Longbing Cao
Yong Feng
Jiang Zhong (Eds.)

Advanced Data Mining and Applications

LNAI 6440

6th International Conference, ADMA 2010
Chongqing, China, November 2010
Proceedings, Part I

Springer

Springer Link

ADMA: International Conference on Advanced Data Mining and Applications

Advanced Data Mining and Applications

6th International Conference, ADMA 2010, Chongqing, China, November 19-21, 2010, Proceedings, Part I

- Editors
- (view affiliations)

- Longbing Cao
- Yong Feng
- Jiang Zhong

Conference proceedings **ADMA 2010**

Part of the Lecture Notes in Computer Science book series (LNCS, volume 6440)

Also part of the Lecture Notes in Artificial Intelligence book sub series (LNAI, volume 6440)

Table of contents

1. Front Matter
   PDF
2. **Data Mining Foundations**
   1. Cost Sensitive Classification in Data Mining
      Zhenxing Qin, Chengqi Zhang, Tao Wang, Shichao Zhang
      Pages 1-11
   2. Web Users Access Paths Clustering Based on Possibilistic and Fuzzy Sets Theory
      Hong Yu, Hu Luo, Shuangshuang Chu
      Pages 12-23

3. Discriminative Markov Logic Network Structure Learning Based on Propositionalization and χ^2-Test
Quang-Thang Dinh, Matthieu Exbrayat, Christel Vrain
Pages 24-35
4. EWGen: Automatic Generation of Item Weights for Weighted Association Rule Mining
Russel Pears, Yun Sing Koh, Gillian Dobbie
Pages 36-47
5. Best Clustering Configuration Metrics: Towards Multiagent Based Clustering
Santhana Chaimontree, Katie Atkinson, Frans Coenen
Pages 48-59
6. On Probabilistic Models for Uncertain Sequential Pattern Mining
Muhammad Muzammal, Rajeev Raman
Pages 60-72
7. Cube Based Summaries of Large Association Rule Sets
Marie Ndiaye, Cheikh T. Diop, Arnaud Giacometti, Patrick Marcel, Arnaud Soulet
Pages 73-85
8. A Perceptron-Like Linear Supervised Algorithm for Text Classification
Anestis Gkanogiannis, Theodore Kalamboukis
Pages 86-97
9. Research on Time Series Forecasting Model Based on Moore Automata
Yixiong Chen, Zhongfu Wu, Zhiguo Li, Yixing Zhang
Pages 98-105
10. A Clustering Algorithm FCM-ACO for Supplier Base Management
Weining Liu, Lei Jiang
Pages 106-113
11. Nearest Neighbour Distance Matrix Classification
Mohd Shamrie Sainin, Rayner Alfred
Pages 114-124
12. Classification Inductive Rule Learning with Negated Features
Stephanie Chua, Frans Coenen, Grant Malcolm
Pages 125-136
13. Fast Retrieval of Time Series Using a Multi-resolution Filter with Multiple Reduced Spaces
Muhammad Marwan Muhammad Fuad, Pierre-François Marteau
Pages 137-148
14. DHPTID-HYBRID Algorithm: A Hybrid Algorithm for Association Rule Mining
Shilpa Sonawani, Amrita Mishra
Pages 149-160
15. An Improved Rough Clustering Using Discernibility Based Initial Seed Computation
Djoko Budiyanto Setyohadi, Azuraliza Abu Bakar, Zulaiha Ali Othman
Pages 161-168
16. Fixing the Threshold for Effective Detection of Near Duplicate Web Documents in Web Crawling
V. A. Narayana, P. Premchand, A. Govardhan
Pages 169-180
17. Topic-Constrained Hierarchical Clustering for Document Datasets
Ying Zhao
Pages 181-192
18. Discretization of Time Series Dataset Using Relative Frequency and K-Nearest Neighbor Approach
Azuraliza Abu Bakar, Almahdi Mohammed Ahmed, Abdul Razak Hamdan
Pages 193-201

19. MSDBSCAN: Multi-density Scale-Independent Clustering Algorithm Based on DBSCAN
Gholamreza Esfandani, Hassan Abolhassani
Pages 202-213
20. An Efficient Algorithm for Mining Erasable Itemsets
Zhihong Deng, Xiaoran Xu
Pages 214-225
21. Discord Region Based Analysis to Improve Data Utility of Privately Published Time Series
Shuai Jin, Yubao Liu, Zhijie Li
Pages 226-237
22. Deep Web Sources Classifier Based on DSOM-EACO Clustering Model
Yong Feng, Xianyong Chen, Zhen Chen
Pages 238-245
23. Kernel Based K-Medoids for Clustering Data with Uncertainty
Baoguo Yang, Yang Zhang
Pages 246-253
24. Frequent Pattern Mining Using Modified CP-Tree for Knowledge Discovery
R . Vishnu Priya, A. Vadivel, R. S. Thakur
Pages 254-261
25. Spatial Neighborhood Clustering Based on Data Field
Meng Fang, Shuliang Wang, Hong Jin
Pages 262-269
26. Surrounding Influenced K-Nearest Neighbors: A New Distance Based Classifier
I. Mendialdua, B. Sierra, E. Lazkano, I. Irigoien, E. Jauregi
Pages 270-277
27. A Centroid *k*-Nearest Neighbor Method
Qingjiu Zhang, Shiliang Sun
Pages 278-285
28. Mining Spatial Association Rules with Multi-relational Approach
Min Qian, Li-Jie Pu, Rong Fu, Ming Zhu
Pages 286-293
29. An Unsupervised Classification Method of Remote Sensing Images Based on Ant Colony Optimization Algorithm
Duo Wang, Bo Cheng
Pages 294-301
30. A Novel Clustering Algorithm Based on Gravity and Cluster Merging
Jiang Zhong, Longhai Liu, Zhiguo Li
Pages 302-309

3. **Data Mining in Specific Areas**

1. Evolution Analysis of a Mobile Social Network
Hao Wang, Alvin Chin
Pages 310-321
2. Distance Distribution and Average Shortest Path Length Estimation in Real-World Networks
Qi Ye, Bin Wu, Bai Wang
Pages 322-333
3. Self-adaptive Change Detection in Streaming Data with Non-stationary Distribution
Xiangliang Zhang, Wei Wang
Pages 334-345
4. Anchor Points Seeking of Large Urban Crowd Based on the Mobile Billing Data

Wenhao Huang, Zhengbin Dong, Nan Zhao, Hao Tian, Guojie Song, Guanhua Chen et al.
Pages 346-357
5. Frequent Pattern Trend Analysis in Social Networks
Puteri N. E. Nohuddin, Rob Christley, Frans Coenen, Yogesh Patel, Christian Setzkorn, Shane Williams
Pages 358-369
6. Efficient Privacy-Preserving Data Mining in Malicious Model
Keita Emura, Atsuko Miyaji, Mohammad Shahriar Rahman
Pages 370-382
7. Analyze the Wild Birds' Migration Tracks by MPI-Based Parallel Clustering Algorithm
HaiMing Zhang, YuanChun Zhou, JianHui Li, XueZhi Wang, BaoPing Yan
Pages 383-393
8. Formal Concept Analysis Based Clustering for Blog Network Visualization
Jing Gao, Wei Lai
Pages 394-404
9. Finding Frequent Subgraphs in Longitudinal Social Network Data Using a Weighted Graph Mining Approach
Chuntao Jiang, Frans Coenen, Michele Zito
Pages 405-416
10. Weigted-FP-Tree Based XML Query Pattern Mining
Mi Sug Gu, Jeong Hee Hwang, Keun Ho Ryu
Pages 417-428
11. Privacy-Preserving Data Mining in Presence of Covert Adversaries
Atsuko Miyaji, Mohammad Shahriar Rahman
Pages 429-440
12. Multiple Level Views on the Adherent Cohesive Subgraphs in Massive Temporal Call Graphs
Qi Ye, Bin Wu, Bai Wang
Pages 441-452
13. Combating Link Spam by Noisy Link Analysis
Yitong Wang, Xiaofei Chen, Xiaojun Feng
Pages 453-464
14. High Dimensional Image Categorization
François Poulet, Nguyen-Khang Pham
Pages 465-476
15. Efficiently Mining Co-Location Rules on Interval Data
Lizhen Wang, Hongmei Chen, Lihong Zhao, Lihua Zhou
Pages 477-488
16. Multiple Attribute Frequent Mining-Based for Dengue Outbreak
Zalizah Awang Long, Azuraliza Abu Bakar, Abdul Razak Hamdan, Mazrura Sahani
Pages 489-496
17. A Top-Down Approach for Hierarchical Cluster Exploration by Visualization
Ke-Bing Zhang, Mehmet A. Orgun, Peter A. Busch, Abhaya C. Nayak
Pages 497-508
18. Distributed Frequent Items Detection on Uncertain Data
Shuang Wang, Guoren Wang, Jitong Chen
Pages 509-520
19. Mining Uncertain Sentences with Multiple Instance Learning
Feng Ji, Xipeng Qiu, Xuanjing Huang
Pages 521-528

Next

Other volumes

1. Advanced Data Mining and Applications
   6th International Conference, ADMA 2010, Chongqing, China, November 19-21, 2010, Proceedings, Part I
2. Advanced Data Mining and Applications
   6th International Conference, ADMA 2010, Chongqing, China, November 19-21, 2010, Proceedings, Part II

About these proceedings

Introduction

With the ever-growing power of generating, transmitting, and collecting huge amounts of data, information overloadis nowan imminent problemto mankind. The overwhelming demand for information processing is not just about a better understanding of data, but also a better usage of data in a timely fashion. Data mining, or knowledge discovery from databases, is proposed to gain insight into aspects ofdata and to help peoplemakeinformed,sensible,and better decisions. At present, growing attention has been paid to the study, development, and application of data mining. As a result there is an urgent need for sophisticated techniques and toolsthat can handle new ?elds of data mining, e. g. , spatialdata mining, biomedical data mining, and mining on high-speed and time-variant data streams. The knowledge of data mining should also be expanded to new applications. The 6th International Conference on Advanced Data Mining and Appli- tions(ADMA2010)aimedtobringtogethertheexpertsondataminingthrou- out the world. It provided a leading international forum for the dissemination of original research results in advanced data mining techniques, applications, al- rithms, software and systems, and di?erent applied disciplines. The conference attracted 361 online submissions from 34 di?erent countries and areas. All full papers were peer reviewed by at least three members of the Program Comm- tee composed of international experts in data mining ?elds. A total number of 118 papers were accepted for the conference. Amongst them, 63 papers were selected as regular papers and 55 papers were selected as short papers.

Keywords

Clustering HPC adaptive algorithms classification data mining data types graphs knowledge discovery machine learning online communities pattern mining sensor data sequences spatial datasets web mining

Editors and affiliations

- Longbing Cao (1)
- Yong Feng (2)
- Jiang Zhong (3)

1. Faculty of Engineering and Information Technology, University of Technology Sydney, , Sydney, Australia
2. College of Computer Science, Chongqing University, , Chongqing, China
3. College of Computer Science, Chongqing University , , Chongqing, China

Bibliographic information

- DOI https://doi.org/10.1007/978-3-642-17316-5
- Copyright Information Springer Berlin Heidelberg 2010
- Publisher Name Springer, Berlin, Heidelberg
- eBook Packages Computer Science
- Print ISBN 978-3-642-17315-8
- Online ISBN 978-3-642-17316-5
- Series Print ISSN 0302-9743
- Series Online ISSN 1611-3349
- Buy this book on publisher's site

An Improved Rough Clustering Using Discernibility Based Initial Seed Computation

Djoko Budiyanto Setyohadi, Azuraliza Abu Bakar, and Zulaiha Ali Othman

Center for Artificial Intelligence Technology University Kebangsaan Malaysia Bangi, Selangor Darul Ehsan, 43000 Malaysia
djokobdy@gmail.com, {aab,zao}@ftsm.ukm

Abstract. In this paper, we present the discernibility approach for an initial seed computation of Rough K-Means (RKM). We propose the use of the discernibility initial seed computation (ISC) for RKM. Our proposed algorithm aims to improve the performance and to avoid the problem of an empty cluster which affects the numerical stability since there are data constellations where $|\underline{Ck}| = 0$ in RKM algorithm. For verification, our proposed algorithm was tested using 8 UCI datasets and validated using the David Bouldin Index. The experimental results showed that the proposed algorithm of the discernibility initial seed computation of RKM was appropriate to avoid the empty cluster and capable of improving the performance of RKM.

Keywords: Discernibility, Initial Seed Computation, Rough K-Means.

1 Introduction

Clustering is a process of classifying objects into classes based on similarities among data. The process of assigning an object to its cluster is fully based on the data similarity; therefore the characteristics of data may influence the clustering result. The performance of K-Means, as the most widely used clustering algorithm, depends on two key points, namely the initial clustering and the instance order [1]; in which initial clustering itself fully depends on the data distribution. Since the characteristic of the data influences the performance of K-means, many improvements of K-means are being developed. Rough K-means clustering (RKM) [2] is one of the well known extended K-means algorithm.

RKM is the clustering algorithm which addresses the problem of vague data. Its capability to cluster vague data comes from the integration of Rough Set Theory in the process of clustering. While in the original K-Means the cluster is viewed as a crisp cluster only, in RKM the cluster is deployed as an interval clustering. Here, the object is divided in the lower approximation where the object is certainly a member of the cluster, and the boundary area where the object is a member of more than one cluster [2]. Looking at its characteristics, RKM can be considered as a powerful algorithm for clustering vague data. Vague data can be clustered in a boundary area which is useful for further processing.

L. Cao, J. Zhong, and Y. Feng (Eds.): ADMA 2010, Part I, LNCS 6440, pp. 161–168, 2010.

Despite its advantages, RKM has a drawback especially on the numerical stability problem [3][4][5]. The problem arises because RKM equation requires that each cluster must have at least a member. This situation is also found in the original K-means and is solved by an initial seed computation [1][6][7][8][9][10]. Unlike in the original K-means, the empty cluster in the RKM will generate the numerical stability problem since there are data constellations where $|\underline{Ck}| = 0$, which refers to the computation of cluster centroid [2][11]. Therefore, several researchers have made improvements on the numerical stability problem [3][4][5].

According to the numerical stability problem, Peters [3] refined RKM by forcing at least one of the objects should be a member of the cluster. Hence one of the objects which is the closest to the centroid of the lower approximation will be assigned to the closest cluster. Miao [4] avoided the empty cluster by using the non-object outlier to the proper cluster and proposed the use of angle measurement to decide the member of clusters. Obviously, all of the previous work on RKM refinement, including that of Zhao [5] and of Lingras [11], focused on the membership function refinement. Although previous researchers had improved the RKM, they ignored the other source of the numerical stability problem i.e. the initial seed, since K-means clustering certainly relies on the chosen initial centroid [1][6][7]. Moreover, when the algorithm is applied, the boundary area should be restricted to avoid a numerical stability problem [3][4][5]. Therefore, to fill the gap of the previous work that heavily focused on refining the membership function to avoid numerical stability this work highlights the initial seed computation to avoid a numerical stability problem.

Many ISCs have been developed since the process of the K-means clustering is deterministic mapping from initial solution to local minima of final result [1][11]. The previous research showed that the use of the ISC did not only improve the performance of K-means but also was able to avoid the empty cluster problem that plagues K-Means. Hence we propose the use of ISC to avoid a numerical stability problem in RKM as an extension of K-means.

In this paper, we review the required characteristics of the previous ISC works, from which we further develop the algorithm based on the discernibility approach of Rough Set Theory which is suitable for the purpose of RKM i.e. processing the vague data. To verify the proposed algorithm, we use David Bouldin (DB) Index, which is a well-known validity measurement in clustering analysis [12].

2 Initial Seed Computation (ISC) on K-Means Clustering

Determining the initial seed points is very important in K-means since the initial centroid will determine the final centroid [1]. The main issue of the initial seed is that the initial centroid should be chosen properly. Currently, there are many studies focusing on the ISC for improving K-means algorithm in order to improve the result of clustering [1][6][7][8][9][10], which is also applicable in RKM to solve its numerical stability problem. Furthermore, the previous characteristics will be discussed below.