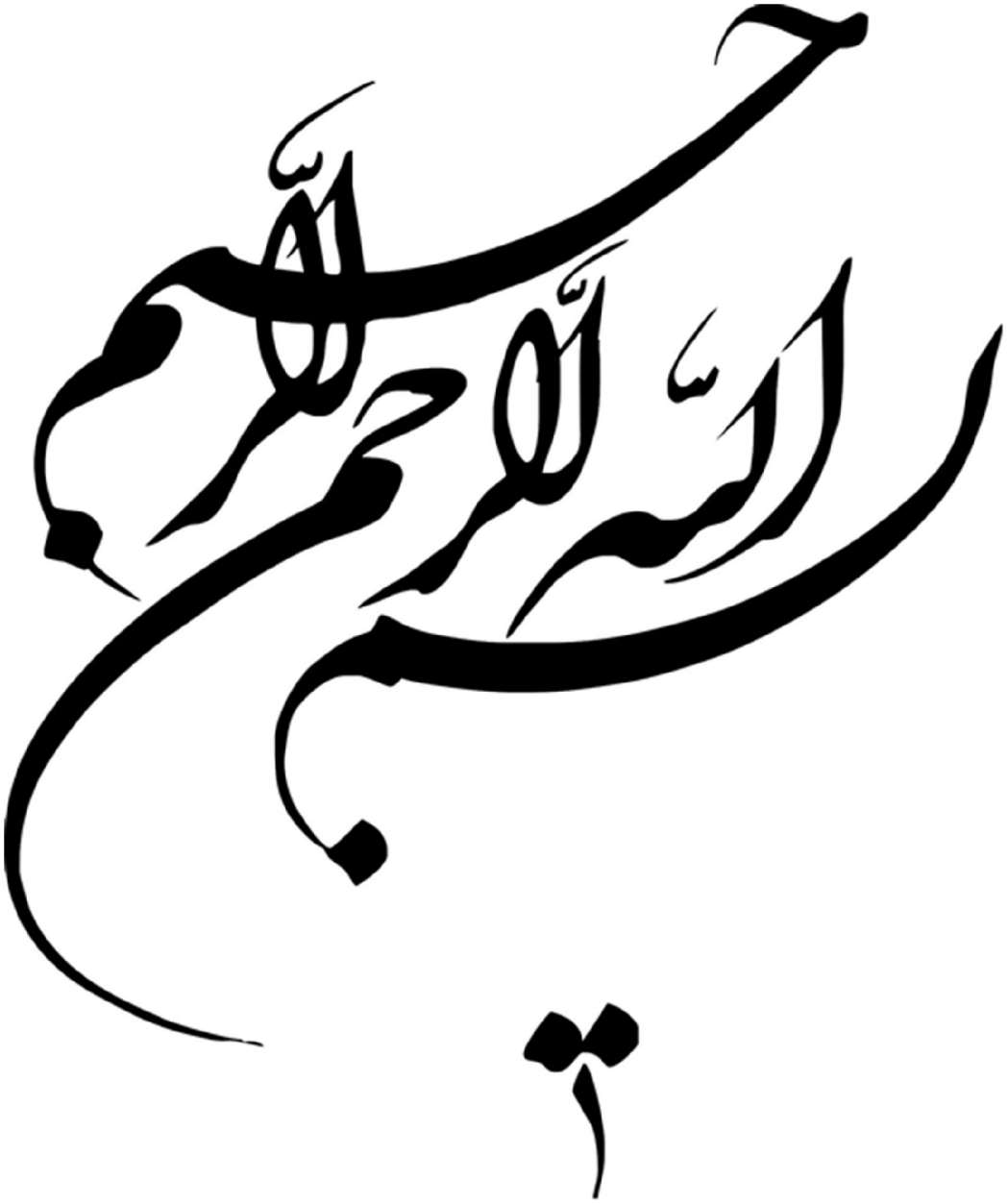
بسم الله الرحمن الرحيم

به نام خدا

دانشکده مهندسی
گروه کامپیوتر
آزمایشگاه شناسایی الگو

پایان‌نامه کارشناسی ارشد

تلفیق توصیفگرهای داده تک کلاسی مبتنی بر بردار پشتیبان

نگارش

نشاط صالحی

استاد راهنما

دکتر هادی صدوقی یزدی

استاد مشاور

دکتر عباس قائمی بافقی

مهر ۹۱

به نام خدا

دانشکده مهندسی گروه کامپیوتر

آزمایشگاه شناسایی الگو

پایان نامه کارشناسی ارشد

عنوان:

تلفیق توصیفگرهای داده تک کلاسی مبتنی بر بردار پشتیبان

دانشجو:

نشاط صالحی

کمیته ممتحنین:

استاد راهنما: **دکتر هادی صدوقی یزدی** امضا:

استاد مشاور: **دکتر عباس قائمی بافقی** امضا:

استاد داور: **دکتر محمدباقر نقیبی سیستانی** امضا:

استاد داور و نماینده کمیته تحصیلات تکمیلی: **دکتر رصا منصفی** امضا:

تاریخ ۱۳۹۱/۰۷/۱۵

تعهدنامه

اینجانب **نشاط صالحی** دانشجوی دوره کارشناسی ارشد **رشته مهندسی کامپیوتر هوش مصنوعی** دانشکده مهندسی دانشگاه فردوسی مشهد نویسنده پایان‌نامه **تلفیق توصیفگرهای داده تک کلاسی مبتنی بر بردار پشتیبان** تحت راهنمایی استاد ارجمند **جناب آقای دکتر هادی صدوقی یزدی و جناب آقای دکتر عباس قائمی بافقی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود و یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه فردوسی مشهد می‌باشد و مقالات مستخرج با نام «دانشگاه فردوسی مشهد» و یا "Ferdowsi University of Mashhad" به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از رساله رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است، اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

نشاط صالحی

تقدیم به

پدرم به استواری کوه

روح پاک مادرم به زلالی چشمه

همسرم به صمیمیت باران

و اساتید گرانقدرم...

تقدیر و تشکر

سپاس و آفرین ایزد جهان آفرین راست.سپاس خدای را که حق ستایش برتر از حد ستایشگران است و نعمت‌هایش فوق اندیشه شمارشگران. سپاس خدای را که پای اندیشه تیزگام در راه شناسایی او لنگ است و سر فکر ژرف رو به دریای معرفتش بر سنگ.

از این رو بر خود لازم می‌دانم که سپاس گویم تمام کسانی را که از آن‌ها آموخته‌ام. تشکر می‌نمایم از استاد گرانقدرم جناب آقای دکتر هادی صدوقی یزدی که فروغ اندیشه‌شان زداینده‌ی ظلمت جهل در حرکت به سوی علم و آگاهی است و نیز ارج می‌نهم راهنمایی‌های استاد بزرگوار جناب آقای دکتر قائمی بافقی، که بزرگوارانه مرا راهنمایی نمودند.

سپاس خود را نثار همه معلمان و اساتیدم می‌دارم که از نخستین مراحل تحصیل تاکنون، مشوق، معلم و راهنمای من در مسیر علم و زندگی بودند.

سپاس می‌گویم خانواده عزیزم را که همیشه پشتیبان و مشوقم بوده‌اند و شرایطی را فراهم آوردند تا در محیطی مطلوب و پر از صفا و آرامش مراتب تحصیل را طی نمایم. در انتها قدردانی می‌کنم زحمات همسرم و جملگی دوستانی‌را که در تهیه این اثر سهمی داشته‌اند، امید که برای پویندگان راه دانش مثمر ثمر واقع گردد.

چکیده

توصیف‌گر داده مبتنی بر بردار پشتیبان (SVDD)، یک طبقه‌بند باناظر تک کلاسه است. هدف این طبقه‌بند مرزی، بهینه کردن حجم دایره (ابرکره) اطراف مجموعه هدف خطی یا غیرخطی می‌باشد. حداقل پیچیدگی زمانی این گونه طبقه‌بندها، $O(n^3)$ است؛ در نتیجه با افزایش تعداد نمونه‌ها، مسئله برای مجموعه داده‌های حجیم کارایی خود را از دست می‌دهد. هدف اصلی این پایان‌نامه، توسعه SVDD، به منظور ایجاد امکان استفاده از آن در کاربردهای حجیم و افزایش سرعت بدون کاهش کارایی یادگیری در توصیف داده‌ها است. این هدف، با بهره‌گیری از روش‌های تلفیق طبقه‌بندها به منظور افزایش سرعت یادگیری و کاهش نسبی میزان خطا در توصیف داده‌ها برآورده شده است؛ که نتیجه‌ی آن تلفیق طبقه‌بندها با وزن دهی مبتنی بر کارایی آن‌ها می‌باشد. در بند اول روش پیشنهادی، طبقه‌بند SVDD، ابتدا با روش AdaBoost ترکیب گردیده و سپس در بند دوم، ضرایب لاگرانژ طبقه‌بند SVDD، با استفاده از روش تخمین مبتنی بر بیشینه کردن امیدریاضی (EM)، محاسبه شده است که نتیجه‌ی آن دو نوع طبقه‌بند ترکیبی همگرا است. دو روش پیشنهادی از نظر سرعت و دقت با سایر الگوریتم‌های تک کلاسه SVDD مبتنی بر کرنل از جمله FSVDD و طبقه‌بند افزایشی (Inc-SVDD) و جعبه‌ابزار LibSVM-SVDD بر روی مجموعه داده‌های استاندارد UCI و مجموعه داده IDS آزمایش و مقایسه شده‌اند. نتایج حاصل، برتری روش‌های پیشنهادی را از نظر کارایی و سرعت در مرحله آموزش و تست نشان می‌دهد.

کلیدواژه: توصیفگر داده مبتنی بر بردار پشتیبان - داده‌های حجیم - ترکیب طبقه‌بندها – روش آدابوست – روش تخمین مبتنی بر بیشینه کردن امیدریاضی(EM).

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فهرست علائم اختیاری

در لیست زیر به بیان متغیرهای به‌کاررفته در روش پیشنهادی در پایان‌نامه اشاره شده است.

عنوان	توضیح
R	شعاع ابرکره مرز توصیفگر داده
a	مرکز ابرکره مرز توصیفگر داده
ξ_i	متغیر لغزش به ازای هر نمونه
C	پارامتر تنظیم نسبت حجم ابرکره و میزان خطا
β_i	ضریب لاگرانژ به ازای هر نمونه
γ_i	ضریب لاگرانژ به ازای هر نمونه
$\varphi : x \rightarrow \varphi(x)$	تابع انتقال به فضای ویژگی (فضای با ابعاد بالاتر)
$K(x_i, x_j)$	تابع کرنل
σ	پارامتر تابع کرنل RBF
z	نمونه مورد آزمون
ω_i	وزن طبقه‌بند یا وزن ترکیب
h	طیقه‌بند
Z_t	فاکتور نرمال‌سازی
f_k	مقدار برازندگی نمونه kام
P_k	احتمال انتخاب
$L(\theta)$	تابع درست‌نمایی
J	تابع حد پایین درست‌نمایی
θ	بردار پارامترهای نامعلوم در تابع چگالی احتمال
$M(\theta)$	تابع نگاشت از فضای پارامتر به خودش
$Jh(\theta)$	ماتریس ژاکوبی $d \times d$
$\tau_c(\theta; y)$	امیدریاضی تابع مشتق دوم لگاریتم درست‌نمایی نسبت به مشاهدات
$N(\mu_i, \Sigma_i)$	تابع توزیع نرمال
M	تعداد ترکیب از چند تابع توزیع نرمال یا چند طبقه‌بند

عنوان	توضیح
μ_i	بردار میانگین تابع توزیع نرمال
$\sum_i$	ماتریس کوواریانس تابع توزیع نرمال
d	بعد فضای ورودی‌ها
$p(X;\theta)$	تابع توزیع نمونه‌های X با پارامترهای θ
$SVDD(x;\theta)$	تابع توزیع SVDD برای نمونه‌های X با پارامترهای θ
$\lvert SV \rvert$	تعداد بردارهای پشتیبان
k	درصدی از تعداد داده‌ها
$f_k(x)$	تابع تصمیم طبقه‌بند
$\delta(C_k \mid x)$	تابع شبه توزیع احتمالی پسین
$F(x_n;\theta)$	تابع ترکیب چند گوسی
$\Lambda(X;\theta)$	تابع درست‌نمایی
$\lambda(X;\theta)$	تابع لگاریتم درست نمایی
$Q(\theta,\theta^g)$	تابع امیدریاضی شرطی تابع درست‌نمایی
NC	تعداد طبقه‌بندها

فهرست اختصارات

اختصار	توضیح
AdaBoost	Adaptive Boosting
EM	Expectation Maximization
FS_SFS	Filtered and Supported Sequential Forward Search
GMM	Gaussian Mixture Model
IDS	Intrusion Detection Systems
KFN-CBD	K-Farthest-Neighbors-Based Concept Boundary Determination
KKT	Karush–Kuhn–Tucker
LIBSVM	Library for Support Vector Machines
NSV	Non Support Vector
QP	Quadratic Programming
RBF	Radial Basis Function
SGD	Stochastic Gradient Descend
SLFNs	Single-hidden Layer Feed forward Networks
SMO	Sequential Minimal Optimization
SVNUB	Support Vector Not Upper Bound
SVDD	Support Vector Data Description
SVM	Support Vector Machine
SVUB	Support Vector Upper Bound
TLFN	Two-hidden Layer Feed-forward Network
UBSVs	Unbounded Support Vectors

فصل ۱) مقدمه

این فصل به معرفی مختصری از مسائل مطرح شده در پایان‌نامه می‌پردازد و شامل موارد زیر می‌باشد:

- ✓ دلایل انتخاب موضوع پایان‌نامه و بیان اهمیت آن
- ✓ تبیین موضوع پایان‌نامه
- ✓ مروری کلی بر مطالب پایان‌نامه

مطالعه این فصل در جهت دادن به سایر مطالب پایان‌نامه مهم بوده و می‌تواند مفید واقع شود.

افزایش روزافزون حجم اطلاعات ذخیره شده و عدم قابلیت استفاده از آرشیوهای عظیم اطلاعات خام در تصمیم‌گیری‌ها، نیاز به فرآیندی که بتواند دانش موجود در این اطلاعات را کشف نماید هر روز نمایان‌تر می‌سازد. شناسایی الگو زمینه تحقیقاتی است که به طراحی سیستم‌هایی برای تشخیص الگوها در داده‌ها و توصیف داده‌ها می‌پردازد و ما را در رسیدن به درکی از ساختار داده‌ها و به دست آوردن دانش و تصمیم‌گیری یاری می‌نماید. توسعه آن به ۱۹۶۰ برمی‌گردد و با ارتباط با رشته‌های دیگر مانند آمار، مهندسی، هوش مصنوعی، علوم کامپیوتر، روانشناسی و... به عنوان زمینه‌ای کاملاً بین رشته‌ای [Web '99] شناخته می‌شود. شناسایی الگو کاربردهای بسیار متنوعی در علوم مختلف مانند بینایی ماشین، سنجش از راه دور، تشخیص‌های پزشکی، داده‌کاوی (در کاربردهایی مانند تحلیل ریسک، ارزیابی اعتبار، تحلیل فروش و ...) و موارد بسیار متعدد دیگر دارد.

یکی از روش‌های شناسایی الگو که در حال حاضر به طور گسترده برای مسئله دسته‌بندی مورد استفاده قرار می‌گیرد، طبقه‌بندهای مبتنی بر بردار پشتیبان[1] است [LEE '07a]. در سال‌های اخیر طبقه‌بندهای مبتنی بر بردار پشتیبان برای شناسایی و دسته‌بندی الگو و در کاربردهایی مانند پردازش تصاویر و ویدئو [Kha '12]، داده‌کاوی، کاربردهای پزشکی [Ji '08] و سیستم‌های تشخیص نفوذ[2] به طور گسترده‌ای مورد استفاده قرار گرفته است [Gha '10]. نتایج حاصل از به‌کارگیری این روش‌ها در کاربردهای متفاوت، حاکی از کارایی بالای آن‌ها است.

این طبقه‌بندها دارای قابلیت‌های ارزشمندی هستند که آن‌ها را نسبت به دیگر روش‌های موجود برتر ساخته است [Tax '99]؛

۱) آموزش خود مشکل بیشینه و کمینه محلی را ندارند.

۲) دسته‌بندی کننده را با حداکثر تعمیم بنا می‌کنند.

۳) ساختار و توپولوژی خود را به صورت بهینه تعیین می‌نمایند

[1] Support Vector Machines
[2] Intrusion Detection Systems

۴) توابع جداساز غیرخطی را به راحتی و با محاسبات کم، با استفاده از توابع کرنل[۳] تشکیل می‌دهند .

۵) طبقه‌بندهای ماشین بردار پشتیبان در انواع دسته‌بندی‌هایی همچون تشخیص نفوذ به شبکه‌های کامپیوتری، تشخیص ارقام دست‌نویس، تشخیص شیء، شناسایی صورت، دسته‌بندی انواع صداها و مانند آن مورد استفاده قرار گرفته است که در مقایسه با روش‌های دیگر از کارایی قابل ملاحظه‌ای برخوردار است.

از جمله معایب این طبقه‌بند عبارتند از:

۱) با افزایش تعداد نمونه‌ها، کارایی این طبقه‌بندها کاهش می‌یابد.

۲) با افزایش تعداد بردارهای پشتیبان، نسبت به سایر طبقه‌بندها، در فاز آزمون[۴]، کندتر عمل می‌کنند؛ چرا که مرحله آموزش این طبقه‌بندها، معادل با حل یک معادله درجه دوم محدب است که این معادله شامل یک ماتریس نیمه‌معین مثبت با تعداد سطرهایی برابر با تعداد نمونه‌های آموزشی می‌باشد.

۳) حساسیت بالا به پارامترها.

۴) در صورتی که اندازه مجموعه داده n و تعداد ویژگی‌های آن d باشد ، در مرحله آموزش، حداقل پیچیدگی زمانی این گونه طبقه‌بندها $O\left(n^3d^3\right)$ و پیچیدگی حافظه‌ای نیز $O\left(n^2d^2\right)$ خواهد بود. در نتیجه با افزایش نمونه‌ها ماتریس نیز بزرگ‌تر می‌گردد و مسئله برای مجموعه داده‌های حجیم کارایی خود را از دست می‌دهد [Tax '99]. برای رفع این مشکل و تسریع این‌گونه طبقه‌بندها راهکارهای گوناگونی ارائه شده است.

[3] kernel
[4] Test phase