ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

GEODESIC BASED HYBRID SIMILARITY CRITERIA FOR APPROXIMATE SPECTRAL CLUSTERING OF LARGE MEDICAL DATASETS

M.Sc. THESIS

Berna YALÇIN

Department of Electronics and Communication Engineering

Biomedical Engineering Programme

JANUARY 2015

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

GEODESIC BASED HYBRID SIMILARITY CRITERIA FOR APPROXIMATE SPECTRAL CLUSTERING OF LARGE MEDICAL DATASETS

M.Sc. THESIS

Berna YALÇIN
(504121402)

Department of Electronics and Communication Engineering

Biomedical Engineering Programme

Thesis Advisor: Assist. Prof. Isa YILDIRIM

JANUARY 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

BÜYÜK MEDİKAL VERİ SETLERİNİN YAKLAŞIK SPEKTRAL ÖBEKLENMESİ İÇİN JEODEZİK TABANLI BENZERLİK ÖLÇÜTLERİ

YÜKSEK LİSANS TEZİ

Berna YALÇIN
(504121402)

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Biyomedikal Mühendisliği Programı

Tez Danışmanı: Assist. Prof. Isa YILDIRIM

OCAK 2015

Berna YALÇIN, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 504121402 successfully defended the thesis entitled **"GEODESIC BASED HYBRID SIMILARITY CRITERIA FOR APPROXIMATE SPECTRAL CLUSTERING OF LARGE MEDICAL DATASETS"**, which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assist. Prof. Isa YILDIRIM** Istanbul Technical University	.....
Co-advisor :	**Assoc. Prof. Kadim TAŞDEMİR** Antalya International University	.....
Jury Members :	**Prof. Ahmet Hamdi KAYRAN** Istanbul Technical University	.....
	Prof. Zehra ÇATALTEPE Istanbul Technical University	.....
	Prof. Aydın AKAN Istanbul University	.....

Date of Submission : **15 December 2014**
Date of Defense : **29 January 2015**

To my family; Alev, Ertuğrul and Akın

FOREWORD

Fifteen months ago I started a project on clustering large datasets thanks to my advisor Assist. Prof. Isa Yıldırım. This project which our small team run in leadership of my co-advisor Assoc. Prof. Kadim Taşdemir is the most enjoyable and efficient as well as strenuous part of in my life. It is inexpressible the long days spent in lab, the hope for good results and the sadness and tiredness with each failed attempt. This thesis is the result of this long process filled with beautiful memories.

So first of all I would like to thank my advisor Assist. Prof. Isa Yıldırım, for your support for my taking part in this project, suggestions and the opportunities you provided me during my education. I am very appreciative for your encouragement and guidance.

I would like to thank my co-advisor Assoc. Prof. Kadim Taşdemir who is my project supervisor. I feel pretty lucky to work with you because of your inspirational solving the different challenges, positive energy, motivation, suggestions and especially your patient and encouragements during the project and my thesis.

To Yaser, I appreciate your fellowship, clustering discussions and general tomfoolery. Metin, thank you for giving me motivation and having accompanied me in the final stages of the writing.

To Cagdas, I want to extend my special thanks for you. I am grateful for your support, understanding and encouragement in every moment of this fifteen months and my life.

I want to thank my parents. I know that your role in my life cannot be expressed by words. This thesis would not have been possible without your help and support. You always show me the right way and constitute a good model in my life. I am tremendously proud of you, and do not know how to thank to you.

Finally, I should mention that this work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant no 112E195, Advanced Similarity Criteria and Quantization Methods For Approximate Spectral Clustering Of Large Datasets.

January 2015 Berna YALÇIN

TABLE OF CONTENTS

Page

FOREWORD ..... **ix**
TABLE OF CONTENTS ..... **xi**
ABBREVIATIONS ..... **xiii**
LIST OF TABLES ..... **xv**
LIST OF FIGURES ..... **xvii**
SUMMARY ..... **xix**
ÖZET ..... **xxi**
1. INTRODUCTION **1**
1.1 Problem and Thesis Contribution 1
1.2 Organization 2
2. A REVIEW OF CLUSTERING METHODS **5**
2.1 Types of Clustering Methods 5
2.1.1 Hierarchical clustering 5
2.1.2 Partitional clustering 5
2.1.2.1 K-means algorithm 6
2.1.2.2 Spectral clustering 9
2.2 Spectral Clustering For Large Datasets 14
3. APPROXIMATE SPECTRAL CLUSTERING (ASC) **17**
3.1 Generating Data Representatives for ASC 19
3.1.1 Neural gas 19
3.1.2 K-means ++ 20
3.1.3 Selective sampling 20
3.2 Similarity Measures for ASC 21
4. PROPOSED GEODESIC BASED HYBRID SIMILARITY MEASURES FOR ASC **27**
5. EXPERIMENTS AND RESULTS **33**
5.1 Datasets 33
5.1.1 PR datasets 33
5.1.2 Medical datasets 33
5.1.3 Large medical datasets 34
5.2 Performance Evaluation 35
5.2.1 ASC results and assessment for PR datasets 35
5.2.2 ASC results and assessment for small/medium medical datasets 45
5.2.3 ASC results and assessment for large medical datasets 47
6. CONCLUSIONS **51**
6.1 Future Work 52
REFERENCES **55**

CURRICULUM VITAE 59

ABBREVIATIONS

ASC	**:** Approximate Spectral Clustering
ARI	**:** Adjusted Rand Index
BMU	**:** Best Matching Unit
CONN	**:** Connectivity Graph
GeoHSC	**:** Geodesic based Hybrid Similarity Criteria
K++	**:** K-means ++
NG	**:** Neural Gas
SC	**:** Spectral Clustering
SOVQ	**:** Sampling or Vector Quantization
SS	**:** Selective Sampling
PR	**:** Pattern Recognition

LIST OF TABLES

Page

Table 5.1 : Properties of datasets..... 34
Table 5.2 : Properties of medical datasets. 34
Table 5.3 : Properties of MR datasets. The datasets with four attributes consist of Flair, T1, post-Gadolinium T1 and T2; the dataset with three attributes is combination of Flair, T1 and T2; the one with two attributes consists of Flair and T2. 34
Table 5.4 : Clustering accuracies for the PR datasets. The ratio of the number of data representatives to the number of data points is 0.1. SS: Selective sampling; NG: neural gas; k++: k-means++ quantization .. 42
Table 5.5 : Adjusted Rand Index (ARI) for the PR datasets in Table 5.4. The bold values show the best ARI values for each dataset and the italic values show the best ARI values for other quantization/sampling(Q/S) methods..... 44
Table 5.6 : Clustering accuracies for the medical datasets. The ratio of the number of data representatives to the number of data points is 0.1. SS: Selective sampling; NG: neural gas; k++: k-means++ quantization..... 46
Table 5.7 : Clustering accuracies for the large medical datasets. The ratio of the number of data representatives to the number of data points is 0.01. k++: k-means++ quantization 48

LIST OF FIGURES

Page

Figure 2.1 : A taxonomy of clustering approach..... 6

Figure 2.2 : (a) A dataset with seven points (A,B,C,D,E,F and G) (b) its two clusters and (c) three clusters according to the dendogram obtained by hierarchical clustering in Fig.2.3. 7

Figure 2.3 : The dendogram obtained by the hierarchical clustering algorithm. Result of clustering is shown in Fig.2.2 (b) for k = 2 and Fig.2.2 (c) for k = 3 7

Figure 2.4 : An example of k-means algorithm. (a) A dataset with three clusters; (b) Initial points for cluster centers and initial clusters; (c) iteration 2 (d) iteration 3 the updated cluster labels and cluster centers; (e) final clustering 8

Figure 2.5 : An example of a simple weighted graph with nine nodes (vertices) and ten edges. 10

Figure 2.6 : An example of a graph with three nodes (vertices) and three edges. (a) An undirected graph; (b) A directed graph 11

Figure 3.1 : Steps of approximate spectral clustering (ASC) algorithm are showed on Lsun dataset obtained from [1]. The dataset has 2 features and 3 clusters. 18

Figure 3.2 : The obtained data representatives for a 2D 3-cluster dataset (Fig. 3.1) (a) data points and (b) prototypes (its red stars) 23

Figure 3.3 : Separation of a dataset into its Voronoi polygons a) Five of prototypes in Fig.3.2 b) The Voronoi polygons with the data points mapped to them of these five prototypes. Prototypes are shown by stars and their data points are shown by the circles. (Each color represents a different prototype and its corresponding data points). . 23

Figure 3.4 : An example of sub-Voronoi polyhedra of a prototype. a) $V_{1,2}$ and $V_{2,1}$ sub-Voronoi polyhedra for w_1, w_2 prototypes b) $V_{1,3}$ and $V_{3,1}$ sub-Voronoi polyhedra for w_1, w_3 prototypes..... 24

Figure 4.1 : A demonstration of difference between geodesic distance and Euclidean distance. The Euclidean approach calculates pairwise (P1 and P2) distance with respect to the data space while geodesic approach calculates this distance through P3 and P4 considering the data manifold..... 28

Figure 4.2 : (a) (Normalized) Euclidean and (b) geodesic distances between the prototypes of the Lsun dataset shown in Fig. 3.1. Some of the distances between clusters are smaller than some within-cluster distances according to the Euclidean distances, whereas three clusters are clearly separated with respect to geodesic approach. 29

Figure 4.3 : Neighborhood graphs obtained by k-nn graph with different values of k and by the CONN graph for data representatives of the dataset in Fig.3.1. The first three graphs are obtained by (mutual) k-nearest neighbor approach: (a) k=7, (b) k=5, (c) k=3, and (d) the last one is obtained by CONN graph. Note that an optimum k should be set to correctly identify submanifolds, where *CONN* enables varying number of neighbors with respect to the data manifold, resulting a clear separation without any user-set parameter. 30

Figure 5.1 : Data representatives of artificial datasets obtained by selective sampling with 0.1 sampling ratio, (a) Lsun (b) Chainlink and (c) Wingnut. 36

Figure 5.2 : Tap: Eigen space representation of the data representatives in Fig. 5.1 (Lsun) and their cluster labels obtained by Euclidean based similarity a) ground truth, b) ASC results. Middle: Eigen space representation of the data representatives and their cluster labels obtained by proposed hybrid criteria c) ground truth, d) ASC results. Bottom: A different view of the figures in the middle e) ground truth, f) ASC results. 37

Figure 5.3 : Tap: Eigen space representation of the data representatives in Fig. 5.1 (Chainlink) and their cluster labels obtained by Euclidean based similarity a) ground truth, b) ASC results. Bottom: Eigen space representation of the data representatives and their cluster labels obtained by proposed hybrid criteria c) ground truth, d) ASC results. 38

Figure 5.4 : Tap: Eigen space representation of the data representatives in Fig. 5.1 (Wingnut) and their cluster labels obtained by Euclidean based similarity a) ground truth, b) ASC results. Bottom: Eigen space representation of the data representatives and their cluster labels obtained by proposed hybrid criteria c) ground truth, d) ASC results. 39

Figure 5.5 : Simulated brain MR images and their clustering results. (a) T1, (b) post-Gadolinium T1, (c) T2 and (d) ground truth show three features and ground truth for SMHG0012 dataset, (e) Euclidean distance based Gaussian kernel based ASC (%83.76), (f) the proposed GeoHSC based ASC (%90.40), (g) ground truth zoom, (h) Euclidean zoom and (i) GeoHSC zoom 49

GEODESIC BASED HYBRID SIMILARITY CRITERIA FOR APPROXIMATE SPECTRAL CLUSTERING OF LARGE MEDICAL DATASETS

SUMMARY

Clustering is the unsupervised classification to group patterns such as data, observations or feature vectors. Due to its extraction of clusters in a fast manner without supervised information, many different methods have been proposed in various applications. Among them, spectral clustering (SC) has been recently popular and successfully used in various areas such as image processing, computer vision and information retrieval, thanks to its ability to find irregularly shaped clusters and its independence from parametric cluster models.

Spectral clustering is a manifold learning algorithm based on eigendecomposition of a graph Laplacian matrix constructed from pairwise similarities of the data points. Although this eigendecomposition produces higher clustering accuracies than the accuracies obtained by traditional methods, it has high computational cost and memory requirement which makes direct use of spectral clustering infeasible for clustering large datasets. To address this challenge, approximate spectral clustering (ASC) methods, which apply spectral clustering on a reduced set of data points (data representatives) selected by sampling or quantization, have been proposed. The ASC not only makes spectral clustering feasible for large datasets but also enables new information types to be used in similarity definition. In this thesis, data representatives are obtained by selective sampling and neural gas as sampling and quantization method respectively because of being the best methods in the literature for sampling and quantization. In addition, k-means++, a clustering algorithm which solves random initialization problem of k-means, is first used to obtain data representatives as a quantization method for ASC.

In order to achieve high clustering accuracies with ASC, an important step is to determine the criterion to define the pairwise similarities of the selected data representatives. Traditionally, an Euclidean distance based Gaussian kernel with a global decay parameter (optimally set by experiments) is used. Alternatively local decay parameters are also used. However, this approach ignores new information types such as data topology, local density distribution and data manifold which are provided by ASC. To utilize all the available information for accurate similarity definition, geodesic based hybrid similarity criteria are proposed in this study.

The geodesic distance between any two data representatives is their shortest path distance depending on a neighborhood graph. In addition to commonly used k-nearest neighbor graph, which requires a user-set parameter k, a weighted Delaunay triangulation (CONN) is employed to determine the optimal number of neighbors with respect to local data characteristics without any user-set parameter. CONN also indicates data topology and detailed local density distribution to be used in similarity definition. Based on these neighborhood graphs, the proposed geodesic distance based

similarities are defined using Euclidean distance, local density distribution by CONN, and their fused approach. Therefore, these proposed similarity criteria represent better pairwise similarities because they use different combination of all available information for ASC.

The clustering performance of the proposed criteria is evaluated by an extensive experimental study. Their advantages are first shown on three artificial datasets with different clustering challenges. Then, they are shown more successful than the commonly used Euclidean based similarity, using ten datasets (including medical data as well) from UCI Machine Learning Repository. Finally the proposed geodesic hybrid similarity criteria based ASC is applied on four sets of brain MR images obtained from MICCAI 2012 challenge on multi-modal brain tumor segmentation, achieving an accuracy of up to 80.06% without any supervised information. This accuracy is much successful than traditional clustering methods (compared to 66.55% obtained by k-means and compared to 76.52% obtained by ASC based on similarity using Euclidean distance) and very close to supervised accuracies existing in the literature. The extensive experiments show the outperformance of the proposed criteria and favor them as a successful approach in clustering both large and small/medium datasets.

BÜYÜK MEDİKAL VERİ SETLERİNİN YAKLAŞIK SPEKTRAL ÖBEKLENMESİ İÇİN JEODEZİK TABANLI BENZERLİK ÖLÇÜTLERİ

ÖZET

Öbekleme, gözlem, veri, özellik vektörü gibi örüntüleri sınıflara ayırmak için kullanılan bir öğreticisiz sınıflandırma yöntemidir. Karar verme, kestirim gibi noktalarda etkili ve hızlı bir süreç olmasından dolayı medikal alandaki örüntü öbekleme uygulamaları hızla artmaktadır. Örüntülerin farklı yapıları olan, düzenli sınırları ve homojen dağılımları olmayan öbekler içermesi nedeniyle çeşitli öbekleme yöntemleri geliştirilmiştir. Bu yöntemler arasında spektral öbekleme (SÖ) parametrik model kullanmayan bir yöntem olmasından dolayı hem düzensiz şekilli öbekleri bulabilmekte hem de parametrik modellere uymayan gerçek öbekleri bulmada daha başarılı olmaktadır. Spektral öbekleme, veri noktaları arasındaki ikili benzerliklerin özdeğer ayrışması tabanlı bir öğrenme yöntemidir. Veri noktaları bu ikili benzerliklerine göre aynı veya farklı öbeklere atanırlar. Bu şekilde öbek içi benzerlik en yüksek öbekler arası benzerlik ise en düşük hale getirilmeye çalışılır. Öbeklere ayırma işlemi verilerin benzerliğine göre yapıldığı için kullanılan benzerlik ölçütleri büyük önem taşımaktadır. Veriler arası benzerlik ne kadar iyi olursa öbekleme başarısı o kadar iyi olacaktır. Veri noktaları arasındaki ikili benzerliklerin öz değer ayrışımı spektral öbeklemenin başarılı bir yöntem olmasını sağlamasına rağmen yüksek hesaplama yükü ve bellek gereksinimi yüzünden büyük veri setlerine doğrudan uygulanması sorun oluşturmaktadır. Bu sorunun çözümü için geliştirilen yaklaşık spektral öbekleme (YSÖ) yöntemleri tüm veri setini kullanmak yerine, azaltılmış veri temsilcileri olarak adlandırılan temsilciler setini kullanır. Bu şekilde hem spektral öbeklemenin büyük veri setlerinde kullanımını mümkün kılar hem de veri setine ait kullanılmayan ve göz ardı edilen çeşitli bilgilerin benzerlik matrisini oluşturmada kullanılmasını sağlar. Yaklaşık spektral öbekleme iki aşamalı bir yöntemdir. İlk aşamasında çeşitli örnekleme veya nicemleme yöntemleriyle veri temsilcileri elde edilirler. Örnekleme yöntemleri veri setinde var olan veri noktalarının bir kısmını veri temsilcileri olarak seçerken, nicemleme yöntemlerinde veri temsilcileri esasında veri setinde var olmayan yeni noktaları oluşturarak elde edilirler. Bu tezde, literatürde en iyi örnekleme ve nicemleme yöntemleri olarak bilinen seçimli örnekleme ve sinir gazı nicemleme yöntemleri kullanılmıştır. Bunlara ek olarak, k-ortalama yönteminin rastgelelikten kaynaklanan ilk merkez atama sorununu çözmek için geliştirilen k-ortalama++ yöntemi de YSÖ nicemlemesinde ilk kez kullanılmıştır. Seçimli örnekleme, ilk olarak veri setini kullanıcının belirlediği sayıda alt öbeğe ayırır (alt öbek sayısı genelde var olan öbek sayısının üç katı olarak seçilir). Bu ayırma işlemi için ilk önce alt öbek sayısı kadar öbek merkezi en uzak en yakın uzaklık algoritmasıyla seçilir. Tüm veri noktaları için öklid uzaklığı kullanılarak en yakın öbek bulunur ve veri noktası alt öbeğe atanır. Bu şekilde bütün alt öbekler oluşturulduktan sonra her öbekten öbek büyüklüğüyle orantılı olarak veri temsilcileri seçilir. Böylece hem veri setindeki bütün öbeklerden veri temsilcisi seçilmesi

sağlanmış hem de veri temsilcilerinin büyük öbekten daha çok küçük öbekten daha az sayıda olacak şekilde dengeli seçilmesi sağlanmış olur. Sinir gazı ise yinelemeli ve sinir öğrenmesi tabanlı bir nicemleme yöntemidir. Bütün veri noktaları için en iyi eşleşen sinir birimi bulunur ve güncellenerek en uygun hale getirilir. Veri temsilcileri olarak sinir birimleri kullanılır. Bu çalışmada kullanılan diğer nicemleme yöntemi k-ortalama++ ise k-ortalama öbekleme algoritması gibi çalışır, tek fark ilk merkezlerini bir olasılık yoğunluk fonksiyonuna göre seçmesidir. Veri temsilcisi seçiminde ise şu şekilde kullanılmıştır. İlk olarak belirlenen veri temsilcisi sayısı kadar öbek merkezi olasılık yoğunluk fonksiyonuna göre seçilir. Daha sonra seçilen bu merkezler için k-ortalama algoritması uygulanır. En son elde edilen merkezler veri temsilcileri olur. Yaklaşık spektral öbeklemenin ikinci aşamasında ise veri temsilcileri üzerinden spektral öbekleme yapılır ve veri temsilcileri ile ilişkili olarak tüm veri noktaları için öbek etiketleri bulunur. Spektral öbekleme yapılırken çeşitli benzerlik ölçütleri kullanılarak veri temsilcilerinin (veri noktaları) ikili benzerliklerini gösteren benzerlik matrisi elde edilir. Bu matrisin veriler arası benzerliği en iyi ifade edecek şekilde elde edilmesi öbekleme performansı açısından oldukça önemlidir. Bunun için çeşitli benzerlik ölçütleri geliştirilmiştir. Geleneksel olarak Öklid tabanlı Gauss kernel benzerlik işlevi kullanılır. Bu işlev global bir ayrıştırma parametresi kullanır ve bu parametrenin en uygun hali her veri seti için deneysel olarak belirlenir. Alternatif olarak yerel ayrıştırma parametresinin kullanılması da mümkündür. Bu yaklaşımlar YSÖ'nün ortaya çıkardığı veri topolojisi, yerel yoğunluk dağılımı ve veri manifoldu gibi veri setine ait yeni bilgilerin kullanımını göz ardı ederler. Bu çalışmada benzerlik tanımını en doğru şekilde yapabilmek için veriye dair elde edilen bütün bilgilerin kullanılmasını mümkün kılan jeodezik tabanlı hibrit benzerlik ölçütleri önerilmiştir. Benzerlik ölçütleri kullanılarak benzerlik matrisi oluşturulduktan sonra bu matris üzerinden öz değer ayrışımı yapılır ve k öbek sayısı için en büyük k öz değerle ilişkili öz vektör bulunur ve bu şekilde veri temsilcilerinin k öz vektörle gösterildiği matris elde edilir. Bu matris üzerinden basit bir öbekleme algoritması (k-ortalama gibi) kullanılarak veri temsilcileri k öbeğe ayrılır. Veri noktaları kendilerini temsil eden veri temsilcilerinin bulunduğu öbeğe atanırlar ve bu şekilde tüm veri seti için öbekleme işlemi gerçekleştirilmiş olur.

Çalışma kapsamında önerilen jeodezik tabanlı hibrit benzerlik ölçütleri yaklaşık spektral öbekleme algoritmasının ikinci aşamasında verilerin ikili benzerliklerini gösteren benzerlik matrisinin elde edilmesi için geliştirilmiştir. Jeodezik uzaklık, iki veri temsilcisi arasındaki uzaklığı komşuluk çizgesindeki en kısa mesafe olarak hesaplar ve bu sebeple komşuluk çizgesi kullanımını gerektirir. Bunun için genelde k en yakın komşuluk çizgesi (k-nn) ve ağırlıklandırılmış Delaunay üçgeni (CONN) kullanılır. K-nn çizgesi kullanıcıdan alınan k parametresi ile komşuluk ilişkisini belirler. Burada verilerin komşu olup olmadığı kararı iki şekilde verilir. İlk yöntemde iki veri noktasından herhangi birisi diğerinin k en yakın komşuluğunda ise veriler bağlanır, diğer yöntemde ise iki veri noktasının her ikisi de birbirlerinin k en yakın komşulukları içerisinde olmaları durumunda bağlanırlar. K-nn çizgesi için temel problem k sayısının kullanıcı tarafından belirlenmesidir. Her veri seti için ve hatta her veri temsilcisi için en uygun k değeri farklıdır ve bütün veri noktaları için aynı k sayısında komşu belirlemek komşuluk çizgesinin doğru oluşturulamamasına sebep olur. CONN ise herhangi bir kullanıcı parametresi kullanmadan yerel veri karakteristikleriyle ilişkili olarak en uygun komşu sayısına karar verir. Bu şekilde her veri noktası ve veri seti için farklı sayıda komşu belirleyerek komşuluk çizgesinin

daha doğru oluşmasını sağlar. CONN veri karakteristikleriyle ilişkili olarak komşu sayısını belirleyebilmek için veri topolojisi ve yerel yoğunluk dağılımı bilgilerini benzerlik tanımında kullanır. Önerilen jeodezik uzaklık tabanlı benzerlik ölçütlerinde CONN ve k-nn komşuluk çizgeleri kullanılmıştır. Bu çizgeler üzerinden jeodezik uzaklık hesabı içinse Öklid uzaklığı ve yerel yoğunluk dağılımını gösteren CONN uzaklığı kullanılmıştır. Önerilen yöntemler yaklaşık spektral öbeklemenin ortaya çıkardığı bütün bilgileri komşuluk çizgesi ve uzaklık hesaplarıyla kullanarak daha iyi bir benzerlik sunumu elde edilmesini sağlamıştır.

Önerilen yöntemlerin öbekleme performansı deneysel olarak değerlendirilmiştir. Öncelikle bu yöntemin avantajlarını göstermek için farklı öbekleme problemlerine sahip üç yapay veriseti (Lsun, Chainlink ve Wingnut) kullanılmış daha sonra UCI Machine Learning Repository'den alınan ve medikal veri setlerini de içeren on veri seti üzerinde önerilen yöntemlerin klasik Öklid tabanlı benzerlikten daha iyi olduğu gösterilmiştir. Son olarak MICCAI 2012 çok modelli beyin tümörü bölütleme yarışmasında kullanılan 4 adet beyin MR görüntüleri seti kullanılmış ve herhangi bir öğreticili yöntem kullanmadan % 80.06 ortalama başarı elde edilmiştir. Bu başarı geleneksel yöntemlerle elde edilen başarılardan daha iyidir (k-ortalama başarısı: % 66.55 ve Öklid uzaklığı kullanan benzerlik tabanlı YSÖ başarısı: % 76.52). Performans değerlendirmesi sonucunda önerilen jeodezik uzaklık tabanlı hibrit benzerlik ölçütlerinin hem küçük/ orta hem de büyük veri setlerinde başarılı bir yaklaşım olduğu görülmüştür.

1. INTRODUCTION

1.1 Problem and Thesis Contribution

The progress of technology and variety in applications such as Internet, digital imaging and social media have produced many high-dimensional and high-volume datasets. The difficulty in analyzing both the high-volume and the variety of these data requires automatic knowledge extraction methods to summarize the data. There are several analysis techniques for this process. In pattern recognition, this analysis is based on predictive modeling which learns the identifications or attitudes of unknown data have tried to predict with using some training data. This process is called *learning*. The learning process is divided into two major types that are (i) supervised (classification) and (ii) unsupervised (clustering). In classification, per-classified (labeled) data have been used to learn the description of classes. A new data has been labeled by using the characteristics of these classes. In case of clustering, there are only unlabeled data and various measures showing the similarity between these unlabeled data used for learning process [2]. Thanks to extracting data characteristics with no a priori information, clustering is an important tool widely used in many areas such as biology, medicine and remote sensing. The use of clustering methods in medicine considerably increases due to rapidly growing amount of data. There have been many applications of these methods in medicine such as clustering of gene expression, diseases risk factor or medical images. For example, clustering methods seek an answer for important questions such as which genes the human genome cause cancer diseases or have in common with other species in bioinformatic research, and they assist medical practitioners in their decision by analyzing of medical images [3, 4] .

The goal of cluster analysis is to partition the objects in a dataset into meaningful classes such that objects in the same class are more similar to each other than to those in other classes. There have been many various parametric or non-parametric methods but parametric methods often suffer from the fact that real datasets often

have non-parametric models. Spectral clustering has overcome this problem due to its ability to cluster without using a parametric model and so it has recently been a popular method. Spectral clustering has a manifold learning algorithm based on eigenvalue decomposition of pairwise similarities of the data points [5–9]. On the one hand its eigenvalue decomposition process provides effective extraction of irregularly shaped clusters, on the other hand it causes its high computational cost ($O(N^3)$, N : number of data points) which makes direct use infeasible for clustering large datasets. To handle this drawback, parallel clustering and approximate spectral clustering (ASC) have been developed. The ASC applies spectral clustering on the data representatives (reduced set of all data) either selected by a data sampling or quantization method [10–13], while the parallel clustering distributes over many computers [14]. The ASC makes not only spectral clustering feasible for large datasets but also different information types (such as data density distribution and topological relations of data representatives possible for similarity definitions between points thanks to data representatives). Although these information types are usually ignored for similarity criteria, they are quite important to determine separation among sub-manifolds [15]. As a similarity measure, Euclidean distance based approach is often used and it may work well for datasets with well-separated clusters but the datasets which have clusters close to each other or inhomogeneous data distribution in the clusters have required using data topology or manifold information [16]. In this study, geodesic based hybrid similarities which use information types including topology, distance and density is proposed and tested on 8 datasets with different characteristics taken from UCI Machine Learning Repository [17]. In addition, a user interface has been designed using MATLAB. This interface allows to use both spectral clustering and approximate spectral clustering as combining similarity measures and sampling/quantization algorithms used in this thesis.

1.2 Organization

This thesis is outlined as follows:

In Chapter 2, we explain clustering analysis and related work. We briefly review clustering algorithms, spectral clustering and its use for large datasets. Then we summarize similarity measures for spectral clustering in previous studies.

In Chapter 3, we describe the ASC algorithm for large datasets. We then discuss the sampling or vector quantization methods used in the first step of ASC and the existing similarity measures for ASC in literature.

In Chapter 4, we propose our geodesic based similarity measures for approximate spectral clustering. We explain approximate spectral clustering and sampling/quantization methods used in our study.

In Chapter 5, we evaluate the performance of our geodesic approach on artificial and real datasets. We show the outperformance of our method using clustering accuracy

Finally, in Chapter 6, we provide our conclusions about future research.

2. A REVIEW OF CLUSTERING METHODS

2.1 Types of Clustering Methods

Clustering methods aim to extract explicit groups in a dataset, without a priori information on their characteristic features. These methods in general are divided into two groups, hierarchical and partitional, as illustrated in Fig. 2.1 [18].

2.1.1 Hierarchical clustering

Hierarchical clustering algorithms generate nested clusters through a series of partitions. These clusters are produced by either agglomerative or divisive. Agglomerative hierarchical clustering firstly assigns each point in the dataset in its own cluster, for example n clusters for n points in the dataset. Then, it merges the most similar pair of clusters at each step. This procedure goes on until all the n clusters are merged in to a single cluster. In contrast, divisive hierarchical clustering starts with one cluster which includes all points in the dataset and recursively divides into smaller clusters. The procedure ends when there are n clusters each of which containing only one data point [19]. Hierarchical clustering produces a dendogram, which presents the arrangement of the clusters, and it can be split at different levels to obtain a desired number of clusters. An illustration is shown in Fig. 2.2 and Fig. 2.3.

Single-Link [20] and Complete-Link [21] are the most well-known hierarchical algorithms. The major difference is that they describe similarity between a pair of clusters with using different way.

2.1.2 Partitional clustering

Partitional clustering methods divide into clusters according to various rules instead of forcing a hierarchical structure. These rules optimize an objective function as being a minimum or maximum depending on the methods. Some methods work iteratively to improve the quality of partitioning where the cluster label of an object is updated

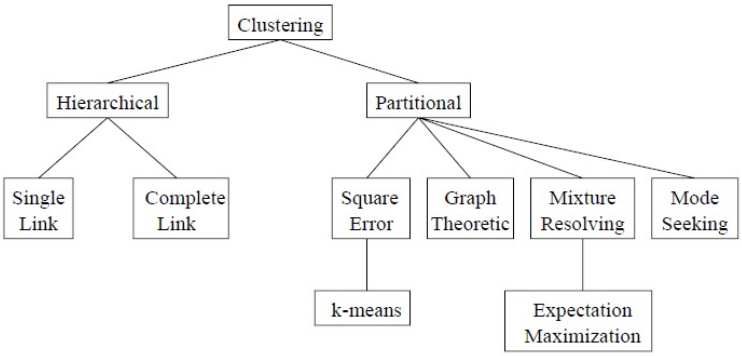

Figure 2.1: A taxonomy of clustering approach.

after each iteration until the algorithm finds the optimum partitions with respect to its objective function. The computation complexity of this algorithm is $Nlog(N)$ for N data points and this makes it feasible for large datasets. K-means is one of the most popular partitioning algorithms although it has several drawbacks such as its need of the number of clusters and depending initial center guess.

Some partitional procedures use the similarity matrix constructed using affinity measures between data points in dataset instead of the data points themselves. In these procedures, the spectrum (eigenvalue and the corresponding eigenvectors) of similarity matrix is used to find a new data representation by a non-linear transformation before clustering. The methods which use these procedures are called spectral clustering methods [2].

2.1.2.1 K-means algorithm

The k-means clustering is well known and preferred due to its easy implementation and accessibility in most clustering packages, and its low computational complexity. This algorithm uses a membership function which allows each data point to be assigned to one cluster such that this function minimizes the distance between the mean of a cluster and the points in that cluster. For a dataset with N samples and K clusters , $X = \{x_i, i = 1,\ldots,N\}$ and $C = \{c_k, k = 1,\ldots,K\}$, the distance between μ_k, (the mean of cluster c_k) and the points in cluster c_k is defined by the squared error as

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \qquad \textbf{(2.1)}$$

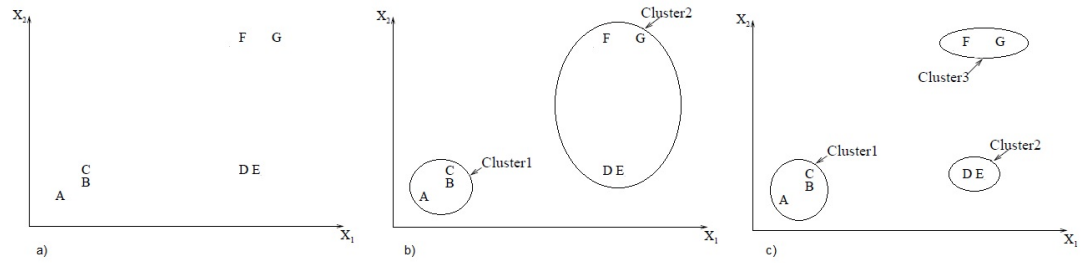

Figure 2.2: (a) A dataset with seven points (A,B,C,D,E,F and G) (b) its two clusters and (c) three clusters according to the dendogram obtained by hierarchical clustering in Fig.2.3.

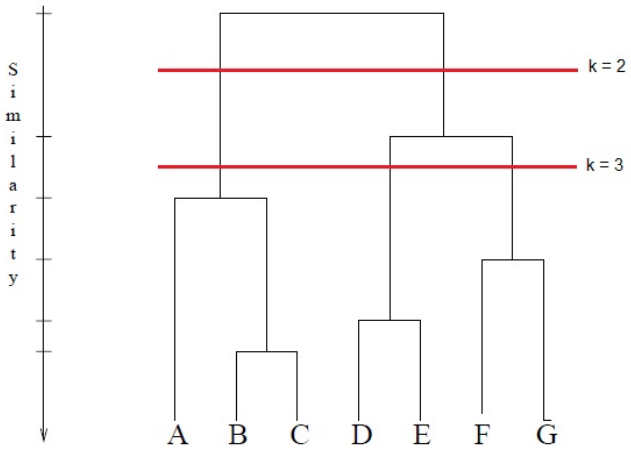

Figure 2.3: The dendogram obtained by the hierarchical clustering algorithm. Result of clustering is shown in Fig.2.2 (b) for k = 2 and Fig.2.2 (c) for k = 3

and k-means minimizes the sum of the squared error over all K clusters as

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{2.2}$$

This algorithm proposed by Forgy [22] is summarized [23] in the following steps:

1. Firstly, k initial centers are selected at random from dataset and each data point is assigned to its closest center.

2. The new cluster centers are found by calculating the mean of each cluster.

3. A new partition is generated by assigning each data point to its nearest cluster center

4. Step 2 and step 3 are repeated until cluster membership stabilizes or stopping criterion is satisfied.

Fig.2.4 shows an example of the K-means algorithm.

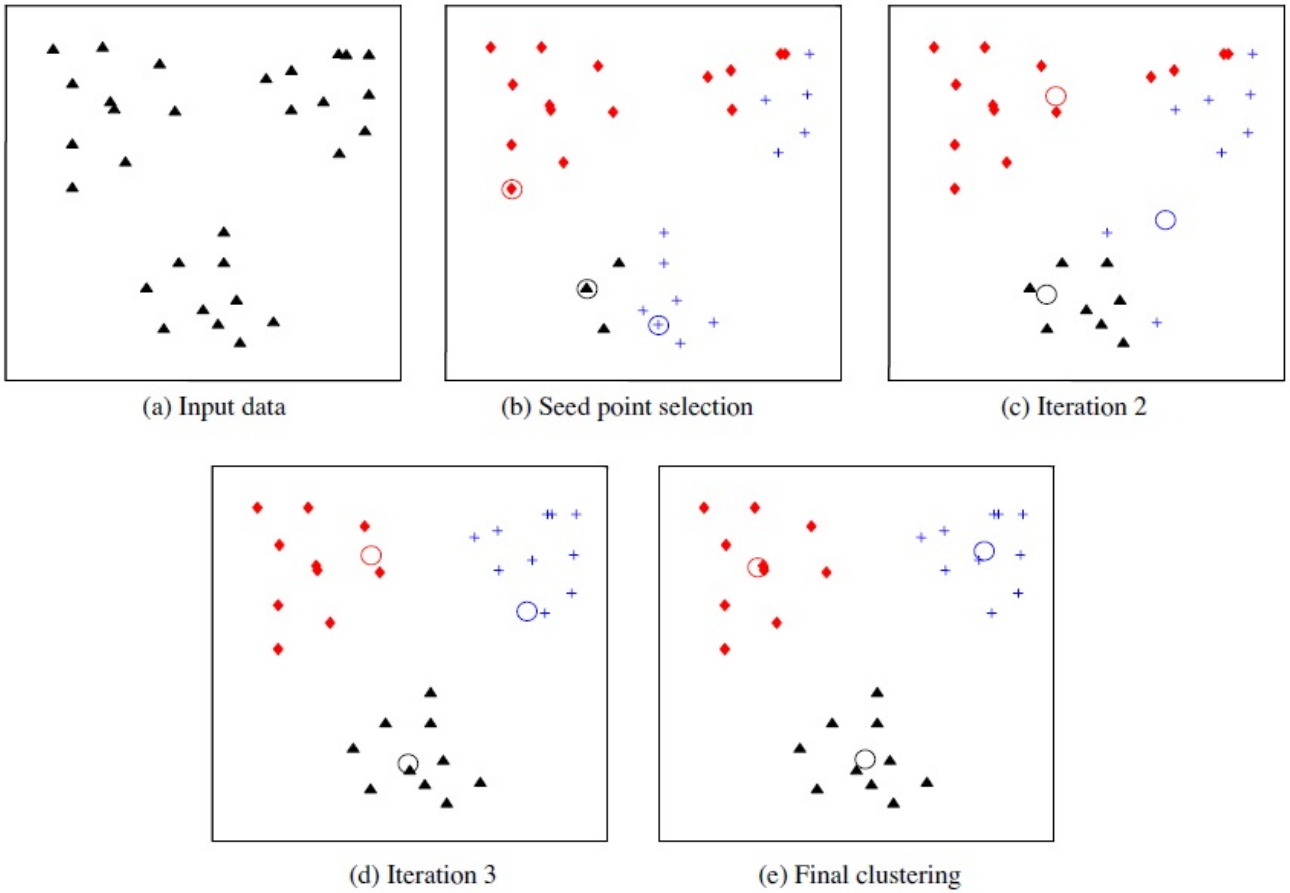

Figure 2.4: An example of k-means algorithm. (a) A dataset with three clusters; (b) Initial points for cluster centers and initial clusters; (c) iteration 2 (d) iteration 3 the updated cluster labels and cluster centers; (e) final clustering

The K-Means algorithm uses three user-specified parameters, number of K, cluster initialization and distance metric. The most important choice is the number of K. There are various methods for this choice, but in general K-means run for different value of K and the value which extracts the most meaningful clusters is selected for K. Cluster initialization is critical for final clustering performance. Unfavorable initialization results in poor clustering. To overcome this, some initialization algorithms are developed such as K-means++ [24]. Finally, different distance metrics such as Euclidean distance, L_1 distance or Mahalanobis distance are used to calculate the distance between cluster centers and data points. They are preferred according to data type, for example K-means with Euclidean distance metric finds spherical or ball-shaped clusters in data while Mahalanobis distance metric has been used to detect hyper-ellipsoidal clusters. In general, Euclidean metric is used due to its lower computational cost.

2.1.2.2 Spectral clustering

Spectral clustering (SC), sometimes called as graph theoretic clustering, uses a graph weighted by pair-wise similarities of the edges connecting the vertices to separate datasets into clusters. This method has a manifold learning algorithm based on eigenvalue decomposition of this graph. Spectral clustering does not make any assumption on the form of the data clusters and it treats the data clustering as a graph partitioning problem. It is very simple to implement because it can be solved efficiently by standard linear algebra methods. It has been popular thanks to its independence from parametric cluster models and its ability to extract irregularly shaped clusters. Spectral clustering involves the following stages; constructing the (Laplacian) graph based on the similarity matrix of the dataset, spectral representation and clustering.

Graph and Similarity Matrix Construction :

A graph which represents data consists of vertices and edges that connect them. Graph theory has been defined diversely, but in the most common of them is that a graph "G" includes a set of V of vertices (nodes) together with a set E of edges (lines) and is shown $G = (V, E)$. Each vertex v_i represents a data point x_i and edges represent neighborhood relationships between data points. In a dataset that consists of $X = \{x_i, \ldots, x_N\}$, similarity $s_{i,j}$ between data points x_i and x_j are found by using

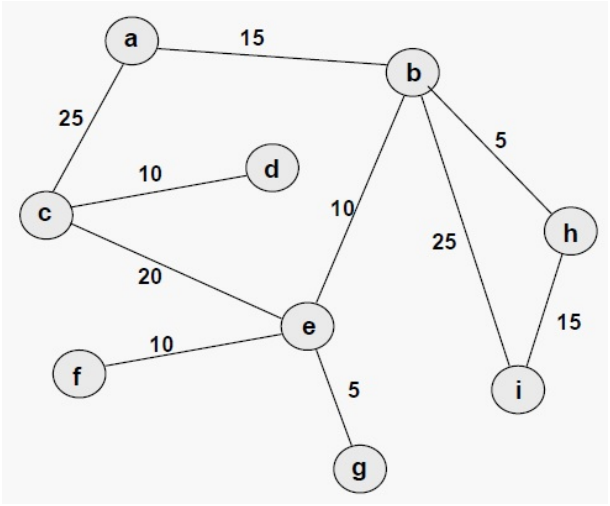

Figure 2.5: An example of a simple weighted graph with nine nodes (vertices) and ten edges.

various similarity measures. If the similarity $s_{i,j}$ is larger than a certain threshold two vertices v_i and v_j are connected and the edge is weighted by $s_{i,j}$. So the weighted adjacency matrix $W = (w_i j)i, j = 1,...,N$ of the graph is obtained. Fig. 2.5 shows an illustration of a weighted-graph of a dataset. Graph may be undirected or directed. In an undirected graph, edges do not orientation while in a directed graph, they have orientation such as from x to y for two points x and y in a dataset. An example is given in Fig.2.6 for an undirected graph and a directed graph.

There are several popular similarity graphs to construct weighted adjacency matrix (similarity matrix). Similarity graphs, which aim to model the local neighborhood relationships between the data points, are regularly used in spectral clustering.

The e-neighbourhood graph: All data points whose pairwise distances are smaller than ε are connected. This graph is usually an unweighted graph and has limited information about the data. It has a drawback such that choosing a useful parameter ε is difficult. If a dataset has points on different scales, the distances between points are different in different regions of the space.

The k-nearest neighborhood graphs: Vertex v_i with vertex v_j are connected provided that v_j is among the k nearest neighbors of v_i. In this method, the neighborhood relationship which is not symmetric produces a direct graph. This graph is converted to an undirected graph by two ways. The first way is that if anyone of v_i and v_j is among the k-nearest neighbors of the other they are connected with an undirected edge. This graph is the *k-nearest neighbor graph*. The second way is that v_i and

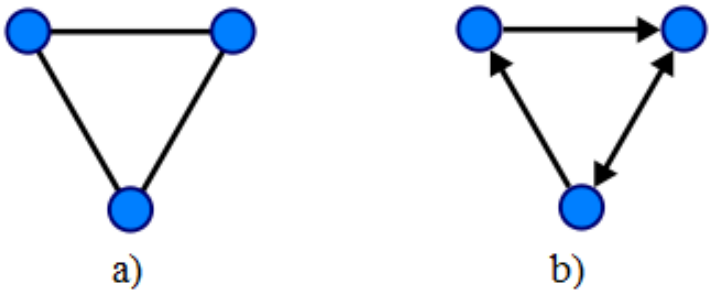

Figure 2.6: An example of a graph with three nodes (vertices) and three edges. (a) An undirected graph; (b) A directed graph

v_j are connected provided that both v_i is among the k-nearest neighbors of v_j and v_j is among the k-nearest neighbors of v_i. This graph is named *the mutual k-nearest neighbor graph.*

The fully connected graph: In this method, all data points have positive similarity with each other are connected. Note that, a local neighborhood based similarity function should be chosen to use this method because the similarity graph should give an opinion about the local neighborhood relationships of data points. One of the similarity function commonly preferred is the distance based Gaussian similarity function in Eq. 2.3. The parameter σ controls the width of neighborhood.

$$s(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{2.3}$$

Spectral Representation and Clustering :

Clustering aims to separate data points in different groups using similarity graph. For this it looks for a partition of the graph such that the edges within a group have high weight (which means that points within the same cluster are similar to each other) and the edges between different groups have a very low weight (which means that points in different clusters are dissimilar from each other). This is a graph partitioning problem and various approaches exist to solve this problem. The minimum cut algorithm solves this problem for a given number of k subsets as choosing a partition $A_1, A_2, \ldots, A_k$ which minimizes

$$cut(A_1, A_2, \ldots, A_k) = \frac{1}{2}\sum_{i=1}^{k} W(A_i, \bar{A_i}) \tag{2.4}$$

Here W is an adjacency matrix, $\bar{A_i}$ is the complement of A and $W(A_i,\bar{A_i}) = \sum_{i\in A, j\in\bar{A}} w_{i,j}$. The minimum cut algorithm often results in clusters of imbalanced sizes. It may divide one individual vertex from the rest of graph but clusters should be reasonably large groups of data points. The two algorithms is common to solve this "reasonably large" problem are RatioCut proposed by Hagen and Hahng and the normalized cut Ncut proposed by Shi and Malik. The RatioCut uses the number of vertices $|A|$ to measure the size of a subset A and it is defined;

$$RatioCut(A_1,A_2,\ldots,A_k) = \frac{1}{2}\sum_{i=1}^{k}\frac{W(A_i,\bar{A_i})}{|A_i|} = \sum_{i=1}^{k}\frac{cut(A_i,\bar{A_i})}{|A_i|} \tag{2.5}$$

The other function Ncut uses the weights of edges $vol(A)$ to measure the size of a subset A and it is defined;

$$Ncut(A_1,A_2,\ldots,A_k) = \frac{1}{2}\sum_{i=1}^{k}\frac{W(A_i,\bar{A_i})}{vol(A_i)} = \sum_{i=1}^{k}\frac{cut(A_i,\bar{A_i})}{vol(A_i)} \tag{2.6}$$

Both algorithms achieve clusters of balanced sizes but they make the simple mincut problem become NP hard. Constructing a graph Laplacian is used to relax this problem. The unnormalized spectral clustering (which uses the unnormalized graph Laplacian) represents relaxing RatioCut and the normalized spectral clustering (which uses the normalized graph Laplacians) represents relaxing Ncut.

The main step for spectral clustering is to construct a graph Laplacian matrix, L. There have been different methods to obtain L and these methods are divided into two groups, unnormalized graph Laplacian and the normalized graph Laplacians. Similarity matrix (weighted adjacency matrix) S and diagonal degree matrix D are used to construct these Laplacians. The degree matrix D is a diagonal matrix with the degrees $d_1,...,d_N$ on the diagonal. The degree of the vertice i is the total similarities of it and is defined as follows

$$d_i = \sum_{j=1}^{N} s_{ij} \tag{2.7}$$

The unnormalized graph Laplacian matrix is defined as

$$L = D - W \tag{2.8}$$

L is symmetric because of the symmetry of W and D. Its the smallest eigenvalue is 0, the corresponding eigenvector is the constant one vector 1.

There two types of normalized graph Laplacians in the literature, L_{sym} and L_{rw} and they are defined as

$$L_{sym} = D^{-1/2} W D^{-1/2} \tag{2.9}$$

$$L_{rw} = D^{-1} W \tag{2.10}$$

The first matrix L_{sym} is a symmetric matrix and the other matrix L_{rw} is closely related to at random walk. Both of them are positive semi-definite and have non-negative real-valued eigenvalues. Their properties are detailed in [8]. Several studies [8, 25] show that there is no clear advantage among different spectral methods as long as a normalized graph Laplacian is considered. Therefore, the normalized Laplacian matrix L_{sym} proposed by Ng et al. [6] can be used. The process from constructing the normalized graph Laplacian to obtain k clusters will be explained according to the algorithm proposed by Ng et al. [6].

After producing the normalized Laplacian matrix L_{norm} (L_{sym}), the k largest eigenvectors $E = \{e_1, e_2, ..., e_k\}$ of L_{norm}, associated with the k greatest eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_k)$. are found. $E \in R^{N \times k}$ is the matrix containing the vectors $e_1, e_2, ..., e_k$ as columns, where N is the number of data point and k is the number of cluster. The matrix $U \in R^{N \times k}$ is obtained from the matrix E by normalizing the rows to norm 1 as follows

$$u_{ij} = \frac{e_{ij}}{\sqrt{\sum_k e_{ik}^2}} \tag{2.11}$$

Finally the N rows of U are clustered with the k-means algorithm into k clusters. In graph Laplacian approach, the main trick is to change the representation of the data points x_i to points u_i and $U \in R^k$ because this change of representation provides enhancement of the cluster-properties in the data. Therefore simple clustering algorithms such as k-means have no difficulties to detect the clusters in this new representation.

2.2 Spectral Clustering For Large Datasets

Although spectral clustering is an efficient and simple way for extracting irregularly shaped clusters, it has a drawback of high computational cost ($O(N^3)$, *N :number of data points*) because of eigendecomposition of the graph Laplacian matrix, *L*. It needs to construct a similarity matrix ($N \times N$) for eigendecomposition. Similarity matrix compares all possible pairs of data points and this process is computationally expensive for a large data set. This makes direct use of spectral clustering infeasible for clustering large datasets. In order to utilize its advantages in large datasets various methods are developed.

One approach is the use of parallel clustering. This method parallelizes both memory use and computation on distributed computers. First *n* data instances are allotted onto *p* distributed machine nodes. The similarities between local data are computed on each node and all are set in a way that uses minimal disk I/O. These two steps are applied on all steps of spectral clustering as parallel eigendecomposition and parallel k-means [14]. Although it speeds up clustering time significantly, it has a drawback of requiring additional resources that should be scaled according to the size of the dataset [26].

Another approach is approximate spectral clustering (ASC). This method applies spectral clustering on the data representatives, either selected by a sampling approach or data quantization from dataset, instead of all data points. In this process, the most critical point is to find a suitable sampling or quantization method to obtain data representatives with similarity measures. Sampling methods select real data among all data points in dataset for data representation while quantization methods use representatives of real data points. Fowlkes et al. [10] use random sampling based on Nystrom method while Bezdek et al. [11] propose a progressive sampling which has a drawback of tendency to oversample. Wang et al. [13] propose selective sampling and point out that it is the best sampling method and has a similar success with k-means quantization. Yan et al. [27] use k-means and random projection trees as quantization methods. They show experimentally that in approximate spectral clustering, the best way is vector quantization with minimum distortion in order to obtain data representatives. Besides these experimental results, theoretical justification for using quantization with minimum distortion exist [28]. Taşdemir [26] researches

alternative quantization methods (neural networks: self-organizing maps (SOM) and neural gas [29]) and compares with k-means. This research shows that neural gas quantization outperforms in approximate spectral clustering. Alternatively, Yalçın and Taşdemir [30] show that k-means ++ [24], a successful variant of k- means with a novel probabilistic approach for initialization, is a good alternative for quantization in ASC.

3. APPROXIMATE SPECTRAL CLUSTERING (ASC)

Approximate spectral clustering (ASC) is one of the methods to provide using spectral clustering for large datasets as mention in section 2.2. It uses data representatives as input data for spectral clustering and applies all steps of spectral clustering on these reducing data points. Thanks to using data representatives instead of all data points the ASC provides an efficient and simple solution. The ASC has three main steps; generating of data representatives by quantization or sampling, spectral clustering of data representatives, assign the labels of the data representatives to their corresponding data points

An ASC algorithm (Fig. 3.1) to find k clusters can be summarized as follows :

1. Produce n data representatives either by vector quantization or sampling for a dataset with N data points,

2. Construct a similarity matrix S based on a user-set similarity criterion and calculate the degree matrix D Eq. (2.7) using the similarity matrix . Similarity matrix shows the pairwise similarities of these n data representatives as section **??**

3. Construct L_{norm} matrix Eq. (2.9) using the similarity matrix S and the degree matrix D and find the k eigenvectors $(e_1, e_2, ..., e_k)$ of L_{norm}, associated with the k greatest eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_k\}$.

4. Construct the $n \times k$ matrix $E = [e_1, e_2, ..., e_k]$ and obtain $n \times k$ matrix U by normalizing the rows of E to have norm 1, i.e. section 2.11

5. Cluster the n rows of U with any simple clustering method such as the k-means algorithm into k clusters. Finally, assign the labels of the n representatives to their corresponding data points.

First and last steps of the ASC algorithm are the different with the SC, while steps 2–6 of ASC algorithm are the same with the traditional spectral clustering defined in [6].

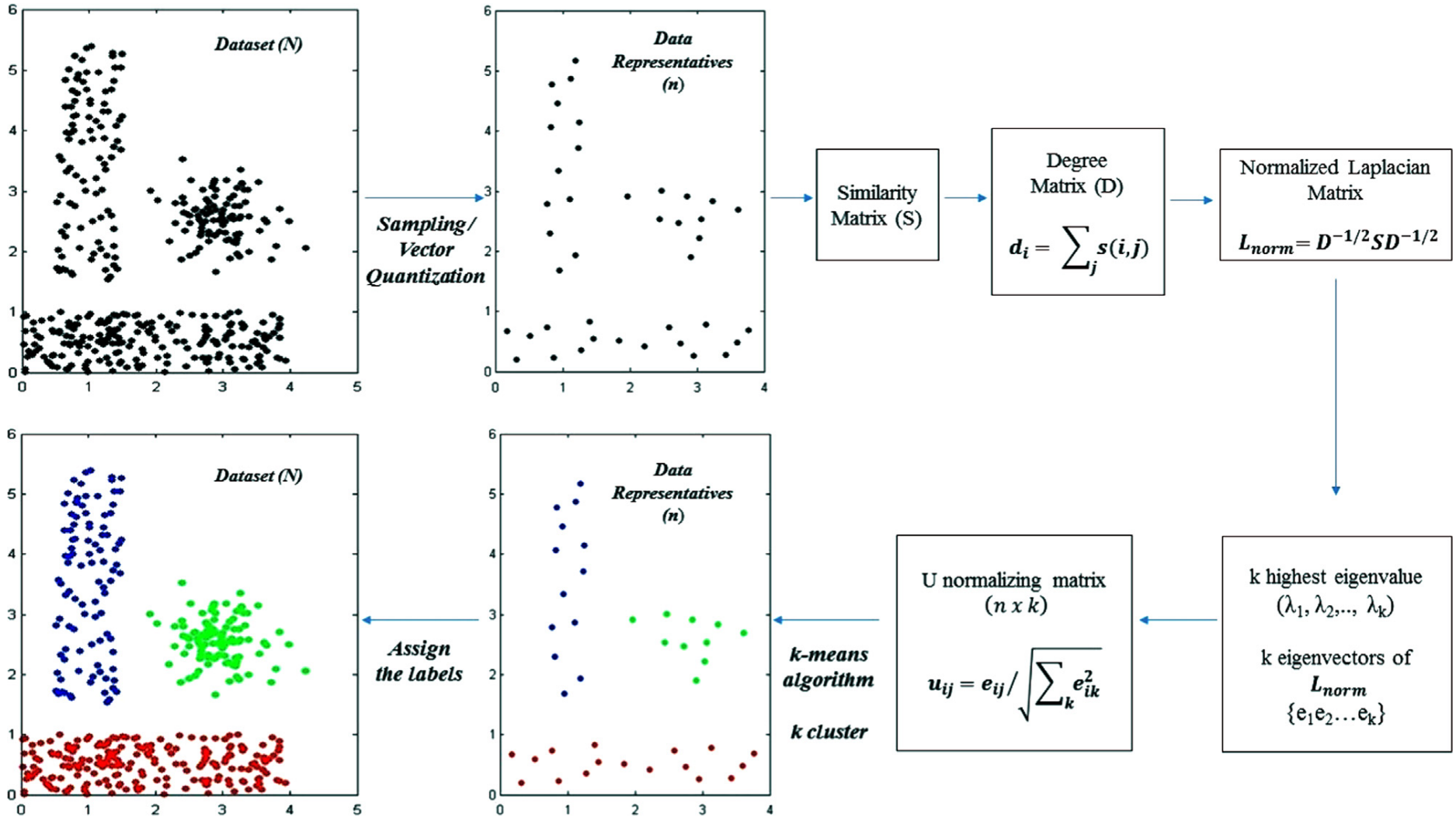

Figure 3.1: Steps of approximate spectral clustering (ASC) algorithm are showed on Lsun dataset obtained from [1]. The dataset has 2 features and 3 clusters.

3.1 Generating Data Representatives for ASC

For the first step, there have been various algorithms as mentioned in section2.2 but in this thesis three different approaches selective sampling, neural gas and k-means++ have been used because selective sampling is the best sampling method [13], neural gas is the best way for vector quantization [26] and k-means++ is a good alternative for quantization [30].

3.1.1 Neural gas

The neural gas [31] is an iterative algorithm based on neural learning and it produces data representatives as preserving data topology by quantization of the data points. It has an adaptive learning process which consists steps of random selection of v, finding the best maching unit (BMU) for v , and update. Firstly, a data points v randomly is selected from the dataset M and then the BMU for v is found by the minimum Euclidean distance in Eq.3.1.

$$||v - w_i|| \leq ||v - w_j|| \tag{3.1}$$

The neighborhood function $h_\tau(w_j)$ in Eq.3.2 is used to determine neighbor data representatives w_j of w_i.

$$h_\tau(w_j) = \exp(-\rho_{w_i}/\tau) \tag{3.2}$$

ρ_{w_j} is the ranks of the distances of neural unit w_j to data sample v and the neighborhood is defined in the data space according to this value. ρ_{w_j} is 0 for the BMU w_i. Finally, the BMU w_i and its neighbor prototypes w_j are adapted by an iterative learning process as in Eq.3.3.

$$w_j(t+1) = w_j(t) + \alpha(t)h_\tau(w_j)(v - w_j(t)) \tag{3.3}$$

In here, $\alpha(t)$ is learning parameter decaying by time and $\alpha(t) \in [0,1]$. The adaptive learning is repeated until a predefined stop criterion is achieved. In each iteration, ρ_{w_j}

is recalculated for each randomly selected v. After learning, the neural units become the data representatives.

3.1.2 K-means ++

K-means is fast and simple clustering algorithm and its goal is to minimize the average squared distance between points in the same cluster. For this, firstly, it chooses initial k cluster centers uniformly at random from dataset and finds the closest center for each data in the dataset. Then centers are updated by minimizing the total squared distance between each point and its closest center. It may stuck in local minimum when random initialization of the centers are adverse and produces a poor clustering. K-means++ [24] is proposed to address the initialization problem in k-means. It uses a probabilistic approach for initial cluster centers instead of randomness. Firstly, a center c_1 is selected uniformly at random from the dataset M. Then, other centers c_i, $(i = 2, \ldots, k)$ are randomly selected from a probability distribution function Eq. (3.4).

$$P\left(\frac{d^2(v, c_v)}{\sum_{v \in M} d^2(v, c_v)}\right) \tag{3.4}$$

In here, $d(v, c_v)$ shows distance of the nearest cluster center c_v to a data point v. After choosing of k initial centers, standard k-means algorithm is executed. In this thesis k-means ++ is used to obtain data representatives because of the outperformance of k-means++ over k-means [24].

3.1.3 Selective sampling

Various sampling methods have been employed for clustering of large datasets to make spectral clustering feasible [10, 27]. Among them, random sampling (RS) is preferred because of being the fastest sampling algorithm despite of its highly error rates. Progressive sampling [32] controlled by statistical divergence test has been more effective than random sampling. Bezdek et al. [11] have been developed eNERF (extended Non-Euclidean Releational Fuzzy c-means) method by using progressive sampling, but Wang et al. [12] show that e-NERF is not useful for large datasets in practice because of samples size required by the divergence test. Therefore, they propose a new method called selective sampling (SS) by combining approaches from [11], [33] with simple random sampling. Besides this Wang et al. [13] show that the

selective sampling outperforms all sampling algorithms for ASC and it has a similar performance with k-means quantization.

The SS algorithm consists of three steps. Firstly, the dissimilarity matrix (D_{NN}) (N is size of dataset) of dataset and the h distinguished objects $p_1, p_2, \cdots, p_h$ are selected from this matrix. The choice of distinguished objects is determined by distinguished features (DF) algorithm in [11]. The first index p_1 is randomly selected from the index set 1,2,...,N. Then, the search array S is generated by Eq. 3.5. For $p_1 = 1$:

$$A = (a_1, a_2, \cdots, a_N) = (d_{1,1}, d_{1,2}, \cdots, d_{1,N}) \tag{3.5}$$

Other distinguished objects (p_i with $i \in [2,h]$) are iteratively selected and the search array A is updated as follows

$$p_i = \arg\max_j a_j \tag{3.6}$$

$$S = (\min(s_1, d_{p_{i-1},N}), \cdots, \min(s_N, d_{p_{i-1},N})) \tag{3.7}$$

so that all distinguished h objects are obtained by using $\max-\breve{}\min$ farthest point strategy for $i = 2, \cdots, h$. In the second step, each vector v_i in the dataset $\{v_1, v_2, \cdots, v_N\}$ is assigned to the nearest distinguished object q using

$$q = \arg\min_j (d_{p_j}, i) \tag{3.8}$$

and the receptive fields (index sets) of the h objects are obtained by adding the corresponding data vector to $R_q = R_q \cup \{i\}$. In the third step, $n = \sum_q n_q$ samples are randomly selected from these subsets R_qs, proportional to the number of samples in $|R_q|$:

$$n_q = n \times |R_q|/N \tag{3.9}$$

3.2 Similarity Measures for ASC

One of the most important steps is to decide similarity criterion to construct similarity matrix and a common approach is using a Gaussian kernel based on the (Euclidean) distances in Eq. 3.10.

$$s_{Euc}(i,j) = \exp(-\frac{d_{Euc}(x_i,x_j)}{2\sigma^2}) \tag{3.10}$$

In here $d_{Euc}(x_i,x_j)$ is the Euclidean distance of data points x_i and x_j and σ is a decaying parameter to be set optimally by experiments [6]. Another similarity criterion common-near-neighbor (CNN) has been defined by Zhang et al. [34] and it considers the number of data points in the intersection of ε-neighborhoods of x_i and x_j as follows

$$s_{CNN}(i,j) = \exp(-\frac{d_{Euc}(x_i,x_j)}{2\sigma^2(CNN(i,j)+1)}) \tag{3.11}$$

The CNN similarity criterion has high computational cost for large datasets although it can perform high accuracies.

Actually, the ASC method has a very important advantage by reveal new information types such as local density and data topology thanks to using data representatives(prototypes) instead of all data points as well as by making the SC feasible for large datasets. Taşdemir and Merényi [35] proposes an alternative similarity criterion the connectivity matrix (CONN) which uses these new information types. This criterion represents the neighborhood relationships of two prototypes w_i and w_j in a set of prototypes $W = \{w_1, w_2 \mathrm{d} \ldots w_n\}$ and it is based on Voronoi polygons of prototypes which separate the data space into groups. These groups are called Voronoi polygons and each prototype is center of group. For five of prototypes in Fig. 3.2 their Voronoi polygons are showed in Fig. 3.3

In the CONN similarity criteria, firstly the best matching (the closest) and the second-best-matching(the second closest) prototypes are found for each data point v in a dataset M respectively by Eq. 3.12 and Eq. 3.13. Each prototype is the closest prototype of data points into its own Voronoi polygon.

$$V_i = \{v \in M : \|v - w_i\| \leq \|v - w_j\|\} \tag{3.12}$$

$$V_{ij} = \{v \in V_i : \|v - w_j\| \leq \|v - w_k\| \forall k \neq i\} \tag{3.13}$$

In here, V_i is the set of data points v for which w_i is the closest prototype and called the Voronoi polyhedron of w_i. V_{ij} is the set of data points v in the V_i for which w_j is the

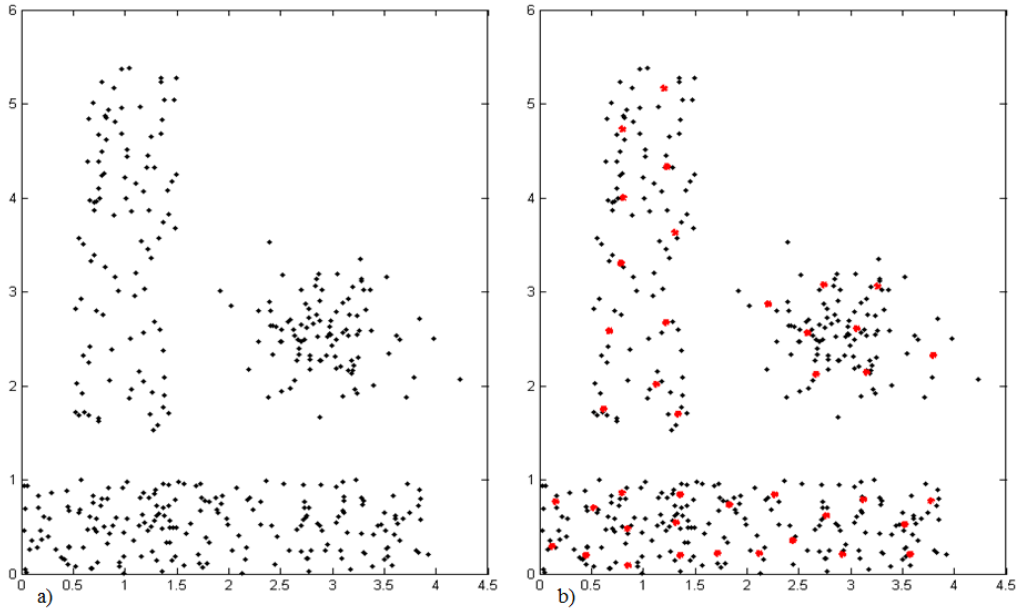

Figure 3.2: The obtained data representatives for a 2D 3-cluster dataset (Fig. 3.1) (a) data points and (b) prototypes (its red stars)

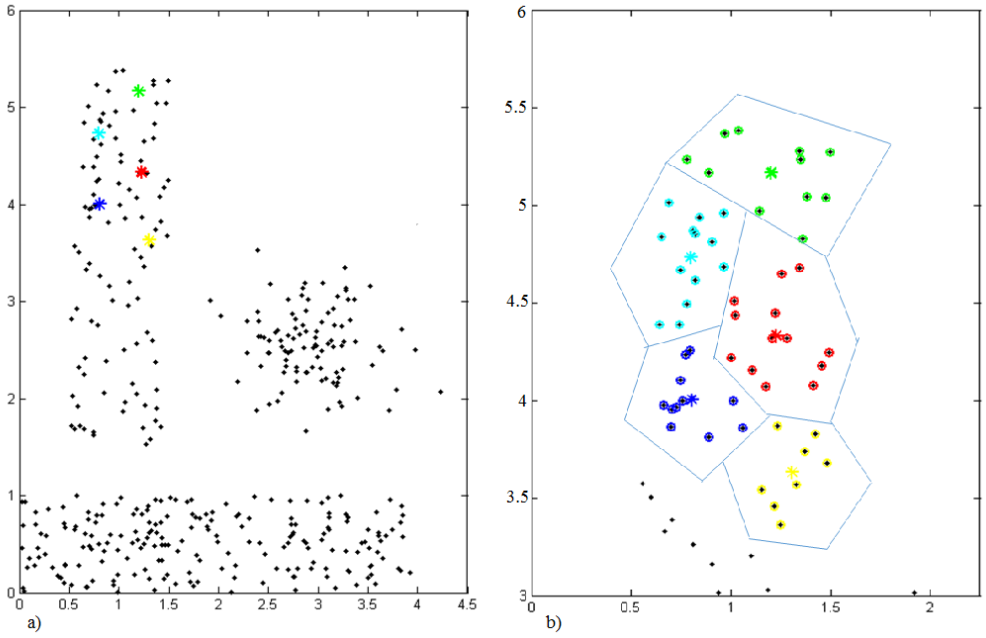

Figure 3.3: Separation of a dataset into its Voronoi polygons a) Five of prototypes in Fig.3.2 b) The Voronoi polygons with the data points mapped to them of these five prototypes. Prototypes are shown by stars and their data points are shown by the circles. (Each color represents a different prototype and its corresponding data points).

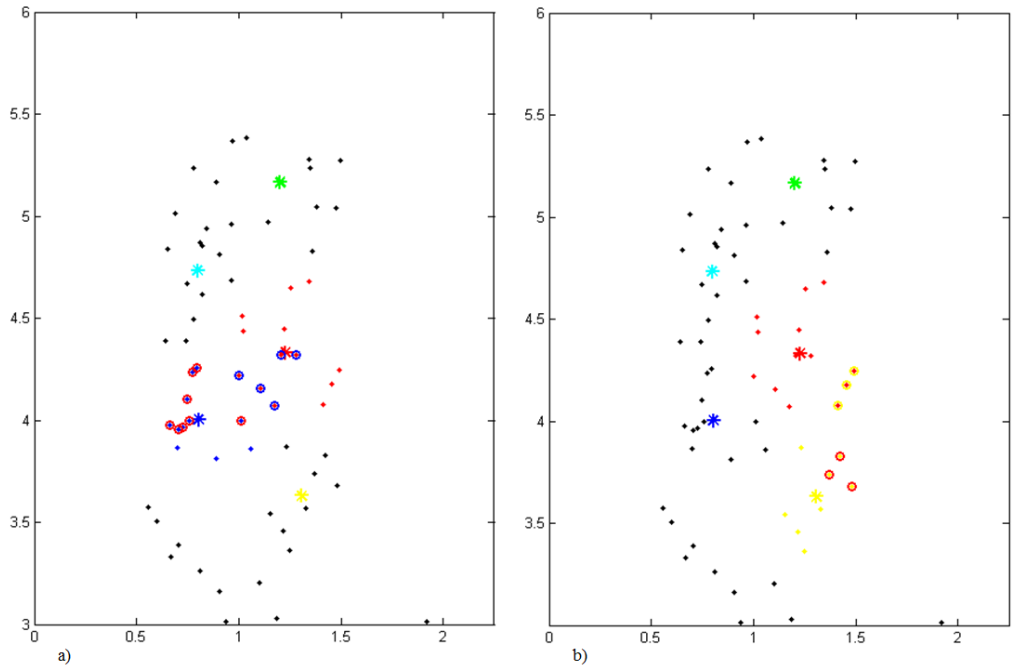

Figure 3.4: An example of sub-Voronoi polyhedra of a prototype. a) $V_{1,2}$ and $V_{2,1}$ sub-Voronoi polyhedra for w_1, w_2 prototypes b) $V_{1,3}$ and $V_{3,1}$ sub-Voronoi polyhedra for w_1, w_3 prototypes.

second closest prototype and called sub-Voronoi polyhedron of w_i . The connectivity matrix CONN is the number of data points in the sub-Voronoi polyhedra $V_{ij} \cup V_{ji}$ as in Eq.3.14. An example of sub-Voronoi polyhedra of a prototype is given in Fig. 3.4.

$$CONN(i, j) = |V_{ij} \cup V_{ji}| \tag{3.14}$$

In Fig. 3.4, red, blue and yellow stars represent respectively w_1, w_2 and w_3 prototypes and the best matching (the closest) prototype for red points(its Voronoi polyhedron) is w_1 represented by red star. In left column (a) V_{12} consists of five red points in blue circles (the closest prototype is w_1 and the second closest prototype is w_2) and V_{21} consists of eight blue points in red circles (the closest prototype is w_2 and the second closest prototype is w_1). In this situation CONN(1,2) = CONN(2,1) = 5 + 8 = 13. In right column (b) V_{13} consists of three red points in yellow circles and V_{31} consists of three yellow points in red circles. In this situation CONN(1,3) = CONN(3,1) = 3 + 3 = 6.

In case CONN (i,j) is used as weighted adjacency matrix, it shows data points distribution within the Voronoi polyhedra of prototypes with respect to the prototypes neighbor according to data manifold. If CONN (i,j) is greater than zero w_i and

w_j are neighbors and CONN (i,j) shows bond strength (similarity degree) of these prototypes. If CONN (i,j) = 0, w_i and w_j are not neighbors and dissimilar each other. It is shown that CONN outperforms distance based approaches [26, 35]. Besides this, Taşdemir [36] proposes a hybrid similarity criterion S_{hyb} in Eq. 3.15 which uses both distance and density information in similarity definition for ASC.

$$s_{hyb}(i,j) = s_{Euc}(i,j) \times \exp\left(\frac{CONN(i,j)}{\max_{i,j} CONN(i,j)}\right) \tag{3.15}$$

S_{hyb} scales the distance based similarity S in Eq. 3.10 by $[1,e]$ according to an exponential based on the local density and topology measure $CONN$. This exponential merging limits the big variation of $CONN(i,j)$, enhancing the distance-based similarity for neighbor prototypes.

4. PROPOSED GEODESIC BASED HYBRID SIMILARITY MEASURES FOR ASC

One of the most important advantages of spectral partitioning is its figuration of data points in another space such that it transforms data points from their own space (data space) to Laplace space (cluster space) using eigendecomposition of a graph Laplacian matrix, L. Thanks to this new figuration makes compact and nested datasets more separated, any simple clustering algorithm such as k-means can achieve a good clustering. The better definition pairwise similarities, the more separated data points according to their clusters in new space. Therefore one of the most critical steps for SC/ASC is that how to define pairwise similarities and construct similarity matrix. In clustering, the similarity is defined on the space spanned by the datasets rather than the manifold itself and the Euclidean distance based Gaussian kernel approach is commonly used as being similarity measure. This approach may work well for datasets which have well-separated clusters, but it has poorly clustering accuracy in case that datasets have clusters close to each other or with inhomogeneous distribution of data points with in the clusters. Accordingly, the use of data topology or manifold information is necessary to effectively extract boundaries of the underlying structures (clusters) in the data for this kind of datasets [16]. The geodesic approach originated from geodesy (the science of measuring the size and shape of the Earth) is a traditional way of calculating topological distances. The geodesic distance indicates the shortest path distance between the two points on the Earth as in Fig. 4.1. This term is generalized for other applications to represent manifold based distances for clustering. To construct the similarity matrix by distance based Gaussian kernel, the use of geodesic based distance approach thanks to its use of data topology information is more successful to show pairwise similarities than the use of euclidean based distance approach for distance information in the distance based Gaussian kernel. The similarity matrix obtained by using these two distance approaches for same set of prototypes of the dataset in Fig. 3.1 have been compared in Fig. 4.2. The three clusters in the similarity matrix obtained using geodesic approach are clearly visible while

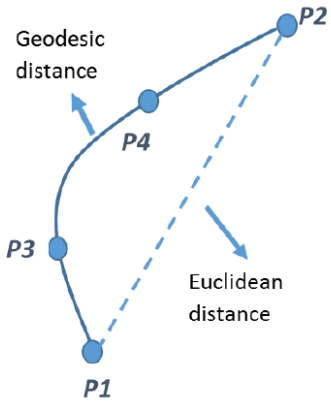

Figure 4.1: A demonstration of difference between geodesic distance and Euclidean distance. The Euclidean approach calculates pairwise (P1 and P2) distance with respect to the data space while geodesic approach calculates this distance through P3 and P4 considering the data manifold.

some within-cluster distances are higher than between-clusters distances although the separation can also seen to a great extent using Euclidean distance.

The constructing neighborhood graph is necessary part in calculating the geodesic distance since two prototypes (or data points) are connected through their neighbors according to geodesic approach and a neighbor graph show the connections between prototypes (or data points). A traditionally way to construct this graph is k nearest neighbor (k-nn) approach mentioned in section **??**. According to this method if any pair of prototype w_i and w_j are in the k-nn set of each other, they are neighbors. In this thesis, the mutual k-nn approach is used and an undirected graph is obtained. The geodesic distance between any two prototypes w_i and w_j is sum of Euclidean distances of their shortest path on this neighbor graph :

$$d_{geoknn}(w_i, w_j) = \sum_{l,m \in SP_{knn}(w_i,w_j)} d_{Euc}(l,m) \tag{4.1}$$

$SP_{knn}(w_i, w_j)$ is the set of edges in the shortest path between w_i and w_j. If w_i and w_j are first neighbors, the geodesic distance is the Euclidean distance, whereas if they are not in the k-nearest neighbor of each other and there is not any path between w_i and w_j, the geodesic distance is infinite. There are two drawbacks of this approach;

1. It is required that optimum number of nearest neighbors k is decided by user for each dataset

2. The optimal k is different for each data representative in the same dataset.

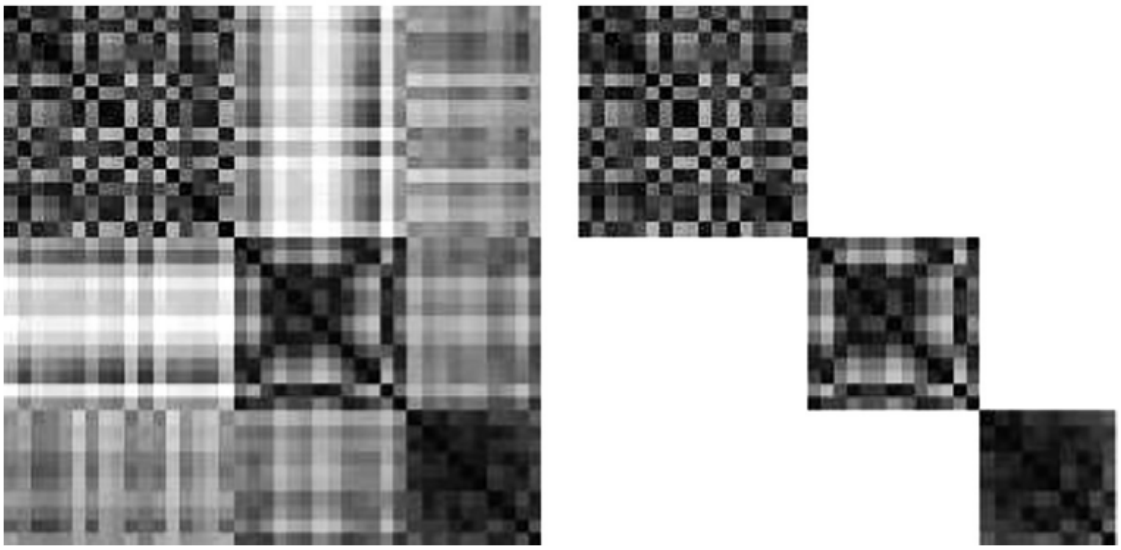

Figure 4.2: (a) (Normalized) Euclidean and (b) geodesic distances between the prototypes of the Lsun dataset shown in Fig. 3.1. Some of the distances between clusters are smaller than some within-cluster distances according to the Euclidean distances, whereas three clusters are clearly separated with respect to geodesic approach.

On the one hand the user-set parameter k causes a useless algorithm, on the other hand inconvenient value of k may result connection between clusters as in Fig.4.3.

To address drawbacks of k-nn neighbor graph, we propose the connectivity graph (CONN) [26] as being an alternative method for neighborhood graph in this thesis. When data topology based CONN graph is used as being a weighted adjacency graph, w_i and w_j prototypes are neighbor if $CONN(i, j) > 0$. Therefore the neighborhood of prototypes are defined with using local data characteristics without any user-set parameter and each prototype has a specific number of neighbors. Thus, the optimal k problem of k-nn approach is solved. The examples of neighborhood graphs obtained by k-nn graph with different values of k and by the CONN graph are showed in Fig. 4.3. k-nn graphs for the same set of data representatives of the dataset in Fig. 3.1 are different from each other for different k values. For k = 3, there is not any connection between clusters, but this results in loose connections within the clusters for data representatives. When k is increased to 5, the number of connections within clusters, but the same time two prototypes belonged to different clusters are connected. When k = 7, the prototypes have tight connections with others in the same cluster, but connection between clusters also increase and so two clusters are tightly connected. Contrary to k-nn graphs, CONN graph more tightly connects data representatives in the

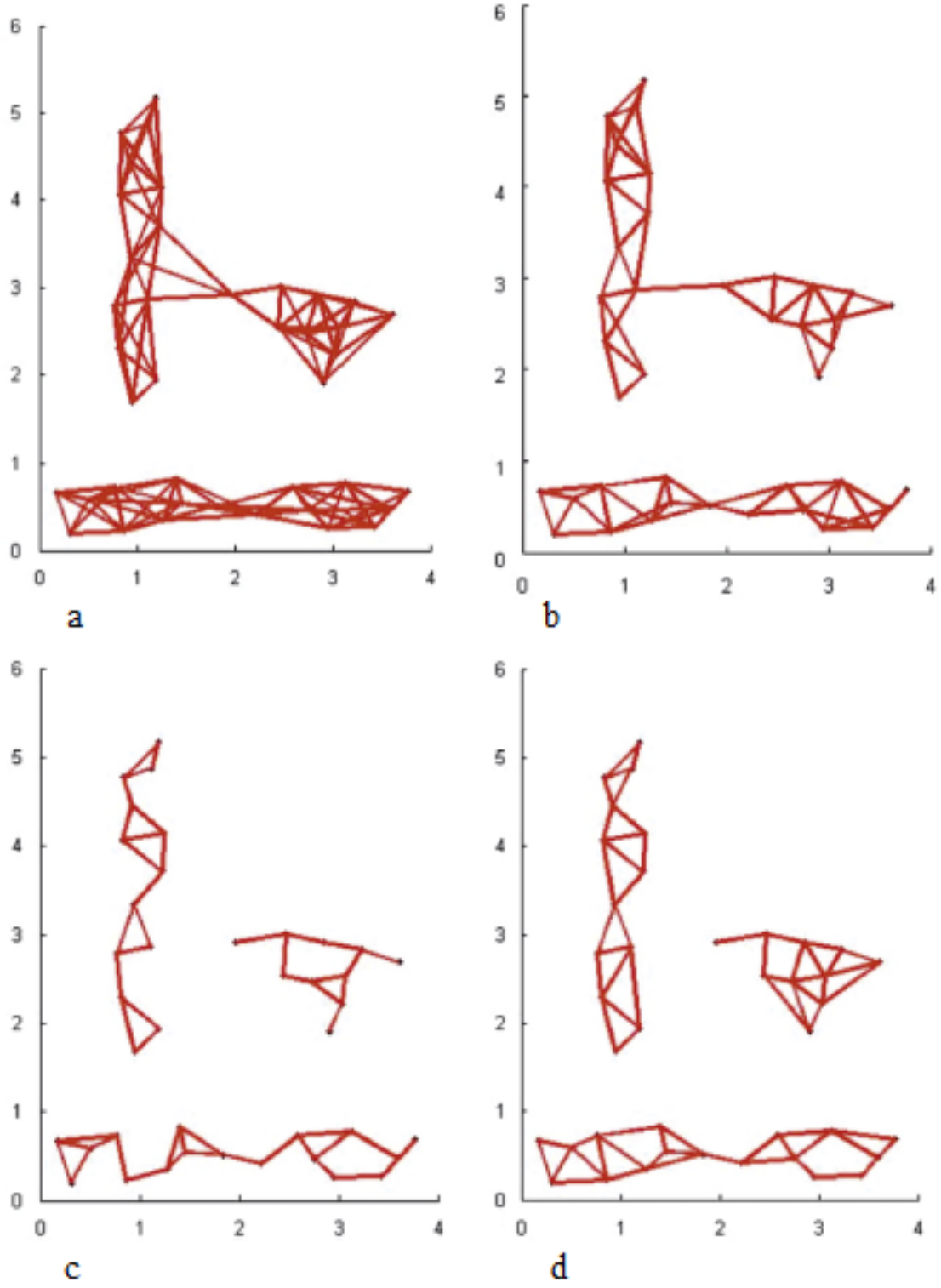

Figure 4.3: Neighborhood graphs obtained by k-nn graph with different values of k and by the CONN graph for data representatives of the dataset in Fig.3.1. The first three graphs are obtained by (mutual) k-nearest neighbor approach: (a) k=7, (b) k=5, (c) k=3, and (d) the last one is obtained by CONN graph. Note that an optimum k should be set to correctly identify submanifolds, where *CONN* enables varying number of neighbors with respect to the data manifold, resulting a clear separation without any user-set parameter.

same cluster without the use any parameter, while it does not produce any connection between-cluster. This is a result of its construction of the neighborhood relations using the data characteristics.

The geodesic distance d_{geoadj} is calculated with using the *CONN* graph and Euclidean distances d_{Euc} as follows

$$d_{geoadj}(w_i, w_j) = \sum_{l,m \in SP_{adj}(w_i,w_j)} d_{Euc}(l,m) \tag{4.2}$$

Here, SP_{adj} shows the set of edges in the shortest path between w_i and w_j and it is found by Euclidean distance and CONN graph. Thanks to the use of CONN graph the correct representation of the data topology on the prototypes, but the local density distribution information is omitted because of the Euclidean based distance. This is one of the most important information revealed by quantization based ASC. The distance between any two prototypes w_i and w_j can be defined by using the local data distribution CONN(i,j) as

$$d_{CONN}(w_i, w_j) = exp\left(-\frac{CONN(i,j)}{\max_{y,z} CONN(y,z)}\right) \tag{4.3}$$

This CONN based distance and CONN graph are used to calculate geodesic distance $d_{geoconn}(w_i, w_j)$ as in Eq.4.4.

$$d_{geoconn}(w_i, w_j) = \sum_{l,m \in SP_{CONN}(w_i,w_j)} d_{CONN}(l,m) \tag{4.4}$$

In here, $SP_{CONN}(w_i, w_j)$ is the set of edges in the shortest path between w_i and w_j. It is found by CONN distance and CONN graph. Although $d_{geoconn}(w_i, w_j)$ uses both the data distribution and data topology, the omitting distances between prototypes is a disadvantage.

To utilize all available information for ASC on the prototype level, all of distance, density and topology should be used. Accordingly, we propose $d_{geohyb}(w_i, w_j)$ which uses CONN neighborhood graph and a hybrid distance approach merging d_{CONN} and d_{Euc}. This geodesic distance is calculated as

$$d_{geohyb}(w_i, w_j) = \sum_{l,m \in SP_{adj}(w_i, w_j)} d_{Euc}(l,m) d_{CONN}(l,m) \quad \textbf{(4.5)}$$

Here, $SP_{hyb}(w_i, w_j)$ is the set of edges in the shortest path between w_i and w_j. It is found by hybrid distance($d_{Euc}(l,m)d_{CONN}(l,m)$) and CONN graph.

All the proposed geodesic distances are used instead of the Euclidean distance in Gaussian kernel based on the distance Eq. (3.10) and geodesic distance based similarity matrices are obtained to construct Laplacian matrix.

5. EXPERIMENTS AND RESULTS

This section shows the clustering results using Geodesic Based Hybrid Similarity Criterion (GeoHSC) on three set of datasets, one well known pattern recognition benchmark in section 5.1.1, one medical datasets in section 5.1.2 and one gray level brain MR images in section 5.1.3. These datasets will be used to compare traditional Euclidean based ASC and proposed GeoHSC based ASC.

5.1 Datasets

5.1.1 PR datasets

In this thesis, we use eight datasets well know pattern recognition benchmark which have different size and are frequently used to compare data analysis techniques. We call PR datasets for them in the remain of thesis. Three of them are artificially generated to have basic clustering challenges [1, 36] and five of them are obtained from UCI Machine Learning Repository [37]. Their properties are listed in Table 5.1 The Lsun dataset is one of the artificial datasets and has three separated clusters with two rectangular (as an L-shape) and one spherical (as a sun inside L-shape) shape, the another artificial dataset Chainlink has two 2D rings entangled in 3D space and the last artificial dataset Wingnut has two clusters very close to each other with heterogeneous within-cluster data distributions. These datasets have basic clustering difficulties. The remained datasets are from UCI Machine Learning Repository and used frequently to demonstrate performance of clustering algorithms. These eight datasets are used as a first step verification for the proposed hybrid similarity criteria.

5.1.2 Medical datasets

The performance of the proposed GeoHSC for ASC of medical datasets are evaluated five medical datasets from UCI Machine Learning Repository. Properties of them are listed in Table 5.2.

Table 5.1: Properties of datasets.

Dataset	Num. of instances	Num. of atributes	Num. of classes
Iris	150	4	3
Lsun	400	2	3
BCWS	699	9	2
Chainlink	1000	3	2
Wingnut	1016	3	2
Yeast	1484	8	10
Statlog	6435	4	6
Pen Digit	10992	16	10

Table 5.2: Properties of medical datasets.

Dataset	Num. of instances	Num. of atributes	Num. of classes
BCWSP	194	33	2
Vertebral	310	6	2
Dermatology	398	34	6
ILP	579	10	2
Biodegration	1055	41	2

5.1.3 Large medical datasets

To show the clustering accuracies of the proposed GeoHSC for ASC of large datasets, four MR images sets from The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) [38] are used. Three of these images sets are simulated as consider [39] and one of them is from glioma patient. All images sets consists of T1, post-Gadolinium T1, and T2 MR images. In addition each simulated images set has two ground truth segmentations, one of which has six clusters and the other has three clusters (1 edemma, 2 active tumor and 3 for everything else) while real images set has only three cluster (1 edemma, 2 active tumor and 3 for everything else) by the manual

Table 5.3: Properties of MR datasets. The datasets with four attributes consist of Flair, T1, post-Gadolinium T1 and T2; the dataset with three attributes is combination of Flair, T1 and T2; the one with two attributes consists of Flair and T2.

Dataset	Num. of instances	Num. of atributes	Num. of classes
SMHG0012	65536 (256x256)	3	2
SMHG0012	65536 (256x256)	3	5
SMLG0013	65536 (256x256)	3	6
SMHG0019	65536 (256x256)	3	6

segmentation. For each set, images are combined with using different attributes (T1, T2, and post-Gadolinium T1). Two of simulated images sets are for low grade (LG) glioma and the other is for high grade glioma, while real images set is obtained from a high grade glioma patient. Properties of all images sets are given in Table 5.3.

5.2 Performance Evaluation

In this section, the results of the proposed four GeoHSC for ASC on datasets given in section 5.1 are shown and these results are evaluated by using various performance measurements. Firstly, the proposed GeoHSC are tested on small and medium-size datasets given section 5.1.1 and 5.1.2 in order to see if their theories are practically supported or not. For the first step of ASC, data representatives of the mentioned datasets are produced by using neural gas and k-means ++ quantization methods and selective sampling method. This step (generating of data representatives) is repeated ten times for each dataset with the aim of elimination of randomness effect since all of three algorithm include steps that randomness is used in. After constructing similarity and Laplacian matrices for each set of data representatives, k-means clustering algorithm is run 20 times because of its randomness. As a result total 200 cluster labels are obtained by a similarity criterion for each data representatives. To compare with proposed similarity criteria traditionally Euclidean distance based similarity (s_Euc), CONN similarity and hybrid Euclidean-CONN similarity (s_{hyb}) are used. The accuracies of clustering results are calculated by the percentage of the ratio of data points which have correct cluster label to all data points. The ASC results of all datasets are shown follow sections.

5.2.1 ASC results and assessment for PR datasets

The proposed method is firstly tested on artificial PR datasets Lsun, Chainlink and Wingnut which have basic clustering challenges. Their data representatives in Fig. 5.1 are obtained by selective sampling with ratio 0.1. Then both Euclidean based ASC and proposed GeoHSC ASC are applied on these data representatives and their Laplacian matrices are obtained. The effects of similarity criterion is shown in Fig. 5.2, Fig. 5.3, Fig. 5.4. It is seen that Laplacian matrices and cluster labels obtained by Euclidean based ASC matrices could not be separated into their real clusters while

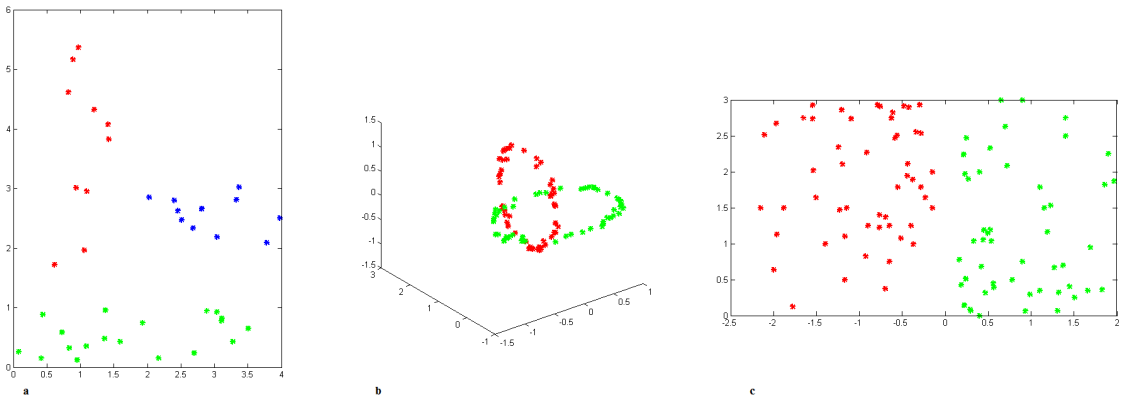

Figure 5.1: Data representatives of artificial datasets obtained by selective sampling with 0.1 sampling ratio, (a) Lsun (b) Chainlink and (c) Wingnut.

Laplacian matrices and cluster labels obtained by the proposed GeoHSC based ASC matrices achieve separation of clusters. The accuracies of Euclidean based ASC for Lsun, Chainlink and Wingnut are %75.50, %68.30 and %91.83 respectively whereas the accuracies of GeoHSC based ASC for Lsun, Chainlink and Wingnut are %100, %100 and %94.38 respectively. Therefore it is concluded that proposed geodesic based similarity criteria represent similarity within clusters and dissimilarity among clusters better than Euclidean distance based Gaussian similarity criteria.

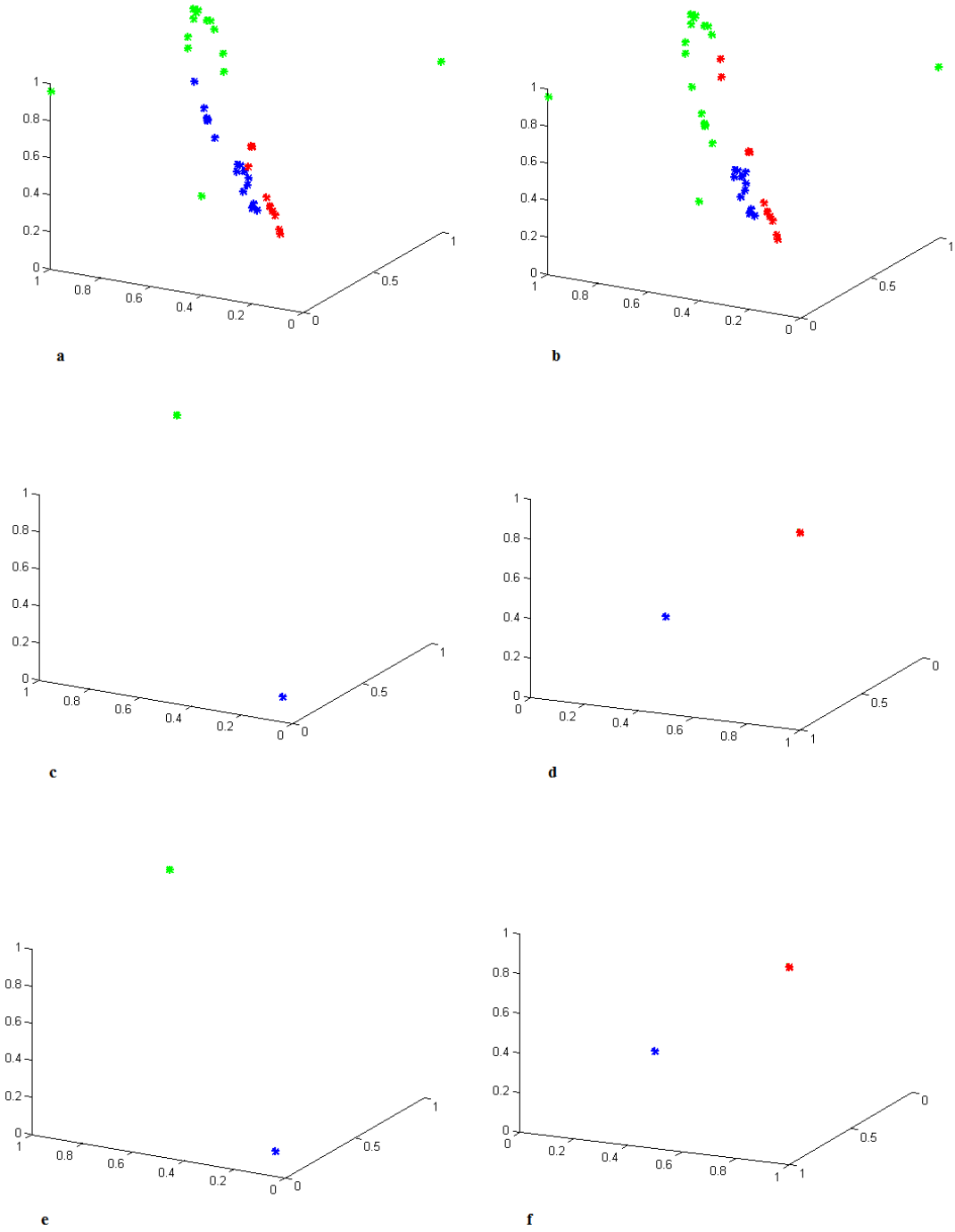

Figure 5.2: Tap: Eigen space representation of the data representatives in Fig. 5.1 (Lsun) and their cluster labels obtained by Euclidean based similarity a) ground truth, b) ASC results. Middle: Eigen space representation of the data representatives and their cluster labels obtained by proposed hybrid criteria c) ground truth, d) ASC results. Bottom: A different view of the figures in the middle e) ground truth, f) ASC results.

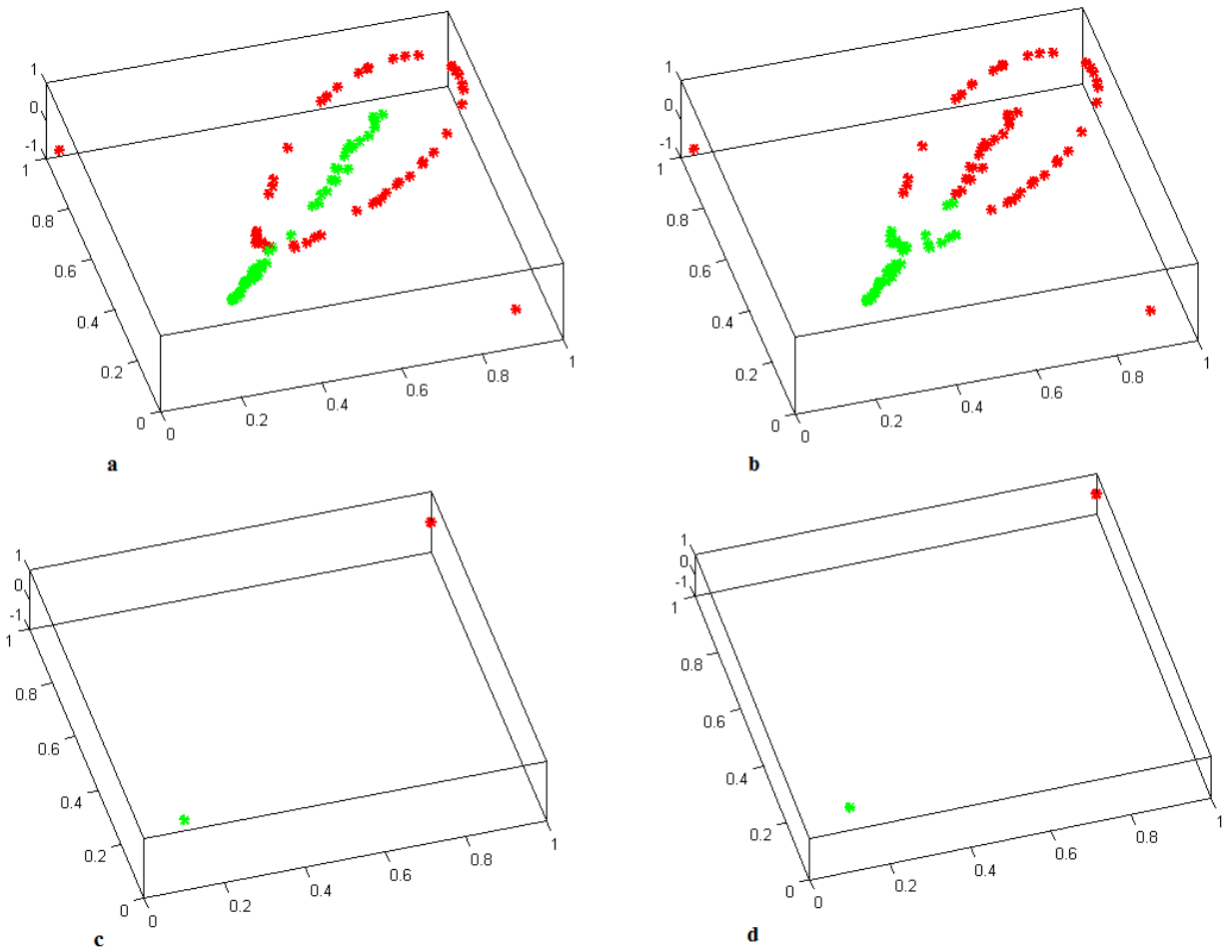

Figure 5.3: Tap: Eigen space representation of the data representatives in Fig. 5.1 (Chainlink) and their cluster labels obtained by Euclidean based similarity a) ground truth, b) ASC results. Bottom: Eigen space representation of the data representatives and their cluster labels obtained by proposed hybrid criteria c) ground truth, d) ASC results.

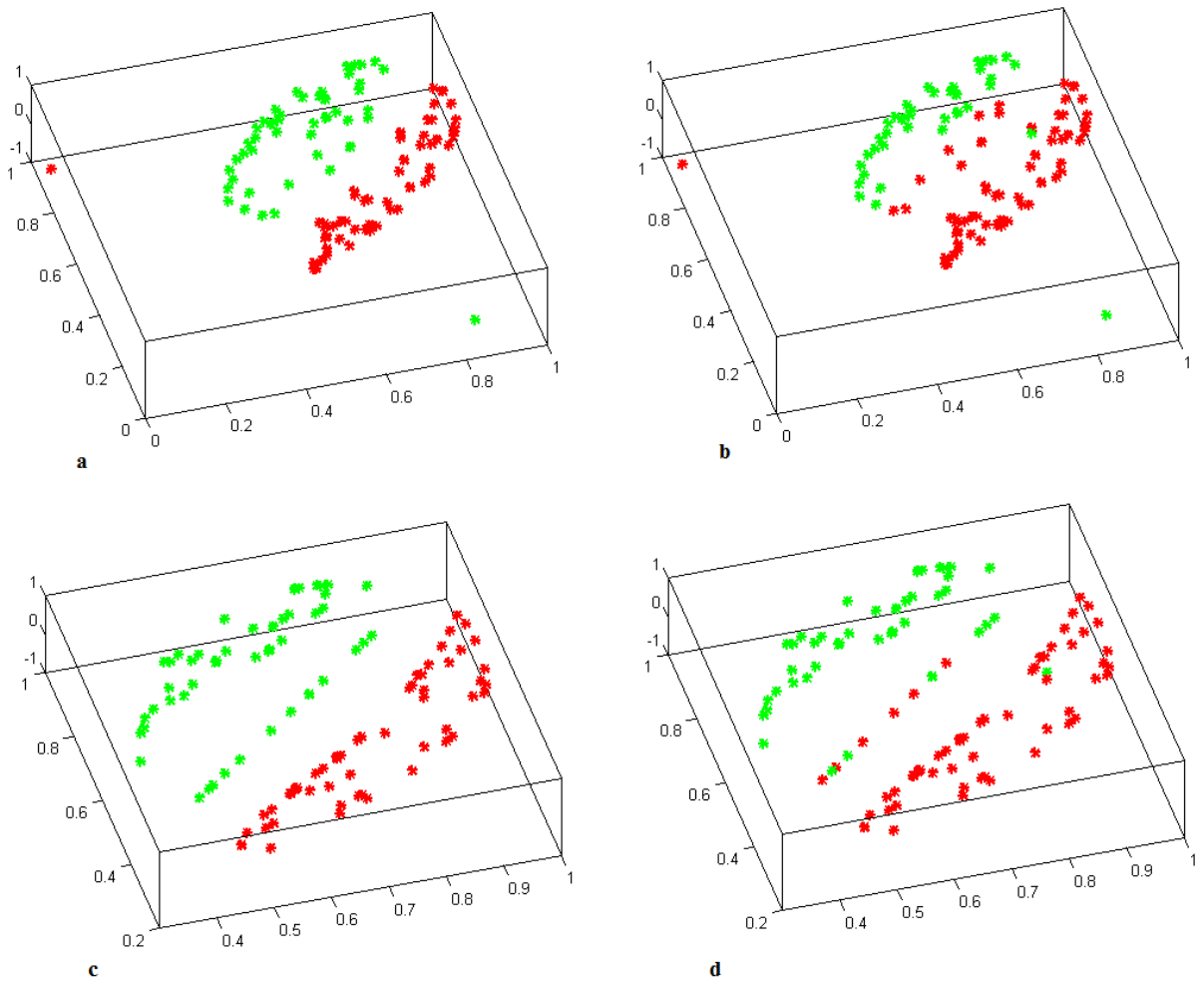

Figure 5.4: Tap: Eigen space representation of the data representatives in Fig. 5.1 (Wingnut) and their cluster labels obtained by Euclidean based similarity a) ground truth, b) ASC results. Bottom: Eigen space representation of the data representatives and their cluster labels obtained by proposed hybrid criteria c) ground truth, d) ASC results.

The ASC results of datasets in Table 5.1 are shown in Table 5.4. Bold values show SOVQ method and similarity criteria combination which has the best accuracies for each dataset as considering all while italic values show similarity criteria which have the best accuracies for other SOVQ methods. In Table 5.4, it is seen that almost all methods except s_Euc have %100 accuracy for Chainlink dataset, so this dataset is ignored for assessment. For the remain seven datasets, the results are evaluated separately in terms of the best both SOVQ and similarity criteria and summarized in below.

For only SOVQ methods:

Neural gas (NG) has the best accuracy for four datasets (Iris: %89.47, Lsun: %99.94, Breast Cancer WS: %96.62, and Yeast: %43.68).K-means++ (k++) outperforms for two artificial datasets (Lsun: %99.94 and Wingnut: %100) and finally selective sampling(SS) achieves the best accuraciy for one dataset(Pen Digits: %73.05).

For only similairty criteria:

The proposed geodesic based similarities totally have the best accuracies for five times (iris, lsun, chainlink, yeast, statlog). CONN similarity outperforms for three artificial datasets (lsun, chainlink, wingnut). Euclidean based similarity has no the best accuracy for any dataset. When the geodesic approaches are evaluated among themselves, s_{geoadj} has the best accuracies for three datasets and outperform the others.

For SOVQ and similarity criteria:

s_Euc produces the best accuracies for three each datasets with the NG (Lsun, Yeast, Chainlink), the k++ (Iris, Wingnut, Statlog) and two datasets with SS (BCWS and Pen Digits). *CONN* obtains the best accuracies for four dataset by NG, for three datasets by k++ and for one dataset by SS. For s_{hyb} which being combination of s_{Euc} and CONN (3.15), NG has the best accuracies four times, SS has the best accuracies three times and the k++ has the best accuracy once. The s_{geoknn} produces the best accuracies for six datasets by NG, two datasets by SS and one dataset by k++. The s_{geoadj}, the $s_{geoconn}$ and the s_{geohyb} have the same number of the best value. They produce the best accuracies for five times by NG, for four times by k++ and for once by SS.

As a result in total, the best SOVQ method is the neural gas which has the greatest accuracies 32 times, the second one is the k-means++ with 20 greatest accuracies

and finally the third is the selective sampling with 11 greatest accuracies. When the computational times of SOVQ methods are considered, it is seen that neural gas is the slowest method to obtain data representatives because of its iterative algorithm while selective sampling is the fastest method thanks to its sampling based algorithm.

According to the average rank of all methods, the combination of k++ and s_{geohyb} has the best result with average rank of 2.12. Although the expected for the best of SOVQ methods is NG, it is surprising that the best of s_{geohyb} because s_{geoadj} is the most frequent winner. The reason of unexpected result is that s_{geohyb} achieves accuracies very close to the winner accuracy. K-means ++ is located between neural gas and selective sampling in terms of both computational cost and accuracy performance such that it is better than NG and SS considering time and accuracy performance respectively. Therefore it may be preferred to obtain data representatives in terms of balance of accuracy and time.

Table 5.4: Clustering accuracies for the PR datasets. The ratio of the number of data representatives to the number of data points is 0.1. SS: Selective sampling; NG: neural gas; k++: k-means++ quantization

Dataset	Q/S Method	time (s)	Similarity criterion						
			s_{Euc}	CONN	s_{hyb}	s_{geoknn}	s_{geoadj}	$s_{geoconn}$	s_{geohyb}
Iris	NG	1.75	63.24 (7.1)	57.45 (9.3)	54.67 (2.3)	**89.47** (2.6)	84.47 (9.6)	86.76 (10.5)	86.69 (10.5)
150, 4D	k++	0.05	67.99 (7.6)	52.53 (5.3)	56.69 (8.9)	61.53 (9.4)	87.08 (1.8)	85.78 (3.3)	*88.63* (3.6)
3 classes	SS	0.01	63.47 (9.2)	68.96 (14.7)	69.01 (14.3)	56.39 (7.4)	80.94 (12.4)	*86.71* (7.8)	80.87 (12.4)
Lsun	NG	10.50	76.85 (1.5)	98.34 (6.8)	99.90 (0.2)	88.07(9.8)	99.54 (1.2)	**99.94** (0.2)	99.93 (0.2)
400, 2D	k++	0.07	75.82 (7.3)	**99.94** (0.2)	99.60 (1.3)	79.58 (16.5)	99.44 (1.2)	89.42 (18.7)	99.74 (0.7)
3 classes	SS	0.02	76.53 (4.7)	92.66 (11.0)	*93.89* (14.0)	79.57 (18.4)	89.21 (12.3)	85.29 (18.7)	89.15 (12.3)
BCWS	NG	10.59	95.84 (0.6)	96.51 (0.8)	**96.62** (0.6)	93.87 (1.2)	94.98 (0.5)	95.04 (0.6)	94.94 (0.4)
699, 9D	k++	0.20	92.12 (1.2)	*96.21* (0.3)	95.74 (0.8)	83.49 (16.3)	91.33 (2.0)	90.83 (3.0)	91.42 (1.9)
2 classes	SS	0.05	*96.04* (0.6)	88.81 (11.6)	92.64 (2.1)	90.00 (6.5)	65.38 (0.5)	65.38 (0.5)	65.38 (0.5)
Chainlink	NG	14.95	66.31 (0.5)	100.0 (0.0)	89.90 (13.7)	100.0 (0.0)	100.0 (0.0)	100.0(0.0)	100.0 (0.0)
1000, 3D	k++	0.21	66.17 (0.9)	100.0 (0.0)	100.0 (0.0)	99.79 (0.7)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
2 classes	SS	0.10	65.77 (0.22)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
Wingnut	NG	15.87	97.58 (0.5)	*99.88* (0.1)	99.64 (0.5)	97.55 (0.6)	98.39 (0.8)	98.41 (0.9)	98.41 (0.9)
1016, 2D	k++	0.18	97.92 (1.0)	**100.0** (0.0)	99.78 (0.3)	98.12 (2.4)	99.32 (0.7)	99.51 (0.7)	99.43 (0.6)
2 classes	SS	0.06	94.18 (1.1)	*98.82* (1.6)	98.06 (1.9)	65.33 (17.7)	91.39 (14.6)	92.81 (15.3)	91.49 (14.6)
Yeast	NG	43.81	43.04 (1.9)	42.31 (4.4)	40.22 (3.6)	34.36 (4.3)	**43.68** (3.2)	43.54 (2.9)	43.67 (2.9)
1484, 8D	k++	0.84	40.53 (1.5)	40.68 (2.8)	38.88 (3.7)	31.59 (0.6)	39.99 (1.8)	39.32 (1.9)	*42.13* (2.4)
10 classes	SS	0.12	36.60 (0.9)	*39.55* (2.8)	35.65 (3.8)	34.52 (2.1)	37.89 (3.0)	31.88 (1.7)	37.84 (2.8)
Statlog	NG	85.87	*66.90* (3.3)	57.84 (14.9)	49.31 (10.6)	65.77 (5.0)	63.40 (5.8)	54.61 (4.8)	63.71 (6.2)
6435,4D	k++	14.06	70.70 (0.7)	65.37 (1.2)	64.42 (2.2)	43.69 (12.9)	**72.14** (1.2)	69.53 (2.6)	71.76 (1.4)
6 classes	SS	1.54	*70.29* (2.1)	61.05 (17.5)	67.38 (2.6)	39.24 (9.5)	64.23 (14.3)	56.26 (9.9)	62.22 (14.5)
Pen Digit	NG	150.95	46.17 (14.9)	63.29 (10.6)	51.07 (12.9)	*68.47* (4.8)	66.86 (5.4)	53.00 (6.3)	67.69 (5.5)
10992, 16D	k++	150.55	*65.96* (9.1)	53.05 (5.9)	47.95 (10.2)	53.27 (6.4)	59.42 (6.4)	58.55 (7.2)	59.77 (6.5)
10 classes	SS	5.10	69.90 (1.4)	59.91 (15.7)	**73.05** (13.7)	58.70 (6.2)	56.49 (8.9)	58.20 (9.5)	56.48 (8.6)
	NG		3.64	3.09	4.33	3.44	2.48	3.08	*2.34*
Average rank	k++		3.88	3.00	4.12	5.62	3.12	3.88	**2.12**
	SS		3.25	*2.75*	2.25	5.12	3.50	4.25	4.00

In addition to accuracy analysis, the Adjusted Rand Index (ARI) is used to evaluate clustering performance. This index is a validation measure of agreement between labels obtained by clustering and ground truth labels for the same data [40]. A greater ARI value indicates a better clustering performance and its maximum value is 1. For multi-class problems, ARI is especially preferred because it observes both the correct separation of data points into different clusters and connection of the same cluster. In this study, the ARI values are calculated for 200 clustering results (combination of a set of data representatives and a similarity criterion) of the PR datasets and their average and standard deviations in Table 5.5. According to these results, the GeoHSC based ASC frequently produce a better clustering than those produced by the others similarities based ASC. The proposed s_{geohyb} similarity criterion has the greatest ARI values for three (Iris, Chainlink and Yeast) and produces ARI values very close to the greatest in other datasets. When the accuracy assessment with the ARI evaluation is compared, both accuracy and ARI favor the same performing metric for 21 cases among 24 cases (3 sampling/quantization methods for 8 PR datasets). A geodesic approach is the winner for four datasets (Iris, Chainlink, Yeast, Statlog) according to both ARI values and accuracies, besides both of them indicate the same proposed geodesic based similarity criteria s_{geohyb} as the best one.

Table 5.5: Adjusted Rand Index (ARI) for the PR datasets in Table 5.4. The bold values show the best ARI values for each dataset and the italic values show the best ARI values for other quantization/sampling(Q/S) methods.

Dataset	Q/S Method	Similarity criterion						
		s_{Euc}	CONN	s_{hyb}	s_{geoknn}	s_{geoadj}	$s_{geoconn}$	s_{geohyb}
Iris	NG	0.492(0.008)	0.465(0.020)	0.502(0.012)	0.453(0.005)	0.694(0.018)	0.699(0.017)	**0.747 (0.008)**
150, 4D	k++	0.446(0,009)	0.446(0,000)	0.439(0.011)	0.451(0.037)	0.693(0.006)	0.666(0.012)	*0.735(0.015)*
3 classes	SS	0.478(0.007)	0.538(0.030)	0.544(0.019)	0.416(0.000)	0.650(0.001)	*0.707(0.007)*	0.648(0.001)
Lsun	NG	0.445(0.008)	0.923(0.083)	0.978 (0.029)	0.665(0.007)	0.955(0.032)	0.924(0.007)	*0.978 (0.054)*
400, 2D	k++	0.457(0.056)	**0.999(0,000)**	0.992(0.014)	0.642(0.010)	0.987(0.004)	0.868(0.001)	0.993(0.013)
3 classes	SS	0.529(0.043)	0.853(0.015)	*0.918 (0.036)*	0.639(0.008)	0.828(0.002)	0.739(0.006)	0.809(0.012)
BCWS	NG	0.793(0.006)	0.868(0.002)	**0.869 (0.000)**	0.689(0.048)	0.807(0.018)	0.790(0.003)	0.788(0,004)
699, 9D	k++	0.705 (0.004)	*0.853 (0.000)*	0.835(0.000)	0.507(0.000)	0.678(0.003)	0.662(0.005)	0.680(0.001)
2 classes	SS	*0.846 (0.003)*	0.645(0.048)	0.749(0.158)	0.654(0.000)	0.000(0,000)	0.000(0.000)	0.000(0.000)
Chainlink	NG	0.106(0.001)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)
1000, 3D	k++	0.104(0.001)	1.000(0.000)	1.000(0.000)	0,992(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)
2 classes	SS	0.112(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)
Wingnut	NG	0.905(0.000)	*0.995 (0.000)*	0.986(0.000)	0.936(0.000)	0.937(0.003)	0.993(0.004)	0.937(0.003)
1016, 2D	k++	0.919(0.000)	**1.000 (0.000)**	0.991(0.001)	0.928(0.000)	0.973(0.000)	0.981(0.000)	0.977(0.000)
2 classes	SS	0.781(0.000)	*0.954 (0.000)*	0.904(0.136)	0.208(0.072)	0.762(0.001)	0.817(0.000)	0.766(0.000)
Yeast	NG	0.154(0.003)	0.132(0.000)	0.123(0.000)	0.110(0.005)	0.156(0.003)	0.157(0.002)	**0.158 (0.002)**
1484, 8D	k++	0.146(0.005)	0.136(0.000)	0.122(0.001)	0.023(0.004)	0.136(0.005)	0.130(0.006)	*0.154 (0.007)*
10 classes	SS	0.124(0.003)	*0.138 (0.002)*	0.125(0.003)	0.055(0.003)	0.115(0.004)	0.101(0.002)	0.118(0.005)
Statlog	NG	*0.521 (0.016)*	0.388(0.003)	0.288(0.003)	0.518(0.011)	0.342(0.020)	0.439(0.004)	0.447(0.005)
6435,4D	k++	0.526 (0.005)	0.495(0.006)	0.484(0.007)	0.222(0.008)	**0.553 (0.005)**	0.523(0.010)	0.546(0.009)
6 classes	SS	*0.526 (0.003)*	0.399(0.001)	0.503(0.009)	0.145(0.000)	0.474(0.004)	0.376(0.001)	0.446(0.003)
Pen Digit	NG	0.299(0.014)	0.474(0.010)	0.366(0.019)	0.575(0.004)	0.419(0.005)	0.566(0.009)	*0.575 (0.011)*
10992, 16D	k++	*0.527 (0.028)*	0.388(0.014)	0.361(0.012)	0.427(0.002)	0.514(0.011)	0.502(0.006)	0.525(0.008)
10 classes	SS	0.553(0.006)	0.473(0.026)	**0.600 (0.008)**	0.470(0.000)	0.454(0.007)	0.454(0.008)	0.453(0.006)
Average	NG	0.516	0.656	0.587	0.618	0.664	0.696	**0.704**
	k++	0.446	0.664	0.653	0.524	0.692	0.666	*0.701*
	SS	0.494	0.625	*0.668*	0.448	0.535	0.524	0.530

5.2.2 ASC results and assessment for small/medium medical datasets

The ASC results of datasets in Table 5.2 are shown in Table 5.6. Bold values show SOVQ method and similarity criteria combination which has the best accuracies for each dataset as considering all while italic values show similarity criteria which have the best accuracies for other SOVQ methods. The results are evaluated separately in terms of the best both SOVQ and similarity criteria. When considering only SOVQ methods, k-means++ (k++) has the best accuracy for three datasets (BCWSP: %72.80, Dermatology: %45.1, ILP: %71.15), neural gas (NG) outperforms for one datasets (Biodegration: %66.05) and finally selective sampling(SS) achieves the best accuracy for one dataset(Vertebral: %72.74). When considering only similarity criteria, the proposed geodesic based similarities totally have the best accuracies for four times (BCWSP, Vertebral, Dermatology and ILP) and CONN similarity outperforms for one datasets (Biodegration). Euclidean based similarity has no the best accuracy for any dataset. When the geodesic approaches are evaluated among themselves, s_{geohyb} has the best accuracies for three datasets and outperform the others. As a result in total, the best SOVQ method is the k-means++ and the best similarity criteria is the s_{geohyb}

Table 5.6: Clustering accuracies for the medical datasets. The ratio of the number of data representatives to the number of data points is 0.1. SS: Selective sampling; NG: neural gas; k++: k-means++ quantization

Dataset	Q/S	Similarity criterion						
	Method	s_{Euc}	CONN	s_{hyb}	s_{geoknn}	s_{geoadj}	$s_{geoconn}$	s_{geohyb}
BCWSP	NG	63.80(0.9)	57.25(2.0)	60.02(2.1)	64.18(0.6)	64.11(0.6)	62.80(0.8)	65.59(1.8)
194,33D	k++	70.07(2.2)	57.48(3.7)	60.5619(3.7)	70.43(2.2)	70.28(2.3)	67.71(3.0)	**72.80(1.7)**
2 classes	SS	59.31(1.9)	55.43(3.2)	56.61(3.0)	61.75(7.8)	57.3(2.9)	55.59(3.2)	56.80(2.7)
Vertebral	NG	55.90(5.1)	68.28(7.3)	61.41(6.6)	58.65(7.1)	59.82(6.6)	60.46(5.4)	58.68(7.2)
310,6D	k++	58.10(3.4)	68.79(7.2)	65.55(6.6)	65.11(3.8)	59.20(5.1)	62.89(5.7)	57.28(4.9)
2 classes	SS	70.95(2.8)	69.27(6.0)	71.06(2.7)	67.32(8.2)	69.65(2.9)	**72.74(1.2)**	70.84(1.5)
Dermatology	NG	27.50(0.5)	36.18(2.7)	33.27(3.7)	27.97(1.4)	41.20(3.6)	43.30(3.5)	44.52(5.8)
358, 34D	k++	27.27(1.1)	37.59(3.4)	34.29(2.7)	29.13(1.9)	45.06(3.2)	45.05(5.3)	**45.1(3.8)**
6 classes	SS	28.03(1.1)	36.13(3.6)	36.55(2.8)	29.83(3.2)	39.98(3.7)	41.57(4.0)	42.41(4.2)
ILP	NG	67.97(1.2)	68.87(5.3)	69.39(3.6)	66.91(1.8)	70.10(1.6)	68.89(4.2)	69.97(1.7)
579,10D	k++	68.25(0.4)	66.75(3.2)	66.87(3.7)	**71.15(0.5)**	69.88(0.4)	65.33(5.5)	69.84(0.5)
2 classes	SS	53.36(1.8)	61.22(2.7)	57.63(5.0)	57.09(6.2)	53.23(1.9)	61.48(2.5)	53.47(2.1)
Biodeg	NG	59.58(1.7)	**66.05(0.2)**	66.00(0.2)	62.86(1.6)	65.99(0.3)	65.99(0.3)	65.99(0.3)
1055, 41D	k++	59.10(1.5)	62.65(1.2)	62.93(1.6)	63.06(0.9)	62.84(1.2)	62.71(1.2)	62.87(1.1)
2 classes	SS	59.41(1.6)	61.95(0.3)	61.68(0.7)	60.11(4.2)	62.02(0.3)	61.81(1.3)	61.99(0.3)

5.2.3 ASC results and assessment for large medical datasets

In this part, the ASC results of four datasets in Table 5.3 are shown in Table 5.7. Bold values show similarity criteria which has the best accuracies for each dataset. The same test parameters in section 5.2.1 and 5.2.2 are valid except that data representatives are obtained by the only use of k-means++ because of being optimum SOVQ method according to results in Table 5.4 and Table 5.6 and unlike experiments for other datasets k-means clustering algorithm is also tested on large medical datasets to compare both traditional ASC and proposed GeoHSC based ASC. The ASC results of similarity criteria are evaluated and one of the proposed GeoHSC s_{geohyb} has the best accuracies for three datasets and the other of the proposed GeoHSC s_{geoknn} has the best accuracies for one dataset. Compared k-means clustering and ASC, they have close clustering success for SMHG0012 datasets with both two classes (healthy, edematous) and five classes (edematous and 4 healthy regions) but s_{geohyb} outperform k-means. It has the lowest accuracy for other two datasets. So we can say that s_{geohyb} outperform both traditional ASC methods and k-means clustering.

T1 (a), post-Gadolinium T1 (b) and T2 (c) simulated images for high-grade glioma subject are shown in Fig.5.5. SMHG0012 dataset given in Table 5.1.3 is obtained by combination of these three features. (d) is ground truth image of SMHG0012 and has six clusters. The cyan color shows brain edema while the other four colors show healthy brain regions. (e) and (f) show the result of ASC based on euclidean distance and the proposed geodesic based hybrid similarity criteria respectively. Euclidean distance based ASC assigns both edema cluster into three clusters(cyan,blue and white) and some of healthy brain regions into edema cluster. The proposed GeoHSC based ASC separates edema into two clusters (cyan and blue), one of them much bigger than the other and decide to be edema for a small healthy region (a few pixel).

In Fig.5.5, for SMHG0012 dataset, the accuracy of Euclidean distance based ASC is %83.76 while the accuracy of ASC based on proposed geodesic based similarity criteria is %90.40. The proposed similarity criteria outperforms with %6.64 difference.

Table 5.7: Clustering accuracies for the large medical datasets. The ratio of the number of data representatives to the number of data points is 0.01. k++: k-means++ quantization

Dataset	Q/S	K-means	Similarity criterion for ASC						
	Method	Clustering	s_{Euc}	CONN	s_{hyb}	s_{geoknn}	s_{geoadj}	$s_{geoconn}$	s_{geohyb}
SMHG0012 65536,3D 5 classes	k++	82.31(0.91)	83.63(2.83)	82.14(0.51)	80.52(3.19)	56.65(9.08)	85.13(3.06)	75.35(2.25)	**87.45(2.26)**
SMHG0015 65536,3D 6 classes	k++	50.82(0.13)	71.35(4.44)	68.13(8.17)	68.27(8.14)	55.14(5.69)	72.35(4.34)	66.41(4.10)	**75.84(4.59)**
SMHG0019 65536,3D 6 classes	k++	51.10(0.19)	70.42(2.13)	68.72(10.25)	65.47(8.96)	51.78(8.60)	72.79(4.30)	68.17(4.82)	**75.20(3.74)**
SMHG0012 65536,3D 2 classes	k++	82.14(0.45)	89.50(3.26)	**97.93(0.14)**	94.33(3.90)	87.89(14.2)	93.72(3.16)	92.04(1.52)	93.97(2.80)

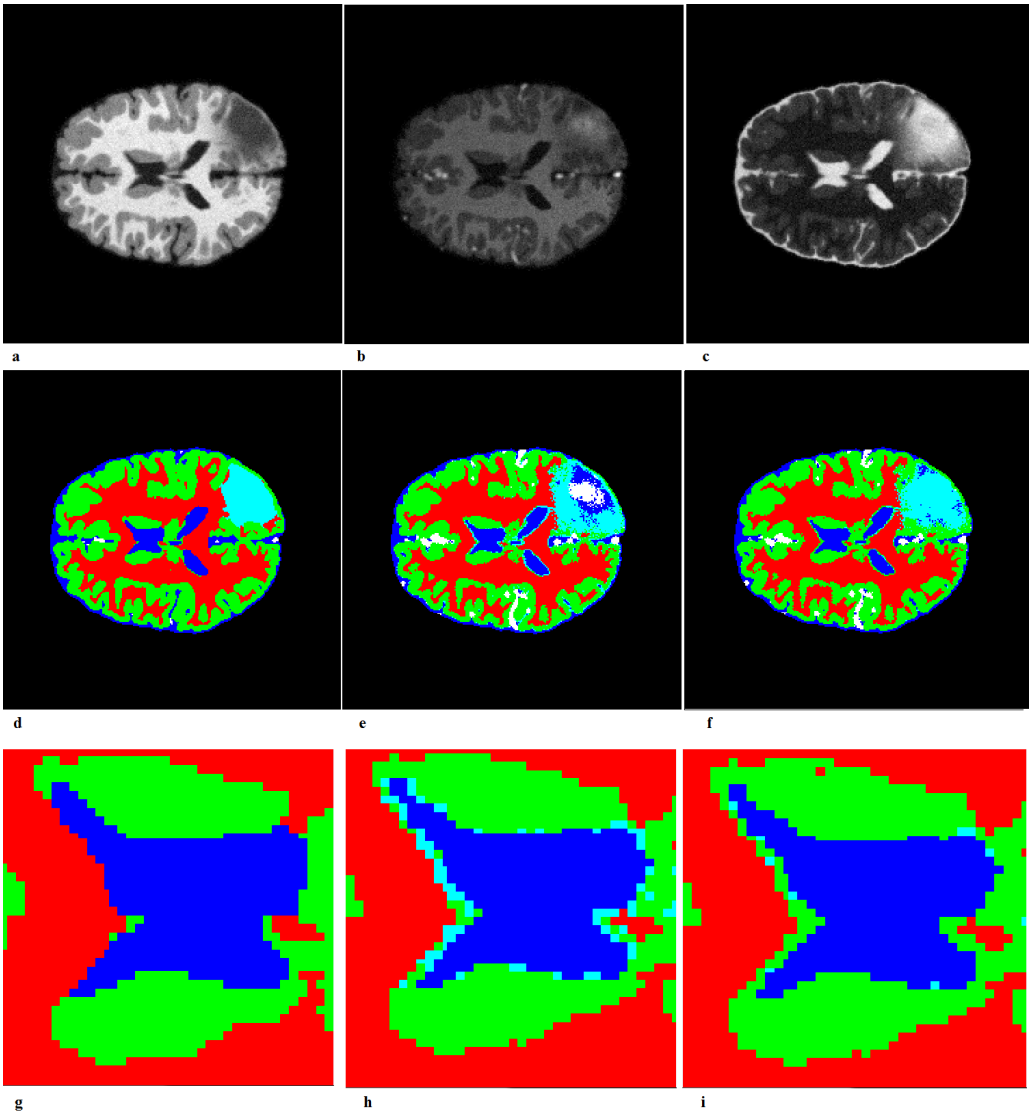

Figure 5.5: Simulated brain MR images and their clustering results. (a) T1, (b) post-Gadolinium T1, (c) T2 and (d) ground truth show three features and ground truth for SMHG0012 dataset, (e) Euclidean distance based Gaussian kernel based ASC (%83.76), (f) the proposed GeoHSC based ASC (%90.40), (g) ground truth zoom, (h) Euclidean zoom and (i) GeoHSC zoom

6. CONCLUSIONS

In this thesis we have presented novel Geodesic Based Hybrid Similarity Criteria, similarity measures which allow the use of information of data topology, local density distribution and distance for approximate spectral clustering (ASC).

ASC is a clustering procedure which enables applying of spectral clustering methods on large datasets. It uses data representatives generated from dataset by a SOVQ method instead of all data points in the dataset. Then cluster labels are obtained by using eigenvalue decomposition of similarity matrix of data representatives. Constructing similarity matrix is one of the most important steps for ASC since it represents similarity within cluster and between clusters. The euclidean distance based Gaussian kernel is a common approach to construct the this matrix. This euclidean distance based method may be insufficient to show relation between data points and cause poorly clustering results because of euclidean distance. ASC enables to reveal new information such as data topology and local density distribution thanks to the use of data representatives. The geodesic based distance takes advantages of these information while the euclidean distance based ignores them.

We proposed geodesic based hybrid similarity measures for ASC with the use of geodesic based distances for distance based Gaussian kernel. Our proposed similarity measures both utilize data topology and combine variety of information available in ASC. Thus it correctly represents similarity matrix for ASC. We used a wide range of datasets which have different size (small, medium and large) and cluster challenges to test our proposed similarity measures. Both sampling (SS) and quantization methods (NG, k++) were used to obtain data representatives for datasets which have small/medium-size while only quantization method (k++) was used for large medical datasets. We concluded that quantization methods have better clustering performance than sampling method because of the construction of these similarities based on data topology and local density distribution while they have slower procedure than sampling method because of being their iterative algorithms.

We used there similarity measures, the euclidean distance based Gaussian kernel (s_{Euc}), CONN similarity (s_{CONN}) and a CONN based hybrid similarity criterion (s_{hyb}) to compare our proposed geodesic based hybrid similarity measures. We had result that our proposed geodesic based similarity criteria outperform the others for both small/medium and large size datasets. It is a difficult problem to determine which similarity criterion will produce the most satisfactory clustering result before clustering, but the proposed geodesic hybrid similarity criterion s_{geohyb} may be preferred when a graph-based measure is used because it the most frequent has the best accuracy even though it is not the best for all datasets. Furthermore s_{geohyb} can be a successful tool for brain segmentation thanks to its high accuracies in experiments on large datasets using medical images simulated MR.

6.1 Future Work

The results of this thesis show that ASC methods have some problems on SOVQ methods and similarity matrices and we try to develop new algorithms which can solve these problems.

Firstly, ASC approaches need an optimum SOVQ algorithm which both run as fast as sampling algorithms and have performance as high as quantization algorithms. Therefore a new approach that combines advantages of both may be a effective solution for ASC.

The other problem is that a wide variety of SOVQ algorithms and similarity measures exist in literature. This makes both selection of SOVQ algorithm /similarity measure and decision of two algorithm used with together difficult. When the results of this thesis are analyzed, it is seen that at each change of the combination of SOVQ method and similarity measure used for ASC different clustering performance are obtained for same dataset. Any combination which has good clustering performance for a dataset may have poorly clustering performance for another dataset. To address this problem, a coefficient may be proposed such that it shows which combination will have the best performance before clustering.

Finally, we have so many cluster labels for each dataset because we use three SOVQ methods and seven similarity measures to cluster each of them. Moreover we run

SOVQ step ten times and clustering step twenty times to get rid of randomness of these algorithms. We use only the best one of all results as cluster label for each dataset and ignore the others. We believe that if a new approach which combines all ASC results by using ensemble methods is proposed, more satisfactory clustering performances may be achieved.

REFERENCES

[1] **Ultsch, A.** (2003). Maps for the visualization of high-dimensional data spaces, *Proc. Workshop on Self organizing Maps*, pp.225–230.

[2] **Jain, A.K.** (2010). Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, ***31***(8), 651–666.

[3] **Remm, M.**, **Storm, C.E. and Sonnhammer, E.L.** (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, *Journal of molecular biology*, ***314***(5), 1041–1052.

[4] **Pham, D.L.**, **Xu, C. and Prince, J.L.** (2000). Current methods in medical image segmentation 1, *Annual review of biomedical engineering*, ***2***(1), 315–337.

[5] **Shi, J. and Malik, J.** (2000). Normalized cuts and image segmentation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, ***22***(8), 888–905.

[6] **Ng, A.Y.**, **Jordan, M.I.**, **Weiss, Y.** *et al.* (2002). On spectral clustering: Analysis and an algorithm, *Advances in neural information processing systems*, ***2***, 849–856.

[7] **Meila, M. and Shi, J.** (2001). A random walks view of spectral segmentation.

[8] **Von Luxburg, U.** (2007). A tutorial on spectral clustering, *Statistics and computing*, ***17***(4), 395–416.

[9] **Filippone, M.**, **Camastra, F.**, **Masulli, F. and Rovetta, S.** (2008). A survey of kernel and spectral methods for clustering, *Pattern recognition*, ***41***(1), 176–190.

[10] **Fowlkes, C.**, **Belongie, S.**, **Chung, F. and Malik, J.** (2004). Spectral grouping using the Nystrom method, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, ***26***(2), 214–225.

[11] **Bezdek, J.C.**, **Hathaway, R.J.**, **Huband, J.M.**, **Leckie, C. and Kotagiri, R.** (2006). Approximate clustering in very large relational data, *International journal of intelligent systems*, ***21***(8), 817–841.

[12] **Wang, L.**, **Bezdek, J.C.**, **Leckie, C. and Kotagiri, R.** (2008). Selective sampling for approximate clustering of very large data sets, *International Journal of Intelligent Systems*, ***23***(3), 313–331.

[13] **Wang, L.**, **Leckie, C.**, **Ramamohanarao, K. and Bezdek, J.**, (2009). Approximate spectral clustering, Advances in Knowledge Discovery and Data Mining, Springer, pp.134–146.

[14] **Chen, W.Y.**, **Song, Y.**, **Bai, H.**, **Lin, C.J. and Chang, E.Y.** (2011). Parallel spectral clustering in distributed systems, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, ***33***(3), 568–586.

[15] **Taşdemir, K.**, **Yalçin, B. and Yildirim, I.** (2014). Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures, *Pattern Recognition*.

[16] **Merényi, E.**, **Tasdemir, K. and Zhang, L.**, (2009). Learning highly structured manifolds: harnessing the power of SOMs, Similarity-Based Clustering, Springer, pp.138–168.

[17] **Asuncion, A. and Newman, D.** UCI machine learning repository, 2007. URL h ttp, *archive. ics. uci. edu/ml/datasets. html S*.

[18] **Jain, A.K.**, **Murty, M.N. and Flynn, P.J.** (1999). Data clustering: a review, *ACM computing surveys (CSUR)*, ***31***(3), 264–323.

[19] **Yan, M.** (2005). Methods of determining the number of clusters in a data set and a new clustering criterion, *Ph.D. thesis*, Virginia Polytechnic Institute and State University.

[20] **Sneath, P.H.**, **Sokal, R.R.** *et al.* (1973). *Numerical taxonomy. The principles and practice of numerical classification.*

[21] **King, B.** (1967). Step-wise clustering procedures, *Journal of the American Statistical Association*, ***62***(317), 86–101.

[22] **Forgy, E.W.** (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications, *Biometrics*, ***21***, 768–769.

[23] **Jain, A.K. and Dubes, R.C.** (1988). *Algorithms for clustering data*, Prentice-Hall, Inc.

[24] **Arthur, D. and Vassilvitskii, S.** (2007). k-means++: The advantages of careful seeding, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics, pp.1027–1035.

[25] **Verma, D. and Meila, M.** (2003). A comparison of spectral clustering algorithms.

[26] **Taşdemir, K.** (2012). Vector quantization based approximate spectral clustering of large datasets, *Pattern Recognition*, ***45***(8), 3034–3044.

[27] **Yan, D.**, **Huang, L. and Jordan, M.I.** (2009). Fast approximate spectral clustering, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp.907–916.

[28] **Belabbas, M.A. and Wolfe, P.J.** (2009). Spectral methods in machine learning and new strategies for very large datasets, *Proceedings of the National Academy of Sciences*, ***106***(2), 369–374.

[29] **Martinetz, T.M.**, **Berkovich, S.G. and Schulten, K.J.** (1993). Neural-gas' network for vector quantization and its application to time-series prediction, *Neural Networks, IEEE Transactions on*, ***4***(4), 558–569.

[30] **Yalcin, B. and Tasdemir, K.** (2014). The use of k-means++ for approximate spectral clustering of large datasets, *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, IEEE, pp.220–223.

[31] **Martinetz, T. and Schulten, K.** (1994). Topology representing networks, *Neural Networks*, ***7***(3), 507–522.

[32] **Provost, F.**, **Jensen, D. and Oates, T.** (1999). Efficient progressive sampling, *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp.23–32.

[33] **Hathaway, R.J.**, **Bezdek, J.C. and Huband, J.M.** (2006). Scalable visual assessment of cluster tendency for large data sets, *Pattern Recognition*, ***39***(7), 1315–1324.

[34] **Zhang, X.**, **Li, J. and Yu, H.** (2011). Local density adaptive similarity measurement for spectral clustering, *Pattern Recognition Letters*, ***32***(2), 352–358.

[35] **Tasdemir, K. and Merényi, E.** (2009). Exploiting data topology in visualization and clustering of self-organizing maps, *Neural Networks, IEEE Transactions on*, ***20***(4), 549–562.

[36] **Tasdemir, K.** (2013). A hybrid similarity measure for approximate spectral clustering of remote sensing images., *IGARSS*, pp.3136–3139.

[37] **Asuncion, A. and Newman, D.**, (2007), UCI machine learning repository.

[38] **Menze, B.**, **Jakab, A.**, **Bauer, S.**, **Kalpathy-Cramer, J.**, **Farahani, K.**, **Kirby, J.**, **Burren, Y.**, **Porz, N.**, **Slotboom, J.**, **Wiest, R.** *et al.* (2014). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS).

[39] **Prastawa, M.**, **Bullitt, E. and Gerig, G.** (2009). Simulation of brain tumors in MR images for evaluation of segmentation efficacy, *Medical image analysis*, ***13***(2), 297–311.

[40] **Santos, J.M. and Embrechts, M.**, (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification, Artificial neural networks–ICANN 2009, Springer, pp.175–184.

CURRICULUM VITAE

Name Surname: Berna YALÇIN

Place and Date of Birth: Tokat 21.09.1988

Adress:Kanuni Sultan Süleyman Cad. Çimen Sk. Sarıyer, İstanbul

E-Mail:h.k.berna@gmail.com

B.Sc.: Electronic Engineering,Ankara University, 2011

PUBLICATIONS/PRESENTATIONS ON THE THESIS

1. Tasdemir, K., **Yalcin, B.**, & Yildirim, I. (2015). Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures. Pattern Recognition, 48(4), 1461-1473.

2. **Yalcin, B.**, & Tasdemir, K. (2014, April). The use of k-means++ for approximate spectral clustering of large datasets. In Signal Processing and Communications Applications Conference (SIU), 2014 22nd (pp. 220-223). IEEE.