یادگیری نیمه‌نظارتی کرنل مرکب با استفاده از روش‌های یادگیری معیار فاصله

طاهره زارع بیدکی*، محمدتقی صادقی و حمیدرضا ابوطالبی

گروه تحقیقاتی پردازش سیگنال، دانشکده مهندسی برق، دانشگاه یزد ،یزد، ایران

چکیده

معیار فاصله، نقشی کلیدی در بسیاری از الگوریتم‌های آموزش ماشین و شناسایی آماری الگو دارد؛ به‌گونه‌ای که انتخاب تابع فاصله مناسب، تأثیر مستقیمی بر عملکرد این الگوریتم‌ها دارد. در سال‌های اخیر، آموزش معیار فاصله با استفاده از نمونه‌های برچسب‌دار و یا دیگر اطلاعات موجود، یکی از حوزه‌های بسیار فعال در حوزه آموزش ماشین شده است. پژوهش‌ها در این راستا، نشان داده است که معیارهای سنجش فاصله مبتنی بر یادگیری، عملکرد بسیار بهتری در مقایسه با معیارهای فاصله مرسوم از قبیل فاصله اقلیدسی دارند. با گسترش این الگوریتم‌ها، نوع مبتنی بر کرنل برخی از این الگوریتم‌ها نیز ارائه شده که در آنها با استفاده از تابع کرنل، نمونه‌ها به‌طور غیر صریح به فضای ویژگی جدیدی با ابعاد بالاتر نگاشت یافته و سپس در این فضای ویژگی جدید، معیار فاصله برای کاربرد مورد نظر آموزش داده می‌شود. برخلاف عملکرد بسیار خوب توابع کرنل در الگوریتم‌های مختلف، یکی از مسائلی که در این الگوریتم‌ها وجود دارد، انتخاب کرنل مناسب و یا پارامترهای مناسب برای یک کرنل مشخص است. استفاده از کرنل مرکب به جای استفاده از یک کرنل به‌تنهایی، بهترین راه حلی است که تاکنون برای این مسئله ارائه شده است. در فرآیند دست‌یابی به کرنل مرکب بهینه نیز، استفاده از الگوریتم‌های یادگیری اهمیت دارد. در این پژوهش، با ادغام این دو فرآیند یادگیری، ساختارهای نیمه نظارتی متفاوتی برای تعیین وزن کرنل‌ها در یک ترکیب کرنلی ارائه می‌شود. کرنل مرکب نهایی برای سنجش فاصله داده‌ها در کاربرد خوشه‌بندی مورد استفاده واقع می‌شود. در ساختارهای نیمه نظارتی بررسی‌شده، سعی بر آن است که در فرآیند بهینه‌سازی با تعیین تابع هدف مناسب، وزن کرنل‌ها به‌گونه‌ای تعیین شود که فاصله زوج‌های مشابه کمینه و فاصله زوج‌های نامشابه بیشینه شود. بررسی عملکرد این ساختارهای پیشنهادی بر روی داده مصنوعی XOR و همچنین مجموعه داده‌های پایگاه داده UCI نشان دهندهٔ مؤثر بودن ساختارهای پیشنهادی است.

واژگان کلیدی: یادگیری معیار فاصله، یادگیری کرنل مرکب، زوج‌های مشابه، زوج‌های نامشابه، یادگیری نیمه‌نظارتی

Semi Supervised Multiple Kernel Learning using Distance Metric Learning Techniques

Tahereh Zare*, Mohammad Taghi Sadeghi and Hamid Reza Abutalebi

Signal Processing Research Group, Electrical Engineering Department, Yazd University, Yazd, Iran

Abstract

Distance metric has a key role in many machine learning and computer vision algorithms so that choosing an appropriate distance metric has a direct effect on the performance of such algorithms. Recently, distance metric learning using labeled data or other available supervisory information has become a very active research area in machine learning applications. Studies in this area have shown that distance metric learning-

* نویسندهٔ عهده‌دار مکاتبات

[DOI: 10.18869/acadpub.jsdp.14.1.53]

based algorithms considerably outperform the commonly used distance metrics such as Euclidean distance. In the kernelized version of the metric learning algorithms, the data points are implicitly mapped into a new feature space using a non-linear kernel function. The associated distance metric is then learned in this new feature space. Utilizing kernel function improves the performance of pattern recognition algorithms, however choosing a proper kernel and tuning its parameter(s) are the main issues in such methods. Using of an appropriate composite kernel instead of a single kernel is one of the best solutions to this problem. In this research study, a multiple kernel is constructed using the weighted sum of a set of basis kernels. In this framework, we propose different learning approaches to determine the kernels weights. The proposed learning techniques arise from the distance metric learning concepts. These methods are performed within a semi supervised framework where different cost functions are considered and the learning process is performed using a limited amount of supervisory information. The supervisory information is in the form of a small set of similarity and/or dissimilarity pairs. We define four distance metric based cost functions in order to optimize the multiple kernel weight. In the first structure, the average distance between the similarity pairs is considered as the cost function. The cost function is minimized subject to maximizing of the average distance between the dissimilarity pairs. This is in fact, a commonly used goal in the distance metric learning problem. In the next structure, it is tried to preserve the topological structure of the data by using of the idea of graph Laplacian. For this purpose, we add a penalty term to the cost function which preserves the topological structure of the data. This penalty term is also used in the other two structures. In the third arrangement, the effect of each dissimilarity pair is considered as an independent constraint. Finally, in the last structure, maximization of the distance between the dissimilarity pairs is considered within the cost function not as a constraint. The proposed methods are examined in the clustering application using the kernel k-means clustering algorithm. Both synthetic (a XOR data set) and real data sets (the UCI data) used in the experiments and the performance of the clustering algorithm using single kernels, are considered as the baseline. Our experimental results confirm that using the multiple kernel not only improves the clustering result but also makes the algorithm independent of choosing the best kernel. The results also show that increasing of the number of constraints, as in the third structures, leads to instability of the algorithm which is expected.

Keyword: Distance Metric Learning, Multiple Kernel Learning, Similarity pairs, Dissimilarity pairs, Semi supervised

۱-مقدمه

عملکرد بسیاری از الگوریتم‌های آموزش ماشین، وابستگی زیادی به معیار مورد استفاده برای سنجش میزان شباهت و یا عدم شباهت الگوهای ورودی دارد [2]،[3]. توابع فاصله، متداول‌ترین معیارهایی هستند که در کاربردهای مختلف مورد استفاده قرار می‌گیرند. به‌عنوان مثال می‌توان به طبقه‌بندی کننده k نزدیک‌ترین همسایه (k-nn[1]) به‌عنوان یک روش نظارتی[2] و الگوریتم خوشه‌بندی k-means به عنوان یکی از روش‌های بدون نظارت[3] اشاره کرد که عملکرد هر دو روش به‌طور قابل توجهی به معیار فاصله انتخاب‌شده وابسته است. یکی از معروف‌ترین معیارهای فاصله که در کاربردهای زیادی از آن استفاده می‌شود، معیار فاصله اقلیدسی است. از نظر تئوری، در شرایط خاصی که احتمالات پیشین طبقات مختلف داده‌ها یکسان باشد، توزیع داده‌ها گوسی بوده و ماتریس‌های کواریانس آنها برابر و قطری باشد، این معیار بهینه است؛ اما پژوهش‌ها نشان داده است که این معیار پاسخ‌گوی داده‌های مختلف در شرایط متفاوت نبوده و در برخی از موارد امکان تفکیک بهینه نمونه‌های متعلق به دسته‌های مختلف با استفاده از این معیار فاصله وجود ندارد. بنابراین انتخاب تابع فاصله مناسب برای بسیاری از کاربردهای عملی ضروری است. بدیهی است که تعیین معیار سنجش فاصله مناسب با سعی و خطا نیز کارآیی لازم را ندارد؛ بر این اساس، در سال‌های اخیر، الگوریتم‌های آموزش معیار فاصله به‌عنوان راه‌حلی بهینه برای انتخاب معیار فاصله مورد توجه قرار گرفته است. نشان داده شده که الگوریتم‌های آموزش معیار فاصله، کارآیی روش‌های مبتنی بر معیار فاصله را بهبود می‌دهند [1]،[4].

الگوریتم‌های ارائه‌شده برای یادگیری معیار فاصله را با توجه به میزان داده‌های آموزشی موجود، می‌توان در سه گروه یادگیری معیار فاصله نظارتی، نیمه‌نظارتی و بدون نظارت تقسیم‌بندی کرد [5].

در دسته الگوریتم‌های نظارتی، پژوهش‌های گسترده‌ای برای آموزش معیار فاصله در طبقه‌بندی کننده

[1] k-nearest neighbor
[2] Supervised method
[3] Unsupervised method

[DOI: 10.18869/acadpub.jsdp.14.1.53]

k نزدیک‌ترین همسایه(k-nn) انجام شده است [6]-[10]. در این الگوریتم‌ها، ابتدا داده‌ها به فضای جدیدی توسط ماتریس فاصله آموزش داده شده، نگاشت می‌شوند؛ به‌گونه‌ای که در این فضا k نزدیک‌ترین همسایه‌ها متعلق به دسته مشابهی هستند. به‌عبارتی این ماتریس نگاشت به‌گونه‌ای آموزش داده می‌شود که فاصله میان k نزدیک‌ترین نمونه آموزشی با برچسب طبقهٔ مشابه کمینه و فاصله این نمونه‌ها با سایر نمونه‌ها از دسته‌های مختلف بیشینه شود. عیب عمده این روش‌ها در نیازمندی به داشتن اطلاعات برچسب تمام نمونه‌هاست که در برخی از کاربردها دسترسی به چنین اطلاعاتی امکان‌پذیر نیست. در بیش‌تر الگوریتم‌های آموزش معیار فاصله برای کاربرد بدون نظارت، هدف جستجوی یک فضای مناسب با ابعاد کمتر برای نگاشت نمونه‌ها از فضای اولیه به یک فضای با ابعاد کمتر است [11]،[12]. به همین دلیل الگوریتم‌های آموزش معیار فاصله بدون نظارت، اغلب الگوریتم‌های آموزش خمینه [1] نیز نامیده می‌شوند. در مقایسه با شیوه‌های آموزشی نظارتی، الگوریتم‌های آموزشی بدون نظارت از قبیل خوشه‌بندی و روش‌های کاهش ابعاد بدون نظارتی، حساسیت بیشتری به تابع فاصله انتخاب‌شده دارند؛ چون این الگوریتم‌ها، مبتنی بر سنجش شباهت یا عدم شباهت داده‌ها بوده و به برچسب داده‌ها دسترسی ندارند. الگوریتم‌های آموزش معیار فاصله، برای چنین کاربردهایی (به‌علت عدم دسترسی به اطلاعات مهمی مانند برچسب طبقه داده‌ها) به‌طورعمومی عملکرد ضعیفی دارند.

دسته دیگری از الگوریتم‌های آموزشی، الگوریتم‌های نیمه‌نظارتی[2] است که در سال‌های اخیر مورد توجه بسیاری از پژوهش‌گران قرار گرفته است. در این الگوریتم‌ها فرض می‌شود که در کنار مجموعه بزرگی از نمونه‌های بدون برچسب طبقه، اطلاعات نظارتی محدودی در دسترس است. بنابراین در الگوریتم‌های آموزش معیار فاصله از این اطلاعات محدود برای آموزش استفاده می‌شود[13]-[15]. در اکثر الگوریتم‌های آموزش معیار فاصله، این اطلاعات محدود در قالب دو مجموعه از زوج‌های مشابه و زوج‌های نامشابه ظاهر می‌شوند که دارای اطلاعاتی ضعیف‌تر از برچسب طبقه داده‌ها بوده، اما به طبیعت مسائل خوشه‌بندی نزدیک‌تر است. زوج‌های مشابه، نمونه‌هایی هستند که به یک دسته تعلق دارند و زوج‌های نامشابه، نمونه‌هایی هستند که به دسته‌های متفاوت تعلق دارند. بر اساس مجموعه زوج‌های ذکرشده، الگوریتم آموزش معیار فاصله می‌کوشد تا معیار فاصله مناسبی را تعیین کند؛ به‌گونه‌ای که فاصله میان زوج نمونه‌های مشابه کم و از طرفی فاصله میان زوج نمونه‌های نامشابه زیاد شود. بر همین اساس، در [16] یک الگوریتم تکراری[3] برای آموزش ماتریس Mahalanobis با درنظرگرفتن قیدهای مرتبط با چنین زوج‌داده‌هایی برای‌کاربرد خوشه‌بندی ارائه شد. در [17] یک الگوریتم غیر تکراری که با نام تحلیل مؤلفه‌های مرتبط (RCA)[4]شناخته می‌شود، ارائه شد که در آن تنها از زوج‌های مشابه برای یادگیری معیار فاصله استفاده می‌شود. ایده اصلی در روش یادشده، اعمال وزن‌های بزرگ‌تر به ویژگی‌های مرتبط و وزن‌های کوچک به ویژگی‌های غیر مرتبط است. شکل تعمیم‌یافته الگوریتم RCA که در آن علاوه‌بر زوج‌های مشابه از زوج‌های نامشابه نیز استفاده می‌شود در [18] ارائه شد.

در بیش‌تر روش‌های ارائه‌شده برای یادگیری معیار فاصله، داده‌ها با استفاده از ماتریس نگاشتی، نگاشتی بطور خطی به فضایی جدید نگاشت یافته و سپس تابع فاصله اعمال می‌شود. عناصر این ماتریس طی فرآیند یادگیری تعیین می‌شوند. حال آن که در برخی از موارد، نگاشت خطی پاسخ‌گوی پیچیدگی ساختار داده‌ها و یا مسئله مورد بررسی نیست؛ به عبارت دیگر، نمی‌توان با استفاده از نگاشت خطی، نمایش مناسبی برای داده‌ها به‌دست آورد. به همین منظور، نوع مبتنی بر کرنل برخی از روش‌های یادگیری معیار فاصله، نیز ارائه شده است [19]-[21]. در [20]، با استفاده از روشی موسوم به kernel-β، از یک کرنل گوسی در شکل کرنلی فاصله اقلیدسی برای خوشه‌بندی داده‌ها استفاده شده است که در آن تحلیل مقادیر ویژه بر روی ماتریس کرنل ذکرشده اعمال‌شده و مقادیر ویژه در یک الگوریتم بهینه‌سازی برای کمینه‌سازی فاصله میان زوج‌های مشابه بهینه شده است. در [22] نیز ماتریس کرنل در چارچوب مسئله یادگیری معیار فاصله به‌گونه‌ای تعیین می‌شود که فاصله زوج‌های مشابه کمینه و فاصله زوج‌های نامشابه بیشینه شود. تابع کرنل را می‌توان تابعی در نظر گرفت که داده‌ها را به‌طور غیرصریح و توسط یک تابع غیرخطی به فضای ویژگی جدیدی نگاشت می‌دهد. تابع کرنل با استفاده از ضرب داخلی داده‌های نگاشت‌یافته، مقدار شباهت هر زوج داده را در فضای ویژگی جدید اندازه می‌گیرد. ماهیت غیرخطی نگاشت ناشی از تابع کرنل امکان تفکیک‌پذیری داده‌های پیچیده‌تر را فراهم

[1] Manifold
[2] Semi-supervised

[3] Iterative
[4] Relevant Component Analysis

[DOI: 10.18869/acadpub.jsdp.14.1.53]

می‌کند؛ اما مسئله اصلی در ارتباط با استفاده از کرنل، انتخاب تابع کرنل بهینه و تنظیم مناسب پارامترهای آن است. درواقع، با استفاده از کرنل‌های متفاوت، می‌توان داده‌ها را به فضاهای ویژگی متفاوتی نگاشت داد. بر این اساس، مهم‌ترین مسئله در روش‌های طبقه‌بندی مبتنی بر کرنل، انتخاب فضای ویژگی با بیشترین قدرت تفکیک‌پذیری است. اگر اطلاعات مربوط به چگونگی پراکندگی داده‌ها در دسترس و یا ابعاد نمونه‌ها کم باشد به‌گونه‌ای که بتوان پراکندگی داده‌ها را مشاهده کرد، بهترین کرنل را بر مبنای پراکندگی داده‌ها تعیین می‌توان کرد؛ اما در بیش‌تر موارد، پراکندگی داده‌ها نامعلوم بوده و یا ابعاد داده‌ها زیاد است؛ به‌گونه‌ای که تعیین کرنل مناسب به‌راحتی امکان‌پذیر نیست. ازاین رو در سال‌های اخیر به استفاده از کرنل مرکب که با ترکیب توابع کرنل حاصل می‌شود به‌جای استفاده از یک تابع کرنل به‌عنوان راه‌حلی برای این مسئله توجه شده است. استفاده از کرنل مرکب و تعیین وفقی پارامترهای مؤثر در ترکیب کرنل‌های پایه، می‌تواند سبب استقلال الگوریتم مورد نظر از چگونگی پراکندگی داده‌ها و تعیین نگاشت بهینه در شرایط متفاوت شود. در روش‌های یادگیری کرنل مرکب (MKL[1])، هدف محاسبه میزان اهمیت هر یک از این فضاهای ویژگی و یا وزن کرنل‌ها در ترکیب کرنلی است که در بیش‌تر موارد از یک الگوریتم تکراری برای محاسبه وزن کرنل‌ها و پارامترهای سامانهٔ یادگیرنده استفاده می‌شود [23]. توابع کرنل و معیارهای سنجش فاصله، هر دو به نوعی شباهت میان نمونه‌ها را اندازه‌گیری می‌کنند و شباهت زیادی در اهداف مورد بررسی در دو روش یادگیری کرنل مرکب و یادگیری معیار فاصله وجود دارد؛ اما با وجود مطالعات فراوانی که در هر دو حوزه یادگیری معیار فاصله و همچنین یادگیری کرنل مرکب صورت گرفته است، کارهای بسیار محدودی برای ادغام این دو حوزه و یا حتی استفاده از الگوریتم‌های یادگیری کرنل مرکب مناسب در شکل کرنلی الگوریتم‌های یادگیری معیار فاصله صورت گرفته است [24]-[28]. در [24] و [25] از ترکیب وزن‌دار کرنل‌ها در چارچوب فرآیند یادگیری معیار فاصله نظارتی استفاده شده است. در [24] روشی ارائه شد که در آن داده‌ها با استفاده از کرنل مرکب (که از ترکیب خطی کرنل‌های پایه حاصل می‌شود) به فضای ویژگی جدیدی نگاشت داده‌شده و سپس در فضای ویژگی جدید از فاصله Mahalanobis برای طبقه‌بندی داده‌ها با استفاده از روش نزدیک‌ترین همسایه استفاده می‌شود. از یک الگوریتم تکراری برای یادگیری وزن کرنل‌های مرکب و همچنین ماتریس نگاشت فاصله Mahalanobis استفاده شده است. در [25] نیز با درنظرگرفتن ترکیب خطی برای کرنل مرکب، از یک الگوریتم یادگیری معیار فاصله با هدف تعیین وزن کرنل ها استفاده شده است. در [26] نیز با درنظرگرفتن دو ساختار برای کرنل مرکب که شامل جمع بدون وزن کرنل‌ها و ساختار الحاقی[2] کرنل‌ها است، شکل کرنلی یک الگوریتم یادگیری معیار فاصله اقلیدسی ارائه شده است. به‌عبارتی در این مقاله ابتدا تجزیهٔ تحلیل مقادیر ویژه بر روی کرنل مرکب اعمال شده و سپس در یک فرآیند آموزشی، وزن ماتریس‌های ویژهٔ به‌دست‌آمده از تجزیه مقادیر ویژه به‌گونه‌ای تعیین می‌شود که فاصله میان زوج‌های مشابه کمینه شود. در [27] و [28] جمع وزن‌دارکرنل‌ها در یک ترکیب خطی در نظر گرفته شده که هدف از الگوریتم یادگیری معیار فاصله، آموزش وزن کرنل‌ها در این ترکیب خطی برای دو کاربرد خوشه‌بندی [27] و طبقه‌بندی با استفاده از طبقه‌بند ماشین بردار پشتیبان (SVM[3]) است [28].

در این پژوهش، ما نیز از روش‌های یادگیری معیار فاصله برای یادگیری کرنل مرکب استفاده کرده و ضمن ارائه توابع هزینه متفاوت، عملکرد این توابع را در یادگیری کرنل مرکب مقایسه می‌کنیم. چنانچه ملاحظه خواهد شد، در این پژوهش، چهار ساختار یادگیری معیار فاصله متفاوت ارائه و بررسی خواهد شد.

سازماندهی این مقاله به قرار زیر است: در بخش ۲ به توضیح اجمالی مفاهیم اصلی یادگیری معیار فاصله می‌پردازیم. مفاهیم مهم مطرح در یادگیری کرنل مرکب نیز در بخش ۳ مورد بررسی قرار خواهد گرفت. در بخش ۴، روش‌های پیشنهادشده برای یادگیری کرنل مرکب در چارچوب مفاهیم یادگیری معیار فاصله با به‌کارگیری توابع هزینهٔ مختلف ارائه خواهد شد. در بخش ۵ به ارزیابی روش‌های پیشنهادشده بر روی داده‌های مصنوعی و واقعی می‌پردازیم. درنهایت در بخش۶، نتایج و پیشنهادها ارائه خواهد شد.

۲-یادگیری معیار فاصله

برخلاف الگوریتم‌های معمول یادگیری که در آن‌ها از اطلاعات مربوط به برچسب هر نمونه آموزشی، در فرآیند یادگیری

[1] Multiple kernel learning

[2] Augmented

[3] Support Vector Machine

[DOI: 10.18869/acadpub.jsdp.14.1.53]

استفاده می‌شود، در یادگیری معیار فاصله به‌طورمعمول از اطلاعات مربوط به برچسب داده‌ها تنها در تشکیل مجموعه‌ای از محدودیت‌ها و یا قیدها به شکل‌های زیر استفاده می‌شود [3]:

- محدودیت یا قید هم‌ارزی: این محدودیت معرف این واقعیت است که یک زوج داده به‌طور معناداری مشابه یکدیگر هستند (متعلق به یک دسته داده). در فرآیند یادگیری فاصله، چنین داده‌هایی بایستی نزدیک به یکدیگر در نظر گرفته شوند.
- محدودیت یا قید غیر هم ارزی: این محدودیت نیز بیان‌کنندهٔ زوج‌داده‌هایی است که به‌طور معناداری نامشابه هستند (متعلق به دسته‌های مختلف داده) و باید در معیار یادگیری فاصله، از یکدیگر دور باشند.

بنابراین معیار فاصله به‌گونه‌ای آموزش داده می‌شود که فاصله بین زوج‌ها در مجموعه قید هم‌ارزی کمینه و فاصله بین زوج‌ها در مجموعه قید غیر هم‌ارزی بیشینه گردد.

فرض کنید $X=\{\mathbf{x}_i\}_{i=1}^{N}$ مجموعه نمونه‌های آموزشی باشند، به‌گونه‌ای که $\mathbf{x}_i \in \mathrm{R}^{n_x}$، N تعداد نمونه‌های آموزشی و n_x بعد بردارهای ویژگی در فضای اولیه است؛ در این‌صورت مجموعه زوج‌های مشابه که مربوط به قید هم‌ارزی است، به‌صورت زیر نمایش داده می‌شود:

$$S=\{(\mathbf{x}_i,\mathbf{x}_j)\mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same class}\} \quad (۱)$$

و مجموعه زوج‌های نامشابه، مربوط به قید غیر هم ارزی نیز به‌صورت زیر خواهد بود:

$$D=\{(\mathbf{x}_i,\mathbf{x}_j)\mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to different classes}\} \quad (۲)$$

حال فرض کنید ماتریس معیار فاصله به وسیله $\mathbf{A}\in \mathrm{R}^{n_x\times n_x}$ نشان داده شود، در این‌صورت مربع فاصله بین دو نمونه $\mathbf{x}$ و $\mathbf{y}$ به‌صورت زیر محاسبه شود:

$$d_{\mathbf{A}}^2(\mathbf{x},\mathbf{y})=\|\mathbf{x}-\mathbf{y}\|_{\mathbf{A}}^2=(\mathbf{x}-\mathbf{y})^T\mathbf{A}(\mathbf{x}-\mathbf{y}) \quad (۳)$$

در رابطهٔ بالا اگر $\mathrm{A}=\mathrm{I}$ که I ماتریس یکانی است، در این صورت معیار فاصله بالا معادل با فاصله اقلیدسی خواهد بود. همچنین اگر ماتریس A به‌صورت یک ماتریس قطری از مقادیر ویژه مربوط به داده‌های ورودی در نظر گرفته شود، رابطه فاصله بالا معادل با فاصله Mahalanobis خواهد بود. به‌طورکلی اگر A یک ماتریس قطری باشد، رابطه فاصلهٔ بالا را این‌چنین می‌توان تفسیر کرد که به هر یک از محورهای مختصات (هر بعد از فضای ورودی) وزن متفاوتی متناسب با مؤلفه متناظر با آن در ماتریس قطری $\mathbf{A}$ اعمال می‌شود. به‌عبارتی ماتریس $\mathbf{A}$ فاصله‌ای از خانواده فاصله‌های Mahalanobis روی فضای R^d را خواهد ساخت. از طرفی دیگر، ماتریس قطری $\mathbf{A}$ در معیار فاصله را می‌توان چنین تفسیر کرد که ابتدا نمونه ورودی $\mathbf{x}$ با رابطه $\mathbf{A}^{1/2}\mathbf{x}$ تغییر مقیاس یافته و سپس فاصله اقلیدسی در فضای جدید اعمال می‌شود که این نتیجه از $\mathbf{A}=\mathbf{A}^{1/2}\mathbf{A}^{1/2}$ برای ماتریس قطری $\mathbf{A}$ به‌دست می‌آید. الگوریتم‌های متفاوتی با درنظرگرفتن قیدهای گوناگون بر روی زوج‌های مشابه و نامشابه، برای یادگیری ماتریس نگاشت $\mathbf{A}$ ارائه شده است. در این مقاله، تمرکز اصلی بر روی یادگیری ماتریس کرنل بهینه به جای ماتریس نگاشت در معیار فاصله است؛ برای این منظور در بخش بعد به معرفی تابع کرنل و الگوریتم‌های یادگیری کرنل مرکب خواهیم پرداخت.

۳-یادگیری کرنل مرکب

کلمه کرنل به‌تنهایی معانی متعددی در ریاضیات و علم آمار دارد؛ اما در کاربرد یادگیری ماشین، منظور از کرنل، کرنل‌های (نیمه) معین مثبت است. تابع کرنل را می‌توان به‌عنوان یک معیار شباهت بین دو نمونه در نظر گرفت که شباهت میان زوج‌نمونه‌ها توسط یک مقدار حقیقی به‌دست آمده از تابع کرنل مشخص می‌شود. تابع کرنل به‌صورت $k(\mathbf{x},\mathbf{y})=\langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$ تعریف می‌شود که در آن $\phi:\mathrm{R}^{n_x}\to H$ نگاشت غیر خطی از فضای اولیه به فضای ویژگی ناشی از کرنل (فضای هیلبرت H) است و $\langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$ نشان‌دهندهٔ ضرب نقطه‌ای بین دو بردار دادهٔ دلخواه $\phi(\mathbf{x})$ و $\phi(\mathbf{y})$ است. بنابراین محاسبه تابع کرنل را می‌توان به مثابه انجام عملیات ضرب نقطه‌ای دو داده در فضای هیلبرت (فضای ویژگی) متناظر با آن کرنل در نظر گرفت و این یکی از مهم‌ترین خصوصیات توابع کرنل است که سبب معرفی ترفند کرنـل[1] شده است. به‌عبارتی ترفند کرنل به این صورت بیان می‌شود: در الگوریتم‌هایی که مبتنی بر نگاشت داده‌های برداری به فضای ویژگی جدید و ضرب نقطه‌ای داده‌های نگاشت‌یافته باشند، می‌توان با استفاده از تابع کرنل مناسب، ضرب داخلی در فضای نگاشت‌یافته را، به‌طورضمنی، با حجم محاسبات به‌مراتب کمتر با ضرب

[1] Kernel Trick

[DOI: 10.18869/acadpub.jsdp.14.1.53]

نقطه‌ای در فضای اولیه و اعمال تابع کرنل پیاده‌سازی کرد. فرض کنید تابع کرنل $k : X \times X \subseteq \mathrm{R}^{n_x} \times \mathrm{R}^{n_x} \to \mathrm{R}$ داده شده باشد و N داده در $X \subseteq \mathrm{R}^d$ نیز موجود باشد. به ماتریس $\mathbf{K}$ که اندازهٔ آن $N \times N$ بوده و هر المان آن با رابطهٔ $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ به‌دست می‌آید، ماتریس گرام حاصل از تابع کرنل k گفته می‌شود. توابع کرنل چندجمله‌ای، گوسی و سیگموئید نمونه‌هایی از پرکاربردترین توابع کرنل هستند.

با توجه به خواص نگاشت کرنلی داده‌ها و ترفند کرنل، نسخهٔ مبتنی بر کرنل الگوریتم‌هایی که در آنها فرآیند آموزش با محاسبهٔ ضرب نقطه‌ای بین نمونه‌های آموزشی انجام می‌پذیرد، قابل ارائه است. به‌عبارت دیگر الگوریتم‌های مبتنی بر کرنل، الگوریتم‌هایی هستند که در آن‌ها تنها به ماتریس کرنل به‌منظور آموزش نیاز است؛ اما انتخاب کرنل مناسب و پارامترهای آن (به‌عنوان مثال درجه چندجمله‌ای در کرنل چندجمله‌ای و یا پارامتر واریانس در کرنل گوسی) یکی از مهم‌ترین مسائل در روش‌های کرنلی است؛ چون تابع کرنل باید متناسب با توزیع داده‌های ورودی انتخاب شود و نوع کرنل انتخاب‌شده تأثیر زیادی در کیفیت عملکرد الگوریتم یادگیری مورد بررسی خواهد داشت. برای تعیین کرنل مناسب و یا پارامترهای آن یکی از روش‌هایی که ممکن است پیشنهاد شود، ارزیابی سامانه با استفاده از روش اعتبارسنجی متقابل است؛ اما استفاده از این روش نه‌تنها زمان‌بر است؛ بلکه تضمینی برای یافتن کرنل بهینه وجود ندارد. از طرفی در برخی از موارد استفاده از یک کرنل، پاسخ‌گوی پیچیدگی داده‌ها یا مسئلهٔ مورد بررسی نبوده و به همین دلیل در سال‌های اخیر استفاده از ترکیب مناسب تعدادی از توابع کرنل یا کرنل مرکب به جای استفاده از یک تابع کرنل پیشنهاد شده است. کرنل‌های مرکب را می‌توان به طرق مختلف و از دیدگاه‌های مختلف مورد بررسی قرار داد. در حالت کلی کرنل مرکب به‌صورت زیر تعریف می‌شود [۲۲]:

$$k_\beta(\mathbf{x}_i, \mathbf{x}_j) = f_\beta\left(\left\{k_m(\mathbf{x}_i^m, \mathbf{x}_j^m)\right\}_{m=1}^{M}\right) \qquad (۴)$$

که در اینجا M تابع کرنل $\{k_m(.,.)\}_{m=1}^{M}$ به‌عنوان کرنل‌های پایه (توابع کرنل اولیه) در نظر گرفته شده و $f_\beta(.)$ تابع ترکیب این کرنل‌ها است که می‌تواند خطی و یا غیر خطی برحسب کرنل‌های پایه باشد. با توجه به تعریف ارائه‌شده برای کرنل مرکب در رابطه (۴)، کرنل مرکب را می‌توان به دو طریق و یا با دو هدف ساخت [23]:

- تولید کرنل مرکب با استفاده از ترکیب توابع مختلف کرنل و یا به‌عبارتی به‌صورت $k_\beta(\mathbf{x}_i, \mathbf{x}_j) = f_\beta\left(\left\{k_m(\mathbf{x}_i, \mathbf{x}_j)\right\}_{m=1}^{M}\right)$: کرنل‌های مختلف را می‌توان معادل با معیارهای متفاوتی برای سنجش شباهت داده‌ها دانست. در این صورت به جای جستجو برای پیداکردن تابع کرنلی که بهترین نمایش از شباهت میان نمونه‌ها را ارائه می‌دهد، می‌توان از یک الگوریتم یادگیری برای انتخاب و یا ترکیب توابع کرنل مختلف (نمایش‌های مختلف از شباهت) استفاده کرد. در این حالت کرنل‌های پایه ممکن است، توابع مختلفی مانند کرنل گوسی، خطی و چندجمله‌ای باشند و یا توابع کرنل با استفاده از یک نوع تابع کرنل، ولی به‌ازای پارامترهای مختلف مانند کرنل گوسی، اما به‌ازای مقادیر واریانس مختلف تولید شوند. زیرنویس m برای توابع کرنل پایه $k_m(.,.)$ به همین دلیل انتخاب شده است.
- تولید کرنل مرکب با استفاده از منابع ورودی متفاوت و یا به‌عبارتی به صورت $k_\beta(\mathbf{x}_i, \mathbf{x}_j) = f_\beta\left(\left\{k(\mathbf{x}_i^m, \mathbf{x}_j^m)\right\}_{m=1}^{M}\right)$: توابع کرنل مختلف ممکن است، به واسطهٔ نمایش‌های مختلف داده‌ها ناشی از منابع ورودی مختلف و یا مودالیتی‌های مختلف ایجاد شوند و از آنجایی که در نمایش‌های مختلف داده‌ها از معیارهای شباهت سنجی متفاوت و درنتیجه کرنل‌های متفاوت استفاده می‌شود؛ در این حالت ترکیب کرنل‌های مختلف راهی برای ترکیب منابع اطلاعاتی مختلف است. استفاده از بالانویس m برای نمونه‌های $\mathbf{x}_i^m$ بیان‌گر فضاهای ورودی مختلف است.

روش‌های بسیار زیادی تاکنون برای ترکیب توابع کرنل پایه ارائه شده است که در آنها از توابع مختلف $f_\beta(.)$ برای ساخت کرنل‌های مرکب و سپس یادگیری این کرنل‌ها استفاده شده است. در این مقاله، ما از مسئله یادگیری معیار فاصله برای یادگیری کرنل مرکب در یک ترکیب خطی، استفاده می‌کنیم. در بخش بعد به توضیح روش‌های بررسی‌شده برای این منظور خواهیم پرداخت.

[DOI: 10.18869/acadpub.jsdp.14.1.53]

۴-ادغام روش‌های یادگیری معیار فاصله و یادگیری کرنل مرکب

همان‌طورکه درقبل اشاره شد، ارتباط تنگاتنگی میان یادگیری ماتریس کرنل و یادگیری معیار فاصله وجود دارد. هر دو مقادیر کرنل و فاصله، شباهت میان زوج‌نمونه‌ها را محاسبه می‌کنند. همچنین در هر دو روش، داده‌ها به فضایی مناسب که توسط کرنل و یا ماتریس معیار تعیین می‌شود، نگاشت می‌شوند. همان‌طورکه در بخش ۲ اشاره شد، در روش‌های یادگیری معیار فاصله، هدف تعیین مقادیر بهینهٔ مجموعه‌ای از پارامترها در معیار فاصله با استفاده از زوج‌های مشابه و یا نامشابه است که این مجموعه از پارامترها در بیش‌تر روش‌ها در قالب ماتریس نگاشت $\mathbf{A}$ لحاظ می‌شود. در این پژوهش، ساختاری متفاوت برای یادگیری معیار فاصله در نظر گرفته شده است که در آن با لحاظ‌کردن جمع وزن‌دار مجموعه‌ای از کرنل‌ها به‌عنوان کرنل مرکب، پارامتر وزن کرنل‌ها در فرآیند یادگیری معیار فاصله آموزش داده می‌شود و نگاشت به فضای مورد نظر توسط ماتریس کرنل حاصل صورت می‌گیرد. بنابراین در این پژوهش، هدف محاسبه وزن کرنل‌های بهینه در چارچوب مسئله یادگیری معیار فاصله است. درواقع در مقایسه با سایر روش‌های یادگیری معیار فاصله، ماتریس $\mathbf{A}$ ماتریس یکانی $\mathbf{I}$ در نظر گرفته می‌شود؛ اما با نگاشت کرنلی بهینه، داده‌ها به فضای مناسبی نگاشت و در آن فضا از معیار فاصله اقلیدسی استفاده می‌شود. یکی از مزایای ادغام کرنل در یادگیری معیار فاصله، استفاده از خاصیت نگاشت غیر خطی ناشی از تابع کرنل است که سبب افزایش قدرت الگوریتم نهایی در تفکیک‌پذیری داده‌های مربوط به دسته‌های مختلف می‌شود.

از طرفی همان‌طورکه درقبل شرح داده شد، یکی از چالش‌های اساسی در برخی از مسائل شناسایی الگو، عدم وجود اطلاعات کافی درخصوص برچسب داده‌های آموزشی است. در سال‌های اخیر تمرکز بسیاری از پژوهش‌گران به روش‌های نیمه‌نظارتی معطوف شده است؛ زیرا جمع‌آوری مجموعه بزرگی از داده‌های برچسب‌دار برای روش‌های نظارتی هزینه زیادی به دنبال داشته و از طرفی استفاده از داده‌های بدون برچسب در روش‌های غیرنظارتی، اغلب موجب افت کیفیت عملکرد سامانه در مقایسه با روش‌های نظارتی می‌شود. ازاین‌رو روش‌های نیمه‌نظارتی در کاربردهای مختلف یادگیری ماشین وارد شده‌اند؛ چون با استفاده از این روش‌ها انعطاف‌پذیری سامانه مورد بررسی افزایش می‌یابد. روش‌های یادگیری معیار فاصله، یکی از حوزه‌هایی است که در آن، روش‌های نیمه‌نظارتی به‌صورت گسترده استفاده شده است. در این روش‌ها، اطلاعات کمتری در مقایسه با روش‌های نظارتی برای یادگیری معیار فاصله مناسب وجود دارد. این اطلاعات می‌تواند به‌صورت مجموعه کوچکی از زوج‌های مشابه و نامشابه در روابط (۱) و (۲) و یا تعداد محدودی نمونه برچسب‌دار باشد که این مجموعه‌ها و یا نمونه‌های برچسب‌دار در مقایسه با نمونه‌های بدون برچسب، خیلی کم است.

در این مقاله، برای ادغام مفاهیم یادگیری کرنل مرکب و یادگیری معیار فاصله، چهار ساختار نیمه‌نظارتی مختلف که در آن‌ها توابع هزینه متفاوتی لحاظ شده است، ارائه و در کاربرد خوشه‌بندی مورد ارزیابی واقع می‌شوند.

فرض کنید $X=\{\mathbf{x}_i\}_{i=1}^{N}$ مجموعه کل داده‌های موجود باشد که $\mathbf{x}_i \in \mathrm{R}^{n_x}$ و N تعداد کل نمونه‌هاست. با درنظرگرفتن چارچوب نیمه‌نظارتی، از میان N نمونه موجود، تنها مجموعه کوچکی از زوج‌های مشابه S و نامشابه D به‌صورت روابط (۱) و (۲) در دسترس است، به‌گونه‌ای که $|S| \mathrm{R}\ N$ و $|D| \mathrm{R}\ N$. در اینجا $|S|$ و $|D|$ به‌ترتیب بیان‌گر تعداد زوج‌ها در مجموعه‌های S و D است. کرنل مرکب در کلیه ساختارها به‌صورت ترکیب خطی کرنل‌های پایه و به‌صورت زیر در نظر گرفته می‌شود:

$$\mathbf{K}_{\beta}=\sum_{m=1}^{M}\beta_m\mathbf{K}_m \qquad (۵)$$

که در رابطهٔ بالا، β_m وزن مربوط به کرنل $\mathbf{K}_m$ است.

برای سادگی، نمودار کلی از ساختار یادگیری کرنل مرکب در چارچوب مسئله یادگیری معیار فاصله در شکل (۱) نشان داده شده است.

در ادامه جزئیات هر ساختار پیشنهادی بررسی می‌شود.

۴-۱-ساختار پیشنهادی ۱:

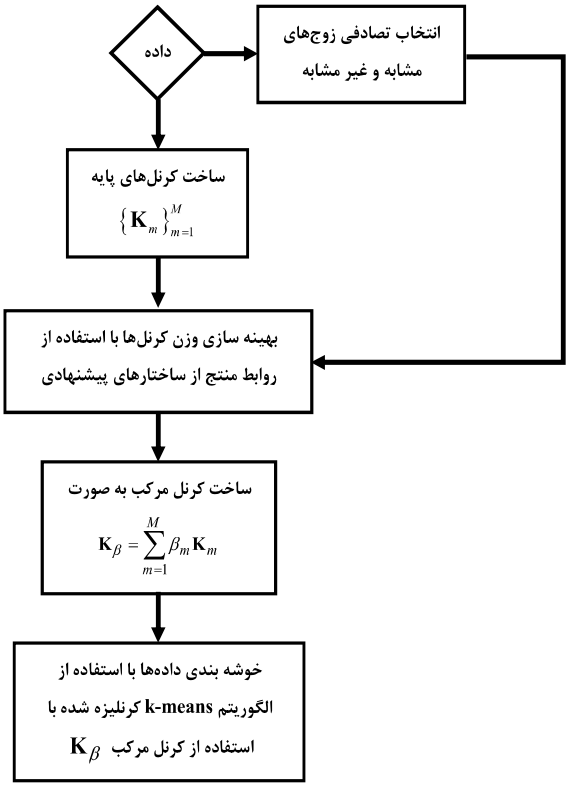
در این ساختار، برای تعیین مقدار بهینه‌ی وزن کرنل‌ها، β_m، تابع هزینه زیر در نظر گرفته می‌شود:

$$\min \frac{1}{|S|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2 \qquad (۶)$$

$$s.t.\ \frac{1}{|D|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D}\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2 \geq c$$

$$\sum_{m=1}^{M}\beta_m=1$$

$$\beta_m \geq 0 \quad \forall m=1,2,...,M$$



(شکل-۱): نمودار کلی ساختار یادگیری کرنل مرکب در چارچوب یادگیری معیار فاصله

(Figure-1): The flowchart of multiple kernel learning method in the distance metric learning framework

که در اینجا Φ بردار ویژگی متناظر با کرنل مرکب $\mathbf{K}_\beta$ است. همچنین $\mathbf{c}$ یک مقدار ثابت است و می‌توان به‌دلخواه مقدار ۱ و یا هر عددی را به آن نسبت داد. همان‌طورکه دیده می‌شود، علاوه‌بر شرط نرم-۱ بر روی وزن‌ها، شرط مثبت‌بودن نیز برای آن‌ها در نظر گرفته شده تا کرنل مرکب شرط نیمه‌معین مثبت‌بودن را برآورده سازد. مفهوم تابع هزینه بالا این است که باید وزن کرنل‌ها به‌گونه‌ای تعیین شود که متوسط فاصله میان زوج‌های مشابه کمینه شود به شرط آنکه متوسط فاصله میان زوج‌های نامشابه از یک حدی که در اینجا $\mathbf{c}$ است، بیشتر شود. تابع لاگرانژ برای مسئله بالا به‌صورت زیر خواهد بود:

$$L = \frac{1}{|S|}\sum_{m=1}^{M}\beta_m \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}\left[(\mathbf{K}_m)_{ii}+(\mathbf{K}_m)_{jj}-(2\mathbf{K}_m)_{ij}\right] \qquad (۷)$$
$$-\lambda_1\left[\left(\frac{1}{|D|}\sum_{m=1}^{M}\beta_m \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D}\left[(\mathbf{K}_m)_{ii}+(\mathbf{K}_m)_{jj}-(2\mathbf{K}_m)_{ij}\right]\right)-c\right]_{\rfloor}$$
$$-\lambda_2\left[1-\sum_{m=1}^{M}\beta_m\right]_{\rfloor}-\sum_{m=1}^{M}\eta_m\beta_m$$

که در رابطهٔ بالا، λ_1، λ_2 و $\{\eta_m\}_{m=1}^{M}$ ضرایب لاگرانژ هستند. با ساده‌سازی رابطه بالا خواهیم داشت:

$$L = \frac{1}{|S|}\sum_{m=1}^{M}\beta_m \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{K}_m(\mathbf{e}_i-\mathbf{e}_j)$$
$$-\lambda_1\left[\left(\frac{1}{|D|}\sum_{m=1}^{M}\beta_m \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D}(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{K}_m(\mathbf{e}_i-\mathbf{e}_j)\right)-c\right]_{\rfloor}$$
$$-\lambda_2\left[1-\sum_{m=1}^{M}\beta_m\right]_{\rfloor}-\sum_{m=1}^{M}\eta_m\beta_m$$
$$=\sum_{m=1}^{M}\beta_m a_m-\lambda_1\left[\sum_{m=1}^{M}\beta_m b_m-c\right]_{\rfloor}$$
$$-\lambda_2\left[1-\sum_{m=1}^{M}\beta_m\right]_{\rfloor}-\sum_{m=1}^{M}\eta_m\beta_m$$

(۸)

که در رابطه بالا $\mathbf{e}_i$ ستون $\mathbf{i}$ ام از ماتریس یکانی با ابعاد $N\times N$ بوده و a_m و b_m مطابق روابط زیر محاسبه می‌شوند:

$$a_m = \frac{1}{|S|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{K}_m(\mathbf{e}_i-\mathbf{e}_j) \qquad (۹)$$

$$b_m = \frac{1}{|D|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{K}_m(\mathbf{e}_i-\mathbf{e}_j) \qquad (۱۰)$$

مسئله بهینه‌سازی بالا را می‌توان به‌صورت زیر بازنویسی نمود:

$$\min_{\boldsymbol{\beta}} \quad \mathbf{a}^T\boldsymbol{\beta}$$
$$\text{s.t.} \quad \mathbf{b}^T\boldsymbol{\beta}\geq c,\ \ \mathbf{1}^T\boldsymbol{\beta}=1,\ \ \boldsymbol{\beta}\geq 0 \qquad (۱۱)$$

که یک مسئله خطی با قیدهای خطی است و از روش‌هایی مانند Simplex قابل حل است [29].

۴-۲-ساختار پیشنهادی ۲:

در این ساختار به تابع هزینه ساختار ۱ یک عبارت جریمه که سعی در برهم‌نزدن ساختار کلی توپولوژی داده‌ها دارد، اضافه می‌شود. درواقع، در این تابع هزینه از ایده گراف برای حفظ توپولوژی داده‌ها در فضای اولیه استفاده شده و قیدهای مسئله مانند ساختار ۱ در نظر گرفته شده است. بر این اساس تابع هزینه به‌صورت زیر است:

$$\min \frac{1}{|S|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2 \qquad (۱۲)$$
$$+\alpha\frac{1}{2N}\sum_i\sum_j\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2 W_{ij}$$
$$s.t.\ \frac{1}{|D|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D}\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2\geq c$$
$$\sum_{m=1}^{M}\beta_m=1$$
$$\beta_m\geq 0\quad \forall m=1,2,...,M$$

که W_{ij} وزن مربوط به هر زوج‌داده بر اساس $\mathbf{k}$ نزدیک‌ترین همسایه آن‌هاست. قابل ذکر است در تابع هزینه بالا، عبارت نخست مطابق ساختار پیشنهادی ۱ سعی در کمینه‌سازی فاصله زوج‌های مشابه دارد حال آن که عبارت دوم برای حفظ

توپولوژی کلی داده‌ها در نظر گرفته شده و بنابراین ارتباطی به مشابه یا نامشابه‌بودن نمونه‌ها ندارد. درواقع این عبارت سعی بر آن دارد که داده‌هایی که در مجاورت یکدیگرند، در فضای جدید نیز تا حد ممکن مجاور یکدیگر باشند. بر این اساس، W_{ij} برای هر زوج‌نمونه به‌صورت زیر محاسبه می‌شود:

$$W_{ij}=\begin{cases}1 & \text{if } \left(\mathbf{x}_i\in N_k(\mathbf{x}_j)\vee \mathbf{x}_j\in N_k(\mathbf{x}_i)\right)\\ 0 & \text{otherwise}\end{cases} \qquad (۱۳)$$

که $N_k(\mathbf{x}_i)$ مجموعه k نزدیک‌ترین همسایه $\mathbf{x}_i$ را نشان می‌دهد. بنابراین ماتریس W یک ماتریس متقارن است که ارتباط میان داده‌ها را نشان می‌دهد. همچنین α ضریبی است که تعیین‌کننده میزان تأثیر جمله جریمه حفظ توپولوژی در مقایسه با جمله اصلی مسئله بهینه‌سازی یعنی کمینه‌سازی متوسط فاصله میان زوج‌های مشابه است. تابع لاگرانژ برای مسئله بهینه‌سازی بالا به‌صورت زیر است:

$$\begin{aligned}L=&\frac{1}{|S|}\sum_{m=1}^{M}\beta_m\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}\left[(\mathbf{K}_m)_{ii}+(\mathbf{K}_m)_{jj}-(2\mathbf{K}_m)_{ij}\right]\\ &+\alpha\frac{1}{2N}\sum_i\sum_j\left[(\mathbf{K}_m)_{ii}+(\mathbf{K}_m)_{jj}-(2\mathbf{K}_m)_{ij}\right]W_{ij}\\ &-\lambda_1\left[\left(\frac{1}{|D|}\sum_{m=1}^{M}\beta_m\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D}\left[(\mathbf{K}_m)_{ii}+(\mathbf{K}_m)_{jj}-(2\mathbf{K}_m)_{ij}\right]\right)-c\right]\\ &-\lambda_2\left[1-\sum_{m=1}^{M}\beta_m\right]-\sum_{m=1}^{M}\eta_m\beta_m\end{aligned} \qquad (۱۴)$$

مسئله بهینه‌سازی بالا را نیز می‌توان به‌صورت زیر بازنویسی کرد که یک مسئله خطی با قیدهای خطی است و به‌کمک روش Simplex قابل حل است.

$$\begin{aligned}&\min_{\boldsymbol{\beta}} \quad (\mathbf{a}+\mathbf{p})^T\boldsymbol{\beta}\\ &\text{s.t.} \quad \mathbf{b}^T\boldsymbol{\beta}\ge c,\ \ \mathbf{1}^T\boldsymbol{\beta}=1,\quad \boldsymbol{\beta}\ge 0\end{aligned} \qquad (۱۵)$$

که در رابطه بالا p به‌صورت زیر تعریف می‌شود:

$$p_{m=}\alpha\frac{1}{2N}\sum_i\sum_j\left[(\mathbf{e}_i-\mathbf{e}_j)^T\mathbf{K}_m(\mathbf{e}_i-\mathbf{e}_j)\right]W_{ij} \qquad (۱۶)$$

۴-۳-ساختار پیشنهادی ۳:

چنانچه ملاحظه شد، در ساختار دوم بیشینه‌سازی متوسط فواصل بین زوج‌های نامشابه، به‌صورت یک قید در مسئله بهینه‌سازی در نظر گرفته شد. در برخی از مراجع از جمله [22]، مسئله بیشینه‌سازی فواصل زوج‌های نامشابه به‌صورت مجموعه‌ای از قیدهای مستقل، به تعداد زوج‌های نامشابه، در نظر گرفته شده است. بنابراین در ساختار سوم فرض شده است که به جای قید متوسط فواصل زوج‌های نامشابه، فاصله بین هر زوج نامشابه به طور مستقل بیشینه شود و بنابراین به‌جای قید $\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D}\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2/|D|\ge c$ از قیدهای $\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2\ge c,\ \forall(\mathbf{x}_i,\mathbf{x}_j)\in D$ استفاده شد.

به‌عبارتی در این ساختار، اثر هر یک از زوج‌های نامشابه به‌صورت مجزا مورد بررسی قرار می‌گیرد. بنابراین:

$$\begin{aligned}&\min\frac{1}{|S|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2\\ &+\alpha\frac{1}{2N}\sum_i\sum_j\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2W_{ij}\\ &s.t.\quad \left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2\ge c,\ \forall(\mathbf{x}_i,\mathbf{x}_j)\in D\\ &\sum_{m=1}^{M}\beta_m=1\\ &\beta_m\ge 0\quad \forall m=1,2,...,M\end{aligned} \qquad (۱۷)$$

در این صورت، تابع لاگرانژ به‌صورت زیر خواهد بود:

$$\begin{aligned}L=&\frac{1}{|S|}\sum_{m=1}^{M}\beta_m\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}\left[(\mathbf{K}_m)_{ii}+(\mathbf{K}_m)_{jj}-(2\mathbf{K}_m)_{ij}\right]\\ &+\alpha\frac{1}{2N}\sum_i\sum_j\left[(\mathbf{K}_m)_{ii}+(\mathbf{K}_m)_{jj}-(2\mathbf{K}_m)_{ij}\right]W_{ij}\\ &-\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D}\rho_{ij}\left[(\mathbf{K}_m)_{ii}+(\mathbf{K}_m)_{jj}-(2\mathbf{K}_m)_{ij}-c\right]\\ &-\lambda_2\left[1-\sum_{m=1}^{M}\beta_m\right]-\sum_{m=1}^{M}\eta_m\beta_m\end{aligned} \qquad (۱۸)$$

که در این رابطه ρ_{ij}ها نیز ضرایب لاگرانژ هستند. با ساده‌سازی تابع لاگرانژ بالا، مسئله بهینه‌سازی حاصله، خطی و مشابه با رابطه (۱۵) خواهد بود با این تفاوت که قید مربوط به بیشینه‌سازی متوسط فاصله میان زوج‌های نامشابه، به بیشینه‌سازی فاصله میان هر یک از زوج‌های نامشابه تبدیل شده است. این مسئله بهینه‌سازی خطی را نیز می‌توان به روش Simplex حل نمود [29]. نکته‌ای که در خصوص این ساختار به نظر می‌رسد این است که برآورده‌سازی هم‌زمان تمام قیود متناظر با زوج داده‌های نامشابه، می‌تواند سبب ایجاد اختلال در فرآیند بهینه‌سازی تابع هزینه شود؛ اما برای مقایسه با سایر روش‌های ارائه‌شده در این مقاله، این ساختار نیز مورد توجه قرار گرفته است که نتایج آن دربعد بررسی می‌شود.

۴-۴-ساختار پیشنهادی ۴:

در مسائل بهینه‌سازی مقید، انتخاب مناسب تابع هدف و قیدهای مربوطه، اهمیت زیادی داشته و تغییر آن‌ها می‌تواند تأثیر زیادی در نتایج نهایی ایجاد کند. در ساختارهای پیشین مورد بحث، بیشینه‌سازی فواصل زوج‌های نامشابه با استفاده از عبارت‌های نامساوی در قالب قیدهایی لحاظ شده بود. از

[DOI: 10.18869/acadpub.jsdp.14.1.53]

دیدگاهی دیگر، مسئله بیشینه‌سازی فواصل زوج‌های نامشابه را می‌توان در تابع هزینه در نظر گرفت. بر این اساس، در ساختار پیشنهادی چهارم، تابع هزینه به‌صورت زیر لحاظ می‌شود:

$$\begin{aligned}&\min_{\boldsymbol{\beta}} \frac{1}{|S|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S}\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2\\&+\alpha\frac{1}{2N}\sum_i\sum_j\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2 W_{ij}\\&-\frac{1}{|D|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D}\left\|\Phi(\mathbf{x}_i)-\Phi(\mathbf{x}_j)\right\|^2\\&s.t.\ \sum_{m=1}^{M}\beta_m=1\\&\beta_m\geq 0\quad \forall m=1,2,...,M\end{aligned} \qquad (۱۹)$$

واضح است بیشینه‌سازی تابع هزینه بالا منوط به بیشترشدن فواصل زوج‌های نامشابه (عبارت سوم در تابع هزینه) است. مسئله بهینه‌سازی بالا را نیز می‌توان به‌صورت زیر بازنویسی کرد:

$$\begin{aligned}&\min_{\boldsymbol{\beta}} \quad (\mathbf{a}+\mathbf{p}-\mathbf{b})^T\boldsymbol{\beta}\\&\text{s.t.} \quad \mathbf{1}^T\boldsymbol{\beta}=1,\quad \boldsymbol{\beta}\geq 0\end{aligned} \qquad (۲۰)$$

مسئله بالا نیز یک مسئله بهینه‌سازی خطی با قیدهای خطی است که به روش Simplex قابل حل است.

پس از معرفی ساختارهای درنظرگرفته‌شده برای یادگیری کرنل مرکب در قالب یک مسئله یادگیری معیار فاصله، در بخش بعد نتایج به‌دست‌آمده از هریک از روش‌های f بالا را مورد بررسی قرار می‌دهیم.

۵-آزمایش‌ها

به‌منظور بررسی کیفیت عملکرد روش‌های یادگیری معیار فاصلهٔ بالا، مسئلهٔ خوشه‌بندی داده‌ها با استفاده از الگوریتم k-means (با فرض معلوم‌بودن k) و به‌عبارت دیگر الگوریتم مبتنی بر کرنل k-means ، مورد بررسی قرار گرفت. در اینجا، ابتدا داده‌های مورد استفاده در این آزمایش‌ها و همچنین معیارهای ارزیابی عملکرد فرآیند خوشه‌بندی معرفی شده و سپس نتایج به‌دست‌آمده از هر روش ارائه و تحلیل می‌شود.

۵-۱-داده‌های استفاده‌شده در آزمایش‌ها

در انجام آزمایش‌ها از دو مجموعه داده‌های مصنوعی و واقعی استفاده شد. مجموعه نخست یعنی داده‌های مصنوعی متعلق به دو دسته متفاوت بوده که با استفاده از توابع گوسی با میانگین‌ها و واریانس‌های متفاوت تولید می‌شود. این داده‌ها با توجه به چگونگی توزیع‌شان به داده‌های XOR شهرت یافته‌اند. مجموعه دوم داده‌ها، داده‌های واقعی جمع‌آوری‌شده در دانشگاه کالیفرنیا در ایروین، UCI[1] است. پایگاه داده UCI شامل مجموعه بزرگی از داده‌ها است که از آن‌ها برای کاربردهای مختلف طبقه‌بندی، خوشه‌بندی و رگرسیون استفاده می‌شود. ما از مجموعه‌های Soybean، Heart، Ionosphere، Wine، Sonar و Iris برای انجام آزمایش‌ها استفاده کرده‌ایم. مشخصات این داده‌ها و همچنین داده‌های تولیدشده XOR در جدول (۱) خلاصه شده است که N تعداد کل نمونه‌ها، n_x بعد بردارهای ویژگی و C تعداد دسته‌ها را نشان می‌دهد.

۵-۲-معیار ارزیابی کیفیت خوشه‌بندی

همان‌طورکه درقبل نیز اشاره شد، کیفیت عملکرد روش‌های پیشنهادی در چارچوب مسئله خوشه‌بندی k-means مورد ارزیابی قرار گرفته است. برای این منظور از شاخص رَند، RI[2]، که مرسوم‌ترین معیار در ارزیابی الگوریتم‌های خوشه‌بندی است، استفاده کرده‌ایم که در ادامه به شرح این معیار خواهیم پرداخت.

(جدول-۱): مشخصات داده‌های XOR و مجموعه داده‌های استفاده شده از پایگاه داده UCI

(Table-1): Main characteristics of the XOR and UCI datasets

Data Set	N	n_x	C
Soybean	47	35	4
Heart	270	13	2
Ionosphere	351	34	2
Wine	178	13	3
Sonar	208	60	2
Iris	150	4	3
XOR	200	2	2

معیار RI

معیار RI یکی از پرکاربردترین معیارهای ارزیابی الگوریتم‌های خوشه‌بندی است و میزان تطابق نتایج حاصل از خوشه‌بندی و دسته‌های واقعی داده‌ها را نشان می‌دهد. فرض کنید N_S تعداد زوج‌هایی باشد که به یک خوشه تعلق داشته و در الگوریتم خوشه‌بندی نیز به یک خوشه مرتبط شوند. همچنین N_D تعداد زوج‌هایی باشد که به دو خوشه متفاوت تعلق داشته و در الگوریتم خوشه‌بندی نیز به دو خوشه متفاوت مرتبط شوند. در این‌صورت RI به‌صورت زیر تعریف می‌شود:

[1] University of California at Irvine
[2] Rand Index

[DOI: 10.18869/acadpub.jsdp.14.1.53]

$$RI = \frac{2(N_S + N_D)}{N(N-1)} \quad (۲۱)$$

زمانی که تعداد خوشه‌ها بیشتر از دو باشد (یعنی k>2)، استفاده از رابطه بالا موجب می‌شود که بر روی زوج‌های متعلق به دسته‌های مختلف در مقایسه با زوج‌های متعلق به یک دسته، تمرکز بیشتری ایجاد شود؛ لذا در چنین شرایطی از رابطه زیر برای محاسبه RI استفاده می‌شود که شانس و یا ضریب یکسانی (۰/۵) را برای هر دو مجموعه در نظر می‌گیرد:

$$RI(C,\hat{C}) = \frac{0.5\times\sum_{i>j}\delta(c_i = c_j \wedge \hat{c}_i = \hat{c}_j)}{\sum_{i>j}\delta(\hat{c}_i = \hat{c}_j)} + \frac{0.5\times\sum_{i>j}\delta(c_i \neq c_j \wedge \hat{c}_i \neq \hat{c}_j)}{\sum_{i>j}\delta(\hat{c}_i \neq \hat{c}_j)} \quad (۲۲)$$

که $\delta(.)$ به این صورت تعریف می‌شود که اگر گزاره داخل آن درست باشد، مقدار δ برابر یک و در غیر این صورت برابر صفر خواهد بود. در رابطه بالا، $\wedge$ برای بیان «و» منطقی استفاده شده است. همچنین $\hat{c}_i$ خوشه‌ای است که در الگوریتم خوشه‌بندی، نمونه $\mathbf{x}_i$ به آن مرتبط شده و c_i بیان‌گر خوشهٔ واقعی این نمونه است. بنابراین مقدار معیار RI بین صفر و یک است، هر قدر این مقدار، به یک نزدیک‌تر شود، بیان‌گر تطابق بیشتر نتایج خوشه‌بندی به‌دست‌آمده با خوشه‌های واقعی است و هر قدر مقدار RI به صفر نزدیک‌تر باشد، بیان‌گر عدم تطابق نتایج با حالت واقعی آن است.

۵-۳-چگونگی انجام آزمایش‌ها و نتایج حاصل

همان‌طورکه درقبل اشاره شد، در اینجا کیفیت عملکرد روش یادگیری معیار فاصله با به‌کارگیری کرنل مرکب مورد بررسی واقع می‌شود. برای این منظور ابتدا هر مجموعه از داده هنجارسازی می‌شود، به‌گونه‌ای که بردارهای ویژگی جدید دارای متوسط صفر و انحراف معیار یک باشند. همان‌طورکه در بخش ۴ در ساختارهای ۱ تا ۴ شرح داده شد، برای یادگیری کرنل مرکب از زوج‌های مشابه و نامشابه استفاده می‌شود. در انجام آزمایش‌ها، از هر مجموعه داده به‌طور تصادفی ده قید که شامل پنج زوج مشابه و پنج زوج نامشابه است استخراج شده و از این زوج‌ها در الگوریتم یادگیری استفاده می‌شود. در الگوریتم‌هایی مانند RCA، روش kernel-β و... که تنها از زوج‌های مشابه استفاده می‌کنند، تمام این ده زوج قید را از میان زوج‌های مشابه انتخاب می‌کنیم. ازآن‌جایی‌که این انتخاب به‌صورت تصادفی صورت می‌گیرد، این فرآیند را بیست بار تکرار می‌کنیم. از طرفی در الگوریتم k-means نیز، نقاط شروع خوشه‌بندی به‌طور تصادفی انتخاب می‌شوند که می‌تواند منجر به نتایج خوشه‌بندی متفاوتی شود؛ ازاین‌رو برای هر مجموعه قیدهای بالا الگوریتم k-means را نیز بیست بار تکرار کرده و متوسط نتایج را گزارش می‌کنیم. برای محاسبه ماتریس W در ساختارهای ۲ تا ۴ از ۶ همسایگی استفاده شده است. با توجه به تکرار آزمایش‌ها، برای نمایش متوسط و انحراف معیار نتایج به‌دست‌آمده از boxplot در نرم‌افزار MATLAB استفاده شده است. برای حل هریک از مسائل بهینه‌سازی خطی به‌دست‌آمده در ساختارهای پیشنهادی از جعبه‌ابزار Mosek[1] در نرم‌افزار MATLAB استفاده شده است.

برای بررسی بیشتر، ساختارهای پیشنهادی را با روش پایه مبتنی بر سنجش فاصله اقلیدسی و برخی از ساختارهای مهم یادگیری معیار فاصله اعم از مبتنی بر کرنل و یا غیر کرنلی که در آن‌ها نیز تاحدودی از زوج‌های مشابه و یا زوج‌های نامشابه در فرآیند یادگیری استفاده می‌شود، مورد مقایسه و ارزیابی قرار داده‌ایم. برای این منظور از روش RCA که در آن به جهت‌های نامرتبط در فضای ورودی وزن کوچک‌تری اختصاص می‌دهد، استفاده شده است. در این الگوریتم تنها از زوج‌های مشابه برای آموزش استفاده می‌شود. برخلاف RCA، الگوریتم پیشنهادشده توسط Xiang در [11] از زوج‌های نامشابه نیز علاوه‌بر زوج‌های مشابه برای آموزش استفاده می‌کند؛ اما هر دو روش RCA و Xiang روش‌های غیر کرنلی هستند پس برای مقایسه بیشتر علاوه‌بر این دو، از روش kernel-β و روش‌های مطرح در [26] و [27] نیز استفاده شده است. در روش kernel-β هدف مقیاس بندی بهینه ماتریس‌های ویژه به‌دست‌آمده از اعمال تحلیل مقادیر ویژه بر روی ماتریس کرنل است. در این روش از یک کرنل گوسی با $\sigma^2 = (\theta / N(N-1))\sum_{i,j=1}^{N}\left\|\phi(\mathbf{x}_i)-\phi(\mathbf{x}_j)\right\|^2$ که در آن θ=5 انتخاب شده‌است، استفاده می‌شود. در روش [25] نیز به ماتریس کرنل مرکب، تحلیل مقادیر ویژه اعمال شده و راستاهایی که ۹۹ درصد تغییرات داده‌ها را شامل شوند، محفوظ نگاه داشته می‌شوند. حال هدف بهینه‌سازی وزن ماتریس‌های ویژه در نظر گرفته شده است. در روش مورد بحث، کرنل مرکب حاصل جمع بدون وزن مجموعه‌ای از کرنل‌های پایه در نظر گرفته شده است. کرنل مرکب نهایی از جمع وزن‌دار ماتریس‌های ویژه‌ی حاصل از اعمال تحلیل مقادیر ویژه بدست می‌آید. در [27] نیز در یک مسئله

[1] http://www.mosek.com

بهینه‌سازی، وزن کرنل‌ها در یک ترکیب خطی به‌گونه‌ای تعیین می‌شود که فاصله زوج‌های مشابه کمینه شود.

بنابراین الگوریتم‌های مقایسه‌شده در این پژوهش عبارتند از:

۱)الگوریتم k-means بدون یادگیری معیار فاصله با نام فاصله اقلیدسی

۲)الگوریتم k-means با استفاده از روش یادگیری معیار فاصله RCA در [۱۷]با نام RCA

۳)الگوریتم k-means با استفاده از روش Xiang در [11] با نام Xiang

۴)الگوریتم مبتنی بر کرنل k-means با استفاده از کرنل به‌دست‌آمده از روش kernel-β در [20] با نام kernel-β

۵)الگوریتم مبتنی بر کرنل k-means با استفاده از روش یادگیری کرنل مرکب در [26] با نام unweighted-S

۶)الگوریتم مبتنی بر کرنل k-means با استفاده از روش یادگیری کرنل مرکب در [27] با نام DMKL

۷)الگوریتم مبتنی بر کرنل k-means با استفاده از ساختار پیشنهادی یک با نام ساختار یک

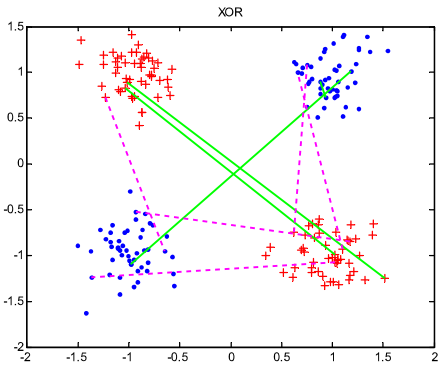
۸)الگوریتم مبتنی بر کرنل k-means با استفاده از ساختار پیشنهادی دو با نام ساختار دو

۹)الگوریتم مبتنی بر کرنل k-means با استفاده از ساختار پیشنهادی سه با نام ساختار سه

۱۰) الگوریتم مبتنی بر کرنل k-means با استفاده از ساختار پیشنهادی چهار با نام ساختار چهار

در انجام آزمایش‌ها از کرنل گوسی و کرنل چندجمله‌ای برای ساخت کرنل‌های پایه استفاده شده است. برای تعیین پارامتر کرنل گوسی، ابتدا پراکندگی داده‌ها از رابطه $\sigma^2 = (1/N(N-1))\sum_{i,j=1}^{N}\left\|\phi(\mathbf{x}_i)-\phi(\mathbf{x}_j)\right\|^2$ محاسبه شده و سپس پارامتر کرنل گوسی به‌صورت $\{$۰/۰۵×σ، ۰/۱×σ، σ، ۵×σ، ۱۰×σ$\}$ محاسبه می‌شود. همچنین درجه کرنل چندجمله‌ای نیز از مجموعه $\{$۲، ۳، ۴، ۵، ۶، ۷، ۸، ۹ و ۱۰$\}$ انتخاب شده است. انتخاب مقادیر پارامتر کرنل‌ها به‌صورت بالا برای ایجاد پوششی مناسب بر روی محدوده پارامتر کرنل‌ها برای استفاده برای انواع داده‌ها است.

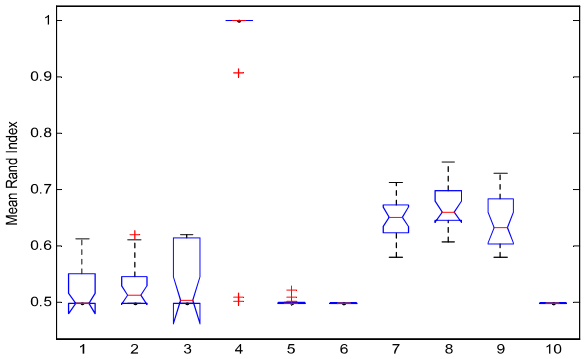
در شکل (۲)، داده XOR و یک مجموعه از زوج‌های نامشابه با خط‌چین و زوج‌های مشابه با خط تیره نمایش داده شده است.



(شکل-۲): نمایش داده XOR با مشخص سازی مجموعه زوج‌های نامشابه (خط چین) و زوج‌های مشابه (خط تیره)

(Figure-2): Original XOR dataset with the pairwise constraints: dissimilarity pairs (dashed pink line) and similarity pairs (solid green line)

همان‌طور که دیده می‌شود توزیع داده XOR به‌گونه‌ای است که در الگوریتم خوشه‌بندی با فاصله اقلیدسی درصد زیادی از داده‌ها به‌اشتباه خوشه‌بندی می‌شوند. نمودار نتایج خوشه‌بندی داده XOR توسط روش‌های مختلف در شکل (۳) آورده شده است.



(شکل-۳): نتایج خوشه‌بندی الگوریتم‌های مختلف بر روی داده XOR. ۱) فاصله اقلیدسی، ۲) RCA [17]، ۳) Xiang [11]، ۴) kernel-β [20]، ۵) unweighted-S [26]، ۶) DMKL [27]، ۷) ساختار یک، ۸) ساختار دو، ۹) ساختار سه،۱۰) ساختار چهار

(Figure-3): Clustering results of different algorithms for XOR dataset. 1) Euclidean, 2) RCA [17], 3) Xiang [11], 4) kernel-β [20], 5) unweighted-*S* [26], 6) DMKL [27], 7) structure 1, 8) structure 2, 9) structure 3, 10) structure 4

از نتایج به‌دست‌آمده برای خوشه‌بندی داده XOR می‌توان دید که کرنل استفاده شده در روش kernel-β بهترین نمایش از داده XOR در فضای جدید را ایجاد می‌کند چرا که نتایج به‌دست‌آمده بیان‌گر خوشه‌بندی دقیق این روش است. همان‌طورکه درقبل نیز ذکر شد، در روش kernel-β از یک کرنل گوسی استفاده شده است که با در نظر گرفتن $\theta=5$ برای تعیین پارامتر کرنل گوسی، به نظر می‌رسد کرنل حاصل بهترین نمایش از داده‌ها در فضای ویژگی را سبب شده است؛

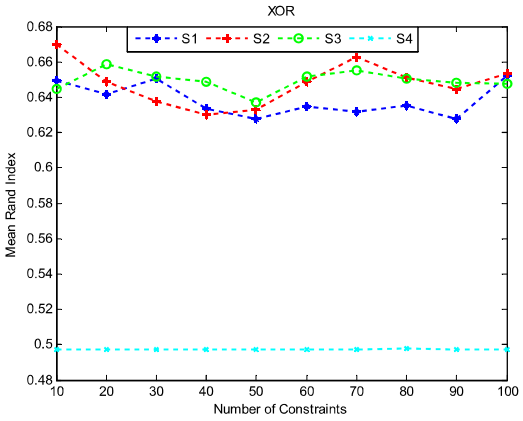
[DOI: 10.18869/acadpub.jsdp.14.1.53]

اما دلیلی برای انتخاب $\theta=5$ ذکر نشده است و برای سایر مقادیر پارامتر θ نتایج ضعیف‌تری حاصل می‌شود. بنابراین انتظار می‌رود استفاده از این کرنل گوسی خاص برای هر مجموعه داده‌ای نتواند منجر به بهترین نتایج شود و بر همین اساس هدف روش‌های یادگیری کرنل مرکب مستقل‌سازی چنین الگوریتم‌هایی در برابر انتخاب یک کرنل ثابت است؛ اما در مجموعه روش‌هایی که کرنل مرکب در آنها استفاده شده است، می‌توان برتری نتایج به‌دست‌آمده از ساختارهای پیشنهادشده در این مقاله را در مقایسه با سایر روش‌ها ملاحظه کرد. در مجموعه ساختارهای پیشنهادشده در این مقاله، ساختار چهارم بدترین جواب را برای خوشه‌بندی داده XOR داشته است که به توزیع داده‌ها در XOR بر می‌گردد. همان‌طورکه در شکل (۲) نشان داده شده است، توزیع داده‌های هر دسته در مجموعه XOR به‌گونه‌ای است که متوسط فاصله زوج‌های نامشابه بیشتر از متوسط فاصله میان زوج‌های مشابه می‌شود و به‌عبارتی اهمیت زوج‌های مشابه در فشرده‌سازی داده‌های هر دسته کم‌رنگ می‌شود. بنابراین الگوریتم، بیش‌تر به سمت بیشینه‌سازی فاصله داده‌های متعلق به دسته‌های مختلف متمایل شده و به‌عبارتی خوشه‌ها در این ساختار از یکدیگر دورتر می‌شوند که با توجه به نوع پراکندگی داده‌های XOR انتظار چنین نتایجی می‌رود.

به‌منظور بررسی میزان وابستگی ساختارهای پیشنهادی به تعداد قیدهای مشابه و نامشابه، آزمایش‌ها برای تعداد قیدهای متفاوت تکرار شد. در شکل (۴)، متوسط RI به‌دست‌آمده از چهار ساختار پیشنهاد شده و به‌ازای تعداد قیدهای متفاوت برای داده XOR نشان داده شده است. همان‌طورکه دیده می‌شود، در اینجا نیز نتایج مربوط به ساختار چهارم ضعیف‌ترین نتایج است.

نتایج به‌دست‌آمده از الگوریتم‌های خوشه‌بندی مختلف برای داده‌های پایگاه داده UCI نیز در شکل (۵) آورده شده است. از نتایج به‌دست‌آمده بر روی داده‌های مختلف این پایگاه داده می‌توان چنین استنباط کرد که روش‌های مبتنی بر کرنل، به‌دلیل نگاشت داده‌ها به فضایی با قدرت تفکیک‌پذیری بیش‌تر، درمجموع عملکرد بهتری در مقایسه با روش‌های غیر کرنلی دارند.

همچنین می‌توان دید که ساختارهای پیشنهادشده در این مقاله در حالت کلی دارای نتایج خوشه‌بندی بهتری و یا واریانس کمتری در مقایسه با سایر روش‌های مبتنی بر کرنل داشته است. نکته قابل ذکر در اینجا، نتایج به‌دست‌آمده از ساختار چهارم برای خوشه‌بندی این داده‌هاست که نتایج متفاوتی را با آنچه در خوشه‌بندی داده XOR مشاهده شد، نشان می‌دهد و این نتیجه تأییدی بر تأثیر ساختار داده‌ها XOR در نتیجه به‌دست‌آمده از مسئله بهینه‌سازی در ساختار چهارم است. متوسط RI برای حالتی که تعداد قیدهای متفاوتی برای خوشه‌بندی داده‌های UCI در نظر گرفته شود نیز در شکل (۶) آورده شده است.



(شکل-۴): متوسط RI روش‌های مختلف بر روی داده XOR. به‌ازای تعداد قیدهای متفاوت. S1) ساختار یک، S2) ساختار دو، S3) ساختار سه، S4) ساختار چهار

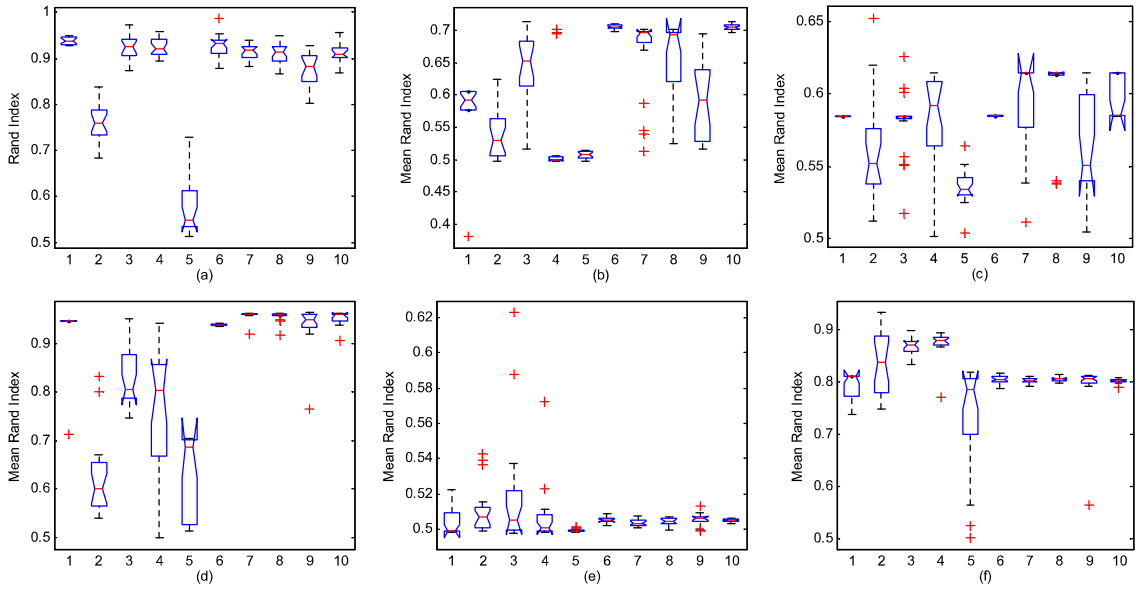
(Figure-4): Average RI of different algorithms for XOR dataset when the different number of constraints is used. S1) structure 1, S2) structure 2, S3) structure 3, S4) structure 4

همان‌طورکه از نتایج مشخص است، ساختارهای پیشنهادشدهٔ نخست، دوم و چهارم تا حدودی نسبت به تعداد زوج قیدهای در نظر گرفته شدهٔ پایدار است؛ اما در ارتباط با ساختار سوم می‌توان دید که با افزایش زوج‌قیدها، نتایج تضعیف می‌شود؛ چون در این ساختار دورسازی داده‌های مربوط به دسته‌های مختلف با بیشینه‌سازی فاصله میان هر زوج نامشابه صورت می‌گیرد که با افزایش تعداد زوج‌های نامشابه، تنظیم وزن کرنل‌ها به‌گونه‌ای که شرط متناظر با تک‌تک قیدهای نامشابه را تأمین کند، بسیار پیچیده شده و موجب سردرگمی الگوریتم می‌شود.

برای بررسی تأثیر مقادیر پارامترهای کرنل‌ها بر روی عملکرد ساختارهای پیشنهادی، از چهارده کرنل جدید که شامل کرنل گوسی با پارامترهای {۰/۱، ۱، ۱۰، ۱۰۰ و ۱۰۰۰} است و همچنین کرنل چندجمله‌ای با درجه {۲، ۳، ۴، ۵، ۶، ۷، ۸، ۹ و ۱۰} استفاده شد. نتایج به‌دست‌آمده از این آزمایش بر روی مجموعه داده‌های منتخب UCI در شکل (۷) گزارش شده است. نتایج نشان می‌دهد که ساختارهای پیشنهادی نسبت به انتخاب پارامترهای کرنل مقاوم بوده و انتخاب کرنل‌های متفاوت، تأثیر ناچیزی در عملکرد این ساختارها دارد.

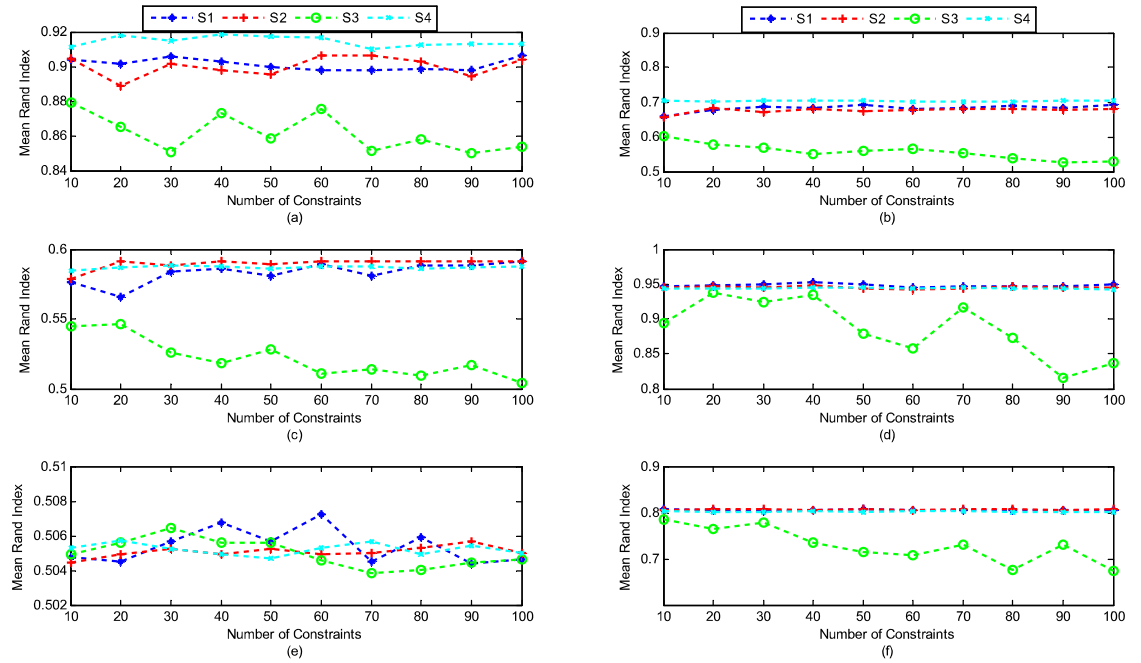
[DOI: 10.18869/acadpub.jsdp.14.1.53]

همچنین به‌منظور بررسی دقیق‌تر عملکرد کرنل مرکب در مقایسه با استفاده از یک کرنل به‌تنهایی، در شکل (۸) نتایج حاصل از الگوریتم مبتنی بر کرنل k-means در شرایطی که تنها از یک کرنل استفاده شود نیز مورد بررسی قرار گرفت و آزمایش‌ها برای استفاده از هر یک از این چهارده کرنل بالا بر روی مجموعه داده‌های منتخب **UCI** گزارش شده است. این نتایج به‌وضوح نشان‌گر برتری عملکرد کرنل مرکب در مقایسه با استفاده از یک کرنل به‌تنهایی است.



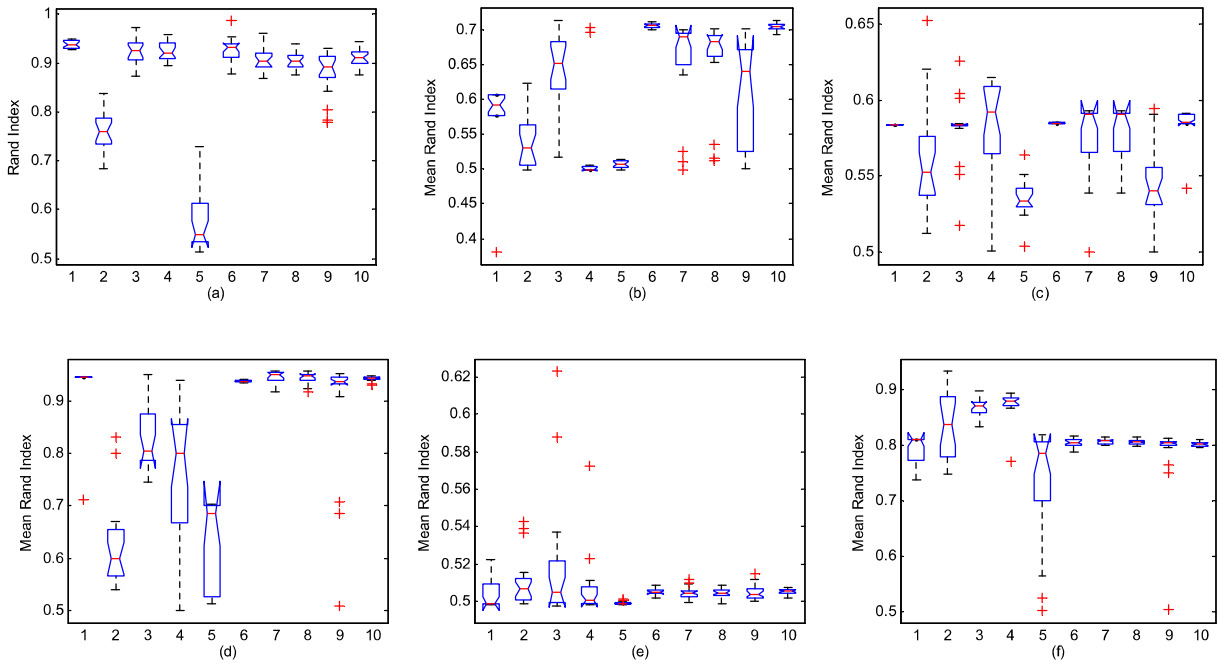
(شکل-۵): نتایج خوشه‌بندی الگوریتم‌های مختلف بر روی داده‌های پایگاه داده UCI. ۱) فاصله اقلیدسی، ۲) RCA [17]، ۳) Xiang [11]، ۴) kernel-β [20]، ۵) unweighted-S [26]، ۶) DMKL [27]، ۷) ساختار یک، ۸) ساختار دو، ۹) ساختار سه،۱۰) ساختار چهار. (a) Soybean، (b) Heart، (c) Ionosphere، (d) Wine، (e) Sonar، (f) Iris

(Figure-5): Clustering results of different algorithms for UCI datasets. 1) Euclidean, 2) RCA [17], 3) Xiang [11], 4) kernel-β [20], 5) unweighted-*S* [26], 6) DMKL [27], 7) structure 1, 8) structure 2, 9) structure 3, 10) structure 4. (a) Soybean, (b) Heart, (c) Ionosphere, (d) Wine, (e) Sonar, (f) Iris



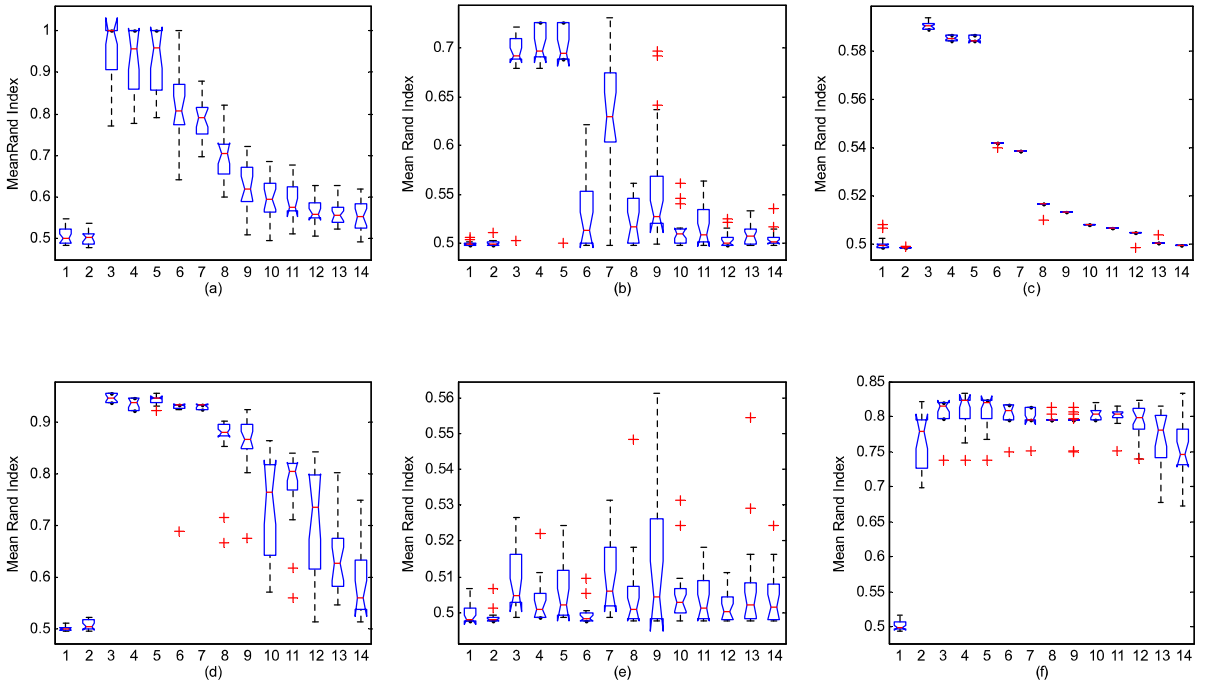
(شکل-۶): متوسط RI روش‌های مختلف بر روی داده UCI. به ازای تعداد قیدهای متفاوت. S1) ساختار یک، S2) ساختار دو، S3) ساختار سه، S4) ساختار چهار. (a) Soybean، (b) Heart، (c) Ionosphere، (d) Wine، (e) Sonar، (f) Iris

(Figure-6): Average RI of different algorithms for UCI dataset when the different number of constraints is used. S1) structure 1, S2) structure 2, S3) structure 3, S4) structure 4. (a) Soybean, (b) Heart, (c) Ionosphere, (d) Wine, (e) Sonar, (f) Iris.



(شکل-۷): نتایج خوشه‌بندی الگوریتم‌های مختلف بر روی داده‌های پایگاه داده UCI. ۱) فاصله اقلیدسی، ۲) RCA [17]، ۳) Xiang [11]، ۴) kernel-β [20]، ۵) unweighted-S [26]، ۶) DMKL [27]، ۷) ساختار یک، ۸)ساختار دو، ۹) ساختار سه،۱۰) ساختار چهار. (a) Soybean، (b)Heart، (c) Ionosphere، (d) Wine، (e)Sonar، (f) Iris

(Figure-7): Clustering results of different algorithms for UCI datasets. 1) Euclidean, 2) RCA [17], 3) Xiang [11], 4) kernel-β [20], 5) unweighted-*S* [26], 6) DMKL [27], 7) structure 1, 8) structure 2, 9) structure 3, 10) structure 4. (a) Soybean, (b) Heart, (c) Ionosphere, (d) Wine, (e) Sonar, (f) Iris



(شکل-۸): نتایج خوشه‌بندی الگوریتم مبتنی بر کرنل k-means با استفاده از یک کرنل از میان مجموعه چهارده کرنل ذکر شده بر روی داده‌های پایگاه داده UCI. (a) Soybean، (b)Heart، (c)Ionosphere، (d) Wine، (e)Sonar، (f) Iris

(Figure-8): Clustering results of kernel k-means using a single kernel among the mentioned fourteen kernels for UCI datasets. (a) Soybean, (b) Heart, (c) Ionosphere, (d) Wine, (e) Sonar, (f) Iris

[DOI: 10.18869/acadpub.jsdp.14.1.53]

۶-نتیجه‌گیری و پیشنهادها

در این مقاله ساختارهای جدیدی برای یادگیری کرنل مرکب با استفاده از مفاهیم اساسی در مسئله یادگیری معیار فاصله مطرح شد که در آنها از زوج‌های مشابه و نامشابه در یک مسئله بهینه‌سازی خطی با قیود خطی برای آموزش وزن کرنل‌ها استفاده می‌شود. ساختارهای پیشنهادشده در این مقاله بر مبنای توابع هدف (هزینه) متفاوتی استوار است که هر کدام در جایگاه خود اهداف خاصی را در یادگیری معیار فاصله دنبال می‌کند. ساختارهای پیشنهادی با برخی از معروف‌ترین الگوریتم‌های یادگیری معیار فاصله اعم از مبتنی بر کرنل و یا غیر کرنلی و همچنین با وجود تنها زوج‌های مشابه و یا زوج‌های نامشابه مورد مقایسه و ارزیابی قرار گرفت. در حالت کلی روش‌های مبتنی بر کرنل، عملکرد بهتری در مقایسه با روش‌های بدون کرنل دارند. روش‌های پیشنهادشده به‌جز ساختار سوم تا حدودی نسبت به تعداد زوج‌قیدهای استفاده شده، پایدار است. در ساختار سوم چون هر زوج نامشابه یک قید به مسئله بهینه‌سازی اضافه می‌کند، با افزایش تعداد این زوج‌ها، ساختار پیشنهادشده ایجاد سردرگمی کرده و توانایی خود در خوشه‌بندی دقیق را در تعداد زوج‌های زیاد از دست می‌دهد. با توجه به نتایج به‌دست‌آمده می‌توان گفت به غیر از موارد خاصی مانند ساختار سوم برای داده‌های Heart، واریانس نتایج در روش‌های پیشنهادی در مقایسه با سایر ساختارهای مورد بررسی کم است که این نشان‌دهندهٔ ناچیزبودن وابستگی ساختارهای پیشنهادی به زوج‌های مشابه و نامشابه انتخاب شده است. یکی دیگر از مزایای ساختارهای پیشنهادی، مقاوم‌بودن این ساختارها در برابر انتخاب پارامترهای کرنل‌های پایه است.

۷-مراجع 7-Refrences

[۱] ا. سجودی شیجانی، " یادگیری به موقع معیار فاصله درمحیطهای غیر ایستا" فصلنامه صنایع الکترونیک، دوره ۶،شماره ۲، تابستان ۱۳۹۴.

[1] O. Sojoodi, "Just-in-time adaptive distance metric learning in nonstationary environments", *Journal of Electronics industries*, Vol. 6, No. 2, 2015.

[2] R. O. Duda, P. E. Hart and D. G.Stork,*Pattern Classification*, 2nd edition, New York: Wiley, 2001.

[3] D. A. Forsyth and J Ponce, *Computer Vision: A Modern Approach*, Englewood Cliffs, NJ: Prentice Hall, 2003.

[4] L. Yang, R. Jin, "Distance metric learning: a comprehensive survey", Technical Report, Michigan State University, 2006.

[5] L. Yang, "An Overview of Distance Metric Learning", Technical Report, Michigan State University, 2007.

[6] C. Domeniconi, J. Peng, D. Gunopulos, "Locally adaptive metric nearest neighbor classification", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 9, pp. 1281-1285, 2002.

[7] C. Domeniconi, D. Gunopulos, "Large margin nearest neighbour classifiers", *IEEE Transactions on Neural Networks*, Vol.16, No.4, pp.899-909, 2005.

[8] D.Wang, J. S. Lim, M.-M. Han, B.-W. Lee, "Learning similarity for semantic images classification", *Neurocomputing*, Vol. 67, pp. 363-368, 2005.

[9] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," In *Advancesin Neural Information Processing Systems (NIPS)*,2005.

[10] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighborclassification,"In*Advancesin Neural Information Processing Systems (NIPS)*,2006.

[11] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information", In *Advancesin Neural Information Processing Systems (NIPS)*, 2003.

[12] S.Shalev-Shwartz, Y. Singer, and A. Y.Ng, "Online and Batch Learning of Pseudo-Metrics", In Proc.*Int. Conf. on Machine Learning (ICML)*, 2004.

[13] N. Kumar, K. Kummamuru, "Semi-supervised clustering with metric learning using relative comparisons", *IEEE Transactions on Knowledge and Data Engineering*, Vol.20, No. 4, 496-503, 2007.

[14] H. Chang, D.-Y. Yeung, "Locally linear metric adaptation with application to semi-supervised

[DOI: 10.18869/acadpub.jsdp.14.1.53]

clustering and image retrieval", *Pattern Recogition*,Vol. 39, No. 7, pp.1253-1264, 2006.

[15] M. Schultz, T. Joachims, "Learning a distance metric from relative comparisons", In *Advancesin Neural Information Processing Systems (NIPS)*,2004.

[16] S. Xiang, F. Nie, C. Zhang, "Learning a Mahalanobis distance metric for data clustering and classification", *Pattern Recognition*, Vol. 41, No. 12, pp. 3600-3612, 2008.

[17] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations,"InProc.*Int. Conf. on Machine Learning (ICML)*, Washington, DC, 2003.

[18] H. Chang, D.-Y. Yeung,, "Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints", *Pattern Recognition*, Vol. 39, No.5, pp. 1007-1010, 2006.

[19] James T. Kwok and IvorW. Tsang, "Learning with idealized kernels", In Proc.*Int. Conf. on Machine Learning (ICML)*, 2003.

[20] D. Y. Yeung and H.Chang, "A kernel approach for semisupervised metric learning", *IEEE Transactions on Neural Networks*,vol. 18, no. 1, pp. 141-149, Jan. 2007.

[21] S. C. H. Hoi, R. Jin, M. R. Lyu, "Learning nonparametric kernel matrices from pairwise constraints", In Proc.*Int. Conf. on Machine Learning (ICML)*,New York, USA, 2007.

[22] M. S. Baghshah and S. B. Shouraki, "Kernel-based metric learning for semi-supervised clustering", *Neurocomputing*, Vol. 73, No. 7-9, pp. 1352-1361, 2010.

[23] M. Gönen, E. Alpaydın, "Multiple kernel learning algorithms", *Journal of Machine Learning Research*, Vol. 12, pp. 2211–2268, 2011.

[24] J. Wang, H. Do, A. Woznica, and A. Kalousis. Metric learning with multiple kernels. In *Advances in Neural Information Processing Systems(NIPS)*, MIT Press, 2011.

[25] F. Yan, K. Mikolajczyk and J. Kittler, "Multiple Kernel Learning via Distance Metric Learning for Interactive Image Retrieval",InProc. *Multiple Classifier Systems*, pp. 147-156, 2011.

[26] T. Zare, M. T. Sadeghi and H. R. Abutalebi, "Semi-supervised Metric Learning Using Composite Kernels", In Proc.*6th Int. Telecommunication symposium (IST)*, 2012.

[27] T.Zare, M. T. Sadeghi, H. R. Abutalebi, "A Novel Multiple Kernel Learning Approach for Semi-Supervised Clustering", In Proc.*the 8th Iranian Conference on Machine Vision and Image Processing* (MVIP), 2013.

[28] T. Zare, M. T. Sadeghi and H. R. Abutalebi, "A Comparative Study of Multiple Kernel Learning Approaches for SVM Classification", In Proc.*6th Int. Telecommunication symposium (IST)*, 2014.

[29] D. Gale, "Linear programming and the simplex method", Notices of the AMS, Vol. 54, No. 3, pp.364–369, 2007.

طاهره زارع بیدکی تحصیلات خود را در مقاطع کارشناسی برق- الکترونیک و کارشناسی ارشد برق- مخابرات به‌ترتیب در سال‌های ۱۳۸۵ و ۱۳۸۸ در دانشگاه یزد به پایان رسانده و در حال حاضر نیز دانشجوی مقطع دکترای مخابرات سیستم در دانشگاه یزد است. زمینه‌های پژوهشی مورد علاقهٔ وی پردازش تصویر، بینایی ماشین و شناسایی آماری الگو است.

نشانی رایانامک ایشان عبارت است از:

t.zare@stu.yazd.ac.ir

محمدتقی صادقی تحصیلات خود را در مقاطع کارشناسی برق- الکترونیک و کارشناسی ارشد برق- مخابرات به‌ترتیب در سال‌های ۱۳۷۰ و ۱۳۷۳ در دانشگاه‌های صنعتی شریف و تربیت مدرس تهران به پایان رسانید. نامبرده از سال ۱۳۷۴ الی ۱۳۷۸ در دانشکده مهندسی برق دانشگاه یزد مشغول به کار بود. پس از آن دوره دکترای مهندسی برق - مخابرات را در دانشگاه ساری انگلستان آغاز کرده و در سال ۱۳۸۱ موفق به اخذ درجه دکتری از آن دانشگاه شد؛ سپس به مدت دو سال در مرکز تحقیقات بینائی ماشین و پردازش سیگنال همان دانشگاه، در زمینهٔ سامانه‌های بازشناسی چهره پژوهش کرده

[DOI: 10.18869/acadpub.jsdp.14.1.53]

و از آن پس تاکنون استادیار دانشکده مهندسی برق و کامپیوتر دانشگاه یزد است. همکاری پژوهشی نامبرده با مرکز تحقیقات بینائی ماشین و پردازش سیگنال دانشگاه ساری همچنان ادامه دارد. زمینه‌های پژوهشی مورد علاقهٔ وی بازشناسی آماری الگو، پردازش تصویر و بینائی ماشین است.
نشانی رایانامک ایشان عبارت است از:
m.sadeghi@yazd.ac.ir

حمیدرضا ابوطالبی دوره کارشناسی و کارشناسی ارشد را به ترتیب در سال‌های ۱۳۷۴ و ۱۳۷۷ در رشته مهندسی برق (مخابرات) در دانشگاه صنعتی شریف گذرانده و مدرک دکترای خود را در سال ۱۳۸۲ در رشته مهندسی برق (مخابرات) از دانشگاه صنعتی امیرکبیر اخذ کرد.
وی در جریان رساله دکترای خویش، به‌مدت یک سال در دوره فرصت مطالعاتی در دانشگاه واترلو کانادا به سر برد. دکتر ابوطالبی در سال ۱۳۸۲ به دانشکده مهندسی برق دانشگاه یزد پیوست و هم‌اکنون به‌عنوان استاد گروه مهندسی مخابرات در این دانشکده مشغول به فعالیت است. وی همچنین در سال ۹۰-۱۳۸۹ یک دوره فرصت مطالعاتی را در مرکز تحقیقاتی Idiap در سوئیس سپری کرد. زمینه‌های علمی مورد علاقهٔ وی پردازش آرایه‌ای سیگنال گفتار، بهسازی گفتار، مکان‌یابی گوینده، فیلترهای وفقی، و تحلیل زمان- فرکانس است.
نشانی رایانامک ایشان عبارت است از:
habutalebi@yazd.ac.ir

[DOI: 10.18869/acadpub.jsdp.14.1.53]