مجله کنترل
جلد ۶، شماره ۱، بهار ۱۳۹۱، صفحه ۱۹-۹

ارائهٔ روشی کارا برای دسته‌بندی مسائل چنددسته‌ای با رویکرد انتخاب دسته‌بند

محمّد علی باقری [۱]، غلامعلی منتظر [۲]

[۱] دانشجوی دکتری مهندسی فنّاوری اطلاعات، گروه مهندسی مهندسی فنّاوری اطلاعات، دانشگاه تربیت مدرس، a.bagheri@modares.ac.ir

[۲] دانشیار، دانشکدهٔ مهندسی فنی و مهندسی، گروه فنّاوری اطلاعات، دانشگاه تربیت مدرس، montazer@modares.ac.ir

(تاریخ دریافت مقاله ۱۳۹۰/۹/۱۲، تاریخ پذیرش مقاله ۱۳۹۱/۱/۱۸)

چکیده: سیستمهای دسته‌بندی شورایی، رویکردی مؤثر در یادگیری ماشین است که در آن با ترکیب نتایج چند دسته‌بند سعی می‌شود تقریب بهتری از یک دسته‌بند بهینه فراهم شود. در حوزهٔ ترکیب خروجی شورای دسته‌بندها، رویکرد «انتخاب دسته‌بند» توجه کمتری را در مقایسه با رویکرد «ادغام دسته‌بند» به خود جلب کرده است. همچنین، اغلب روشهای موجود در این حوزه، هزینهٔ محاسباتی بالایی دارند. در این مقاله، روشی مؤثر در دسته بندی مسائل چنددسته‌ای بر اساس رویکرد انتخاب ویژگی پیشنهاد شده است. روش پیشنهادی ابتدا با استفاده از ۱۴ مجموعه دادهٔ تراز از پایگاه UCI آزمون شده و پس از اثبات توانایی آن، برای شناسایی رایحه‌های سه نوع شیرین‌بیان به کار گرفته شده است. مقایسه نتایج روش پیشنهادی و روشهای دیگر سیستمهای شورایی بر اساس دو معیار صحّت شناسایی و زمان محاسباتی، کارایی بهتر روش را در دسته بندی مسائل چنددسته‌ای نشان می‌دهد.

کلمات کلیدی: دسته‌بندی، مسألهٔ چنددسته‌ای، سیستم شورایی، انتخاب دسته‌بند، شناسایی بو.

An Efficient Multiclass Classification Method Based on Classifier Selection Technique

Mohammad Ali Bagheri, Gholamali Montazer

Abstract: Individual classification models have recently been challenged by ensemble of classifiers, also known as multiple classifier system, which often shows better classification accuracy. In terms of merging the outputs of an ensemble of classifiers, classifier selection has not attracted as much attention as classifier fusion in the past, mainly because of its higher computational burden. In this paper, we propose a novel technique for improving classifier selection. In our method, the simple divide-and-conquer strategy is adapted in that a complex classification problem is divided into simpler binary sub-classification problems. We conduct extensive experiments on a series of multi-class datasets from the UCI (University of California, Irvine) repository and on odor database. The experimental results demonstrate the advanced performance of the proposed method.

Keywords: classification, multi-class, ensemble system, classifier selection, odor recognition.

۱- مقدمه

یکی از رویکردهای مؤثر برای حل مسائل دسته‌بندی پیچیده (از جمله مسائل چنددسته‌ای)، طراحی شورایی از دسته‌بندهای پایه و سپس ترکیب خروجی آنها است. . این رویکرد بیشتر با نامهای «سیستم دسته‌بندی شورایی» یا «سیستم دسته بند چندگانه» خوانده می‌شود. نتایج نظری [۱, ۲] و تجربی [۳, ۴] پژوهشها نشان می‌دهد سیستمهای شورایی اغلب نتایج دسته‌بندی بهتری را در مقایسه با دسته‌بندهای پایه در شورا به‌دست می‌دهند..

نویسنده عهده دار مکاتبات: غلامعلی منتظر

[DOR: 20.1001.1.20088345.1391.6.1.2.4]

فرایند طراحی سیستم دسته‌بند چندگانه به‌طور کلی شامل دو مرحله است [۵]: الف) ساخت شورایی از دسته‌بندها، ب) طراحی قاعدهٔ ترکیب (ادغام)[1]، که بهینه‌سازی تصمیم نیز نامیده می‌شود [۶, ۷]. دو رویکرد کلّی در بهینه‌سازی تصمیم وجود دارد: ادغام دسته‌بند[2] و انتخاب دسته‌بند[3] [۸]. در ادغام دسته‌بند، هر دسته‌بند از همهٔ داده‌های فضای ویژگی استفاده می‌کند. در این حالت، فرایند ادغام شامل ترکیب دسته‌بندهای پایه و ضعیف‌تر برای دست‌یابی به دسته‌بندی باکارایی بیشتر است. رأی‌گیری ساده یا وزن‌دارِ خروجی دسته‌بندها در اغلب روشهای مطرح سیستمهایی شورایی مانند *Bagging* ، *AdaBoost* و روش زیرفضای تصادفی[4] استفاده می‌شود.

رویکرد «ادغام دسته‌بند» براساس این فرضیه است که هر دسته‌بند خطای مستقلی دارد و باعث می‌شود که دسته‌بندها خطاهای یکدیگر را بپوشانند. با این حال، از آنجا که تمامی دسته‌بندها در شورا به مسألهٔ مشخصی پاسخ می‌دهند، ایجاد استقلال در خطای دسته‌بندها مشکل خواهد بود [۹]. در نتیجه، ترکیب خروجی دسته‌بندها نتایج دسته‌بندی بهتری را تضمین نمی‌کند. علاوه براین، استفاده همزمان از چندین دسته‌بند برای دسته‌بندی یک الگوی آزمون لازم است.

در انتخاب دسته‌بند، هر دسته‌بند در شورا بخشی از فضای ویژگی را به‌خوبی یاد می‌گیرد و لذا دسته‌بندی داده‌های آن بخش از فضا را انجام می‌دهد. از این‌رو، در رویکرد انتخاب، معمولاً یک دسته‌بند برای تعیین دستهٔ الگوی آزمون استفاده می‌شود [۸]. روشهای مبتنی بر رویکرد انتخاب، به عنوان مثال «انتخاب پویایی دسته‌بند براساس صحّت همسایگی»[5]، به علّت تخمین کارایی هر دسته‌بند در ناحیهٔ همسایگی الگوی آزمون، اغلب هزینهٔ محاسباتی بالایی دارند. همچنین، این روشها به پارامترهای روش مانند تعداد نزدیک‌ترین k همسایه و تابع فاصله وابستگی زیادی دارند [۱۰].

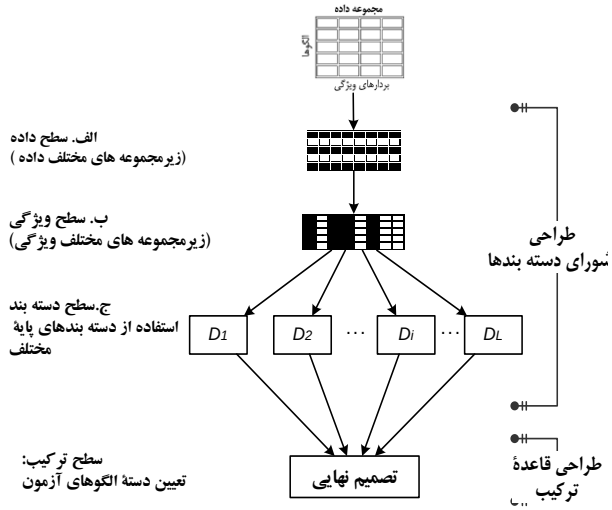
در این مقاله، روش انتخاب دسته‌بند جدیدی برای مسائل دسته‌بندی چنددسته‌ای ارائه شده‌c است: بدین منظور برای هر زوج دسته، یک دسته‌بند دودویی[6] با استفاده از داده‌های آموزش آن دو دسته ایجاد می‌شود. سپس، برای دسته‌بندی یک الگوی آزمون، ابتدا دو دسته‌ای که بیشترین احتمال را دارند تعیین می‌شود. براساس این دو دسته، یکی از دسته‌بندهای دودویی انتخاب می‌شود که دستهٔ نهایی الگوی آزمون را تعیین کند.

ادامهٔ این مقاله به شرح ذیل تنظیم شده است: ساختار سیستمهای شورایی در بخش دوّم به اختصار معرفی شده است. درادامه، روش پیشنهادی در انتخاب دسته‌بند در بخش سوّم شرح داده شده است. در بخش چهارم، نتایج تجربی استفاده از روش پیشنهادی با استفاده از ۱۴ مجموعه دادهٔ تراز آورده شده، سپس حل مسئله تشخیص الگوهای بویایی در بخش پنجم آمده و سرانجام بخش ششم و هفتم به بحث و نتیجه‌گیری مقاله اختصاص یافته است.

۲- سیستمهای دسته‌بند چندگانه برای ادغام دسته‌بندها

یکی از رویکردهای مؤثر برای حل مسائل دسته‌بندی پیچیده (از جمله مسائل چنددسته‌ای)، طراحی شورایی از دسته‌بندهای پایه و سپس ترکیب خروجی آنها است. این رویکرد، در منابع یادگیری ماشینی با نامهای مختلفی مانند سیستم دسته بند چندگانه، کمیتهٔ یادگیرنده‌ها[7]، نظریهٔ اجماع[8] و سیستم دسته‌بندی شورایی نیز خوانده می‌شود و یکی از چهار مسیر پیشگام در حوزهٔ یادگیری ماشینی را شکل داده است [۱۱]. با این حال، هنگامی یادگیری شورایی از یادگیری بهترین دسته‌بند پایه بهتر است که دسته‌بندهای پایه دارای کارایی قابل قبول بوده و «گوناگون در خطا»[9] باشند [۱۲]. دو دسته‌بند، زمانی گوناگونی در خطا دارند که نمونه‌هایی که به صورت نادرست دسته‌بندی می‌کنند متفاوت باشند. تفاوت در موارد خطای دسته‌بندهای پایه باعث می‌شود که دسته‌بندها خطاهای یکدیگر را بپوشانند و به همین علت گوناگونی در خطا، از نکات اساسی در موفقیت سیستم دسته‌بندی چندگانه است.

فرایند طراحی سیستم شورایی به‌طور کلی شامل دو مرحله است [۵]: الف) ساخت شورای دسته‌بندها، ب) طراحی قاعدهٔ ترکیب (شکل ۱)؛ که در بخشهای زیر شرح داده شده است.



شکل ۱. ساختار سیستمهای دسته‌بند چندگانه (شورایی)

۲-۱. طراحی شورای دسته‌بندها

به طور کلّی سه رویکرد مختلف برای ایجاد شورای دسته‌بندها وجود دارد: رویکرد اوّل استفاده از الگوریتمهای یادگیری مختلف مانند شبکهٔ

[1] combination rule
[2] classifier fusion
[3] classifier selection
[4] Random Subspace Method (RSM)
[5] dynamic classifier selection with local accuracy (DCS-LA)
[6] binary classifier
[7] committees of learners
[8] consensus theory
[9] error diverse

عصبی، درخت تصمیم و غیره و یا تغییر پارامترهای یک نوع دسته بند (مانند وزنهای اولیه مختلف در شبکهٔ عصبی) است. رویکرد دیگر که توجه بیشتری را به خود جلب کرده است، استفاده از مجموعه‌های دادهٔ مختلف برای آموزش دسته‌بندهای پایه است. این مجموعه‌ها، اغلب توسط روشهای «نمونه‌برداری مجدد»[1] از مجموعهٔ آموزش حاصل می‌شود؛ به‌طوری‌که مجموعه‌های آموزش مختلف با انتخاب تصادفی از نمونه‌های آموزش تولید می‌شوند و هر مجموعه به عنوان ورودی یک دسته‌بند استفاده می‌شود. روشهای bagging [۱۳] و boosting [۱۴] محبوب‌ترین روشها با این رویکرد هستند. رویکرد سوم که از رویکردهای نوین طراحی شورا محسوب می‌شود، استفاده از زیرمجموعه‌های ویژگی مختلف و آموزش دسته‌بندهای پایه با این زیرمجموعه‌ها است که «شورای انتخاب ویژگی»[2] گفته می‌شود. به بیان دیگر، در رویکرد دوّم هر مجموعه داده شامل برخی نمونه‌های آموزش است که تمام ویژگی‌های مسأله را دارا است؛ درحالی‌که در رویکرد شورای انتخاب ویژگی، هر مجموعه داده شامل تمام نمونه‌های آموزش است که از برخی ویژگی‌های مسأله برخوردار است.

هدف شورای انتخاب ویژگی، یافتن زیرمجموعه‌های ویژگی است که صحّت دسته‌بندی خوبی داشته و تا حدممکن گوناگون باشند. «هُو»[3] نشان داد درحالی‌که اغلب روشهای دسته‌بندی با مشکل ابعاد زیاد داده مواجهند، روش زیرفضای تصادفی می‌تواند از این موضوع بهره‌مند شود [۱۵].

۲-۲. طراحی قاعدهٔ ترکیب

پس از ایجاد شورایی از دسته‌بندهای پایه، گام بعدی، طراحی قاعدهٔ ترکیب خروجی آنها است؛ که «بهینه‌سازی تصمیم» نیز خوانده می‌شود. دو رویکرد اصلی برای بهینه‌سازی تصمیم وجود دارد: ادغام دسته‌بند و انتخاب دسته‌بند [۵-۶]. در ادغام دسته‌بند، هر دسته‌بند از همهٔ داده‌های فضای ویژگی استفاده می‌کند. در این حالت، فرایندِ ادغام شامل ترکیب دسته‌بندهای پایه و ضعیف‌تر برای دست‌یابی به دسته‌بند باکارایی بیشتر است. رأی‌گیری ساده یا وزن‌دار، قواعد جبری (مانند ضرب، جمع، بیشینه و کمینه)، انتگرال فازی، و قاعدهٔ ترکیب دمپستر-شفر مرسوم‌ترین توابع ترکیب ادغام دسته‌بندها هستند. اغلب روشهای مطرح سیستمهایی شورایی مانند Bagging ، AdaBoost و روش زیرفضای تصادفی بر اساس این رویکرد طراحی شده‌اند.

ایدهٔ استفاده از دسته‌بندهای مختلف برای تعیین دستهٔ الگوهای آزمون به سال ۱۹۷۸ میلادی بازمی‌گردد [۱۶]. بااین حال تمایل در بکارگیری رویکرد انتخاب دسته‌بند در سیستمهای شورایی توّسط «وودز»[4] و همکارانش در سال ۱۹۹۷ میلادی شکل گرفت [۸]. رویکرد انتخاب دسته‌بند فرض می‌کند که هر دسته‌بند یک ناحیهٔ انتخاب (ناحیهٔ رقابت[5]) در فضای ویژگی مسأله دارد. ازاین‌رو، هدف انتخاب دسته‌بند تخصیص هر دسته‌بند به یک ناحیهٔ رقابت در فضای مسأله است [۶, ۸, ۱۷]. در این رویکرد، دو شیوهٔ انتخاب دسته‌بند ارائه شده است: انتخاب ایستای دسته‌بند[6] و انتخاب پویای دسته‌بند[7]. در شیوهٔ ایستا، یک دسته‌بند دسته‌بند برای تعیین دستهٔ تمام الگوهای آزمون استفاده می‌شود. در روشهای مبتنی بر شیوهٔ پویا، ناحیهٔ رقابت در هنگام فرایند دسته‌بندی الگوی آزمون و با در نظر گرفتن مشخصات الگوی آزمون تعیین می‌شود. از آنجا که الگوهای مختلف ویژگی‌های دسته‌بندی متفاوتی دارند، استفاده از دسته‌بندهای مختلف نتایج بهتری را در مقایسه با شیوهٔ ایستا می‌دهند.

بر اساس صحّت محلّی الگوی آزمون، «وودز» و همکارانش یکی از معروف‌ترین روشهای انتخاب دسته‌بند را ارائه کردند که DCS-LA [8] خوانده می‌شود [۸]. در این روش، ابتدا مجموعه‌ای از دسته‌بندهای پایه ایجاد می‌شود. این الگوریتم، برای هر الگوی آزمون، ناحیهٔ محلّی را به‌صورت k نزدیک‌ترین همسایه در میان الگوهای آموزش، تعریف می‌کند. سپس، صحّت محلّی هر دسته‌بند (نسبت الگوهایی که در ناحیهٔ محلّی الگوی آزمون درست دسته‌بندی شده‌اند) تخمین زده می‌شود و دسته‌بندی که بیشترین صحّت دسته‌بندی را در آن ناحیه داشته انتخاب می‌شود.

با این حال، هر دو رویکرد سیستمهای شورایی برای ترکیب خروجی دسته‌بندها (ادغام یا انتخاب دسته‌بند)، هزینهٔ محاسباتی بالایی دارند. در رویکرد ادغام دسته‌بند، فرایند دسته‌بندی زمان‌بر خواهد بود؛ چرا که سیستم شورایی شامل تعداد زیادی دسته‌بند است و برای تعیین دستهٔ هر الگوی آزمون، خروجی تمام دسته‌بندهای پایه در شورا باید ترکیب شوند. همچنین در رویکرد انتخاب دسته‌بند، تخمین فضای انتخاب و صحّت محلّی هر دسته‌بند، موجب طولانی شدن زمان دسته‌بندی خواهد شد.

در این مقاله، روش جدیدی در انتخاب دسته‌بند برای مسائل دسته‌بندی چنددسته‌ای پیشنهاد شده است. ایدهٔ اصلی این روش، انتخاب دسته‌بند بر اساس فراوانی الگوهای دسته‌های مختلف در همسایگی الگوی آزمون است. براین اساس، گزینه‌های ممکن در تعیین برچسب الگوی آزمون از چند دسته به دو دسته‌ای که بیشترین احتمال را دارند محدود می‌شود. سپس، مسألهٔ دودسته‌ای با استفاده از یک دسته‌بند دودویی که مرزهای تصمیم دقیق‌تری دارد حل می‌شود. در این روش، نیازی به فرایند زمان‌بر تخمین کارایی هر دسته‌بند در همسایگی هر الگوی آزمون نیست. همچنین تعداد و پیچیدگی دسته‌بندهای مسأله به مراتب کاهش می‌یابد. روش دسته‌بندی پیشنهادی، در بخش بعد به تفصیل معرفی شده است.

[1] resampling
[2] ensemble feature selection
[3] Ho
[4] Woods

[5] region of competence
[6] static classifier selection (SCS)
[7] dynamic classifier selection (DCS)
[8] dynamic classifier selection by local accuracy (DCS-LA)

۳. مراحل روش پیشنهادی انتخاب دسته‌بند

۳-۱ . مرحلهٔ آموزش

در مرحلهٔ آموزش، برای هر دو دسته در مسأله چنددسته‌ای، یک دسته‌بند دودویی با استفاده از داده‌های دو دستهٔ متناظر ایجاد می‌شود. در این گام، دو نسخه از روش پیشنهادی ارائه شده است. در نسخهٔ اوّل، هر دسته‌بندِ دودویی با استفاده از همهٔ ویژگی‌های موجود، آموزش می‌بینند. در نسخهٔ دوّم، فرض می‌شود برای تمایز هر دو دستهٔ خاص، یک مجموعه ویژگی قابلیت تمایز بیشتری می‌تواند ایجاد کند. لذا، سعی می‌شود بهترین زیرمجموعهٔ ویژگی برای تمایز میان هر دو دستهٔ مسأله انتخاب شود. در این پژوهش، روش «انتخاب ویژگی با الگوریتم ژنتیک»[1] به عنوان روش انتخاب ویژگی به‌کار گرفته شده است؛ و سپس آموزش هر دسته‌بند دودویی با زیرمجموعهٔ ویژگی متفاوت انجام شده است. این فرایند برای همهٔ $c(c-1)/2$ دستهٔ مسأله انجام شده است که c تعداد دسته‌های مسأله را نشان می‌دهد. از این رو، در پایان مرحلهٔ آموزش، دسته‌بندهایی ایجاد شده‌اند که بین هر دو دسته، قابلیت تمایز بیشتری را ایجاد می‌کنند.

انتظار می‌رود دسته‌بندهایی که با این روش ایجاد شده‌اند، مرزهای تصمیمی با پیچیدگی خطی کمتر (مرزهای تصمیم ساده‌تری) در مقایسه با مرز تصمیم مسألهٔ چنددسته‌ای داشته باشند. این موضوع به این علت است که برای یک مسأله دو دسته‌ای، یک دسته باید تنها از دستهٔ دیگر جدا شود؛ درحالی‌که در مسأله چنددسته‌ای، هر دسته باید به‌طور همزمان از تمام دسته‌های دیگر جدا شود.

۳-۲. مرحلهٔ آزمون

در این مرحله، برای هر الگوی ناشناختهٔ آزمون، $\boldsymbol{x}^*$، ابتدا یک ناحیهٔ محلّی تعیین می‌شود. این ناحیه، مشابه الگوریتم *DCS-LA* براساس نزدیک‌ترین k همسایه از میان الگوهای آموزش تعریف می‌شود. سپس، تعداد الگوهای هر دسته در این ناحیه شمارش شده و دو دسته‌ای که در این ناحیه بیشترین تعداد الگوها را داشتند به‌عنوان محتمل‌ترین دسته‌ها تعیین می‌شوند. در این روش، عمل پیش‌بینی دو دستهٔ محتمل‌تر، به‌کمک تابعی به‌نام «حدس اولیه» انجام می‌شود. برای هر نمونهٔ آزمون، اگر یکی از دو دستهٔ پیش‌بینی شده، دستهٔ واقعی نمونهٔ آزمون باشد حدس اولیه صحیح بوده است. براساس این دو دسته، یکی از دسته‌بندهای دودویی که در مرحلهٔ آموزش ایجاد شده بود، انتخاب می‌شود که دستهٔ نهایی الگوی $\boldsymbol{x}^*$ را تعیین کند. شکل ۲، شبه‌کد روش پیشنهادی را نشان می‌دهد.

ورودی

■ الگوهای آموزش، $S = \{\{\boldsymbol{x}_n, \omega_n\} \mid \omega_n \in \{\omega_1, \omega_1, .., \omega_c\}\}$، که ω_n نشان‌دهندهٔ دستهٔ صحیح هر الگو است.

■ الگوریتم دسته‌بندی دودویی

مرحلهٔ آموزش

به ازای $i = 1, \dots, c$ انجام بده

به‌ازای $j = i+1, \dots, i$ انجام بده

۱. دسته‌بندهای دودویی $D_{i,j}$ را با استفاده از داده‌های آموزش متناظر، $S_{i,j} = \{\{x_t, \omega_t\} \mid \omega_t \in \{\omega_i, \omega_j\}\}$، ایجاد کن:

۱. الف. آموزش هر دسته‌بند با تمام ویژگی‌های موجود

۱.ب. با استفاده از GAFS بهترین زیرمجموعهٔ ویژگی برای تمایز دو دسته را بیاب و آموزش دسته‌بند $D_{i,j}$ با آن زیرمجموعه انجام بده

۲. دسته‌بند $D_{i,j}$ را به شورا اضافه کن

مرحلهٔ آزمون: به ازای هر الگوی آزمون $\boldsymbol{x}^*$

۱. ناحیهٔ محلّی الگوی $\boldsymbol{x}^*$ را به‌صورت k نزدیک‌ترین همسایه از داده‌های آموزش تعریف کن.

۲. دو دسته‌ای که در این ناحیه بیشترین تعداد الگوها را دارند، به‌عنوان محتمل‌ترین دسته‌های ممکن تعیین کن: $\omega^* = \{\omega_m, \omega_n\};\ m, n: 1, \dots, c;\ m \neq n$

۳. دسته‌بند $D_{m,n}$ را فرابخوان و دستهٔ الگوی $\boldsymbol{x}^*$ را تعیین کن.

شکل ۲. شبه کد الگوریتم پیشنهادی

۴- آزمایش روش با مجموعه داده‌های تراز

در این بخش، برای آزمون کارایی روش پیشنهادی، آزمایش تجربی روش و مقایسهٔ آن با روشهای دیگر آورده شده است. روش پیشنهادی مقاله (نسخهٔ ۱ و نسخهٔ ۲) با مدل دسته‌بند منفرد[2]، روش *DCS-LA*، و دو روش مطرح سیستمهای شورایی (*Bagging* و *RSM*) مقایسه شده‌اند. در ادامهٔ این بخش، ابتدا تنظیمات آزمایشها بیان شده و سپس نتایج تجربی آزمایشها ارائه شده است. روشهای ذکر شده، با استفاده از ۱۴ مجموعه دادهٔ مخزن دادهٔ *UCI*[3] [۱۸] آزمون شده است. این مجموعه‌های داده به وفور در پژوهشهای یادگیری ماشین برای ارزیابی الگوریتمهای مختلف به‌کارگرفته شده است. جدول ۱ مشخصات هر مجموعه داده شامل تعداد دسته‌ها، تعداد الگوها و تعداد ویژگی‌های هر مجموعه را نشان می‌دهد.

۴-۱. تنظیمات آزمایش

در این مقاله، از ماشین بردار پشتیبان (*SVM*) به‌عنوان دسته‌بند پایه در سیستم شورایی استفاده شده است. براساس نتایج آزمایشهای اولیه، اندازهٔ شورا (تعداد دسته‌بندهای پایه در شورا) برای سیستمهای شورایی (*DCS-LA*، *Bagging* و *RSM*) ۲۵ تعیین شده است؛ همچنین نتایج پژوهشهای مختلف نشان می‌دهد این مقدار بهترین نتیجهٔ دسته‌بندی را موجب می‌شود [۱۹]. در روش *DCS-LA* و روش پیشنهادی مقاله، مقدار k برای تعیین ناحیهٔ محلّی (ناحیهٔ همسایگی) از میان چهار مقدار انتخاب

[1] Genetic Algorithm Feature Selection (GAFS)

[2] single classifier

[3] UCI machine learning repository

شده‌اند: ۴، ۷، ۱۱ و ۱۵. برای هر مجموعه داده، نتایج دسته‌بندی حاصل از بهترین مقدار k آورده شده است.

جدول ۱. مشخصات مجموعه داده‌های مورد استفاده

	مجموعه داده	تعداد دسته‌ها	تعداد الگوها	تعداد ویژگی‌ها
۱	Abalone	۳	۴۱۷۷	۸
۲	Balance	۳	۶۲۵	۴
۳	Car	۴	۱۷۲۸	۶
۴	Cmc	۳	۱۴۷۳	۹
۵	Derm	۶	۳۶۶	۳۴
۶	Ecoli	۸	۳۳۶	۷
۷	Glass	۷	۲۱۴	۱۰
۸	Iris	۳	۱۵۰	۴
۹	Lymph	۴	۱۴۸	۱۸
۱۰	Pendigits	۱۰	۱۰۹۹۲	۱۶
۱۱	Sat	۶	۶۴۳۵	۳۶
۱۲	Wine	۳	۱۷۸	۱۳
۱۳	Yeast	۱۰	۱۴۸۴	۸
۱۴	Zoo	۷	۱۰۱	۱۶

همهٔ آزمایشها با استفاده از نرم‌افزار *Matlab* نسخهٔ ۷/۹ انجام شده است. برای پیاده‌سازی *SVM*، از بستهٔ نرم‌افزاری *LibSVM* (نسخهٔ ۳/۱) [۲۰] با هستهٔ خطی[1] استفاده شده است. پیاده‌سازی روشهای دیگر نیز براساس شبه‌کد آنها انجام شده است.

به‌عنوان یک قاعدهٔ کلّی، روش مرسوم *Holdout* (تقسیم الگوها به دو مجموعهٔ آموزش و آزمون) برای مجموعه داده‌های بزرگ (مجموعه‌های با بیش از ۴۰۰۰ تعداد الگو) استفاده شده است. در این موارد، نمونه‌های مسأله به دو مجموعهٔ آموزش (۷۰٪) و آزمون (۳۰٪) تقسیم شده‌اند. برای سایر مجموعه داده‌ها، روش تقسیم ۵-بخشی داده[2] به‌کارگرفته شده است.

در نسخهٔ دوّم روش پیشنهادی، ارزیابی مجموعه ویژگی براساس تقسیم دادهٔ آموزش به دو مجموعهٔ آموزش (۶۰٪) و اعتباریابی (۴۰٪) انجام شده است. مجموعهٔ اعتباریابی، برای ارزیابی اولیهٔ زیرمجموعهٔ ویژگی به‌کار گرفته می‌شود. در روش انتخاب ویژگی با الگوریتم ژنتیک، اندازهٔ جمعیت اولیه برابر ۱۵، ضریب تقاطع برابر با ۵۰٪ و ضریب جهش برابر با ۱۰٪ تعیین شده اند. فرایند تکراری روش، شامل حداکثر ۱۰ تکرار (نسل) است. بررسی‌های اولیه نشان داده است که صحت دسته‌بندی پس از ۱۰ تکرار بهبود نمی‌یابد. علاوه‌براین، اگر صحّت دسته‌بندی در ۳ تکرار متوالی بهبود نیابد، فرایند متوقف می‌شود. در ابتدا، مقادیر ژنهای هر کروموزوم به طور تصادفی برابر ۰ یا ۱ انتخاب شده است. میزان موفقیت (مقدار برازش) هر کروموزوم (مجموعهٔ ویژگی) برابر صحّت دسته‌بندی آن بر روی مجموعهٔ اعتباریابی است. در انتهای فرایند، کروموزوم با بیشترین مقدار برازش به‌عنوان زیرمجموعهٔ ویژگی بهینه انتخاب شده است.

۴-۲. نتایج آزمایشها

جدول ۲، صحّت دسته‌بندِ منفرد، روشهای *DCS-LA* ، *Bagging*، و *RSM* و نیز نتایج روش پیشنهادی (نسخهٔ ۱ و ۲) را نشان می‌دهد. در این جدول، میانگین و انحراف استاندارد صحّت دسته‌بندی حاصل از ۲۰ بار تکرار هر روش برای هر مجموعه داده آورده شده است. برای هر مجموعه داده، بهترین نتیجه به‌دست آمده پررنگ شده است. ستون آخر جدول ۲، میانگین صحّت «حدس اوّلیه» را نشان می‌دهد. صحّت دسته‌بندی این تابع برابر با نسبت نمونه‌های آزمون است که یکی از دو دستهٔ پیش‌بینی‌شده، دستهٔ واقعی نمونهٔ آزمون باشد.

برای مقایسهٔ آماری روش پیشنهادی با سایر روشها، از آزمون t زوجی یک‌طرفه[3] در سطح اطمینان ۹۵٪ استفاده شده است. جدول ۳ نتیجهٔ مقایسهٔ آماری روش پیشنهادی با روشهای دیگر را نشان می‌دهد. در این جدول، برتری روش در یک مجموعه داده در مقایسه با روش دیگر با علامت (+) و ضعف آن روش با علامت (-) نشان داده شده است؛ همچنین اگر نتیجهٔ روش پیشنهادی تفاوت معناداری با روش دیگر نداشته باشد، هیچ علامتی درج نشده است. مقایسه نتایج هر روش با نسخهٔ ۱ و نسخهٔ ۲ روش پیشنهادی با علامت ممیز (/) جدا شده است. دو ردیف آخر جدول ۳، مقایسهٔ «برد- باخت-تساوی»[4] روش پیشنهادی (نسخهٔ ۱ و ۲) را با دیگر روشها براساس آزمون آماری نشان می‌دهد. اولین مقدار این شاخص، تعداد مجموعه داده‌هایی را نشان می‌دهد که نتایج دسته‌بندی روش پیشنهادی برتری معناداری با روش مقابل دارد؛ به طور مشابه، دومین مقدار این شاخص، تعداد مجموعه داده‌هایی است که نتایج دسته‌بندی روش مقابل برتری معناداری با روش پیشنهادی دارد و سومین مقدار این شاخص، تعداد مجموعه داده‌هایی است که تفاوت معناداری بین روش پیشنهادی و روش مقابل وجود ندارد.

۴-۲-۱. مقایسهٔ دو نسخهٔ روش پیشنهادی

بررسی نتایج حاصل از پیاده‌سازی دو نسخهٔ روش پیشنهادی در جدول ۲ نشان می‌دهد که هر دو نسخه، نتایج مشابهی را می‌دهند؛ با این حال، نسخهٔ اول نتایج کلّی بهتری را به‌دست می‌دهد. نتیجهٔ مهم حاصل از این مقایسه این است که برای مجموعه داده‌های نسبتاً کوچکتر، نسخهٔ اوّل نتایج بهتری را ایجاد می‌کند. نتایج ضعیف نسخهٔ دوّم، می تواند ناشی از ورآموزی[5] الگوریتم ژنتیک در انتخاب زیرمجموعهٔ ویژگی بر روی الگوهای آموزش باشد؛ چرا که تعداد الگوها برای ایجاد دسته‌بندهای دودویی که خوب آموزش دیده باشند، کم است.

[1] linear kernel
[2] 5-fold cross validation
[3] one-tailed paired *t-test*
[4] win–loss–tie
[5] overfitting

جدول ۲. میانگین و انحراف معیار صحّت دسته‌بندی روش پیشنهادی (نسخهٔ ۱ و۲) و سایر روشها با استفاده از مجموعه‌داده‌های تراز

	مجموعه داده	single SVM	DCS-LA	Bagging	RSM	DCS-DQ.v1	DCS-DQ .v2	حدس اولیه
۱	Abalone	63.97±0.21	57.71±1.35	62.40±1.46	62.58±1.55	**64.24**±0.99	63.86±0.72	91.08
۲	Balance	91.45±0.72	78.91±0.53	90.84±0.81	78.51±2.03	**93.46**±0.67	92.64±1.06	98.72
۳	Car	84.90±0.37	70.02±0.00	84.91±0.34	70.02±0.00	**87.00**±0.30	86.31±0.45	99.18
۴	Cmc	51.03±0.63	51.62±0.35	51.44±0.45	47.43±1.28	51.67±0.87	**52.24**±0.55	82.13
۵	Derm	95.97±0.69	93.69±0.60	96.63±0.48	**97.94**±0.44	96.39±0.42	97.66±0.90	99.78
۶	Ecoli	**80.16**±0.75	67.63±1.15	76.43±0.32	76.64±1.23	80.02±1.51	78.07±2.30	94.58
۷	Glass	64.17±2.26	60.56±0.94	59.71±2.54	56.87±1.80	**64.37**±1.34	63.93±1.89	89.43
۸	Iris	97.47±1.01	95.27±1.05	97.50±0.99	95.40±0.91	**98.13**±0.70	96.60±1.24	99.96
۹	Lymph	82.96±1.79	78.48±0.74	82.53±1.43	**82.97**±1.08	80.09±4.40	80.62±5.11	96.21
۱۰	Pendigits	97.66±0.30	97.47±0.28	98.09±0.25	96.47±0.36	**98.95**±0.13	98.76±0.12	99.84
۱۱	Sat	86.24±0.72	87.28±0.64	87.00±0.64	87.08±0.64	88.97±0.47	**89.67**±0.54	98.36
۱۲	Wine	95.00±0.90	93.21±0.81	94.80±1.28	**96.41**±0.75	93.22±1.22	92.17±1.09	95.95
۱۳	Yeast	54.41±0.46	50.73±0.98	50.94±0.49	43.30±1.73	**56.31**±0.67	55.70±0.42	83.21
۱۴	Zoo	**95.07**±1.41	89.24±1.00	94.70±1.41	93.54±2.22	94.05±1.40	83.22±2.51	98.3

جدول ۳. نتایج مقایسهٔ روش پیشنهادی (نسخهٔ ۱ و۲) با سایر روشها براساس آزمون آماری

	مجموعه داده	single SVM	DCS-LA	Bagging	RSM
۱	Abalone	/	+/+	+/+	+/+
۲	Balance	+/+	+/+	+/+	+/+
۳	Car	+/+	+/+	+/+	+/+
۴	Cmc	+/+	/	/+	+/+
۵	Derm	+/+	+/+	-/+	-/
۶	Ecoli	/-	+/+	+/+	+/+
۷	Glass	/	+/+	+/+	+/+
۸	Iris	+/-	+/+	+/	+/+
۹	Lymph	/	+/	/	/
۱۰	Pendigits	+/+	+/+	+/+	+/+
۱۱	Sat	+/+	+/+	+/+	+/+
۱۲	Wine	-/-	/-	-/-	-/-
۱۳	Yeast	+/+	+/+	+/+	+/+
۱۴	Zoo	/-	+/-	+/-	+/-
	تساوی / باخت / برد (نسخهٔ ۱)	**۸ / ۱ / ۴**	**۱۲ / ۰ / ۲**	**۱۰ /۲/ ۲**	**۱۱ /۲/ ۱**
	تساوی / باخت / برد (نسخهٔ ۲)	**۷ /۳/ ۴**	**۱۰ /۲/ ۲**	**۱۰ / ۲ / ۲**	**۱۰ / ۲ / ۲**

برتری روش پیشنهادی در یک مجموعه داده در مقایسه با روش دیگر با علامت (+) و باخت روش با علامت (–) نشان داده شده است. اگر نتیجهٔ روش پیشنهادی تفاوت معناداری با روش دیگر نداشته باشد، هیچ علامتی درج نشده است. مقایسه نتایج هر روش با نسخهٔ ۱ و ۲ روش پیشنهادی با علامت ممیز (/) جدا شده است.

این نتیجه با نتایج پژوهشهای [۲۱, ۲۲, ۲۳] همخوان است که نشان می‌دهد الگوریتم ژنتیک در انتخاب زیرمجموعه ویژگی تمایل به ورآموزی دارد. برای نمونه برای مجموعه دادهٔ *Zoo*، نسخهٔ دوم روش پیشنهادی نتایج بدتری را در مقایسه با نسخهٔ اوّل ارائه می‌کند که این موضوع ناشی از تعداد کم نمونه‌های هر دسته است. به علت برتری نسخهٔ اوّل الگوریتم، در ادامه، سایر روشها با نسخه اوّل روش پیشنهادی مقایسه شده است.

۴-۲-۲. مقایسه با دسته‌بند منفرد

همانگونه که در جدول ۳ نشان داده شده است، برای ۸ مجموعه دادهٔ مورد استفاده، نتایج نسخهٔ اول روش پیشنهادی برتری معناداری را نسبت به دسته‌بند منفرد نشان می‌دهد. از طرفی، دسته‌بند منفرد *SVM* تنها در یک مجموعه داده صحّت بیشتری را نسبت به روش پیشنهادی نشان می‌دهد. نتیجهٔ قابل توجه دیگر این که بهبود صحّت دسته‌بندی برای مجموعه داده‌های بزرگ‌تر بیشتر است. برای نمونه، برای مجموعه داده‌های *Car* ، *Balance*، و *Sat* بهبود صحّت دسته‌بندی بیش از ۲٪ است.

۴-۲-۳. مقایسه با رویکرد انتخاب دسته‌بند بر اساس صحّت محلی

مقایسهٔ صحّت دسته‌بندی نسخهٔ اول روش پیشنهادی و روش DCS-LA در جدول ۲ نشان می‌دهد که صحّت دسته‌بندی در ۹ مجموعه دادهٔ آزمایش بیش از ۲٪ بهبود یافته است. بیشترین میزان بهبود صحّت دسته‌بندی ۱۶/۹۸٪ برای مجموعه دادهٔ Car است. به‌طور کلی، نسخهٔ اول روش پیشنهادی برای تمام ۱۴ مجموعه دادهٔ موجود نتایج بهتری را نشان می‌دهد که برای ۱۲ مجموعه داده تفاوت معنادار آماری نیز وجود دارد.

۴-۲-۴. مقایسه با روشهای شورایی ادغام دسته‌بندها

نتایج جدول ۳ نشان می‌دهد برای ۱۲ مجموعه دادهٔ آزمایش، هیچکدام از روشهای ادغام دسته‌بندها (Bagging و RSM) از نظر آماری بهبود معناداری را درمقایسه با نسخهٔ اول روش پیشهادی نشان نمی‌دهند. از طرفی، در حداقل ۱۰ مجموعه داده، نسخهٔ اول روش پیشنهادی برتری معناداری را نسبت به روشهای ادغام دسته‌بندها به‌دست می‌دهد.

دربرخی مجموعه داده‌ها، RSM بیش از ۱۶٪ صحّت دسته‌بندی کمتری را نشان می‌دهد. مجموعه داده‌هایی که با روش RSM صحّت دسته‌بندی خوبی را نشان نمی‌دهد احتمالا دارای ویژگی‌هایی هستند که وابستگی کمی با هم دارند و وجود آنها در هر مجموعه ویژگی الزامی است. نتایج پژوهش [۲۴] نیز مؤید این مطلب است.

۴-۳. مقایسهٔ هزینهٔ محاسباتی

بهترین راه مقایسه برای مقایسهٔ هزینهٔ محاسباتی روش‌های مختلف، زمان اجرای هر روش در شرایط مشابه است. در این پژوهش، تمامی آزمایشها با مشخصات سخت افزاری ذیل انجام شده است: رایانه‌ای با پردازندهٔ *Intel Core 2 Duo 2.26GHz* و *2GB* حافظهٔ *RAM*. جدول ۴ زمان متوسط اجرای هر روش برای هر مجموعه داده را نشان می‌دهد.

جدول ۴. زمان متوسط اجرای هر روش (به ثانیه) برای هر مجموعه داده

مجموعه داده	single SVM	DCS-LA	Bagging	RSM	DCS-DQ.v1	DCS-DQ .v2
Abalone	1.86	12.70	3.29	8.12	1.02	221.24
Balance	0.05	1.34	0.26	0.92	0.06	35.59
Car	0.54	7.09	2.34	4.89	0.45	231.25
Ecoli	0.05	0.96	0.26	0.47	0.14	100.15
Glass	0.05	0.99	0.34	0.69	0.07	8.87
Iris	0.01	0.17	0.04	0.07	0.01	0.56
Wine	1.24	1.92	12.11	36.59	1.46	215.88
Yeast	0.63	13.23	3.22	7.76	0.50	40.05
Zoo	0.03	0.47	0.14	0.25	0.05	6.77
Sat	70.03	488.82	402.05	499.97	60.82	2911.30
Pendigits	44.23	894.23	124.96	147.51	34.99	3432.66
Lymph	0.03	0.29	0.11	0.17	0.03	1.87
Cmc	3.10	50.12	18.10	26.92	3.42	157.15
Derm	0.11	1.17	1.28	1.91	0.17	12.84
مجموع	**121.96**	**1473.50**	**568.50**	**736.24**	**103.19**	**7376.18**

همانگونه که در جدول ۴ نشان داده شده، دسته‌بند منفرد *SVM* کمی سریع‌تر از نسخهٔ ۱ روش پیشنهادی برای دسته‌بندی مجموعه داده‌های کوچک است. با این حال، زمانی که مجموعه داده بزرگ‌تر می‌شود و به‌ویژه تعداد دسته‌های مسأله افزایش می‌یابد، شرایط تغییر می‌کند. در این مجموعه داده‌ها (مانند *Abalone*، *Car, Sat, Pendigits* و*Cmc*)، روش پیشنهادی سریعتر از دسته‌بند منفرد عمل می‌کند. در مقایسه با روشهای شورایی ادغام دسته‌بند و انتخاب دسته‌بند، تفاوت قابل ملاحظه‌ای در زمان اجرای روشها وجود دارد: نسخهٔ اول روش پیشنهادی، حدود ۵ بار سریع‌تر از *Bagging* و *RSM* و حدود ۱۰ بار سریع‌تر از *DCS-LA* است.

۵- اعمال روش پیشنهادی در شناسایی الگوهای بویایی

در این بخش روش پیشنهادی مقاله برای تشخیص رایحه‌های بو به‌کار گرفته شده است.مجموعه دادهٔ الگوهای بویایی، داده‌های تجربی مربوط به سه نوع شیرین‌بیان[1] است [۲۵] که امکان دسترسی به آن از طریق وب وجود دارد [۲۶]. نمونه‌های مختلف در معرض ۱۲ حسگر اکسید فلز[2] قرار گرفته و مقادیر سیگنال خروجی هر حسگر (پاسخ گذرای حسگر) به‌عنوان داده‌های مجموعه گردآوری شده‌اند. این مجموعه داده شامل ۱۸ نمونه است که هر نمونه خود شامل ۱۲ سیگنال (به تعداد حسگرها) است. لذا این مجموعه داده دارای سه بعد «حسگر × زمان × نمونه» است. بُعد حسگر شامل ۱۲ عنصر (به تعداد حسگرها)، بُعد زمان شامل ۲۴۱ عنصر (مقادیر ۲۴۰ ثانیه سیگنال خروجی حسگر همراه با مقدار مبنای حسگر $x_S(0)$)، و بُعد نمونه‌ها شامل ۱۸ عنصر (به‌تعداد

[1] licorice
[2] metal oxide sensor (MOS)

نمونه‌ها) است. نمونه‌های شیرین‌بیان شامل سه نوع خوب، بد و فاسدشده هستند که ۶ نمونه از هر نوع وجود دارد [۲۵].

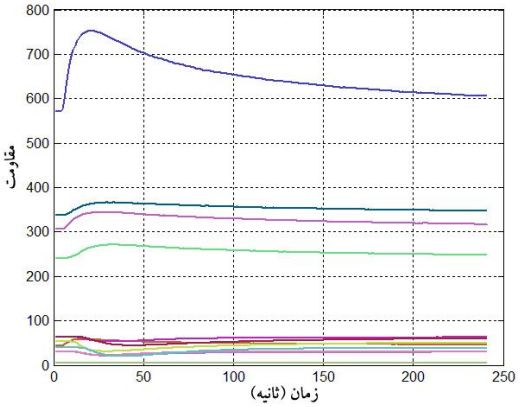
۵-۱. پیش‌پردازش سیگنالهای بو

۵-۱-۱. تنظیم مبنا

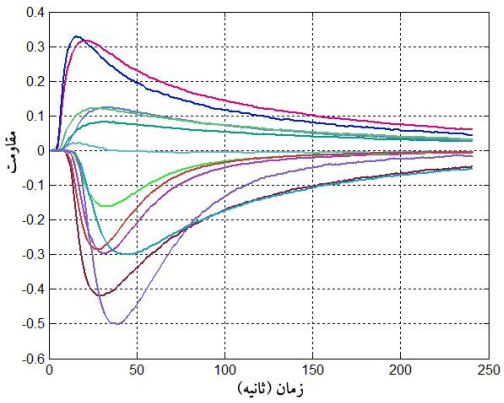
در این پژوهش از روش کسری برای تنظیم مبنای پاسخ حسگر، طبق رابطهٔ زیر، استفاده شده است که درنتیجه پاسخی بدون بُعد و هنجارشده خواهیم داشت [۲۷].

$$y_s(t) = \frac{x_s(t) - x_s(0)}{x_s(0)}$$

که $x_s(0)$ پاسخ مبنا، $x_s(t)$ پاسخ حسگر و $y_s(t)$ پاسخ تبدیل یافتهٔ حسگر است. شکل (۳) نتایج تنظیم مبنای پاسخ حسگرها برای نمونهٔ اوّل همراه با پاسخ اولیهٔ حسگرها برای این نمونه نشان می دهد.



الف)



ب)

شکل ۳. الف) پاسخ اولیهٔ ۱۲ حسگر به نمونهٔ اوّل و ب) پاسخ حسگرها پس از تنظیم مبنا به روش کسری به نمونهٔ اوّل

شکل ۴ سیگنال پاسخ ۱۲ حسگر را به نمونهٔ اوّل نشان می‌دهد. سیگنالهای پاسخ حسگرها برای نمونه‌های دیگر نیز کاملاً مشابه نمونهٔ اوّل است. مطابق شکل، سیگنال پاسخ حسگر ۱ برای نمونهٔ اول تکرارپذیری پاسخ را نشان نمی‌دهد و همچنین همراه با نویز زیادی است. از این‌رو، سیگنالهای پاسخ حسگر اوّل به ۱۸ نمونهٔ مجموعه دادهٔ شیرین بیان در پردازشهای بعدی مسأله حذف شده است.

۵-۲. استخراج ویژگی

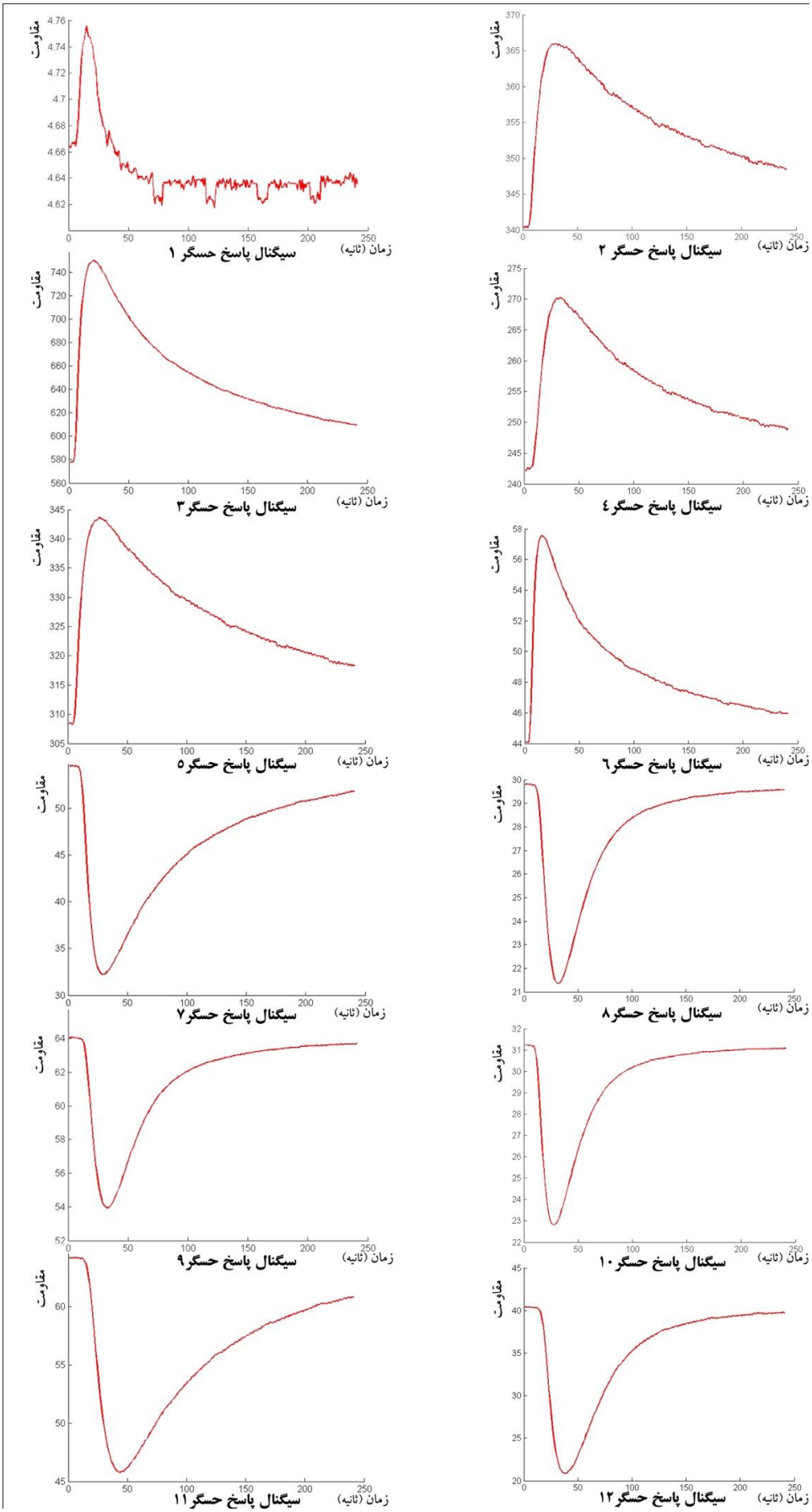
در پیش‌پردازش و استخراج ویژگی از سیگنال اولیهٔ حسگرهای گاز در بینی الکترونیکی، دو رویکرد وجود دارد: استفاده از سیگنال پاسخ در حالت مانا[1] و استفاده از پاسخ گذرای[2] سیگنال [۲۸]. روش مانا، ذخیرهٔ مقدار سیگنال پاسخ در حالت نهایی و سپس پردازش آن با روشهای مرسوم است. این روش بخشهای گذرای سیگنال که ممکن است اطلاعات زیادی را دربرداشته باشند، درنظر نمی‌گیرد و فرایند دسته‌بندی پیچیده‌تر خواهد بود [۲۹]. روش نوین در استخراج ویژگی‌های بهتر استفاده از پاسخ گذرای حسگر است. در این روش علاوه‌بر کاهش فرایند دریافت داده، موجب افزایش طول عمر حسگر نیز می‌شود [۳۰]. در سالهای اخیر برخی پژوهشها روشهای مختلفی را برای استخراج ویژگی از سیگنال حسگر به‌کار برده‌اند. مهمترین ویژگی‌هایی که در پژوهشهای مختلف با استفاده از پاسخ گذرای سیگنال استخراج شده است شامل: میانگین مقادیر بازه‌های زمانی [۳۱, ۳۲] ؛ تقسیم‌بندی زمانی پنجره‌ای[3] سیگنال حسگر [۳۳, ۳۴] ؛ مساحت زیرمنحنی (انتگرال) پاسخ حسگر [۳۱] ؛ بیشترین مقدار پاسخ حسگر [۳۵] ؛ مربع خطای پاسخ گذرای بوی نمونهٔ آزمون با پاسخ گذرای نمونهٔ مرجع [۳۶]، تبدیل *Padé-Z* سیگنال [۳۷] و نیز استفاده از تحلیل موجک سیگنال حسگر [۳۰, ۳۸] است.

در این مقاله دو ویژگی از سیگنال گذرای حسگرها استخراج شدند که عبارتند از: ۱) مقدار اکسترمم (بیشینه یا کمینهٔ) سیگنال؛ ۲) میانگین مقادیر پاسخ حسگر. از این رو ابعاد فضای ویژگی‌ها برابر ۲۲ (دو برابر تعداد حسگرها) خواهد بود. استخراج ویژگی از پاسخ گذرای حسگرها، پس از تنظیم مبنا انجام شده است.

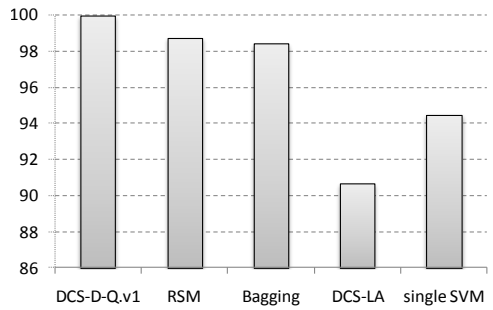
۵-۳. نتایج دسته‌بندی برای مجموعه دادهٔ بویایی

اعتباریابی آزمایشها با استفاده از روش اعتباریابی متقابل k بخشی[4] انجام شده است. از آنجا که تعداد نمونه‌های مجموعه دادهٔ شیرین بیان کم است، مقدار $k = ۱۸$ انتخاب شده است که برابر با تعداد نمونه‌ها است. این حالت که اعتباریابی متقابل *LOO* [5] نیز خوانده می‌شود، در هر بار تعداد (k-۱) نمونه برای آموزش و ۱ نمونه برای آزمون دسته‌بند استفاده می‌شود. مشابه آزمایشهای پیش، دسته‌بندهای پایهٔ شورا در دو آزمایش مختلف *MLP* و *SVM* (با پارامترهای یکسان) تعیین شدند. شکل ۵ میانگین صحّت دسته‌بندی روشهای مورد مطالعه را برای مجموعه دادهٔ بویایی نشان می‌دهد. همانگونه که شکل نشان می‌دهد، روش پیشنهادی قابلیت تشخیص ۱۰۰٪ را برای الگوهای بویایی به‌دست آورده است.

[1] steady state
[2] transient
[3] windowed time slicing
[4] K-fold cross-validation
[5] leave-one-out



شکل ۴. سیگنالهای پاسخ ۱۲ حسگر مسأله به نمونهٔ اول



شکل ۵. نتایج صحّت دسته‌بندی روش پیشنهادی برای الگوهای بویایی

۶. بحث

در روش پیشنهادی، برای دسته‌بندی صحیح نمونهٔ آزمون توّسط دسته‌بند دودویی، ابتدا لازم است دو دستهٔ محتمل در ناحیهٔ محلی به درستی پیش‌بینی شوند. به بیان دیگر، کارایی نهایی روش به کارایی حدس اوّلیه وابسته است. واکاوی مجموعه داده‌هایی که روش پیشنهادی نتیجهٔ بهتری در مقایسه با روشهای دیگر به‌دست می‌دهد، به روشنی مؤید این موضوع است. در این مجموعه‌های داده، صحّت دسته‌بندی حدس اولیه یا نزدیک به ۱۰۰٪ است یا بیش از ۲۰٪ بیشتر از صحّت دسته‌بندی نهایی روش است. این موضوع نشان می‌دهد کارایی پیش‌بینی دو دستهٔ محتمل با توجه به نوع پیچیدگی داده مناسب بوده است. از طرفی، اگر حدس اولیه کارایی خوبی نداشته باشد، صحّت نهایی دسته‌بندی روش پیشنهادی نیز مطلوب نخواهد بود. برای نمونه، برای مجموعه دادهٔ *Wine* صحّت دسته‌بندی تابع حدس اولیه کمتر از روشهای دیگر است؛ به‌عبارتی می‌توان گفت به‌طور کلّی، دو دستهٔ محتمل برای نمونه‌های این مجموعه داده به درستی تشخیص داده نشده‌اند؛ لذا روشهای دیگر، کارایی بهتری را نسبت به روش پیشنهادی نشان می‌دهند.

علاوه بر صحّت دسته‌بندی و کاهش زمان محاسباتی، روش پیشنهادی مزیتهای دیگری نیز دارد. یکی از مزیتهای آن، کارایی بهتر یادگیری در زمان اضافه شدن یک دستهٔ جدید به مسأله است. در روش پیشنهادی، زمانی که لازم است یک دستهٔ جدید فرا گرفته شود، تنها نیاز به آموزش C دسته‌بند دودویی جدید است؛ به‌طوری که دسته‌بندهای موجود تغییر نمی‌کنند.

۷- نتیجه‌گیری

در این مقاله، روشی جدید برای دسته‌بندی مسائل چنددسته‌ای بر اساس رویکرد انتخاب دسته‌بند ارائه شده است. در این روش، همانند سایر روشهای انتخاب دسته‌بند، همسایگی (ناحیهٔ محلّی) هر الگو تعریف می‌شود. امّا به جای ارزیابی دسته‌بندهای مختلف براساس صحّت دسته‌بندی در آن ناحیه، فراوانی الگوهای دسته‌های مختلف شمارش می‌شود. دو دسته‌ای که بیشترین الگو را در ناحیهٔ همسایگی داشته باشند، به‌عنوان محتمل‌ترین الگوهای نمونهٔ آزمون انتخاب می‌شوند. سپس یکی از دسته‌بندهای دودویی برای تعیین برچسب نهایی نمونهٔ آزمون انتخاب می‌شود. مشخصاً این فرایند بسیار سریع‌تر از ارزیابی صحّت هر دسته‌بند شورا در آن ناحیه است. نتایج ارزیابی روش پیشنهادی بر روی ۱۴ مجموعه دادهٔ تراز و مجموعه دادهٔ تجربی الگوهای بویایی، مزیت روش را در مقایسه با دسته‌بند منفرد، روشهای *DCS-LA*، *Bagging*، و *RSM* نشان می‌دهد: روش پیشنهادی، صحّت دسته‌بندی بیشتر و سرعت محاسباتی کمتری در مقایسه با روشهای ذکر شده دارد.

شایان ذکر است که هدف این پژوهش، یافتن پارامترهای بهینه برای الگوریتمهای دسته‌بندی نیست. برای برخی مجموعه‌های داده، تغییر پارامترهای دسته‌بند ممکن است نتایج دسته‌بندی را به میزان قابل توجهی بهبود دهد؛ بااین حال مقایسهٔ نتایج نشان می‌دهد حتی اگر دسته‌بند پایه قوی نباشد، روش پیشنهادی با ایجاد مرزهای دقیق بین هر دو دسته، کارایی پایین دسته‌بند پایه را جبران می‌کند.

مراجع

[1] L. Hansen and P. Salamon, "Neural network ensembles," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 12, pp. 993-1001, Oct. 1990.

[2] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in Advances in Neural Information Processing Systems, vol. 7, D. Touretzky and T. Leen, Eds. Cambridge, MA: MIT Press, 1995, pp. 231-238.

[3] S. Hashem, "Optimal linear combinations of neural networks: an overview,," Neural Networks, vol. 10, pp. 599–614, 1997.

[4] L. I. Kuncheva, M. Skurichina, and R. P. W. Duin, "An experimental study on diversity for bagging and boosting with linear classifiers," Information Fusion, vol. 3, pp. 245-258, 2000.

[5] L. Y. Yang, Z. Qin, and R. Huang, "Design of a multiple classifier system," in IEEE Proceedings of the Third International conference on Machine Learning and Cybernetics, Shanghai, 2004, pp. 3272–3276.

[6] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. New York, NY: Wiley, 2004.

[7] T. K. Ho, "T. K. Ho. Multiple classifier combination: Lessons and the next steps," in Hybrid Methods in Pattern Recognition, A. Kandel and H. Bunke, Eds.: World Scientific Publishing, 2002, pp. 171–198.

[8] K. Woods, W. P. J. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 405–410, 1997.

[9] A. Sharkey, N. Sharkey, U. Gerecke, and G. Chandroth, "The "Test and Select" Approach to Ensemble Combination," in Multiple Classifier Systems, vol. 1857: Springer Berlin / Heidelberg, 2000, pp. 30-44.

[10] E. M. Dos Santos, R. Sabourin, and P. Maupin, "A dynamic overproduce-and-choose strategy for the selection of classifier ensembles," Pattern Recognition, vol. 41, pp. 2993-3009, 2008.

[DOR: 20.1001.1.20088345.1391.6.1.2.4]

[11] T. G. Dietterich, "Machine learning research: Four current directions," Artificial Intell. Mag., vol. 18, pp. 97-136, 1997.
[12] R. Polikar, "Ensemble based systems in decision making," IEEE Circuits and Systems Magazine, vol. 6, pp. 21-45, 2006.
[13] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, pp. 123-140, 1996.
[14] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in Proceeding of the Thirteenth International Conference on Machine Learning, 1996, pp. 148-156.
[15] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattem Analysis and Machine Intelligence, vol. 20, pp. 832-844, 1998.
[16] B. V. Dasarathy and B. V. Sheela, "A composite classifier system design: Concepts and methodology," Proceedings of the IEEE, vol. 67, pp. 708-713, 1979.
[17] G. Giacinto and F. Roli., "An approach to the automatic design of multiple classifier systems," Pattern Recognition Letters, vol. 22, pp. 25–33, 2001.
[18] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases, Department of Information and Computer Sciences, University of California, Irvine," 1998. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
[19] E. Bauer and R. Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," Machine Learning, vol. 36, pp. 105-139, 1999.
[20] C.-C. Chang and C.-J. Lin. (2001). LIBSVM: A library for support vector machines. [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
[21] J. Loughrey and P. Cunningham, "Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets," in Research and Development in Intelligent Systems XXI, M. Bramer, F. Coenen, and T. Allen, Eds.: Springer London, 2005, pp. 33-43.
[22] H. Zhang and G. Sun, "Feature selection using tabu search method," Pattern Recognition, vol. 35, pp. 701-711, 2002.
[23] H. Fröhlich, O. Chapelle, and B. Schölkopf, "Feature Selection for Support Vector Machines by Means of Genetic Algorithms," in Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, 2003, vol. 142, pp. 148-154.
[24] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A Comparison of Decision Tree Ensemble Creation Techniques," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 173-180, 2007.
[25] T. Skov and R. Bro, "A new approach for modelling sensor based data," Sensors and Actuators B, vol. 106 pp. 719–729, 2005.
[26] T. Skov and R. Bro. (2004). Three-way electronic nose data. [Online]. Available: http://www.models.kvl.dk /3Dnosedata.
[27] K. Arshak, E. Moore, G. M. Lyons, F. Harris, and S. Clifford, "A review of gas sensors employed in electronic nose application," Sens. Rev, vol. 24, pp. 181–198, 2004.
[28] F. Hossein-Babaei and V. Ghafarinia, "Compensation for the drift-like terms caused by environmental fluctuations in the responses of chemoresistive gas sensors," Sensors and Actuators B: Chemical, vol. 143, pp. 641-648, 2010.
[29] R. Gutierrez-Osuna and H. T. Nagle, "A method for evaluating data preprocessing techniques for odor classification with an array of gas sensors," IEEE Trans. Syst. Man, Cybern. B, Cybern, vol. 29, pp. 626-632, 1999.
[30] E. Phaisangittisagul and H. T. Nagle, "Sensor Selection for Machine Olfaction Based on Transient Feature Extraction," IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, vol. 57, pp. 369-378, 2008.
[31] S. Balasubramanian, S. Panigrahi, C. M. Logue, H. Gu, and M. Marchello, "Neural networks-integrated metal oxide-based artificial olfactory system for meat spoilage identification," Journal of Food Engineering, vol. 91, pp. 91-98, 2009.
[32] W. Li, H. Leung, C. Kwan, and B. R. Linnell, "E-Nose Vapor Identification Based on Dempster–Shafer Fusion of Multiple Classifiers," IEEE Transactions on Instrumentation and Measurement, vol. 57, pp. 2273-2282, 2008.
[33] B. G. Kermani, On Using Artificial Neural Networks And Genetic Algorithm for Electronic Nose. PhD Dissertation, North Carolina State University, 1996.
[34] B. G. Kermani, S. S. Schiffman, and H. T. Nagle, "Performance of the Levenberg–Marquardt neural network training method in electronic nose applications," Sensors and Actuators B, vol. 110 pp. 13–22, 2005.
[35] J. Fu, G. Li, Y. Qin, and W. J. Freeman, "A pattern recognition method for electronic noses based on an olfactory neural network," Sensors and Actuators B, vol. 125 pp. 489–497, 2007.
[36] F. Hossein-Babaei, M. Hemmati, and M. Dehmobed, "Gas diagnosis by a quantitative assessment of the transient response of a capillary-attached gas sensor," Sensors and Actuators B, vol. 107 pp. 461–467, 2005.
[37] F. Hossein-Babaei, S. M. Hosseini-Golgoo, and A. Amini, "Extracting discriminative information from the Padé-Z-transformed responses of a temperature-modulated chemoresistive sensor for gas recognition," Sensors and Actuators B: Chemical, vol. 142, pp. 19-27, 2009.
[38] Y. Yin, H. Yu, and H. Zhang, "A feature extraction method based on wavelet packet analysis for discrimination of Chinese vinegars using a gas sensors array," Sensors and Actuators B: Chemical, vol. 134, pp. 1005-1009, 2008.