یادگیری متریک نیمه نظارتی در فضای لایه‌ای با بهره‌گیری دقیق‌تر از دانش پیشین

زهره کریمی، سعید شیری قیداری و روح‌اله رمضانی

چکیده: یادگیری متریک نیمه‌نظارتی مبتنی بر منیفلد در سال‌های اخیر بسیار مورد توجه واقع شده است. این رویکردها، منظم‌سازی مبتنی بر فرض هموار‌بودن داده‌ها روی منیفلد را اعمال می‌کنند، هرچند در معرض دو چالش قرار دارند: ۱) شباهت بین دسته‌های مختلف، تقاطع منیفلدها با یکدیگر را ایجاد می‌کند که با فرض هموار‌بودن برچسب در این نواحی در تناقض است. ۲) دسته‌بند 1NN که برای تعیین برچسب داده‌ها در مسایل یادگیری متریک اعمال می‌شود با وجود تعداد کم داده‌های برچسب‌دار دقت مناسب را ندارد. در این مقاله روشی برای یادگیری متریک نیمه‌نظارتی با فرض قرارگیری داده‌ها در فضای لایه‌ای ارائه شده که در آن از دانش پیشین موجود که همان فرض هموار‌بودن داده‌ها روی هر منیفلد است به صورت دقیق‌تر بهره‌برداری شده است. در مرحله یادگیری متریک، فرض هموار‌بودن در نواحی تقاطع اعمال نشده و در مرحله دسته‌بندی، داده‌های برچسب‌دار در نقاط داخلی منیفلدها بر اساس فرض هموار‌بودن توسعه داده شده است. تفکیک نقاط تقاطع منیفلدها از سایر نقاط بر مبنای رفتار متمایز لاپلاسین تابع هموار روی هر منیفلد در نقاط داخلی نسبت به سایر نقاط صورت می‌گیرد. آزمایش‌ها نشان‌دهنده دقت خوب روش پیشنهادی نسبت به روش‌های مشابه است.

کلیدواژه: **یادگیری متریک نیمه‌نظارتی، فضای لایه‌ای، لاپلاسین، فرض هموار‌بودن.**

۱- مقدمه

مطالعات سال‌های اخیر نشان داده که یادگیری متریک، نقشی اساسی در یادگیری ماشین از جمله دسته‌بندی، خوشه‌بندی و بازیابی اطلاعات دارد [۱] تا [۳]. از آنجا که هر وظیفه[1]، فضای معنایی خود را جهت تعریف فاصله‌ها دارد، فاصله اقلیدسی اغلب نمی‌تواند این فضا را به درستی نشان دهد و انتخاب مناسب متریک تأثیر زیادی در نتیجه به دست آمده دارد. هدف یادگیری متریک، یادگیری یک تابع فاصله با مقدار حقیقی است که نقاط مشابه را به یکدیگر نزدیک نگه داشته و نقاط نامشابه را از یکدیگر دور می‌کند. اغلب روش‌های یادگیری متریک یک تابع فاصله متریک را از دانشی که به شکل محدودیت‌های دوتایی یا سه‌تایی در دسترس هستند یاد می‌گیرند. این محدودیت‌ها از داده‌های برچسب‌دار موجود استنتاج می‌شوند. از آنجا که برچسب‌گذاری داده‌ها هزینه زیادی داشته اما داده‌های بدون برچسب به میزان زیادی در دسترس هستند در سال‌های اخیر رویکردهای یادگیری متریک نیمه‌نظارتی مطرح شده است. تمرکز ما بر وظیفه دسته‌بندی و بر رویکردهایی است که فرض قرارگیری داده روی یک یا چند منیفلد را دارند. این رویکردها فرض هموار‌بودن روی منیفلد را به مسأله بهینه‌سازی یادگیری متریک اعمال می‌کنند که بیانگر این است که داده‌های نزدیک به یکدیگر با احتمال بالایی برچسب یکسان دارند. هرچند این روش‌ها در مورد داده‌هایی که در فضای لایه‌ای قرار دارند در معرض چالش‌هایی قرار دارد. فضای لایه‌ای فضایی است که در آن داده‌ها روی چند منیفلد قرار می‌گیرند که ممکن است با یکدیگر اشتراک داشته و ابعاد ذاتی آنها متفاوت باشد. این چالش ها عبارتند از:

- شباهت داده‌های نزدیک به یکدیگر که در روش‌های موجود اعمال شده است در نواحی تقاطع منیفلدها لزوماً برقرار نیست.
- دسته‌بندی که از متریک یاد گرفته شده استفاده کرده و جهت تعیین برچسب داده‌ها به کار می‌رود به خاطر تعداد کم داده‌های برچسب‌دار دقت لازم را ندارد.

در این مقاله، یک رویکرد یادگیری متریک نیمه‌نظارتی برای دسته‌بندی داده‌هایی که در فضای لایه‌ای قرار دارد پیشنهاد شده که از دانش پیشین موجود که همان فرض هموار‌بودن در فضای لایه‌ای است به صورت دقیق‌تر در موارد ذیل بهره‌برداری می‌کند:

- عدم اعمال فرض هموار‌بودن در نواحی تقاطع منیفلدها
- انتشار داده‌های برچسب‌دار به نزدیک‌ترین نقاط به این داده‌ها در نواحی داخلی منیفلدها، بر اساس فرض هموار‌بودن روی منیفلد

موارد مذکور بر مبنای تفکیک نقاط داخلی منیفلد از سایر نقاط است که با توجه به تئوری موجود در این زمینه انجام می‌شود.

در ادامه، ابتدا در بخش ۲ پژوهش‌های مرتبط بررسی گردیده و سپس در بخش ۳ جزئیات روش پیشنهادی شرح داده شده است. در بخش ۴ نتایج آزمایش‌های انجام‌شده و در بخش ۵ نتایج و کارهای پیشنهادی آمده است.

۲- پژوهش‌های مرتبط

در سال‌های اخیر روش‌های یادگیری متریک زیادی پیشنهاد شده [۱]، [۳] و [۴] که این روش‌ها در سه گروه بدون ناظر، نظارتی و نیمه‌نظارتی قرار می‌گیرد.

رویکردهای یادگیری متریک بدون ناظر سعی در حفظ توپولوژی و یا ساختار هندسی داده در حد امکان دارند. این رویکردها به فرم مسأله کاهش بعد و/ یا یادگیری منیفلد مطرح می‌شوند. هدف اکثر رویکردهای یادگیری منیفلد، یادگیری بازنمایی بعد پایین داده است به نحوی که فاصله بین داده‌ها حفظ شود. معروف‌ترین رویکردهای اولیه در این زمینه

این مقاله در تاریخ ۵ مهر ۱۳۹۶ دریافت و در تاریخ ۳۰ اردیبهشت ماه ۱۳۹۷ بازنگری شد.

زهره کریمی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران، (email: z_karimi@aut.ac.ir).

سعید شیری قیداری (نویسنده مسئول)، دانشکده علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران، (email: shiry@aut.ac.ir).

روح‌اله رمضانی، دانشکده ریاضی و علوم کامپیوتر، دانشگاه دامغان، دامغان، ایران، (email: r_ramezani@du.ac.ir).

1. Task

شامل LLE [۵]، ISOMAP [۶]، LTSA[1] [۷] و Eigenmap [۸] هستند. توسعه‌هایی از این روش‌ها در سال‌های اخیر پیشنهاد شده است. Wang و همکاران (۲۰۱۸) LTSA را با بهره‌گیری از منظم‌سازی نرم اتمی[2] انجام می‌دهند تا همسایگی مناسب را در زمانی که برخی داده‌ها کامل نیستند بیابند [۹]. در [۱۰] نگاشت از فضای تانژانت با ابعاد پایین به فضای نمونه با ابعاد بالا و نگاشت معکوس آن به صورت هم‌زمان یاد گرفته می‌شود که عمل یادگیری منیفلد و نیز بازیابی داده‌های گمشده را به صورت هم‌زمان انجام می‌دهد.

تلاش اکثر روش‌های یادگیری منیفلد در راستای بازکردن[3] منیفلد است و فرض می‌شود بعد ذاتی منیفلد به صورت دانش پیشین در دسترس است. Piterlis و همکاران یادگیری یک منیفلد را به صورت یافتن یک اطلس[4] یا مجموعه‌ای از نمودارهای[5] متقاطع بیان نموده است که هم‌زمان پارامترهای پیوسته تعریف نمودارها و نحوه تخصیص نقاط به نمودارها را می‌یابد [۱۱]. روش بیان‌شده امکان بازنمایی منیفلدهای بسته را نیز دارد. در سال‌های اخیر جهت بازنمایی، داده‌های مولتی‌منیفلد با ابعاد ذاتی متفاوت ارائه شده است. Wang و همکاران ابتدا بعد ذاتی هر داده را تخمین زده و سپس داده‌ها را به خوشه‌هایی تقسیم می‌کنند به نحوی که تعداد ابعاد ذاتی داده‌های هر خوشه یکسان باشد. سپس منیفلدهای هر خوشه را یاد گرفته و یک مدل متغیر نهان را برای هر منیفلد مقداردهی اولیه می‌کند و نهایتاً یک مدل آمیخته سلسله‌مراتبی برای منیفلدها می‌سازد [۱۲].

در روش‌های نظارتی یادگیری متریک تلاش بر آن است که نمونه‌های یک دسته تا حد امکان نزدیک به یکدیگر و نمونه‌های دسته‌های متفاوت دور از یکدیگر قرار گیرند. روش LMNN از معروف‌ترین روش‌ها است که از اطلاعات محلی مجموعه آموزش استفاده می‌کند [۱۳]. Der و سایرین یک مدل متغیر پنهان را پیشنهاد داده‌اند و الگوریتم EM را جهت تخمین پارامتر به کار برده‌اند [۱۴]. ملاک‌های تئوری اطلاعات مانند دیورژانس لگاریتم دترمینان در برخی رویکردهای یادگیری متریک نظارتی نیز به کار رفته و به نتایج خوبی دست یافته‌اند.

رویکردهای نظارتی در صورت عدم وجود داده برچسب‌دار کافی در معرض بیش‌برازش قرار دارد و رویکردهای بدون ناظر از دانش موجود در داده‌های برچسب‌دار که می‌تواند متریک مناسب برای کاربرد خاص را ارائه کند بهره‌برداری نمی‌نماید. در ادامه، رویکردهای نیمه‌نظارتی که مرتبط با این مقاله هستند بررسی شده‌اند.

رویکردهای نیمه‌نظارتی بر اساس فرضی که در مورد داده‌های بدون برچسب دارند در دو دسته رویکردهای مبتنی بر فرض خوشه و مبتنی بر فرض منیفلد قرار می‌گیرند. در رویکردهای مبتنی بر منیفلد، ماتریس مجاورت، W، که میزان نزدیکی تمام داده‌ها با یکدیگر را اندازه‌گیری می‌کند تعریف شده و یک جمله منظم‌سازی در جملات بهینه‌سازی یادگیری متریک سعی در حفظ این مجاورت دارد. یادگیری متریک منظم‌شده لاپلاسین (LRML) از داده‌های بدون برچسب بدین روش بهره گرفته و به خوشه‌بندی و بازیابی تصاویر اعمال شده است [۱۵]. در ماتریس مجاورت LRML، اگر x_i در بین k نزدیک‌ترین همسایه x_j باشد $W_{ij}=1$ و در غیر این صورت برابر صفر است. یادگیری متریک تنک[6] نیمه‌نظارتی ماتریس مجاورت را با استفاده از انتشار مجاورت یاد می‌گیرد [۱۶]. یک ماتریس شاخص همسایگی در این روش تعریف می‌شود که فاصله بین زوج‌های مشابه را تا حد ممکن کم کرده و فاصله بین زوج‌های متفاوت را تا حد امکان زیاد می‌کند. باغشاه و همکاران، یک روش نیمه‌نظارتی یادگیری متریک مبتنی بر منیفلد پیشنهاد داده‌اند که بازنمایی خطی محلی در فضای ورودی را حفظ می‌کند [۱۷].

تمام روش‌های مذکور، نقاطی را که بر مبنای فاصله اقلیدسی نزدیک به یکدیگر قرار دارند مشابه یکدیگر در نظر گرفته و با تعریف محدودیت‌های محلی یا جملات منظم‌سازی سعی در حفظ این شباهت دارند. این رویکردها از دانش پیشین مربوط به شباهت داده‌های بدون برچسب بهره گرفته و دانش مربوط به عدم شباهت را در نظر نمی‌گیرند. Niu و همکاران، یک روش یادگیری متریک نیمه‌نظارتی مبتنی بر تئوری اطلاعات (SERAPH) از طریق منظم‌سازی انتروپی پیشنهاد داده که برخی محدودیت‌های رویکردهای مبتنی بر منیفلد را مرتفع نموده است [۱۸]. در این روش، ابتدا احتمال شرطی نسبت‌دادن برچسب‌ها با داشتن داده با فاصله ماهالونوبیس پارامتری شده و سپس اصل بیشینه انتروپی روی داده‌های برچسب‌دار و اصل کمینه انتروپی روی داده‌های بدون برچسب با استفاده از منظم‌سازی انتروپی اعمال شده است. SERAPH بر مبنای فرض جداسازی با چگالی پایین است.

Wang و همکاران، رویکرد یادگیری متریک نیمه‌نظارتی منظم‌شده (RSSML)[7] را بر اساس هر دو فرض هموار‌بودن روی منیفلد و فرض خوشه پیشنهاد داده‌اند [۱۹]. ماتریس مجاورت در این رویکرد بر اساس تابع هسته گوسی تعریف گردیده و چگالی محلی در آن با پنجره پارزن[8] [۲۰] تخمین زده شده است. هدف اصلی این روش، کمینه‌کردن فاصله بین نقاط نزدیک در نواحی با چگالی زیاد جهت اعمال فرض جداسازی در چگالی پایین است.

رویکردهای مذکور در مواجهه با داده‌هایی که روی چند منیفلد متقاطع قرار دارد با چالش مواجه هستند. از آنجا که فرضیات منیفلد و جداسازی در چگالی کم در نواحی اشتراک منیفلدها می‌تواند نقض شود، در این مقاله، فرض هموار‌بودن روی چند منیفلد در یادگیری متریک اعمال می‌شود. بدین معنا که تابع هدف به صورت قطعه‌ای روی هر منیفلد هموار است و دانش پیشین مرتبط با شباهت داده‌های برچسب‌دار و بدون برچسب در نواحی صحیح (و نه در مورد تمام داده‌ها) اعمال خواهد شد.

۳- روش پیشنهادی

در مسأله مورد نظر ما، مجموعه داده $\chi=\{x_i\}_{i=1}^{l+u}\in\Re^{(l+u)\times d}$ شامل داده‌های برچسب‌دار $\chi_l=\{x_i\}(1\le i\le l)$ و داده بدون برچسب $\{x_i\}(1\le i\le u)$ $(l+u=n)$ هستند. $C_l=\{c_i\}_{i=1}^{l}$ مجموعه برچسب داده‌های برچسب‌دار و در یک دسته‌بندی دودویی $c_i\in\{+1,-1\}$ است. داده‌ها از چند منیفلد $\Omega_i(1\le i\le n_\Omega)$ که می‌توانند با یکدیگر اشتراک داشته باشند نمونه‌برداری شده است به طوری که برچسب نمونه‌ها روی هر منیفلد به صورت هموار تغییر می‌کند. n_Ω تعداد منیفلدها و نامشخص است. هدف، یافتن تابع دسته‌بندی f جهت برچسب‌گذاری داده‌های بدون برچسب است. منیفلدهای زیربنایی داده با گراف مجاورتی که از داده‌ها ساخته شده بازنمایی می‌شود. ماتریس لاپلاسین به صورت

1. Local Tangent Space Alignment
2. Nuclear Norm
3. Unwrapping
4. Atlas
5. Charts
6. Sparse
7. Regularized Semi-Supervised Metric Learning
8. Parzen Window

جدول ۱: نمادهای استفاده‌شده در مقاله.

نماد	مفهوم
$\chi = \{x_i\}_{i=1}^{l+u} \in \Re^{(l+u)\times d}$	مجموعه داده‌ها
n	برابر کل تعداد داده‌ها $n = l+u$
χ_l	مجموعه داده‌های برچسب‌دار
χ_u	مجموعه داده‌های بدون برچسب
$C_l = \{c_i\}_{i=1}^{l}$	مجموعه برچسب داده‌های برچسب‌دار
L	لاپلاسین گراف، زمانی که تمرکز روی پارامتر آن است از L_ω استفاده می‌شود.
D	ماتریس درجه گراف
W	ماتریس وزن‌های گراف
ω	پارامتر هسته گاوسی
g	تابع قطعه‌ای هموار روی منیفلدهای متقاطع
g_i	تابع هموار روی i امین منیفلد
Γ_ω	عملگر لاپلاسین با پارامتر ω
Δ	عملگر لاپلاس- بلترامی
K	تعداد نزدیک‌ترین نقاط در گراف nn
f	تابع دسته‌بندی
$n_{non\text{int}}$	تعداد نقاط غیر داخلی
M	ماتریس مثبت نیمه‌معین
D_M	تابع فاصله ماهالونوبیس پارامتری‌شده با ماتریس M
$W^{(1)}$	ماتریس وزن مربوط به همسایه‌های داده‌های برچسب‌دار
$W^{(2)}$	ماتریس وزن مربوط به همسایه‌های تمام داده‌ها
X_b	مجموعه داده‌های برچسب‌گذاری شده شامل و داده‌های برچسب‌گذاری شده اولیه و داده‌های برچسب‌گذاری شده با عمل انتشار برچسب

$L = D - W$ تعریف شده که در آن W ماتریس مجاورت گراف و D ماتریسی قطری است که $D(i,i) = \sum_j w_{ij}$ می‌باشد. هر عنصر ماتریس مجاورت برای داده‌هایی که در k نزدیک‌ترین همسایگی یکدیگر هستند به صورت $w_{ij} = \exp(-\|x_i - x_j\|^2 / 2\omega)$ تعریف شده و برای سایر داده‌ها برابر صفر است.

هدف، یادگیری فاصله متریک ماهالونوبیس به فرم ذیل است

$$D_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)} \qquad (1)$$

که در آن M ماتریس مثبت نیمه‌معین است. همانند LRML و RSSML مسأله تخمین پارامتر در مدل ما به عنوان مسأله بهینه‌سازی در فضای ماتریس‌های مثبت نیمه‌معین فرمول‌بندی شده است. همانند اکثر مطالعات مرتبط با یادگیری متریک با هدف دسته‌بندی، تلاش ما در بهبود دسته‌بندی نزدیک‌ترین همسایه است [۳] و [۲۱]. مؤلفه‌های اصلی مدل پیشنهادی عبارتند از ۱) اعمال محدودیت‌های محلی به صورت دقیق برای یادگیری متریک داده‌هایی که روی منیفلدهای متقاطع قرار دارند. ۲) انتشار داده‌های برچسب‌دار برای غلبه بر کمبود داده‌های برچسب‌دار. در ادامه، ابتدا شهود روش پیشنهادی و سپس جزئیات آن ارائه شده است.

۳-۱ شهود

مدل ما بر اساس شهودهای زیر است که از فرض هموار بودن در فضای لایه‌ای استنتاج می‌شود: ۱) بر اساس فرض هموار بودن روی چند منیفلد، برچسب هر داده با احتمال بالایی برابر با برچسب نزدیک‌ترین همسایه‌ها است. این فرض برای تمام داده‌ها به جز برای داده‌های ناحیه اشتراک منیفلدها برقرار است و به بیان دیگر در نقاط داخلی منیفلدها این فرض برقرار می‌باشد. در واقع، اعمال فرض هموار بودن در نواحی اشتراک منیفلدها دانش پیشین اشتباهی را در مورد داده‌هایی که در فضای لایه‌ای قرار دارد به مدل اعمال می‌کند. ۲) داده‌های بدون برچسبی که به اندازه کافی نزدیک به داده‌های برچسب‌دار هستند و در نواحی داخلی منیفلدها واقع شده‌اند با احتمال بالایی برچسب یکسانی با داده‌های برچسب‌دار همسایه خود دارند.

جهت اعمال موارد مذکور نیاز به تمایز بین نقاط داخلی منیفلد از سایر نقاط است. فرض کنید $\bar{\Omega}$ مجموعه $\Omega_i (1 \le i \le n_\Omega)$ باشد. تابع قطعه‌ای هموار $g : \bar{\Omega} \to \Re$ بدین صورت تعریف می‌شود: اگر g_i تابع محدودشده به منیفلد iام باشد، g_i در نقاط داخلی منیفلد iام، C۲- پیوسته است. اگر نقاط به صورت یکنواخت از یک منیفلد هموار نمونه‌برداری شده باشد در نقاط داخلی منیفلد هنگامی که تعداد نقاط، n، به سمت بی‌نهایت میل کرده و ω با نرخ مناسب به سمت صفر میل کند، عملگر لاپلاسین Γ_ω به عملگر لاپلاس- بلترامی همگرا می‌شود [۸]

$$\Gamma_\omega g(x) = \Delta g(x) + O(1) \qquad (2)$$

که در آن Δ عملگر لاپلاس- بلترامی روی منیفلد است. لاپلاسین گراف اعمال‌شده به تابع g به این صورت محاسبه می‌شود

$$L_\omega g(x) = \sum_{j=1}^{n} w_{ij}(\omega)[g(x) - g(x_j)] \qquad (3)$$

$L_\omega g$ فرم ناپیوسته Γ_ω است. در بخش‌های بعد Lg و w_{ij} به ترتیب به جای $L_\omega g$ و $w_{ij}(\omega)$ استفاده خواهد شد مگر در صورتی که تمرکز ما روی پارامتر ω باشد.

نتایج به دست آمده از مطالعات نظری در سال‌های اخیر نشان داده که نقاط غیر داخلی شامل نقاط اشتراک منیفلدها، مرز و لبه‌ها جنبه‌های مهمی از داده هستند که در بسیاری از پژوهش‌ها در نظر گرفته نشده‌اند. برای نقطه x در همسایگی $\in$ نقاط غیر داخلی، (۴) برقرار است [۲۲]

$$\Gamma_\omega g(x) = \frac{1}{\sqrt{\omega}} \frac{\pi}{2} \partial_{\vec{n}} g(x) + O(\frac{1}{\sqrt{\omega}}) \qquad (4)$$

که در آن $\partial_{\vec{n}}$ نرمال واحد به سمت خارج[1] در نقطه x، $\partial_{\vec{n}} g(x)$ مشتق جهت‌دار g در جهت $\partial_{\vec{n}}$ و g تابع قطعه‌ای هموار شرح داده شده در بخش ۳-۱ است. از این ویژگی برای دسته‌بندی داده‌ها در فضای لایه‌ای استفاده شده است [۲۳].

بر طبق معادله بیان‌شده، $L_\omega g$ از مرتبه $O(1/\sqrt{\omega})$ است که بزرگ‌تر از $L_\omega g$ در نقاط داخلی است که برای مقادیر کوچک ω از مرتبه $O(1)$ است. بنابراین مقادیر بزرگ‌تر $L_\omega g$، بیانگر نقاط نزدیک نقاط غیر داخلی است.

تعریف تابع g به دلیل کمبود دانش قبلی در خصوص این که چه داده‌ای متعلق به کدام منیفلد است و نیز بعد بالای داده ورودی ساده نیست. هرچند با توجه به فرض هموار بودن روی چند منیفلد، تابع دسته‌بندی روی منیفلد، f، یک تابع قطعه‌ای هموار است، لذا پیشنهاد می‌گردد که با محاسبه لاپلاسین تابع f، نقاط داخلی از نقاط غیر داخلی تمیز داده شود.

جهت وضوح بیشتر، نمادهای استفاده‌شده مقاله در جدول ۱ آمده است.

1. Outward

۳-۲ مسأله بهینه‌سازی

مسأله بهینه‌سازی ارائه‌شده شامل دو تابع هزینه است: تابع هزینه اول بر اساس محدودیت‌های دوتایی و تابع هزینه دوم بر اساس محدودیت‌های سه‌تایی است. با داشتن زوج‌های (x_i, x_j) یادگیری متریک سعی در یافتن ماتریس مثبت نیمه‌معین M دارد به نحوی که x_i و x_j را به یکدیگر نزدیک نگه دارد و بر مبنای محدودیت‌های سه‌تایی (x_i, x_j, x_k)، تلاش می‌کند فاصله بین x_i و x_j را بزرگ‌تر از فاصله بین x_i و x_k داشته باشد. در توابع $\ell_{pull}(M)$ و $\ell_{push}(M)$ به ترتیب نقض موارد مذکور جریمه می‌شود. هدف، کمینه‌سازی جملات زیر است

$$\ell_{pull}(M) + \lambda \ell_{push}(M) \qquad (۵)$$

پارامتر وزن‌دهی λ اهمیت جملات مذکور نسبت به یکدیگر را مشخص می‌کند. بر طبق [۱۳] برای هر x_i، x_j و x_k به ترتیب همسایه هدف و همسایه impostor گفته می‌شود و تمایز روش ما با این روش در تعریف این دو نوع همسایگی است. برای هر نقطه در داخل منیفلد، k_l همسایه با برچسب یکسان به عنوان همسایه‌های هدف و k_l نزدیک‌ترین همسایه با برچسب غیر یکسان به عنوان همسایه impostor در نظر گرفته شده است.

برای محاسبه $\ell_{pull}(M)$، دو ماتریس وزن $W^{(۱)}$ و $W^{(۲)}$ تعریف می‌شود: ماتریس وزن $W^{(۱)}$ به صورت ماتریسی که عناصر درایه‌های همسایه‌های هدف داده‌های برچسب‌دار آن یک و سایر عناصر آن صفر است تعریف می‌شود. ماتریس وزن $W^{(۲)}$ ماتریسی است که در آن k نزدیک‌ترین همسایه هر داده به عنوان همسایه‌های هدف تعریف شده و عناصر درایه‌های همسایه‌های هدف به صورت زیر تعریف می‌شود

$$\exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2w}\right) \qquad (۶)$$

تابع هزینه $\ell_{pull}(M)$ به صورت زیر تعریف می‌شود

$$\ell_{pull}(M) = \frac{1}{2}\sum_{i,j\in X_b} W_{ij}^{(1)} D_{ij}^1 + \frac{1}{2}\sum_{i,j\in X_b} W_{ij}^{(2)} D_{ij}^2 =$$
$$\frac{1}{2}\sum\sum W_{ij}^{(2)} \left\|U^T(x_i - x_j)\right\|^2 = \qquad (۷)$$
$$tr(U^T XL'XU) = tr(XL'XUU^T) = tr(XL'XM)$$

$$L' = L^{(1)} + L^{(2)}$$
$$L^{(m)} = D_w^{(m)} - W^{(m)} \qquad (۸)$$

$$D_w^{(m)}(j,j) = \sum_i w_{ij}^{(m)} \qquad (۹)$$

نماد tr نشان‌دهنده عملگر trace است.

جهت اعمال حاشیه بزرگ بین imposterها و دورترین همسایه هدف، باید نامساوی زیر برقرار باشد

$$D_{ij}^2 + 1 < D_{ik}^2 \qquad (۱۰)$$

تابع هزینه $\ell_{push}(M)$ به صورت زیر تعریف می‌شود

$$\ell_{push}(M) = \sum_{i,j,k} [1 + D_{ij}^2 - D_{ik}^2]_+ \qquad (۱۱)$$

نماد $[z]_+$ به صورت $\max\{z, 0\}$ تعریف می‌شود.

مسأله بهینه‌سازی کلی به صورت زیر است

$$\min tr(XL'XM) + \gamma \sum [1 + D_{ij}^2 - D_{ik}^2]_+$$
$$\text{s.t.} M \succ= 0 \qquad (۱۲)$$

که عملگر $\succ=$ بیانگر محدودیت مثبت نیمه‌معین بودن است. مسأله بهینه‌سازی مذکور به فرم زیر بازنویسی می‌شود:

$$\min tr(XL'XM) + \gamma \sum_{i,j,k=1}^{l} \xi_{ijk}$$
$$\text{s.t.}$$
$$D_{ij}^2 - D_{ik}^2 \geq 1 - \xi_{ijk} \qquad (۱۳)$$
$$\xi_{ijk} \geq 0$$
$$M \succ= 0$$

مسأله بهینه‌سازی مذکور از نوع برنامه‌ی نیمه‌معین[1] است که می‌تواند با حل‌کننده‌های[2] همه‌منظوره حل شود [۲۴]، هرچند این روش حل نسبت به تعداد محدودیت‌ها مقیاس‌پذیر نیست. Weinberger و همکاران یک روش حل کارا بر اساس نزول گرادیان پیشنهاد کرده‌اند که سریع‌تر از solverهای معمول است [۱۲]. ما از رویکردی مشابه جهت یادگیری متریک استفاده کرده‌ایم بدین صورت که در تکرار tام، محدودیت‌هایی که ارضا نمی‌شوند مشخص شده $(\psi(t))$ و متریک کنونی $(M(t))$ جهت ارضای آنها بروز می‌شود. جهت گرادیان $(\nabla F(t))$ بر اساس این محدودیت‌ها محاسبه می‌شود. $F(t)$ به صورت زیر تعریف می‌شود

$$F(t) = \min tr(XL'XM(t)) +$$
$$\gamma \sum_{i,j,k\in\psi(t)} [1 + D_{ij}^2(t) - D_{ik}^2(t)]_+ \qquad (۱۴)$$

رویه مذکور تا زمان ارضای تمام محدودیت‌ها یا رسیدن تعداد تکرارها به تعداد تکرار از قبل مشخص شده ادامه می‌یابد.

۳-۳ اعمال دسته‌بند نزدیک‌ترین همسایه با انتشار برچسب داده‌ها در سطح مکان

جهت دسته‌بندی، متریک یاد گرفته شده به یک دسته‌بند مبتنی بر فاصله که اغلب دسته‌بند نزدیک‌ترین همسایه (NN) است اعمال می‌شود [۳]. دسته‌بند نزدیک‌ترین همسایه یک روش ناپارامتری دسته‌بندی و بسیار توانمند است. هرچند تعداد کم داده‌های برچسب‌دار در دسته‌بند NN که جهت تعیین برچسب داده‌ها در رویکردهای مبتنی بر متریک به کار می‌رود مسأله‌ساز است، لذا مدلی برای انتشار داده‌های برچسب‌دار پیشنهاد می‌شود. انگیزه اعمال این مدل، استفاده بیشتر از دانش پیشین در دسترس است. از آنجا که فرض هموار‌بودن داده‌ها روی هر منیفلد در فضای لایه‌ای برقرار است، می‌توان از این فرض جهت افزایش تعداد داده‌های برچسب‌دار با اطمینان بالا استفاده کرد. تلاش رویکردهای موجود، در اعمال این فرض فقط به داده‌های بدون برچسب و در مرحله یادگیری متریک بوده است. از فرض هموار‌بودن برچسب‌ها روی هر منیفلد، چنین استنتاج می‌شود که نزدیک‌ترین داده بدون برچسب به هر داده برچسب‌دار در نقاط داخلی منیفلد با احتمال بالایی برچسب یکسان با داده برچسب‌دار مورد نظر دارد. لذا در این مرحله، k_l نزدیک‌ترین داده برچسب‌دار به داده‌های برچسب‌دار موجود در نواحی داخلی منیفلدها به نقاط برچسب‌دار

ورودی: $\chi = \{x_i\}_{i=1}^{l+u}$، C_l، $n_{non\text{int}}$، k_l و η، stepsize

1. Semi-Definite Program
2. Solvers

خروجی: M

گام ۱: مقداردهی اولیه

– تخمین برچسب‌ها با استفاده از منظم‌سازی مینفلد

– محاسبه لاپلاسین برچسب‌های تخمین زده شده با استفاده از (۳) و مشخص‌کردن نقاط داخلی منیفلدها

گام ۲: یادگیری متریک

– $\psi(t)$ را مشخص کن.

– اگر $\psi(t)$ تهی است خاتمه.

– $\nabla F(t)$ را محاسبه کن.

– متریک کنونی را به روز رسانی کن:

$$M(t+1) = M(t) - stepSize \times \nabla F(t)$$

– $M(t+1)$ را به زیرفضای شامل ماتریس‌های مثبت نیمه‌معین تصویر کن.

– $M(t)$ را برابر $M(t+1)$ قرار بده.

گام ۳: دسته‌بندی

– فرض هموار‌بودن را به نزدیک‌ترین همسایه‌های داده‌های برچسب‌دار اعمال کن تا X_b برچسب‌گذاری شود.

– دسته‌بند ۱NN را به باقیمانده داده‌های بدون برچسب اعمال کن.

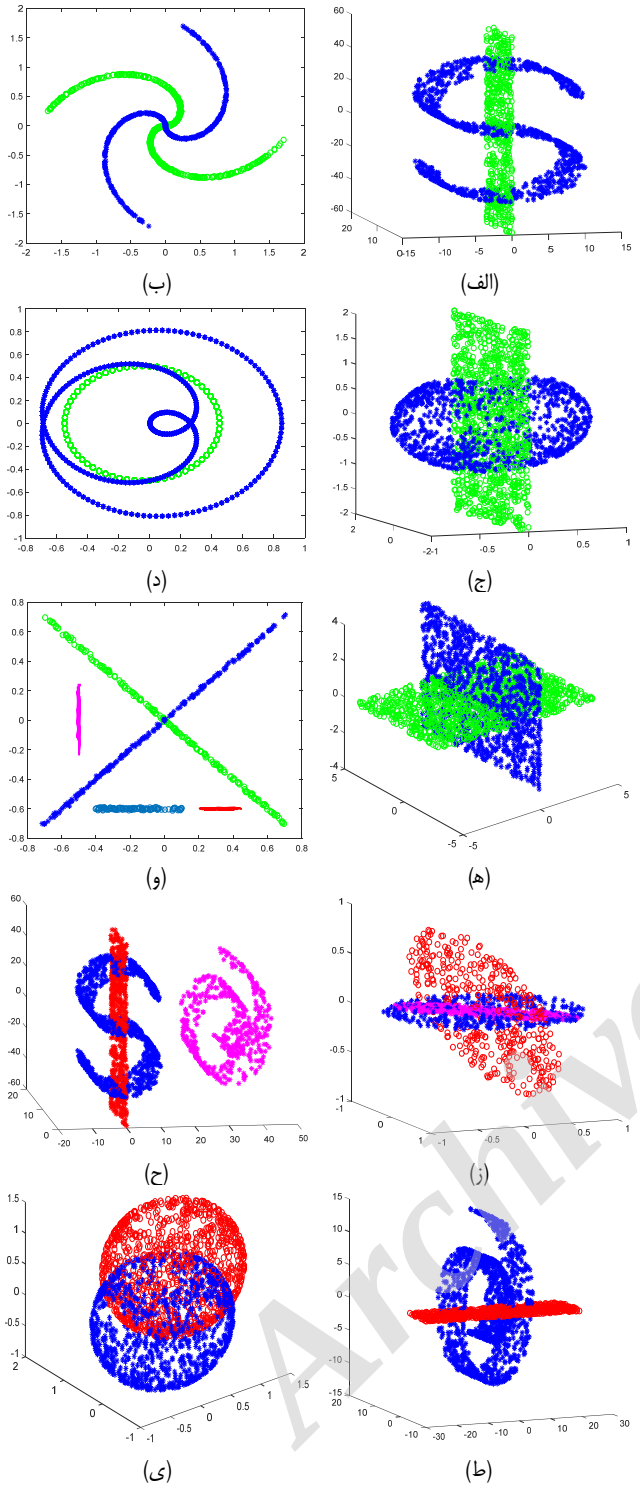
شکل ۱: گام‌های الگوریتم یادگیری متریک نیمه‌نظارتی در فضای لایه‌ای.

موجود اضافه می‌شوند و این مجموعه جدید داده‌های برچسب‌دار، X_b نام‌گذاری می‌شود. تعیین برچسب داده‌ها با دسته‌بند ۱NN با استفاده از این مجموعه داده برچسب‌دار جدید انجام می‌شود. جزئیات الگوریتم در شکل ۱ ارائه می‌شود.

٤- آزمایش‌ها

الگوریتم پیشنهادی با فاصله اقلیدسی، دو روش یادگیری متریک نیمه‌نظارتی بروز شامل RSSML و Seraph و یک روش دسته‌بندی نیمه‌نظارتی، MR[1] [۲۵] مقایسه شده است. RSSML یک روش یادگیری متریک نیمه‌نظارتی بر اساس فرض‌های هموار‌بودن روی منیفلد و جداسازی در چگالی کم است. Seraph یک روش یادگیری متریک نیمه‌نظارتی بر مبنای تئوری اطلاعات است که سعی کرده بر محدودیت‌های روش‌های مبتنی بر منیفلد غلبه کند. انواع متفاوتی از این الگوریتم ارائه شده که برخی از آنها برای بهره‌گیری از فرض منیفلد توسعه داده شده‌اند. MR فرمی از منظم‌سازی مبتنی بر تیخونوف[2] است که در روش پیشنهادی برای جداسازی نقاط داخلی منیفلد از سایر نقاط به کار رفته است. تمام آزمایش‌ها ۳۰ مرتبه با داده‌های برچسب‌دار متفاوت انتخاب‌شده به صورت تصادفی تکرار گردیده و میانگین درصد خطا (نسبت تعداد داده‌های صحیح دسته‌بندی شده به تعداد کل داده‌ها) در این ۳۰ بار اجرا و نیز انحراف معیار درصد خطا گزارش شده است. کوچک‌بودن مقدار انحراف معیار بیانگر پایداربودن روش در برابر تغییر داده‌های برچسب‌دار است.

پارامترهای MR شامل λ_A و λ_I به ترتیب از مجموعه $\{$۰٫۰۰۰۱, ۰٫۰۱,۰٫۰۲,۰٫۰۳,۰٫۰۴,۰٫۰۵,۰٫۰۶,۰٫۰۷,۰٫۰۸,۰٫۰۹,۰٫۱,۰٫۲,۰٫۳,۰٫۴,۰٫۵, ۰٫۶,۰٫۷,۰٫۸$\}$ و $\{$۰٫۹,۱,۵,۱۰,۱۵,۲۰$\}$ انتخاب گردیده است. تابع هزینه مربعی به MR اعمال شده و پارامتر K در RSSML، MR و الگوریتم پیشنهادی برابر ۱۰ است. تمام انواع Seraph شامل $Seraph_{None}$، $Seraph_{Post}$، $Seraph_{Proj}$، $Serpah_{Hyper}$، $Seraph_{Mani}$، $Seraph_{ManiPost}$ و $Serpah_{ManiHyper}$ اجرا شده و پارامترهای آن بر اساس [۱۶] تنظیم شده است. چهار تنظیم اول بیان‌شده بر اساس فرض منیفلد بوده و سایر موارد بر اساس ترکیبی از روش‌های مبتنی بر منیفلد و مبتنی بر انتروپی است.



شکل ۲: مجموعه داده‌های مصنوعی، (الف) Dollar Sign (DS)، (ب) Two Spirals (TS1)، (ج) Surface-Sphere (SS)، (د) Roll Curve and Circle (RCC)، (ه) Two Intersecting Planes (TP1)، (و) Five Segments (FS)، (ز) Three Planes (TP2)، (ح) Dollar-Sign and Plane and Roll (DPR)، (ط) Roll and Plane (RP) و (ی) Two Spheres (TS2).

1. Manifold Regularization
2. Tikhonov

جدول ۲: درصد خطا و انحراف معیار روی مجموعه داده‌های مصنوعی.

روش / مجموعه داده	DS	SS	TP۱	TP۲	TS۱	RCC	FS	RP	TS۲	DPR
روش پیشنهادی	۱۴٫۸۵ (۰٫۰۹)	۱۱٫۰۲ (۰٫۰۴)	۵٫۹۳ (۰٫۰۲)	۲٫۷۵ (۰٫۰۱)	۰٫۷۳ (۰٫۰۱)	۳٫۷۸ (۰٫۰۱)	۱٫۵۷ (۰٫۰۲)	۰٫۴۰ (۰٫۰۰)	۱٫۴۶ (۰٫۰۱)	۲٫۵۸ (۰٫۰۳)
MR	۲۴٫۸۵ (۹٫۰۲)	۱۴٫۵۸ (۵٫۶۹)	۶٫۴۵ (۲٫۴۶)	۲۲٫۵۵ (۰٫۰۳)	۴۶٫۱۱ (۰٫۰۶)	۳۶٫۶۴ (۶٫۵۰)	۴۰٫۱۶ (۰٫۰۵)	۴٫۱۰ (۰٫۰۰)	۱۴٫۴۸ (۰٫۰۲)	۱۷٫۱۶ (۰٫۰۲)
RSSML	۲۶٫۰۵ (۰٫۰۷)	۱۵٫۸۲ (۰٫۰۵)	۱۲٫۷۷ (۰٫۴۷)	۱۴٫۸۰ (۰٫۰۲)	۲٫۰۵ (۰٫۰۱)	۵٫۸۶ (۰٫۰۱)	۳٫۸۶ (۰٫۰۲)	۲٫۸۶ (۰٫۰۰)	۷٫۶۱ (۰٫۰۱)	۱۲٫۹۵ (۰٫۰۱)
فاصله اقلیدسی	۳۹٫۸۳ (۰٫۰۵)	۲۴٫۰۷ (۰٫۰۴)	۱۵٫۳۰ (۰٫۰۳)	۱۶٫۷۲ (۰٫۰۲)	۱٫۹۷ (۰٫۰۱)	۵٫۳۹ (۰٫۰۱)	۳٫۶۸ (۰٫۰۲)	۳٫۱۷ (۰٫۰۰)	۷٫۵۰ (۰٫۰۱)	۱۲٫۸۱ (۰٫۰۱)
$SERAPH_{None}$	۲۴٫۶۳ (۰٫۰۹)	۲۲٫۶۰ (۰٫۱۰)	۲۱٫۷۲ (۰٫۱۱)	۱۹٫۵۹ (۰٫۰۵)	۱٫۹۸ (۰٫۰۱)	۵٫۶۱ (۰٫۰۱)	۴٫۰۴ (۰٫۰۲)	۱۹٫۱۵ (۰٫۰۳)	۲۰٫۴۱ (۰٫۰۱)	۳۶٫۱۳ (۰٫۰۴)
$SERAPH_{Post}$	۲۴٫۷۵ (۰٫۰۹)	۲۹٫۵۱ (۰٫۱۳)	۲۵٫۵۹ (۰٫۱۴)	۲۷٫۲۵ (۰٫۱۰)	۳٫۰۶ (۰٫۰۵)	۶٫۱۶ (۰٫۰۱)	۳٫۹۸ (۰٫۰۲)	۱۶٫۴۵ (۰٫۰۴)	۲۸٫۶۹ (۰٫۰۲)	۳۶٫۲۱ (۰٫۰۴)
$SERAPH_{proj}$	۲۴٫۶۳ (۰٫۰۹)	۲۳٫۵۳ (۰٫۱۱)	۲۱٫۹۴ (۰٫۱۱)	۱۹٫۶۲ (۰٫۰۵)	۱٫۹۸ (۰٫۰۱)	۵٫۶۱ (۰٫۰۱)	۴٫۰۴ (۰٫۰۲)	۱۹٫۰۶ (۰٫۰۳)	۲۰٫۳۱ (۰٫۰۱)	۳۶٫۱۳ (۰٫۰۴)
$SERAPH_{Hyper}$	۲۴٫۷۱ (۰٫۰۹)	۳۱٫۲۴ (۰٫۱۳)	۲۵٫۵۹ (۰٫۱۴)	۲۷٫۴۹ (۰٫۱۱)	۴٫۱۲ (۰٫۰۸)	۶٫۱۶ (۰٫۰۱)	۳٫۹۸ (۰٫۰۲)	۱۶٫۶۶ (۰٫۰۴)	۲۸٫۷۹ (۰٫۰۱)	۳۶٫۲۱ (۰٫۰۴)
$SERAPH_{mani}$	۲۷٫۱۴ (۰٫۱۲)	۲۲٫۲۳ (۰٫۱۰)	۲۰٫۵۱ (۰٫۱۱)	۱۹٫۵۸ (۰٫۰۵)	۱٫۹۹ (۰٫۰۱)	۵٫۶۱ (۰٫۰۱)	۴٫۱۸ (۰٫۰۲)	۱۲٫۸۱ (۰٫۰۱)	۲۰٫۳۲ (۰٫۰۱)	۴۵٫۶۴ (۰٫۱۲)
$SERAPH_{Post+mani}$	۲۲٫۴۰ (۰٫۱۰)	۲۰٫۱۲ (۰٫۰۹)	۱۶٫۱۹ (۰٫۰۹)	۵۳٫۸۵ (۰٫۰۶)	۱٫۹۷ (۰٫۰۱)	۵٫۵۰ (۰٫۰۱)	۴٫۱۴ (۰٫۰۲)	۲۰٫۶۶ (۰٫۰۱)	۲۰٫۹۳ (۰٫۰۲)	۳۵٫۵۲ (۰٫۰۴)
$SERAPH_{proj+mani}$	۲۷٫۱۹ (۰٫۱۲)	۲۳٫۲۱ (۰٫۱۱)	۲۰٫۱۵ (۰٫۱۱)	۱۹٫۶۳ (۰٫۰۵)	۱٫۹۹ (۰٫۰۱)	۵٫۶۱ (۰٫۰۱)	۴٫۱۸ (۰٫۰۲)	۱۲٫۸۲ (۰٫۰۱)	۲۰٫۳۲ (۰٫۰۱)	۴۵٫۶۴ (۰٫۱۲)
$SERAPH_{Hyper+mani}$	۲۳٫۹۶ (۰٫۰۹)	۲۰٫۰۳ (۰٫۰۸)	۱۶٫۲۱ (۰٫۰۹)	۵۳٫۹۶ (۰٫۰۶)	۱٫۹۷ (۰٫۰۱)	۵٫۴۸ (۰٫۰۱)	۴٫۱۵ (۰٫۰۲)	۲۰٫۶۶ (۰٫۰۱)	۲۱٫۰۵ (۰٫۰۲)	۳۵٫۵۲ (۰٫۰۴)

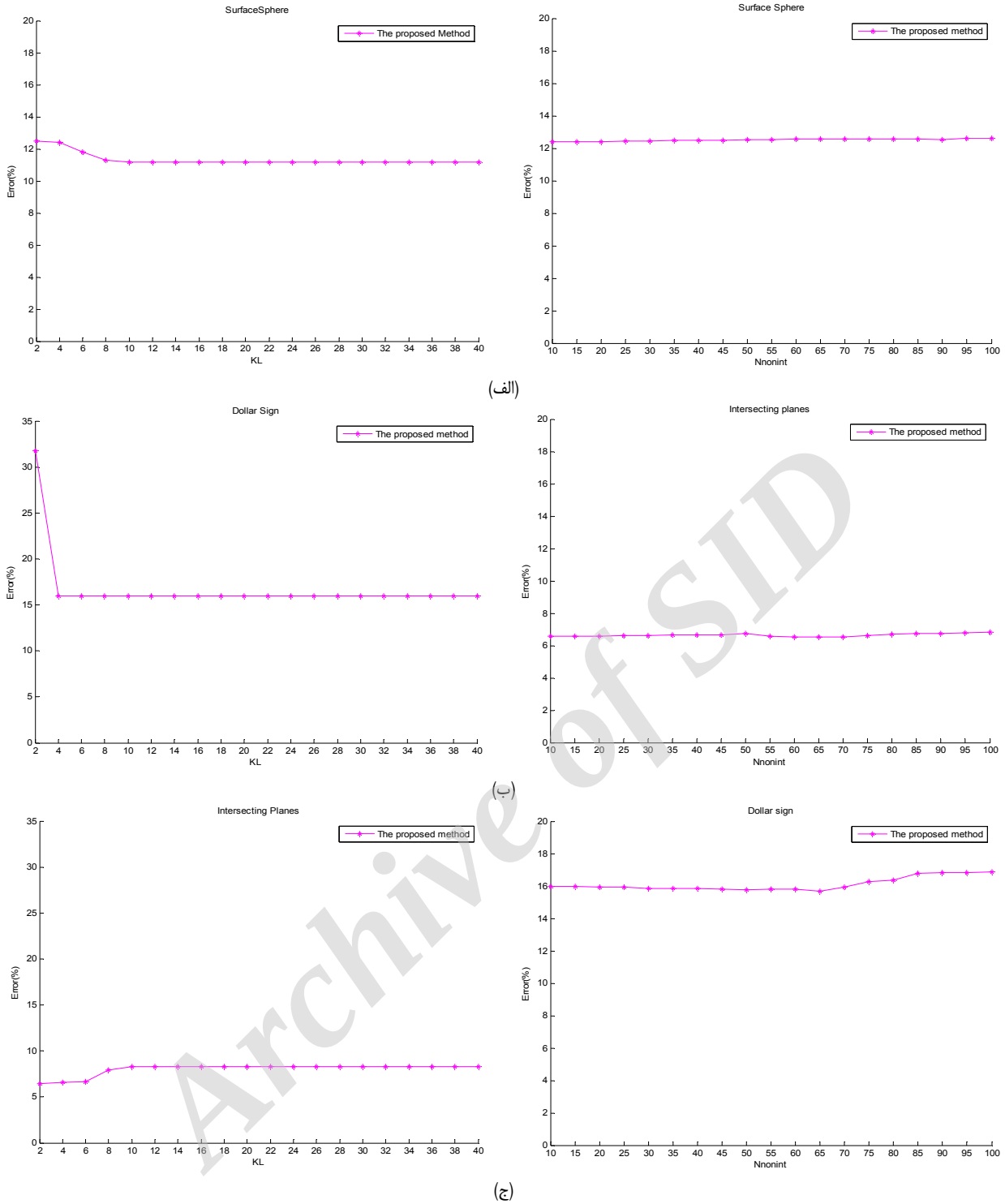
جدول ۳: مقایسه درصد خطا و تعداد داده‌های برچسب‌دار رویکرد پیشنهادی با رویکرد BOOTSTRAPPING استاندارد.

مجموعه داده	خطا				تعداد داده‌های برچسب‌دار			
	روش پیشنهادی	MR	MRBS بعد از اولین تکرار	MRBS بعد از دومین تکرار	MRBS بعد از سومین تکرار	MR و روش پیشنهادی	MRBS بعد از اولین تکرار	MRBS بعد از دومین تکرار	MRBS بعد از سومین تکرار
DS	۱۸٫۴۶	۱۰٫۴۲	۴۷٫۶۰	۳۳٫۳۳	۵۶٫۲۶	۱۰	۱۵۲	۴۹۱	۵۶۵
SS	۱۴٫۷۵	۱۷٫۹۵	–	–	–	۲۰	–	–	–
TP۱	۹٫۱۴	۶٫۲۰	–	–	–	۲۰	–	–	–
TP۲	۱۴٫۵۴	۲۲٫۲۵	۶۶٫۶۶	۱۰۰٫۰۰	–	۱۲۰	۱۹۳	۴۶۹	–
TS۱	۲٫۹۰	۲۴٫۵۰	۶۶٫۲۰	۶۰٫۷۰	۵۶٫۱۰	۱۰۰	۲۸۴	۳۰۶	۳۱۵
TS۲	۶٫۸۳	۱۶٫۸۰	۶۱٫۳۰	۵۹٫۰۰	۵۶٫۷۰	۲۰۰	۳۸۴	۵۲۲	۵۹۷
RCC	۶٫۹۴	۲۴٫۹۸	۷۵٫۱۳	۷۲٫۳۲	۷۲٫۳۲	۲۵۶	۳۶۶	۴۶۳	۵۰۱
FS	۶٫۸۳	۱۵٫۴۳	۸۵٫۷۱	۱۰۰٫۰۰	–	۷۰	۹۶	۴۹۵	–
RP	۲٫۵۶	۳٫۵۵	۳۳٫۵۵	۳۳٫۴۵	۳۳٫۳۰	۲۰۰	۱۱۴۲	۱۲۶۸	۱۳۰۷
DPR	۲٫۵۶	۱۴٫۹۵	۷۵٫۰۰	۱۰۰٫۰۰	–	۲۰۰	۲۰۰	۴۷۶	۱۱۷۰

نتیجه آزمایش روی ۱۰ مجموعه داده مصنوعی استاندارد آمده است و شکل ۲ مجموعه داده‌ها را نشان می‌دهد. در این مجموعه داده‌ها، $w=۰٫۵$ ، $n=۲۰۰۰$ و ۱۰٪ داده‌ها برچسب‌دار است و تعداد داده‌های هر دو دسته با یکدیگر برابر هستند. عرض هسته در MR برابر ۰٫۵ در نظر گرفته شده است. میانگین نرخ درصد خطا به ازای ۳۰ بار اجرای مستقل و انحراف معیار نتایج به دست آمده از اعمال یک دسته‌بند ۱NN با استفاده از معیارهای فاصله بیان‌شده در جدول ۲ آمده است. آزمون تی-استیودنت با سطح معناداری ۰٫۰۵ نشان می‌دهد که روش پیشنهادی از سایر روش‌ها بهتر است.

۴-۱ بحث

همان طور که نتایج به دست آمده در جدول ۲ نشان می‌دهد با وجود موفقیت MR در بسیاری از داده‌ها، این روش روی داده‌هایی که روی منیفلدهای متقاطع قرار دارند به واسطه انتشار برچسب اشتباه به نتایج



شکل ۳: نمودار درصد خطا به ازای تغییر مقدار پارامترهای $n_{non\text{int}}$ (ستون سمت راست) و k_l (ستون سمت چپ) روی مجموعه داده‌های مصنوعی، (الف) Surface-sphere، (ب) Intersecting planes و (ج) Dollar sign.

خوبی نمی‌رسد. همچنین در برابر تغییر داده‌های برچسب‌دار نسبت به سایر روش‌ها ناپایدارتر (انحراف معیار بالاتر) است و فاصله اقلیدسی نیز به واسطه قرارگیری داده‌های روی منیفلد و نیز اشتراک منیفلدها کارایی لازم را ندارد. روش پیشنهادی بهتر از RSSML و Seraph که تقاطع منیفلدها را در نظر نمی‌گیرند عمل می‌کند.

در ادامه، جهت تحلیل دقیق‌تر، دقت روش پیشنهادی با روش معمول bootstrapping (MRBS)[1] مقایسه شده است. در روش معمول در هرمرحله داده‌های دسته‌بندی شده با اطمینان بالا به داده‌های برچسب‌دار در دسترس اضافه می‌شود. MR به عنوان دسته‌بند اولیه انتخاب شده و دقت این دسته‌بند در چهار تکرار اول و نیز تعداد داده‌های برچسب‌دار در این تکرارها در جدول ۳ آمده است. همان گونه که مشخص است تکرار عمل اضافه‌نمودن داده‌هایی که با اطمینان بالایی درست دسته‌بندی شده‌اند در اکثر موارد موجب خطای بسیار زیادی شده و هر چند تعداد داده‌های برچسب‌دار در گام‌های دو به بعد زیاد می‌شود اما خطای دسته‌بندی افزایش می‌یابد.

1. MR Bootstrapping

٤-۲ بررسی تأثیر تغییر مقدار پارامترها

در این بخش، تأثیر انتخاب پارامترها را روی دقت روش پیشنهادی بررسی می‌کنیم. دو پارامتر اصلی تعداد نقاط غیر داخلی و k_l در روش پیشنهادی وجود دارد.

میانگین درصد خطای روش پیشنهادی روی ۳۰ اجرای مستقل به ازای مقادیر مختلف پارامترها روی چند مجموعه داده بررسی شده است. در هر زمان، مقدار یک پارامتر تغییر داده شده و مقدار سایر پارامترها برابر با مقدار پیش‌فرض تنظیم شده است. ستون سمت راست در شکل ۳ نمودار تغییر درصد خطا به ازای تغییر مقدار پارامتر $n_{non\,\mathrm{int}}$ از ۱۰ تا ۱۰۰ را نشان می‌دهد. از آنجا که در برخی تنظیمات تعداد نقاط با مقادیر لاپلاسین تابع غیر صفر کمتر از مقدار مشخص‌شده $n_{non\,\mathrm{int}}$ است، بیشینه مقدار $n_{non\,\mathrm{int}}$ در نظر گرفته شده است. نمودار ارائه‌شده حاکی از آن است که در تمام مجموعه داده‌ها برای مقادیر زیادی از $n \ll n_{non\,\mathrm{int}}$ کارایی رویکرد پیشنهادی مستقل از مقدار $n_{non\,\mathrm{int}}$ است. نمودار تغییر مقدار درصد خطا به ازای تغییر مقدار k_l در ستون سمت چپ شکل ۳ آمده است. روش پیشنهادی در زمانی که مقدار k_l خیلی کوچک نباشد مستقل از این پارامتر است.

۵- نتیجه‌گیری

در این مقاله، یک رویکرد یادگیری متریک نیمه‌نظارتی با فرض قرارگیری داده در فضای لایه‌ای ارائه شد. با اعمال فرض هموار‌بودن فقط در نواحی داخلی منیفلدها، رویکردهای مبتنی بر منیفلد موجود جهت یادگیری متریک در فضای لایه‌ای توسعه داده شده است. همچنین رویکرد جدیدی جهت اعمال فرض هموار‌بودن به نزدیک‌ترین همسایه داده‌های برچسب‌دار تعریف شده است. موارد مذکور بر مبنای بهره‌گیری دقیق‌تر از دانش پیشین قرارگیری داده‌ها روی چند منیفلد بوده و منجر به کارایی بهتر رویکرد یادگیری متریک پیشنهادی در مقایسه با روش‌های مشابه شده است.

مراجع

[1] J. Lu, X. Zhou, Y. P. Tan, Y. Shang, and J. Zhou, "Neighborhood repulsed metric learning for kinship verification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 331-345, Feb. 2014.

[2] J. Yu, M. Wang, and D. Tao, "Semisupervised multiview distance metric learning for cartoon synthesis," *IEEE Trans. on Image Processing*, vol. 21, no. 11, pp. 4636-4648, Nov. 2012.

[3] B. Kulis, "Metric learning: a survey," *Foundations and Trends in Machine Learning*, vol. 5, no. 4, pp. 287-364, 2012.

[4] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *J. of Machine Learning Research*, vol. 13, no. 1, pp. 1-26, Jan. 2012.

[5] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsuper-learning of low dimensional manifolds," *the J. of Machine Learning Research*, vol. 4, pp. 119-155, 12 Jan. 2003.

[6] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 22 Dec. 2000.

[7] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM J. on Scientific Computing*, vol. 26, no. 1, pp. 313-338, 2004.

[8] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373-1396, 1 Jun. 2003.

[9] J. Wang, X. Sun, and J. Du, "Local tangent space alignment via nuclear norm regularization for incomplete data," *Neurocomputing*, vol. 273, pp. 141-151, 17 Jan. 2018.

[10] M. A. Carreira-Perpin and Z. Lu, "Manifold learning and missing data recovery through unsupervised regression," in *Proc. IEEE 11th Int. Conf. on Data Mining, ICDM'11*, pp. 1014-1019, Vancouver, Canada, 11-14 Dec. 2011.

[11] N. Pitelis, C. Russell, and L. Agapito, "Learning a manifold as an atlas," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'13,* pp. 1642-1649, Portland, OR, USA, 23-28 Jun. 2013.

[12] X. Wang, T. Peter, and A. F. Mark, "Multiple manifolds learning framework based on hierarchical mixture density model," in *Proc. Joint European Conf. on Machine Learning and Knowledge Discovery in Databases,* pp. 566-581, Antwerp, Belgium, 15-19 Sept. 2008.

[13] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The J. of Machine Learning Research*, vol. 10, pp. 207-244, 12 Jan. 2009.

[14] M. Der and L. K. Saul, "Latent coincidence analysis: a hidden variable model for distance metric learning," in *Proc. 25th Int. Conf. on Neural Information Processing Systems*, vol. 2, pp. 3230-3238, Lake Tahoe, NV, US, 3-6 Dec. 2012.

[15] S. C. Hoi, W. Liu, and S. F. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Trans. on Multimedia Computing, Communications, and Applications,* vol. 6, no. 3, pp. 18-43, Aug. 2010.

[16] W. Liu, S. Ma, D. Tao, J. Liu, and P. Liu, "Semi-supervised sparse metric learning using alternating linearization optimization," in *Proc. Int. 16th Conf. on Knowledge Discovery and Data Mining, ACMSIGKDD'10*, pp. 1139-1148, Washington, DC, US, 25-28 Jul. 2010.

[17] M. Baghshah and S. Shouraki, "Semi-supervised metric learning using pairwise constraints," in *Proc. of the 21st Int. Joint Conf. on Artificial Intelligence*, pp. 1217-1222, Pasadena, CA, US, 11-17 Jul. 2009.

[18] G. Niu, B. Dai, M. Yamada, and M. Sugiyama, "Information-theoretic semisupervised metric learning via entropy regularization," *Neural Computation*, vol. 26, no. 8, pp. 1717-1762, Aug. 2012.

[19] Q. Wang, P. C. Yuen, and G. Feng, "Semi-supervised metric learning via topology preserving multiple semi-supervised assumptions," *Pattern Recognition*, vol. 46, no. 9, pp. 2576-2587, Sept. 2013.

[20] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley and Sons, 2001.

[21] L. Yang and R. Jin, *Distance Metric Learning: A Comprehensive Survey*, Michigan State Universiy, 2006.

[22] M. Belkin, Q. Que, Y. Wang, and X. Zhou, "Toward understanding complex spaces: graph laplacians on manifolds with singularities and boundaries," in *Proc. 25th Annual Conf. on Learning Theory, COLT'12,* pp. 1-26, Edinburgh, Scotland, 2012.

[23] Z. Karimi and S. Shiry Ghidary, "Semi-Supervised Classification in Stratified Spaces by Considering Non-interior Points Using Laplacian Behavior," *Neurocomputing*, vol. 239, pp. 223-231, 24 May 2017.

[24] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49-95, Mar. 1996.

[25] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples," *J. of Machine Learning Research*, vol. 7, pp. 2399-2434, Nov. 2006.

زهره کریمی در رشته مهندسی کامپیوتر در گرایش نرم‌افزار در سال ۱۳۸۵ از دانشگاه شهید بهشتی فارغ التحصیل شد و مدرک کارشناسی ارشد خود را در سال ۱۳۸۸ از دانشگاه صنعتی شریف در همان رشته اخذ کرد. او دکتری خود را در رشته هوش مصنوعی در دانشگاه صنعتی امیرکبیر دریافت نموده است. زمینه‌های پژوهش وی شامل یادگیری ماشین، داده‌کاوی و یادگیری مبتنی بر منیفلد است.

سعید شیری قیداری استاد دانشکده علوم کامپیوتر دانشگاه صنعتی امیرکبیر هستند. ایشان مدرک کارشناسی خود را در رشته مهندسی الکترونیک و کارشناسی ارشد خود را در مهندسی کامپیوتر از دانشگاه صنعتی امیرکبیر اخذ نموده است. وی در سال ۱۳۸۱ دکتری خود را از دانشگاه کوبه ژاپن دریافت نموده و از سال ۱۳۸۳ استادیار دانشگاه صنعتی امیرکبیر است. زمینه‌های تحقیقاتی مورد علاقه وی رباتیک، بینایی ماشین، علوم شناختی و مدل‌سازی مغز است.

روح‌اله رمضانی در رشته آمار در سال ۱۳۸۳ از دانشگاه بین‌المللی امام خمینی (ره) فارغ‌التحصیل شد و مدرک کارشناسی ارشد خود را در سال ۱۳۸۵ از دانشگاه صنعتی امیرکبیر در همان رشته اخذ کرد. زمینه‌های پژوهش وی شامل داده‌کاوی، پژوهش‌های آماری، کنترل کیفیت آماری و قابلیت اطمینان است. وی، هم اکنون عضو هیأت علمی دانشگاه دامغان است.