Abstract. In this paper we propose the basic principles and philosophy for a collaborative methodology for performing data mining work. Such a methodology allows the data mining effort to be expended at very different locations communicating via a web-based tool. The aim of the methodology is to enable information and knowledge sharing, as well as the freedom to experiment with any problem solving technique. The data mining work methodology follows and extends the CRISP-DM methodology.

1 Introduction

The leading edge of European Data Mining expertise is currently spread over a number of Data Mining Units consisting of research laboratories and companies across Europe. Each one of these Data Mining Units has developed areas of specific expertise to solve particular data mining problems. Frequent on site collaboration, often required in solving data mining problems, is made difficult by geographical distance and costs. In this report, we propose both a Data Mining methodology and a system outline for remote collaborative data mining projects: RAMSYS1.

According to the proposed methodology, each Data Mining (DM) project will be developed by a number of Data Mining Units – or nodes – in a network of expertise. Each node represents a work site operated by an expert, an expert team, or a technical support team. All the participating nodes in the project will work as a team. Some network members form the Management Committee which has special obligations, including managing the interface with the client, defining project evaluation criteria, and performing solution selection.

1 RApid collaborative data Mining SYStem.
2 Guiding Principles

The RAMSYS Methodology is intended for distributed teams so that they may collaborate effectively on Data Mining projects. Physically, the elements of the team may be distant from each other, as on site collaboration may not always be possible. Remote collaboration, however, should be exploited to the maximum possible extent, taking advantage of the differing expertise of the members in the projects. The following principles guide the design of the methodology.

- **Light management.** Problem definition and objectives should be clear from the beginning of the project to all the participants. Consequently it is the task of the Management Committee for each problem to ensure that information flows within the network and that a good solution is provided. However, the Management Committee will not control directly the work of each team.

- **Start any time.** From time to time it may be necessary in solving a data-mining problem that extra expertise is necessary, or becomes available. Consequently all the problem information necessary to start problem solving should be available at all times. This includes problem definition, data, evaluation criteria and any problem specific knowledge already produced by project participants.

- **Stop any time.** Problem solving should be conducted by each team so that a working solution is available whenever the Management Committee issues a stop signal. One approach would be that participants follow a sort of “simplicity first” strategy (Holte 93) for modelling: simpler models are tried first, and more complex models are compared to the simpler ones.

- **Problem Solving Freedom.** Members of the network have varied expertise, techniques, and tools. It is hoped that these complement each other in the data mining process. Each team is in the best position to decide which approach to follow for the given problem. The Management Committee may give specific suggestions but doesn’t prescribe problem-solving approaches. In this way, experts are not constrained as to which techniques to employ. This principle is analogous to the ones of the so-called “open-source movement” [Cioffi 01].

- **Knowledge sharing.** As each modeller experiments and produces new knowledge on the problem, this should be immediately shared with all the participants on the data-mining problem.

- **Security.** The data and information relating to any data-mining problem is likely to contain sensitive information, which may not be revealed outside the project. The lightly managed network structure must ensure that access to information project information is strictly controlled and monitored.

- **Better Solutions.** Through the combination of different solutions arising from different nodes, it is envisaged that not only a range of solutions to the data

---

2 The main focus of the problem solving is the Modelling phase. The modes typically form the basis to a solution to the data mining problem (see section 6.3 for further discussion about modelling).
mining problem are produced, but also that the combination of the solutions – in some way – allows a better solution to be arrived at.

3 Problem Information

Central to the Data Mining Problem Solving process is Problem Information collection, development and maintenance. What is the vital information to be kept to enable and boost modelling? Where and how should all this information be kept? In the RAMSYS Methodology, Problem Information corresponds to the current best understanding of the problem. This consists of the problem definition, the data, hypotheses about the data and the models, and the current validity status of any such hypotheses. All this information is kept in the Information Vault.

Table 1: Information Vault contents.

- Problem Definition
  - The business understanding and the data mining view of the problem.
- Distilled Knowledge from Related problems
  - Similar problems addressed in the past, applied solutions, literature and produced knowledge.
- Evaluation Criteria Definition
  - How is the solution to be evaluated?
- Data
  - The collected database (DB₀).
  - Meta-data.
  - Quality of data (e.g. missing values, inconsistent records)
  - Derived data and information on how to repeat the derivation process from DB₀.
- Hypothesis Investment Account
  - Problem specific knowledge derived during problem solving.

The Hypothesis Investment Account

An hypothesis, in this context, is defined as an unproven statement about the problem, or the problem solution. The intention is that some of the hypotheses become corroborated and are taken to be “facts”. During the problem solving activity, team members may suggest hypotheses. These can be statements about the data like “we could benefit from having attribute X discretised by method B” or about the modelling as in “Applying clustering before building the decision trees improves accuracy”. Hypotheses like these should be clearly stated and made public that each hypothesis can then be actively refuted, or corroborated by any team member. The Hypothesis Investment Account (HIA) contains all the hypothesis threads for one
When the project is completed the HIA can be analysed to extract reusable DM problem solving knowledge in the form of lessons learnt.

Other operations can be added to the tracking mechanism: e.g. hypothesis refinement, hypothesis generalisation. The HIA is in some aspects similar to a bug tracking mechanism.

Table 2: Hypothesis Investment Account operations.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis statement</td>
<td><em>where someone can state an hypotheses, propose an idea, argue why it should be valid and perhaps suggest some experiments that might help determine the validity of the hypothesis.</em></td>
</tr>
<tr>
<td>Refutation</td>
<td><em>where someone may argue that a proposed hypotheses is false, provide experimental or theoretical results that refute the hypotheses.</em></td>
</tr>
<tr>
<td>Corroboration</td>
<td><em>where someone may provide evidence that supports the validity of an hypotheses.</em></td>
</tr>
<tr>
<td>Proof</td>
<td><em>If an hypotheses is proved true.</em></td>
</tr>
<tr>
<td>Refinement</td>
<td><em>a given hypotheses may be specialised.</em></td>
</tr>
<tr>
<td>Generalisation</td>
<td><em>a given hypotheses may be generalised.</em></td>
</tr>
</tbody>
</table>

Data

The database for any one data mining project does not remain static. Often an initially collected data base (DB₀) is subject to transformations suggested by the data modellers. At each moment, there is a current version of the database (DB_current) resulting from applying selected transformations to DB₀. The task of maintaining the database and applying the transformations is given to the Data Master. Suggested transformations are posted to the HIA as hypothesis relating the transformation with the expected result (improved accuracy, model simplification, etc.). Vigorous debate about the usefulness of any transformation will be encouraged, and potentially the transformation will be agreed amongst those in the network. At this time they are submitted to the Data Master to be applied to DB_current. All the modellers can then use the new database and test the validity of each hypothesis, or post new hypotheses based on the same transformation.

All data transformations will be applied in a monotonic manner, such that only the addition of data columns is allowed. This will ensure that the data used in any analyses, prior to the transformation, will be able to be re-run on the current database. Also, any analysis effort eventually spent on any version of the database is not invalidated by any transformation.
Data base transformations should be suggested in a formal language, such as SQL. Selected transformations are stored in the Information Vault encoded in the same language. It is assumed that modellers can then apply further transformations on the current database so that it can be exported as their new local version of the data.

One further important facet of data mastering and transformations is that all transformations will also be applied (as necessary) to the agreed evaluation data set. This way, modellers can access the evaluation data set having the same structure as the database used for generating the modellers. This should reduce the effort required for both the modellers, and the evaluation of models.

The Data Master is also in charge of maintaining the meta-data and any relevant information about the collected data.

4 Restrictions on Problem Information flow

The above section suggests some restrictions on Problem Information flow that come as a consequence of the Guiding Principles of the Methodology. We summarise those restrictions as follows:

- **Data is stored in one node of the project network.** Data storing and maintenance should be centralised in the sense that only one node has the official version of the data set. Data can be mirrored in other nodes for efficiency reasons, but should always have the data storage node as reference. The person in charge of maintaining data at the data node is the *Data Master*.

- **All participants have remote access to the data and transformations to it at any time.** This is a consequence of *start any time principle*. Any authorised participant should have, at any time, access to the best current understanding of the problem, including, problem definition, related work, data and, new findings within the project.

- **Data transfer operations should ensure security to the desired level.** Data must be safe from sniffers or other network spies. This might be achieved by using encryption, or virtual private networks.

- **The Problem definition should be always available to all participants.** Also a consequence of the start any time principle.

- **The Model Evaluation Criteria should be clearly stated.** Modellers must be aware of how models will be assessed and selected.

- **Participants should share problem Specific Knowledge (PSK).**

- **Shared PSK is maintained at a central point.** Similarly to data, it is important that the knowledge about the problem being solved is readily available at some point of the network. This is ensured by centralising (physically or virtually) this information.

- **Shared PSK should be available at any time.** Also a consequence of the start any time principle.
5 The Work Model

Data mastering is centralised. Having only one source of data, secure and permanently available, ensures that modelling can start at any time and prevents unnecessary replication of work. The modelling – or problem solving effort -- however, should not be centralised. The space of applicable models is far too vast to be completely explored by one node’s or one expert’s favourite techniques. One of the proposed strengths of the RAMSYS model is the relative freedom given to modellers in the project to try their own approaches. This sort of freedom is sympathetic to university oriented experts, but can also be productive for company developers. The obvious difficulty is that a project has a client – or customer – and a solution must be provided. This requires the participants to schedule their individual and combined efforts. How can an unstructured network of modellers be controlled? The answer may lie in a mixture of modelling challenge, and modellers’ collaboration.

Data analysis challenges are becoming more and more popular in the scientific community. International congresses (KDD99, KDD2000, PKDD2000), and at least one network of excellence (COIL challenge) organise scientific contests that motivate hundreds of researchers and technology users from different fields and backgrounds to participate – both from industry and academia. The result of this diversity is that very good models appear, and frequently science takes a leap forward. The working assumption is that “open competition enhances the quality of the produced solutions”. In the RAMSYS methodology modellers have access to same Problem Information. They can communicate and learn with each other by utilising the Hypothesis Investment Account. We also imagine less structured discussion forums for modellers. A Management Committee, formed by some elements of the network, is responsible for managing the interface with the Client and scheduling delivery. This Committee is then responsible for setting up the challenge relative to the Data Mining project in hand: defining the success criteria, receiving and selecting submissions. Modellers can compete with each other, but they can also collaborate in the traditional way. Collaboration may be suggested by the Management Committee or freely engaged by competing nodes.

Modes of collaboration

Early in the SolEuNet Project Ashwin Srinivasan identified the following three models of working:

- **Co-operate**: where groups with complementary skills analyse a problem jointly.
- **Challenge**: where groups analyse a problem competing with each other.
- **Single**: a single group analyses a problem getting advice from other groups.
The cooperative model has the advantage of having the problem analysed by different groups with different expertise and skills. Distance, however, makes it difficult for sharing data and ideas. Management is also difficult. Tasks have to be assigned and scheduled. Coordination is necessary.

The challenge mode has the advantage of requiring little communication: ideas don’t have to be shared (although they can be shared). Interdependency between groups is small or nonexistent. Another advantage is that many autonomous solutions are proposed. Evaluation criteria are defined up-front what makes the objectives clearer from the start. Problems are that the little sharing of ideas may under-exploit the potential of the groups. The challenge still has to be managed, although management effort is perceived to be much less than with the cooperative model.

The single model is by far the easiest one to manage. However, it is also the one that provides the least number of solutions.

A fourth possible mode of collaboration is:

- A Blend of Co-operate and Challenge.

In this mode the problem is analysed by many groups that work independently and competing with each other but that have means of sharing ideas. This way, all data modellers share the current best understanding of the data mining problem. To be successful, the blend work mode requires a support system and a methodology that promotes and allows rapid information transfer and remote access to problem information. The support system also aids in the evaluation and selection of the most appropriate models for the construction of the final solution.

One advantage of the blend mode with support is that it clearly separates modelling and evaluation. Modellers benefit of central data management, but have to conform to the defined evaluation criteria. The disadvantages of blend with support are that it demands better scientific practice by the modellers, knowledge produced during problem solving must be shared in real time and it requires a support system.

In the following sections we describe the proposed support methodology and system.

The Methodology: refining CRISP-DM

The data mining workflow --from data and problem to model and solution -- has been presented many times in the literature (Fayyad et al.; Adriaans et al.; Dialogis). In this section we review the CRISP-DM methodology and propose necessary refinements for the RAMSYS approach.

The CRISP-DM methodology identifies six phases in a DM project:

- Business Understanding.
- Data Understanding.

---

Each of the phases is divided into a number of generic tasks. The RAMSYS methodology can be regarded as a refinement of CRISP-DM, where some of the generic tasks have to be carried out in a collaborative mode while accounting for the possibility of the 'remote' restriction.

**Business Understanding and Deployment**

The first and last phases, Business Understanding and Deployment, suffer no additions under the RAMSYS methodology. These have to be done on site, and in the traditional way, by an expert or team of experts. In Table 3 and Table 4 we give a brief outline of these two phases.

**Table 3: Business Understanding phase: generic tasks and outputs (in italics).**

- Determine business objectives
  - *Background*
  - *Business objectives,*
  - *Business success criteria*
- Assess situation
  - *Inventory of resources*
  - *Requirements assumptions and constraints*
  - *Risks and contingency*
  - *Terminology*
  - *Costs and benefits*
- Determine DM goals
  - *DM goals,*
  - *DM success criteria*
- Produce project plan
  - *Project plan*
  - *Initial assessment of tools and technologies*

The Business Understanding phase contributes to the Information Vault with general information about the business that can be condensed in one report, problem definition (DM goals) and success criteria.

**Table 4: Deployment phase: generic tasks and outputs (in italics).**

- Plan deployment
  - *Deployment plan*
- Plan monitoring and maintenance
  - *Monitoring and maintenance plan*
Data Understanding and Data Preparation

From the Data Understanding phase the meta-data and information about data quality is produced. The Data Exploration task of this phase may result in the statement of some hypotheses. These can be considered in the Data Preparation phase that follows.

Table 5: Data Understanding phase: generic tasks and outputs (in italics).

- Collect initial data: *Initial data collection report*
- Describe data: *Data description report*
- Explore data: *Data exploration report*
- Verify data quality: *Data quality report*

The initial data set (DB0) is built in the Data Preparation phase. Data selection and derivation should already be tracked as hypotheses that can be validated at a later stage. At the same time, data transformation operations should be recorded in an executable form as detailed in section 0.

Table 6: Data Preparation phase: generic tasks and outputs (in italics).

- Data set
  - Data set description
- Select data
  - Rationale for inclusion/exclusion
- Clean data
  - Data cleaning report
- Construct Data
  - Derived attributes
  - Generated records
- Integrate data
  - Merged data
- Format data
  - Reformatted data

Modelling

When there is enough Information in the Vault, modellers can start their work. The “start anytime principle” is enforced by ensuring that the vault contains the current
shared view on the problem and its data. Modellers should bear in mind the evaluation criteria, and aim to use a simplicity-first approach [Holte 93]. This enables that the Management Committee can, after an initial period, ask for the current best models from each of the nodes, evaluate them, and deliver them if appropriate. This is what we call the “stop anytime principle”.

**Table 7:** Modelling phase: generic tasks and outputs (in italics).

- Select Modelling Technique
  - Modelling technique
  - Modelling assumptions
- Generate test design
  - Test design
- Build model
  - Parameter settings
  - Models
  - Model description
- Assess model
  - Model assessment
  - Revised parameter setting

As discussed in section 5.1, the Work model will typically be a blend of Collaborative and Challenge modes. However, the RAMSYS methodology can still be followed in pure Collaborative, Challenge or Single mode of work.

The generic tasks of the Modelling phase may be carried out within each node of the network, or within clusters of collaborating nodes. Generated hypothesis can be stated in the HIA. During Modelling, modellers can suggest new data transformations to the data master. After acceptance, these transformations are performed and stored into the Information Vault.

Modelling stops when the Management Committee issues a stop signal. After that, modellers submit their best models for selection by the Committee. The submission step is only useful if there is a challenge component in the work model.

**Table 8:** Modelling phase: additions to the CRISP-DM methodology.

- Model Submission
  - Submission in the required form

**Evaluation**

All submitted models must be subject to an evaluation and selection process. Selection is done by the Management Committee according to the previously defined evaluation criteria. It is assumed that the Management Committee performs some evaluation tests only after modelling is done. This prevents test-fitting modelling.

**Table 9:** Evaluation phase: generic tasks and outputs (in italics).
• Evaluate results
  • Assessment of DM with respect to the Business success criteria
  • Approved models
• Review process
  • Review of process
• Determine next steps
  • List of possible actions
  • Decision

6 The system

The RAMSYS methodology will be supported by a number of tools referred to as RAMSYS (the system). Such tools are described below.

• The Data Master support tool: used by the Data Master to maintain the current version of the database, transformations and information on the data such as meta-data and data quality information. Other problem information (except the Hypothesis Investment Account) should also be stored here. The tool also serves modellers that have access to the data via http or ftp or secure versions of these protocols. Modellers can also submit data base transformations or new problem information.

• The Hypothesis Investment Account and modeller communication support tool: used by modellers via web. It might work in a way somewhat similar to a bug tracking tool (Openticket). Other types of communication between modellers can also be provided (e.g. discussion forums).

• The model submission and model evaluation support tool: used by the Management Committee to store submitted models, evaluate them and select one (or more). Modellers can submit models via web. The tool could also be extended to assist modellers in model evaluation before submission.

7 Conclusion

The proposed methodology intends to allow the collaborative work of remotely placed data miners in a disciplined manner in what respects the flow of information while allowing the free flow of ideas for problem solving. The discipline is achieved by following the CRISP-DM methodology and by forcing/strongly inviting collaborators to publish their ideas among each other. However some other problems arise that may compromise some of the basic principles. Important questions are:

• Rationality: Is it rational to adopt the challenge/blend mode? Isn’t it always more efficient to plan first, assign tasks and work in a classical team mode?
• **Productivity**: The challenge mode of working may need more experts than the collaborative mode. This may decrease productivity (resources/output), but it may also lead to better quality output -- solutions.

• **Credit assignment**: How is credit assigned when several experts are involved in one project? They are all authors of the solution even if they have not contributed to any part of the final chosen solution? How, then, should any payment be allocated?

• **Delivery guaranteed**: What happens if no node submits a solution? Should there be an assigned default node, with the special responsibility of providing a solution on schedule, which will compete with other solutions if they come?

• **Security**: how can security of such a web-intensive process be guaranteed?

Some of these issues are already being addressed in the first proposed web-based implementation of RAMSYS, and projects are already being carried out under RAMSYS principles [Voss et al. 01 (LSO), Voss et al. 01 (IDDM)]. We believe this on-going work will contribute to the refinement and clarification of some of the problems while allowing the development of a real usable collaborative remote data mining system.

**References**

6. Openticket: [http://openticket.point-one.net](http://openticket.point-one.net)