عملکرد روش‌ پیشنهادی برای انتخاب ویژگی ها مبتنی بر تکنیک های خوشه بندی و الگوریتم‏های بهینه سازی مورد ارزیابی قرار می‌گیرد.

برای پیاده‌سازی روش‌های انتخاب ویژگی از زبان برنامه‌نویسی متلب استفاده شده است. همچنین تمام آزمایش‌ها بر روی یک سیستم با پردازنده Corei3 2.3 GHz و 2 گیگابایت حافظه داخلی (RAM) اجرا شده‌اند.

## مجموعه‌های داده‌ای

در بعضی از این مجموعه‌های داده‌ای، چندین مقدار از دست رفته وجود دارد. برای غلبه بر این مشکل، مقادیر از دست‌ رفته در این ویژگی‌ها، با میانگین داده‌های موجود متناظر با آن ویژگی که در دسترس است، جایگذاری می‌شود[[6](#_ENREF_6)].

جدول ‏4‑1- مشخصات مجموعه‌های داده‌ای

|  |  |  |  |
| --- | --- | --- | --- |
| ‏**Dataset** | **Features** | **Classes** | **Patterns** |
| Wine | 13 | 3 | 178 |
| Hepatitis | 19 | 2 | 155 |
| WDBC | 30 | 2 | 569 |
| Ionosphere | 34 | 2 | 351 |
| Spambase | 57 | 2 | 4601 |
| Sonar | 60 | 2 | 208 |
| Arcene | 10000 | 2 | 900 |
| Colon | 2000 | 2 | 62 |

## طبقه‌بندی کننده‌ها

برای نشان دادن قابلیت تعمیم روش‌ پیشنهادی در طبقه‌بندی کننده‌های مختلف، در آزمایش‌ها از چهار طبقه‌بندی کننده ماشین بردار پشتیبان[[1]](#footnote-1)(SVM)، درخت تصمیم[[2]](#footnote-2)(DT)، بیز ساده[[3]](#footnote-3)(NB)، و K-نزدیک‌ترین همسایه[[4]](#footnote-4)(KNN) استفاده ‌شده است.

 برای پیاده‌سازی طبقه‌بندی‌ کننده‌های ذکر شده از ابزار وکا( WEKA ) [[7](#_ENREF_7)]، استفاده‌ شده است. به این منظور، برای پیاده‌سازی طبقه‌بندی کننده‌های SVM، DT، NB و KNN به ترتیب SMO، J48 (نسخه پیاده‌سازی از الگوریتم C4.5)، Naïve Bayes و IBk به‌کار رفته‌اند. همچنین از مقادیر پیش‌فرض در ابزار وکا، برای تنظیم پارامتر‌های این طبقه‌بندی کننده‌ها استفاده‌ شده است.

## نتایج عملی

برای بررسی کارایی روش‌های پیشنهادی بر روی مجموعه‌های داده‌ای مختلف با استفاده از طبقه‌بندی کننده‌های معرفی‌ شده در بخش قبل، آزمایش‌هایی بر روی آن‌ها انجام ‌شده است. در آزمایش‌های صورت گرفته بر روی روش پیشنهادی، مجموعه داده‌ای به‌طور تصادفی به داده‌‌های آموزشی، داده‌های اعتبارسنجی و داده‌های آزمایشی تقسیم می‌شوند. برای این کار 50 درصد از مجموعه داده‌ای به‌عنوان داده آموزشی، 25 درصد به‌عنوان داده اعتبارسنجی و 25 درصد مابقی نیز به‌عنوان داده آزمایشی در نظر گرفته می‌شود. همچنین در تمام آزمایش پس از مشخص‌ شدن مجموعه‌های آموزشی، آزمایشی و اعتبارسنجی، هر روش انتخاب ویژگی ده بار اجرا شده و از میانگین ده اجرای مختلف برای مقایسه روش‌های مختلف استفاده می‌شود.

ابتدا طبق فرمول ارسال شده که در همین فایل توضیح دادم تکه کد نوشته شود و جایگزین شود سپس یکبار با نسبت داده اموزش 70، تست 15 و اعتبار 15 از سیستم خروجی گرفته شود و جداگانه گزارش شود و یکبار هم بدون تکه کد جدید یعنی با همان کد قبلی ولی روی همین نسبتها خروجی گرفته شود و گزارش شود و در هر مرحله خروجی گرفتن توضیحات مربوطه ارسال شود. یکبار هم همین مراحل بالا با نسبت داده اموزش 33، تست 33 و اعتبار 33 روی هر دو حالت فرمول جدید و قدیم خروجی گرفته شود و گزارش شود.

یکبار هم با فرمول جدید و تکه کد جدید با نسبت داده اموزش 70 تست 30 گزارش نویسی شود.

توضیحات اضافه تر مربوط به کد جدید که کجا قرار گیرد و با مرحله نرمالسازی و بدون نرمالسازی در همین فایل در ادامه بصورت هایلایت مشخص شده است.

برای ارزیابی از طبقه بندی کننده نزدیک ترین همسایه (KNN) استفاده می شود.

در واقع ابتدا با استفاده از روش ارائه شده ویژگی های مناسب انتخاب و سپس با استفاده از این طبقه بندی کننده ها، دقت طبقه بندی بررسی می شود.

 روش پیشنهادی دارای سه مرحله است.

1. بازنمایی گرافی
2. خوشه بندی ویژگی ها
3. انتخاب ویژگی های نهایی از هر خوشه

در ابتدا و قبل از شروع مراحل روش پیشنهادی باید دیتاست خوانده شود. کد این قسمت به صورت زیر است:

\*\*\*\*\*\*\*\*\* کد نوشته شده هست.

در مرحله اول باید شباهت بین ویژگی ها را حساب کرد. لازم به ذکر است برای افزایش دقت محاسبه شباهت، قبل از محاسبه، دیتاست نرمال سازی می شود. کد این قسمت به صورت زیر می باشد.

کد نوشته شده است.

 در بخش محاسبه شباهت یعنی قسمت پایین بجاش فرمول زیر پیاده سازی شود

|  |  |
| --- | --- |
| $$Sim\left(S\_{i},S\_{j}\right)=\frac{Cov\left(S\_{i}\right)Cov(S\_{j})}{\sqrt{Var\left(S\_{i}\right)Var(S\_{j})}}$$توضیحات فرمول بالا برای مشخص شدن هدف:در رابطه بالا $Cov(S\_{i})$ نشان دهنده کواریانس بردار ویژگی $x\_{i}$ و همچنین $Var(S\_{i})$ نشان دهنده محاسبه واریانس برای بردار ویژگی $x\_{i} $ می باشد. همان طور که در این رابطه دیده می شود اگر دو ویژگی کاملا مشابه هم باشند در این حالت میزان شباهت برابر 1 و یا 1- می شود و دو ویژگی که کاملا مستقل از همدیگر باشند در این حالت میزان شباهت آنها برابر صفر خواهد بود. |  |

% Calculate Similarity between Data

Similarity=zeros(NumberFeature, NumberFeature);

for i=1:NumberFeature

 for j=1:NumberFeature

 Similarity(i,j)= abs(exp(-1.\*( sqrt(sum((Data(:,i)-Data(:,j)))))));

 end

end

همچنین خود شباهت محاسبه شده در این قسمت نیز نرمال سازی می شود:

اگر شباهت بدست امده با فرمول جدید در محاسبه دقت و زمان اجرا بهتر از حالت قبلی است پس مرحله نرمالسازی حذف شود یعنی که در گرفتن خروجی فرمول جدید یکبار با نرمالسازی بعد از شباهت و یکبار بدون نرمالسازی چک شود هر کدام از نتایج بهتر بود گذاشته شود طبعا باید زمان اجرا کمتر شود و دقت بیشتر شود. همچنین نتایج بر اساس فرمول جدید و قبلی مقایسه شود هر کدام از مراحل محاسبه شباهت که باعث شد زمان اجرا کمتر و دقت بیشتر شود همان در نتایج نهایی لحاظ شود و در توضیحات جدید برای گرفتن نتایج نیز ذکر شود که از کدام فرمول جدید یا قدیم و نرمالسازی استفاده شده یا خیر.

\*\*\*\*\*\* کد نوشته شده.

در نتیجه خروجی مرحله اول روش پیشنهادی یک ماتریس است که شباهت دو به دوی ویژگی ها را نمایش می دهد. برای نمونه اگر ما 100 ویژگی داشته باشیم خروجی مرحله اول یک ماتریس 100\*100 است که هر درایه آن شباهت ویژگی ها رو نشان می دهد. هر درایه این ماتریس یک عدد بین صفر و یک را نشان می دهد که هر چه عدد به 1 نزدیک تر باشد یعنی شباهت آن دو ویژگی بیشتر است.

در مرحله دوم باید با استفاده از گراف ساخته شده در مرحله قبل ویژگی ها به چند دسته خوشه‌بندی شوند. در واقع در این مرحله مشخص می شود کدام ویژگی ها شبیه به هم هستند و در یک خوشه قرار دارند.

پس از تعیین شدن خوشه هر ویژگی، در مرحله سوم ابتدا ارزش هر ویژگی انتخاب می شود.

پس از این قسمت حال از هر خوشه، ویژگی که بیشترین ارزش را داشته باشد به عنوان نماینده آن خوشه انتخاب می شود.

پس از این مرحله و انتخاب ویژگی های نهایی نوبت ارزیابی کارایی روش انتخاب ویژگی است. برای این کار باید یک دیتاست جدید با استفاده از ویژگی های انتخاب شده ساخته شود:

و سپس باید با استفاده از یکی از طبقه بندهای مشخص شده، عملکرد طبقه بند ارزیابی شود.

در اینجا منظور از طبقه بندها توضیحات زیر می باشد:

برای نشان دادن قابلیت تعمیم روش‌ پیشنهادی در طبقه‌بندی کننده‌های مختلف، در آزمایش‌ها از چهار طبقه‌بندی کننده ماشین بردار پشتیبان[[5]](#footnote-5)(SVM)، درخت تصمیم[[6]](#footnote-6)(DT)، بیز ساده[[7]](#footnote-7)(NB)، و K-نزدیک‌ترین همسایه[[8]](#footnote-8)(KNN) استفاده ‌شده است.

 برای پیاده‌سازی طبقه‌بندی‌ کننده‌های ذکر شده از ابزار وکا( WEKA ) [[7](#_ENREF_7)]، استفاده‌ شده است. به این منظور، برای پیاده‌سازی طبقه‌بندی کننده‌های SVM، DT، NB و KNN به ترتیب SMO، J48 (نسخه پیاده‌سازی از الگوریتم C4.5)، Naïve Bayes و IBk به‌کار رفته‌اند. همچنین از مقادیر پیش‌فرض در ابزار وکا، برای تنظیم پارامتر‌های این طبقه‌بندی کننده‌ها استفاده‌ شده است.

1. Support vector machines [↑](#footnote-ref-1)
2. Decision tree [↑](#footnote-ref-2)
3. Naive bayes [↑](#footnote-ref-3)
4. K-nearest neighbors [↑](#footnote-ref-4)
5. Support vector machines [↑](#footnote-ref-5)
6. Decision tree [↑](#footnote-ref-6)
7. Naive bayes [↑](#footnote-ref-7)
8. K-nearest neighbors [↑](#footnote-ref-8)