case-base. The acquisition of cases can be helped by KA methods, and can also be coupled with ML systems in hybrid approaches.

All the three fields have striven towards comprehensibility: KA, by maintaining the original expression of knowledge given by the expert; CBR by delivering their solutions in a case format familiar to the user, and ML by exploiting knowledge descriptions easily translatable into natural language.

On the **application** front, the number of fielded applications exploiting automatically learned knowledge is increasing rapidly. Among the application domains that most did benefit from ML are *information extraction from text,* design of *softbots* for the Web, *medicine, molecular biology, telecommunications, banking* and *commerce.*

In several among the above mentioned domains also KA has been fruitfully applied, alone or in co-operation with automated knowledge acquisition, for instance, design of *Smart interfaces* (for semantic Web access and intelligent information integration and retrieval), design of *Agent systems* (exploitation of domain/task knowledge, knowledge sharing), and *Web services.* Further application domain for KA have been *Digital libraries* and *Knowledge management* (knowledge organisation, structuring and integration).

The same sharing of goals and tasks occurs for CBR: in fact, *classification and pattern recognition, diagnostic reasoning* and *user profiling* have also been targeted by ML, whereas legal reasoning, design, and automated programming have been more specific of the CBR approach.

3. Future

Among the research themes approached today, and mentioned in Section 2, some specific topics seems to have the potential to persist and to gain further relevance in the medium term future. Among these **trends** we may notice the more human-oriented aspects of Data Mining, whose current popularity will most likely dominate for another few years; so ML will continue to see an emphasis on scalability and user issues such as *data visualisation, results explanation,* and *model validation and interpretation.*

In parallel, development in *Statistical Learning Theory* (and statistics in general) will influence ML even more strongly. Statistical data analysis considers problems which often show similarities to problems studied by the machine learning and data mining community. A number of powerful learning tools (such as Naïve Bayesian classifiers or Support Vector Machines) which emerged from a statistical background are now well known in the Machine Learning community. Probabilistic knowledge representation schemes (Bayesian networks) provide interesting new learning tasks which are now being studied in machine learning.

We are currently moving away from worst-case analyses of learning algorithm (which had their background in computer science) towards average- or actual-case analyses which take the properties of the given learning problem, or of the focused learning algorithm, into account and explain the behaviour of the learner accurately. The issue of complexity is strictly related to the one of knowledge representation. *Ontologies,* still a focus of research in KA, can help in understanding and standardising domains, especially those requiring large ontologies, together with methods to establish mapping between ontologies. Links with other areas, such as EuroWordnet of the analysis of thesauri, will become important. On the other hand, Machine Learning can help with (semi-)automated ontology construction. Another relevant issue, in all approaches to learning, will be determination of the *granularity* of problem-solving methods and ontologies.

Evaluation of hypotheses, or models, is an issue that is gaining importance. In particular, for data mining tasks methods are required that return provably accurate hypotheses after only a reasonable amount of database queries. Although the known statistical sampling methods do not solve this problem well enough, they provide the starting point from which better solutions have to be sought.
Many research efforts will still be devoted to various aspects of dealing with *textual data:* knowledge acquisition from texts, full text document retrieval, digital libraries, Web navigation, information extraction, and so on. Knowledge acquisition from *text* will continue to be pursued, in particular for information extraction. To this aim, a tighter integration between *Natural language* understanding and KA is sought. Great efforts will be devoted to KA on the *Web* (Web Agents and Wrappers, Brokering services). Specific ontologies could be built up for Web communities.

Finally, use of ML to implement *adaptivity* will mostly be found in Intelligent agent and robotics applications. From the point of view of the methodologies, beyond the classical

symbolic induction of decision trees and rules, Machine Learning is the more and more exploiting *neural* and *evolutionary* approaches, or a combination thereof.

The above mentioned trends, however, only partly cover the goals that the European Machine Learning community consider pre-eminent among its **desiderata** for the next few years. Two topics that are both trends and desired goals are *scalability* and *automation of the pre- and post-processing* phases in Data Mining. Concerning scalability, it is not just a question to handle large data sets, but also to handle more complex learning problems. Current application efforts tend to focus on "bashing the problem into feature vectors" so that existing ML tools can be applied. We need to study these applications and ask what tools would have made it easier to solve the problem directly, without so much reformulation and feature tweaking. This may involve capturing the qualitative structure of the problem. Another actual application is the *exploration of parameter spaces* of complicated processes which can not be modelled mathematically and contain "dangerous" or non-optimal regions as for example in biotechnology, energy production and chemical technology thus simulating human skill acquisition in the process control. In general risk parameters have to be included in this case and risk has to be minimised instead of error.

Topics not pursued currently in KA but likely to be important in the future are *acquisition and representation of images (including diagrams),* knowledge content *updating* and directed *forgetting*. Hot topics include *Knowledge management* (maintenance, mining) in particular development of *Enterprise Ontologies* and *integration with Intranets.* There is an emerging concern with knowledge management suggesting that case-based reasoning techniques and data mining techniques should be integrated with other forms of knowledge management to provide a solution to the problems inherent in it, offering thus occasion to the development of multistrategy systems. An emerging growth area for CBR, of central relevance to issues like corporate memory, is the development of ideas on the maintenance of *multiple case-bases* for a single application. CBR systems in large organisations will increasingly allow multiple users to input cases into the case-base. A major problem is how to deal with this type of distributed CBR; in particular, how a case-base can be appropriately updated in this type of environment. Implicitly, this involves changes and expansions of current ideas of learning in CBR systems. *Formal methods* will be required to support knowledge validation and verification in cases where correctness of systems is of great importance and justifies the effort.

In learning agents, the key challenge is to understand how complex problems can be divided into simpler ones while ensuring that there is some way of combining locally optimal solutions to the simpler problems into a globally optimal (or at least very good) solution to the original problem. This leads us to the issue of *hierarchical* and *multiagent* learning. To design Web agents involves the problems of distributed systems and integration of information gathered from heterogeneous sources (Intelligent Information Integration). For *life-long learning,* the key issue is to develop long-life agents capable of accumulating knowledge and reuse it to guide further learning. Learning by Exploration or *Active Learning* is the goal directed generation of training sets for exploring regions of interest in the feature or process parameter space, for example given by complicated

numerical constraints. It can be used for example to compute decision functions for constraint satisfaction or even (approximately) solving constraint nets for spatial problems like spatial inference and rule learning where no simple analytical solution can be constructed. This can also help to provide approximate solutions to numerous well known problems, e.g. action and motion planning for robots (with obstacle avoiding) where constraints are given by numerical equations/inequalities containing trigonometric functions.

The issue of *reusability* also concerns the intervention of the *human expert/user,* which should be facilitated as much as possible, as well as his/her work capitalised for the future; this goal can be articulated in a number of subgoals: design of tools to configure, reuse and reconfigure Knowledge Base Systems, tools for user-friendly knowledge maintenance, and for allowing the user to transfer to the learning system as much as possible of his/her background knowledge. To this aim, problem solving methods are to be designed in such a way that complex problem solvers can be easily assembled from them. The issue of re-usability is especially important and difficult for CBR: past case adaptation and the development of suitable tool supporting it is currently hampering the construction of more advanced CBR systems. Solving the adaptation problem and learning adaptation knowledge remain thus serious issues for future research.

A key challenge is to find easy and expressive ways of capturing *"contextual"* knowledge so that the learning process is efficient and effective. Examples of directions worth pursuing: (a) using belief network notation to express qualitative prior knowledge about the structure of a process, (b) developing a language for describing complex applications involving time-series and spatial data, (c) methods for deriving useful features from these representations, (d) incorporating prior knowledge with training data.

The trend to learn from textual data should be extended to include learning from *images,* from *spatial* and *temporal* data, from *multi-media* data and hypertexts. There will also be requirements for *spatial* and *temporal* knowledge acquisition, and, in general, design of ontologies for indexing large multimedia databases.

Machine learning should address new fields, where the construction of software systems for *"difficult-to-program applications"* could bring substantial improvements; these applications include robotics, process control, planning, scheduling. To reach this goal, we need to represent the "context" in which the learning system will operate so that we can be optimising the right measure of performance and handling the interactions between system components. Induction of (recursive) functional programs from example computations or finite initial programs can be used - especially if combined with analogical reasoning and program transformation - to model human acquisition of programming and other cognitive skills and - seen from the point of view of applications - as a tool for assisting (functional) programmers.

Even though some *modelling of human learning* already occurred, much more is required in order to grasp the very foundations of learning, and to design more effective artificial learning systems. This line of research is important to realise cognitive architectures for

learning agents. This issue involves also *user modelling,* based on both rules and cases. In fact, several systems have begun to develop tutoring techniques that rely on cases, either to deliver materials or to characterise user profiles. This appears to be a low-knowledge-engineering solution to the difficult problem of user modelling which have led to some initial successes.

An interesting goal is to *embed a learning component* is performance systems, in such a way that learning could become a side-effect of other (routine) tasks. Finally, a further related topic will be the control of processes, whose state is not completely known (see "partial observability"), which is almost always the case in practical applications, where not every parameter of the process can be measured, where the measurement itself may prone to inaccuracies, or where the process can only be described incompletely by a finite list of parameters (e.g. chemical processes).

Transversal to all fields is the development of *hybrid systems,* which combine multiple techniques, for instance, neural networks, genetic algorithms, rule-based symbolic induction, KA and CBR. Extensive investigation are required to assess whether such combinations will be winning for the future.

In order to reach the above goals, some **methodological advances** are necessary. The first one is taming the *complexity* of the learning tasks, especially when dealing with structured or otherwise complex data. One of the key could be *constructive induction* or *abstraction,* and another one *parallelism* and data and/or task *partition.* Effective exploitation of results from AI and Cognitive Science on knowledge representation and reasoning could help substantially.

Two kinds of learning systems development seem necessary: a *"horizontal"* general-purpose system that can be easily adapted to a wide range of applications, and an application-oriented *"vertical"* one, specialised for a given application field. A possibility could be to develop "generic" learning architectures, which can then be instantiated into specialised systems for different applications. In KA and CBR, *Knowledge management* will require sophisticated organisation and structuring tools, as well as effective visualisation techniques.

Finally, among the most promising **fields of applications**, we can mention *Natural Language* Processing, *Biology* and *Ecology, Banking, E-commerce, Geographical Information Systems* (GIS), *Telecommunications* and *Security* (for instance, intrusion detection), *World Wide Web.* Sectors of industry that are “underrepresented” in current applications are *administration and government* and *computing services*. Being CBR one of the most successfully methodology integrated into internet applications (e.g., help-desk systems and on-line purchasing systems) this development is expected to gather greater momentum with the more systematic use and dispersal of CBR in such technologies.

This may be due to lack of familiarity with learning methods in these areas rather than an intrinsic obstacle. We therefore may expect a rise of applications in these areas.

4. External factors

Apart from scientific and technical factors, several external factors have an effect on the application of ML, CBR and KA technology. Factors that restrict the potential impact of ML are legal conditions, lack of standardisation, lack of training and lack of innovative, paradigmatic applications that exemplify the potential of ML technology. Important *legal conditions* are *privacy regulations* (that prohibit the collection of data about individuals as units, for example in e-commerce) and legal *ownership* of the results of machine learning. A lack of *standardisation* of data, terminology and methodology restricts the possibilities for exchanging and combining data and results (although progress has been made recently in the context of the CRISP DM project). Machine Learning is not yet well-embedded in European higher *education*. With a few exceptions, ML is offered as a specialisation course to students in Computer Science. However, the main application of ML is now in data analysis and this is a topic that is not part of most computer science curricula. Important application areas are dominated by educational programmes in engineering and business and management sciences where ML, CBR and KA are not part of the curricula. Considering the discrepancy between the scientifically advanced state of the field and the relatively small number of applications, there seems to be a lack of innovative *application paradigms,* in particular of learning systems that are embedded in other software systems.

5. Actions

A suitable and timely action could be the launch of a Europe-wide R&D project, involving all the methodological components relevant to manual and automated knowledge acquisition, with the aim of exploiting both the available variegated researcher competencies, and the complementarities of methods. This project could provide the ground to make the (semi-)automated construction of intelligent systems a European reality. The intrinsic difficulty of automating learning requires a major effort to let the field take off. On the other hand, making knowledge acquisition and discovery practical could have a major impact on tomorrow's society.

In parallel an effort should be made to demonstrate the potential of machine learning to potential users of the technology in the form of (relatively small) innovative applications and to produce educational materials that support education to facilitate teaching ML, CBR and KA outside computer science.