علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۶، شماره ۱، ۱۳۹۷
صفحات ۱- ۸
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

دسته‌بندی نیمه نظارتی منیفلدهای متقاطع بر مبنای تمایز نقاط داخلی منیفلدها از سایر نقاط

زهره کریمی[۱] سعید شیری قیداری[۲] محمد رحمتی[۱] روح‌اله رمضانی[۳]

[۱]دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران
[۲]دانشکده ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران
[۳]دانشکده ریاضی و علوم کامپیوتر، دانشگاه دامغان، دامغان، ایران

چکیده

دسته‌بندی نیمه نظارتی مبتنی بر منیفلد در سال‌های اخیر مورد توجه بسیاری از پژوهشگران واقع شده است. رویکردهای موجود از فاصله اقلیدسی به صورت محلی برای تقریب فاصله روی منیفلدها و اعمال فرض هموار بودن روی منیفلد استفاده می‌کنند. در فضایی که چند منیفلد با یکدیگر اشتراک دارند این تقریب در نواحی اشتراک صحیح نبوده و باعث انتشار اشتباه برچسب‌ها می‌شود. در این مقاله الگوریتمی بر مبنای تفکیک نقاط داخلی منیفلد از سایر نقاط جهت دسته‌بندی نیمه‌نظارتی روی منیفلدهای متقاطع جهت یادگیری دسته‌بند مبتنی بر اتصالات مطمئن‌تر در گراف ارائه‌کننده‌ی داده پیشنهاد شده است. الگوریتم پیشنهادی وزن یال‌های گراف ارائه‌کننده‌ی منیفلد را جهت انتشار برچسب اصلاح می‌کند. در مقایسه با رویکردهای دسته‌بندی نیمه نظارتی روی چند منیفلد، رویکرد پیشنهادی بر مبنای این فرض‌های محدودکننده نیست: مشخص بودن ابعاد ذاتی منیفلدها، نیاز به تعداد خیلی زیاد داده‌های بدون برچسب جهت تخمین منیفلدها و انتساب خصوصیات همسایگی مشابه به تمام نقاط. آزمایش‌ها روی مجموعه داده‌های مصنوعی و واقعی نشان‌دهنده‌ی دقت خوب روش پیشنهادی نسبت به روش‌های مشابه است.

کلمات کلیدی: دسته‌بندی نیمه نظارتی، منیفلدهای متقاطع، لاپلاسین، فرض هموار بودن.

۱- مقدمه

دسته‌بندی نیمه نظارتی مبتنی بر فرض قرارگیری داده‌ها روی منیفلد در سال‌های اخیر بسیار مورد توجه قرار گرفته است [۶، ۷، ۱۰ و ۱۶]. این روش‌ها، فرض هموار بودن داده‌ها روی منیفلد را دارند بدین معنا که داده‌های "مجاور[۱]" یکدیگر با احتمال بالایی برچسب یکسان دارند. مجاورت براساس فاصله‌ی اقلیدسی محلی تعریف می‌شود. این مشخصه برای روش‌هایی که فرض قرارگیری داده‌ها روی یک منیفلد را دارند مناسب است، هر چند زمانی که داده‌ها روی چند منیفلد قرار دارند، در نظر گرفتن مجاورت با توجه به فاصله‌ی اقلیدسی محلی مسأله‌ساز است. از آن‌جا که فرض هموار بودن روی داده‌های هر منیفلد و نه لزوماً روی تمام داده‌ها برقرار است. داده‌های منیفلدهای گوناگون ممکن است شباهت زیادی به یکدیگر داشته باشند، به بیان دیگر منیفلدها ممکن است با یکدیگر تقاطع داشته باشند در این صورت تعریف مجاورت براساس فاصله‌ی اقلیدسی محلی صحیح نیست.

در سال‌های اخیر رویکردهای دسته‌بندی مبتنی بر فرض قرارگیری داده روی چند منیفلد پیشنهاد شده‌اند [۲۱]. هر چند، این رویکردها محدودیت‌های زیر را دارند: (۱) فرض می‌کنند تعداد ابعاد ذاتی منیفلدها از ابتدا شناخته شده است [۲۳]؛ از آن‌جا که منیفلدی که هر داده متعلق به آن است از ابتدا شناخته شده نیست و منیفلدهای گوناگون می‌توانند ابعاد ذاتی متفاوت داشته باشند، این فرض صحیح نیست. نشان داده شده که تشخیص اشتباه ابعاد ذاتی منیفلدها تاثیر زیادی در این روش‌ها دارد [۲۲]. علاوه بر آن، بعد ذاتی در یک منیفلد در نقاط مرزی

متفاوت از سایر نقاط است [۳]. (۲) نیاز به تعداد بسیار زیاد داده‌های بدون برچسب دارند [۱۲]: هر چند داده‌های بدون برچسب به وفور در دسترس است پردازش حجم بسیار بالای آن، الگوریتم مربوطه را ناکارا می‌سازد. (۳) تمایزی بین نقاط داخلی منیفلد و سایر نقاط جهت تعیین همسایگی ندارند [۱۱]: همان‌گونه که بیان شد بعد ذاتی منیفلد در نقاط مختلف متفاوت بوده و تعیین همسایگی براساس فاصله‌ی اقلیدسی در تمام نقاط مناسب نیست.

در این مقاله روشی نیمه نظارتی با فرض قرارگیری داده روی منیفلدهای متقاطع پیشنهاد شده که محدودیت‌های مذکور را مرتفع نموده است. این روش از ویژگی لاپلاسین تابع قطعه‌ای هموار روی منیفلدها جهت تشخیص نقاط داخلی منیفلدها از سایر نقاط استفاده می‌کند. در نقاط داخلی منیفلد، استفاده از فاصله‌ی اقلیدسی جهت اعمال فرض هموار بودن روی منیفلد صحیح است و در سایر نقاط ضریب اطمینانی پیشنهاد می‌شود که اطمینان به فاصله‌ی اقلیدسی را نسبت به نقاط داخلی کم می‌سازد. مسأله‌ی بهینه‌سازی ابتدا به صورت دو مولفه‌ای روی تابع دسته‌بندی و وزن یال‌های گراف ارائه شده و سپس به مساله‌ای یک مولفه‌ای تبدیل شده و با استفاده از رویکردی شبیه $EM^{۲}$ حل شده است. رویکرد پیشنهادی روی مجموعه داده‌های مصنوعی و واقعی ارزیابی شده و نتایج بدست آمده حاکی از دقت بهتر روش پیشنهادی نسبت به روش‌های موجود است.

در ادامه در بخش ۲ پژوهش‌های مرتبط شرح داده شده، در بخش ۳ جزئیات رویکرد پیشنهادی و در بخش ۴ نتایج آزمایش‌ها خواهد آمد. در بخش ۵ نتیجه‌گیری ارائه می‌شود.

۲- پژوهش‌های مرتبط

رویکردهای دسته‌بندی نیمه نظارتی مبتنی بر فرض قرارگیری داده‌ها روی منیفلد بسیار مورد توجه محققان قرار گرفته است [۵]. در این روش‌ها، در ابتدا گرافی که نشان‌دهنده‌ی منیفلد زیربنایی داده است ساخته شده و سپس انتشار برچسب از داده‌های برچسب‌دار به داده‌های بدون برچسب انجام می‌شود. در برخی رویکردها تابع دسته‌بندی جهت دسته‌بندی داده‌های بدون برچسبی که در مرحله‌ی آموزش دیده نشده‌اند یاد گرفته می‌شود [۱]. بسیاری از تحقیقات روی نحوه‌ی انتشار برچسب یا یادگیری تابع دسته‌بندی متمرکز شده‌اند که سعی در اعمال فرض هموار بودن برچسب‌ها روی منیفلد دارند، هر چند پژوهش‌های اخیر نشان داده است که گراف ارائه‌کننده‌ی منیفلد نیز نقش بسیار تعیین‌کننده‌ای در کارایی دسته‌بند بکار رفته دارد [۸].

لذا رویکردهایی جهت بازنمایی مناسب داده با گراف مناسب ارائه شده است [۱۸] این رویکردها فرض قرارگیری داده روی یک منیفلد را دارند در حالی که داده‌های واقعی روی منیفلدهای متقاطع قرار دارند. در سال‌های اخیر رویکردهایی جهت دسته‌بندی نیمه نظارتی با فرض هموار بودن برچسب روی منیفلدهای متقاطع پیشنهاد گردیده است [۹، ۱۲]. Goldberg و همکاران از فاصله‌ی هلینگر به جای فاصله‌ی اقلیدسی جهت تشخیص تغییر در بعد، جهت[۳] و چگالی داده‌ها برای ساخت گراف استفاده کرده است و بعد خوشه‌بندی طیفی با قیود اندازه را به گراف بدست آمده اعمال می‌کند تا در هر خوشه به تعداد کافی داده‌ی برچسب‌دار و بدون برچسب جهت اطمینان از بهبود عملکرد رویکردهای نیمه نظارتی نسبت به رویکرد نظارتی وجود داشته باشد [۱۲]. فاصله‌ی هلینگر وزن یال‌های موجود در نواحی اشتراک را کم می‌کند؛ هر چند محاسبه‌ی آن نیاز به تعداد خیلی زیاد داده‌ی بدون برچسب دارد که الگوریتم مورد نظر را ناکارا می‌سازد. این الگوریتم از رویکرد حریصانه جهت نمونه‌برداری از داده‌های بدون برچسب استفاده می‌کند که دقت روش را کم می‌سازد.

Xing و همکاران جهت مرتفع نمودن این مسأله، وزن یال‌ها را براساس زاویه‌ی بین فضاهای تانژانت محلی نقاط محاسبه کرده و سپس انتشار برچسب را روی گراف حاصل اعمال می‌کند [۲۳]. محاسبه‌ی زاویه‌ی بین فضای تانژانت محلی نقاط بر مبنای این فرض است که تعداد ابعاد ذاتی منیفلدها از ابتدا شناخته شده است. در حالی که منیفلدها می‌توانند ابعاد ذاتی متفاوت داشته باشند و تعلق هر داده به هر منیفلد و نیز بعد ذاتی منیفلدها از ابتدا شناخته شده نبوده و خود یک مسأله‌ی تحقیقاتی است.

Geng و همکاران فرض می‌کنند داده‌ها روی ترکیب محدب چند منیفلد از پیش مشخص شده قرار دارند و سپس سعی در یافتن وزن هر یک از این منیفلدها در ساخت منیفلد نهایی بهینه دارد [۱۱]. در این روش وزن هر یک از منیفلدها در تمام داده‌ها یکسان فرض شده است در حالی‌که از آن‌جا که داده‌ها روی چند منیفلد قرار دارند وزن آن در ساخت داده‌های هر منیفلد متفاوت است. به عبارت دیگر، ویژگی‌های همسایگی در اطراف تمام نقاط در این روش یکسان فرض شده است که مناسب نیست.

رویکردهای دسته‌بندی با ناظر [۱۳، ۲۰]، کاهش بعد [۱۷]، استخراج ویژگی [۱۵، ۲۴] و یادگیری منیفلد [۱۹] نیز با فرض قرارگیری داده روی چند منیفلد در سال‌های اخیر پیشنهاد شده است که منطبق بر هدف این مقاله که دسته‌بندی نیمه نظارتی است، نیست. در این مقاله، یک روش دسته‌بندی نیمه نظارتی مبتنی بر فرض قرارگیری داده‌ها روی منیفلدهای متقاطع ارائه می‌شود که محدودیت‌های روش‌های مذکور در آن مرتفع می‌گردد.

۳- روش پیشنهادی

در این بخش ابتدا تعریف مسأله ارائه شده و سپس جزئیات رویکرد پیشنهادی بیان می‌گردد.

۳-۱- تعریف مسأله

مسأله‌ی موردنظر به این‌صورت است: مجموعه داده‌ی $\chi = \{x_i\}_{i=1}^{l+u} \in \Re^{(l+u)\times d}$ شامل داده‌های برچسب‌دار $\chi_l = \{x_i\}(1<=i<=l)$ و داده‌ی بدون برچسب $(l+u=n)\ \{x_i\}(1<=i<=u)$ هستند. $C_l = \{c_i\}_{i=1}^{l}$ مجموعه‌ی برچسب داده‌های برچسب‌دار است، در یک دسته‌بندی دودویی است. داده‌ها از چند منیفلد $\Omega_i(1<=i<=n_\Omega)$ که می‌توانند با یکدیگر اشتراک داشته باشند نمونه‌برداری شده است، به طوری‌که 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 تعریف شده که در آن W ماتریس مجاورت گراف و D ماتریسی قطری است که $D(i,i)=\sum_j w_{ij}$.

فرض کنید $\overline{\Omega}$ مجموعه‌ی $\Omega_i(1<=i<=n_\Omega)$ باشد، تابع قطعه‌ای هموار بدین صورت تعریف می‌شود: اگر g_i تابع محدود شده به منیفلد iم باشد، g_i در نقاط داخلی منیفلد iم، C^2- پیوسته است. اگر نقاط به صورت یکنواخت از یک منیفلد هموار نمونه‌برداری شده باشد در نقاط داخلی منیفلد هنگامی که n به سمت بی‌نهایت میل کرده و t با نرخ مناسب به سمت صفر میل کند، عملگر لاپلاسین Γ_t به عملگر لاپلاس- بلترامی همگرا می‌شود [۲]:

$$\Gamma_t\, g(x) = \Delta g(x) + O(1) \qquad (۱)$$

t پارامتر هسته گاوسی است که در محاسبه وزن یال‌ها بکار رفته است و Δ عملگر لاپلاس - بلترامی روی منیفلد است. لاپلاسین گراف اعمال شده به تابع g به این صورت محاسبه می‌شود:

$$L_t\, g(x) = \sum_{j=1}^{n} w_{ij}(t)[g(x) - g(x_j)] \qquad (۲)$$

جدول ۱- نمادهای استفاده شده در مقاله

نماد	مفهوم
$\chi = \{x_i\}_{i=1}^{l+u} \in \Re^{(l+u)\times d}$	مجموعه داده‌ها
d	تعداد ابعاد داده‌ی ورودی
n	برابر کل تعداد داده‌ها n=l+u
χ_l	مجموعه‌ی داده‌های برچسب‌دار
χ_u	مجموعه‌ی داده‌های بدون برچسب
$C_l = \{c_i\}_{i=1}^{l}$	مجموعه‌ی برچسب داده‌های برچسب‌دار
L	لاپلاسین گراف، زمانی که تمرکز روی پارامتر آن است از L_t استفاده می‌شود.
D	ماتریس درجه‌ی گراف
W	ماتریس وزن‌های گراف
T	پارامتر هسته‌ی گاوسی
g	تابع قطعه‌ای هموار روی منیفلدهای متقاطع
g_i	تابع هموار روی iمین منیفلد
Γ_t	عملگر لاپلاسین با پارامتر t
Δ	عملگر لاپلاس- بلترامی
γ_A	ضریب پیچیدگی تابع در فضای امبینت[۴]
γ_I'	ضریب پیچیدگی تابع در منیفلدهای متقاطع
H_k	فضای هیلبرت هسته‌ی بازتولید با هسته‌ی مرتبط K
$K: \chi \times \chi \to \Re^+$	هسته‌ی مرسر
V	تابع هزینه
K	تعداد نزدیک‌ترین نقاط در گراف نزدیک‌ترین همسایگی
W'	وزن جدید اصلاح شده با توجه به تمایز بین نقاط داخلی و غیرداخلی منیفلد از یکدیگر
f	تابع دسته‌بندی
n_{dec}	تعداد نقاط غیرداخلی
c_{dec}	ضریب اطمینان وزن‌ها
$H_{n\times n}$	ماتریسی که اتصالات داخلی را مشخص می‌کند

$L_t g$ فرم ناپیوسته‌ی Γ_t است. در بخش‌های بعد Lg و w_{ij} به‌ترتیب به‌جای $L_t g$ و (t) استفاده خواهد شد مگر در صورتی که تمرکز ما روی پارامتر t باشد. جهت وضوح بیشتر، نمادهای استفاده شده در مقاله در جدول ۱ آمده است.

۳-۲- الگوریتم پیشنهادی

در این بخش، روش پیشنهادی جهت دسته‌بندی نیمه نظارتی داده‌هایی که روی چند منیفلد متقاطع قرار دارد ارائه می‌شود. رویکرد پیشنهادی مبتنی بر قائل بودن اهمیت بیشتر برای اتصالات بین نقاط داخلی منیفلدها است از آن‌جا که در این نقاط فاصله اقلیدسی محلی با فاصله روی منیفلد برابر است.

لذا رویکرد پیشنهادی به‌صورت یک مسأله‌ی بهینه‌سازی دو مولفه‌ای روی وزن یال‌های گراف و نیز برچسب داده‌ها فرموله می‌شود؛ این مسأله‌ی بهینه‌سازی همزمان وزن یال‌های گراف اولیه را در جهت سازگاری با منیفلدهای متقاطع تغییر داده و همچنین تابع دسته‌بندی را می‌یابد.

$$\min_{f\in H_k, W'>=0} \frac{1}{l}\sum_{i=1}^{l} V(f, x_i, c_i) + \gamma_A \| f \|_k^2 + \gamma_I' \sum_{i,j=1}^{n} w_{ij}' (f(x_i) - f(x_j))^2$$

$$\text{s.t}\begin{cases} 0 <= w_{ij}' < w_{ij}, x_i isno \tan int\ eriorpo\ int \\ w_{ij}' = w_{ij}, otherwise \end{cases} \qquad (۳)$$

که در آن فضای هیلبرت هسته‌ی بازتولید[۵] (RKHS) و V تابع هزینه مانند هزینه‌ی هینگ[۶] یا هزینه‌ی مربع است، جمله‌ی جریمه هموار نبودن در فضای امبینت[۷] است. پارامترهای γ_A، جهت ایجاد توازن بین تابع هزینه و جملات منظم‌سازی امبینت و ذاتی[۸] در امتداد منیفلدهای متقاطع بکار می‌رود. هموار بودن دسته‌بند در امتداد منیفلدهای متقاطع از داده‌های برچسب‌دار در آخرین جمله تخمین زده می‌شود. W_{ij}' عنصر سطر iم و ستون jم ماتریس وزن W' است. جهت تمایز نقاط داخلی منیفلد از سایر نقاط تاکنون رویکردهای متفاوتی پیشنهاد شده است: هلینگر و همکاران از فاصله هلینگر برای کشف تغییرات در چگالی، بعد یا جهت داده استفاده نموده‌اند [۱۲]. Xing و همکاران شباهت هندسی نقاط با استفاده از زاویه‌ی اصلی[۹] بین زوایای تانژانت محلی بین نقاط را محاسبه کرده‌اند [۲۳]. محاسبه‌ی فاصله هلینگر بین نقاط نیاز به تعداد بسیار زیاد داده داشته و محاسبه زاویه‌ی اصلی بین فضاهای تانژانت محلی مبتنی بر این فرض است که بعد ذاتی منیفلدها از قبل شناخته شده است، در حالی‌که مشخص نیست داده‌ها مربوط به کدام منیفلد هستند. در این مقاله رویکردی جهت تمایز نقاط داخلی منیفلدها از سایر نقاط با بهره‌گیری از رفتار لاپلاسین تابع در نزدیکی نقاط غیرداخلی پیشنهاد شده است که هر دو مسأله ذکر شده در آن مرتفع گشته است.

نتایج بدست آمده از مطالعات نظری در سال‌های اخیر نشان داده است که نقاط غیرداخلی شامل نقاط اشتراک منیفلدها، مرز و لبه‌ها جنبه‌های مهمی از داده هستند که در بسیاری از پژوهش‌ها در نظر گرفته نشده‌اند. برای نقطه‌ی x در همسایگی نقاط غیرداخلی، رابطه‌ی (۴) برقرار است:

$$\Gamma_t g(x) = \frac{1}{\sqrt{t}}\frac{\pi}{2}\partial_{\vec{n}}\, g(x) + O(\frac{1}{\sqrt{t}}) \qquad (۴)$$

که در آن $\partial_{\vec{n}}$ نرمال واحد به سمت خارج[۱۰] در نقطه‌ی x، مشتق جهت‌دار g در جهت $\partial_{\vec{n}}$ و g تابع قطعه‌ای هموار شرح داده شده در بخش ۳-۱ است.

بر طبق معادله بیان شده، از مرتبه‌ی است که بزرگ‌تر از در نقاط داخلی است که برای مقادیر کوچک t از مرتبه $O(1)$ است. بنابراین مقادیر بزرگ‌تر بیانگر نقاط نزدیک نقاط غیرداخلی است. علاوه بر آن، نشان داده شده که اگر داده‌ها از توزیع غیرنرمال نمونه‌برداری شده باشند، تحت فرضیات ضعیفی روی تابع چگالی احتمال، رفتار متفاوت تابع لاپلاسین در نقاط داخلی و غیرداخلی منیفلد همچنان برقرار است. در این مقاله از این ویژگی برای اصلاح همسایگی در گراف نزدیک‌ترین همسایگی استفاده می‌شود.

تعریف تابع g به‌دلیل عدم دانش قبلی در خصوص این‌که چه داده‌ای متعلق به کدام منیفلد است و نیز بعد بالای داده‌ی ورودی ساده نیست. هرچند، با توجه به فرض هموار بودن روی چند منیفلد، f یک تابع قطعه‌ای هموار است لذا پیشنهاد می‌گردد که با محاسبه‌ی لاپلاسین تابع f، نقاط داخلی از نقاط غیرداخلی تمیز داده شود. لذا W' می‌تواند به عنوان تابعی از f تعریف شود. ضریب اطمینانی به وزن‌های یال‌های گراف اولیه نسبت داده شده و W' به‌صورت زیر تعریف می‌شود:

$$w'_{ij}(f) = w_{ij} \times conf_{ij}(f),$$
$$conf_{ij}(f) = (c_{dec} - 1) \times H_{i,j}(f) + 1,$$
$$H_{i,j}(f) = \begin{cases} \text{if } L_t f(x_i) \text{ or } L_t f(x_j) \text{ belongs to} \\ 1 \quad \text{the top } n_{dec} \text{ data points with} \\ \quad \text{largest value of } L_t f \\ \quad 0 \quad otherwise \end{cases} \tag{۵}$$

که در آن $0 < c_{dec} < 1$ ضریب اطمینان، n_{dec} عدد از پیش تعریف شده بیانگر تعداد نقاط غیرداخلی و ماتریسی است که در آن اتصالات بین نقاط غیرداخلی با ۱ و اتصالات بین نقاط داخلی با ۰ مشخص شده است. برابر اطمینان به است که اگر x_i یا x_j نقطه غیرداخلی باشند برابر c_{dec} و در غیر این‌صورت برابر ۱ است.

براساس تعریف مذکور، مسأله‌ی بهینه‌سازی (۳) می‌تواند به صورت مسأله‌ی بهینه‌سازی یک مولفه‌ای ارائه گردد:

$$\min_{f \in H_k} \frac{1}{l} \sum_{i=1}^{l} V(f, x_i, y_i) + \gamma_A \| f \|_k^2$$
$$+ \gamma'_I \sum_{i,j=1}^{n} w'_{ij}(f) (f(x_i) - f(x_j))^2$$
Subject to
$$w'_{ij}(f) = w_{ij} \times conf_{ij}(f),$$
$$conf_{ij}(f) = (c_{dec} - 1) \times H_{i,j}(f) + 1,$$
$$H_{i,j}(f) = \begin{cases} \text{if } L_t f(x_i) \text{ or } L_t f(x_j) \text{ belongs to} \\ 1 \quad \text{the top } n_{dec} \text{ data points with} \\ \quad \text{largest value of } L_t f \\ \quad 0 \quad otherwise \end{cases} \tag{۶}$$
$$i, j = 1, \ldots, n$$

از آن‌جا که مسأله‌ی بهینه‌سازی مذکور محدب نیست به شیوه‌ای شبیه EM حل می‌شود، هر چند رویکرد پیشنهادی برای آن احتمالاتی نیست. متغیرهای مخفی عناصر ماتریس H هستند. الگوریتم پیشنهادی، دو مرحله‌ی نسبت دادن نقاط به نقاط داخلی یا غیرداخلی در گام E و بازتخمین تابع دسته‌بند در گام M را متناوباً تکرار می‌کند.

در گام E، به‌جای در نظرگرفتن مقادیر میانگین روی توزیع متغیرهای مخفی، مقدار متغیرهای مخفی براساس مقدار کنونی تابع f، در نظر گرفته می‌شود.

$$Q(f, f') = \frac{1}{l} \sum_{i=1}^{l} V(f, x_i, y_i) + \gamma_A \| f \|_k^2 \tag{۷}$$
$$+ \gamma'_I \sum_{i,j=1}^{n} w'_{ij}(f) (f(x_i) - f(x_j))^2$$

در حقیقت، در این مرحله در نظر گرفته شده و W' براساس رابطه‌ی (۵) محاسبه می‌شود، در حالی‌که نقاط غیرداخلی، نقطه‌ی با بزرگ‌ترین مقدار لاپلاسین محاسبه شده براساس رابطه‌ی (۲) است. f'

در گام M، f، تابع دسته‌بند جدیدی که مسأله‌ی بهینه‌سازی زیر را کمینه می‌کند محاسبه می‌شود.

$$obj(f) = \frac{1}{l} \sum_{i=1}^{l} V(f, x_i, y_i) + \gamma_A \| f \|_k^2 \tag{۸}$$
$$+ \gamma'_I \sum_{i,j=1}^{n} w'_{ij}(f') (f(x_i) - f(x_j))^2$$

مسأله‌ی بهینه‌سازی فوق محدب بوده و راه‌حل آن به فرم زیر است [۱]:

$$f^*(x) = \sum_{i=1}^{n} \beta_i K(x_i, x) \tag{۹}$$

که در آن، هسته‌ی مرتبط با RKHS و با توجه به نوع تابع هزینه محاسبه می‌شود. برای تابع هزینه‌ی حداقل مربعات، به صورت زیر محاسبه می‌شود [۱]:

$$\beta = (JK + \gamma_A l I + \frac{l \gamma_{I'}}{n^2} L' K)^{-1} Y \tag{۱۰}$$

که در آن، $K_{n \times n}$ ماتریس گرام روی کل داده‌های برچسب‌دار و بدون برچسب، Y یک بردار برچسب بعدی $Y = [c_1, c_2, \ldots, c_l, 0, \ldots, 0]$، J یک ماتریس قطری به صورت است. L' ماتریس لاپلاسین گراف ساخته شده با ماتریس مجاورت است.

الگوریتم پیشنهادی در شکل ۱ نشان داده شده است. جهت مقداردهی اولیه‌ی f سه گزینه پیشنهاد می‌شود: (۱) برچسب‌های دردسترس برای داده‌های برچسب‌دار و صفر برای داده‌های بدون برچسب، (۲) نتیجه‌ی دسته‌بندی منظم‌سازی منیفلد [۱] و (۳) برچسب‌گذاری تصادفی. f با نسبت دادن هر کدام از این سه گزینه مقداردهی اولیه شده و گام ۲ الگوریتم به‌صورت مستقل برای هر کدام اجرا می‌شود.

۴- آزمایش‌ها

الگوریتم پیشنهادی با روش‌های MR [۱] و EMR [۱۱] که به ترتیب روش‌های دسته‌بندی نیمه نظارتی مبتنی بر فرض قرارگیری داده روی یک منیفلد و ترکیبی از چند منیفلد هستند مقایسه شده است. ارزیابی روی مجموعه داده‌های مصنوعی و واقعی انجام شده است. مقدار پارامتر n_{dec} از مجموعه‌ی انتخاب شده است که در آن $n_{Adaptive}$ برابر با تعداد نقاطی است که در هر تکرار، مقدار لاپلاسین تابع آن‌ها غیرصفر است.

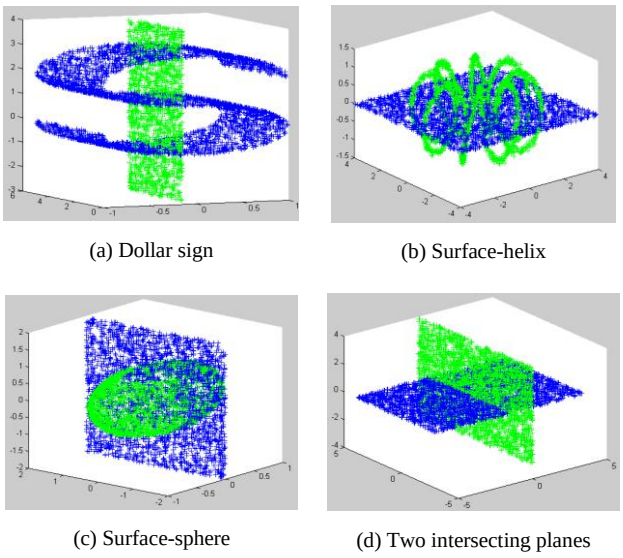
ورودی: $c_{dec}, n_{dec}, C_l, \chi = \{x_i\}_{i=1}^{l+u}$
خروجی: f
گام ۱: f را مقدار دهی اولیه کن.
گام ۲: گام‌های زیر را تا زمانی که تابع هدف (۳) کاهش نمی‌یابد یا تا زمان رسیدن به عدد از پیش تعیین شده‌ی بیشینه تکرارها، تکرار کن:
۱- لاپلاسین f را با رابطه‌ی (۲) محاسبه کن.
۲- تعداد n_{dec} نقطه‌ی با بیشترین مقدار لاپلاسین تابع را به عنوان نقاط غیرداخلی مشخص کن.
۳- W' را با رابطه‌ی (۵) محاسبه کن.
۴- f^* را با رابطه‌ی (۹) محاسبه کن.
۵- $f = f^*$.

شکل ۱- گام‌های الگوریتم پیشنهادی

۴-۱- مجموعه داده‌های مصنوعی

در این بخش، نتیجه‌ی آزمایش روی چهار مجموعه داده‌ی مصنوعی سه بعدی آمده است. شکل ۲ مجموعه داده‌های مصنوعی را نشان می‌دهد: (۱) مجموعه داده‌ی علامت دلار شامل منیفلد "S" که با منیفلد "|" اشتراک دارد، (۲) مجموعه داده‌ی سطح - مارپیچ شامل منیفلدهای مارپیچ ۱ بعدی و سطح، (۳) منیفلدهای متقاطع سطح -کره و (۴) دو سطح متقاطع.

در این مجموعه داده‌ها، مقدار پارامترهای K، t و عرض هسته به ترتیب برابر ۵، ۱۰۰ و ۰٫۶ در نظر گرفته شده است. ۲۰۰۰ داده از هر مجموعه داده تولید شده است و تعداد داده‌های برچسب‌دار برابر ۴۰ است که به‌طور یکسان بین داده‌های هر دسته تقسیم شده است. میانگین خطا و انحراف معیار ۱۰ بار اجرا در جدول ۲ آمده است. همان‌طور که مشخص است الگوریتم پیشنهادی بهتر از سایر الگوریتم‌ها عمل می‌کند.



(a) Dollar sign (b) Surface-helix

(c) Surface-sphere (d) Two intersecting planes

شکل ۲- مجموعه داده‌های مصنوعی

جدول ۲- نرخ خطا و انحراف معیار روی مجموعه داده‌های مصنوعی

مجموعه داده	خطا به درصد (انحراف معیار)		
	MR	EMR	روش پیشنهادی
Dollar sign	۲۲٫۸۸ (۳٫۳۴)	۲۵٫۷۰ (۳٫۴۰)	**۲۰٫۴۱ (۲٫۷۹)**
Surface-helix	۳۰٫۱۹ (۵٫۷۷)	۳۳٫۳۱ (۵٫۹۲)	**۲۸٫۶۷ (۴٫۵۱)**
Surface-sphere	۱۱٫۳۹ (۳٫۷۸)	۱۲٫۱۱ (۳٫۹۶)	**۱۰٫۳۲ (۳٫۳۳)**
Two intersecting planes	۴٫۷۶ (۱٫۵۵)	۶٫۱۴ (۲٫۱۰)	**۴٫۴۵ (۱٫۴۶)**

۴-۲- مجموعه داده‌های واقعی

در این بخش نتیجه‌ی ارزیابی روی چند مجموعه داده‌ی شناخته شده‌ی واقعی آمده است: BCI [۱۴]، Sonar و Ionosphere از مجموعه داده‌های UCI، WebKB[11]، مجموعه داده‌ی دسته‌بندی متن و CBCL [۴]، مجموعه داده‌ی دسته‌بندی چهره. این مجموعه داده‌ها بارها در ارزیابی روش‌های نیمه نظارتی بکاررفته است [۵، ۹، ۲۱]. مشخصات این مجموعه داده‌ها در جدول ۳ آمده است.

تعداد همسایه‌ها در گراف همسایگی در هر گراف برابر با حداقل تعدادی که گراف را متصل می‌سازد تنظیم شده است و این تعداد در گراف‌های اولیه‌ی بکار رفته در EMR نیز در نظر گرفته شده است تا مقایسه‌ی انجام شده عادلانه باشد. روش پیشنهادی با روش M2SGMM نیز که روشی با فرض قرارگیری داده روی چند منیفلد است و از معیار زاویه بین فضای تانژانت نقاط برای تعیین وزن یال‌های گراف استفاده می‌کند مقایسه شده است [۲۳]. میانگین دقت دسته‌بندی و انحراف معیار ۱۰ بار اجرا به ازای برچسب‌گذاری تصادفی ۱۰ درصد از داده‌ها در جدول ۴ آمده است.

جدول ۳- ویژگی‌های مجموعه داده‌های واقعی

مجموعه داده	تعداد داده	تعداد ابعاد
CBCL	۳۰۰۰	۳۶۱
WebKB (Page)	۱۰۵۱	۳۰۰۰
BCI	۴۰۰	۱۱۷
Ionosphere	۳۵۱	۳۴
Sonar	۲۰۸	۶۰

جهت ارزیابی دقیق‌تر، دقت دسته‌بندی روش‌های مختلف به ازای درصد متفاوت داده‌های برچسب‌دار در شکل ۳ با یکدیگر مقایسه شده است.در ادامه، نتایج ارائه شده به صورت جزئی‌تر بررسی می‌شود.

جدول ۴- نرخ خطا و انحراف معیار روی مجموعه داده‌های واقعی

مجموعه داده	خطا به درصد (انحراف معیار)			
	MR	M2SGMM	EMR	روش پیشنهادی
CBCL	۷٫۷۵ (۰٫۷۱)	۹٫۰۷ (۰٫۰۱)	۷٫۳۳ (۰٫۹۶)	**۶٫۸۸ (۰٫۵۷)**
WebKB (Page)	۳٫۵۶ (۱٫۳۴)	۵٫۰۸ (۰٫۰۲)	۴٫۲۳ (۰٫۵۹)	**۳٫۴۹ (۱٫۳۲)**
WebKB (Page+Link)	۳٫۸۲ (۱٫۶۱)	۳٫۱۳ (۰٫۷۷)	۳٫۳۲ (۱٫۲۹)	**۲٫۸۹ (۱٫۳۲)**
BCI	۴۲٫۱۵ (۲٫۰۸)	۴۷٫۸۳ (۳٫۰۶)	۰۰۴۲. (۲٫۳۱)	**۴۱٫۱۵ (۲٫۲۲)**
Sonar	۳۳٫۴۶ (۴٫۷۹)	۳۸٫۹۹ (۴٫۹۸)	۳۱٫۶۸ (۴٫۰۳)	**۳۱٫۶۸ (۴٫۰۳)**
Ionosphere	۱۵٫۱۲ (۳٫۶۴)	۱۵٫۱۴ (۰٫۰۴)	۱۹٫۵۴ (۳٫۸۱)	**۱۴٫۳۶ (۳٫۱۲)**

- **تحلیل نتایج**

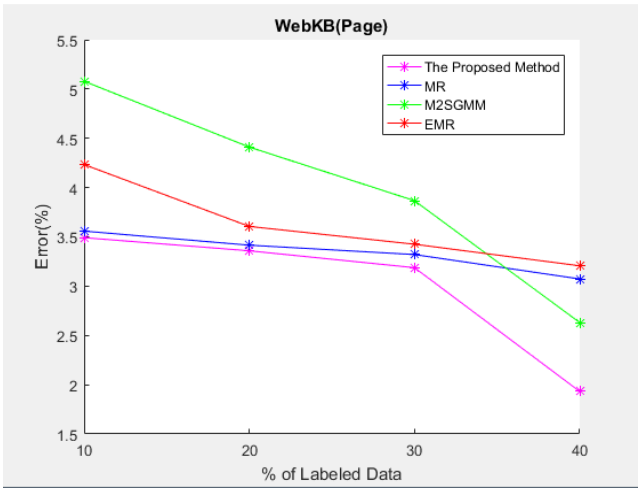
نتایج بدست آمده حاکی از دقت مناسب روش پیشنهادی نسبت به سایر رویکردها است؛ جهت تحلیل دقیق‌تر برتری‌های روش پیشنهادی توجه به نکات زیر حائز اهمیت است:

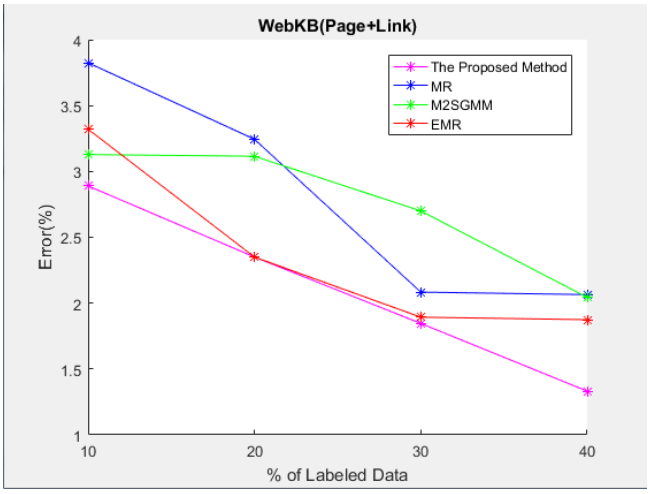
(۱) در روش M2SGMM فرضیات محدود کننده‌ی زیر در نظر گرفته شده است:

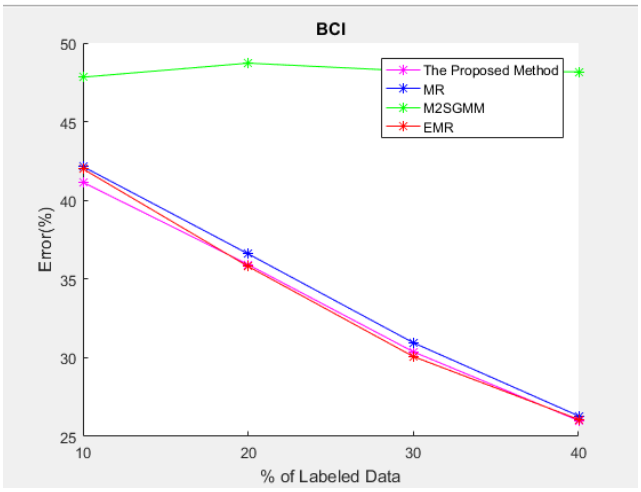
- تمام منیفلدها بعد ذاتی یکسان برابر d_{in} را دارند که فرضی واقعی نیست.
- قابل اعمال به داده‌های با ابعاد بالا به صورت مستقیم نبوده و ابتدا گام پیش‌پردازشی کاهش بعد روی داده‌ها اعمال می‌شود؛ از آن‌جا که زاویه بین فضاهای تانژانت نقاط نزدیک به هم خیلی شبیه به یکدیگر است، زاویه بین فضاهای تانژانت بین بلوک‌های داده (و نه تمام داده‌ها) که منطبق با بلوک‌های محلی داده‌ی بدست آمده از مدل‌های آمیخته گاوسی نیمه نظارتی است محاسبه می‌شود. پردازش داده‌ها به مدل‌های آمیخته‌ی گاوسی نیمه نظارتی در داده‌های با ابعاد بالا به دلیل تعداد زیاد بهینه‌های محلی در عمل کارایی لازم را ندارد لذا ابتدا داده‌ها با روش PCA کاهش بعد داده شده و سپس زاویه‌ی بین فضاهای تانژانت مولفه‌های گاوسی محاسبه می‌شود.

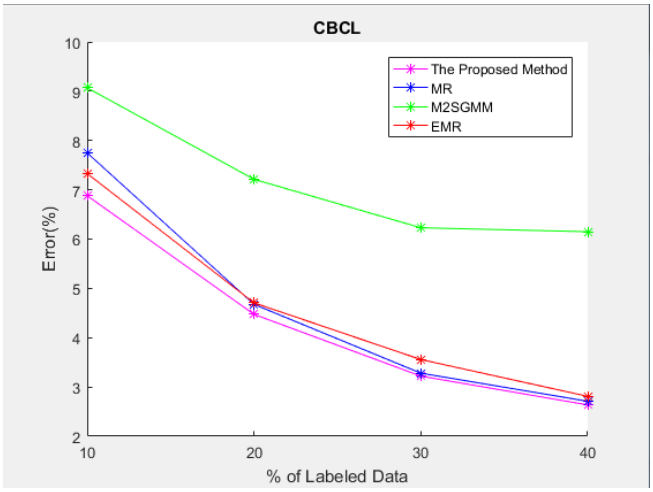
روش کاهش بعد انجام شده، تعداد ابعاد کاهش یافته و نیز تعداد ابعاد ذاتی منیفلدها از پارامترهای تاثیرگذار هستند که روش ساده‌ای جهت برآورد آن‌ها وجود ندارد. این پارامترها در روش پیشنهادی تاثیرگذار نیستند. علاوه بر آن، اعمال روش کاهش بعد غیرخطی به داده‌هایی که روی منیفلدهای متقاطع قرار دارد مناسب نیست.

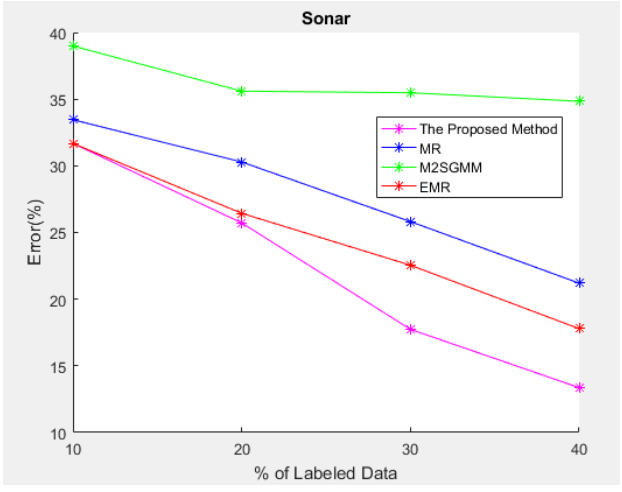
(۲) در روش EMR، فرض بر این است که داده‌ها از ترکیب محدب چند منیفلد از پیش تعیین شده نمونه‌برداری شده و ضرایب ترکیب محدب برای تمام داده‌ها یکسان در نظر گرفته شده است در صورتی که انتظار می‌رود این ضرایب برای داده‌های منیفلدهای گوناگون و نیز نقاط تقاطع منیفلدها متفاوت باشد؛ روش پیشنهادی با تفکیک نقاط داخلی از غیرداخلی این فرض را به نحو مناسب‌تری پوشش می‌دهد.

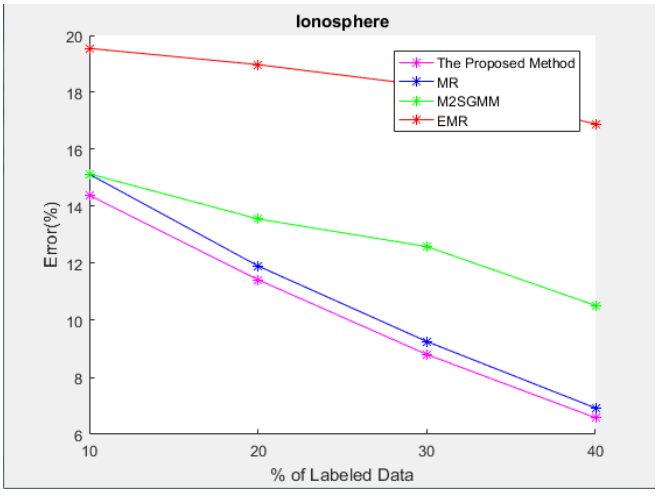












شکل ۳- دقت دسته‌بندی به ازای درصد متفاوت داده‌های برچسب‌دار

ادامه شکل ۳- دقت دسته‌بندی به ازای درصد متفاوت داده‌های برچسب‌دار

۵- نتیجه‌گیری

در این مقاله، یک روش نیمه نظارتی جدید مناسب برای داده‌هایی که روی منیفلدهای متقاطع قرار دارند ارائه شد. روش ارائه شده بر مبنای تئوری موجود در زمینه‌ی تمایز رفتار لاپلاسین تابع هموار روی هر منیفلد در نقاط داخلی و غیرداخلی منیفلدها است و با تعریف ضریب اطمینان برای وزن یال‌های گراف اولیه بر مبنای مقدار لاپلاسین تابع، الگوریتمی با بهره‌گیری از این تئوری ارائه شده است. نتایج بدست آمده حاکی از مناسب بودن این رویکرد جهت دسته‌بندی منیفلدهای متقاطع است. جهت ادامه‌ی تحقیقات، کاربرد رویکرد پیشنهادی در کاربردهای نیمه نظارتی مدنظر خواهد بود.

تشکر و قدردانی

بدینوسیله از آقای Xianglei Xing جهت در اختیار قرار دادن کد الگوریتم M2SGMM قدردانی می‌شود.

مراجع

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

[2] M. Belkin, Q. Que, Y. Wang, and X. Zhou, "Toward Understanding Complex Spaces: Graph Laplacians on Manifolds with Singularities and Boundaries," in *The 25th Annual Conference on Learning Theory (COLT 2012)*, Edinburgh, Scotland, 2012.

[3] K. M. Carter, R. Raviv, and A. O. Hero, "On local intrinsic dimension estimation and its applications," *IEEE Transactions on Signal Processing*, vol 58, no. 2, pp. 650-663, 2010.

[4] S. Chen, S. Li, S. Su, Q. Tian, and R. Ji, "Online MIL tracking with instance-level semi-supervised learning," *Neurocomputing 139*, pp. 272-288, 2014.

[5] M. Z. X. L. Z. Z. Z. B. H. Fan, "A Regularized Approach for Geodesic Based Semi-Supervised Multi-Manifold Learning," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2133-2147, 2014.

[6] Q. Gao, Y. Huang, X. Gao, W. Shen, and H. Zhang, "A novel semi-supervised learning for face recognition," *Neurocomputing 152*, pp. 69-76, 2015.

[7] B. Geng, D. Tao, C. Xu, Y. Yang, and X.-S. Hua, "Ensemble manifold regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, no. 6, pp. 1227-1233, 2012.

[8] A. B. Goldberg, X. Zhu, A. Singh, Z. Xu, and R. Nowak, "Multi-manifold semi-supervised learning," *in International Conference on Artificial Intelligence and Statistics*, 2009.

[9] M. H. Rohban, and H. R. Rabiee, "Supervised neighborhood graph construction for semi-supervised classification," *Pattern Recognition 45*, no. 4, pp. 1363-1372, 2012.

[10] Y. Wang, S. Chen, H. Xue, and Z. Fu, "Semi-supervised classification learning by discrimination-aware manifold regularization," *Neurocomputing 147*, pp. 299-306, 2015.

[11] Y. Wang, Y. Jiang, Y. Wu, and Z.-H. Zhou, "Spectral clustering on multiple manifolds," *IEEE Transactions on Neural Networks 22*, no. 7, pp. 1149-1161, 2011.

[12] X. Xing, Y. Yu, H. Jiang, and S. Du, "A multi-manifold semi-supervised Gaussian mixture model for pattern classification," *Pattern Recognition Letters 34*, no.16, pp. 2118-2125, 2013.

[13] W. Yang, C. Sun, and L. Zhang, "A multi-manifold discriminant analysis method for image feature extraction," *Pattern Recognition 44*, no. 8, pp. 1649-1657, 2011.

[14] B. Li, D. Huang, and C. L. K. Wang, "Feature extraction using constrained maximum variance mapping," *Pattern Recognition. 2008 Nov 1;41(11)*, pp. 3287-94, 2008.

[15] L. Ma, M. C. Melba, Y. Xiaoquan, and G. Yan, "Local-manifold-learning-based graph construction for semisupervised hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing 53*, no. 5, pp. 2832-2844, 2015.

[16] D. Meng, Y. Leung, T. Fung, and Z. Xu, "Nonlinear dimensionality reduction of data lying on the multicluster manifold," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 38*, no. 4, pp. 1111-1122, 2008.

[17] Y. Su, S. Li, S. Wang, and Y. Fu, "Submanifold decomposition," *IEEE Transactions on Circuits and Systems for Video Technology, 24(11)*, pp. 1885-1897, 2014.

[18] J. Wang, H. Bensmail, N. Yao, and X. Gao, "Discriminative sparse coding on multi-manifolds," *Knowledge-Based Systems. 2013 Dec 1;54*, pp. 199-206, 2013.

[19] H.-B. Huang, H. Hong, and F. Tao, "Hierarchical manifold learning with applications to supervised classification for high-resolution remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing 52.3 (2014)*, pp. 1677-1692, 2014.

[20] T. N. Lal, S. Michael, H. Thilo, W. Jason, B. Martin, B. Niels, and S. Bernhard, "Support vector channel selection in BCI," *IEEE transactions on biomedical engineering 51*, no. 6, pp. 1003-1010, 2004.

[21] S. de, A. R. Celso, O. R. Solange, and E. B. Gustavo, "Influence of graph construction on semi-supervised learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Heidelberg, 2013.

[22] W. Chen, Y. Shao, N. Deng, and Z. Feng, "Laplacian least squares twin support vector machine for semi-supervised classification," *Neurocomputing. 2014 Dec 5;145*, pp. 465-76, 2014.

[23] P. PA, "CBCL Face Database," USA. MIT Center For Biological and Computation Learning, 2001.

[24] O. Chapelle, B. Schölkopf, and Z. Alexander, Semi-supervised learning, Cambridge: MIT press, 2006.

زهره کریمی در رشته مهندسی کامپیوتر در گرایش نرم‌افزار در سال ۱۳۸۵ از دانشگاه شهید بهشتی فارغ‌التحصیل شد و مدرک کارشناسی‌ارشد خود را در سال ۱۳۸۸ از دانشگاه صنعتی شریف در همان رشته اخذ کرد. او هم‌اکنون در مقطع دکتری در رشته هوش مصنوعی در دانشگاه صنعتی امیرکبیر مشغول به تحصیل است. زمینه‌های پژوهش وی شامل یادگیری ماشین و داده‌کاوی است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

z_karimi@aut.ac.ir

سعید شیری قیداری استاد دانشکده علوم کامپیوتر دانشگاه صنعتی امیرکبیر هستند. ایشان مدرک کارشناسی خود را در رشته مهندسی الکترونیک و کارشناسی‌ارشد 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.

آدرس پست‌الکترونیکی ایشان عبارت است از:

shiry@aut.ac.ir

محمد رحمتی استاد دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیرکبیر می‌باشند. ایشان دکترای خود را در سال ۱۳۷۳ از دانشکاه کنتاکی آمریکا در رشته مهندسی برق و کامپیوتر اخذ نموده‌اند. مسئولیت‌های مختلف ایشان در دانشکده شامل ریاست دانشکده، مدیر تحصیلات تکمیلی، و معاون پژوهشی دانشکده بوده است. ایشان در حال حاضر مدیر گروه هوش مصنوعی دانشکده مهندسی کامپیوتر و فناوری اطلاعات می‌باشند. زمینه تحقیقاتی مورد علاقه ایشان یادگیری ماشین و شناسایی الگو، پردازش تصویر و بینایی ماشین می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

rahmati@aut.ac.ir

روح‌الله رمضانی در رشته آمار در سال ۱۳۸۳ از دانشگاه بین‌المللی امام خمینی (ره) فارغ‌التحصیل شد و مدرک کارشناسی‌ارشد 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، هم‌اکنون عضو هیئت علمی دانشگاه دامغان است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

r_ramezani@du.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۶/۰۳/۱۶

تاریخ اصلاح: ۱۳۹۶/۱۲/۰۱

تاریخ قبول شدن: ۱۳۹۷/۰۲/۰۳

نویسنده مرتبط: زهره کریمی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران.

[1]Local
[2]Expectation-Maximization
[3]Orientation
[4]Ambient Space
[5]Reproducing Kernel Hilbert Space
[6]Hinge
[7]Ambient
[8]Intrinsic
[9]Principal
[10]Outward
[11]http://vikas.sindhwani.org/manifoldregularization.html.