فضای جدید مبتنی بر انطباق منحنی جهت طبقه بندی داده ها

مصطفی قاضی زاده
دانشگاه فردوسی مشهد

هادی صدوقی یزدی
دانشگاه فردوسی مشهد

محمود نقیب زاده
دانشگاه فردوسی مشهد

naghibzadeh@um.ac.ir,sadoghi@sttu.ac.ir, GhazizadehAhsaee.most@stu-mail.um.ac.ir

چکیده: در این مقاله یک فضای جدید مبتنی بر انطباق منحنی(CFS)[1]، جهت نگاشت داده های جداپذیر غیر خطی به داده های جداپذیر خطی ارائه گردیده است. نگاشتهای کوادراتیک یا خطی ، داده ها را به فضای جدید به گونه ای نگاشت می دهند که طبقه بندی داده ها بهتر انجام پذیرد. اساس روش ارائه شده ، انطباق منحنی، سطح و یا حجم بر روی داده های تعلیم می باشد. در این روش خط ، سطح و یا حجم منطبق شده ، به عنوان محورهای فضای جدید در نظر گرفته می شوند. در انتها با استفاده از طبقه بند خطی Adaline، طبقه بندی انجام گرفت که نتایج با یک طبقه بند غیر خطی SVM مقایسه شده است و نتایج از خوب بودن روش ارائه شده حکایت دارد.

واژه های کلیدی: فضای مبتنی بر انطباق منحنی، فضای مبتنی بر فاصله، انطباق، طبقه بندی

۱- مقدمه

طبقه بندی یکی از زمینه های تحقیقاتی مهم و پر کاربرد می باشد. توابع تفکیک کننده خطی (NDF)[2] که برای تشخیص الگوهای مشخصی تعلیم داده می شوند ، به عنوان یکی از روشهای پر کاربرد در زمینه طبقه بندی استفاده می شود. شبکه عصبی (NN) و ماشین های بردار پشتیبان (SVM) از ابزارهای NDF می باشند. SVM [۱] در تعلیم سیستمها بسیار قوی می باشد و در بین محققان یکی از ابزارهای شناخته شده و محبوب به شمار می آید، چرا که از ماشینهای کرنل در خطی سازی استفاده می کند و خصوصیات عمومی سازی خوبی فراهم می کند.

نتیجه استفاده از کرنل ها این است که الگوریتم در فضای تبدیل یافته ، می تواند یک ابرسطح با ماکزیمم حاشیه را منطبق کند. این تبدیل ممکن است غیر خطی بوده و فضای جدید ممکن است تعداد ابعاد زیادی داشته باشد . گرچه این طبقه بند یک ابرسطح در فضای تبدیل یافته با تعداد ابعاد زیاد می باشد، اما ممکن است در فضای ورودی اصلی غیر خطی باشد.به عنوان مثال اگر کرنل مورد استفاده RBF گوسی باشد ، فضای تبدیل یافته ، فضای هیلبرت خواهد بود که دارای بینهایت بعد خواهد بود. طبقه بندهای مبتنی بر مارجین به خوبی منظم شده اند به گونه ای که بینهایت بعد تاثیر منفی بر روی جواب نخواهند داشت.

طبیعتا روشهای مبتنی بر کرنل [۲-۵] نگاشتی از فضای ورودی به فضایی با تعداد ابعاد زیاد(HDF)[3] می باشند که در خطی سازی می توانند مفید باشند. برخی کرنل ها به همین منظور ارائه شده اند مثل کرنل چند جمله ای، گوسی و spline [۶]، اما این کرنل ها هیچ تضمینی بر خطی سازی در فضای با تعداد ابعاد بالا ندارند. این مساله ما را به این سمت می کشاند تا به دنبال ارائه یک فضای جدید باشیم که در آن با استفاده از طبقه بند خطی بتوان الگوها را از هم جدا کرد که ما نام آن را فضای مبتنی بر انطباق منحنی می نامیم (CFS) چرا که از مفهوم انطباق منحنی برای ایجاد فضای جدید بهره گرفته ایم.

رویکرد ارائه شده در مقایسه با روشهای مبتنی برکرنل ، گسترش فضای واقعی به فضای با تعداد ابعاد زیاد را در نظر ندارد بلکه نگاشت از فضای n بعدی به فضای با m بعد می باشد که m تعداد کلاسها است.

در ادامه مقاله به موضوعات زیر پرداخته می شود. در بخش ۲ روش ارائه شده مطرح می گردد و در بخش ۳ با ذکر نمونه داده هایی به تشریح چگونگی عملکرد روش ارائه شده می پردازیم . در بخش ۴ جهت مقایسه کارخود ، روش ارائه شده را بر روی داده های واقعی اعمال کرده و در فضای CFS با استفاده از یک طبقه بند خطی داده ها را طبقه بندی کرده ایم، سپس همان داده های فضای ورودی را با استفاده از یک طبقه بند غیر خطی (Multi class SVM) طبقه بندی نموده ایم و در نهایت نتایج مربوط به میزان دقت[4] در هر دو طبقه بندی را مقایسه کرده ایم و در فصل ۵ به نتیجه گیری پرداخته ایم.

۲- فرموله سازی فضای انطباق مبتنی بر منحنی

تعاریف:

$$\{x_{ij} \in [x_{11}, x_{21},..., x_{k_1 1}, x_{12},..., x_{k_2 2},...,$$
$$x_{1j}, x_{2j},..., x_{k_j j}], j = 1,...,m\}$$

نمونه i آم از کلاس j ام با n بعد می باشد.

---

[1] Curve Fitting Space

[2] Nonlinear Discriminant Functions

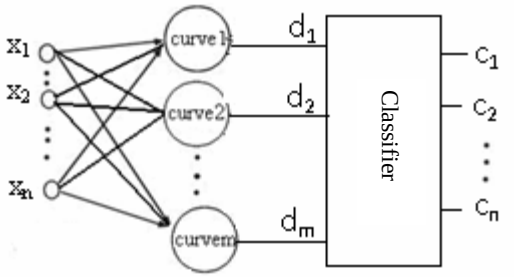
[3] High Dimensional Feature

[4] Accuracy

Cj منحنی ، ابر صفحه و یا فضای منطبق شده بر مجموعه $\{(X_{ij}=,i=1,\ldots,k_j)\}$است که yj برچسب j ام داده های تعلیم بوده و k_j تعداد نمونه های با برچسب j می باشد. به طور کلی نگاشت از فضای با m الگو(کلاس) از داده های n بعدی ، به فضای m بعدی به صورت زیر انجام می شود:

$$\varphi: X = \{x_1, x_2, \ldots, x_n\} \in \mathcal{R}^n \rightarrow \varphi(X) \in \mathcal{R}^m$$

که φ تابعی است که به فاصله بین داده ها تا منحنی ، ابر صفحه و یا حجم منطبق شده بر الگوها وابسته است. این تبدیل در شکل ۱ قابل مشاهده است که X الگویی است در فضای ورودی و $D = \{d_1, d_2, \ldots, d_m\}$ یک عنصر از فضای تبدیل یافته و c_i کلاس i ام می باشد.



شکل۱: طبقه بندی با استفاده از انطباق منحنی

۲-۱ نحوه عملکرد در برخورد با ابر صفحه

زمانی که یک ابر صفحه بر مجموعه ای از داده ها منطبق می شود، ، منظور از فاصله(d)،فاصله نقطه $X_i = \{x_{1i}, x_{2i}, \ldots, x_{ni}\}$ از آن سطح می باشد.

۲-۲ نحوه عملکرد در برخورد با ابر کره

زمانی که یک ابر کره (C) بر مجموعه داده ها منطبق گردیده است فاصله یک نقطه از پوسته ابرکره محاسبه می شود.

۲-۳ نحوه عملکرد در برخورد با منحنی های چند جمله ای در فضای دو بعدی

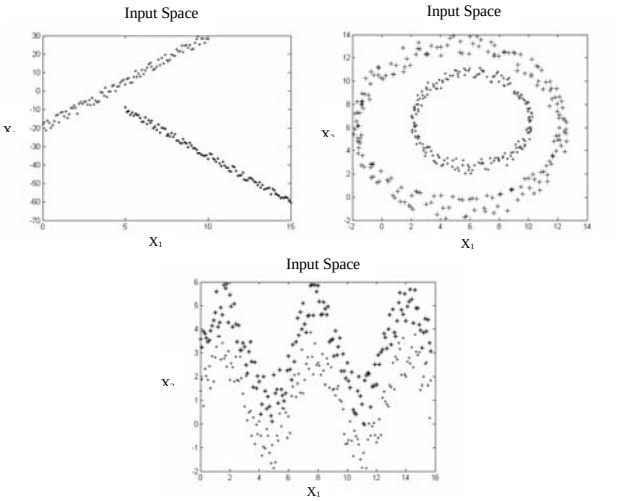
اگر در فضای دو بعدی یک منحنی چند جمله ای به مجموعه ای از داده ها منطبق شود، جهت به دست آوردن فاصله مینیمم عمومی ، ابتدا تمامی مینیمم ها محاسبه شده و سپس با جایگذاری مقادیر به دست آمده ، چک می شود که کدام یک مینیمم عمومی است و آن انتخاب می شود.

۲-۴ استفاده از SVDD برای منطبق کردن فضا به دور داده ها

در استفاده از SVDD[۷] ، به ازای هر کلاس از مجموعه داده، از یک حجم که توسط SVDD به دست می آید بهره جسته ایم. به عنوان مثال اگر دو کلاس داده داشته باشیم دو SVDD استفاده می کنیم. سپس به منظور چک کردن اینکه یک داده به یک کلاس متعلق می باشد یا خیر، فاصله آن را تا پوسته بیرونی حجم مربوطه محاسبه کرده و سپس این فاصله را به عنوان معیار جهت طبقه بندی استفاده کرده ایم. به عبارت دیگر یک داده به یک کلاس متعلق است اگر فاصله آن داده در فضای کرنل به آن کلاس کمتر از فاصله آن تا بقیه کلاسها باشد.

۳- نتایج تجربی

در این بخش ابتدا نحوه تبدیل با استفاده از منطبق کردن دایره بر روی داده های ساده نشان داده می شود و در انتهای این بخش روش ارائه شده با داده های واقعی تست می شود. در ادامه این بخش فرض بر این است که برچسب کلاسها از قبل مشخص است. پیاده سازی مسائل فوق در زبان برنامه نویسیMatlab انجام گرفته است.



شکل۲ : کلاسهای دایره ای *(b)* کلاسهای خطی *(a)*

تابع سینوسی *(c)*

۳-۱ نحوه عملکرد برای داده های با توزیع ساده

ابتدا روش ارائه شده بر روی یکسری داده های با توزیع ساده توصیف می شود.داده های مورد استفاده، در شکل۲ (a-c) نشان داده شده اند.

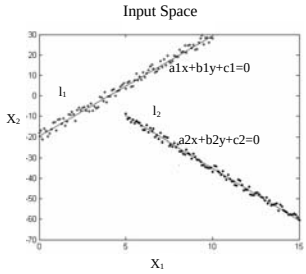
در شکل ۲(a) دو کلاس با شکل خطی نشان داده شده است که به صورت خطی جداپذیرند. اما در شکلهای دیگر(b-c) کلاسها جداپذیر خطی نمی باشند و ما آنها را به دو دسته تقسیم کرده ایم:۱)شکل مبتنی بر دایره وکره(b) ۲)توابع سینوسی (c)

شکل ۲(a) را درنظر بگیرید که شامل دو کلاس خطی می باشد.در ابتدا با استفاده از روش منطبق سازی با کمترین مربعات خطا ، به هر کلاس یک خط منطبق می شود. که نتیجه در شکل ۳ قابل مشاهده است.

سپس برای نگاشت آنها به فضای جدید از فرمولی مثل فرمول زیر جهت محاسبه فاصله هرنقطه(x_0, y_0)تاهرخط بهره گرفته می شود.در فرمول زیر$d(X_0,l_1)$ فاصله بین نقطه $X_0= (x_0, y_0)$ تا خطd_1 می باشد.

$$d(X_.,l_1)=\frac{a_1x_.+b_1y_.+c_1}{\sqrt{a_1^2+b_1^2}}$$

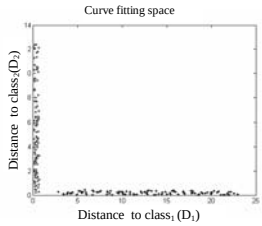
$$d(X_.,l_2)=\frac{a_2x_.+b_2y_.+c_2}{\sqrt{a_2^2+b_2^2}}$$



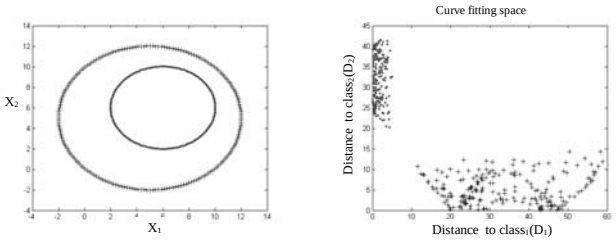
شکل ۳: انطباق خط بر کلاسهای داده

از فواصل به دست آمده توسط دو فرمول فوق برای نگاشت داده ها به فضای دو بعدی ،مانند آنچه در شکل ۴ نشان داده شده است، استفاده شده است و همان طور که در شکل مشخص می باشد این نگاشت منجر به فضایی می شود که داده های فوق در آن به صورت خطی جداپذیر می باشند.

شکل ۲(b) را به عنوان مثال دوم در نظر بگیرید.این شکلها مثالی از ابر کره در فضای دو بعدی می باشند.در اینجا ابتدا یک منحنی دایره شکل به کلاس منطبق می گردد. سپس فاصله بین منحنی دایره ای و هر داده محاسبه می شود. نتایج درشکل ۵ (a و b) مشاهده می شود. محور افقی، فاصله بین هر داده ها و کلاس داخلی(دایره قرمز رنگ داخلی) را نشان می دهد و محور عمودی، فاصله بین هر داده و کلاس خارجی (دایره آبی رنگ خارجی) می باشد.



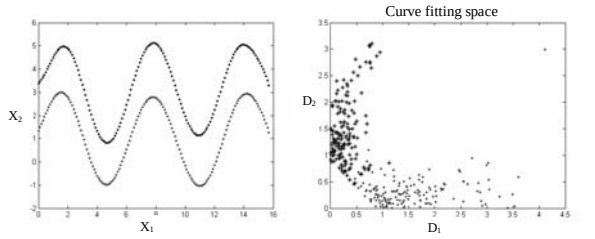
شکل ۴: نگاشت شکل ۳ به فضای ارائه شده



شکل ۵ : خروجی نگاشت *b)* دایره های منطبق شده *a)*

۳-۲ تبدیل فضا برای کلاسهای پیچیده تر

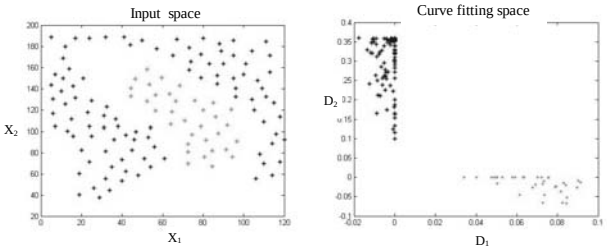
توزیع داده های کلاسها ممکن است پیچیده به نظر برسد، اما روش ارائه شده سعی می کند به سادگی آنها را به فضای جدیدی ببرد که در آن فضا داده های کلاسها به صورت خطی جداپذیر باشند (شکل ۶).



شکل ۶: انطباق منحنی به داده های هرکلاس و خروجی

۳-۴ تبدیل فضا با استفاده از SVDD

همان طورکه قبلا بیان شد، SVDD یک طبقه بند یک کلاسه می باشدکه در طبقه بندی داده های حجمی کاربرد دارد. در اینجا مجموعه داده ای در شکل۷(a) نشان داده شده است که توسط روش بیان شده برروی داده هایش نگاشت انجام می شود. خروجی در شکل۷(b) آورده شده است.



شکل ۷: نتیجه نگاشت به فضای ارائه شده با استفاده از *SVDD*

داده ها در فضای ارائه شده *b)* **داده های ورودی *a)***

۴- کار با مجموعه داده های واقعی

در این بخش روش ارائه شده ، برای مجموعه داده های واقعی به کار برده شد تا نتایج واقعی تری از طبقه بندی با استفاده از روش فوق به نمایش گذاشته شود. مجموعه داده های انتخابی،از [۸] گرفته شده اند و شامل داده های breast-cancer-wisconsin، Iris ، Ionosphere، Wine با به ترتیب ۱۱،۵،۳۵و ۱۵ویژگی وبا ۲،۳،۲و ۳ کلاس از داده ها می باشند.

۴-۱ مقایسه روش ارائه شده با SVM چند کلاسه

برای نشان دادن کارآیی روش ارائه شده ، ابتدا داده ها به فضای جدید نگاشت شدند و سپس از یک طبقه بند خطی ساده (Adaline) جهت طبقه بندی داده ها بهره گرفته شد. در این مرحله ،از ۵۰٪ داده ها برای تعلیم استفاده شد. از طبقه بند SVM مبتنی بر کرنل هم برای طبقه بندی داده های فضای ورودی بهره گرفته شد تا نتایج طبقه بندی با استفاده از روش ارائه شده با نتایج یک طبقه بند مبتنی بر کرنل مقایسه شود. از بین کرنل های موجود ، کرنل با بهترین کارآیی جهت انجام

طبقه بندی با استفاده از SVM به ازای هر مجموعه داده ، به کاربرده شد. هر آزمایش ۳۰ بار تکرار انجام شد و متوسط خروجی ۳۰ اجرا برای مقایسه در نظر گرفته شده است. نکته قابل توجه اینکه در هر اجرا داده های تعلیم و تست به صورت تصادفی انتخاب شده اند.

طبقه بند Adaline از بردار وزن W ، برای جدا سازی کلاسها در فضای جدید از یکدیگراستفاده می کندکه W از رابطه (۲۵) به دست می آید:

$$W=DX(XX')^{-1} \qquad (۲۵)$$

AC = (تعداد کل داده ها) / (تعداد داده های درست طبقه بندی شده)

نتایج مقایسه در جدول ۱ آورده شده است. همان طور که مشاهده می شود، در روش ارائه شده فقط در حالتی که ابرکره به داده های Wine منطبق می شود، دقت پایین تر است و این نشان می دهد که اگر برای منطبق کردن از منحنی،ابرسطح و... مناسبی استفاده نشود ممکن است به نتایج مورد نظر نرسیم.

جدول ۱. مقایسه دقت طبقه بند خطی *Adaline* در *CFS* و طبقه بند *SVM* درفضای ورودی (متوسط دقت در ۳۰ اجرا)

متوسط دقت در ۳۰ اجرا	SVM	روش ارائه شده با استفاده از SVDD	روش ارائه شده با منطبق کردن دایره
Cancer	96.38%	97.25 %	96.76%
Wine	76.55%	100%	30.80%
Iris	95.20%	96.46%	93.44%
Ionosphere	85.89%	99.72%	72.27%

۵- نتیجه گیری

در این مقاله یک فضای تبدیل جدید ارائه گردید که از دیگر فضاهای تبدیل برای طبقه بندی داده ها، ساده تر می باشد. ایده اصلی استفاده از فاصله داده ها تا منحنی، ابرسطح و... منطبق شده بر داده های هر کلاس می باشد. نتایج نشان می دهد که این فضای تبدیل برای مجموعه های داده به خوبی کار می کند.

مراجع

[1] V. Vapnik, “The Nature of Statistical Learning Theory”. New-York: Springer-Verlag, 1995.

[2] S. Abe,” Support Vector Machines for Pattern Classification”, Springer-Verlag London Limited,2005.

[3] J. Shawe-Taylor, N. Cristianini, “Kernel Methods for Pattern Analysis”, Cambridge University Press, 2004.

[4] B. Scholkopf, Alexander J. Smola, “Learning with Kernels”, Massachusetts Institute of Technology,2002.

[5] T.Huang, V. Kecman, I. Kopriva, “Kernel Based Algorithms for Mining Huge Data Sets”, Springer-Verlag Berlin Heidelberg, 2006.

[6] J.H. Friedman “Multivariate adaptive regression splines”. Annals of Statistics, 19, 1–141, 1991.

[7] D. M. J. Tax, R. P. W. Duin, "Support Vector Data Description". Kluwer Academic Publishers, Machine Learning 54, pp. 45-66, 2004.

[8] UC Irvine Machine Learning Repository ,http://archive.ics.uci.edu/ml/