Medikal Verilerin Jeodezik Tabanlı Yaklaşık Spektral Öbekleme Yöntemiyle Analizi
Analysis of Medical Datasets by Using Geodesic Based Approximate Spectral Clustering

Berna YALÇIN[1], Kadim TAŞDEMİR[2], İsa YILDIRIM[1]
[1] Elektronik ve Haberleşme Mühendisliği Bölümü
İstanbul Teknik Üniversitesi
yalcinbern@itu.edu.tr, isayildirim@itu.edu.tr
[2] Bilgisayar Mühendisliği Bölümü
Uluslararası Antalya Üniversitesi
kadim.tasdemir@antalya.edu.tr

Özetçe **—Tıp alanında, medikal veri analizinin hızlı ve doğru bir şekilde yapılması teşhis ve tedavi açısından oldukça önemlidir. Teknolojinin gelişmesiyle birlikte çeşitliliği ve boyutları artan verisetlerini değerlendirmek zorlaşmakta ve ulaşmak istenen sonuçla ilgili tavsiye ya da karar verebilen yöntemlere ihtiyaç duyulmaktadır. Öbekleme yönteminin, eğitim verisine ihtiyaç duymayan bir öğreticisiz sınıflandırma yöntemi olmasının yanı sıra karar verme, kestirim gibi noktalarda etkili ve hızlı bir süreç olmasından dolayı medikal alandaki uygulamaları hızla artmaktadır. Yaklaşık Spektral Öbekleme (YSÖ) yöntemi parametrik bir model kullanmaması sayesinde hem düzensiz şekilli öbekleri bulabilmekte hem de parametrik modellere uymayan gerçek öbekleri bulmada daha başarılı olmaktadır. Verilerin ikili benzerliklerini gösteren benzerlik matrisini kullanan YSÖ için çeşitli benzerlik ölçütleri önerilmiştir. Bu çalışmada jeodezik uzaklık tabanlı benzerlik ölçütleri kullanan yaklaşık spektral öbekleme ile medikal verisetlerinin analizi gerçekleştirilmiştir. Yöntemin sonuçları geleneksel k-ortalama ve Euclid tabanlı yaklaşık spektral öbekleme sonuçlarıyla karşılaştırılmıştır.**

Anahtar Kelimeler—yaklaşık spektral öbekleme, jeodezik benzerlik, medikal veri analizi, MRI.

Abstract**—Fast and accurate analysis of medical data is of great importance for diagnosis and treatment. In line with the technological developments, the size and diversity of these data have been increasing, which in turn makes their assessment difficult. Therefore, there is an ever growing need for automated decision support systems. To this end, clustering based applications are rapidly increasing in medicine thanks to their limited user interaction, no requirement of labeled training samples, and their effectiveness for decision and estimation. Among clustering methods, approximate spectral clustering (ASC), which depends on a pairwise similarity matrix of data points, has been recently popular and successful thanks to its ability to find irregularly shaped clusters and its independence from parametric cluster models. Various similarity criteria have been proposed to represent pairwise similarities. In this study, medical datasets have been analyzed by approximate spectral clustering with our geodesic distance based similarity criteria. Experimental results indicate that our approach performs better than traditional k-means and Euclidean based approximate spectral clustering.**

978-1-4673-7765-2/15/ $31.00 ©2015 IEEE

Keywords—Approximate spectral clustering, geodesic similarity criteria, medical data analysis, MRI.

I. GİRİŞ

Hastalığın doğru teşhis edilmesi ve uygun tedavi yöntemine karar verilmesi açısından toplanan verilerin doğru analiz edilmesi oldukça önemlidir. Öbekleme algoritmaları tanı koyma, anatomik yapı değerlendirme, manyetik rezonans (MR), bilgisayarlı tomografi (BT) görüntülerinin bölütlenmesi gibi amaçlarla biyomedikal uygulamalarda son yıllarda sıklıkla kullanılmaktadır [1]. Öbekleme yöntemleri arasında spektral öbeklemenin (SÖ) parametrik bir model kullanmaması sayesinde, genelde parametrik modellere uymayan gerçek verilerin öbeklenmesinde daha başarılı olduğu gösterilmiştir [2], [3]. Verilerin ikili benzerliklerinin özdeğer ayrışımını kullanan spektral öbekleme, özdeğer ayrışımının gerektirdiği yüksek hesaplama yükü ve hafıza gereksinimi sebebiyle büyük veri setlerine uygulanamaz. Büyük veri setlerindeki bu sorunun çözümü için yaklaşık spektral öbekleme (YSÖ) yöntemleri kullanılır [4], [5]. YSÖ, iki adımlı bir algoritmadır. İlk adımda örnekleme veya nicemleme yöntemleriyle veri setinden veri temsilcileri elde edilir, ikinci adımda ise veri temsilcileri spektral öbekleme yöntemi ile öbeklere ayrılırlar ve tüm veri setine veri temsilcileriyle ilişkili olarak öbek etiketleri atanır [4], [5]. 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 kullanılan Euclid uzaklığının yanısıra YSÖ' nün ortaya çıkardığı veri topolojisi, yerel yoğunluk dağılımı ve veri manifoldu gibi veri setine ait bilgilerin kullanımını sağlayan benzerlik ölçütleri de önerilmiştir [5]–[7].

Bu çalışmada jeodezik tabanlı yaklaşık spektral öbekleme algoritması kullanılarak farklı özelliklere sahip medikal

verilerin analizlerinin yapılması amaçlanmıştır. Çalışma kapsamında jeodezik tabanlı yaklaşık spektral öbekleme sonuçları, Euclid tabanlı yaklaşık spektral öbekleme yöntemi ve geleneksel k-ortalama yöntemi ile karşılaştırılacaktır.

Öncelikle Bölüm II' de YSÖ kısaca anlatılacak ve Bölüm III' te YSÖ' de kullanılan benzerlik ölçütlerinden bahsedilecektir. Bölüm IV' te ise farklı veri setleri üzerindeki başarılar gösterilecek ve sonuçlar değerlendirilecektir. Son olarak Bölüm V' te çalışmanın vargıları sunulacaktır.

II. Yaklaşık Spektral Öbekleme

Spektral öbekleme (SÖ), veri noktaları arasındaki ikili benzerliklerin özdeğer ayrışması tabanlı bir öğrenme yöntemidir. Özdeğer 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, 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 edilir. İkinci aşamada ise veri temsilcilerinin ikili benzerliklerini gösteren benzerlik matrisi spektral öbekleme ile öbeklere ayrılır. YSÖ' nün basamakları şu şekildedir:

- N örnekli bir veri kümesi için örnekleme/nicemleme yöntemi ile n veri temsilcisi elde edilir.
- Veri temsilcileri için S benzerlik matrisi, L Laplace matrisi hesaplanır ve öz değer ayrışımı gerçekleşir.
- Öz değer matrisi üzerinden k-ortalama ile k öbek elde edilir.
- Her veri temsilcisinin öbek etiketi temsil ettiği veri örneklerine atanır.

Bu çalışmada veri temsilcilerinin elde edilmesinde, hem performans hem de hesaplama yükü açısından YSÖ için uygun olduğu gösterilen k-ortalama ++ nicemleme yöntemi kullanılmıştır [8].

III. Benzerlik Ölçütleri

Yaklaşık spektral öbeklemede veri noktaları 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. Bu aşamada genellikle (1)' de gösterilen Euclid uzaklık tabanlı Gauss işlevi kullanılır.

$$s_{Euc}(w_i, w_j) = exp\left(-\frac{d_{Euc}(w_i, w_j)}{2\sigma^2}\right) \quad (1)$$

Burada $d_{Euc}(w_i, w_j)$ ise iki veri temsilcisi w_i, w_j arasındaki Euclid uzaklığını, σ ise Gauss parametresini gösterir. Bu benzerlik ölçütü iyi ayrılmış öbeklere sahip veri setlerinde yüksek performans gösterirken, homojen olmayan öbek içi veri dağılımına veya iç içe geçmiş öbeklere sahip veri kümelerinde düşük performansa sahip olabilmektedir [6]. Yerel veri karakteris-tiklerini gösteren ağırlıklandırılmış Delaunay üçgeni tabanlı CONN benzerliği alternatif benzerlik ölçütü olarak sunulmuştur [5].

$$CONN(i,j) = |\{v \in (V_{ij} \cup V_{ji})\}| \quad (2)$$

$$V_{ij} = \{v \in V_i : \|v - w_j\| \leq \|v - w_k\| \forall k \neq i\} \quad (3)$$

$$V_i = \{v \in M : \|v - w_i\| \leq \|v - w_j\|\} \quad (4)$$

Bu yöntemde benzerlik hesabı için her v veri örneğine en yakın ve ikinci en yakın veri temsilcileri bulunur. Veri temsilcilerinin ortak alt manifoldlarında bulunan veri örneği sayısı veri temsilcileri arasındaki benzerliği gösterir.

Alternatif olarak önerilen jeodezik uzaklık tabanlı benzerlik ölçütleri ise herhangi iki veri temsilcisi arasındaki uzaklık için bir komşuluk çizgesi üzerinden en kısa yolu hesaplar. Komşuluk çizgesinin veri manifoldunu gösterecek şekilde oluşturulması önemlidir. CONN ile komşuluk çizgesi oluşturulduğunda bu çizge yerel veri karakteristiklerini gösterir ve bu durumda her temsilci için özel bir komşu sayısı belirler. Bu şekilde komşuluklar üzerinden daha etkin bir jeodezik uzaklık hesaplar. w_i ve w_j temsilcileri $CONN(i,j) > 0$ ise komşudurlar. CONN üzerinden jeodezik uzaklık (5)' deki gibi hesaplanır.

$$d_{geoadj}(w_i, w_j) = \sum_{l,m \in SP_{adj}(w_i, w_j)} d_{Euc}(l, m) \quad (5)$$

Burada $SP_{adj}(w_i, w_j)$, CONN' a göre i ve j arasındaki en kısa yolda bulunan kenarların kümesidir. İki temsilci arasındaki uzaklık CONN' a göre hesaplanabilir ve jeodezik uzaklık (7) ile elde edilir.

$$d_{CONN}(w_i, w_j) = exp\left(-\frac{CONN(i,j)}{\max_{y,z} CONN(y,z)}\right) \quad (6)$$

$$d_{geoconn}(w_i, w_j) = \sum_{l,m \in SP_{CONN}(w_i, w_j)} d_{CONN}(l, m) \quad (7)$$

Eşitlik (8)' deki jeodezik uzaklık ölçütü ise yaklaşık spektral öbeklemede veri temsilcileriyle ortaya çıkan tüm bilgileri Euclid ve CONN birleşimi ile kullanır.

$$d_{geohyb}(w_i, w_j) = \sum_{l,m \in SP_{adj}(w_i, w_j)} d_{Euc}(l, m) d_{CONN}(l, m) \quad (8)$$

Tablo I: k-ortalama öbekleme yöntemi ve farklı benzerlik ölçütleri için YSÖ yöntemi ile elde edilen doğru öbekleme yüzdeleri (%). Her veri kümesi için örnek sayıları, öznitelik boyutları (B) ve sınıf sayıları veri kümesinin altında belirtilmiştir. Ayrıca her bir veri kümesi için en yüksek başarı kalın fontla ifade edilmiştir

Veri Kümesi	k-ortalama	YSÖ için Benzerlik Ölçütleri				
		s_{Euc}	CONN	s_{geoadj}	$s_{geoconn}$	s_{geohyb}
BCWSP 194,33B 2 sınıf	63.40 (0.001)	70.07 (2.2)	57.48 (3.7)	70.28 (2.3)	67.71 (3.0)	**72.80 (1.7)**
Vertebral 310,6B 2 sınıf	67.25 (0.0)	58.10 (3.4)	**68.79 (7.2)**	59.20 (5.1)	62.89 (5.7)	57.28 (4.9)
Dermatology 358, 34B 6 sınıf	29.51 (0.0)	27.27 (1.1)	37.59 (3.4)	45.06 (3.2)	45.05 (5.3)	**45.1 (3.8)**
ILP 579,10B 2 sınıf	**71.35 (0.0)**	68.25 (0.4)	66.75 (3.2)	69.83 (0.4)	65.33 (5.5)	69.84 (0.5)
Biodegration 1055, 41B 2 sınıf	58.86 (0.0)	59.10 (1.5)	62.65 (1.2)	62.84 (1.2)	62.71 (1.2)	**62.87 (1.1)**
SMHG0012 65536,3B 5 sınıf	82.31 (0.91)	83.63 (2.83)	82.14 (0.51)	85.13 (3.06)	75.35 (2.25)	**87.45 (2.26)**
SMHG0015 65536,3B 6 sınıf	50.82 (0.13)	71.35 (4.44)	68.13 (8.17)	72.35 (4.34)	66.41 (4.10)	**75.84 (4.59)**
SMHG0019 65536,3B 6 sınıf	51.10 (0.19)	70.42 (2.13)	68.72 (10.25)	72.79 (4.30)	68.17 (4.82)	**75.20 (3.74)**

IV. Deney ve Sonuçlar

Jeodezik tabanlı yaklaşık spektral öbekleme algoritmasının medikal veriler üzerindeki performansını ölçmek için farklı özelliklere sahip yapay ve gerçek veri kümeleri kullanılmıştır. Yapay veri kümeleri The Multimodal Brain Tumor Image Segmentation Benchmark (BRAST) veri tabanından alınan farklı özelliklere sahip tümörlü ve ödemli bölge içeren MR görüntülerinden oluşmaktadır [9]. Öbekleme için T1, post-Gadolinium T1, ve T2 ağırlıklı görüntüler öznitelik olarak kullanılmıştır. Çeşitli hastalıklarla ilgili olarak yapılan tetkikleri içeren gerçek veri kümeleri ise (BCWSP, Vertebral, Dermatology, ILP, Biodegration) UCI Machine Learning Repository [10]' den alınmış olup detaylı bilgi için [10]' a bakılabilir.

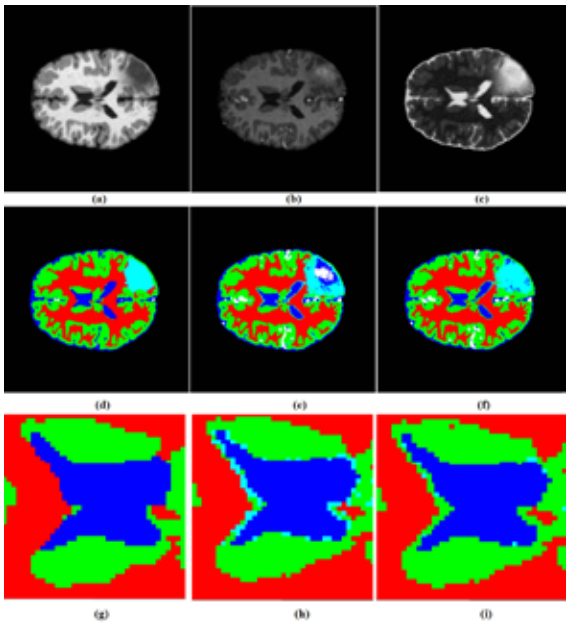
Tablo I' de veri kümelerinin örnek sayısı, öznitelik boyutu ve sınıf sayısı gösterilmiştir. YSÖ' nün ilk aşamasında veri temsilci sayısı gerçek veri kümeleri için veri örnek sayısının onda biri olacak şekilde; yapay veri kümeleri içinse veri örnek sayısının binde biri olacak şekilde alınmış ve her veri kümesi için nicemleme işlemi 10 kez tekrarlanmış, her veri temsilci seti YSÖ ikinci aşamasında 5 farklı benzerlik ölçütünün her birisi için 20 kez öbeklenmiş ve oluşan 200 öbeklemenin ortalama başarı değerleri hesaplanmıştır. k-ortalama içinse öbekleme işlemi 20 kez tekrarlanmış ve 20 öbeklemenin ortalama başarısı hesaplanmıştır. Sonuçlar Tablo I' de gösterilmiştir.

Medikal verilerde hangi öbekleme yaklaşımının kullanılması gerektiğinin analizi için başarı sonuçları değerlendirildiğinde, geleneksel k-ortalama öbekleme yönteminin sekiz veri kümesinin yalnızca birisinde (ILP) en yüksek başarıya ulaştığı görülmektedir. Yaklaşık spektral öbekleme yönteminde ise geleneksel Euclid benzerlik ölçütü hiçbir veri kümesi için en başarılı sonucu elde edemezken, CONN benzerlik ölçütü bir veri seti (Vertebral) için en iyi sonucu vermiştir. Jeodezik tabanlı benzerlik ölçütleri incelendiğinde ise sekiz veri kümesinin altısında s_{geohyb} ölçütünün en yüksek başarıya sahip olduğu görülmüştür.Bunun yanı sıra Euclid benzerlik ölçütü genel olarak jeodezik tabanlı benzerlik ölçütlerinden daha düşük başarı göstermiştir.

Şekil 1' de ödemli bölge içeren 5 öbekli SMHG0012 veri kümesi için öz nitelik olarak kullanılan T1 (a), post-Gadolinium T1 (b) ve T2 (c) görüntüleri, doğru sınıf etiketleri (d) bilgisi, geleneksel Euclid tabanlı YSÖ sonucu (e) ve jeodezik tabanlı (s_{geohyb}) YSÖ sonucu (f) gösterilmiştir. Turkuaz ödemli bölgeyi gösterirken diğer renkler beynin sağlıklı bölgelerini göstermektedir. Buna göre Euclid tabanlı YSÖ yaklaşımı ödemli bölgeyi üç öbeğe (turkuaz, mavi, beyaz) ayırırken jeodezik tabanlı YSÖ yaklaşımı ise ödemli bölgeyi iki öbeğe (turkuaz ve mavi) ayırmıştır. Bunun yanısıra gerçekte ödemli olan bölgede sağlıklı (mavi ve beyaz) olarak işaretlenen bölüm Euclid tabanlı YSÖ' de jeodezik tabanlı YSÖ 'ye göre oldukça büyüktür. Şekil 1' de (g), (h) ve (i)' de (d), (e) ve (f) görüntülerinin yakınlaştırılmış hali verilmiştir. Bu görüntüler incelendiğinde gerçekte sağlıklı olan bölgede ödem olarak işaretlenen piksel sayısının Euclid tabanlı YSÖ ile yapılan öbeklemede jeodezik tabanlı YSÖ ile yapılan öbeklemeye göre daha fazla olduğu görülmektedir. SMHG0012 veri kümesi için Şekil 1' de gösterilen Euclid tabanlı YSÖ başarısı % 83.76 iken jeodezik tabanlı YSÖ başarısı % 90.40 'tır. Jeodezik tabanlı YSÖ başarıyı % 6.4 artırmıştır.

V. Vargı

Bu çalışmada, tıbbi alanlardaki teşhis ve tedavi süreçlerine destek olmak için jeodezik tabanlı yaklaşık spektral öbekleme (YSÖ) algoritmasının medikal veriler üzerindeki etkinliği değerlendirilmiştir. İstatistiksel ve tıbbi görüntüleme yöntemleriyle elde edilen gerçek ve yapay veri kümeleri üzerindeki denemelerde, jeodezik tabanlı yaklaşık spektral öbekleme yönteminin genel olarak geleneksel k-ortalama öbekleme ve Euclid tabanlı yaklaşık spektral öbekleme yöntemlerinden daha başarılı olduğu görülmüştür. Jeodezik tabanlı benzerlik ölçütleri kendi aralarında değerlendirildiğinde ise s_{geohyb} en başarılı ölçüt olmuştur.



Şekil 1: Birinci satır: 2B MR beyin görüntüleri (a) T1, (b) post-Gadolinium T1, (c) T2. İkinci satır: (d) doğru sınıf etiketleri, (e) Euclid tabanlı YSÖ için öbekleme sonucu, (f) Jeodezik tabanlı YSÖ için öbekleme sonucu. Üçüncü satır: (g) doğru sınıf etiketleri yakın plan, (h) Euclid tabanlı YSÖ yakın plan, (i) Jeodezik tabanlı YSÖ yakın plan

VI. TEŞEKKÜR

Bu çalışma 112E195 nolu "Büyük Veri Setlerinin Yaklaşık Spektral Öbeklenmesi için İleri Benzerlik Kriterleri Ve Nicemleme Yöntemleri"isimli araştırma projesi kapsamında TÜBİTAK tarafından desteklenmektedir. Ayrıca Kadim Taşdemir, FP7 Marie Curie Career Integration Grant kapsamında desteklenmektedir.

KAYNAKÇA

[1] Rajalakshmi, N. and Lakshmi Prabha, V. (2015), MRI brain image classification a hybrid approach. Int. J. Imaging Syst. Technol., 25: 226-244. doi: 10.1002/ima.22140

[2] Kannan, Ravi, Santosh Vempala, and Adrian Vetta. "On clusterings: Good, bad and spectral." Journal of the ACM (JACM) 51.3 (2004): 497-515.

[3] Verma, Deepak, and Marina Meila. "A comparison of spectral clustering algorithms." University of Washington Tech Rep UWCSE030501 1 (2003): 1-18.

[4] Wang, Liang, et al. "Approximate spectral clustering." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2009. 134-146.

[5] Taşdemir, Kadim. "Vector quantization based approximate spectral clustering of large datasets." Pattern Recognition 45.8 (2012): 3034-3044.

[6] Taşdemir, Kadim, Yalçin, Berna and Yildirim, Isa . "Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures." Pattern Recognition 48.4 (2015): 1461-1473.

[7] Tasdemir, Kadim, Moazzen, Yaser and Yildirim, Isa . "Geodesic based similarities for approximate spectral clustering." Pattern Recognition (ICPR), 2014 22nd International Conference on. IEEE, 2014.

[8] Yalcin, Berna, and Tasdemir, Kadim. "The use of k-means++ for approximate spectral clustering of large datasets." Signal Processing and Communications Applications Conference (SIU), 2014 22nd. IEEE, 2014.

[9] Menze, Bjoern, Reyes, Mauricio and Leemput, Koen Van . "The Multimodal Brain TumorImage Segmentation Benchmark (BRATS)." (2014).

[10] Asuncion, Arthur, and Newman, David. "UCI machine learning repository." (2007).