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

Tomáš Maršálek

December 29, 2014

1 Introduction

Purpose of JSAI KDD Challenge is to gather many data analysts and collectively try to discover knowledge in a single data set[2]. Data set for challenge of year 2001 was provided by a medical doctor Prof. Shusaku Tsumoto (Shimane Medical Univ.), who is the domain expert on the meningoencephalitis diagnosis and on the supervisory board of this challenge.

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 (which is symbol ?).

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 (early_symptoms, examination, laboratory,

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

treatment).

5 Modeling

WEKA of University of Waikato was used as a data mining tool for this analysis for its simplicity, powerfulness and vast set of implemented classification algorithms. Outstanding performance of algorithms is not needed due to size of data set.

For the analysis we are interested only in classification rules which can be reproduced by humans, eg. to give a physician set of simple rules to determine the cause of the illness and not to give him complex neural network which he can not remember even though complex models will predict more accurately. However because of the small sample size and huge amount of attributes this problem is prone to overfitting. Therefore it doesn't even make sense to compare complex classificators such as neural networks to simple rules and pruned decision trees.

We first use algorithm One Rule which gives us insight and the most general rule (no overfitting) for the predicted attribute. Success rate of this classifier for this data set is often nothing to write home about though. Next we use decision list algorithm PART[1] which is an algorithm that builds C4.5 decision tree to a limited depth. Result of such a classifier is a set of simple rules which is perfect for our analysis. Simpler rules that do not overfit as much are obtained with Ridor decision tree algorithm. The most complex algorithm we use is J48, which is WEKA's implementation of famous C4.5 decision tree algorithm. All of the decision tree based algorithms are pruned (reducedErrorPruning = true).

We compare the models not by absolute error, but rather by Kappa statistic, which tells us how much better relatively is the model to Zero Rule algorithm (taking the majority class).

5.1 Associations between attributes

Dependency between attributes varies to some degree. Some of the attributes directly imply other, some of them are completely independent. By using association analysis, we can find rules that show us dependencies.

For that we use Apriori algorithm (FilteredAssociator with filter numerictonominal in WEKA).

6 Results

After playing with the decision tree based classifiers and repeated reevaluation after reprocessing data we got a set of rules that we will present. Huge number of rules were obviously result of overfitting which we tried to mitigate with post-prunning of the trees. We provide decision rules found for each stage of diagnosis that we found general enough to be used by domain experts.

6.1 Early symptoms

In a stage when a patient has not been examined by physician and only early obvious symptoms can be observed, we haven't found any general rules that would distinguish cause between bacterial and viral based only on symptoms of cold, headache, fever, nausea, loss of consciousness, seizure and onset.

It follows that we haven't found any significant general rule to classify culture of bacteria in case of bacterial cause based solely on early symptoms. The most general rule was by One Rule algorithm with success rate of 24.24%.

Prognosis can't be predicted in this diagnosis stage either. No general rule was found.

6.2 Physical examination

The patient was taken to a physician and routine diagnosis tests were performed.

No general rule was found to predict diagnosis, however the closest one without overfitting is the following set of rules by algorithm PART with overall success rate of 70.71% and Kappa statistic 0.2882.

```
SEX = F: VIRUS (44.0/5.0)

FEVER <= 11 AND
```

```
KERNIG <= 0 AND
FOCAL = - AND
AGE <= 62: VIRUS (19.0/1.0)

LASEGUE <= 0 AND
BT <= 39.7 AND
BT > 37.6: BACTERIA (14.0)

LOC_DAT = +: BACTERIA (10.0/3.0)
```

Algorithm Ridor gives simpler rules for the cost of success rate (67.14%) and kappa statistic (0.1844).

```
Diag2 = VIRUS  (140.0/42.0)
           Except (AGE > 45.5) and (BT > 37.85) and (FEVER <= 6.5)
                   => Diag2 = BACTERIA  (9.0/0.0) [1.0/0.0]
```

In order to predict whether culture is found, the best we can do is to use PART to get success rate of 73.57% and Kappa of 0.20 at the cost of obvious overfitting.

```
LOC_DAT = - AND
ONSET = ACUTE AND
COLD <= 8 AND
LOC <= 0 AND
FOCAL = - AND
LASEGUE <= 0 AND
SEX = F AND
AGE <= 51: F (22.0)

LOC_DAT = - AND
ONSET = ACUTE AND
COLD <= 8 AND
LASEGUE <= 0 AND
LOC <= 0 AND
FOCAL = - AND
SEX = M: F (33.0/4.0)

SEIZURE <= 1 AND
GCS > 10 AND
LOC_DAT = - AND
```

```
LASEGUE <= 0 AND
ONSET = ACUTE AND
COLD <= 4 AND
SEX = M: F (7.0)

SEIZURE <= 1 AND
GCS > 10 AND
ONSET = SUBACUTE: F (7.0/1.0)

SEIZURE <= 1 AND
GCS > 10 AND
LASEGUE > 0 AND
LOC_DAT = -: F (7.0)

SEIZURE <= 1 AND
GCS <= 10: F (6.0)

SEIZURE <= 1 AND
ONSET = ACUTE AND
BT > 39.6: T (5.0)

SEIZURE > 1: F (5.0)

ONSET = ACUTE AND
KERNIG > 0 AND
LOC <= 0: F (4.0)

ONSET = ACUTE AND
COLD > 4 AND
SEX = M: T (6.0)

FEVER <= 10 AND
ONSET = ACUTE AND
FEVER > 2 AND
BT <= 39.3: T (11.0/1.0)

ONSET = ACUTE AND
KERNIG <= 0 AND
FEVER > 0: F (12.0)
```

```
SEX = F AND
STIFF <= 0 AND
HEADACHE <= 9: T (3.0/1.0)

SEX = F: F (5.0)

FEVER <= 1: T (4.0)
```

7 Laboratory tests

Results of laboratory examination are known and we have more attributes to use to predict diagnosis, culture and prognosis.

Cause of meningoencephalitis can finally be generally predicted using attributes Cell_Poly and Cell_Mono without overfitting. PART found three simple rules with succes rate of 95.71% and Kappa 0.899.

```
Cell_Poly <= 220 AND
Cell_Mono > 12: VIRUS (96.0/1.0)

Cell_Poly > 3: BACTERIA (37.0)

CT_FIND = abnormal: BACTERIA (4.0)
```

The same result is confirmed by Ridor algorithm with 93.57% success rate and Kappa 0.844.

```
Diag2 = BACTERIA  (140.0/98.0)
           Except (Cell_Poly <= 220.5) and (Cell_Mono > 7.5)
                 => Diag2 = VIRUS  (67.0/1.0) [31.0/1.0]
```

If we want more detailed prediction of diagnosis, PART can give us an answer with success rate 72.14% and Kappa 0.590.

```
Cell_Poly <= 220 AND
FEVER <= 20 AND
Cell_Mono > 8 AND
LOC_DAT = -: VIRUS (50.0/7.0)

Cell_Poly <= 36 AND
Cell_Mono > 8: VIRUS(E) (15.0/2.0)
```

```
Cell_Mono <= 36 AND
AGE > 29: ABSCESS (5.0)

EEG_FOCUS = + AND
NAUSEA <= 1: BACTERIA (5.0/2.0)

EEG_FOCUS = - AND
AGE > 27: BACTERIA (15.0/2.0)

: BACTE(E) (4.0/2.0)
```

Prognosis is best predicted by pruned C4.5 decision tree (J48 with pruning) but gives us nothing statistically significant (Kappa 0).

8 After treatment

At this stage all of the attributes are known.

The best way to predict diagnosis is already known from laboratory tests. We haven't found a general way to predict the culture and prognosis of the patient.

8.1 Associations

The following couple of rules were found by first observing graphs of dependencies in WEKA and then subsets of attributes with some noticeable dependencies were fed to association analysis.

1. `COURSE(Grouped)=p 23 ==> EEG_WAVE=abnormal 21  conf:(0.91)`
   Patients witch positive EEG_WAVE are likely to have aftermath

2. `Diag2=BACTERIA CULT_FIND=T 15 ==> SEX=M 14  conf:(0.93)`
   Males are more prone to bacterial infection than females

3. `DIAG=VIRUS 68 ==> CT_FIND=normal 63  conf:(0.93)`
   Viral infections are likely not to be found by CT scan

4. `DIAG=VIRUS 68 ==> C_COURSE=negative 63  conf:(0.93)`
   `DIAG=BACTERIA 24 ==> C_COURSE=negative 22  conf:(0.92)`
   If the diagnosis is not one of ABSCESS, BACTE(E), TB(E) or VIRUS(E) then the prognosis is good

9 Conclusion

We performed a simple analysis of meningoencephalitis data set and found a diagnostic rule based on attributes Cell_Poly and Cell_Mono. We found that the problem is very prone to overfitting and for the medical domain what we need is the exact opposite - preferably rules which can be remembered by humans.

We also observed some dependencies in the data. More of these examples can be found by using larger set of data mining methods. More of these rules can be found by association rules between binned attributes and closely looking at correlations between targeted and measured attributes. We gave some of these rules as an example, but we are sure there is much more of them. Simply more thorough digging through data and reevaluation is needed. Not all of the found rules are significant though.

The analysis can be extended by finding dependencies between numerical attributes by using regression analysis. We did not perform that here.

References

[1] FRANK, E. – WITTEN, I. H. Generating Accurate Rule Sets Without Global Optimization. In SHAVLIK, J. (Ed.) *Fifteenth International Conference on Machine Learning*, s. 144–151. Morgan Kaufmann, 1998.

[2] WASHIO, T. JSAI KDD Challenge 2001: JKDD01. In *JSAI Workshops'01*, s. –1–1, 2001.