Data Mining: Stage 1 Report

Yimeng Lu • 10.01.2017

Overview

Tasks

Mining frequent patterns and association rules from given papers

Steps

- Apriori algorithm for frequent item sets
- Generate association rules
- Rule evaluation
- Examine rule with facts

Tools

- Mlxtend, pandas

Frequent item sets-1

Pre-processing

- One author’s name written in different ways

antecedants	consequents	support
(scholkopf)	(thomas hofmann)	0.008306
(arthur gretton)	(bernhard scholkopf)	0.008306
(thomas hofmann)	(scholkopf)	0.006645

Generate rules

- Find frequent item sets
- Generate rules based on confidence
- Sort the rules by support first, then by lift
- Keep the highest ones

Frequent item sets-2

Top rules 1

antecedants	consequents	support	confidence	lift
(arthur gretton)	(bernhard scholkopf)	0.008306	0.800000	43.781818

Arthur Gretton is a Reader (Associate Professor) with the Gatsby Computational Neuroscience Unit, CSML, UCL, which he joined in 2010. He received degrees in physics and systems engineering from the Australian National University, and a PhD with Microsoft Research and the Signal Processing and Communications Laboratory at the University of Cambridge. He worked from 2002-2012 at the MPI for Biological Cybernetics and from 2009-2010 at the Machine Learning Department, Carnegie Mellon University.

6.6 Bernhard Schölkopf

Personal

Born February 20, 1968, Stuttgart, Germany;
three children with the Spanish illustrator Ana Martín Larrañaga

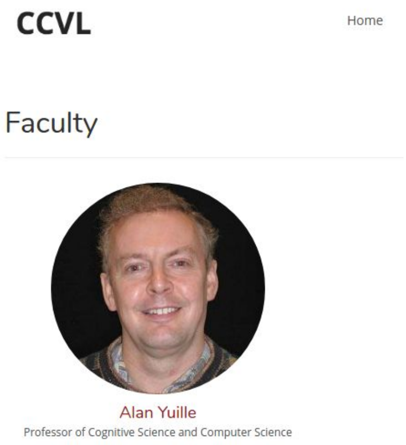
Employment

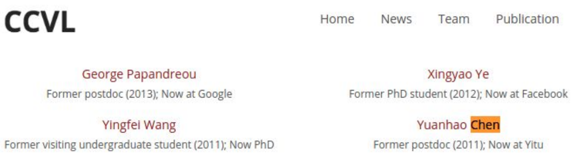
since 2011	Director at the Max Planck Institute for Intelligent Systems (Managing Director since 1.5.2011)
2001 – 2010	Director at the Max Planck Institute for Biological Cybernetics (Managing Director 1.8.2006–31.7.2009)
2000 – 2001	Group leader at the biotech startup Biowulf Technologies, New York
1999 – 2000	Researcher at Microsoft Research Ltd., Cambridge
1997 – 1999	Researcher at GMD (German National Research Center for Computer Science), Berlin

Frequent item sets-

Top rules 2

(alan yuille)	(yuanhao chen)	0.006645	0.750000	150.500000





Apply same idea to reference papers...

Possible generated rules

- Papers in same fields as frequent sets
- Recommended papers to read based on rules

Generated rules

antecedants	consequents	support	confidence	lift
(7746FE50)	(7A61221C)	0.023810	0.833333	35.000000
(7A61221C)	(7746FE50)	0.023810	0.833333	35.000000
(7D849A60)	(7D8197BF)	0.015873	0.750000	34.363636

A global geometric framework for nonlinear dimensionality reduction
JB Tenenbaum, V De Silva, JC Langford - science, 2000 - science.sciencemag.org
Abstract Scientists working with large volumes of high-dimensional data, such as global climate patterns, stellar spectra, or human gene distributions, regularly confront the problem of dimensionality reduction: finding meaningful low-dimensional structures hidden in their high-dimensional observations. The human brain confronts the same problem in everyday perception, extracting from its high-dimensional sensory inputs—30,000 auditory nerve ...
☆ 99 被引用次数：10658 相关文章 所有 98 个版本 Web of Science: 4846

Nonlinear dimensionality reduction by locally linear embedding
ST Roweis, LK Saul - science, 2000 - science.sciencemag.org
Abstract Many areas of science depend on exploratory data analysis and visualization. The need to analyze large amounts of multivariate data raises the fundamental problem of dimensionality reduction: how to discover compact representations of high-dimensional data. Here, we introduce locally linear embedding (LLE), an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embeddings of high- ...
☆ 99 被引用次数：11453 相关文章 所有 89 个版本 Web of Science: 5233

Large margin methods for structured and interdependent output variables
I Tsochantaridis, T Joachims, T Hofmann… - Journal of machine …, 2005 - jmlr.org
Abstract Learning general functional dependencies between arbitrary input and output spaces is one of the key challenges in computational intelligence. While recent progress in machine learning has mainly focused on designing flexible and powerful input representations, this paper addresses the complementary issue of designing classification algorithms that can deal with more complex outputs, such as trees, sequences, or sets. ...
☆ 99 被引用次数：1920 相关文章 所有 30 个版本 Web of Science: 644 »

[PDF] Max-margin Markov networks
B Taskar, C Guestrin, D Koller - Advances in neural information …, 2004 - papers.nips.cc
Abstract In typical classification tasks, we seek a function which assigns a label to a single object. Kernel-based approaches, such as support vector machines (SVMs), which maximize the margin of confidence of the classifier, are the method of choice for many such
☆ 99 被引用次数：1425 相关文章 所有 24 个版本 »

Thank you