Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 8, Number 4, Winter 2019, pp. 98-109
Received: 01/13/2019, Revised: 08/16/2019, Accepted: 10/20/2019

Using Low-Rank Approximation In Order To Improve the Efficiency of the Support Vector Machine and Applications

Mohsen Esmaeilbeigi[1*], Omid Chatrabgoun[2]
1*- Department of Mathematics, Malayer University, Malayer, Iran.
2- Department of Statistics, Malayer University, Malayer, Iran.
[1*]m.esmaeilbeigi@malayeru.ac.ir, [2]o.chatrabgoun@malayeru.ac.ir,

Corresponding author address: Mohsen Esmaeilbeigi, Faculty of Mathematical Sciences and Statistics, Malayer University, Malayer, Iran.

Abstract- Support vector machine is one of the most powerful tools in the field of supervised machine learning to classify the existed data. In the data that the linear support vector machine does not have the required efficiency in their classification, using the kernel-based support vector machine which is based on the use of feature space instead of the original data is considered. As a result of this structure, nonlinear classification can be provided. One of the challenges in this approach is to increase the computational complexity and ultimately increase in the required time for classification. As such, it is not particularly useful for large data sets. This increasing in computational time is mainly due to the appearance of the kernel in solving the quadratic optimization problem, which we will be able to overcome this problem using the presented low rank approximation in this paper. In this technique, using a truncated Mercer series of the kernel, the quadratic optimization problem in the kernel-based support vector machine is replaced with a much simpler optimization problem. In the new presented approach, the required vector computations and matrix decompositions will be much faster such that these changes lead to faster resolution of the quadratic optimization problem and increase efficiency. Finally, the results of experiments show that using a low rank kernel-based approximation of support vector machine, while keeping the classification performance in an acceptable range, the computational time has been significantly reduced.

Keywords- Support vector machine, Kernel based SVM, Mercer series, Low rank approximation.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۸، شماره ۴، زمستان ۱۳۹۸ ، صفحه ۱۰۹-۹۸
دریافت: ۱۳۹۷/۱۰/۲۳، بازنگری: ۱۳۹۸/۰۵/۲۵، پذیرش: ۱۳۹۸/۰۷/۲۸

استفاده از تقریب رتبه پایین به منظور بهبود کارایی ماشین بردار پشتیبان مبتنی بر هسته‌ها و کاربردهای آن

محسن اسماعیل بیگی*[۱]، امید چترآبگون[۲]
*۱- گروه ریاضی، دانشکده علوم ریاضی و آمار، دانشگاه ملایر، ملایر، ایران.
۲- گروه آمار، دانشکده علوم ریاضی و آمار، دانشگاه ملایر، ملایر، ایران.
[1]m.esmaeilbeigi@malayeru.ac.ir, [2]o.chatrabgoun@malayeru.ac.ir

* نشانی نویسنده مسئول: محسن اسماعیل بیگی، ملایر، کیلومتر ۴ جاده اراک، دانشگاه ملایر، دانشکده علوم ریاضی و آمار

چکیده- ماشین بردار پشتیبان یکی از ابزارهای توانمند در زمینه یادگیری ماشین با ناظر در طبقه‌بندی داده‌ها می‌باشد. در مواجهه با داده‌هایی که ساختار ماشین بردار پشتیبان خطی در طبقه‌بندی آنها از کارایی لازم برخوردار نیست، استفاده از ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها مدنظر می‌باشد. در رویکرد مبتنی بر هسته‌ها به دلیل استفاده از فضای ویژگی داده‌ها به جای خود داده‌های اصلی امکان طبقه‌بندی غیرخطی فراهم می‌آید. یکی از چالش‌های موجود در این رویکرد افزایش پیچیدگی‌های محاسباتی و در نهایت افزایش زمان لازم برای طبقه‌بندی است. عمدتاً این افزایش زمان محاسباتی به دلیل ظاهر شدن هسته در حل مسئله بهینه‌سازی درجه دوم است که با استفاده از تقریب رتبه پایین ارائه شده در این مقاله قادر خواهیم بود بر این مشکل غلبه کنیم. در این تکنیک با به کارگیری سری تقریبی قطع شده از هسته موجود، مسئله بهینه‌سازی درجه دوم در ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها با یک مسئله بهینه سازی با ساختار ساده‌تر جایگزین می‌گردد. در این رویکرد ، حاصلضرب‌های بردار-ماتریس و تجزیه‌های ماتریسی مورد نیاز بسیار سریع‌تر انجام خواهد شد. این تغییرات منجر به حل سریعتر مسئله بهینه‌سازی درجه دوم موجود و افزایش کارایی در طبقه‌بندی می‌گردد. نهایتاً نتایج عددی ارائه شده در طبقه‌بندی برخی داده‌های کاربردی با استفاده از تقریب رتبه پایین ماشین بردار پشتیبان مبتنی بر هسته‌ها نشان می‌دهد که ضمن حفظ عملکرد طبقه‌بندی در حد قابل قبول، زمان محاسباتی به‌طور قابل توجهی کاهش یافته است.

واژه‌های کلیدی: طبقه‌بندی نظارت‌شده، ماشین بردار پشتیبان، بسط مرکز، تقریب رتبه پایین

۱- مقدمه

به مجموعه‌ای از روش‌های قابل اعمال بر پایگاه داده‌های بزرگ و پیچیده به منظور کشف الگوهای پنهان و جالب توجه نهفته در میان داده‌ها، داده‌کاوی گفته می‌شود. علم میان‌رشته‌ای داده‌کاوی، پیرامون ابزارها، روش‌ها و تئوری‌هایی است که برای آشکارسازی الگوهای موجود در داده‌ها مورد استفاده قرار می‌گیرند و گامی اساسی در راستای کشف دانش محسوب می‌شود [۱]. در واقع با کمک داده‌کاوی که شامل دو بخش یادگیری با ناظر[۱] و بدون ناظر[۲] است می‌توان به اطلاعات مهمی در داده‌های موجود دست یافت [۲]. یادگیری با ناظر در فرم گسسته آن به طبقه‌بندی[۳] و در فرم پیوسته به رگرسیون[۴] معروف است. فرم گسسته آن یعنی طبقه‌بندی دارای اشکال متنوعی می‌باشد که طبقه‌بندی دودویی یکی از مهمترین آنها است. فرایندهای گاوسی و شبکه‌های عصبی از ابزارهای معروفی هستند که برای طبقه‌بندی به کار می‌آیند. این

روش‌ها برای طبقه‌بندی دودویی نیز به خوبی مورد استفاده قرار گرفته‌اند [۳-۶].

یکی از مهمترین ابزارها برای طبقه‌بندی دودویی استفاده از ماشین بردار پشتیبان است [۷]. در واقع بسیاری از روشهای قبلی نظیر فرایند گاوسی و شبکه‌های عصبی سعی می کنند تا خطای طبقه-بندی را کاهش دهند در حالی که هدف استفاده از ماشین بردار پشتیبان کاهش ریسک عملیاتی طبقه‌بندی است و این ریسک عملیاتی را می‌توان مدل‌سازی نمود. نسخه‌های متفاوتی از ماشین بردار پشتیبان طی سالیان اخیر استفاده شده است که از این دسته می‌توان به موارد مطرح شده در منابع[۸ و ۹] اشاره کرد.

در مواجهه با داده‌هایی که طبقه‌بندی آنها با استفاده از ابرصفحه-ها امکان‌پذیر باشد، استفاده از ماشین بردار پشتیبان خطی معمول می‌باشد. این در حالی است که در بسیاری از داده‌های موجود با توجه به ماهیت آنها، استفاده از ساختار ماشین بردار خطی برای طبقه‌بندی آنها از کارایی لازم برخوردار نیست. در چنین مواقعی استفاده از ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها به دلیل استفاده از فضای ویژگی داده‌ها مدنظر می‌باشد. به عبارتی به جای اینکه از فضای خود داده‌ها برای طبقه‌بندی استفاده شود، از یک ویژگی در داده‌ها مثل فاصله آنها از یکدیگر یا توابعی از این فاصله استفاده می‌شود. بنابراین در رویکرد ماشین بردار پشتیبان مبتنی بر هسته‌ها امکان طبقه‌بندی غیرخطی فراهم می‌آید. رویکرد ماشین بردار پشتیبان مبتنی بر هسته‌ها در فصل دوم این مقاله معرفی می‌گردد و برای کسب اطلاعات بیشتر در این زمینه می‌توان به منابع [۸ و ۱۰] مراجعه کرد. لازم به ذکر است که ظرفیت‌های موجود در ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها عموما به ویژگی‌های مفید هسته‌ها مانند امکان تعامل با داده‌های در ابعاد بالا و امکان مواجهه با داده‌هایی که به صورت غیرخطی طبقه‌بندی می‌شوند، بر می‌گردد. به‌علاوه وجود برخی پارامترهای قابل تنظیم در هسته‌ها، باعث انعطاف‌پذیری بیشتر ماشین بردار پشتیبان مبتنی بر هسته‌ها در طبقه‌بندی می‌باشد. برای آشنایی بیشتر با هسته‌ها و ویژگی‌های آنها می‌توان به [۱۱] مراجعه کرد.

رویکرد استفاده از هسته‌ها در طبقه‌بندی به کمک ماشین بردار پشتیبان، علی‌رغم افزایش قابلیت‌های این روش، باعث افزایش پیچیدگی‌های محاسباتی و در نهایت افزایش زمان لازم برای طبقه‌بندی می گردد [۱۲]. این پیچیدگی‌های محاسباتی به دلیل حضور هسته در مسئله بهینه‌سازی درجه دوم موجود در ساختار ماشین بردار پشتیبان است. در حقیقت ماتریس هسیان موجود در این مسئله بسیار شلوغ[۵] و بدوضع می‌باشد. از سوی دیگر هر چه هسته به‌کار رفته در ساختار ماشین بردار پشتیبان از ظرفیت بالاتری جهت طبقه‌بندی داده با ماهیت غیرخطی برخوردار باشد، عمدتا این چالش‌های محاسباتی افزایش می‌یابد. به‌علاوه حل مسئله بهینه‌سازی موجود، بیشترین حجم محاسباتی را در فرایند یادگیری ماشین بردار پشتیبان مبتنی بر هسته‌ها شامل می‌گردد و نقش اصلی را در زمان محاسباتی و کارایی این روش ایفا می‌کند.

بنابراین ارائه یک ساختار کارآمد که در حداقل زمان ممکن منجر به دستیابی به جواب مسئله بهینه‌سازی موجود در ماشین بردار پشتیبان گردد، بسیار حائز اهمیت است. در این زمینه تلاش‌های فراوانی صورت گرفته است که بطور مثال در منبع [۱۳] روش‌های کاهش بعد داده‌ها برای کاهش پیچیدگی‌های محاسباتی در ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها استفاده شده است. اگرچه کاهش بعد داده‌ها به کاهش چالش‌های محاسباتی کمک می کند، اما این کاهش بعد، می‌تواند منجر به از دست‌دادن اطلاعات زیادی در مورد هسته‌ها گردد و عملکرد طبقه بندی را با چالش مواجه نماید. همچنین در منبع [۱۴] این امر با انتخاب تصادفی از ستون‌های ماتریس متناظر با هسته موردنظر در ساختار ماشین بردار پشتیبان انجام شده است. اگرچه ساختار ارایه شده می‌تواند باعث کاهش قابل توجه پیچیدگی‌های محاسباتی گردد، اما انتخاب‌های تصادفی از میان ستونهای ماتریس می تواند با چالش‌های زیادی همراه باشد. به‌علاوه در [۱۵] از روشهای تنک-سازی ماتریس هسته استفاده شده است که بر اساس آن ماتریس چگال مورد استفاده به یک ماتریس تنک تبدیل می‌شود. این روش نیز باعث کاهش دشواری‌های محاسباتی می‌گردد، اما به‌دلیل چالش‌های موجود در فرایند جایگزینی ماتریس چگال با ماتریس تنک، برخی مواقع عملکرد طبقه‌بندی روش چندان مطلوب نیست.

در این مقاله نشان خواهیم داد چگونه با استفاده از تقریب رتبه پایین هسته موجود، بر اساس بسط سری مرکر هسته در ساختار ماشین بردار پشتیبان، امکان جایگزینی مسئله بهینه‌سازی درجه دوم اصلی با یک مسئله بهینه‌سازی درجه دوم با ساختار بسیار ساده‌تر مهیا می‌باشد. تقریب رتبه پایین به کار رفته مبتنی بر بسط مرکر[۶] هسته، امکان جایگزینی هسته مورد نظر را با یک سری متناهی فراهم می‌سازد. در واقع به جای هسته رتبه کامل اصلی از یک ساختار رتبه پایین جایگزین استفاده می‌شود. جملات این بسط با استفاده از مقادیر ویژه و توابع ویژه عملگر هیلبرت-اشمیت متناظر با هسته بکار رفته در ساختار ماشین بردار پشتیبان می-باشد. نام گذاری رتبه پایین برای ساختار تقریب فوق به دلیل استفاده از سری مرکر قطع شده می‌باشد که از رتبه پایین‌تری در تقریب هسته برخوردار است و این در حالی است که استفاده از خود هسته اصلی باعث ایجاد یک رویکرد رتبه کامل می‌گردد. با

توجه به آنکه بر اساس نرخ کاهشی موجود در مقادیر ویژه، تنها جملات اولیه بسط مرکز از نقش اساسی در تقریب هسته برخوردارند بنابراین تقریب رتبه پایین بدست آمده به اندازه کافی دقیق خواهد بود. به عبارت دیگر، بسط مرکز قطع شده M جمله-ای بهترین تقریب M جمله‌ای از منظر کمترین مربعات را برای هسته در فضای تقریب متناظر با آن فراهم می‌نماید [۱۱].

بنابراین در روش مورد استفاده در این مقاله، ماتریس هسیان جدید موجود در مسئله بهینه‌سازی درجه دوم دارای ساختار بسیار ساده و تنک می‌باشد و این در حالی است که ماتریس هسیان در فرم رتبه کامل هسته کاملا چگال بود. بنابراین در ساختار جدید، محاسبات برداری و تجزیه‌های ماتریسی بسیار سریع‌تر و با چالش محاسباتی کمتر انجام خواهد شد که این تغییرات منجر به حل سریع‌تر مسئله بهینه‌سازی درجه دوم و کاهش زمان طبقه‌بندی می‌گردد، طوری‌که نهایتاً باعث افزایش کارایی طبقه‌بندی خواهد شد. البته در گذشته تلاش‌هایی برای تقریب رتبه پایین هسته با استفاده از روش‌های دیگر صورت گرفته است که از میان آنها می توان به منابع [۱۶ و ۱۷] اشاره کرد. در این روش‌ها به طور مستقیم از تجزیه‌های ماتریسی استفاده شده است که می‌تواند با دشواری‌های محاسباتی همراه باشد. این در حالی است که ساختار تقریب به کار رفته در این مقاله، بر مبنای قضیه مرکز قطع شده، بهترین تقریب از منظر کمترین مربعات را فراهم می‌نماید بدون اینکه نیازی به تجزیه مستقیم ماتریسی داشته باشد. به‌علاوه در تعدادی از منابع مانند [۱۹-۲۱] سعی شده با استفاده از روش-های برنامه‌نویسی و ارایه الگوریتم‌های کامپیوتری بهبود یافته کارایی روش‌های ماشین بردار پشتیبان را افزایش دهند. اما در این مقاله به کمک یک ساختار مدلسازی جدید که مبتنی بر تغییرات در ریاضیات مسئله بهینه سازی موجود در ساختار بردار پشتیبان است، این کار را انجام خواهیم داد.

برای آن‌که امکان مقایسه حالت رتبه پایین ماشین بردار پشتیبان مبتنی بر هسته با حالت رتبه کامل فراهم گردد، برخی نتایج عددی و مثال‌های کاربردی در این مقاله ارایه شده است. نتایج به-دست آمده نشان می‌دهد که تقریب رتبه پایین به‌کار رفته در ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها، منجر به سادگی محاسبات، کاهش زمان محاسباتی و نهایتاً افزایش کارایی در طبقه‌بندی می‌گردد. به‌علاوه با توجه به معیارهای مطرح شده، علی‌رغم استفاده از ساختار تقریبی جایگزین، عملکرد مطلوب طبقه‌بندی حفظ می‌گردد.

ساختار این مقاله به صورت زیر سازمان‌دهی شده است: ابتدا ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها شرح داده می‌شود و سپس ضمن اشاره به خواص مناسب هسته‌ها، چالش‌های محاسباتی استفاده از هسته‌ها ذکر می‌گردد. در قسمت بعد سعی می‌شود با به کار بردن تقریب رتبه پایین هسته مورد نظر، فرم ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها بازنویسی شود و چالش‌های محاسباتی آن مرتفع گردد. در قسمت نتایج عددی ابتدا با یک مثال عددی، عملکرد ماشین بردار پشتیبان مبتنی بر هسته‌ها را شرح خواهیم داد و مولفه‌های آن را بررسی خواهیم کرد و همچنین با حالت رتبه کامل هسته مقایسه خواهیم نمود. در قسمت بعد با ارایه دو مثال کاربردی، عملکرد روش رتبه پایین را در مقایسه با حالت رتبه کامل بررسی خواهیم نمود و در آخر نیز به بحث و نتیجه‌گیری خواهیم پرداخت.

۲- ماشین بردار پشتیبان مبتنی بر هسته

برای یک مسئله طبقه‌بندی دودویی فرض می کنیم که $\{X_i,y_i\}$ برای $i=1,2,...,N$ داده‌های آموزشی هستند طوریکه $X_i\in\Omega\subset\mathbb{R}^d$ و $y_i\in\{-1,+1\}$ است. همان‌طور که ذکر شد، روش ماشین بردار پشتیبان برای طبقه‌بندی خطی، سعی دارد که با ساختن یک ابرصفحه (که عبارت است از یک معادله خطی)، داده‌ها را از هم تفکیک کند. در واقع روش طبقه‌بندی بر مبنای ماشین بردار پشتیبان خطی سعی می‌کند بهترین ابرصفحه‌ای را پیدا کند که با رعایت حداکثر فاصله ممکن، داده‌های مربوط به دو طبقه را از هم تفکیک نماید. حال اگر معادله این ابرصفحه به صورت $w^Tx+b=0$ درنظر گرفته شود که در آن w بردار ضرایب و b عرض از مبدا ابرصفحه تفکیک کننده مورد نظر است، این پارامترها باید طوری بهینه شوند که ریسک عملیاتی در این طبقه بندی حداقل گردد. به عبارتی ایده اصلی آن است که یک جدا کننده مناسب انتخاب شود، یعنی جدا کننده‌ای که بیشترین فاصله را با نقاط همسایه از هر دو طبقه دارد. این جواب در واقع بیشترین مرز را با نقاط مربوط به دو طبقه مختلف دارد و می‌تواند با دو ابرصفحه موازی که حداقل از یکی از نقاط دو طبقه عبور می‌کنند، کران‌دار شود. این نقاط یا به عبارتی بردارهای واقع بر روی ابرصفحه‌های موازی، بردارهای پشتیبان نام دارند. رابطه این دو ابرصفحه موازی که مرز جداکننده را تشکیل می‌دهند به صورت $w^Tx+b=1$ و $w^Tx+b=-1$ است.

نکته قابل توجه آن است که اگر داده‌های آموزشی به صورت خطی تفکیک‌پذیر باشند، می‌توان دو ابرصفحه مرزی را به‌گونه‌ای انتخاب کرد که هیچ داده‌ای بین آنها نباشد و سپس فاصله بین این دو ابرصفحه موازی را به حداکثر رساند. با به کارگیری قضایای هندسی، فاصله این دو ابرصفحه عبارت است از $2/||w||$ که به منظور دستیابی به حداکثر فاصله ممکن باید $||w||$ را به حداقل رساند. همچنین باید از قرار گرفتن نقاط داده در ناحیه درون مرز

جلوگیری کرد که برای این کار یک قید به تعریف اضافه می‌شود، طوریکه برای هر i ، با اعمال محدودیت‌های که در ادامه می‌آید اِطمینان حاصل می‌شود که هیچ نقطه‌ای در داخل مرز قرار نمی‌گیرد:

برای داده‌های مربوط به طبقه اول:

$$w^T x_i + b \geq 1 \quad (۱)$$

و برای داده‌های مربوط به طبقه دوم:

$$w^T x_i + b \leq -1 \quad (۲)$$

می‌توان قیود (۱) و (۲) را با ضرب y_i در طرفین هر دو رابطه به یک صورت واحد و به فرم رابطه (۳) نمایش داد:

$$y_i(w^T x_i + b) \geq 1, \quad 1 \leq i \leq n, \quad (۳)$$

و این مسئله بهینه سازی مورد نظر است که در آن $\|w\|$ باید مینیمم گردد. حل این مسأله بهینه‌سازی با توجه به آنکه وابسته به مقدار دقیق $\|w\|$ است، دشوار می‌باشد. چرا که از منظر محاسباتی، حل یک مسأله بهینه‌سازی غیرمحدب بسیار دشوارتر از حل یک مسأله بهینه‌سازی محدب است. خوشبختانه می‌توان مسأله را با جایگزینی $\frac{1}{2}\|w\|^2$ به جای $\|w\|$ بدون آنکه تغییری در جواب مسأله ایجاد شود، حل نمود. این مسأله از نوع مسائل بهینه سازی درجه دوم بوده و یک مسأله محدب است. پس مسأله بهینه‌سازی عبارتست از به حداقل رساندن $\frac{1}{2}\|w\|^2$ با در نظر گرفتن محدودیت مطرح شده در رابطه (۳) که در آن ضریب $\frac{1}{2}$ جهت سادگی محاسبات استفاده می‌شود. این مسئله بهینه‌سازی با شرط مورد نظر به کمک ضرایب لاگرانژ α_i $(\alpha_i \geq 0)$ به عنوان یک مسئله بهینه‌سازی اولیه غیر مقید در رابطه (۴) آمده است:

$$l_P = \frac{1}{2} w^T w + \sum_{i=1}^{N} \alpha_i \big(1 - y_i(w^T x_i + b)\big), \quad (۴)$$

که در آن $w^T w = ||w||^2$ است. در این مسئله بهینه سازی به منظور بهینه کردن پارامترهای موجود باید به دنبال یک نقطه زینی باشیم، زیرا l_P باید نسبت به b و w مینیمم گردد در حالی که نسبت به α باید ماکزیمم شود و در واقع با یک مسئله بهینه-سازی چند هدفه مواجهه هستیم. لذا به منظور تبدیل آن به یک مسئله بهینه‌سازی تک هدفه که به مسئله دوگان مسئله بهینه-سازی اولیه مشهور است، سعی می‌کنیم که آن را بازسازی کنیم. برای این منظور با مشتق‌گیری از l_P در (۴) نسبت به w و برابر صفر قرار دادن آن، رابطه (۵) را خواهیم داشت:

$$w + \sum_{i=1}^{N} \alpha_i(-y_i)x_i = 0, \quad (۵)$$

که نتیجه می‌دهد

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (۶)$$

همچنین با مشتق گیری نسبت به b و برابر صفر قرار دادن آن داریم:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad (۷)$$

که این قید، به قید بایاس معروف است. با جایگزینی مقدار بدست آمده برای w در رابطه (۶) و قرار دادن آن در l_P، به دوگان مسأله اولیه، یعنی l_D، دست خواهیم یافت:

$$\begin{aligned} l_D = \frac{1}{2}\sum_{i=1}^{N} \alpha_i y_i x_i^T \sum_{j=1}^{N} \alpha_i y_i x_j \\ + \sum_{i=1}^{N}\left(1 - y_i\left(\sum_{i=1}^{N} \alpha_i y_i x_j^T x_i + b\right)\right) \\ = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i , \end{aligned} \quad (۸)$$

لذا تابع هدف تنها یک تابع از $\alpha_i \geq 0$ نسبت به قید بایاس است و این تعریف مسأله بهینه‌سازی تک هدفه دوگان مسئله اصلی است. این مسأله، یک مسأله بهینه‌سازی درجه دوم است و همواره یک مقدار بیشینه برای α_i ها وجود خواهد داشت. هر چند با اعمال یک ضریب جریمه C برای داده‌هایی که از طبقه‌بندی مورد نظر تخطی نمایند، ضریب لاگرانژ باید در شرط $0 \leq \alpha_i \leq C$ صدق کند. ماشین بردار پشتیبانی که در این شرایط صدق کند، به ماشین بردار پشتیبان با حاشیه نرم مشهور است و در قالب کلی مسأله بهینه‌سازی درجه دوم در قالب رابطه (۹) بیان می‌گردد:

$$\begin{aligned} &\max_{\alpha}\left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j X_i^T X_j\right) \\ &\text{subject to} \sum_{i=1}^{N} \alpha_i y_i = 0, \\ &\qquad 0 \leq \alpha_i \leq C, \end{aligned} \quad (۹)$$

لازم به ذکر است که پارامتر C یک پارامتر آزاد است که باید توسط کاربر یا بر اساس یک فرایند استاندارد نظیر اعتبارسنجی متقابل تعیین گردد.

در مواجهه با داده‌هایی که طبقه‌بندی آنها با استفاده از ابرصفحه‌ها امکان‌پذیر باشد، استفاده از ماشین بردار پشتیبان خطی معمول می‌باشد. این در حالی است که در بسیاری از داده‌های موجود با توجه به ماهیت آنها، استفاده از ساختار ماشین بردار پشتیبان خطی برای طبقه‌بندی آنها از کارایی لازم برخوردار نیست. در چنین مواقعی استفاده از ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها به دلیل استفاده از فضای ویژگی داده‌ها مد نظر می باشد. در حقیقت بر اساس قضیه کاور[۷] در منبع [۲۲]، زمانی که با استفاده از ابرصفحه‌ها در فضای ورودی داده‌ها قادر به تفکیک

خطی داده‌ها نباشیم، این اطمینان وجود دارد که با استفاده از یک نگاشت ویژگی مناسب، Φ، داده‌ها را به فضای ویژگی جدید انتقال دهیم به‌گونه‌ای که در این فضا امکان تفکیک خطی با استفاده از ابرصفحه‌ها امکان‌پذیر باشد. لازم به ذکر است که داده X_i در فضای ورودی، با ویژگی آن یعنی Φ_{X_i} در فضای ویژگی جایگزین می‌گردد. بنابراین ابرصفحه جدا کننده داده‌ها به صورت رابطه (۱۰) تعیین می‌گردد:

$$\Phi_X^T w + b = 0, \qquad (۱۰)$$

و مساله دوگان مطرح در رابطه (۹) برای ساختار ماشین بردار پشتیبان به فرم رابطه (۱۱) تبدیل می گردد:

$$\max_{\alpha}\left(\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \Phi_{X_i}^T\Phi_{X_j}\right) \quad \text{subject to} \sum_{i=1}^{N}\alpha_i y_i = 0\,, \quad 0 \le \alpha_i \le C. \qquad (۱۱)$$

لازم به ذکر است، دستیابی به فضای ویژگی داده‌ها در چهارچوب فضای هیلبرت هسته باز تولید امکان‌پذیر است. به عبارت دیگر، نگاشت ویژگی معرفی شده به صورت $\Phi: \Omega \mapsto \mathcal{H}_K(\Omega)$ در نظر گرفته می‌شود، به طوری که تحت نگاشت

$$X \longmapsto \Phi_X = K(.,X), \qquad (۱۲)$$

مجموعه داده‌ها از فضای Ω به فضای ویژگی $\mathcal{H}_K(\Omega)$ انتقال می‌یابد. جایی‌که $\mathcal{H}_K(\Omega)$ فضای هیلبرت هسته بازتولید نظیر به هسته K می‌باشد. برای کسب اطلاعات بیشتر در زمینه فضای هیلبرت هسته بازتولید می‌توان به منبع [۱۱] مراجعه کرد.

با توجه به خصوصیات فضای هیلبرت هسته بازتولید، ضرب داخلی در فضای ویژگی بدست آمده، $\mathcal{H}_K(\Omega)$، با استفاده از هسته K به سادگی امکان‌پذیر است [۱۱]، یعنی:

$$K(X,Z) = \Phi_X^T \Phi_Z. \qquad (۱۳)$$

حال به کمک رابطه (۱۳)، مساله دوگان بهینه‌سازی درجه دو در ساختار ماشین بردار پشتیبان مبتنی بر فضای ویژگی به صورت رابطه (۱۴) خواهد بود:

$$\max_{\alpha}\left(\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(X_i,X_j)\right) \quad \text{subject to} \sum_{i=1}^{N}\alpha_i y_i = 0\,, \quad 0 \le \alpha_i \le C, \qquad (۱۴)$$

با توجه به نقش و حضور پررنگ هسته‌ها، رویکرد فوق تحت عنوان ماشین بردار پشتیبان مبتنی بر هسته‌ها نامگذاری می‌گردد و امکان طبقه‌بندی غیرخطی داده‌ها را فراهم می‌نماید. لازم به ذکر است که ظرفیت‌های موجود در ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها عموما به ویژگی‌های مفید هسته‌ها مانند امکان تعامل با داده‌های در ابعاد بالا و امکان مواجهه با داده‌هایی که به صورت غیر خطی طبقه‌بندی می‌شوند، بر می‌گردد. به‌علاوه وجود برخی پارامترهای قابل تنظیم در هسته‌ها، باعث انعطاف‌پذیری بیشتر ماشین بردار پشتیبان مبتنی بر هسته‌ها در طبقه‌بندی می‌شود. عموما استفاده از سه هسته معروف در ساختار ماشین بردار پشتیبان، بیشتر از هسته‌های دیگر مرسوم است. این سه هسته عبارتند از هسته چندجمله‌ای، هسته گاوسی و هسته سیگموئید که در جدول ۱ ساختار آنها مشخص شده است و هر کدام دارای یک پارامتر می‌باشند که باعث انعطاف‌پذیری بیشتر آنها می‌شود.

جدول۱: هسته‌های معروف در ساختار ماشین بردار پشتیبان

هسته	ساختار
چند جمله ای	$K(x,z) = (1 + x^T z)^{\epsilon}$
گاوسی	$K(x,z) = e^{-\epsilon\|x-z\|^2}$
سیگموئید	$K(x,z) = \tanh(1 + \epsilon x^T z)$

جدا از مزیت‌های متعدد، رویکرد استفاده از هسته‌ها در طبقه‌بندی به کمک ماشین بردار پشتیبان باعث افزایش پیچیدگی‌های محاسباتی و نهایتا افزایش زمان لازم برای طبقه‌بندی و عدم کارایی مناسب می‌گردد. این پیچیدگی‌های محاسباتی به دلیل حضور هسته در حل مسئله بهینه‌سازی درجه دوم موجود در ساختار ماشین بردار پشتیبان است. در حقیقت ماتریس هسیان موجود در این مسئله بسیار شلوغ و بدوضع می‌باشد. از سوی دیگر هر چه هسته بکار رفته در ساختار ماشین بردار پشتیبان از ظرفیت بالاتری جهت طبقه‌بندی داده‌ها با ماهیت غیرخطی برخوردار باشد، عمدتا این چالش‌های محاسباتی افزایش می‌یابد. به‌علاوه این مسئله بهینه‌سازی بیشترین حجم محاسبات را در ماشین بردار پشتیبان مبتنی بر هسته‌ها را شامل می‌گردد و نقش اصلی را در زمان محاسباتی این روش ایفا می‌کند. در قسمت بعدی سعی می‌شود با تکنیک تقریب رتبه پایین هسته به این چالش‌های مهم در محاسبات عددی ساختار ماشین بردار پشتیبان بپردازیم.

۳- تکنیک تقریب رتبه پایین در ماشین بردار پشتیبان مبتنی بر هسته

بر اساس قضیه مرکر [۲۳] هر هسته معین مثبت به کمک بسط سری نامتناهی موجود در رابطه (۱۵) قابل نمایش است:

$$K(x,z) = \sum_{n=1}^{\infty}\lambda_n\,\varphi_n(X)\,\varphi_n(Z), \qquad (۱۵)$$

که در این بسط λ_n و φ_n به ترتیب مقادیر ویژه مثبت و توابع ویژه متعامد یکه نظیر به عملگر هیلبرت اشمیت هسته K هستند.

لازم به ذکر است که عملگر هیلبرت اشمیت مورد اشاره، $\kappa: L_2(\Omega) \to L_2(\Omega)$، در رابطه (۱۶) تعریف می‌گردد:

$$\left(\kappa f\right) = \int_{\Omega} K(x,z)\, f(z) dz\,, \qquad (۱۶)$$

و مقادیر ویژه و بردارهای ویژه به کار رفته در بسط مرکر از حل مسئله مقدار ویژه ارایه شده در رابطه (۱۷) حاصل می‌شود،

$$\left(\kappa\varphi\right)(X) = \lambda\varphi(X). \qquad (۱۷)$$

حال از آنجایی که کار کردن با تعداد جملات نامتناهی از سری مرکر در سیستم‌های محاسباتی امکان‌پذیر نیست و با توجه به نرخ کاهشی موجود در مقادیر ویژه (که تنها جملات اولیه این بسط از نقش اساسی در تقریب هسته K برخوردارند) از تقریب رتبه پایین ارایه شده در رابطه (۱۸) برای تقریب هسته K استفاده می‌کنیم. در واقع اگر N تعداد داده‌های موجود باشد آنگاه برای M خیلی کوچکتر از N داریم:

$$K(x,z) = \sum_{n=1}^{M} \lambda_n\, \varphi_n(X)\, \varphi_n(Z) \qquad (۱۸)$$

یاد آوری می‌گردد بر اساس قضیه سری قطع شده مرکر [۱۱]، اگر K یک هسته معین مثبت با بسط مرکر ارائه شده در رابطه (۱۵) بوده و سری قطع شده مرکر M جمله‌ای به صورت (۱۸) مدنظر باشد، آنگاه سری قطع شده فوق بهترین تقریب M-جمله‌ای از منظر کمترین مربعات را در $L_2(\Omega)$ برای هسته K فراهم می‌نماید. حال با در نظر گرفتن نکات فوق، امکان تقریب ماتریس K در مساله دوگان در رابطه (۱۴) فراهم می باشد. بدین منظور ابتدا برخی روابط را در ادامه معرفی خواهیم کرد:

$$\boldsymbol{\varphi}(\mathrm{X}) = \begin{pmatrix} \varphi_1(X) \\ \vdots \\ \varphi_M(X) \end{pmatrix}, \qquad \emptyset = \begin{pmatrix} \boldsymbol{\varphi}(X_1)^T \\ \vdots \\ \boldsymbol{\varphi}(X_N)^T \end{pmatrix}_{\boldsymbol{N\times M}}$$

همچنین

$$\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_M \end{pmatrix}_{M\times M}$$

با در نظر گرفتن تعریف‌های فوق و با استفاده از سری مرکر قطع شده M جمله‌ای، ماتریس K در مساله دوگان بهینه‌سازی درجه دو به فرم رابطه (۱۹) تقریب زده می‌شود:

$$\boldsymbol{K} \approx \emptyset \wedge \emptyset^{\boldsymbol{T}} \qquad (۱۹)$$

با توجه به آنکه تنها از M جمله اول بسط مرکر استفاده شده است، از نماد $\approx$ به جای نماد $=$ استفاده کرده‌ایم. اما خوشبختانه در بسیاری از هسته‌ها به ازای مقادیر خیلی کم M می‌توان تقریب‌های بسیار دقیقی از هسته K بدست آورد، لذا در ادامه بحث بجای نماد $\approx$ از نماد $=$ استفاده می‌کنیم.

بر اساس آنچه که بدست آوردیم، قصد داریم با استفاده از تقریب جدید بدست آمده برای K، به فرم بازسازی شده مسئله دوگان بهینه سازی درجه دو دسترسی یابیم. بدین منظور در ابتدا با اندکی تغییرات می توان مسئله بهینه‌سازی در رابطه (۱۴) را به فرم ماتریسی در رابطه (۲۰) بازنویسی کرد،

$$\begin{aligned} &\min_{\alpha} \left(\frac{1}{2}\alpha^T D_y K D_y \alpha - e^T\alpha\right) \\ &\text{subject to } y^T\alpha = 0, \\ &\qquad \alpha \in [0,C]^N, \end{aligned} \qquad (۲۰)$$

که در آن D_y یک ماتریس قطری با درایه‌های y روی قطر اصلی است و e یک بردار شامل درایه‌های یک است. حال بر اساس تقریب بدست آمده در رابطه (۱۰) می‌توانیم ماتریس K را به صورت ارایه شده در رابطه (۲۱) بنویسیم،

$$\boldsymbol{K} \approx \left(\Lambda^{\frac{1}{2}}\emptyset\right)^T \left(\Lambda^{\frac{1}{2}}\emptyset\right), \qquad (۲۱)$$

سپس با در نظر گرفتن $V = D_y\emptyset\,\Lambda^{\frac{1}{2}}$ و با توجه به اینکه $VV^T \approx$ $D_y\boldsymbol{K}D_y$ می‌باشد، مساله بهینه‌سازی فوق به صورت ارایه شده در رابطه (۲۲) خواهد بود:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \left(\frac{1}{2}(V^T\boldsymbol{\alpha})^T(V^T\boldsymbol{\alpha}) - \boldsymbol{e}^T\boldsymbol{\alpha}\right) \\ &\text{subject to } \boldsymbol{y}^T\boldsymbol{\alpha} = 0, \\ &\boldsymbol{\alpha} \in [0,C]^N, \end{aligned} \qquad (۲۲)$$

در ادامه، ماتریس $V^T\boldsymbol{\alpha}$ را که شامل درایه‌های مجهول بردار $\boldsymbol{\alpha}$ می‌باشد را به شکل رابطه (۲۳) در نظر می‌گیریم:

$$V^T\boldsymbol{\alpha} = I_M\boldsymbol{\beta}, \qquad (۲۳)$$

که I_M یک ماتریس همانی از مرتبه M و $\beta \in \mathbb{R}^M$ می باشد.

با این ملاحظات مساله دوگان بهینه سازی درجه دو به صورت ارایه شده در رابطه (۲۴) قابل بازنویسی است،

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\alpha}} \frac{1}{2}\begin{pmatrix}\boldsymbol{\beta}^T & \boldsymbol{\alpha}^T\end{pmatrix}\begin{pmatrix} I_M & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta} \\ \boldsymbol{\alpha}\end{pmatrix} - \begin{pmatrix}\mathbf{0} & \boldsymbol{e^T}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta} \\ \boldsymbol{\alpha}\end{pmatrix} \\ &\text{subject to } \begin{pmatrix} \mathbf{0} & \boldsymbol{y^T} \\ -I_M & \boldsymbol{V^T}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta} \\ \boldsymbol{\alpha}\end{pmatrix} = 0, \\ &\qquad \boldsymbol{\alpha} \in [0,C]^N,\ \beta \in \mathbb{R}^M \end{aligned} \qquad (۲۴)$$

قابل توجه است که، اگر چه این سیستم از مرتبه بالاتر $M+N$ می‌باشد (در حالی که دستگاه اصلی از مرتبه N بود) اما ماتریس هسیان بدست آمده در ساختار جدید مساله بهینه‌سازی درجه دوم بسیار تنک می‌باشد. این در حالی است که ماتریس هسیان برای هسته K در فرم قبلی که از این پس آن را فرم رتبه کامل می-نامیم، کاملا چگال بود. بنابراین در ساختار جدید، حاصلضرب بردار –ماتریس و تجزیه‌های ماتریسی مورد نیاز در فرایند حل مساله بهینه‌سازی درجه دو بسیار سریعتر انجام خواهد شد، که هر دوی این تغییرات منجر به حل سریعتر مسئله بهینه‌سازی درجه دو می‌گردد و در نهایت باعث کاهش زمان محاسباتی برای طبقه‌بندی

با استفاده از ماشین بردار پشتیبان مبتنی بر هسته‌ها می‌گردد.

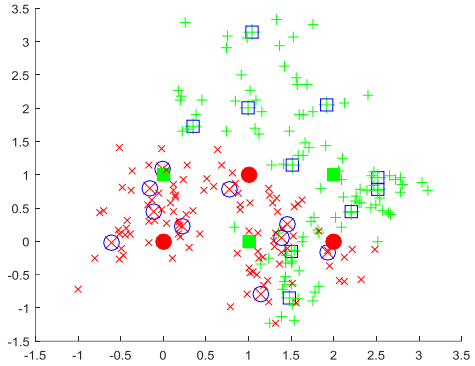
۴- نتایج عددی

در این قسمت نتایج حاصل از طبقه‌بندی داده‌ها با ماشین بردار پشتیبان مبتنی بر هسته‌ها در حالت تقریب رتبه پایین مطرح شده در این مقاله را مورد بررسی قرار داده و عملکرد آن را با حالت رتبه کامل مقایسه خواهیم کرد. در اینجا ۱۰۰ نمونه از دو جامعه نرمال آمیخته دو متغیره (جمعا ۲۰۰ نمونه) را در نظر می‌گیریم. جامعه اول ترکیب خطی سه متغیر نرمال با ضرایب مساوی و مراکز $\{(1,0),(0,1),(1,2)\}$ است در حالی که همه متغیرهای نرمال موجود در این ترکیب غیرخطی دارای ماتریس واریانس-کوواریانس همانی می‌باشند. برای جامعه دوم نیز به همین شکل عمل شده است، با این تفاوت که مراکز آن به صورت $\{(0,0),(1,1),(2,0)\}$ است. همان‌طور که در شکل ۱ نشان داده شده است نقاط جامعه اول با + و نقاط شبیه‌سازی شده از جامعه دوم با × نشان داده شده است. قابل ذکر است که این داده‌ها به دلیل ماهیت غیرخطی که دارند قابل طبقه‌بندی به کمک ماشین بردار پشتیبان خطی نمی‌باشند و به همین دلیل سراغ ماشین بردار پشتیبان مبتنی بر هسته‌ها خواهیم رفت.

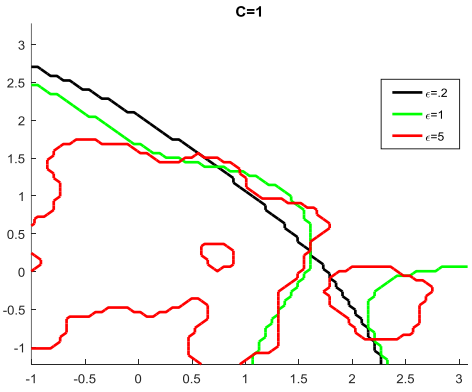
پیش از آنکه به بررسی عملکرد ماشین بردار پشتیبان مبتنی بر هسته‌ها با رتبه پایین در مقایسه با حالت رتبه کامل آن بپردازیم لازم است تاثیر پارامترهای موجود بر عملکرد ماشین بردار پشتیبان در حالت رتبه پایین ارایه شده در این مقاله بررسی گردد. به این منظور ساختار ماشین بردار پشتیبان مبتنی بر تقریب رتبه پایین هسته‌ها روی این نقاط شبیه‌سازی شده به منظور طبقه-بندی آنها اعمال می‌گردد. در اینجا به منظور طبقه‌بندی مبتنی بر تقریب رتبه پایین هسته‌ها از هسته گاوسی استفاده می‌کنیم که در جدول ۱ مطرح شد و دارای پارامتر ϵ است. همچنین در نهایت برای تست کردن طبقه‌بندی ایجاد شده به کمک ساختار ماشین بردار پشتیبان مبتنی بر هسته با تقریب رتبه پایین، ۱۰ نمونه از هر جامعه را (جمعا ۲۰ نمونه) در نظر می‌گیریم. این نقاط برای جامعه اول با علامت ⊗ و برای جامعه دوم با علامت ⊞ در شکل ۱ مشخص هستند.

همان‌طور که گفته شد، هسته گاوسی دارای پارامتر ϵ است که به نظر می‌رسد تاثیر بسیاری بر طبقه‌بندی داشته باشد. از سوی دیگر پارامتر C به عنوان یک کران بالا برای α در رابطه (۱۴)، می‌تواند بر عملکرد طبقه‌بندی تاثیرگذار باشد. در شکل ۲ با در نظر گرفتن $C=1$، تاثیر پارامتر ϵ بر عملکرد طبقه‌بندی بررسی شده است. همچنین در شکل ۳ با در نظر گرفتن $\epsilon=1$، تاثیر پارامتر C بر عملکرد طبقه‌بندی در نظر گرفته شده است.

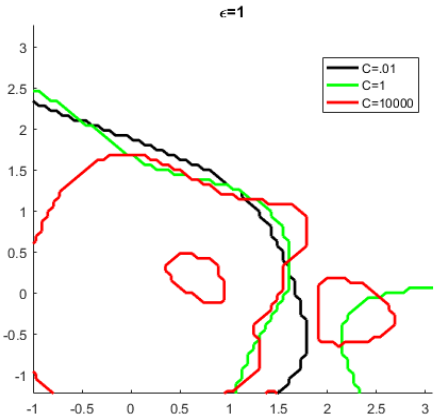
همان‌طور که در شکل های ۲ و ۳ مشاهده می‌شود، انتخاب این پارامترها در عملکرد طبقه‌بندی بسیار تاثیرگذار هستند طوری‌که مقادیر بزرگ پارامتر ϵ باعث ایجاد رویکرد محلی‌تر در طبقه‌بندی می‌گردد در حالی که مقادیر کوچک آن از نگاه محلی کمتری برخوردار بوده و نواحی طبقه‌بندی شده را بزرگتر در نظر می‌گیرد. رفتار مشابهی نیز برای پارامتر C رخ می‌دهد.



شکل ۱: داده های آموزشی دو جامعه نرمال آمیخته با ماتریس های کوواریانس همانی

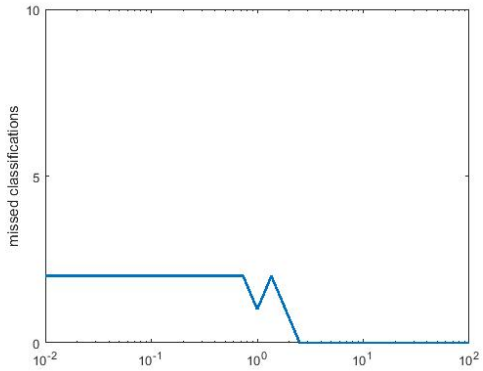


شکل ۲: بررسی تاثیر پارامتر ϵ با $C=1$ در طبقه‌بندی به کمک ماشین بردار پشتیبان مبتنی بر هسته با تقریب رتبه پایین

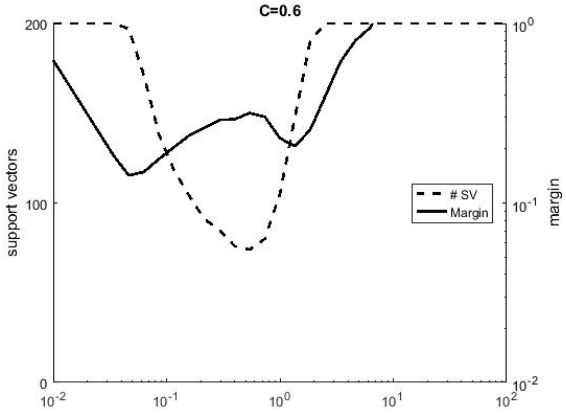


شکل ۳: بررسی تاثیر پارامتر C با $\epsilon=1$ در طبقه‌بندی به کمک ماشین بردار پشتیبان مبتنی بر هسته با تقریب رتبه پایین

ملاک دیگری که به کمک آن می‌توان عملکرد ماشین بردار پشتیبان مبتنی بر تقریب رتبه پایین هسته را در طبقه‌بندی ارزیابی نمود تعداد نقاط تستی می‌باشد که به درستی طبقه‌بندی نشده است. در واقع ابزار طبقه‌بندی مطرح شده نتوانسته است نقاط تست را به درستی طبقه‌بندی نماید. این موضوع در شکل ۴ برای C=0.6 بر حسب تغییرات پارامتر ϵ نشان داده شده است. همچنین تعداد نقاط بردار پشتیبان و حاشیه $\frac{1}{||w||}$ در طبقه‌بندی برای C=0.6 بر حسب تغییرات پارامتر ϵ در شکل ۵ نشان داده شده است. قابل توجه است همان‌طور که در فرم بهینه‌سازی درجه دوم مطرح شد، هر چقدر حاشیه مورد نظر بزرگتر باشد سطح اطمینان در طبقه‌بندی افزایش می‌یابد به‌علاوه تعداد نقاط پشتیبان کمتر منجر به افزایش کارایی در طبقه بندی می‌گردد. در شکل‌های ۶ و ۷ نیز تعداد نقاط پشتیبان و حاشیه مورد نظر برای $\epsilon = 0.01$ و مقادیر متفاوت C نشان داده شده است.

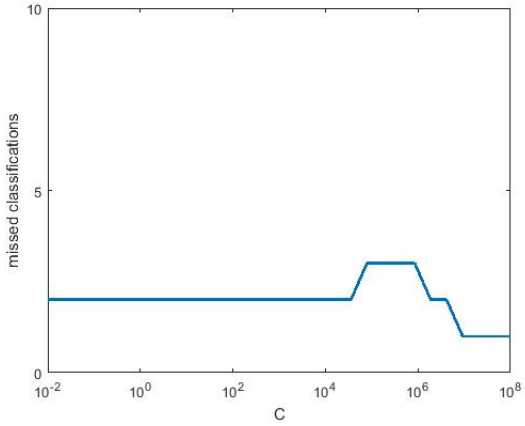


شکل۴: تعداد نقاط تست که درست طبقه‌بندی نشده اند برای C=0.6 بر حسب تغییرات پارامتر ϵ

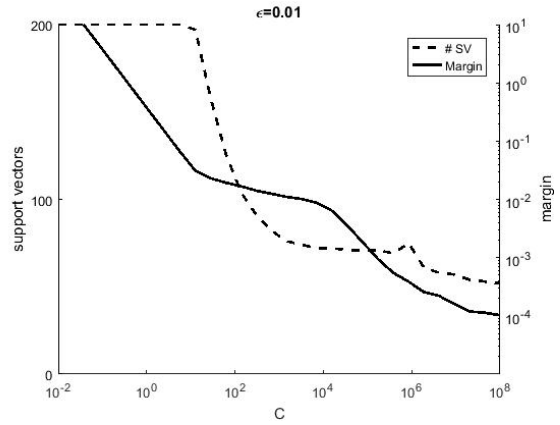


شکل۵: تعداد نقاط بردار پشتیبان و حاشیه $\frac{1}{||w||}$ در طبقه‌بندی برای C=0.6 بر حسب تغییرات پارامتر ϵ

با توجه به شکل‌های ۴ و ۵، بزرگی حاشیه ایجاد شده معیار مناسبی برای تعیین مقدار بهینه برای پارامتر ϵ نیست. در واقع هنگامی که پارامتر ϵ به سمت صفر میل می‌کند حاشیه طبقه-بندی بزرگ می‌شود، در حالی‌که با کوچک شدن پارامتر ϵ، طبقه-بندی در تشخیص نقاط تست عملکرد نامناسبی دارد. از سوی دیگر، اگر چه به حداقل رساندن تعداد بردارهای پشتیبانی از نقطه نظر محاسباتی مطلوب است، اما با توجه به شکل های ۴ و ۵ این معیار نیز در تعیین پارامتر بهینه چندان کارآمد نیست.



شکل۶ : تعداد نقاط بردار پشتیبان و حاشیه $\frac{1}{||w||}$ در طبقه‌بندی برای ϵ=0.01 بر حسب مقادیر متفاوت C

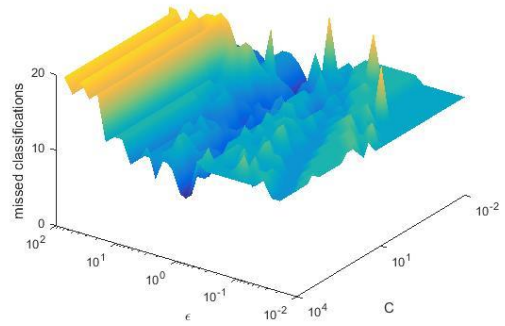


شکل ۷: تعداد نقاط بردار پشتیبان و حاشیه $\frac{1}{||w||}$ در طبقه‌بندی برای $\epsilon = 0.01$ برای بر حسب تغییرات پارامتر C

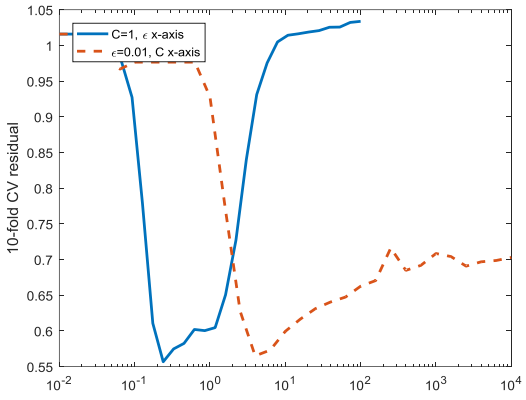
همچنین در شکل های۶ و۷ پارامتر ϵ ثابت در نظر گرفته شده و به ازای مقادیر متفاوت C، تعداد نقاط تستی که درست طبقه‌بندی نشده‌اند (شکل ۶) و حاشیه طبقه‌بندی به همراه تعداد نقاط بردار پشتیبان (شکل ۷) گزارش شده است. وضعیت در اینجا نیز چندان متفاوت نیست و نمی‌توان از نتایج بدست آمده به عنوان معیاری برای تعیین پارامترهای بهینه استفاده کرد. به نظر می‌رسد که یک طبقه‌بندی خوب برای مقادیر کوچک C اتفاق می‌افتد و این در حالی است که تعداد بردارهای پشتیبان کمتر به ازای مقادیر بزرگ C بدست می‌آید. همچنین مقدار بزرگ C باعث کاهش سرعت طبقه‌بندی می‌گردد زیرا به زمان بیشتری برای حل بهینه‌سازی معادله درجه دوم واقع در ساختار ماشین بردار پشتیبان مبتنی بر تقریب رتبه هسته‌ها نیاز دارد و فضای جستجوی بزرگتری را می-طلبد.

۵- روش اعتبارسنجی متقابل برای یافتن پارامترهای بهینه

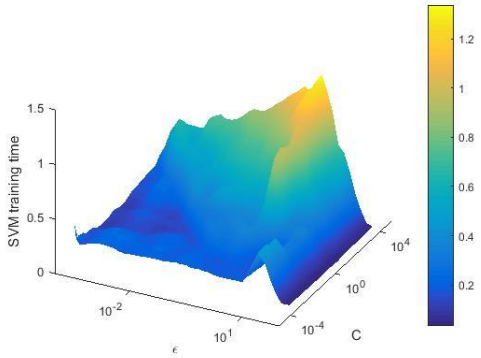
بر اساس نتایج بدست آمده واضح است که پارامترهای موجود بر عملکرد طبقه‌بندی تاثیر بسزایی دارد. از سوی دیگر استفاده از حاشیه طبقه‌بندی و تعداد نقاط بردار پشتیبان به تنهایی نمی‌توانند معیارهای مناسبی برای تعیین بهینه پارامترها در طبقه‌بندی باشند. به همین منظور از روش اعتبار سنجی متقابل با تعداد لایه مناسب برای پیدا کردن پارامتر بهینه مورد نظر استفاده می‌کنیم. اعتبار سنجی متقابل[۸] یک روش ارزیابی مدل است که تعیین می‌نماید نتایج یک تحلیل آماری بر روی یک مجموعه داده تا چه اندازه قابل تعمیم و مستقل از داده‌های آموزشی است. این روش به‌طور ویژه در کاربردهای پیش‌بینی مورد استفاده قرار می‌گیرد تا مشخص شود مدل موردنظر تا چه اندازه در عمل مفید خواهد بود. به‌طور کلی یک دور از اعتبارسنجی ضربدری شامل افراز داده‌ها به دو زیرمجموعه مکمل، انجام تحلیل بر روی یکی از آن زیرمجموعه‌ها (داده‌های آموزشی) و اعتبارسنجی تحلیل با استفاده از داده‌های مجموعه دیگر است (داده‌های اعتبارسنجی یا آزمایش). برای کاهش پراکندگی، عمل اعتبارسنجی چندین بار با افرازهای مختلف انجام و از نتایج اعتبارسنجی‌ها میانگین گرفته می‌شود. در اعتبارسنجی متقابل با k لایه، داده‌ها به k زیرمجموعه افراز می‌شوند. از این k زیرمجموعه، هر بار یکی برای اعتبارسنجی و k-1 تای دیگر برای آموزش بکار می‌روند. این روال k بار تکرار می‌شود و همه داده‌ها دقیقاً یک بار برای آموزش و یک بار برای اعتبارسنجی بکار می‌روند. در نهایت میانگین نتیجه این k بار اعتبارسنجی به‌عنوان یک تخمین نهایی برگزیده می‌شود. به‌طور معمول از روش اعتبارسنجی پنج لایه یا ده لایه در پژوهش‌های مدل‌سازی و پیش‌بینی استفاده می‌شود. در اینجا برای تعیین پارامتر بهینه از روش اعتبارسنجی متقابل ده لایه استفاده می‌کنیم. نتایج حاصل از روش اعتبارسنجی متقابل برای پیدا کردن پارامتر بهینه در شکل ۸ ارایه شده است. همچنین در شکل ۹ یکی از پارامترها ثابت نگه داشته می‌شود و دیگری در محدوده $[10^{-2},10^{2}]$ تغییر می‌کند تا روش اعتبارسنجی متقابل به منظور پیدا کردن پارامتر روی آن اعمال گردد. بر اساس این شکل طبقه‌بندی صحیح هنگامی رخ می‌دهد که پارامتر ϵ کاهش می‌یابد و این معادل با وقتی است که پارامتر C در حال افزایش است. همچنین زمان لازم برای آموزش داده‌ها در تقریب رتبه پایین ماشین بردار پشتیبان بر حسب پارامترهای موجود در شکل ۱۰ نشان داده شده است که به‌خوبی نشان می دهد انتخاب این پارامترها بر زمان آموزش در تقریب رتبه پایین نیز تاثیرگذار است. بعد از آنکه به بررسی عملکرد ماشین بردار پشتیبان مبتنی بر هسته با تقریب رتبه پایین پرداختیم، به بررسی قابلیت این ساختار جدید در کاهش زمان طبقه بندی داده‌ها در مقایسه با روش رتبه کامل خواهیم پرداخت. در این ساختار جدید، معادله بهینه‌سازی درجه دوم موجود در ماشین بردار پشتیبان از ساختار بسیار ساده‌تری برخوردار خواهد بود که این امر زمان لازم برای آموزش داده‌ها در طبقه‌بندی ماشین بردار پشتیبان مبتنی بر هسته در حالت تقریب رتبه پایین را در مقایسه با حالت رتبه کامل کاهش خواهد داد. این موضوع در شکل ۱۱ به‌خوبی نشان داده شده است.



شکل ۸: پارامترهای ماشین بردار پشتیبان بر حسب خطای طبقه‌بندی با روش اعتبارسنجی متقابل

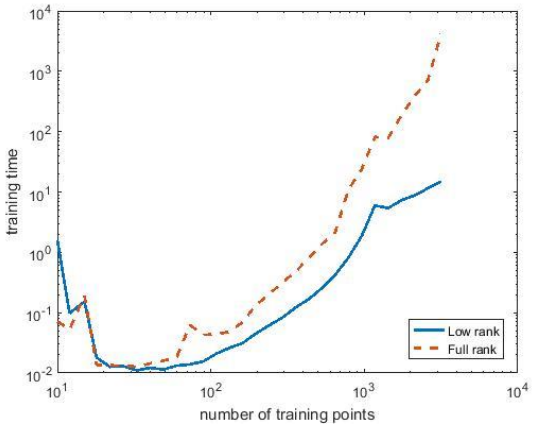


شکل ۹: روش اعتبار سنجی متقابل به منظور بهینه کردن پارامترها



شکل ۱۰: زمان آموزش داده‌ها در تقریب رتبه پایین ماشین بردار پشتیبان بر حسب پارامترهای موجود

به عبارتی برای نشان دادن این موضوع تعداد نقاط متفاوتی در نظر گرفته شده است که از ۱۰ نقطه تا ۱۰۰۰۰ نقطه متغیر می‌باشد. برای تعداد نقاط کم، این اختلاف زمانی ناچیز است هر چند هنوز تقریب رتبه پایین نسبت به حالت رتبه کامل عملکرد بهتری دارد، در حالی که برای تعداد نقاط زیاد این اختلاف زمانی بسیار زیاد است و این موضوع عملکرد بسیار خوب تقریب رتبه پایین هسته در مقایسه با حالت رتبه کامل را نشان می‌دهد. در این محاسبات پارامتر $\epsilon=0.01$، و $C=1$ در نظر گرفته شده است. همچنین تعداد جملات به کار رفته در تقریب رتبه پایین هسته (M) برابر با ۰,۱ تعداد نقاط آموزشی قرار داده شده است.



شکل ۱۱. زمان آموزش (به ثانیه) بر حسب تعداد نقاط در ماشین بردار پشتیبان تقریب رتبه کامل و رتبه پایین

۶- کاربردهایی از تقریب رتبه پایین ماشین بردار پشتیبان مبتنی بر هسته

۶-۱- کاربرد روش در تشخیص حضور افراد

تشخیص دقیق حضور یا عدم حضور افراد در فضاهای موجود در ساختمان‌ها می‌تواند موجب صرفه‌جویی قابل توجهی در میزان مصرف انرژی گردد. زمانی که اطلاعات حضور افراد به عنوان ورودی در اختیار سیستم‌های کنترل گرمایش، سرمایش و تهویه هوای ساختمان‌ها قرار گیرد، موجب صرفه‌جویی بین ۳۰ تا ۸۰ درصدی در مصرف انرژی می‌گردد [۲۴].

امروزه با افزایش قابلیت حسگرها به همراه توسعه ظرفیت‌های محاسباتی در سامانه‌های اتوماسیونی، تعیین دقیق میزان حضور افراد می‌تواند تاثیر قابل توجهی در کاهش مصرف انرژی در تهویه مطبوع و روشنایی ساختمان‌ها داشته باشد.

در این قسمت از روش طبقه‌بندی ماشین بردار پشتیبان مبتنی بر هسته‌ها در حالت رتبه کامل و رتبه پایین برای تشخیص حضور افراد در فضاهای موجود در ساختمان‌ها استفاده می‌کنیم. داده‌های مورد استفاده با استفاده از حسگرهایی با دقت مناسب تهیه شده است. این داده‌ها در یک اتاق کار در دانشگاه مونیز در کشور بلژیک تدوین شده است. این داده‌ها شامل اطلاعاتی از میزان نور، دما، رطوبت، نسبت رطوبت و CO_2موجود در هوا در زمان‌های مختلف در خلال ماه فوریه می‌باشد. لازم به ذکر است، نسبت رطوبت که به عنوان یکی از ویژگی‌ها در داده‌ها مدنظر قرار گرفته است با استفاده از دمای اندازه‌گیری شده و رطوبت نسبی قابل محاسبه است، جزییات این امر در منبع [۲۵] ذکر شده است.

برای آن‌که تاثیر تقریب رتبه پایین در کاهش زمان محاسباتی ماشین بردار پشتیبان مبتنی بر هسته‌ها به خوبی مشخص گردد، طبقه‌بندی داده‌ها را با تعداد داده‌های مختلف انجام می‌دهیم. به این منظور تعداد داده‌ها را به ترتیب برابر ۱۰۰۰، ۲۰۰۰ و ۳۰۰۰ در نظر گرفته‌ایم که هر کدام از داده‌ها به ترتیب دارای ۵ ویژگی دما، رطوبت، نور، میزان CO_2 و نسبت رطوبت می‌باشند. حضور یا عدم حضور افراد در اتاق کار نیز به ترتیب با $+1$ و -1 در داده-های هدف مشخص شده است. این داده‌ها در [۲۵] موجود می-باشند.

نتایج حاصل از طبقه‌بندی داده‌ها به کمک ماشین بردار پشتیبان در حالت رتبه کامل و رتبه پایین در جدول شماره ۲ با یکدیگر مقایسه شده‌اند و خطای طبقه‌بندی و زمان آموزش این مدل‌ها در حالت رتبه پایین و رتبه کامل گزارش شده است. از نتایج به‌دست آمده مشخص است که روش رتبه پایین به طور قابل توجهی زمان محاسباتی روش را کاهش می‌دهد، این در حالی است که عملکرد طبقه‌بندی روش همچنان در حد مطلوب باقی می‌ماند. لازم به ذکر است خطای طبقه‌بندی تعداد داده‌هایی را مشخص می‌کند که به درستی طبقه‌بندی نشده‌اند. با توجه به تعداد اندک داده‌هایی که به درستی طبقه‌بندی نشده‌اند (به نسبت کل داده‌های موجود)، عملکرد طبقه‌بندی در هر دو حالت رتبه کامل و رتبه پایین مطلوب است. به‌علاوه، نتایج ارایه شده در جدول شماره ۲ نشان می‌دهد با افزایش تعداد داده‌ها، روش ماشین بردار پشتیبان رتبه پایین در کاهش زمان آموزش و حفظ عملکرد طبقه‌بندی همچنان از موفقیت قابل توجهی برخوردار است.

جدول ۲ : مقایسه عملکرد روش رتبه پایین و رتبه کامل در تشخیص حضور افراد

نام مدل	تعداد داده‌ها	زمان آموزش (ثانیه)	خطای طبقه‌بندی
رتبه پایین	۱۰۰۰	۱,۸۸	۲۱
رتبه کامل	۱۰۰۰	۱۰,۶۹	۱۲
رتبه پایین	۲۰۰۰	۹,۴۲	۷۱
رتبه کامل	۲۰۰۰	۷۲,۴۷	۵۵
رتبه پایین	۳۰۰۰	۲۶,۷۵	۹۲
رتبه کامل	۳۰۰۰	۲۱۵,۶۳	۷۶

۶-۲- کاربرد روش در تشخیص پست‌های الکترونیک اسپم

به سوءاستفاده از ابزارهای الکترونیکی نظیر پست الکترونیک برای ارسال پیام به تعداد زیاد و به صورت ناخواسته اسپم می‌گویند. با توجه به هزینه اندک این روش نسبت به پست سنتی که در گذشته برای ارسال پلاک به پلاک تبلیغات مورد استفاده قرار می‌گرفت و همچنین ناقص بودن قوانین بین‌المللی برای محدود کردن آنها، این قبیل پست‌های الکترونیک در سطح وسیعی ارسال می‌شوند. پست الکترونیکی تجاری ناخواسته یا همان اسپم در اواسط سال ۱۹۹۰ وارد اینترنت شد و کم کم تبدیل به یک معضل بزرگ شد، به صورتی که امروزه با یک تخمین محافظه‌کارانه ۸۰ تا ۸۵ درصد پست‌های الکترونیک را در بر می‌گیرد و در بعضی منابع از ۹۰ درصد بالاتر می‌رود [۲۶]. اگرچه ابزارهای متنوعی به صورت عمومی برای شناسایی اسپم ها وجود دارد، اما طراحی ابزارهای شخصی سازی شده برای شناسایی اسپم ها می تواند از کارایی بالاتری برخوردار باشد.

در این قسمت از روش طبقه‌بندی ماشین بردار پشتیبان مبتنی بر هسته‌ها در حالت رتبه کامل و رتبه پایین برای تشخیص پست های الکترونیک اسپم و غیر اسپم استفاده می‌کنیم. با استفاده از این ابزار قادر خواهیم بود یک سیستم تشخیص اسپم به صورت شخصی سازی شده تهیه نماییم.

اسپم ها عمدتا توسط مراکزی که پست‌های الکترونیک تجاری را ارسال می کنند و یا از سوی افرادی که هرزنامه منتشر می کنند تولید می گردد. از سوی دیگر اکثر پست‌های الکترونیک غیر اسپم از طرف همکاران و پست‌های الکترونیک شخصی ارسال می گردند. بر این اساس قادر خواهیم بود ویژگی هایی را تعریف نماییم که در نوع پست الکترونیک دریافتی تاثیرگذارند، به عنوان نمونه هر پست الکترونیک که در بر دارنده نام فرد دریافت کننده باشد می تواند یک پست الکترونیک غیر اسپم باشد. در این مطالعه داده‌های مورد استفاده به صورت شخصی سازی شده و بر اساس اطلاعات یک کاربر مشخص تهیه و تدوین شده است. این داده‌ها در بردارنده ۴۶۰۰ نمونه می باشد که هر کدام از آنها دارای ۵۷ ویژگی است. اغلب این ویژگی ها شامل اطلاعاتی از درصد حضور برخی کلمات و کاراکترهای مشخص در متن پست الکترونیک دریافتی است، که میزان فراوانی آنها را در متن مشخص می کند. برخی دیگر از این ویژگی ها به میزان حضور و نحوه حضور دنباله های حروف بزرگ به هم پیوسته در متن اشاره دارد. در حقیقت ۴۸ ویژگی اول به درصد حضور برخی کلمات مشخص در متن اختصاص دارد، ۶ ویژگی بعدی میزان حضور برخی کاراکتر های خاص را تعیین می کند و سه ویژگی آخر به بررسی دنباله های حروف بزرگ به‌هم پیوسته موجود در متن اشاره دارد. جزییات این امر در [۲۷] ارایه شده است. تعیین اینکه ایمیل دریافتی اسپم یا غیر اسپم می باشد نیز به ترتیب با $+1$ و -1 در داده‌های هدف مشخص شده است. این داده‌ها در [۲۷] موجود می‌باشند.

عملکرد روش رتبه پایین و رتبه کامل در تشخیص پست‌های الکترونیک اسپم در جدول شماره ۳ با یکدیگر مقایسه شده است. برای آنکه امکان مطالعه هر چه بهتر روش فراهم گردد، نتایج عددی با تعداد داده‌های متفاوت ارزیابی و گزارش شده است. نتایج به دست آمده به خوبی قابلیت روش رتبه پایین را در کاهش زمان آموزش ماشین بردار پشتیبان نشان می دهد. این در حالی است که عملکرد طبقه‌بندی همچنان در حد مطلوب باقی مانده است. مشابه نتایج ارایه شده در جدول شماره ۲، با افزایش تعداد داده‌ها عملکرد مطلوب روش رتبه پایین در مقایسه با روش رتبه کامل همچنان حفظ می‌گردد. اما کاهش ایجاد شده در زمان آموزش ماشین بردار پشتیبان در حالت رتبه کامل و رتبه پایین در جدول‌های شماره ۲ و ۳ با یکدیگر تفاوت دارد. در کاربرد ارایه شده در تشخیص حضور افراد ، استفاده از روش رتبه پایین حدود ۸ برابر زمان آموزش را بهبود داده است. اما در کاربرد روش ماشین بردار پشتیبان در تشخیص اسپم، زمان آموزش توسط روش رتبه پایین فقط در حدود ۲ برابر بهبود یافته است. این تفاوت در عملکرد، از تفاوت در تعداد ویژگی‌های موجود در این داده‌ها نشات می‌گیرد. به عنوان نمونه تعداد ویژگی‌ها در تشخیص حضور افراد تنها شامل ۵ ویژگی بوده در حالی که تعداد ویژگی‌ها در تشخیص پست‌های الکترونیک اسپم برابر با ۵۷ ویژگی می‌باشد. به عبارت دیگر افزایش قابل توجه تعداد ویژگی‌ها می‌تواند بر عملکرد روش تاثیرگذار باشد، اما همچنان روش رتبه پایین از مزیت محاسباتی قابل توجهی نسبت به روش رتبه کامل برخوردار است.

جدول ۳ : مقایسه عملکرد روش رتبه پایین و رتبه کامل در تشخیص پست‌های الکترونیک اسپم

نام مدل	تعداد داده‌ها	زمان آموزش (ثانیه)	خطای طبقه‌بندی
رتبه پایین	۲۲۰۰	۱۵,۲۱	۳۶۴
رتبه کامل	۲۲۰۰	۵۳,۷۴	۲۹۴
رتبه پایین	۳۴۰۰	۷۱,۳۲	۴۷۶
رتبه کامل	۳۴۰۰	۱۹۳,۲۱	۳۸۱
رتبه پایین	۴۶۰۰	۲۳۸,۳۷	۶۸۲
رتبه کامل	۴۶۰۰	۵۰۱,۵۷	۵۵۴

۷- بحث و نتیجه گیری

در این مقاله، به‌دلیل ضعف ماشین بردار پشتیبان خطی در طبقه-بندی داده‌های غیرخطی، استفاده از ساختار ماشین بردار پشتیبان

مبتنی بر هسته‌ها به دلیل استفاده از فضای ویژگی داده‌ها به جای استفاده از داده‌های اصلی مدنظر قرار گرفت. در حقیقت بر اساس این رویکرد جدید، امکان طبقه‌بندی غیرخطی داده‌ها فراهم می-آید. همان‌طور که مطرح شد، یکی از چالش‌های موجود در این رویکرد افزایش پیچیدگی‌های محاسباتی و نهایتاً افزایش زمان لازم برای طبقه‌بندی بود طوری‌که برای مجموعه داده‌های بزرگ از کارایی لازم برخوردار نیست. عمدتاً این افزایش زمان محاسباتی به دلیل ظاهر شدن هسته در حل مسئله بهینه‌سازی درجه دوم بود که با استفاده از تقریب رتبه پایین ارائه شده در این مقاله قادر خواهیم بود این مشکل را تا حد زیادی مرتفع نماییم. در این تکنیک با بکارگیری سری تقریبی قطع شده از هسته موجود، مسئله بهینه‌سازی درجه دوم در ساختار ماشین بردار پشتیبان مبتنی بر هسته‌ها با یک مسئله بهینه‌سازی بسیار ساده‌تر جایگزین شد. در رویکرد جدید ارائه شده، محاسبات برداری و تجزیه‌های ماتریسی مورد نیاز بسیار سریع‌تر انجام می‌شود که این تغییرات منجر به حل سریع‌تر مسئله بهینه‌سازی درجه دوم موجود و افزایش کارایی گردید. نهایتا نتایج عددی ارائه شده در طبقه‌بندی داده‌ها در برخی مسایل کاربردی با استفاده از تقریب رتبه پایین ماشین بردار پشتیبان مبتنی بر هسته‌ها نشان داد که ضمن حفظ عملکرد طبقه‌بندی، زمان محاسباتی بطور قابل توجهی کاهش می‌یابد.

مراجع

[1] T. Hastie, R. Tibshirani, and J. Friedman, Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer Series in Statistics, Springer-Verlag, New York, 2009.

[2] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, 3rd Edition., Elsevier Inc., Waltham, USA, 2012.

[3] H. Nickisch, C. E. Rasmussen, Approximations for Binary Gaussian Process Classification, Journal of Machine Learning Research, Vol. 9, pp. 2035-2078, 2008.

[4] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[5] R. A. Araújo, A. L. I. Oliveira and S. Meira, A morphological neural network for binary classification problems, Engineering Applications of Artificial Intelligence, Vol. 65, pp. 12-28, 2017.

[6] G. Qian, L. Zhang, A simple feedforward convolutional conceptor neural network for classification, Applied Soft Computing, Vol. 70, pp. 1034-1041, 2018.

[7] L. Wang, Support Vector Machines: Theory and Applications, Springer, Berlin, 2005.

[8] I. Steinwart, A. Christmann, Support Vector Machines, Information Sciences and Statistics, Springer-Verlag, New York, 2008.

[9] V. Vapnik, The nature of statistical learning theory, Springer-Verlag, New York, 2013.

[10] B. Scholkopf and A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, 2002.

[11] G. Fasshauer, M. McCourt, Kernel-based Approximation Method using Matlab, World Scientific Publishing, 2016.

[12] L. Chih-Jen, Support Vector Machines and Kernel Methods: Status and Challenges, Talk at K. U. Leuven Optimization in Engineering Center, 2013.

[13] X. Nguyen, L. Huang and A. D. Joseph, Support Vector Machines, Data Reduction, and Approximate Kernel Matrices, Springer-Verlag, Berlin Heidelberg, 2008.

[14] F. Bach, Sharp analysis of low-rank kernel matrix approximations, JMLR: Workshop and Conference Proceedings, Vol. 30, pp. 1–25, 2013.

[15] M. S. Andersen, L. Vandenberghe, Support vector machine training using matrix completion techniques, Technical report, University of California, Los Angeles, 2010.

[16] K. Zhang, L. Lan, Z. Wang, F. Moerchen, Scaling up Kernel SVM on Limited Resources: A Low-rank Linearization Approach, Appearing in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS), La Palma, Canary Islands, 2012.

[17] L. Lan, Z. Wang, S. Zhe, W. Cheng, J. Wang and K. Zhang, Scaling Up Kernel SVM on Limited Resources: A Low-Rank Linearization Approach, IEEE Transactions on Neural Networks and Learning Systems, Vol. 30, No. 2, pp. 369-378, 2018.

[18] K. Chen, R. Li, Y. Dou, Z. Liang and Q. Lv, Ranking Support Vector Machine with Kernel Approximation, Computational Intelligence and Neuroscience, Vol. 2017, 4629534. doi:10.1155/2017/4629534, 2017.

[19] S. Fine, and K. Scheinberg, Efficient SVM training using low-rank kernel representations, Journal of Machine Learning Research, Vol. 2, pp. 243–264, 2002.

[20] G. W. Flake, and S. Lawrence, Efficient SVM regression training with SMO, Machine Learning, Vol. 46, pp. 271–290, 2002.

[21] C. Yang, R. Duraiswami and L. S. Davis, Efficient kernel machines using the improved fast Gauss transform, Advances in Neural Information Processing Systems, Vol. 17, pp. 1561–1568, 2004.

[22] T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Trans. Electron, Vol. 14, pp. 326–334. 1965

[23] J. Mercer, Functions of positive and negative type, and their connection with the theory of integral equations, Philos. T. R. Soc. Lond. Vol. 209, pp. 441–458, 1909.

[24] J. Brooks, S. Kumar, S. Goyal, R. Subramany and P. Barooah, Energy-efficientcontrol of under-actuated HVAC zones in commercial buildings, Energy and Buildings. Vol. 93, pp. 160–168, 2015.

[25] L. M. Candanedo, V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO2measurements using statistical learning models, Energy and Buildings. Vol. 112, pp. 28–39, 2016.

[26] Wikipedia contributors. "Email spam." Wikipedia, The Free Encyclopedia. Wikipedia, 2019.

[27] C. L. Blake, C. J. Merz. UCI repository of machine learning databases, URL https://archive.ics.uci.edu/ml/support/spambase, 1998.

پاورقی‌ها:

[1] Supervised Learning

[2] Unsupervised Learning

[3] Classification

[4] Regression

[5] Full Matrix

[6] Mercer’s Theorem

[7] Cover’s Theorem

[8] Cross Validation