در این بخش، عملکرد روش‌ پیشنهادی برای انتخاب ویژگی ها در کلان داده ها با استفاده از پردازش موازی مدل نگاشت و کاهش در بستر هدوپ مبتنی بر تکنیک های خوشه بندی و الگوریتم‏های بهینه سازی مورد ارزیابی قرار می‌گیرد. به این منظور، روش‌ پیشنهادی با جدیدترین روش‌های انتخاب ویژگی ها در کلان داده ها که در فرایند انتخاب ویژگی از الگوریتم‌های تکاملی استفاده می‌کنند، مقایسه می‌شود. روش‌های انتخاب ویژگی برای مقایسه با روش‌های پیشنهادی، عبارت‌اند از:

**الگوریتم ژنتیک ترکیبی بر مبنای جستجوی محلی برای انتخاب ویژگی**[[1]](#footnote-1)(**HGAFS**)[[1](#_ENREF_1)]

**الگوریتم بهینه‌سازی کلونی مورچه‌های ترکیبی برای انتخاب ویژگی**[[2]](#footnote-2)(**ACOFS**)[[2](#_ENREF_2)]

**الگوریتم بهینه‌سازی ازدحام ذرات برای انتخاب ویژگی[[3]](#footnote-3)(PSOFS)**[[3](#_ENREF_3)]

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

## مجموعه‌های داده‌ای

در این پایان‌نامه، از چندین مجموعه داده‌ای، با مشخصات مختلف جهت ارزیابی روش‌های پیشنهادی و مقایسه عملکرد آن‌ با سایر روش‌های انتخاب ویژگی استفاده ‌شده است. این مجموعه‌های داده‌ای شامل Wine، Hepatitis، Wisconsin Diagnostic Breast Cancer(WDBC)، Ionosphere، Spambase، Sonar، Arcene و Colon هستند. تمام این مجموعه‌های داده‌ای به ‌جز آخرین مجموعه داده‌ای (Colon) از پایگاه داده دانشگاه کالیفرنیا [[4](#_ENREF_4)]، انتخاب‌ شده است. همچنین، مشخصات و جزییات مجموعه داده‌ای Colon توسط آقای آلون[[4]](#footnote-4) و همکارانش [[5](#_ENREF_5)] ارائه‌ شده است. مشخصات کلی این مجموعه‌های داده‌ای در جدول 4-1 نشان داده ‌شده است..

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[[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.

## تنظیم پارامتر‌ها

جدول 4-2 مقادیر پارامتر‌های مختلف را برای روش‌های پیشنهادی نشان می‌دهد.

جدول ‏4‑2- تنظیم پارامتر‌های روش پیشنهادی

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Notation** | **Value** |
| Number of iteration |  | 50 |
| Population size |  | Number of features |
| Size of reduced features |  | 100 |
| Threshold for remove edges |  | 0.5 |
| Number of selected feature from each cluster |  | 2 |

## 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.

به ‌منظور بررسی آماری نتایج به ‌دست‌ آمده، از آزمون فریدمن استفاده می‌شود. لازم به ذکر است، در این بخش روش پیشنهادی با استراتژی ترمیم تصادفی و ترمیم امتیازی به ترتیب، با عناوین GCPSO\_Random و GCPSO\_Score نامیده شده‌اند.

### اندازه زیرمجموعه ویژگی انتخاب‌ شده

در این بخش روش‌های مختلف از نظر تعداد ویژگی‌هایی که به‌عنوان زیرمجموعه نهایی انتخاب می‌کنند مقایسه می‌شوند. جدول 4-3 میانگین اندازه زیرمجموعه ویژگی انتخاب‌ شده توسط هر کدام از روش‌ها را بر روی مجموعه داده‌های مختلف نشان می‌دهد.

جدول ‏4‑3- میانگین اندازه زیرمجموعه‌ انتخاب شده توسط روش‌ پیشنهادی در مقایسه با سایر روش‌ها

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Feature Selection Method** | | | | |
| **GCPSO** | **HGAFS** | **ACOFS** | **PSOFS** | **All features** |
| Wine | 6.2 | 4.7 | 4.9 | 7.3 | 13 |
| Hepatitis | 6.8 | 5.2 | 5.1 | 9.4 | 19 |
| WDBC | 8.2 | 6.2 | 6.4 | 13.7 | 30 |
| Ionosphere | 6.8 | 6.5 | 6.7 | 15.2 | 34 |
| Spambase | 7.8 | 6.9 | 6.8 | 17.1 | 57 |
| Sonar | 8.2 | 7.3 | 7.1 | 21.6 | 60 |
| Arcene | 11.4 | 8.4 | 8.6 | 57.6 | 279 |
| Colon | 10.6 | 12.8 | 11.9 | 96.9 | 2000 |
| Average | 8.25 | 7.25 | 7.19 | 29.85 | 311.50 |

#### 

### دقت طبقه‌بندی

برای بررسی قابلیت تعمیم روش‌های انتخاب ویژگی مختلف از چهار طبقه‌بندی کننده SVM، DT، NB و KNN استفاده‌ شده است. تمام نتایج این جدول‌ها از میانگین دقت طبقه‌بندی در 10 اجرای مستقل برای روش‌های مختلف به‌ دست ‌آمده است.

جدول ‏4‑4- میانگین و واریانس دقت طبقه‌بندی در روش پیشنهادی در مقایسه با سایر روش‌های انتخاب ویژگی بر روی طبقه‌بندی کننده SVM. در اینجا Acc دقت طبقه‌بندی و Std انحراف معیار در ده اجرای مستقل را نشان می‌دهد.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** |  | **Feature Selection Method** | | | | | |
| **GCPSO\_Random** | **GCPSO\_Score** | **HGAFS** | **ACOFS** | **PSOFS** | **All features** |
| Wine | Acc (%) | 94.26 | 93.11 | 92.78 | 93.44 | 94.42 | 96.38 |
| Std | 1.57 | 1.55 | 1.98 | 1.92 | 1.48 | 0.83 |
| Hepatitis | Acc (%) | 84.14 | 87.54 | 81.50 | 82.63 | 83.95 | 74.33 |
| Std | 1.48 | 1.68 | 1.57 | 1.46 | 1.42 | 2.60 |
| WDBC | Acc (%) | 94.76 | 95.38 | 93.15 | 92.76 | 94.81 | 96.11 |
| Std | 1.13 | 1.13 | 1.44 | 1.47 | 1.25 | 0.52 |
| Ionosphere | Acc (%) | 90.49 | 91.25 | 87.97 | 89.49 | 90.75 | 86.46 |
| Std | 1.33 | 1.19 | 1.63 | 1.52 | 1.36 | 1.10 |
| Spambase | Acc (%) | 88.22 | 89.76 | 85.04 | 87.30 | 88.69 | 88.39 |
| Std | 1.22 | 1.11 | 1.38 | 1.10 | 1.18 | 1.22 |
| Sonar | Acc (%) | 81.96 | 84.36 | 76.33 | 79.29 | 78.72 | 76.05 |
| Std | 1.20 | 1.53 | 2.04 | 2.02 | 1.89 | 1.72 |
| Arcene | Acc (%) | 61.16 | 61.55 | 60.06 | 61.68 | 62.52 | 56.91 |
| Std | 2.39 | 2.32 | 2.42 | 2.56 | 2.34 | 1.77 |
| Colon | Acc (%) | 82.85 | 83.33 | 83.81 | 83.80 | 85.23 | 67.14 |
| Std | 1.82 | 2.26 | 2.60 | 2.36 | 2.55 | 2.47 |
| Average | Acc (%) | 84.73 | 85.78 | 82.58 | 83.79 | 84.88 | 80.22 |
|  | Std | 1.51 | 1.79 | 1.88 | 1.80 | 1.68 | 1.58 |

جدول‏4‑5- میانگین و واریانس دقت طبقه‌بندی در روش پیشنهادی در مقایسه با سایر روش‌های انتخاب ویژگی بر روی طبقه‌بندی کننده DT. در اینجا Acc دقت طبقه‌بندی و Std انحراف معیار در ده اجرای مستقل را نشان می‌دهد.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** |  | **Feature Selection Method** | | | | | |
| **GCPSO\_Random** | **GCPSO\_Score** | **HGAFS** | **ACOFS** | **PSOFS** | **All features** |
| Wine | Acc (%) | 92.29 | 92.45 | 91.30 | 91.96 | 92.78 | 92.45 |
| Std | 1.56 | 1.48 | 2.00 | 1.84 | 1.44 | 1.06 |
| Hepatitis | Acc (%) | 83.57 | 84.52 | 80.74 | 81.50 | 84.14 | 72.82 |
| Std | 1.33 | 1.22 | 1.57 | 1.39 | 1.42 | 2.36 |
| WDBC | Acc (%) | 93.72 | 94.40 | 91.96 | 91.37 | 94.08 | 94.86 |
| Std | 1.40 | 1.56 | 1.14 | 1.46 | 1.23 | 1.07 |
| Ionosphere | Acc (%) | 90.24 | 90.83 | 86.71 | 88.56 | 88.14 | 83.61 |
| Std | 1.47 | 1.15 | 1.16 | 1.56 | 1.42 | 2.07 |
| Spambase | Acc (%) | 85.93 | 88.48 | 83.01 | 84.90 | 86.36 | 84.94 |
| Std | 1.79 | 1.82 | 1.67 | 1.90 | 1.89 | 1.72 |
| Sonar | Acc (%) | 82.53 | 83.37 | 79.99 | 80.83 | 76.75 | 74.08 |
| Std | 1.60 | 1.62 | 2.32 | 1.60 | 2.15 | 2.21 |
| Arcene | Acc (%) | 58.43 | 60.38 | 57.39 | 59.93 | 61.97 | 53.30 |
| Std | 2.34 | 2.25 | 2.39 | 2.45 | 2.45 | 1.14 |
| Colon | Acc (%) | 81.90 | 83.33 | 80.48 | 82.85 | 84.76 | 63.80 |
| Std | 1.93 | 2.01 | 2.62 | 2.45 | 2.50 | 1.82 |
| Average | Acc (%) | 83.57 | 84.72 | 81.44 | 82.73 | 83.62 | 77.44 |
|  | Std | 1.67 | 1.63 | 1.85 | 1.83 | 1.81 | 1.68 |

جدول‏4‑6- میانگین و واریانس دقت طبقه‌بندی در روش پیشنهادی در مقایسه با سایر روش‌های انتخاب ویژگی بر روی طبقه‌بندی کننده NB. در اینجا Acc دقت طبقه‌بندی و Std انحراف معیار در ده اجرای مستقل را نشان می‌دهد.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** |  | **Feature Selection Method** | | | | | |
| **GCPSO\_Random** | **GCPSO\_Score** | **HGAFS** | **ACOFS** | **PSOFS** | **All features** |
| Wine | Acc (%) | 92.12 | 92.45 | 94.09 | 91.14 | 92.78 | 95.40 |
| Std | 1.51 | 1.48 | 1.70 | 1.60 | 1.44 | 1.22 |
| Hepatitis | Acc (%) | 84.33 | 85.65 | 80.18 | 82.51 | 84.17 | 72.82 |
| Std | 1.33 | 1.14 | 1.35 | 1.37 | 1.49 | 2.52 |
| WDBC | Acc (%) | 92.89 | 95.64 | 91.60 | 92.95 | 94.81 | 94.91 |
| Std | 1.34 | 1.26 | 1.02 | 1.62 | 1.25 | 1.29 |
| Ionosphere | Acc (%) | 89.99 | 90.83 | 86.13 | 88.65 | 88.06 | 85.62 |
| Std | 1.57 | 1.15 | 0.66 | 1.67 | 1.21 | 1.00 |
| Spambase | Acc (%) | 85.29 | 88.48 | 82.11 | 84.45 | 85.58 | 89.03 |
| Std | 1.93 | 1.82 | 1.58 | 1.99 | 2.01 | 0.81 |
| Sonar | Acc (%) | 82.24 | 83.37 | 80.70 | 79.99 | 80.13 | 76.61 |
| Std | 1.63 | 1.62 | 2.08 | 1.57 | 1.34 | 1.60 |
| Arcene | Acc (%) | 60.64 | 62.20 | 61.42 | 63.11 | 62.55 | 56.03 |
| Std | 1.96 | 2.23 | 2.31 | 2.65 | 2.58 | 1.71 |
| Colon | Acc (%) | 81.42 | 81.90 | 79.52 | 82.37 | 83.33 | 67.62 |
| Std | 2.04 | 2.21 | 2.59 | 2.34 | 2.74 | 2.58 |
| Average | Acc (%) | 83.61 | 85.06 | 81.96 | 83.14 | 83.92 | 79.75 |
|  | Std | 1.66 | 1.61 | 1.66 | 1.85 | 1.75 | 1.59 |

جدول ‏4‑7- میانگین و واریانس دقت طبقه‌بندی در روش پیشنهادی در مقایسه با سایر روش‌های انتخاب ویژگی بر روی طبقه‌بندی کننده KNN. در اینجا Acc دقت طبقه‌بندی و Std انحراف معیار در ده اجرای مستقل را نشان می‌دهد.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** |  | **Feature Selection Method** | | | | | |
| **GCPSO\_Random** | **GCPSO\_Score** | **HGAFS** | **ACOFS** | **PSOFS** | **All features** |
| Wine | Acc (%) | 94.26 | 96.39 | 92.29 | 94.09 | 93.76 | 92.78 |
| Std | 1.57 | 1.30 | 2.02 | 1.60 | 1.55 | 1.17 |
| Hepatitis | Acc (%) | 84.71 | 87.35 | 81.88 | 82.06 | 84.14 | 72.07 |
| Std | 1.44 | 1.58 | 1.64 | 1.35 | 1.64 | 2.38 |
| WDBC | Acc (%) | 94.76 | 95.28 | 92.27 | 93.05 | 94.29 | 95.12 |
| Std | 1.13 | 1.07 | 1.36 | 1.61 | 1.21 | 1.08 |
| Ionosphere | Acc (%) | 90.83 | 91.75 | 87.13 | 89.32 | 90.14 | 85.62 |
| Std | 1.07 | 0.97 | 1.49 | 1.57 | 1.35 | 1.00 |
| Spambase | Acc (%) | 88.58 | 90.29 | 84.29 | 86.57 | 88.18 | 85.94 |
| Std | 1.36 | 1.11 | 1.39 | 1.27 | 1.20 | 1.71 |
| Sonar | Acc (%) | 83.23 | 84.78 | 76.61 | 80.41 | 78.72 | 73.37 |
| Std | 1.53 | 1.57 | 1.84 | 2.05 | 1.89 | 2.14 |
| Arcene | Acc (%) | 64.99 | 63.63 | 61.68 | 62.00 | 61.74 | 53.95 |
| Std | 2.53 | 2.57 | 2.44 | 2.74 | 2.40 | 1.13 |
| Colon | Acc (%) | 83.80 | 85.71 | 81.85 | 83.33 | 83.33 | 64.76 |
| Std | 2.14 | 1.97 | 2.53 | 2.53 | 2.53 | 2.00 |
| Average | Acc (%) | 85.64 | 86.89 | 82.25 | 83.85 | 84.28 | 77.95 |
|  | Std | 1.59 | 1.51 | 1.83 | 1.84 | 1.72 | 1.57 |

|  |
| --- |
|  |
|  |
|  |
|  |

### زمان اجرا

جدول 4-9 میانگین زمان اجرا در ده اجرای مستقل در روش‌های مختلف انتخاب ویژگی را نشان می‌دهد.

جدول ‏4‑9- میانگین زمان اجرا (بر حسب دقیقه) در روش پیشنهادی در مقایسه با سایر روش‌ها

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **Feature Selection Method** | | | | |
| **GCPSO\_Random** | **GCPSO\_Score** | **HGAFS** | **ACOFS** | **PSOFS** |
| Wine | 0.12 | 0.16 | 0.53 | 0.41 | 0.25 |
| Hepatitis | 0.16 | 0.21 | 1.28 | 1.17 | 0.38 |
| WDBC | 4.16 | 4.23 | 5.36 | 4.81 | 7.43 |
| Ionosphere | 2.86 | 2.91 | 3.12 | 2.64 | 5.63 |
| Sonar | 1.13 | 1.19 | 1.27 | 1.05 | 1.08 |
| Arcene | 9.81 | 9.89 | 7.84 | 6.32 | 11.57 |
| Colon | 0.91 | 0.96 | 0.93 | 0.81 | 374.82 |
| Average | 2.73 | 2.79 | 2.90 | 2.45 | 57.30 |

1. Rostami, M., Berahmand, K. and Forouzandeh, S., 2021. A novel community detection based genetic algorithm for feature selection. Journal of Big Data, 8(1), pp.1-27.
2. Paniri, M., Dowlatshahi, M.B. and Nezamabadi-pour, H., 2020. MLACO: A multi-label feature selection algorithm based on ant colony optimization. Knowledge-Based Systems, 192, p.105285.
3. Sharif, M., Amin, J., Raza, M., Yasmin, M. and Satapathy, S.C., 2020. An integrated design of particle swarm optimization (PSO) with fusion of features for detection of brain tumor. Pattern Recognition Letters, 129, pp.150-157.
4. Asuncion, A. and D. Newman, UCI repository of machine learning datasets. Availablefrom: <http://archive.ics.uci.edu/ml/datasets.html>, 2007.
5. Alon, U., et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In Proceedings of the national academy of sciences, USA, 1999. 96: p. pp. 6745–6750.
6. Theodoridis, S. and K. Koutroumbas, Pattern Recognition. Academic Press, Oxford., 2008.
7. Hall, M., et al., The WEKA data mining software. Available from: <http://www.cs.waikato.ac.nz/ml/weka>.

1. Local search based hybrid genetic algorithm for feature selection [↑](#footnote-ref-1)
2. Hybrid ant colony optimization algorithm for feature selection [↑](#footnote-ref-2)
3. Particle swarm optimization for feature selection [↑](#footnote-ref-3)
4. Alon [↑](#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)