Table of Contents

1 Introduction

1.1 Microarray Data Analysis 1

1.2 Coherent Pattern Identification: A Clustering Problem 2

1.3 Existing Approaches 3

1.4 Motivation 3

1.5 Aim of our work 4

1.5.1 Objectives of our work 4

1.6 Organization of the dissertation 4

2 Background of the work

2.1 Proximity measurement for gene expression data 5

2.2 Comparison of Proximity measures 7

2.3 Cluster Validity Measures 7

2.4 Discussion 8

3 Existing Work done on Clustering

3.1 Partitional Clustering 9

3.1.1 K-Means Algorithm 9

3.1.2 QT Clustering 10

3.2 Hierarchical Clustering 10

3.2.1 UPGMA 10

3.3 Density based Clustering method 11

3.3.1 DHC 11

3.4 Model based methods 11

3.4.1 SOM 12

3.4.2 SOTA 12

3.5 Graph based Clustering 13

3.5.1 CLICK 13

3.5.2 CAST 14

3.6 Discussion 15

4 The Proposed Approach

4.1 Concepts and definitions in E-CAST 16

4.2 Overview of E-CAST algorithm 17

4.2.1 Threshold Calculation 17

4.2.2 Cluster Formation 18

4.3 Phase 1: The proposed graph based clustering algorithm 19

4.3.1 Algorithm Description 20

4.4 Phase 2: Embedded Cluster Detection 22

4.4.1 Algorithm description 23

4.5 Discussion 24

5 Performance Evaluation

5.1 Environment 25

5.2 Dataset description 25

5.3 Results 26

6 Conclusions and Future Work

6.1 Conclusions 48

6.2 Future work 48

References