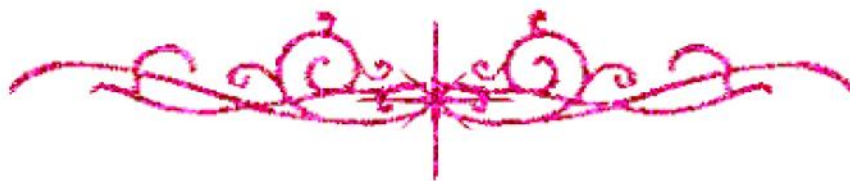


بسم الله الرحمن الرحيم





شبكة المعلومات الجامعية التوثيق الالكتروني والميكروفيلم



جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات



يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



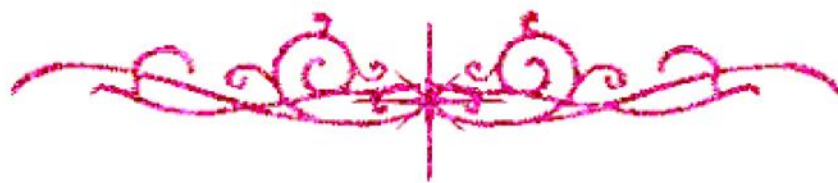


بعض الوثائق الأصلية تالفة





بالرسالة صفحات
لم ترد بالأصل



Ain Shams University
Faculty of Computer and Information Sciences
Information Systems Department



An Effective Approach for Processing Data Streams Queries

A thesis submitted in partial fulfillment of the requirements for
the degree of Ph.D in Computer and Information Sciences

To

Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

By

Fatma Mohamed Mahmoud Najib

Assistant Lecturer, Department of Information Systems
Faculty of Computer and Information Sciences, Ain Shams University

Under Supervision of

Prof. Dr. Tarek Fouad Gharib

Professor, Head of Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University

Prof. Dr. Nagwa Badr

Professor, Dean of Faculty of Computer and Information Sciences,
Faculty of Computer and Information Sciences, Ain Shams University

Assoc. Prof. Rasha Ismail

Associate Professor, Vice Dean for Graduate Studies and Research,
Faculty of Computer and Information Sciences, Ain Shams University

October – 2020

Acknowledgement

First of all thanks to Allah for giving me the opportunity and the capability to finish this thesis. Great thanks to everyone who assists and encourages me to complete this work.

I would like to deeply thank Prof. Dr. Tarek Gharib for his leadership, support and encouragement all time.

Also, I would like to thank Prof. Dr. Nagwa Badr for her supervision, instructions and help throughout this thesis.

I would also like to thank Assoc. Prof. Rasha Ismail for her guidance, assistance and support through the work.

Finally, all thanks to my father for being the reason to any success in my life. I would also like to thank every member of my family for their encouragement.

ABSTRACT

Most of recent applications such as sensor networks generate continuous and time varying data which are called data streams. Additional constraints are faced for efficient query processing of such data streams that have uncertain nature and require fast and timely processing. Traditional query processing techniques of static data process the whole data without partitioning them, which is not applicable to data streams. Applying data clustering is demanded as a preprocessing step of data streams. Also, data streams are often suffer from incompleteness and high dimensionality. So, in this thesis, we introduce a framework for efficiently answering incomplete high dimensional data streams queries.

The proposed framework handles the incompleteness issue by estimating missed values based on the corresponding nearest-neighbors' intervals. The continuous clustering mechanism is adopted and extended to accurately handle the incomplete data streams. The performance of incomplete data clustering was improved by 20% compared to the alternative approaches using two different data sets.

The proposed framework provides an improved subspace clustering to deal with high dimensional data streams. The experimental results using two datasets proved the efficiency of the proposed framework on average by 7.9% over the comparing algorithms for clustering such incomplete high dimensional data streams. The query processing performance of the incomplete high dimensional data streams was improved by 62% using two different data sets over the compared algorithms due the proposed clustering improvements of such data.

Table of Contents

Acknowledgement	i
Abstract	ii
List of Figures.....	v
List of Tables	vi
List of Abbreviations	viii
List of Publications	ix
Chapter 1 Introduction.....	2
1.1 Problem Definition.....	3
1.2 Objective	4
1.3 Contributions	4
1.4 Thesis Organization.....	5
Chapter 2 Background.....	7
2.1 Fuzzy C-means (FCM) Algorithm.....	7
2.2 Incomplete Data Imputation.....	8
2.3 High Dimensional Data Clustering	9
2.4 Data Streams Processing and Querying.....	10
Chapter 3 Related Works	12
3.1 Incomplete Data Imputation and Clustering	12
3.2 High Dimensional Data Processing	15
3.3 Data Streams Processing and Querying.....	16
Chapter 4: The Proposed Framework.....	19
4.1 Introduction	19
4.2 The Proposed Framework Phases.....	20
Chapter 5 The Proposed Fuzzy C-means Clustering for Incomplete Data Streams.....	28
5.1 Introduction	28
5.2 The Proposed Fuzzy c-means clustering algorithm for Incomplete Data streams (FID).....	29
5.3 Experimental Evaluation	32
5.4 Results and Analysis.....	34
5.4.1. The Average Misclassification	34
5.4.2. The Average Results of the Confusion Matrix Measurements	39
5.5 Summary	48

Chapter 6 The Proposed Subspace Clustering For Incomplete High Dimensional Data Streams.....	50
6.1 Introduction	50
6.2 The Proposed Subspace Clustering for Incomplete High Dimensional Data Streams (SIHD) Algorithm	51
6.2.1 1-D Subspaces Generation.....	52
6.2.2 Maximal Subspace Generation	53
6.2.3 Incomplete Stream Elements Imputation.....	53
6.2.4 Complete Data Streams Clustering.....	54
6.3 Experimental Evaluation	56
6.4 Results and Analysis.....	57
6.4.1. The Average Misclassification	58
6.4.2. The Average Results Based On Confusion Matrix Measurements.....	62
6.5 Summary	72
Chapter 7: The Proposed Incomplete High Dimensional Data Streams Queries Processing.....	74
7.1 Introduction	74
7.2 The Proposed Incomplete High dimensional Data streams Query processing (IHDQ) Algorithm....	75
7.3 Experimental Evaluation	78
7.4 Results and Analysis.....	80
7.4.1 Clustering Results.....	80
7.4.2 Query Performance Results	84
7.5 Summary	95
Chapter 8 Conclusion and Future Work.....	97
References	99
ملخص الرسالة	109

LIST OF FIGURES

Fig. 4.1	The proposed framework	20
Fig. 7.1	Average accuracy over data set1 (uniform)	82
Fig. 7.2	Average accuracy over data set2 (uniform)	83
Fig. 7.3	Average accuracy over data set1 (non-uniform)	84
Fig. 7.4	Average execution time using the proposed IHDQ and the SC algorithms	86
Fig. 7.5	Average throughput using the proposed IHDQ and the SC algorithms	87
Fig. 7.6	The average execution time over data set1 (uniform)	89
Fig. 7.7	The average execution time over data set2 (uniform)	90
Fig. 7.8	The throughput over data set1 (uniform)	91
Fig. 7.9	The throughput over data set2 (uniform)	92
Fig. 7.10	The average execution time over data set1(non-uniform)	93
Fig. 7.11	The average throughput over data set1(non-uniform)	94

LIST OF TABLES

Table 5.1	Data sets description	32
Table 5.2	Average Misclassifications And Standard Deviation Of The Misclassifications Over Data Set1 (Uniform)	36
Table 5.3	Average Misclassifications And Standard Deviation Of The Misclassifications Over Data Set2 (Uniform)	36
Table 5.4	Average Misclassifications And Standard Deviation Of The Misclassifications Over Data Set1 (Non-Uniform)	37
Table 5.5	Average Misclassifications And Standard Deviation Of The Misclassifications Of FID Over Data Set1 Using Different Distributions	38
Table 5.6	Average Confusion Matrix Measurements Over Data Set1 (Uniform)	40
Table 5.7	Average Confusion Matrix Measurements Over Data Set2 (Uniform)	42
Table 5.8	Average Confusion Matrix Measurements Over Data Set1 (Non-Uniform)	45
Table 5.9	Average Confusion Matrix Measurements Of FID Over Data Set1 Using Different Objects Distribution	47
Table 6.1	Data sets description	56
Table 6.2	Average misclassifications and standard deviation of the misclassifications over data set1 (uniform)	59
Table 6.3	Average misclassifications and standard deviation of the misclassifications over data set2 (uniform)	60
Table 6.4	Average misclassifications and standard deviation of the	61

	misclassifications over data set1 (non-uniform)	
Table 6.5	Average misclassifications and standard deviation of the misclassifications over data set1 using different data division	62
Table 6.6	Average confusion matrix measurements over data set1 (uniform)	63
Table 6.7	Average confusion matrix measurements over data set1 (uniform)	66
Table 6.8	Average confusion matrix measurements over data set1 (uniform)	69
Table 6.9	Average confusion matrix measurements over data set1 using different data distribution	71
Table 7.1	Data Sets Description	78

LIST OF ABBREVIATIONS

DBMS	Database management systems
DMSC	The discriminative multi-view subspace clustering algorithm
DS	The dense segments algorithm
DSMS	Data streams management systems
FCM	The fuzzy c-means algorithm
FID	The proposed fuzzy c-means clustering for incomplete data streams framework
FNM-T	The fuzzy non-metric model for data with tolerance
GBDC-P2P	The gossip-based distributed clustering for P2P networks algorithm
GFCM	The Grey based Fuzzy c-Means algorithm
IC	The improved clustering algorithm
IHDQ	The proposed incomplete high dimensional data streams query processing framework
IKFCM	The interval kernel fuzzy c-means clustering for incomplete data algorithm
KNN-FS	The KNN-based feature selection algorithm
LADEL	The label-aware distributed ensemble learning algorithm
LP	The Lattice Partition algorithm
MIDC	The multiple imputation of discrete combinations algorithm
MIGEC	The multiple imputation using gray-system-theory and entropy based on clustering algorithm
MMIC	The model based missing value imputation using correlation method
MRBIG	The map reduced enhanced bitmap index guided algorithm
MSC_IAS	The multi-view subspace clustering with intactness-aware similarity algorithm
OBAL	The online batch-based active learning algorithm
PS	The pointed subspace algorithm
SC	The smart scheme method
SIHD	The proposed subspace clustering for incomplete high dimensional data streams framework
SIM	The shape interaction matrix approach
SNMF	The symmetric non-negative matrix factorization approach
SSR	Sparse sample self-representation
VQ_fk_nps	The Gaussian kernel function and the nearest prototype strategy

LIST OF PUBLICATIONS

- | |
|---|
| [1] Najib, F.M., Ismail, R.M., Badr, N.L. and Gharib, T., 2020. Clustering based approach for incomplete data streams processing. Journal of Intelligent & Fuzzy Systems, 38 (3), 3213-3227. |
| [2] Najib, F.M., Ismail, R.M., Badr, N.L. and Gharib, T., 2020. Incomplete high dimensional data streams clustering. Journal of Intelligent & Fuzzy Systems, 39 (3), 4227-4243. |
| [3] Najib, F.M., Ismail, R.M., Badr, N.L. and Gharib, T., 2020. An efficient approach for query processing of incomplete high dimensional data streams. In: The 7th International Conference on Advanced Machine Learning and Technologies and Applications (AMLTA2021). (Accepted) |