

بسم الله الرحمن الرحيم

Sudan University of Science and Technology
College of Graduate Studies
College of Computer Science and Information Technology

Comparison of Two Clustering Techniques

Case Study: Breast Cancer Data

مقارنة بين تقنيتين من التقنيات العنقودية
دراسة الحالة: بيانات سرطان الثدي

A Thesis Submitted To Fulfill the Partial Requirements Of
Master Degree in Computer Science
(BSc. (Computer Science) Bayan College of Science and Technology 2009)

Submitted By:
Mustafa Ahmed Mohammed Zain Ahmed

Supervisor:
Dr. Mohamed Elhafiz Mustafa Musa

June 2013

Abstract

The main goal of this thesis is to study two clustering methods practically. The aims of practical study are to find out the capabilities of these two methods in clustering the breast cancer data set.

Many clusters have been generated using both k-means and Two-step.

Extensive comparisons have been conducted. The main conclusion is that these two methods generate different clusters.

The main reason could be the difference in their strategies. Further, studies are needed to find more reasons.

الخلاصة

الهدف الرئيسي لهذه الأطروحة هو دراسة طريقتان التجميع عمليا. أهداف الدراسة العملية هو لمعرفة قدرات هاتين الطريقتين في تجميع مجموعة البيانات لسرطان الثدي.

لقد تم إنشاء العديد من المجموعات باستخدام كل من k-means و Two-step.

أجريت مقارنات واسعة النطاق، والإستنتاج الرئيسي هو أن هاتين الخوارزميتان تولدان مجموعات مختلفة.

السبب الرئيسي يمكن أن يكون الفرق في إستراتيجياتهما، بالإضافة على ذلك، هناك حاجة لدراسات أكثر للعثور على المزيد من الأسباب.

Acknowledgements

Firstly, I thank God for giving me the ability to complete this thesis and to learn a little more about one small aspect of His universe.

This thesis could not have been done without the help of a few individuals who deserve thanks and credit. First of all, I owe a great deal of thanks to my supervisor, Dr. Mohamed El-Hafiz, for his sharp insight, expertise and enthusiasm. Because of his various responsibilities, his time was often in high demand, but he provided me with all the time that I needed. His continuing encouragement and support has always inspired me and boosted my confidence in myself before my work.

I owe a great thanks to my country Sudan and of course to my university Sudan University for a huge knowledge I learned during my semesters.

Of course, I am grateful to my parents without whom none of my university education would have come to existence. They have always been there for me and supported me in every aspect of life. Their love, care and confidence have always motivated me and gave me hope through the good times and in those more difficult moments.

To my brothers and friends who have also been very supportive and caring.

"Thank you all ".

Contents

Abstract.....	ii
الخلاصة	iii
Acknowledgements.....	iv
Contents.....	v
List of Tables.....	vii
List of Figure.....	viii
1 Introduction	
1.1 Research Overview	2
1.2 Aims and objectives	2
1.3 Research Methodology.....	3
1.4 Thesis outline	3
2 Literature Review	
2.1 Introduction.....	5
2.2 Related Work	5
2.3 Clustering Methods	6
2.3.1 Hierarchical Methods.....	7
2.3.2 Partitioning Methods.....	8
2.4 K-means Algorithm.....	9
2.5 Two-Step Algorithm	11
3 Result and Discussion	
3.1 Introduction.....	15
3.2 Data Descriptions	15
3.3 Data Pro-processing	15
3.4.1 K-means.....	18

3.4.2	Two-Step	19
3.5	Experiments results comparisons	20
3.6	Discussion.....	37
4	Conclusion and Future Works	
4.1	Conclusion	44
4.2	Future Works	44
	Bibliography.....	46
	Appendix A.....	49
	Appendix B.....	65

List of Tables

3.1	Description of data	16
3.2	Total number of records in each cluster “k-means”	18
3.3	Total number of records in each cluster “Two-step”	19
3.4	Intersections between K-Means and Two-Step “10 Clusters”	20
3.5	Intersections between K-Means and Two-Step “8 Clusters”	21
3.6	Percentage of intersections between Means and Two-Step “8 Clusters”	22
3.7	Intersections between Two-Step and K-Means “8 Clusters”	22
3.8	Percentage of intersections between Two-Step and K-Means“8 Clusters”	23
3.9	Intersections between K-Means and Two-Step “9 Clusters”	24
3.10	Percentage of intersections between K-Means and Two-Step “9 Clusters”	24
3.11	Intersections between Two-Step and K-Means “9 Clusters”	25
3.12	Percentage of intersections between Two-Step and K-Means “9 Clusters”	25
3.13	Intersections between K-Means and Two-Step “10 Clusters”	26
3.14	Percentage of intersections between K-Means and Two-Step “10 Clusters”	27
3.15	Intersections between Two-Step and K-Means “10 Clusters”	27
3.16	Percentage of intersections between Two-Step and K-Means “10 Clusters”	28
3.17	Intersections between K-Means and Two-Step “11 Clusters”	29
3.18	Percentage of intersections between K-Means and Two-Step “11 Clusters”	30
3.19	Intersections between Two-Step and K-Means “11 Clusters”	31
3.20	Percentage of intersections between Two-Step and K-Means “11 Clusters”	32
3.21	Intersections between K-Means and Two-Step “12 Clusters”	33
3.22	Percentage of intersections between K-Means and Two-Step “12 Clusters”	34
3.23	Intersections between Two-Step and K-Means “12 Clusters”	35
3.24	Percentage of intersections between Two-Step and K-Means “12 Clusters”	36
3.25	Content of intersecction between cluster-10 and cluster-3.....	39
3.26	Description of content of intersection between cluster-10 and cluster-3	40
3.27	Content of intersection between cluster-3 and cluster-7	41
3.28	Description of content of intersection between cluster-3 and cluster-7	42

List of Figures

2.1	Hierarchical Clustering: Agglomerative versus Divisive Methods.....	8
2.2	K-means Algorithm	10
3.1	K-means clusters	18
3.2	Two-Step clusters.....	19
3.3	Intersections between K-Means and Two-Step when 10 Clusters.....	21
3.4	Intersections between K-Means and Two-Step when 8 clusters	23
3.5	Intersections between K-Means and Two-Step when 9 clusters	26
3.6	Intersections between K-Means and Two-Step when 10 clusters	28
3.7	Intersections between K-Means and Two-Step when 11 clusters	32
3.8	Intersections between K-Means and Two-Step when 12 clusters	36
3.9	Intersections between K-Means and Two-Step when 10 clusters	38