Lecture Notes in Networks and Systems 8

H. S. Saini
Rishi Sayal
Sandeep Singh Rawat *Editors*

Innovations in Computer Science and Engineering

Proceedings of the Fourth ICICSE 2016

Springer

Lecture Notes in Networks and Systems

Volume 8

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Advisory Board

More information about this series at http://www.springer.com/series/15179

H.S. Saini · Rishi Sayal · Sandeep Singh Rawat
Editors

Innovations in Computer Science and Engineering

Proceedings of the Fourth ICICSE 2016

Springer

Editors
H.S. Saini
Guru Nanak Institutions
Ibrahimpatnam, Telangana
India

Sandeep Singh Rawat
Guru Nanak Institutions
Ibrahimpatnam, Telangana
India

Rishi Sayal
Guru Nanak Institutions
Ibrahimpatnam, Telangana
India

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-10-3817-4 ISBN 978-981-10-3818-1 (eBook)
DOI 10.1007/978-981-10-3818-1

Library of Congress Control Number: 2017932092

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The volume contains 41 papers presented at the 4th International Conference on Innovations in Computer Science and Engineering (ICICSE 2016) held during 22–23 July, 2016 at Guru Nanak Institutions Campus in association with CSI Hyderabad Chapter.

The focus of the 4th ICICSE 2016 is to provide an opportunity for all the professionals and aspiring researchers, scientists, academicians and engineers to exchange their innovative ideas and new research findings in the field of computer science and engineering. We have taken an innovative approach to give an enhanced platform for these personnel, participants, researchers, students and other distinguished delegates to share their research expertise, experiment breakthroughs or vision in a broad criterion of several emerging aspects of computing industries. The conference received an overwhelming response in terms of number of submissions from different fields pertaining to innovations in the field of computer science in main tracks and special session. After a rigorous peer-review process through our program committee members and external reviewers, we accepted 41 submissions with an acceptance ratio of 0.38.

ICICSE 2016 was inaugurated by Dr. Anirban Basu, President, Computer Society of India. The guest of honors were Dr. A. Govardhan, Principal, JNTUCEH, Ms. G. Sandhya, Senior Program Manager Microsoft India, Mr. Raju Kanchibotla, Southern, Regional Vice President, CSI and Dr. Raghav Kune, Scientist, ADRIN and ISRO.

We take this opportunity to thank all keynote speakers and special session chairs for their excellent support to make ICICSE 2016 a grand success. We would like to thank all reviewers for their time and effort in reviewing the papers. Without their commitment it would not have been possible to have the important ‘referee’ status assigned to papers in the proceedings. The quality of these papers is a tribute to the authors and also to the reviewers who have guided any necessary improvement. We are indebted to the program committee members and external reviewers who have contributed towards excellent reviews and in a very short span of time. We would also like to thank CSI Hyderabad Chapter having coming forward to support us to organize this mega event.

We would also like to thank the authors and participants of this conference. Special thanks to all the volunteers for their tireless efforts. All the efforts are worth and would please us all, if the readers of these proceedings and participants of this conference rate the papers and the event inspiring and enjoyable.

Finally, we place our special sincere thanks to the press, print and electronic media for their excellent coverage of this conference.

Ibrahimpatnam, India

H.S. Saini
Rishi Sayal
Sandeep Singh Rawat

Organizing Committee

Patrons

Sardar Tavinder Singh Kohli
Sardar Gagandeep Singh Kohli

Conference Chair

Dr. H.S. Saini

Conference Co-Chairs

Dr. Veeranna
Dr. D.D. Sharma
Dr. S. Sreenatha Reddy
Dr. Rishi Sayal

Convenors

Dr. S. Masood Ahamed
Prof. V. Deva Sekhar
Dr. K. Madhusudana
Dr. Sandeep Singh Rawat
Ms. Thayyaba Khatoon

Co-Convenors

Dr. V. Sathiyasuntharam
Mr. S. Madhu
Mr. Lalu Nayak
Mrs. Subbalakshmi
Mrs. D. Sirisha
Mr. D. Saidulu
Mr. Ch. Ravindra

Conference Committee

Dr. Rishi Sayal
Mr. B. Sudhakar
Mr. M. Bharath
Mr. B. Nandan
Mr. Manikrao Patil

Publicity Chair International

Dr. D.D. Sharma
Mr. Imran Quereshi
Ms. E. Swetha Reddy
Ms. Thayyaba Khatoon
Mr. V. Poorna Chandra
Mr. B. Venkateswarlu

Publicity Chair National

Prof. V. Deva Sekhar
Mr. Y. Ravi Kumar
Ms. Kanchanlatha
Mr. D. Kiran Kumar
Ms. B. Mamatha

Program and Publication Chair

Ms. Thayyaba Khatoon
Mr. T. Ravindra
Mrs. K. Prasunna
Mr. K. Suresh
Mr. Devi Prasad Mishra
Mr. Nusrath Khan

Accommodation Committee

Dr. S. Masood Ahamed
Mr. A. Ravi
Mr. A. Vinay Sagar
Mr. B. Sudhakar
Mr. A. Srinivas
Mr. A. Ugendar

Advisory Board-International/National, Technical Program Committee

Dr. San Murugesan, Australia
Dr. Hemant Pendharkar, USA
Dr. Chandrashekar Commuri, USA
Dr. Muzammil H. Mohammed, Saudi Arabia
Dr. William Oakes, USA
Dr. Sartaj Sahni, USA
Dr. Jun Suzuki, USA
Dr. Prabhat Kumar Mahanti, Canada
Mrs. Sunitha B., Melbourne, Australia
M. Siva Ganesh, USA
Dr. Maliyanath Sundaramurthy, USA
Dr. Raj Kamal, India
Prof. Bipin V. Meheta, India
Dr. A. Damodaram, India
Dr. Amirban Basu, India
Dr. P.S. Avadhani, India
Dr. D.C. Jinwala, India
Dr. Aruna Malapadi, India
Mr. Ravi Sathanapalli, India

Dr. Sohan Garg, India
Dr. C. Shoba Bindu, India
Mr. Raju Kancibhotla, India
Prof. Rajkumar Buyya, Australia
Dr. Anuj Sharma, USA
Dr. Stephanie Farell, USA
Dr. Arun Somani, USA
Prof. Pascal Lorenz, France
Dr. Vamsi Chowdavaram, Canada
Mr. M. Kiran, CTS, New Jersey, USA
Dr. Lakshmivarahan, USA
Dr. S.R. Subramanya, USA
Dr. Sitalakshmi Venkataraman, Australia
Prof. Avula Damodaram, India
Dr. A. Govardhan, India
Dr. V. Kamakshi Prasad, India
Mr. H.R. Mohan, India
Dr. D.V.L.N. Somayajulu, India
Dr. Naveen Kumar, India
Dr. Uday Bhaskar Vemulapati, India
Dr. R.B.V. Subramanyam, India
Dr. Vijaylakshmi, India
Dr. K.P. Supreethi, India
Mr. Ramanathan, India

A Note from the Organizing Committee

Welcome to the 4th International Conference on Innovations in Computer Science & Engineering, India. On behalf of the entire organizing committee, we are pleased to welcome you to ICICSE-2016.

ICICSE, as the conference in the field, offers a diverse program of research, education, and practice-oriented content that will engage computer science engineers from around the world. The two-day core of the meeting is anchored by the research paper track. This year, the research paper track received 151 submissions. The papers underwent a rigorous two-phase peer-review process, with at least two program committee members reviewing each paper. The program committee selected 41 papers. All members of the program committee attended the meeting. These papers represent world-wide research results in computer science engineering.

Planning and overseeing the execution of a meeting of ICICSE is an enormous undertaking. Making ICICSE-2016 happen involved the combined labor of more than 50 volunteers contributing a tremendous amount of time and effort. We offer our sincere thanks to all the committee members and volunteers, and encourage you to take the opportunity to thank them if you meet them here. We would also like to thank all our sponsors who helped in making this event accessible to the Computer Science Engineering community.

Finally, we would like to thank the editorial board of Springer for agreeing to publish the proceedings and the staff at the editorial office for all their help in the preparation of the Proceedings.

Dr. H.S. Saini
Professor and Managing Director

Dr. Rishi Sayal
Professor and Associate Director

Dr. Sandeep Singh Rawat
Professor and Head-CSE

Contents

About the Editors

Dr. H.S. Saini, Managing Director of Guru Nanak Institutions, obtained his Ph.D. in Computer Science. He has over 24 years of experience at university/college level in teaching UG/PG students and has guided several B.Tech., M.Tech., and Ph.D. projects. He has published/presented high-quality research papers in international, national journals and proceedings of international conferences. He has two books to his credit. Dr. Saini is a lover of innovation and is an advisor for NBA/NAAC accreditation process to many institutions in India and abroad.

Dr. Rishi Sayal, Associate Director of Guru Nanak Institutions Technical Campus, has done B.E. (CSE), M.Tech. (IT), Ph.D. (CSE), LMCSI, LMISTE, MIEEE, MIAENG (USA). He has completed his Ph.D. in Computer Science and Engineering in the field of data mining from prestigious and oldest Mysore University of Karnataka state. His research work is titled "Innovative Methods for Robust and Efficient Data Clustering". He has published 22 research papers in international journals and conferences to support his research work (one paper being awarded the best paper in ICSCI 2008 supported by Pentagram Research Centre, India). He has over 25 years of experience in training, consultancy, teaching, and placements. His major accomplishments: Coordinator of professional chapter including IEEE, reviewer of IJCA (International Journal of Computer Association, USA). He is member of international association of engineers and guided more than 20 PG students. His current areas of research interest include data mining, network security, and databases.

Dr. Sandeep Singh Rawat obtained his Bachelor of Engineering in Computer Science from National Institute of Technology, Surat (formerly REC, Surat) and his Masters in Information Technology from Indian Institute of Technology, Roorkee. He was awarded Doctorate in Computer Science and Engineering by University College of Engineering, Osmania University, Hyderabad in 2014. He has been working at Guru Nanak Institutions Hyderabad since 2009 and he has 12 years of teaching and 2 years of industrial experiences. His current research interests include data mining, grid computing and data warehouse technologies. He is a life member of technical societies like CSI, ISTE, and member of IEEE.

Comparative Study of Techniques and Issues in Data Clustering

Parneet Kaur and Kamaljit Kaur

Abstract Data mining refers to the extraction of obscured prognostic details of data from large databases. The extracted information is visualized in the form of charts, graph, tables and other graphical forms. Clustering is an unsupervised approach under data mining which groups together data points on the basis of similarity and separate them from dissimilar objects. Many clustering algorithms such as algorithm for mining clusters with arbitrary shapes (CLASP), Density peaks (DP) and k-means are proposed by different researchers in different areas to enhance clustering technique. The limitation addressed by one clustering technique may get resolved by another technique. In this review paper our main objective is to do comparative study of clustering algorithms and issues arising during clustering process are also identified.

Keywords Data mining · Database · Clustering · k-means clustering · Outliers

1 Introduction

Data mining is used for analyzing huge datasets, finds relationships among these datasets and in addition the results are also summarized which are useful and understandable to the user. Today, large datasets are present in many areas due to the usage of distributed information systems [1]. Sheer amount of data is stored in world today commonly known as big data. The process of extracting useful patterns of knowledge from database is called data mining. The extracted information is visualized in the form of charts, graphs, tables and other graphical forms. Data

P. Kaur (✉) · K. Kaur
Department of Computer Engineering and Technology, Guru Nanak Dev University, Amritsar, India
e-mail: parneetbriar23@yahoo.in

K. Kaur
e-mail: kamal.aujla86@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Lecture Notes in Networks and Systems 8, DOI 10.1007/978-981-10-3818-1_1

mining is also known by another name called KDD (Knowledge Discovery from the Database). The data present in database is in structured format whereas, data warehousing may contain unstructured data. It is comparatively easier to handle static data as compared to dynamically varying data [2]. Reliability and scalability are two major challenges in data mining. Effective, efficient and scalable mining of data should be achieved by building incremental and efficient mining algorithms for mining large datasets and streaming data [1]. In this review paper our main objective is to do the comparative study of clustering algorithms and to identify the challenges associated with them.

2 Clustering in Data Mining

Clustering means putting objects having similar properties into one group and the objects with dissimilar properties into another. Based on the given threshold value the objects having values above and below threshold are placed into different clusters [1]. A cluster is group of objects which possess common characteristics. The main objective in clustering is to find out the inherent grouping in a set of unlabeled data [2]. Clustering is referred to as unsupervised learning technique because of the absence of classifiers and their associated labels. It is a type of learning by observation technique [3]. Clustering algorithm must satisfy certain requirements such as, it should be scalable, able of dealing with distinct attributes, capable of discovering arbitrary shaped clusters, and must possess minimal requirements for domain knowledge to determine input parameters. In addition, it should deal with noise and outliers [4–6].

2.1 Partitioning Clustering

In partitioning methods the instances are relocated and are moved from one cluster to another by starting the relocation from initial partitioning. The number of clusters to be formed is user defined. Examples of partitioning algorithms include Clustering Large Datasets Algorithm (CLARA) and k-means [7].

2.2 Density Based Clustering

These methods are based upon density and the cluster grows till the time the density does not exceed some threshold value. Density Based Spatial Clustering of Applications with Noise (DBSCAN) approach is a density based technique which is based on the idea that the least number of data points (Minpts) must be present around a point in its neighbourhood with radius (ε) [2].

2.3 Model Based Clustering

Model based approaches exaggerate the fit among the dataset and few mathematical models. The mathematical model generates data and then the original model is discovered from the data. The recovered model defines clusters and assigns documents to clusters [8].

2.4 Hierarchical Clustering

In such methods the data set is decomposed into a hierarchy. The decomposition can be done in agglomerative or divisive manner. Agglomerative approach is a bottom up technique where initially each data object is present in a single group whereas divisive is top down approach in which initially all the clusters are present in one cluster and then with every iteration this cluster is splitted into tiny clusters and the process continues until each data point is present within a single cluster. This kind of decomposition is represented by a tree structure called as dendrogram [9].

3 Literature Survey

Different types of mining algorithms have been proposed by distinct researchers. Selecting appropriate clustering algorithm however, depends on the application goal and algorithm's compatibility with the dataset. This section illustrates issues that may arise during the formation of clusters and different approaches to tackle with these problems.

3.1 Identification of Formation of Clusters

Very few techniques are available which can automatically detect the number of clusters to be formed. Some of the techniques rely on the information provided by the user while some use cluster validity indices which are very costly in terms of time required for computation. Some statistics such as Pseudo-F statistic and the Cubic Clustering Criterion (CCC) are used for identifying the cluster number [3]. Hao Huang et al. [4] designed an approach which is used for clustering clusters having arbitrary shapes (CLASP) that shrinks the size of dataset. CLASP is very effective and efficient algorithm which automatically determines the number of clusters and also saves computational cost. Zhensong Chen et al. [10] presented an

approach for image segmentation, based on density peaks (DP) clustering. This method possesses many advantages in comparison to current methods and can predict the cluster number, based on the decision graph and defines the correct cluster centers. Christy et al. [6] proposed two algorithms, detection of outliers on the basis of clusters and on the basis of distance, which uses outlier score for the detection and then removal of the outliers.

3.2 Clustering Large Datasets

Significant accuracy in clustering can be achieved by using Constrained Spectral Clustering (CSC) algorithms. However, to handle large and moderate datasets the existing CSC algorithms are inefficient. Clustering Large Applications (CLARA) is the best partitioning technique designed for large datasets which has less computation time [7]. Ahmad Chih-Ping Wei et al. [7] gives the comparative study of algorithms which cluster complex datasets. As the number of clusters increase, Clustering Large Applications based on Randomized Search (CLARNS) performs best in case of execution time and produces good quality clusters. In large datasets, CLARA gives better clustering results whereas Genetic Algorithm based clustering-Random Respectful Recombination (GAC-R) performs efficient clustering only in case of small datasets.

3.3 Large Computational Time

As compared to the traditional clustering algorithms like k-means, hierarchical clustering algorithms have many advantages but such algorithms may suffer from high computational cost [1]. Density based outlier detection algorithms also suffer from the problem of large computation time. High computation time is a major barrier in case of density based outlier detection algorithms although, they have number of advantages. Such algorithms have a less obvious parallel structure. So to resolve the problem of time and cost some algorithms are proposed by different researchers. Spectral clustering algorithms can easily recognize non-convex distribution and are used in segmentation of images and many more fields. Such clustering often costs high computation time when they deal with large images. So to solve this problem Kai. Li et al. [5] proposed an algorithm based on spectral clustering which performs segmentation of images in less computational time.

3.4 Efficient Initial Seed Selection

K-means algorithm is the crucial clustering algorithm used for mining data. The centers are generated randomly or they are assumed to be already available. In seed based integration, small set of labeled data (called seeds) is integrated which improves the performance and overcome the problem of initial seed centers. Iurie Chiosa et al. [11] has proposed novel clustering algorithm called Variational Multilevel Mesh Clustering (VMLC) which incorporates the benefits of variational algorithms and hierarchical clustering algorithms. The selection of seeds to be selected initially is not predefined. So to solve this problem, a multilevel clustering is built which offers certain benefits by resolving the problems present in variational algorithms and performs the initial seed selection. Another problem that the clusters have non optimal shapes can be solved by using greedy nature of hierarchical approaches.

3.5 Identification of Different Distance and Similarity Measures

For measuring the distance some standard equations are used in case of mathematical attributes like Euclidean, Manhattan and other maximum distance. These three special cases belong to Minkowski distance. Euclidean distance (ED) is the measure which is usually used for evaluating similarity between two points. It is very simple and easy metric, but it also possesses some disadvantages like it is not suitable in case of time series application fields and is highly susceptible to outliers and also to noise [12]. Usue Mori et al. [12] has proposed a multi-label classification framework which selects reliable distance measure to cluster time series database. Appropriate distance measure is automatically selected by this framework. The classifier is based on characteristics describing important features of time series database and can easily predict and discriminate between different set of measures.

4 Summary of Clustering Approaches

This section summarizes the clustering approaches which are reviewed in the above section. It is very clear from the table that the limitation addressed by one technique may get resolved by another (Table 1).

Table 1 Clustering techniques

Author (Year)	Clustering technique	Benefits	Limitations
Hao Huang (2014)	Algorithm for mining clusters with arbitrary shapes (CLASP)	Less computational cost	Efficiency reduces while clustering large datasets
Zhensong Chen (2015)	DP clustering algorithm	Defines correct cluster centres	More computational time
Chih-Ping Wei (2000)	Clustering large datasets (CLARA)	Produce small, distinct and symmetric clusters	Overlapping of clusters
A. Christy (2015)	Cluster based outlier detection, distance based outlier detection	Removes noise	Poor feature selection
Kai Li (2012)	Image segmentation algorithm based on spectral clustering	Recognize non convex distribution in images	High computation time
Iurie Chiosa (2008)	Variational multilevel mesh clustering	Produces clusters having optimal shape	More complexity and overhead

5 Conclusion

This paper describes the comparative study of clustering techniques such as CLARA, K-means, CLASP and SHRINK which are used by researchers in different application examined at different levels of perception. This paper highlights the concerned issues and challenges present in different clustering algorithms. The issue arising in one approach is resolved by other approach. Fuzzy logic is good for handling uncertainties and due to parallel nature, neural networks are good at handling real time applications. By doing hybridization of neural networks and fuzzy techniques we can obtain efficient results in detection of outliers. We have concluded that algorithms like CLARA are used for clustering large datasets efficiently, but some asymmetric clustering algorithms like CLASP, efficiently cluster simple datasets but do not give expected outputs in case of mixed and tightly coupled datasets. They are less accurate and efficient for clustering large datasets. Therefore, the technique based on the neural networks should be proposed to improve clustering and for increasing the efficiency in the asymmetric clustering algorithms.

References

1. R. Mythily, Aisha Banu, ShriramRaghunathan, Clustering Models for Data Stream Mining, Procedia Computer Science, Volume 46, 2015, Pages 619–626, ISSN 1877-0509.
2. Amineh Amini, Teh Ying, Hadi Saboohi, On Density-Based Data Streams Clustering Algorithms: A Survey, Journal of Computer Science and Technology, January 2014, Volume 29, Issue 1, pp 116–141.

3. Parul Agarwal, M. Afshar Alam, Ranjit Biswas, Issues, Challenges and Tools of Clustering Algorithm, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.
4. Hao Huang, Yunjun Gao, Kevin Chiew, Lei Chen, Qinming He, "Towards effective and efficient mining of arbitrary shaped clusters", ICDE, 2014, 2014 IEEE 30th International Conference on Data Engineering (ICDE), 2014 IEEE 30th International Conference on Data Engineering (ICDE) 2014, pp. 28–39.
5. Kai Li, Xinxin Song, "A Fast Large Size Image Segmentation Algorithm Based on Spectral Clustering," 2012 Fourth International Conference on Computational and Information Sciences, pp. 345–348.
6. A. Christy, G. Meera Gandhi, S. Vaithya subramanian, Cluster Based Outlier Detection Algorithm for Healthcare Data, Procedia Computer Science, Volume 50, 2015, Pages 209–215, ISSN 1877-0509.
7. Chih-Ping Wei, Yen-Hsien Lee and Che-Ming Hsu. Department of Information, Empirical Comparison of Fast Clustering Algorithms for Large Data Sets, Proceedings of the 33rd Hawaii International Conference on System Sciences – 2000.
8. Zhang Tie-jun, Chen Duo, Sun Jie, Research on Neural Network Model Based on Subtraction Clustering and Its Applications, Physics Procedia, Volume 25, 2012, Pages 1642–1647, ISSN 1875-3892.
9. Pedro Pereira Rodrigues, Joao Gama, Joao Pedro Pedroso, "Hierarchical Clustering of Time-Series Data Streams," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 5, pp. 615–627, May, 2008.
10. Zhensong Chen, Zhiquan Qi, Fan Meng, Limeng Cui, Yong Shi, Image Segmentation via Improving Clustering Algorithms with Density and Distance, Procedia Computer Science, Volume 55, 2015, Pages 1015–1022, ISSN 1877-0509.
11. Iurie Chiosa, Andreas Kolb, "Variational Multilevel Mesh Clustering," Shape Modeling and Applications, International Conference on, pp. 197–204, 2008 IEEE International Conference on Shape Modeling and Applications, 2008.
12. Usue Mori, Alexander Mendiburu, and Jose A. Lozano, Member, Similarity Measure Selection for Clustering Time Series Databases, IEEE Transactions on Knowledge and Data Engineering, 2015.

Adaptive Pre-processing and Regression of Weather Data

Varsha Pullabhotla and K.P. Supreethi

Abstract With the evolution of data and increasing popularity of IoT (Internet of Things), stream data mining has gained immense popularity. Researchers and developers are trying to analyze data patterns obtained from various devices. Stream data have several characteristics, the most important being its huge volume and high velocity. Although, a lot of research is being conducted in order to develop more efficient stream data mining techniques, pre-processing of stream data is an area that is under-studied. Real time applications generate data which is rather noisy and contain missing values. Apart from this, there is the issue of data evolution, which is a concern when dealing with stream data. To deal with the evolution of data, the proposed solution offers a hybrid of preprocessing techniques which are adaptive in nature. As a result of the study, an adaptive preprocessing and learning approach is implemented. The case study with sensor weather data demonstrates the results and accuracy of the proposed solution.

Keywords Stream mining · Data evolution · Adaptive pre-processing

1 Introduction

In the present day scenario, there are many applications in our day to day lives ranging from social networks, health monitors, telecommunications, network monitoring tools, sensor devices (in manufacturing, industrial pumps etc.) and such which continually generate huge volume of data at high velocity. These data streams evolve over time. Thus, there is a need for adaptivity of predictive models

V. Pullabhotla (✉) · K.P. Supreethi
Computer Science and Engineering Department,
Jawaharlal Nehru Technological University Hyderabad, Hyderabad, Telangana, India
e-mail: varshapull28@gmail.com

K.P. Supreethi
e-mail: supreethi.pujari@jntuh.ac.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Lecture Notes in Networks and Systems 8, DOI 10.1007/978-981-10-3818-1_2

to adapt to the evolution and change in environment of data streams. Recently, a lot of research and study is being carried out for such adaptive learning [1–3].

In real applications, pre-processing of data is a very important step of the data mining process as real data often comes from complex environments and can be noisy and redundant. In adaptive learning literature, the data pre-processing gets low priority in comparison to designing adaptive predictors. As data is continually changing, adapting only the predictor model is not enough to maintain the accuracy over time.

A good way to approach the above problem would be to tie the adaptivity of the preprocessor with the predictor. This can be accomplished in two ways. The first approach is to put aside a validation set, and use this validation set to optimize the pre-processing parameters and keep the pre-processing fixed in the model. The other approach would be to retrain the preprocessor afresh every time the learner is retrained. This approach requires the preprocessor to be synchronized with the learner.

In this paper, the aim is to present an implementation that can achieve adaptive pre-processing to get accurate output from adaptive learning. The pre-processing algorithm used is the "Multivariate Singular Spectrum Analysis". The learner algorithm used is the "K Nearest Neighbor" algorithm. These algorithms coupled with the Fixed window strategy [4] produce the adaptive pre-processing and learner framework.

The remainder of the paper is structured as follows: Sect. 2 presents the surveyed related work. Section 3 presents the proposed method for adaptivity. Section 4 shows the experimental results and performance evaluation. Finally, in Sect. 5, the conclusions drawn are presented.

2 Related Work

There has been a considerable amount of research and study conducted to address the issue of adaptive pre-processing along with adaptive learning. The issue of adaptive pre-processing while learning from a continuously evolving stream of data was raised in [4]. A framework that connects adaptive pre-processing to online learning scenarios was proposed. A prototype was developed to enable adaptive pre-processing and learning for stream data.

There has been the use of Genetic algorithm (GA) proposed by Wei Li to improve adaptive pre-processing to accomplish better results from adaptive learning [5].

Adaptive pre-processing of data streams has also been used with clustering algorithms. A pre-processing technique called equi-width cubes splits data space into a number of cubes, depending upon the data dimension and memory limit [6]. The new data which arrives is incorporated into one of the cubes. The algorithm computes a cluster center from previous chunk to create a new chunk. This new chunk is then sent to the clustering algorithm. This algorithm makes sure that the

data will not occupy all the available memory space and prevents loss of data due to the rate at which it arrives.

Adaptive pre-processing has been addressed in stationary online learning [7] for normalization of the input variables in neural networks. This was carried out so the input variables would fall into the range [−1, 1]. This proposed approach relates scaling of input features with scaling of the weights. However, the pre-processor is not adaptive.

3 Proposed Method

The framework of the proposed method is described in Fig. 1. The proposed Multivariate Singular Spectrum Analysis (MSSA) and K Nearest Neighbor approaches are applied to streaming weather data. Streaming weather data for the city of Hyderabad, India is used. The approach is carried out in two stages. In the first stage the MSSA algorithm which is used for pre-processing is trained with historical weather data for the city. The K-Nearest Neighbor algorithm is used for prediction. This model is trained using the output generated by the pre-processing algorithm. A stream of weather data is passed to this model, however, the results are not satisfactory.

The second stage involves applying the Fixed window strategy to the stream of weather data and retrain the preprocessor from scratch using the results obtained. The output obtained from the preprocessor results in the decomposition of the original time series into a stream of data without any noise. This output is passed to two K-nearest neighbor learner models: A model which is already trained using historical weather data and the other which is trained using the output obtained after retraining the pre-processor.

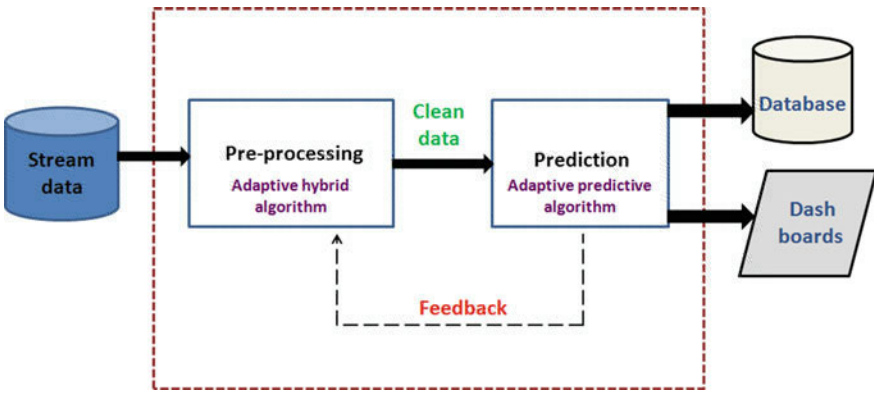

Fig. 1 Proposed system

4 Experimental Results

The proposed method is applied to streaming weather data for the city of Hyderabad, India. The performance measurement used to determine the accuracy of the prediction is the Root Mean squared error. This measure is meant to be used to understand how accurate the weather prediction for the next hour will be.

In this study, m = 3 and N = 10 are used, where N represents the sample size of the stream and m is the number of lags considered where the covariance is positive (this is determined using the autocorrelation function. We see a positive correlation at lag 3). The Root mean square error for the adaptive pre-processing and non-adaptive predictor is 0.29154. This error isn't considered too high and thus prediction is considerably accurate.

In the case of adaptive pre-processing and adaptive prediction, the RMSE is relatively low (0.014) and thus results in accurate prediction. However, this predictor is not trained by historical data as it is adaptive in nature. Thus this adaptive predictor has to be periodically re-trained for every 10 historical values to ensure that the predictor maintains its accuracy.

5 Conclusion and Future Scope

In this study, it has been demonstrated that the proposed approach, adaptive MSSA–KNN, could yield significantly higher prediction accuracy of weather data variables such as Temperature and Humidity than that of the non-adaptive KNN method. Adaptive MSSA–KNN results in a significant improvement prediction of weather data with RMSE of 0.014 and non-adaptive method results in RMSE of 0.29.

Future work would be focused on applying an incremental model instead of using a replacement model for adaptive pre-processing. Another area to work on would be to focus on passing multiple streams of weather data from different cities at once (Table 1).

Table 1 Prediction accuracies with adaptivity

Data	Adaptive pre-processing	Adaptive learning	RMSE
Stream weather data	No	No	4.008
Stream weather data	Yes	No	0.29154
Stream weather data	Yes	Yes	0.014

References

1. A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda.: New Ensemble Methods for Evolving Data Streams. In: Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 139–148, 2009
2. E. Ikonomovska, J. Gama, and S. Dzeroski.: Learning Model Trees from Evolving Data Streams. In: Data Mining Knowledge Discovery, vol. 23, no. 1, pp. 128–168, 2011
3. P. Kadlec and B. Gabrys.: Architecture for Development of Adaptive on-Line Prediction Models. In: Memetic Computing, vol. 1, no. 4, pp. 241–269, 2009
4. Indrė Žliobaitė and Bogdan Gabrys.: Adaptive Pre-processing for Streaming Data. In: IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 2, February 2014
5. Ketan Desale and Roshani Ade.: Preprocessing of Streaming Data using Genetic Algorithm. In: International Journal of Computer Applications (0975–8887) Volume 120–No.17, June 2015
6. Piotr Duda, Maciej Jaworski, and Lena Pietruczuk.: On Pre-processing Algorithms for Data Stream, L. Rutkowski et al. (Eds.): ICAISC 2012, Part II, LNCS 7268, pp. 56–63, 2012. Springer-Verlag Berlin Heidelberg 2012
7. H. Ruda.: Adaptive Preprocessing for on-Line Learning with Adaptive Resonance Theory (Art) Networks. In: Proc. IEEE Workshop Neural Networks for Signal Processing (NNSP), 1995

A Comparative Analysis for CBIR Using Fast Discrete Curvelet Transform

Katta Sugamya, Suresh Pabboju and A. Vinaya Babu

Abstract A Content Based Image Retrieval is proposed using two techniques in order to show a comparative analysis. The comparative analysis points out it's overall performance depends on the type of techniques used to extract multiple features and similarity metrics between the query image and images database. The first method uses colour histogram to extract colour features and the second method uses the Fast Discrete Curvelet Transform (FDCT) for the same process. In the first method, based on the colour features the query and database images were compared by using chi-square distance. Colour-histograms for both images were obtained and the images with most similarities are displayed (five images in this case). In the second method, instead of one feature (colour in the first case), a set of features are taken into consideration for calculating the feature vector. Once computation of feature vector is done, database and query images are compared to find out the top five similar images and results are displayed to the user.

Keywords Fast discrete curvelet · Transform (FDCT) · Inverse fast fourier transform (IFFT)

1 Introduction

With the enormous growing of digital data, it has become one of the active research area in the field of machine learning. The usage of multimedia caused an explosively growing of digital data giving people more ways to get those data.

K. Sugamya (✉) · S. Pabboju
Department of IT, Chaitanya Bharathi Institute of Technology, Hyderabad, India
e-mail: sugamya.cbit@gmail.com

S. Pabboju
e-mail: plpsuresh@gmail.com

A. Vinaya Babu
Department of CSE, JNTU, Hyderabad, India
e-mail: avb1222@jntuh.ac.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Lecture Notes in Networks and Systems 8, DOI 10.1007/978-981-10-3818-1_3