Collaborative Data Mining and Data Exchange: A Case Study

Olga Štěpánková, Jiří Kléma, Petr Mikšovský

Department of Cybernetics, CTU Prague,
Technicka 2, 166 27 Prague 6, Czech Republic,
{step,klema,miksovsp}@labe.felk.cvut.cz

Abstract. The paper wraps up our experience gained during a collaborative data mining project solved using RAMSYS methodology. We pay special attention to some difficulties, which have appeared during the collaborative data mining and we try to identify their reasons. Finally, we raise several suggestions how to ensure efficient function of organizational memory necessary to support concise and transparent information exchange among all participating partners.

1 Introduction

The aim of data mining is to extract non-trivial and actionable knowledge from large or very large databases [10]. No one doubts that data mining is a very complex activity. Many different technologies have to be combined to accomplish its goal – experience is needed in handling large databases, in usage of various machine learning techniques, in statistics, etc. An institution handling a data mining project has to establish a team consisting of several specialists, usually. If the institution is a virtual one [6], i.e. its members do not share the location of their office but their relation is expressed in more subtle or indirect way, it can easily happen, that the members of the team are distributed on very distant places all over the world and connected by Internet, only. Is it possible to achieve useful collaboration on a data mining project even under these conditions?

This question is crucial for the SolEuNet project [11], which aims to develop virtual enterprise producing data mining and decision support services. The SolEuNet project partners belong both to academic institutions and to companies across Europe. Each data mining process follows the CRISP-DM principles [4] dividing the DM process into six interrelated phases: Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment. Moreover, efficient cooperation among the partners has to be ensured. To minimize management and coordination efforts RAMSYS methodology [7] has been designed. It is based on the following six principles: *"Light management, Start and stop at any time, Problem solving freedom, Knowledge sharing* and *Security.*" In SolEuNet, the groupware system called ZENO [9] was used to implement a tool supporting RAMSYS methodology. This paper wraps up experience gained when using this tool for the Spa DM project.

2 SPA Data Mining Project: Resource Allocation in Spa Facilities

A company running a spa facility offers a rich set of health procedures to heal medical problems of the patients who are arriving into the health farm for a restricted period. Obviously, each patient is supposed to obtain an individual treatment, i.e. a set of procedures assigned to him/her by the spa physician, who bases his recommendation on results of careful inspection of the patient upon his arrival. But the written recommendation of the spa physician is not enough to ensure that the patient really gets exactly the recommended procedures. The second necessary condition is that necessary resources (personnel/technical equipment) are available in appropriate quantity.

All over the fact that the groups of patients occupying the spa are changing frequently, the spa aims to be able to ensure the appropriate individual treatment for each of its patients. How can such a goal be achieved? It is vital for the spa administration to know in advance the total requirements of a group of patients for all individual procedures offered by the spa. That is why the company running the spa administrative system decided to start a data mining project (SPA Project) to be solved within the SolEuNet Project using the available history data as the basic data source.

2.1 SPA Data Mining Task and CRISP Phases

The SPA project happened to be one of the first collaborative DM projects where RAMSYS methodology could be verified. Four teams (CTU, KUL, BRI and LIACC) took part in this exercise – the first three started almost simultaneously, while LIACC joined 6 months later. These teams used 2 basic communication channels: e-mail and ZENO. ZENO played a role of organizational memory – place, where intermediate products of CRISP phases achieved by the partner teams were made public to be shared with all the participating partners. Some of these results were used later by another participating team – in this way collaborative DM really took place. This happened e.g. in the case of **data preprocessing**: its significant part was ensured by the CTU team using the data preprocessing tool SumatraTT [2].

In the **modeling phase**, all the participating teams decided to approach a prediction goal on their own using different ML tools. The intermediate results have been published on ZENO and this on-line information source highly supported competition among the partners. Two modeling directions have been considered in the project:

- *The individual centered approach* starts by predicting all procedures to be prescribed to a single patient. The total for one week is obtained as a sum of predictions for individual patients actually present.
- *The aggregated approach* tries to predict usage of resources for a whole group of patients at once.

Table 1 provides condensed **evaluation** of the results obtained by the participating teams. Only after a careful inspection we have found out that objectivity of the result comparison is hindered by the fact that the DM goal has been slightly modified or

improved by some of the teams. Modification of a DM goal is understood as a natural part of the CRISP methodology, which counts with loops. Any DM process has to be understood as pursuit of a moving target – to be successful one has to modify his/her goal according to the obtained results. At the present state of the RAMSYS implementation, any information concerning the goals considered by different participating teams had to be mined from the text reports provided by these teams with significant efforts. To overcome this problem we suggest simple refinement of RAMSYS structure in the next section 2.2.

Let us illustrate the upper mentioned claim on the SPA example: at early stages of business understanding phase it was not discussed how far in advance the prediction has to be generated. Everyone could thus assume that the prediction for the week no. W has to be available just at the beginning of that week. This timing gives a chance to use information contained in the data collected during the previous week ($W - 1$) to improve the prediction accuracy for the week W. Later on it was revealed by the domain expert that the prediction should be available at least four weeks in advance. The team, which joined the problem solving party as the last one, omitted this information due to the fact it was buried deep among other less important details. Thus they have been solving a modified goal without making this explicit - this finally caused certain incompatibility of the results obtained by different teams.

Table 1. The results reached by the individual teams. The individual cells show the total number of procedures (max. 35) that satisfy the specified condition. The first column shows how often the result exceeds the customer's required relative error (RE). The second column shows number of procedures for which the given team delivered the best result. The last column shows for how many times the given team was "close" to the best one.

Team	RE>20%	Best	RE-RE(best)<5%
BRI	5	2	18
KUL	12	1	14
LIACC	2	29	33
CTU	9	3	19

2.2 RAMSYS Principles and Organizational Memory

The main aim of the six RAMSYS principles (*Light management, Start any time, Stop at any time, Problem solving freedom, Knowledge sharing* and *Security*) is to ensure cooperative environment supporting competition and creativity of partners while not restricting their academic freedom. How did we succeed to follow these principles during the SPA experiment? Most of the RAMSYS principles caused no problems. The only exception is the **knowledge sharing** principle. It proved to be the most difficult part in the considered project. All the partners did their best to provide all necessary information, everybody made available his/her results (e.g. ideas, evaluation results) as well as the modified, extended or transformed datasets on ZENO. But sometimes the provided description happened to be difficult to follow and the rest of partners preferred not to rely on the results of others but did most of the work on their own. We believe that one of the reasons is lack of tools and standards

supporting this aspect of DM process. The organizational memory has to have transparent structure, which ensures direct access to knowledge characterizing the considered task. The core knowledge to be shared has to specify precise description of:

1. all the considered DM goals (new goals are being suggested often during the course of the project, the domain expert can assign them with his priorities etc.),
2. scripts designed for preprocessing the treated data.

The first item can be handled easily if the structure of ZENO for RAMSYS is extended by a direct access to a place to present all the considered DM goals. This **Review of DM goals** could be situated e.g. in Business understanding section. Each goal has to be presented in a clear structured way including a unique name, time the goal was defined, set of its considered input attributes (and used preprocessing), evaluation of the goal importance given by the domain expert, as well as the list of all generated models. Appropriate extensions of PMML could be applied for that purpose. The next section is devoted to problems of knowledge sharing in the data preprocessing tasks.

3 Centralized Support and RAMSYS

Collaborative data mining has to be situated in distributed problem solving environment supporting sharing and re-use of resource-intensive results. Processes involved in data transformation and model evaluation certainly belong among resource-intensive activities. If the collaborating subjects want to share knowledge it has to be expressed in a form understandable to all of them. It is very important to ensure standardization of knowledge to be exchanged. While PMML becomes a standard data mining model representation, the standard description of data pre-processing tasks is still missing. This is a serious drawback: if we want to support *start-any-time* and *stop-any-time* RAMSYS principles we necessarily have to be able to reconstruct problem-solving path of any of collaborating subjects.

3.1 Data Transformation Standard

Knowledge sharing would be significantly improved if all data transformations applied in a certain DM or DSS problem are described in a single format shared by all partners. Let us call it **Data Transformation Mark-up Language** (DTrML). Its obvious advantage is transparent description of all data modifications and transformations applied when solving the given task. Moreover, as soon as DTrML becomes operational, we get a powerful universal data pre-processing tool, which can ensure the centralized data transformation phase of collaborative DM (see Fig. 2B). No doubt, such a centralization results in efficient use of available resources (time and computational capacity).

We believe that design of SumatraTT [2], a data pre-processing tool developed at CTU, contains the basic components of DTrML including the metadata concept for data transformation description [1]. In this way, it integrates descriptive metadata

with the operational ones. The metadata is stored in XML format (similarly as PMML), which can help in the case of conversion into another format. Moreover, the resulting kind of standardization brings, similarly as Java does, self-documenting effect which simplifies human understanding. To support the last claim let us describe the Sumatra-based data transformation. There are input and output data sources the structure and location of which is described in metadata. Then a certain sequence of transformation templates is applied. Every transformation template consists of a documentation describing what input/output is expected and how to set-up parameters – it is not usable without such documentation.

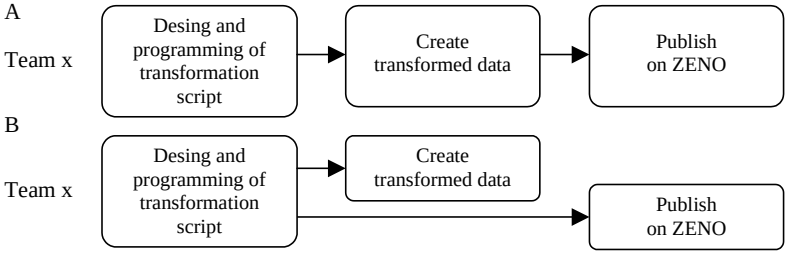

Fig. 2. Two options for exchange of data transformation results. The case A represents situation where each collaborating subject prepares data by a different data transformation tool. It means that they can share/publish transformed data sets only. Whereas in the case B the standardized data transformation script can be shared/published. It is not only more operational but besides others it saves disk space on ZENO.

3.2 Centralized Model Evaluation

The SPA experiment has been reviewed in [3] from the point of view of model evaluation, which authors divide into 5 sequential steps: tune, build, predict, evaluate and publish. The authors have noted number of problems: "*... the evaluation criterion may change, people tent to report the measures that their tools compute but are reluctant to implement measures themselves ...*" and they present centralized model evaluation as the remedy for these problems. The centralized model evaluation can appear in 4 options depending on the specific point in the tune ... publish sequence when the centralized approach is started. In the option 1, the only centralized step is the last one "publish" – this evaluation mode was applied in the SPA experiment and it proved to cause some misunderstandings among the cooperating subjects. That is why more complex options seem preferable. The authors conclude "*In the longer run, under assumption that PMML is general enough to describe any kind of model that could be submitted and that interpreters are available, it seems desirable to shift to Option 3 (which ensures the prediction on test data centrally)*." On the first glance, this mode seems simple: the collaborating subjects send their working solution in the form of the PMML description to a single node which ensures centrally their use for prediction, evaluation and finally publish the results. This sounds easy until we realize that most often the models created as the result of DM do not rely directly on the original data. They use data, which are pre-processed, aggregated, complemented by new derived attributes, transformed to fit the requirements of the model and of the

DM task. Consequently, the working solution must include DTrML part beside the PMML parts. That is why development of DTrML has to be considered as an integral part of the plan to incorporate centralized model evaluation process into RAMSYS.

4 Conclusions

Our experience in the SPA project pointed to the problems of knowledge sharing in all steps of the collaborative DM process. Existence of tools and standards, which support organizational memory seem to be a prerequisite for reaching the expected effects of collaborative data mining and efficient collaboration. We are suggesting additional tools and standards which range from a simple knowledge structure designed in the Section 2.2 up to very ambitious goal to standardize language used for description of data transformations suggested in 3. Introduction of DTrML can have significant impact on re-use of the obtained results as it simplifies creation of a universal functional repository of solved cases containing solutions which can be reconstructed or re-applied on new data at any time. The suggested standardization will simplify transfer of DM results into the practice.

5 References

1. Aubrecht, P., Kouba, Z.: Meta-Data Driven Data Transformation. Proc. of the 5th World Muti-conference on Systemics, Cybernetics and Informatics, 2001.
2. Aubrecht, P., Zelezny, F., Miksovsky, P., Stepankova, O.: SumatraTT: Towards a Universal Data Preprocessor. Proc. of the 16th European Meeting on Cybernetics and Systems Research – Vienna 2002, pp. 818-823, Austrian Society for Cybernetic Study.
3. Blockeel, H., Moyle, S.: Collaborative data mining needs centralized model evaluation, submitted to DMLL-2002 at ICML-2002
4. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R.: CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM consortium, 2000.
5. Jorge, A., Moyle, S., Richter G., Voß, A.: Remote Collaborative Data Mining Through Online Knowledge Sharing, submitted to PROVE-2002.
6. Lavrac, N., Urbancic, T., Orel, A.: Virtual Enterprise For Data Mining And Decision Support: A Model For Networking Academia And Business, submitted to PROVE-2002.
7. Moyle, S., Jorge, A.: RAMSYS – A Methodology for Supporting Rapid Remote Collaborative Data Mining Projects, IDDM'01, ECML/PKDD Workshop notes, 2001.
8. Stepankova, O., Lauryn, S., Aubrecht, P., Klema, J., Miksovsky, P., Novakova, L., Palous, J.: Data Mining for Resource Allocation: A Case Study. In: Intelligent Methods for Quality Improvement in Practice, Prague, CTU FEE, Department of Cybernetics, pp. 94-105, 2002.
9. Voss, A., Gartner, T., Moyle, S.: Zeno for Rapid Collaboration in Data Mining Projects. Proceedings of the ECML/PKDD Workshop on Integration of Data Mining, Decision Support and Meta-Learning (pp. 43-51). ECML/PKDD'01 Workshop notes, 2001.
10. Witten, I., Frank, E.: Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.
11. Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise, SolEuNet pages available at http://soleunet.ijs.si/.