JSAI KDD Challenge 2001 (JKDD01) in WEKA
4IZ451 - Knowledge discovery in databases

Tomáš Maršálek

December 27, 2014

1 Introduction

Given a set of medical data of meningoencephalitis diagnoses we try to discover predictive rules which could be used by domain experts to find out the cause of the illness in various stages of diagnosis process (early symptoms, physical examination and after laboratory results). We are also interested to find similar rules to find out the culture of bacteria if the bacteria were the cause in the first place. Lastly we are interested in prognosis of the patient based on mentioned three stages of diagnosis process.

The analysis is performed using CRISP-DM methodology, which is a industry standard method for data mining designed to perform vast data mining tasks faster and more effectively while avoiding basic mistakes.

2 CRISP-DM

CRISP-DM (stands for CRoss Industry Standard Process for Data Mining) consists of six stages from goal definition to results interpretation and deployment of resulting model. The stages are:

1. **Business understanding** - The problem should be sufficiently understood so that it even makes sense to define goals
2. **Data understanding** - Data should be understood so that we understand meaning and quality of it
3. **Data preparation** - Models operate on data in a certain form. We preprocess the data to make the models digest the data properly and make the modeling as efficient as possible
4. **Modeling** - Various algorithms are performed to classify, segment, cluster, ... preprocessed data
5. **Evaluation** - Review the process done so far from step one because after obtaining more knowledge about the problem spending time working on it we should go back and see if we might have missed something or do something differently
6. **Deployment** - Presenting result of the analysis to the client if the goal was only to perform an analysis or implement the model programmatically to be used repeatedly by the client.

3 Goals

The data mining challenge asks the analysts to find factors for

1. diagnosis
2. detection of bacteria or virus and culture of bacteria
3. prognosis of the patient

4 Data

Data set consists of 140 cases of patients with finding of severe inflammation in dura mater (covering membrane of a brain).

The cases are described by 38 attributes from which we try to predict diagnosis **DIAG** and grouped diagnosis **DIAG2** for the first goal of analysis. For the second goal we are want to predict attributes **CULT_FIND** and **CULTURE** and to predict the prognosis of a patient we will use attributes **C_COURSE** and **COURSE**.

Attributes known before physical examination - early symptoms

attribute	**explanation**
COLD	Days of symptoms of common cold
HEADACHE	Days of symptoms of headache
FEVER	Days of symptoms of fever
NAUSEA	Days of nausea
LOC	Days since loss of consciousness occured
SEIZURE	Days since convulsion or epilepsy observed
ONSET	-

Attributes known after physical examination

attribute	explanation
BT	Body temperature
STIFF	Neck stiffness
KERNIG	Kernig sign
LASEGUE	Lasegue sign
GCS	Glasgow Coma Scale
LOC_DAT	Grouped loss of consciousness

Attributes known after laboratory tests

attribute	explanation
WBC	White Blood Cell Count
CRP	C-Reactive Protein
ESR	Blood Sedimentation Test
CT_FIND	Grouped CT Findings
EEG_WAVE	Grouped EEG Wave findings
EEG_FOCUS	Focal sign in EEG
CSF_CELL	Cell count in cerebulospinal fluid
Cell_Poly	Polynuclear cell count in CSF
Cell_Mono	Mononuclear cell count in CSF
CSF_GLU	Glucose in CSF
CULT_FIND	Whether bacteria or virus is specified or not
CULTURE	The name of bacteria or virus

4.1 Data preprocessing

Proprocessing steps for this analysis included finding columns with unknown values and converting them into values recognizable by WEKA as unknowns.

Analysis is performed in stages as depicted in tables of attributes above. For each stage only attributes known in this or earlier stages are used to predict attributes of following stages or goals. Therefore we end up with four attribute-filtered data sets.

5 Modeling

WEKA of University of Waikato was used as a data mining tool for this analysis for its simplicity and powerfulness.

Attributes after beginning of treatment

attribute	**explanation**
THERAPY2	categorical type of therapy determined after diagnosis
CSF_CELL3	CSF after 3 days of treatment
CSF_CELL7	CSF after 7 days of treatment
C_COURSE	Clinical course at discharge. Symptoms after discharge
COURSE	Grouped C_COURSE. If symptoms found or not
RISK	Risk factor