Chapter 18

COLLABORATIVE DATA MINING WITH RAMSYS AND SUMATRA TT

Prediction of resources for a health farm

Olga Štěpánková, Jiří Kléma, and Petr Mikšovský

Abstract: This chapter catalogs the experience gained during a collaborative data mining project solved using the RAMSYS methodology. The data mining project aimed to produce a system for planning the allocation of resources in a spa (health farm). The chapter discusses and describes how past data can be used as a source for data mining leading to the discovery of models useful for the prediction of resource requirements. Data preprocessing using the SumatraTT tool is emphasized. Difficulties which appeared during the collaborative data mining process are highlighted, and their reasons are identified. The chapter concludes with several suggestions for effective knowledge management supporting concise and transparent information exchange among all participating partners.

1. INTRODUCTION

Spa facilities offer various health procedures to treat medical problems of patients who attend the health farm for a treatment in a given time period. Each patient obtains individual treatment – a set of procedures, assigned to the patient by the spa physician. The physician's recommendation is based on results of a careful examination of the patient upon his/her arrival. However, this recommendation alone cannot ensure that the patient will indeed receive the recommended procedures; this could be achieved only if all the necessary resources of the spa were available in the sufficient quantities for the patient's treatment (human resources in terms of skilled personnel, as well as technical resources such as a bath tube or diathermia).

The health farm aims to provide appropriate individual treatment for each of its patients. It is vital for the spa administration to know in advance (before a new group of patients arrives) what the total amount of required procedures in the considered period will be. Such prediction of resource requirements cannot be based
on data from patient’s medical history, as it is not available to the spa administration before the patient enters the health farm. The only patient data available for the prediction process is the data the spa administration receives as a part of the patient’s application for the treatment (e.g., sex, age, planned length of stay, type of disorder, etc.). Lauryn v. o. s., a company developing information systems for health farms – referred to as the SPA company – and the SolEuNet project partners agreed to cooperate on a data mining project (abbreviated as the SPA project in the rest of this chapter) with the following objectives:

- to study the possibility of predicting, a few weeks in advance, the overall number of prescriptions for specific health procedures within a specific time period,
- to identify previously unknown groups of clients (e.g., women 50 to 60 years old having the same disorders) exhibiting characteristic behaviors or requirements for procedures; for each group, predict the corresponding treatment (the set of all procedures).

The SPA project searched for non-trivial and actionable knowledge from a large database using the CRISP-DM methodology (Chapman, et al., 2000) in the context of the SolEuNet consortium of collaborating teams that are distributed across distant places in Europe and connected only by the Internet (see Chapter 5). To support the remote collaborative data mining approach, a prototype implementation of a system for supporting collaborative data mining, ZENO-for-RAMSYS, was introduced. The supporting methodology RAMSYS (see Chapter 6 or (Moyle, et al., 2000)) was implemented using a groupware system called ZENO (Voß, et al., 2001). The SPA project served as a test bed for the development of the collaborative data mining methodology. The following sections discuss the process of the SPA project solution in terms of the methodologies mentioned above.

2. CRISP-DM PHASES OF THE SPA PROJECT

Four remote teams participated in this data mining problem; they will be denoted as T1, ..., T4 in the rest of this chapter. Team T1, residing in the same country as the SPA company, maintained contact with the domain expert within the SPA company. This team is referred to as the contacting team.

The CRISP-DM methodology defines the data mining process as consisting of six interrelated phases (see Chapters 1 and 6). The initial phases of Business and Data Understanding and some of the Data Preparation phase are usually carried out in close cooperation with a data owner/provider. For the SolEuNet data mining project only one of the collaborating teams, namely the contacting team, was in personal touch with the SPA company domain expert. Consequently, this team was mainly responsible for conveying the information gained in these early data mining phases to the other teams. This was done by sharing all the information on ZENO-for-RAMSYS in the form of reports. The three other teams contributed data mining results by publishing their own data analyses and visualizations. A detailed report on the process of solving the SPA project appeared in (Štěpánková, et al., 2002a, Štěpánková, et al., 2002b), reviewing the results obtained by all partners as well as the time of their delivery.