4th Credit Project Page

Created by Jiang, Meng on Jun 04, 2017

4th Credit Project for CS 412 Summer 2017

Project goal:

Given a real dataset, (1) conduct data preprocessing and data cleaning, (2) define at least three data mining tasks, and (3) design and implement data mining methods: the project should go through the Knowledge Discovery in Databases (KDD) process.

Submission: project report (.docx) + code package (.zip) + output (.zip)

Evaluation:

1. Project organization: It contains all the steps of the KDD process: data preprocessing, data cube construction, data mining task definition, data mining method design, model evaluation, discussions and future work. (5%)
2. Data preprocessing and cube construction: How is the raw dataset transformed into an understandable format? How are the **dimensions and values** extracted from the dataset? Are basic data cube operations doable on the cube? (30%)
3. Task definition: Define at least **three different** data mining tasks (**frequent pattern mining, classification, clustering, prediction, recommendation, outlier detection, etc.**) with input and output as "Given … (concrete description of task-relevant data), find/classify/recommend/detect …". (10% * 3 = 30%)
4. Data mining method design and model evaluation. (30%)
5. Discussions on how to use knowledge that has been discovered in real life and future work. (5%)

Data sets:

KDD 2015-2016 papers: PDF files and TXT files.

https://www.dropbox.com/s/aoq4duz5vzjnfuz/pdf.zip?dl=0

https://www.dropbox.com/s/ne7wdm92jft0vnf/txt.zip?dl=0

224 KDD'15 papers and 206 KDD'16 papers.

Tips:

1. How to extract dimensions and values from research papers?

Probably you can read papers and manually turn research text into dimension-value structures as follows.

Ex.

Paper	Dimension	Value	Paper	Dimension	Value
kdd15-p9	PROBLEM	prediction	kdd15-p9	DATASET	netflix
kdd15-p9	METRIC	prediction accuracy	kdd15-p9	DATASET	yahoo music
kdd15-p9	PROBLEM	uncertainty	kdd15-p9	PROBLEM	recommender systems
kdd15-p9	PROBLEM	over-fitting	kdd15-p9	PROBLEM	music recommendation
kdd15-p9	METHOD	Bayesian matrix factorization	kdd15-p9	PROBLEM	book recommendation
kdd15-p9	PROBLEM	prohibitive cost of inference	kdd15-p9	PROBLEM	movie recommendation
kdd15-p9	METHOD	scalable distributed Bayesian matrix factorization	kdd15-p9	PROBLEM	news recommendation
kdd15-p9	METHOD	stochastic gradient mcmc	kdd15-p9	PROBLEM	partner recommendation
kdd15-p9	METHOD	distributed stochastic gradient langevin dynamics	kdd15-p9	PROBLEM	recommender systems
kdd15-p9	METHOD	mcmc	kdd15-p9	DATASET	netflix
kdd15-p9	METHOD	gibbs sampling	kdd15-p9	PROBLEM	recommender
kdd15-p9	METHOD	stochastic gradient descent	kdd15-p9	DATASET	netflix movie rating
kdd15-p9	PROBLEM	prediction	kdd15-p9	METHOD	bayesian probabilistic matrix factorization
kdd15-p9	METRIC	accuracy	kdd15-p9	METHOD	posterior distribution
kdd15-p9	METHOD	gibbs sampling	kdd15-p9	METHOD	bayesian analysis
kdd15-p9	METRIC	prediction error	kdd15-p9	METRIC	confidence intervals
kdd15-p9	METHOD	distributed stochastic gradient descent	kdd15-p9	METRIC	robustness
kdd15-p9	METRIC	rmse			

See kdd15-p9-annotated.pdf

OR https://www.dropbox.com/s/8jl9v8g01ncs7f1/kdd15-p9-annotated.pdf?dl=0

1. Show me some examples of data mining tasks that can be done on the dataset.

Ex.

(1) Given a table of Paper-Problem and a table of Paper-Method, find association rules like ProblemàMethod or frequent (Problem, Method) pairs/itemsets; categorized as “frequent pattern mining”.

(2) Given a table of Paper-Author and a table of Paper-Problem, for a specific problem, classify authors into “experts” or “non-experts”; categorized as “binary classification”.

(3) Given a table of Paper-Author, Paper-Problem, and Paper-Method, group papers into K clusters by different metrics of paper-paper similarity and compare the results.

No labels