Introduction to Data Mining

Lab 5: Putting it all together

5.1. The data mining process

In the fifth class, we are going to look at some more global issues about the data mining process. (See the lecture of class 5 by Ian H. Witten, [1][1]). We are going through four lessons: the data mining process, Pitfalls and pratfalls, and data mining and ethics.

According to [1], the data mining process includes steps: ask a question, gather data, clean the data, define new features, and deploy the result. Write down the brief for these steps:

- Ask a question: Think about the question that we want to answer. We need to know what we want to know from the data.

- Gather data: Classify data to use classification techniques in data mining. We nee expert judgements on the data, expert classifications or correct results.

- Clean data: Real data is mucky. That's going to be a painstaking matter of looking through it and looking for anomalies. So, clean it.

- Define new features: feature engineering

- Deploy the result: technical implementation

Alternatively, according to (Han and Kamber, 2011), the data mining process is treated as a knowledge discovery (KDD) process including an iterative sequence of 7-steps. Please list them all in the below:

- Data cleaning: Clean missing values, noisy data.

- Data Integration: is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse). Data integration using Data Migration tools, Data Synchronization tools and ETL(Extract-Load-Transformation) process.

- Data Selection: is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. For this we can use Neural network, Decision Trees, Naive bayes, Clustering, and Regression methods.

- Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

[1] http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/

- Data mining: is defined as techniques that are applied to extract patterns potentially useful. It transforms task relevant data into patterns, and decides purpose of model using classification or characterization.

- Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. It find interestingness score of each pattern, and uses summarization and Visualization to make data understandable by user.

- Knowledge Representation: presenting the results in a way that is meaningful and can be used to make decisions.

5.2. Pitfalls and pratfalls

Follow the lecture in [1] to learn what are pitfalls and pratfalls in data mining.

Do experiments to investigate how **OneR** and **J48** deal with missing values.

Write down the results in the following table:

Dataset	**OneR's classifier model and performance**	**J48's classifier model and performance**
weather-nominal.arff (original)	outlook: sunny -> no overcast -> yes rainy -> yes (10/14 instances correct) Correction: 43%	outlook = sunny | humidity = high: no (3.0) | humidity = normal: yes (2.0) outlook = overcast: yes (4.0) outlook = rainy | windy = TRUE: no (2.0) | windy = FALSE: yes (3.0) Number of Leaves : 5 Size of the tree : 8 Correction: 50%
weather-nominal.arff (with missing values)	outlook: sunny -> yes overcast -> yes rainy -> yes ? -> no (13/14 instances correct) Correction: 93%	J48 pruned tree ------------------ : yes (14.0/5.0) Number of Leaves : 1 Size of the tree : 1 Correction: 50%

Remark: how do OneR and J48 deal with missing values?

- OneR uses the fact that a value is missing as significant, as something we can branch on.

- J48 does not have a branch that corresponds to a missing value.

5.3. Data mining and ethics

Reading

5.4. Association-rule learners

Do experiments to investigate how **Apriori** and **FP-Growth** generate association rules for datasets **vote.arff**

Dataset	**Apriori based association rules**	**FP-Growth based association rules**
Vote.arff	1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 <conf:(1)> lift:(1.63) lev:(0.19) [84] conv:(84.58) 2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 <conf:(1)> lift:(1.63) lev:(0.18) [76] conv:(76.47) 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 <conf:(1)> lift:(1.62) lev:(0.19) [80] conv:(40.74) 4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 <conf:(1)> lift:(1.62) lev:(0.18) [77] conv:(39.01) 5. physician-fee-freeze=n 247 ==> Class=democrat 245 <conf:(0.99)> lift:(1.62) lev:(0.21) [93] conv:(31.8) 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 <conf:(0.98)> lift:(1.77) lev:(0.2) [85] conv:(22.18) 7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 <conf:(0.98)> lift:(1.76) lev:(0.2) [88] conv:(18.46) 8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 198 <conf:(0.98)> lift:(1.72) lev:(0.19) [82] conv:(14.62) 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 <conf:(0.97)> lift:(1.57) lev:(0.17) [71]	1. [el-salvador-aid=y, Class=republican]: 157 ==> [physician-fee-freeze=y]: 156 <conf:(0.99)> lift:(2.44) lev:(0.21) conv:(46.56) 2. [crime=y, Class=republican]: 158 ==> [physician-fee-freeze=y]: 155 <conf:(0.98)> lift:(2.41) lev:(0.21) conv:(23.43) 3. [religious-groups-in-schools=y, physician-fee-freeze=y]: 160 ==> [el-salvador-aid=y]: 156 <conf:(0.97)> lift:(2) lev:(0.18) conv:(16.4) 4. [Class=republican]: 168 ==> [physician-fee-freeze=y]: 163 <conf:(0.97)> lift:(2.38) lev:(0.22) conv:(16.61) 5. [adoption-of-the-budget-resolution=y, anti-satellite-test-ban=y, mx-missile=y]: 161 ==> [aid-to-nicaraguan-contras=y]: 155 <conf:(0.96)> lift:(1.73) lev:(0.15) conv:(10.2) 6. [physician-fee-freeze=y, Class=republican]: 163 ==> [el-salvador-aid=y]: 156 <conf:(0.96)> lift:(1.96) lev:(0.18) conv:(10.45) 7. [religious-groups-in-schools=y, el-salvador-aid=y, superfund-right-to-sue=y]: 160 ==> [crime=y]: 153 <conf:(0.96)> lift:(1.68) lev:(0.14) conv:(8.6) 8. [el-salvador-aid=y, superfund-right-to-sue=y]: 170 ==> [crime=y]: 162 <conf:(0.95)> lift:(1.67) lev:(0.15) conv:(8.12) 9. [crime=y, physician-fee-freeze=y]: 168 ==> [el-salvador-aid=y]: 160 <conf:(0.95)> lift:(1.95) lev:(0.18) conv:(9.57)

	conv:(9.85) 10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210 <conf:(0.96)> lift:(1.7) lev:(0.2) [86] conv:(10.47)	10. [el-salvador-aid=y, physician-fee-freeze=y]: 168 ==> [crime=y]: 160 <conf:(0.95)> lift:(1.67) lev:(0.15) conv:(8.02)