Chapter XVIII
Data Mining in Atherosclerosis Risk Factor Data

Petr Berka
University of Economics, Prague, Czech Republic;
Academy of Sciences of the Czech Republic, Prague, Czech Republic

Jan Rauch
University of Economics, Prague, Czech Republic;
Academy of Sciences of the Czech Republic, Prague, Czech Republic

Marie Tomečková
Academy of Sciences of the Czech Republic, Prague, Czech Republic

ABSTRACT

The aim of this chapter is to describe goals, current results, and further plans of long-time activity concerning application of data mining and machine learning methods to the complex medical data set. The analyzed data set concerns a longitudinal study of atherosclerosis risk factors. The structure and main features of this data set, as well as methodology of observation of risk factors, are introduced. The important first steps of analysis of atherosclerosis data are described in details together with a large set of analytical questions defined on the basis of first results. Experience in solving these tasks is summarized and further directions of analysis are outlined.

INTRODUCTION

Atherosclerosis is a slow, complex disease that typically starts in childhood and often progresses when people grow older. In some people it progresses rapidly, even in their third decade. Many scientists think it begins with damage to the innermost layer of the artery. Atherosclerosis involves the slow buildup of deposits of fatty substances, cholesterol, body cellular waste products, calcium, and fibrin (a clotting

material in the blood) in the inside lining of an artery. The buildup (referred as a plaque) with the formation of the blood clot (thrombus) on the surface of the plaque can partially or totally block the flow of blood through the artery. If either of these events occurs and blocks the entire artery, a heart attack or stroke or other life-threatening events may result.

People with a family history of premature cardiovascular disease (CVD) and with other risk factors of atherosclerosis have an increased risk of the complications of atherosclerosis. Research shows the benefits of reducing the controllable risk factors for atherosclerosis: high blood cholesterol, cigarette smoking and exposure to tobacco smoke, high blood pressure, diabetes mellitus, obesity, physical inactivity.

Atherosclerosis-related diseases are a leading cause of death and impairment in the United States, affecting over 60 million people. Additionally, 50% of Americans have levels of cholesterol that place them at high risk for developing coronary artery disease. Similar situation can be observed in other countries. So the education of patients about prevention of atherosclerosis is very important.

In the early seventies of the twentieth century, a project of extensive epidemiological study of atherosclerosis primary prevention was developed under the name National Preventive Multifactorial Study of Hard Attacks and Strokes in the former Czechoslovakia. The aims of the study were:

- Identify atherosclerosis risk factors prevalence in a population generally considered to be the most endangered by possible atherosclerosis complications, i.e. middle aged men.
- Follow the development of these risk factors and their impact on the examined men health, especially with respect to atherosclerotic cardiovascular diseases.
- Study the impact of complex risk factors intervention on their development and cardiovascular morbidity and mortality.
- 10–12 years into the study, compare risk factors profile and health of the selected men, who originally did not show any atherosclerosis risk factors with a group of men showing risk factors from the beginning of the study.

Men born between 1926 and 1937 living in centre of the capital of the Czechoslovakia -Prague - were selected from election lists in year 1975. The invitation for examination included a short explanation of the first examination purpose, procedure and later observations and asked for co-operation. At that time, no informed signature of the respondent was required. Entry examinations were performed in the years 1976–1979 and 1,419 out of 2,370 invited men came for the first examination and risk factors of atherosclerosis were classified according to the well defined methodology. The primary data covers both entry and control examination. 244 attributes have been surveyed with each patient at entry examination and there are 219 attributes, which values are codes or results of size measurements of different variables. 10,610 control examination were further made, each examination concerns 66 attributes. Some additional irregular data collections concerning these men were performed. Study is named STULONG – LONGitudinal STUdy - and continues for twenty years.

The observation resulted into data set consisting of four data matrices that are suitable for application of both classical statistical data analysis method and for application of data mining and machine learning. The project to analyze these data by methods of data mining started by setting large set of analytical questions. The goal of this chapter is to describe first steps in application of data mining and machine learning methods to the STULONG data. We also summarize the additional analyzes inspired by set of analytical questions and we introduce further planned work.