Intelligent Data Analysis 12 (2008) 145–146
IOS Press

Guest Editorial

Philosophies and methodologies for knowledge discovery

At its inception in the mid 1990's, Knowledge Discovery in Databases (KDD) was portrayed/defined [1] as a new cohesive discipline, formed from the confluence of statistics, machine-learning and information systems, with the aim being "to discover by automatic means useful new knowledge *from large and complex data stored in databases*". Data mining (DM), often taken to be synonymous to KDD, or a technical component of KDD, is strictly speaking more general than KDD since the italicized part of the above definition of KDD is dropped.

However, defining KDD/DM as a new discipline does not mean that such a disciple exists as a coherent entity, or will ever exist. The diverse range of developments under the KDD/DM banner in the last 10 years do not give the impression of KDD/DM being a coherent discipline. Quite the converse! The dichotomy in DM between those in the machine-learning camp and those in the statistics camp seems to persist. DM seems to have been pursued by the machine-learning community into ever more specialized specific applications, e.g. text, screen, image, ..., mining, but with no general intellectual framework being constructed. The statistical community seems almost to have left the KDD/DM theatre of operations, possibly due to their historical antipathy to the multiple-comparison paradigm that is implicit within KDD/DM, and seem to have focused their communal efforts, such as they are, into the development of R, the open-source descendent of S and S-Plus.

Similarly, there seems to have been a continuing dichotomy between KDD/DM over whether the database/warehouse structure is fundamental or not. Researchers from the IS discipline continue to regard the DBMS/warehouse structures as relevant to KDD/DM, and have generally limited their DM techniques to SQL-like analytical tools, mostly making use of the materialized data hypercube, or other simple DM-algorithms with scale linearly with the size of the database.

It is ironic that KDD, concerned as it is with knowledge, has no theory of knowledge as a part of it, and lacking such a feature KDD has remained essentially incoherent. Furthermore there has never been a general framework enunciated to guide the selection of appropriate DM-models, let alone one which would support the hope for DM to be an automatic process.

It was the above list of concerns about the continuing lack of coherence and integration of KDD/DM that led to the DEXA Workshops of 2005/2006 on "Philosophies and Methodologies for Knowledge Discovery" (PMKD), with the hope of addressing some of these core inadequacies in KDD/DM. The five papers in this special issue derive from the papers presented at these PMKD Workshops but have been much extended, completely rewritten, or in some cases are the result of an integration of ideas from more than one workshop paper. We feel that these five papers do address, each in their own way, the

[1] U. Fayyad, R. Uthurusamy and P. Smyth, The KDD Process for Extracting useful Knowledge from Volumes of Data, *Communications of the ACM* **39**(11) (1996), 27–34.

conceptual issues/problems which lie at the heart of the KDD/DM "discipline", and each has its own perspective on how KDD/DM can be moved towards the status of a truly coherent discipline.

The call for PMKD'05[2] had the aim of addressing the internal consistency and coherency issues of KDD/DM as a discipline, at what Rennolls and Al-Shawabhkeh term the technical/technological level. The call for PMKD'06[3] aimed to widen the scope to the technological/strategic level, so as to examine the relationships between KDD/DM and the wider KM (Knowledge Management) corporate environment within which KDD/DM is conducted, with particular themes on Deployment, Development, and Decision Support Systems.

At the technical/technology level, Rennolls and Al-Shawabhkeh highlight the fact that most KDD/DM systems do not fully exploit machine-learning search techniques, particularly in the exploitation of multiple-comparison searches. They also suggest that linked ontologies of data and models would provide the required bridge to support the choice of suitable DM-model, and that a theory of knowledge is needed for KDD/DM in order to meaningful interpretation of results. Vityaev and Kovalerchuk adopt the 'representational'[4] principle of measurement theory and take 'meaningfulness'[5] of DM-models to be crucial. They adopt a first order logic (FOL) as their modelling framework, and superimpose a probabilistic infrastructure to support the inference process. They then an ontology of models that may be exploited to yield a "Discovery" system for DM. Vityaev and Kovalerchuk demonstrate the use of these methods on an important range of practical problem areas, including financial analysis, and image analysis, and show the logical structure of their approach offers considerable advantages over an attribute-value-language (AVL) approach which might use neural network or decision-tree DM-models. Charest et al. make use of a formal OWL-DL ontology, in conjunction with the use of case-based reasoning (CBR), to design and build a prototype DM-Assistant with the aim of providing decision support with the knowledge products of the DM process.

At the technology/strategic level, Rennolls and Al-Shawabhkeh argue that knowledge representations are needed which can provide a consistent enterprise-wide knowledge discovery and communications framework. They suggest that visual/graphical representations are most likely to provide the required *lingua-franca*, and that Bayesian Belief Networks would provide a suitable formal framework for the communication and sharing of aims, prior knowledge, and discovered knowledge. Pechenizkiy et al. argue that DM, by analogy with IS, should be more directed in its future research activities by matters of utility and relevance, and therefore that the business-end users should be regarded as primary drivers. Rennolls and Al-Shawabhkeh indicate that this has to some extent already happened and the result has been the recent development of specific-business-problem analytics, presented as Business Intelligence (BI).

The primary questions about KDD/DM as an integrated and coherent generic discipline remain. Some suggested routes forward are discussed, but there is still much to be done if the original hopes and expectations for KDD/DM as a coherent and integrated discipline are to be fully realized.

Evgenii Vityaev and Keith Rennolls
Guest Editors

[2] http://cms1.gre.ac.uk/conferences/iufro/DEXA05_PMKD1/.

[3] http://cms1.gre.ac.uk/conferences/iufro/DEXA06_PMKD2/.

[4] Note that Vityaev and Kovalerchuk use the adjective Relational to describe their methodology, since it is both representational and relational, mathematically speaking.

[5] A DM-model is 'meaningful' if its truth value is invariant under the scale transformations appropriate to the variables in which the model is expressed.