第二节

写好一篇科技论文

发表高被引论文

1. Extensions to the k-means algorithm for clustering large data sets with categorical values, Data mining and knowledge discovery, 被引用次数：3788
2. Automated variable weighting in k-means type clustering, IEEE TPAMI, 被引用次数：1100
3. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data, IEEE TKDE, 被引用次数：865
4. A fuzzy k-modes algorithm for clustering categorical data, IEEE TFS, 被引用次数：671
5. A fast clustering algorithm to cluster very large categorical data sets in data mining, Dmkd, 被引用次数：989
6. Clustering large data sets with mixed numeric and categorical values, PAKDD, 被引用次数：822

目录

沟通技能
Communication Skills

- 交流和沟通能力是人的一种基本技能，对个人的健康生活和事业发展有重大影响。
- 交流和沟通能力是从事科学与技术研究的人的必备技能。
- 交流和沟通能力是人一生中要不断提升的技能。
- 为了更好地发布我们的科研成果，需要研究生掌握沟通技能，论文写作是重要的技能。

沟通的目的
Purpose of Communication

- 把你的想法和事情传给其他人，包括
  1. Events 事件
  2. Facts 事实
  3. Ideas 想法
  4. Methods 方法
  5. Results 结果
  6. Work 科研工作
  7. Thoughts 思想
  8. Emotions 情感
  9. …

基本要求

- 正确 Correct
- 准确 Accurate
- 简洁 Concise, Succinct
- 逻辑性强 Logical
- 信息明晰 Informative
- 一致性 Consistency
- 吸引人 Attractive, Interesting

为什么要学会英文读写技能?

- 英语是国际学术界的通用语言
  - 科技书籍、学术刊物、国际会议交流语言多数用英语
  - 计算机领域的新成果和新技术往往用英文发表
  - 工具书、技术手册、说明书等都用英文
  - 国际交流的通用语言
  - 网上资料英文也很多
  - 程序注释、文档用英文写

英语沟通的五种技能

- Listening 听
- Speaking 说
- Reading 读
- Writing 写
- Vocabulary 词汇量

科研论文

- 科研人员、研究生发表科研成果的主要手段。
- **Publish or perish.**
  - 发论文或自我毁灭

Information Sciences 466 (2018) 129–151

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

I-nice: A new approach for identifying the number of clusters and initial cluster centres

Md Abdul Masud[a], Joshua Zhexue Huang[a,*], Chenghao Wei[a], Jikui Wang[a], Imran Khan[b], Ming Zhong[a]

[a] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[b] Shenzhen Key Laboratory of Computational Intelligence, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

ARTICLE INFO

Article history:
Received 7 January 2017
Revised 4 July 2018
Accepted 11 July 2018
Available online 20 July 2018

Keywords:
Clustering algorithm
Initial cluster centres
Number of clusters

ABSTRACT

This paper proposes I-nice, which is a new method for automatically identifying the number of clusters and selecting the initial cluster centres in data. The method mimics a human being in observing peaks of mountains in field observation. The clusters in a dataset are considered as the hills in a field terrain. The distribution of distances between the observation point and the objects is computed. The distance distribution is modelled by a set of Gamma mixture models (GMMs), which are solved with the expectation-maximization (EM) algorithm. The best-fitted model is selected with an Akaike information criterion variant (AICc). In the I-niceSO algorithm, the number of components in the model is taken as the number of clusters, and the objects in each component are analysed with the *k*-nearest-neighbour method to find the initial cluster centres. For complex data with many clusters, we propose the I-niceMO algorithm, which combines the results of multiple observation points. Experimental results show that the two algorithms significantly outperformed two state-of-the-art methods (Elbow and Silhouette) in identifying the correct number of clusters in data. The results also show that I-niceMO improved the clustering accuracy and efficiency of the *k*-means clustering process.

1. Introduction

Clustering is one of the key techniques in data analysis. It is the process of dividing the data of objects into a set of clusters in which the objects in the same clusters are close to each other according to a similarity measure, whereas the objects in different clusters are far from each other. One problem in cluster analysis is that the number of clusters in the data to be analysed must be known in advance because many clustering algorithms require the number of clusters as an input parameter to run the algorithms. However, the number of clusters that exist in real data is usually unknown. Therefore, a number is often guessed in practical cluster analysis, which often results in unsatisfactory results. Although several methods for estimating the number of clusters in data have been developed [13,21,42,45,47], they either produce incorrect results or are difficult to use in real applications. Therefore, finding the correct number of clusters from real data remains a classical problem in cluster analysis. It is also an active research topic.

* Corresponding author.
E-mail addresses: masud@szu.edu.cn (M.A. Masud), zx.huang@szu.edu.cn (J.Z. Huang), chenghao.wei@yahoo.com (C. Wei), wjkweb@163.com (J. Wang), imran.khan@sustc.edu.cn (I. Khan), mingz@szu.edu.cn (M. Zhong).

https://doi.org/10.1016/j.ins.2018.07.034
0020-0255/

科技论文类型

- 研究论文
- 研究快报
  - Letters, Communications, correspondence
- 研究简报
  - Notes，Erratum （勘误）
- 讲座性论文 （Tutorial Paper ）
  - A tutorial paper provides novel and original insights and reflection on the use of one or several methods of modeling, design, analysis or synthesis in an acceptable format that can be used for guided and self-instruction.

- 学术论文
  - 研究成果论文 Research papers
  - 综述论文 Survey papers
  - 技术回顾论文 Technical review papers
  - 科学报导 Scientific reports

What is good academic writing?

- Writing that meets the expectations of the audience.
  - The most important audience is the reviewers of your writing who comment on your paper and make recommendations on whether your paper should be accepted for publication or not.
  - When you are writing a paper for publication, it is important to keep in mind the criteria that the reviewers will use in assessing your work.

What is a good scientific paper?

- A good and interesting topic
- Good research problems
- Innovative solutions
- Impressive and significant results
- Good and logical structures
- Excellent presentation

论文内容要保持平衡
Good Balance of a Paper

- Balance in two senses:
  - Physically, it refers to the distribution of information in your text;
  - Intellectually, expectation that you will present both sides of an issue.
- Balance in presentation
  - You need to decide roughly how much space (how many words) to allocate to the various sections of the paper.

论文内容要保持平衡
Good Balance of a Paper

- Balance in argument (论证、论据)
  - Presenting an ‘even-handed’ argument.
  - Making a strong logical case to persuade your reader to accept your point is really only half the picture.
  - Effective argumentation involves
    1. anticipating possible objections to your reasons or evidence, （预见反对意见）
    2. showing that you have considered those objections,
    3. using counter-arguments - the process known as *refutation*. （反驳）

论文写作

- 写一篇论文，概述MapReduce和Non-MapReduce分布式计算框架支持复杂算法大数据计算的步骤、特点和优缺点。
- Non-MapReduce参考文献：
  - Xudong Sun; Yulin He; Dingming Wu; Joshua Zhexue Huang, Survey of Distributed Computing Frameworks for Supporting Big Data Analysis, Big Data Mining and Analytics, 2023,6(2):154-169
  - Xudong Sun; Zhao Lingxiang; Chen Jiaqi; Cai Yongda; Wu Dingming, Joshua Zhexue Huang, Non-MapReduce computing for intelligent big data analysis, ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE, 2024
  - Xudong Sun, Dingming Wu, Yongda Cai, Lingxiang Zhao,Changda Xiao, Joshua Zhexue Huang. MapReduce vs. Non-MapReduce Efficiency and Scalability in Big Data Computing. World Congress: "System Theory, Algebraic Biology, Artificial Intelligence". 2023 from https://publ.icgbio.ru/wp-content/uploads/2023/12/SELECTED-WORKS-Congress26-300623-244-280.pdf
  - Alternate framework for distributed computing tames Big Data's ever growing costs (2023, February 23) retrieved 9 September 2024 from https://techxplore.com/news/2023-02-alternate-framework-big.pdf
- MapReduce 参考文献：
  - 自己查找

目录

Select a research topic

- 领域 Research area
- 方向 Research direction
- 题目 Research topic
- 问题 Research problems

Example:

Computer Science → Data mining →

k-means clustering →Initial center selection, k selection, etc.

Determine research area and direction

- Determine research area
  - Personal interests, specialty and knowledge background
- Determine research direction
  - Personal interests
  - Importance
  - Whether it is a hot research direction
  - Potential applications
  - Funding
- When research direction is determined, do not change easily

Research Topic

- Given by a supervisor
- Select by yourself
  - Literature survey and review (latest ones)
  - Listen to talks
  - Attend conferences and seminar
  - Search Web （Call for Papers）
- In topic selection, consider your capability and research environment and conditions. Evaluate the topic seriously to reduce the risk.

Research problems

- Each research problem defines a specific research task. The result should be able to publish one paper.
- Well-defined research problem is a precondition for the success of a research
- Define a research problem
  - Specificity
  - Meaningfulness
  - Real life research problem (not imagination)
  - Challenge
  - Can be solved in certain time （3-6 months)

How to determine a research problem

- New problems
  - E.g., big data computing, AI in face recognition
- Old problems without good solutions
  - E.g., feature selection in k-means clustering
- Identify problems from published papers
- Identify problems from real applications
  - e.g., categorical data clustering

Importance of research problems

- Is the problem of current interest? Is it topical (热点问题)?
- Is the problem likely to continue into the future?
- Will more information about the problem have practical applications?
- Will more information about the problem have theoretical importance?
- How large is the population affected by the problem?
- How important, influential, or popular is this population?
- Would this study substantially revise or extend existing knowledge?
- …

目录

Preparation for Research Work

- Problem statement
- Research plan
- Working paper
- PPT presentation document

Problem Statement

- Define the exact problem\problems you want to solve
- Your motivation (Why you want to solve this problem?)
- Example
  - Current distributed computing technologies for big data analysis are not efficient because of the high overhead of data communication among the nodes.
  - Define a new distributed computing framework that can reduce communication costs.

Prepare a Working Paper

- Title
- Outline
- References
- Word or Latex

Reference Papers

- 20-30-50+
- Survey papers (ACM Computing Survey)
- Technical review papers
- Papers relevant to your problems
- Your own papers or papers from your team

Current Methods and Shortcomings

- Current methods used to solve the problems
- The shortcomings of the current methods
- How to identify the shortcomings?
  - Think about special situations where the current methods may not work

Your Research and Contribution

- Develop new methods
- Enhance existing methods
- Algorithm design
- Algorithm analysis
  - Accuracy
  - Scalability
  - Easy to use (number of parameters the user needs to assign)

Implementation

- Quick coding
  - MATLAB or other high level language
  - Quick test
- Coding in C, C++, Java
  - Smart data structure
  - Scalability

Experiments

- Objectives
- Data
- Evaluation methods
- Result analysis

Synthetic Data and Real Data

- Synthetic data
  - Demonstrate the properties of the new algorithm
  - Data generation
- Real data
  - Demonstrate the performance of the new algorithm in solving real world problems
  - Selection of real world data sets

Result Analysis

- General analysis
  - Accuracy
  - Scalability
- Detail analysis
  - Advantages
  - Disadvantages

目录

Write a Paper

- When getting satisfactory experiment results, paper writing can start
- Opportunity driven
  - Conference call for papers
  - Journal special issue call for papers
  - Submission deadline
- Start with the working paper if you have.

Preparation

- Determine the publication place
- Read the requirements
  - Scope
  - Format
  - Style
- Download the template
- Study some papers in the publication

Basic Structure of a Scientific Paper

1. Title
2. Affiliations
3. Abstract
4. Key words
5. Introduction
6. Related work
7. Background and preliminaries
8. Method(s) and analysis
9. Experiments
   - Environment
   - Settings
   - Evaluation Methods
   - Results
   - Discussions
10. Conclusions
11. Acknowledgement
12. References
13. Appendix

Paper Title

- Informative
- Attractive
- Specific
- Short
- Use key words

Good Examples

- Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values
- Automated Variable Weighting in k-Means Type Clustering

Bad Examples

- Study on Knowledge Management System Based on Ontology
- Genetic Framework and Key Issues for Updates Propagation between Heterogeneous Spatial Databases
- Research on Incomplete Data Analysis Approach Based on the Extended Rough Set Theory

Abstract

- A summary of your paper
- Short (200 words)
- Main content
  - What is the contribution of the paper
  - What are the major characteristics of the contribution
  - What are the problems the new contribution can solve
  - What have the experiment results shown

Introduction

- Research importance
- Research problems
- Your motivations
- Current situations
- Your contributions
- Your results
- Organization of the paper

Related Work

- Related work can be a part of introduction or a separate section
- Summarize other works aiming at solving the same or similar problem, including your previous work if you have
- Make statements on the shortcomings of the existing work, e.g.,
  - Slow in handling large data
  - In some situations that the problem cannot be solved

Your Contribution

- Some preliminary introduction may be needed in order to make your contribution more readable or easy-to-understand
  - Definitions
  - Notations
  - Basic principle

Your Contribution (continue)

- Present your contribution concisely and clearly, using
  - Simple description
  - Formula
  - Figures
  - Flow chart
- Idea and then process or algorithm
- Too much mathematics can confuse referees

Evaluation Methods

- In algorithmic research, new algorithms need to be analyzed
  - Big O analysis for scalability
  - Convergence proof for iterative algorithms
- Evaluation methods
  - Comparison study on new algorithms and other algorithms
- Evaluation methods themselves are research topics

Experiments

- Objectives
- Data used
- Experiment conditions
- Experiment results
- Result analysis

Conclusions and Future Work

- Summarize the major contribution of the paper
- Points some research problems and directions for further research

Acknowledgement

- Funding support
- Those who helped in implementation
- Those who helped in experiments
- Those who gave useful comments on improving the research work and paper writing
- Those who helped correct your English

References

- Coverage
- Relevance
- Recency
- Easy access source
- Formats
- Your own work

Mathematics

- Mathematical formulation can help present contributions concisely
- Make the contribution correct
- Develop the algorithm
- However, too much and complicated mathematics should be avoided
  - Notations, explanation, careful in writing the formulas

Figures and Tables

- Simple and clear
- Captions
- Explanation in the text
- Numbering
- Size
- Color

Preparation for Submission

- Manuscript
  - Format
  - Proofreading
  - Technical editor
  - Packaging for online submission
- Submission letter
- Contact information

目录

Review Process

- Your paper is assigned to one associated editor
- Each associated editor is responsible for papers on certain topics
- The associated editor assigns each submission to 3-5 reviewers
- In the first round of review, it will take 3-6 months to receive reviewers' reports

Criteria

- Relevancy
- Originality
- Technical soundness (Quality)
- Presentation
  - Organization
  - Introduction
  - References
  - Language (readability)

Reviewer Recommendation

- Accept without change
- Accept with minor revision
- Resubmission after a major revision
- Reject

Associated Editor Recommendation

- According to the reviewers' reports, the associated editor will make a recommendation on your submission
  - Accept without change
  - Accept with minor revision
  - Resubmission after a major revision
  - Reject
- The recommendation will send to you together with reviewers' comments

Acceptance Rate

- Top journals and conferences
  - 15%-30% or lower
- Other SCI journals
  - Less than 50%

What Kind of Papers are Publishable?

- Interesting research problems
- Innovative methods
- Clear and accurate process
- Sufficient mathematics and formalization (not too much)
- Interesting and comprehensive experiment results
- Well written and presented
  - A good research work may not be accepted if it is not well presented

目录

Actions on Rejection

- Don’t simply think the reviewers were biased and discriminated to you or think someone told bad things about your work to reviewers.
  - Please don’t hate reviewers

Actions on Rejection

- Carefully analyze reviewers' comments. There must be some reason to reject your paper, find this reason and analyze
  - If it is a fundamental problem, then forget this paper and do something else
  - If reviewers misunderstood your work or it was your own mistake you can correct, revise your paper and submit it to another journal

Actions on Acceptance

- Very happy but you have job to do
  - Revise your paper according to reviewers' comments
  - Carefully proofread your final submission to correct errors as many as possible. When your paper is printed, your error will be there forever
  - Submit your final version on time

Actions on Revisions

- Analyze the reviewers' negative comments
- If it is reviewer's bias, try to address them in a positive way. (don't use sentences to offend reviewers because your paper will be reviewed again by them)
- Write a separate letter to explain how you revised your paper to address each comment
- Some general comments can be ignored.

Resubmission

- Package your revised manuscript together with the separate cover letter telling your revision
- In your cover letter to the association editor, you have to indicate that this is a resubmission with your submission number
- Online submission

Final Publication

- Carefully proofread your final submission to correct errors as many as possible. When your paper is printed, your error will be there forever
- Submit your final version on time
- Sign a copyright form

目录

Summary

- A well written paper is a necessary condition for publication
- It is not a simple task to get a paper published in top international journals
- One needs help in English writing
- Read more good papers to learn and mimic
- Don’t copy other people’s work

总结

- 研究方法和论文写作是研究生阶段所要学习掌握的重要技能，正确的研究方法可以使研究工作事半功倍，取得所期望的成果，而写作技能是研究成果得以顺利发表的必要条件。
- 研究方法和论文写作有规律可寻

Thank you !