مطالعات علوم کاربردی در مهندسی

دوره ۹، شماره ۱، بهار ۱۴۰۲

صفحات ۲۰-۱

Online ISSN: ۲۴۷۶-۴۵۰۷

Print ISSN: ۲۵۳۸-۲۰۵۵

www.irijournals.com

بهبود الگوریتم ژنتیک با اضافه کردن توابع چند پارامتری در ساختار ژن ها

الهام السادات احمدی

کارشناسی ارشد نرم افزار کامپیوتر

چکیده

در سال‌های اخیر با توجه به افزایش ارتباط روزافزون بین ماشین و انسان، توجه زیادی به تکنولوژی قلم دیجیتال در موبایل-های note و تبلت‌ها شده است. افزایش استفاده از این تکنولوژی، نیاز به ایجاد صفحه کلیدهای مجازی برپایه‌ی قلم را در پی داشته است. علیرغم تلاش‌های انجام شده در زبان انگلیسی برای ایجاد این کیبورد مجازی، نقص چنین سیستم‌هایی در زبان-های فارسی و عربی آشکار است. هدف پژوهش حاضر، بهبود الگوریتم ژنتیک برای بهبود عملکرد یک الگوریتم تشخیص نویسه به عنوان مطالعه موردی است. مطالعه موردس استفاده شده بازشناسی برخط دست نوشته‌ی فارسی با کمک یک الگوریتم یادگیری سراسری مبتنی بر برنامه نویسی ژنتیک با تعریف ویژگی‌های متمایزکننده‌ی کاراکتر فارسی و عربی می‌باشد. بدنه‌ی اصلی حروف و نوع استروک‌های هر حرف توسط یک الگوریتم یادگیری سراسری مبتنی بر برنامه نویسی ژنتیک تشخیص داده می‌شود. پس از تشخیص بدنه‌ی اصلی حروف، با توجه به نوع استروک قرارگرفته بعد از بدنه‌ی اصلی و با توجه به مکان قرارگیری بخش استروک حرف، تشخیص نهایی حرف توسط یک DFA انجام می گیرد. الگوریتم پیشنهادی، حروف و اعداد مجزای فارسی را با ۹۷,۵۲٪ و دنباله‌ای از حروف و اعداد پیوسته‌ی فارسی را با ۹۲,۴۳٪ تشخیص می‌دهد.

واژه‌های کلیدی: قلم دیجیتال، برنامه نویسی ژنتیک، نمایش ژن، بهبود الگوریتم، ماشین آتاماتای متناهی

مقدمه

به طور کلی مشکلات برنامه نویسی ژنتیک را می توان بصورت زیربیان کنیم. فضای بسیار وسیع مساله (تعداد زیاد توابع، متغیرها، ثابت ها ، عملگرها و ... و ترکیب های مختلف آنها) وجود دارد. تعداد کروموزومها در هر نسل(محاسبه شایستگی تمام عناصر زمان بر است)زیاد است. زمان زیاد برای تست اینکه جواب یک کروموزوم (برنامه) به ازای ورودی های متفاوت (که در مواردی بسیار زیاد است)، چقدر به جواب واقعی نزدیک است (محاسبه شایستگی کروموزوم) نیاز است.الگوریتم ژنتیک استفاده شده برایبافتن جواب بایستی تعداد نسلهای بسیار زیادی کار کند. برایبک مساله ساده ممکن است نیاز باشد که الگوریتم ژنتیک چند صد هزار نسل کار کند. با در نظر گرفتن موارد فوق می بینیم که برنامه نویسی ژنتیک برای پیدا کردن جواب های مناسب، نیاز به کامپیوترهای بسیار سریع دارد.

در این مقاله قصد داریم ساختار داده و الگوریتم هایی برای تعریف ، استفاده از توابع چند پارامتری تعریف کنیم تا ضمن کاهش پیچیدگی ژن ها توان محاسباتی و قدرت تمایز آنها را افزایش دهیم. لذا عملیاتی مانند فاصله اقلیدسی در الگوریتم الگوریتم برنامه نویسی ژنتیک جدید تنها در یک گره صورت میگیرد.

الگوریتم پیشنهادی را روی پایگاه داده های مختلف تصویر اعمال کرده و کارایی آن را مورد بررسی قرار خواهیم داد.

فرضیه های پژوهش

۱- روش ماتریسی نمایش ژن و محاسبه fitness مبتنی بر آن برای ژن‌های الگوریتم برنامه نویسی ژنتیک می تواند منجر به بهبود سرعت همگرایی الگوریتم شود

۲- می‌توان ایجاد ژن‌ها را به گونه ای بازنویسی کرد که در کلاسترهای مختلف امکان ایجاد ژن تکراری نباشد. بدین وسیله فضای مساله به چند زیر مساله مجزا افراز میشود که قابلیت اجرا روی کلاسترهای مختلف را دارا خواهد شد.

۳- در هر کلاستر با cache کردن جواب‌ها می‌توان سرعت محاسبات را پائین آورد. این امر میتواند با اصلاح نحوه نمایش ژن‌ها انجام شود.

الگوریتم برنامه نویسی ژنتیک

با الهام گرفتن از نظریه انتخاب طبیعی داروین که مبتنی برعمل تولید مثل موجودات واصل"بقای مناسبترین" که گونه های زیست شناختی را قادر می سازد با شرایط محیطی خود را وفق دهند می باشد پروفسور جان هالند از دانشگاه میشیگان " ا لگوریتم های ژنتیک " را برای رشته های دودوئی با طول ثا بت را پایه گذاری کرد. (" وفق پذیری در طبیعت و سیستمهای مصنوعی ۱۹۷۵) در این مقاله "هالند" نشان داد مسائل زیادی در سیستمهای وفقی این قابلیت را دارند که به صورت وا ژه های ژنتیک بیان شوند و توسط الگوریتم های ژنتیک که روند تکاملی داروین را شبیه سازی می کنند به صورت موازی حل شوند. کار در این زمینه توسط افراد مختلفی دنبال شد تا ا ینکه " جان کوزا " در سال ۱۹۹۲ مفهوم " برنامه نویسی ژنتیک " را معرفی کرد که در این روش عناصر برنامه جایگزین رشته های دودوئی میشوند.

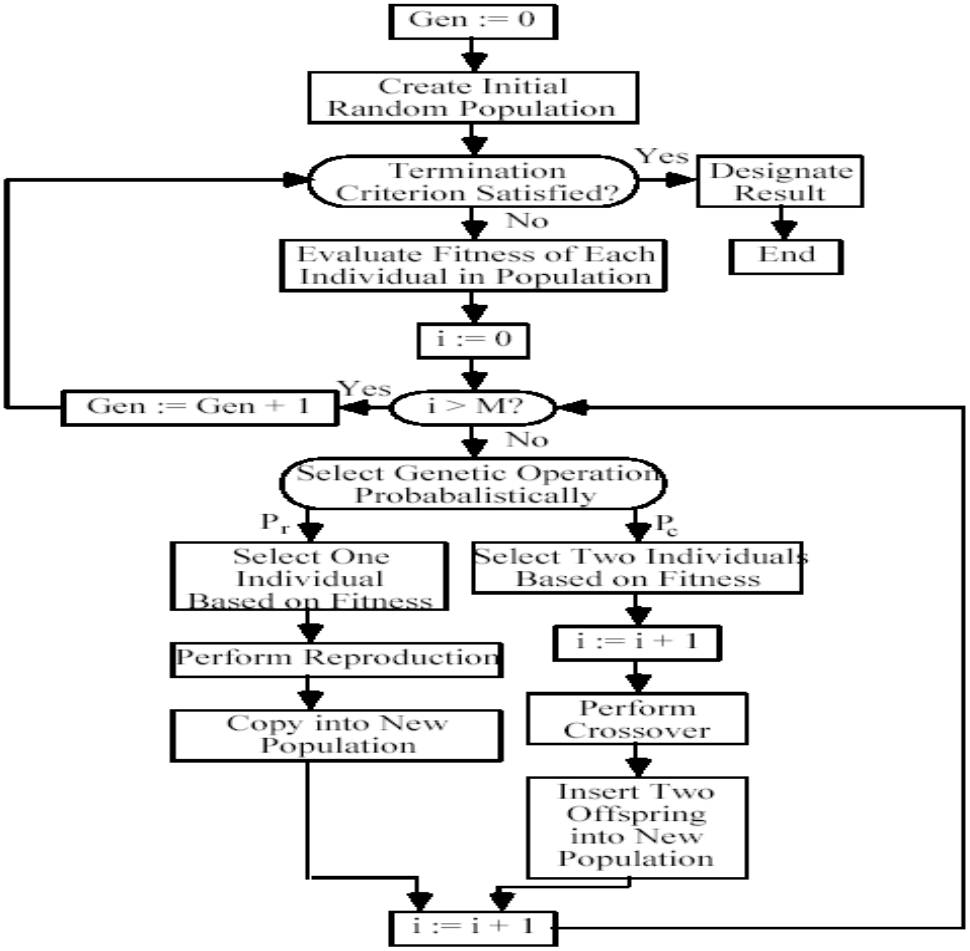
بیان الگوریتم برنامه نویسی ژنتیک

فرآیند برنامه نویسی ژنتیک توسط هر زبان برنامه نویسی که قابلیت بیان و ارزیابی ترکیب ا ز توا بع را داشته باشد قابل پیاده سازی ا ست هر چند که زبان لیسپ بیشترین قا بلیت را دارد. زبان لیسپ تنها از دو نوع نهاد تحت عنوان اتم ولیست تشکیل شده است. اتمها می توانند مقادیر ثابت مانند 7 یا مقادیر متغیر مانند TIME باشند. لیستها در زبان لیسپ مجموعه ای مرتب از اجزا درون پرانتز می باشد مانند (2 1 +). کامپایلر و سیستم عامل زبان لیسپ به گونه ای کار می کند که هر عبارتی را می بیند آنرا ارزیابی می کند. اتمهای ثابت به خود مقادیر ارزیابی می شوند در حالیکه اتمهای متغیر به مقادیر جاریشان ارزیابی می شوند. هنگامی که لیستی توسط زبان لیسپ مشاهده می شود ، اولین عنصر درون لیست به عنوان تابع و سایر عناصر بعنوان آرگومانهای تابع فرض می شوند.

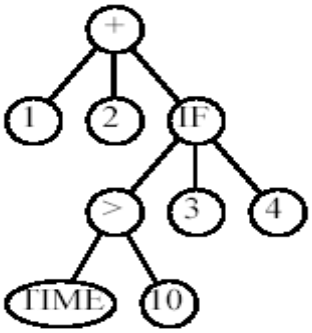
به عنوان مثا ل عبارت ((3 4 (IF (> TIME 10) 1 2 +) را در نظر بگیرید. در زیر عبارت (TIME > 10) تابع > بر روی اتم متغیر TIME و اتم ثابت 10 اعما ل می شود. سپس این زیر عبارت به مقدار درست یا نادرست

دوره ۹، شماره ۱، بهار ۱۴۰۲، صفحات ۲۰-۱

با توجه مقدار جاری اتم متغیر TIME ارزیابی می شود .مقدار منطقی ای که توسط زیر عبارت (TIME 10 <) برگشت داده می شود اولین آرگومان تابع IF فرض می شود. تابع IF تابعی با سه آرگومان می باشد. ا ین تابع مقدار ارزیابی شده دومین آرگومان را در صورتی که اولین آرگومان آن صحیح باشد باز می گرداند در غیر ا ین صورت مقدار ارزیابی شده آرگومان سوم برگشت داده می شود. هر عبارت زبان لیسپ را می توا ن به صورت ساختار درختی نمایش داد. درخت متناظر با عبارت بیان شده درشکل ۱ نشان داده شده ا ست.



شکل۱-: فلوچارت برنامه نویسی ژنتیک



شکل۲ : نمودار درختی یک عبارت لیسپ

دقت شود که این فرم درختی عبارت لیسپ معادل درخت پارسری می باشد که سایر کامپایلرها برای ارائه یک برنامه کامپیوتری می سازند.

الگوریتم برنامه نویسی ژنتیک

وفق پذیری شامل یک سری تغییرات در ساختار برنامه می باشد به گونه ای که در محیط خود بهتر عمل کند. یادگیری گونه ای از وفق پذیری می باشد که در آن هدف حل یک مسئله است. در ادامه به بررسی موارد زیر خواهیم پرداخت.

- ساختارهایی که عمل وفق پذیری را انجام می دهند.
- ساختارهای اولیه
- مقدار تناسب که ساختارها را مورد ارزیابی قرار می دهد.
- عملیاتی که جهت تغییر ساختارها استفاده می شود.
- انتخاب پاسخ
- شرط خاتمه
- پارامترهای کنترل

تناسب

به هر عضو در یک نسل مقداری عددی که نشان دهنده تراکنش آن با محیطش می باشد نسبت داده می شود. تناسب الهام گرفته شده از انتخاب داروینی از طبیعت است. در این بخش به معرفی چهار نوع به نامهای تناسب خام ، استاندارد ، تنظیم شده و نرمال شده خواهیم پرداخت.

تناسب خام اندازه گیری از تناسب است که در ذات خود مسئله نهفته است و این بستگی به مسئله دارد. برای خیلی از مسائل می تواند مجموع فواصل (فرضا خطا) انتخاب شود که از رابطه زیر بدست می آید.

$$r(i,t)=\sum_{j=1}^{N_c}\left|S(i,j)-C(j)\right|$$

که در آن $S(i,j)$ مقدار بازگشتی از عضو i ام برای مورد تناسب j ام و C (j) مقدار صحیح مورد تناسب j ام می باشد و t نشاندهنده زمان می باشد.

برای سایر مسائل تناسب خام می تواند زمان اندازه گیری شده ، سود و... باشد. بنابراین تناسب خام برای مسائلی که تناسب خطا است هر چه کمتر باشد بهتر است و برای مسائلی که تناسب سود باشد هر چه بیشتر باشد بهتر می باشد.

تناسب استاندارد ، تناسب خام را به گونه ای مطرح می کند که مقادیر کمتر بهتر باشند. اگر کمتربودن مقدار تناسب خام نشاندهنده بهتر بودن باشد این دو مقدار با یکدیگر برابرند در غیر این صورت از رابطه زیر بدست می آید.

$$s(i,t)=r_{\max}-r(i,t)$$

که $r_{\max}$ نشاندهنده مقدار تناسب خام ماکزیمم می باشد.

تناسب تنظیم شده از رابطه زیر بدست می آید:

$$a(i,t)=1/(1+s(i,t))$$

مقدار تناسب تنظیم شده بین صفر و یک می باشد.بر خلاف تناسب استاندارد برای مقادیر بیشتر بهتر می باشد. اگر مقدار $r_{\max}$ مشخص نباشد این بخش می تواند حذف شده و مقدار تناسب تنظیم شده مستقیما از تناسب خام بدست آید.

تناسب نرمال شده از رابطه زیر بدست می آید:

دوره ۹، شماره ۱، بهار ۱۴۰۲، صفحات ۲۰-۱

$$n(i,t) = a(i,t) \Big/ \sum_{k=1}^{M} a(k,t)$$

تناسب نرمال شده دارای خواص زیر است:

بین صفر و یک قرار دارد.

برای عضوهای بهتر این مقدار بهتر است.

مجموع مقادیر تناسبهای نرمال شده یک می باشد.

عملیاتی جهت تغییر ساختارها

عملیات اولیه ای که جهت تغییر ساختارها استفاده می شوند به شرح زیر می باشند:

عمل خود تولید بر اساس نسبت تناسب داروینی

عمل تولید مثل

انتخاب پاسخ

بهترین عضو جمعیت در هنگام خاتمه برنامه بعنوان جواب انتخاب می شود و دلیلی ندارد که این انتخاب بهترین جواب ممکن باشد.

شرط خاتمه

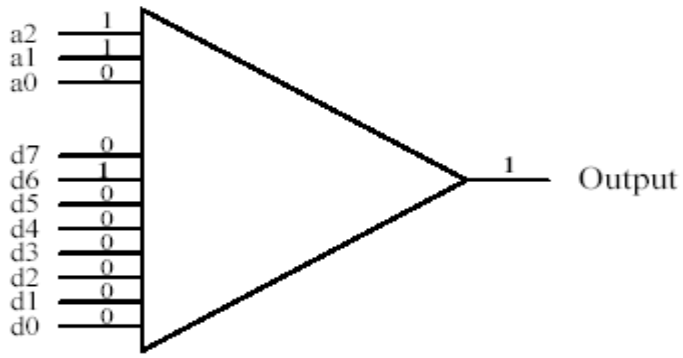
یک برنامه ژنتیک هنگامی خاتمه می یابد که یا تا یک تعداد نسل مشخص برنامه اجرا شده باشد و یا تناسب استاندارد مسئله به صفر ویا همسایگی صفر رسیده باشد.

پارامترهای کنترل

دو پارامتر مهم عبارتند از اندازه جمعیت وتعداد نسلی که برنامه باید اجرا شود.پارامتر بعدی انتخاب درصدی از جمعیت برای عمل تولید مثل می باشد. فرضا ۹۰ ٪ از جمعیت از عمل تولید مثل و ۱۰ ٪ باقیمانده از عمل خود تولید داروینی انتخاب می شود.پارامتر بعدی می تواند انتخاب فرضا ۹۰ ٪ از نودهای داخلی جهت عمل تولید مثل و ۱۰ ٪ از نودهای برگ انتخاب شود. با این عمل ساختارهای بزرگتری با یکدیگر ترکیب می شوند. سایر پارامترها می تواند انتخاب محدوده عمق برای نسل اولیه ونسلهای بعدی باشد.

مالتی پلکسر-۱۱ بولی

این مسئله یکی از مسائل کلاسیک در زمینه برنامه نویسی ژنتیک می‌باشد. مالتی پلکسر- N شامل k خط آدرس a_i و ۲ خط داده d_i می باشد که $N = k + 2$. مقدار تابع مالتی پلکسربولی مقدار بولی صفر یا یک خط دادهای می باشد که توسط k خط آدرس انتخاب و به خروجی منتقل می شود. شکل ۳-یک مالتی پلکسر-۱۱ را با ورودی ۱۱۰۰۱۰۰۰۰۰۰ وخروجی ۱ نشان می دهد.



شکل ۳-: مالتی پلکسر ۱۱ تائی

دوره ۹، شماره ۱، بهار ۱۴۰۲، صفحات ۲۰-۱

برای حل یک مسئله با برنامه نویسی ژنتیک پنج گام اصلی باید طی شود که بشرح زیر می باشد.

۱. مجموعه پایانه ها
۲. مجموعه توابع
۳. تابع تناسب
۴. پارامترهای راه انداز برنامه
۵. شرط خاتمه برنامه

اولین گام مهم در حل برنامه نویسی ژنتیک انتخاب پایانه ها می باشد. در این مسئله این مجموعه برابر با خطـوط آدرس و داده خواهد بود. البته در ابتدای مسئله تفاوتی بین خطوط آدرس وداده قائل نمی شویم.

$$T = \{ A0 , A1 , A2 , D0 , D1 , ... , D7 \}$$

دومین گام مهم انتخاب مجموعه توابع مناسب جهت حل مسئله می باشد. برای این مسئله این مجموعه بصورت زیر می با شد:

$$F = \{ AND , OR , NOT , IF \}$$

که به ترتیب دارای ۱،۲،۲ و۳ آرگومان می باشد. تابع IF بکار رفته قبلا مورد بررسی قرار گرفته بود. این مجموعه توابع هـر دو شرط بسته بودن وکافی بودن را در بر می گیرد.

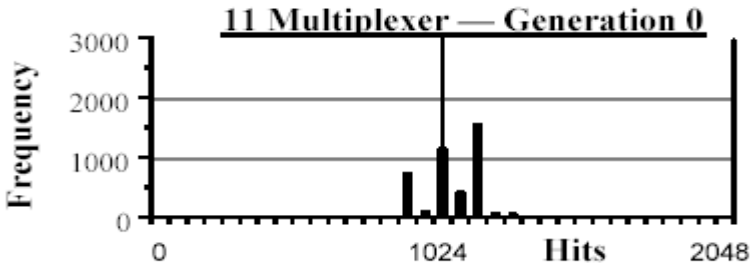
فضای جستجوی مسئله تمام عباراتی است که می تواند از ترکیب مجموعه توابع وپایانه ها بدست آید .

سومین گام مهم انتخاب تابع تناسب می باشد. همانطور که $2^{11} = 2048$ ترکیب مختلـف از آرگومانهـا ی مسـئله بدسـت می آید. بنابراین این عدد می تواند تخمینی مناسب جهت تناسب خام باشد یعنی هر تعداد جـواب صـحیحی کـه توسـط یـک عبارت پیدا شود بعنوان تناسب خام ارزیابی شده که این مقدار می تواند بین ۰ و ۲۰۴۸ باشد. عبارتی که تمامی ۲۰۴۸ مورد را صحیح ارزیابی کند جواب مسئله است. تناسب استاندارد این مسئله برابر با تفاضل تعداد صحیح بدست آمده از تعداد مـاکزیمم حالات می باشد.

چهارمین گام مهم انتخاب پارامترهای برنامه می باشد. فرضا می توان تعداد اعضاء نسل اول را ۴۰۰۰ در نظر گرفت.

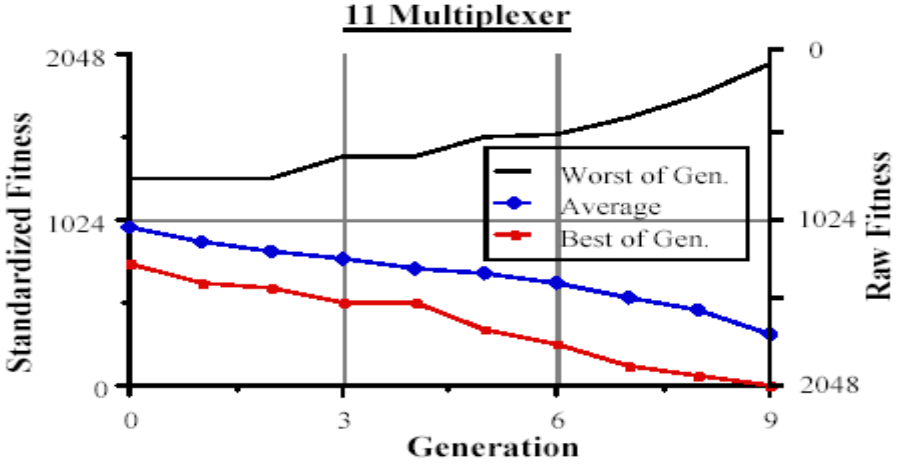
پنجمین گام مهم برنامه انتخاب شرط خاتمه می باشد که در این مسئله رسیدن به تناسب خام ۲۰۴۸ و یا اجرای برنامه تـا یک تعداد نسل می با شد.عمل پردازش با ایجاد تصادفی نسل صفر آغاز شد. اکثر عضو های ایجاد شده در این نسل دارای مقدار تناسب پایینی بودند. بعضی از آنها مقادیر ثابتی بودند مانند ((NOT A0) AND A0) در برخـی دیگـر مقدار ورودی با خروجی یکسان بود مانند ((NOT (NOT A1)) برخی از اعضاء ناکارا بودند ماننـد (OR D7 D7) و در برخی دیگر جای آرگومانهای داده وآدرس با هم جابجا شده بود مانند (IF D0 A0 A2) . هر چنـد کـه حتـی در چنین جمعیتی برخی از اعضاء نسبت به مابقی دارای تناسب بالاتری بودند. در واقـع ۲۳ عضـو از ۴۰۰۰ عضـو نسـل اولیـه دارای تناسب خام ۱۲۸۰ بودند. (IF A0 D1 D2) یکی از این موارد می باشد. تناسـب اسـتاندارد متوسـط نسـل اول ۹۸۵ بود .

هیستوگرام برخورد یکی از روشهای سودمند جهت نشان دادن اندازه برخورد می باشـد. (منظـور از انـدازه برخـورد تعـداد موارد صحیح تشخیص داده می باشد.) محور افقی این نمودار تعداد برخوردها را نشان می دهد ومحـور عمـودی نشـان دهنـده تعداد اعضایی می باشد که این میزان برخورد را داشته اند. در روی محور افقی بسـته هـای ۶۴ تـائی از مقـدار تناسـب در نظـر گرفته شده است. شکل ۵-۲ نمودار برخورد در نسل صفر را نشان می دهد.



شکل ۴-: نمودار هیستوگرام اولین نسل

سپس با استفاده از قوانین نسبت تناسب داروین و عمل تولید مثل نسلهای بعدی را بوجود می آوریم. با تکرار این عمل مشاهده شد که نهایتا پس از ۹ نسل تکرار به تناسب استاندارد صفر دست پیدا کردیم. شکل ۵-۳ تناسب استاندارد را برای بهترین ، بدترین ومیانگین اعضاء نشان می دهد.

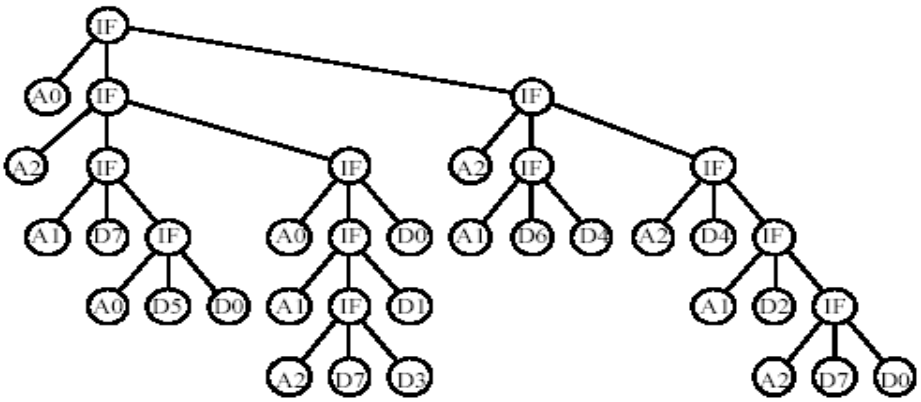


شکل ۵-: مقایسه نسل های مختلف

عضوی که در نسل ۹ تمامی حالات را درست ارزیابی کرد بصورت می باشد:

(IF A0 (IF A2 (IF A1 D7 (A0 D5 D0))
(IF A0 (IF A1 (IF A2 D7 D3) D1) D0))
(IF A2 (IF A1 D6 D4)
(IF A2 D4 (IF A1 D2 (IF A2 D7 D0)))))

شکل ۶ درخت حاصله را نشان می دهد.



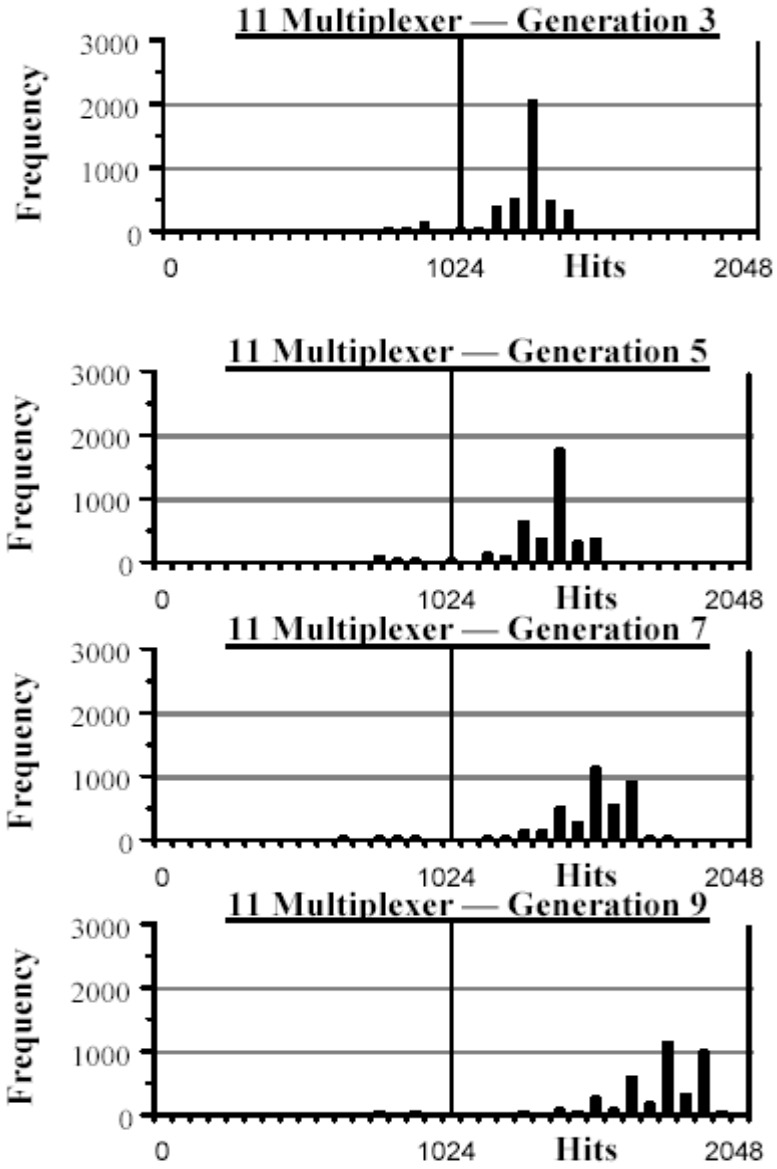
شکل ۶-: درخت جواب مسئله

دوره ۹، شماره ۱، بهار ۱۴۰۲، صفحات ۲۰-۱

اگر عبارت لیسپ حاصله را ساده نماییم به عبارت زیر دست پیدا می کنیم:

(IF A0 (IF A2 (IF A1 D7 D5) (IF A1 D3 D1))
(IF A2 (IF A1 D6 D4) (IF A1 D2 D0)))

که نشان می دهد این عبارت بخوبی قادر است کلیه حالات را بدرستی ارزیابی کند. شکل ۵-۵ نمودار هیستوگرام مسئله را برای نسلهای ۷،۵،۳ و ۹ نشان می دهد.



شکل -۷ : نمودار هیستوگرام برای نسلهای مختلف

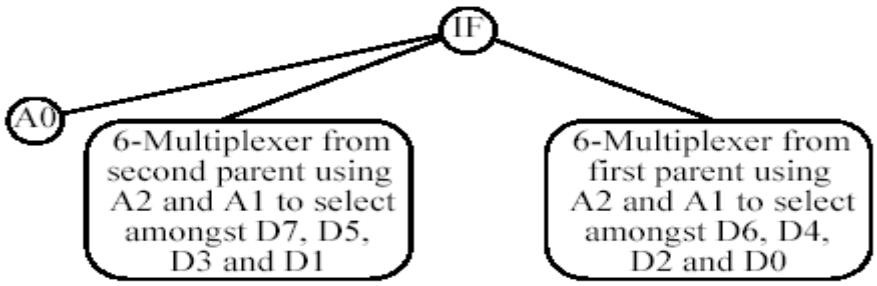
حال بد نیست نگاهی به والدینی که بهترین حالت را بوجود آورده اند نگاهی داشته باشیم.اولین والد ۱۷۹۲ مورد را بدرستی ارزیابی می کرد.

(IF A0 (IF A2 D7 D3)
(IF A2 (IF A1 D6 D4)
(IF A2 D4 (IF A1 D2 (IF A2 D7 D0))))))

اگر A0 صفر باشد این عبارت بخوبی تمامی حالات باقیمانده را درست ارزیابی می کند. که در واقع خطـوط داده زوج را کـاملا درست ارزیابی کرده است. والد دوم ۱۹۲۰ را بدرستی ارزیابی می کرد و در واقع خطوط داده فرد را بخوبی پوشش می داد.

(IF A2 (IF A1 D7 (IF A0 D5 D0))
(IF A0 (IF A1 (IF A2 D7 D3) D1) D0))

شکل ۸ بخشهای به ارث گرفته شده درخت نهایی را نشان می دهد.



شکل ۸-: بخشهای به ارث گرفته شده از والدین

دنباله های استقرائی

در این قسمت ساختارهای مورد استفاده برای حل یک دنباله استقرائی مورد بررسی قرار می گیرد. دنباله زیر را در نظر بگیرید:
, ...۱،۱۵،۱۲۹،۵۴۷،۱۵۹۳،۳۷۱۱،۷۴۶۵،۱۳۵۳۹

ما بدنبال تابعی می باشیم که این دنباله را برای ما ایجاد نماید.مجموعه پایانه ها را می توان بصورت زیر تعریف کرد:

$T = \{ J , 0 , 1 , 2 , 3 \}$

که در آن T متغیر دنباله و اعداد صحیح هم جهت بدست آوردن مقادیر صحیح دنباله استفاده می شود. مجموعـه توابـع هـم بصورت زیر تعریف می شود:

$F = \{ + , - , * \}$

تا بع تناسب در این مسئله برای یک تعداد خاصی از دنباله فرضا ۲۰ بکار گرفته می شود. تابع تناسب خام بکار رفته برابر با مجموع قدر مطلق تفاضل بین اعداد بدست آمده از تابع بدست آمده و جواب نهایی می باشـد. پـس از ۴۳ نسـل تـابع حاصـله بدست آمد که پس از ساده شدن بصورت زیر درآمد:

$$1+2j+3j^2+4j^3+5j^4$$

فرمهای مفهومی

درختهای تصمیم گیری را می توان به راحتی توسط برنامه نویسـی ژنتیـک حـل کرد.مثـال معـروف لـذت از ورزش را در نظـر بگیرید. با انتخاب مجموعه پایانه ها برابر با $\{۰،۱\}$ که نشاندهنده لذت بردن یا نبردن از ورزش می باشـد و مجموعـه توابـع بـا خصیصه ها می توان مسئله را حل کرد. تعداد آرگومانهای هر تابع برابر تعداد مقادیری است که هر خصیصه مـی توانـد بگیـرد. تابع تناسب هم برابر تعداد مثال آموزشی می باشد که به خوبی پوشش داده شود.

سایر عملگرها

در این بخش به معرفی سایر عملگرهایی که در برنامه نویسی ژنتیک استفاده می شود می پردازیم.

پارامترهای اجرا

جدول ۱- پیکربندی اجرای سیستم و نتایج اجرای سیستم را نشان می دهد را نشان می‌دهد.

جدول۱: پیکربندی و نتایج اجرا

پارامترهای سیستم	مقادیر
Config file	برنامه نویسی ژنتیک موازی شده
Fitness function	رگرسیون
Generational user function	ندارد
Data set	مجموعه کاراکترهای فارسی
Parallel mode	فعال
Parallel workers	۱۰
Fitness cache	Enabled
Start time	07-Feb-2016 04:07:58
Stop Serial time	07-Feb-2016 04:28:36
Stop Parallel time	07-Feb-2016 04:15:22
Merged independent runs	۳
Run Serial timeout (sec)	۲۱
Run Parallel timeout (sec)	۷

همانطور که از جدول فوق مشخص است، الگوریتم در روش جدید و بکارگیری ژن های جدید روی یک خوشه متشکل از ۱۰ پایانه اجرا شده است. زمان اجرای الگوریتم بعد از ۳ بار اجرا به طور میانگین ۷ دقیقه شده است. به علاوه دقت الگوریتم که مطلوب اولیه تحقیق است بالا رفته است.

جدول ۲: پارامترهای الگوریتم برنامه نویسی ژنتیک

اندازه جمعیت	۲۵۰
ماکزیمم نسل	۱۵۰
گزارش دهی در هر چند نسل	۸
متغیر های ورودی	به اندازه ویژگی ها
تعداد نمونه های آموزش	۷۰۰
اندازه تورنمنت	۲۵
درصد ژن های نخبه در هر اجرا	۰/۷
مرتب سازی ژن ها بر اساس الفبا	بله
Probability of pareto tournament	۰/۷
ماکزیمم ژن ها	۶
ماکزیمم عمق درخت ژن ها	۴
ماکزیمم تعداد گره ها	بی نهایت
احتمال وجود عدد ثابت در درخت ژن	0.1 Integer 0.5
احتمال ترکیب	0.84

	High level 0.2, Low level 0.8
احتمال جهش	0.14 Subtree 0.9, Input 0.05, Perturb ERC 0.05
معیار چیپیدگی	عبارت محاسبه
مجموعه توابع	TIMES MINUS PLUS TANH MULT3 ADD3

با استفاده از پارامترهای فوق نتایج بدست آمده به خوبی توانست مدل زیر را محاسبه کند:

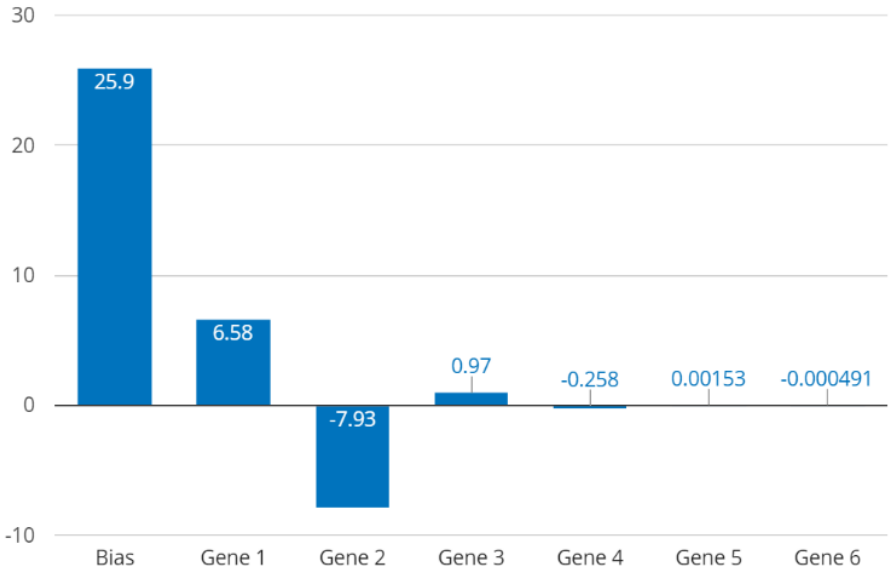
$$y = 6.58\ x2 - 1.35\ x1 - 0.258\ x4 - 0.97\ \tanh(x1 - 1.0\ x3) - 4.91e\text{-}4\ \tanh(x3) - 4.91e\text{-}4\ x32\ x4 - 4.91e\text{-}4\ x2\ (2.0\ x3 + x4) + 0.00153\ x1\ x3\ x4 + 25.9$$

از ایستگاه های مختلف بهترین ژن‌های بدست آمده در جدول زیر نشان داده شده است:

جدول۳: ژن های بدست آمده از اجراهای مختلف

ژن	مقدار
بایاس	25.9
ژن ۱	6.58 x1 + 6.58 x2
ژن ۲	-7.93 x1
ژن ۳	-0.97 tanh(x1 - 1.0 x3)
ژن ۴	-0.258 x4
ژن ۵	0.00153 x1 x3 x4
ژن ۶	- 4.91e-4 tanh(x3) - 4.91e-4 x32 x4 - 4.91e-4 x2 (2.0 x3 + x4)

ارزش هر ایستگاه در شکل یک نشان داده شده است. ارزش هر ایستگاه برازندگی ژن محاسبه شده توسط آن ایستگاه است.



شکل-۹: ارزش ایستگاه های مختلف در محاسبه ژن‌ها

کارایی روش ارایه شده

کارایی روش ارایه شده در محاسبه داده‌های آموزش و تست در جداول زیر آمده است. در جداول زیر مخفف ها عبارتند از:

- R^2 - goodness of fit (coefficient of determination).
- RMSE - root mean squared error.
- MAE - mean absolute error.
- SSE - sum of squared errors.
- MSE - mean squared error.

جدول ۴: معیارهای ارزیابی برای داده های آموزشی

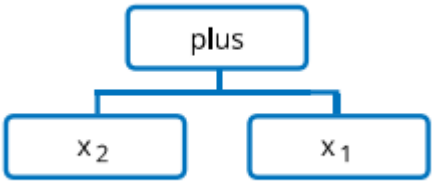
Metric	Value
R^2	0.99211
RMSE	0.2587
MAE	0.20563
SSE	46.8473
Max. abs. error	0.71151
MSE	0.066925

جدول ۵: معیارهای ارزیابی برای داده های تست

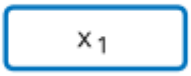
Metric	Value
R^2	0.99039
RMSE	0.29024
MAE	0.22923
SSE	25.1877
Max. abs. error	0.71283
MSE	0.08424

خروجی سیستم‌های مختلف

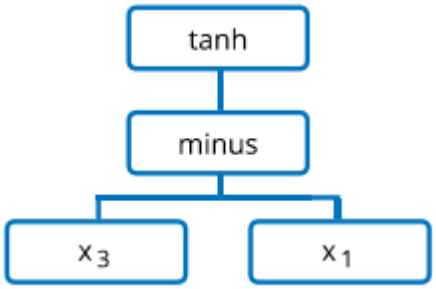
ژن های محاسبه شده بر اساس روش نمایش جدید در هر یک از سیستم های تحت تست در شکل های زیر آمده است. بـرگ-های هر درخت ویژگی‌ها را نشان می‌دهد و گره‌ها عملگرهای محاسباتی را. همانطور که مشخص شده است، هر ایستگاه مـوازی پیچیدگی متفاوتی ارایه داده است.



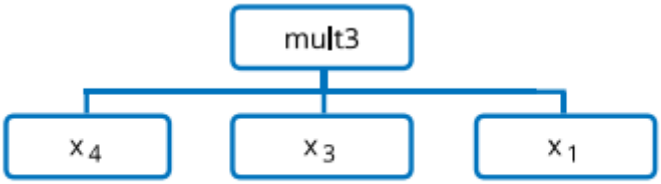
شکل ۱۰: ژن ایستگاه شماره ۱



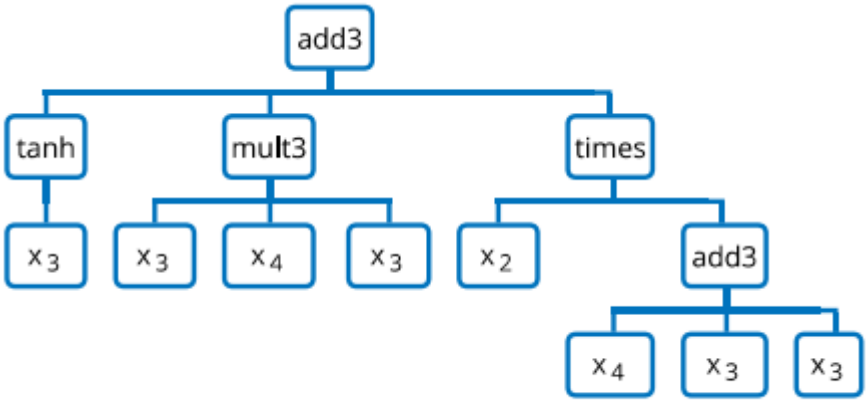
شکل ۱۱: ژن ایستگاه شماره ۲



شکل ۱۲: ژن ایستگاه شماره ۴



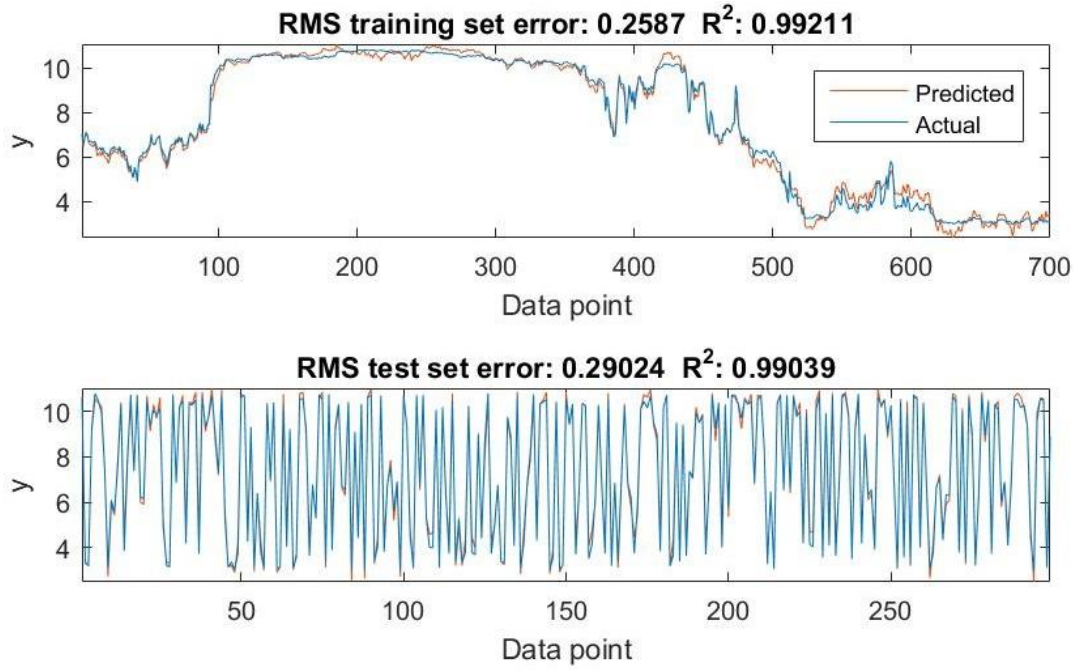
شکل ۱۳: ژن ایستگاه شماره ۵



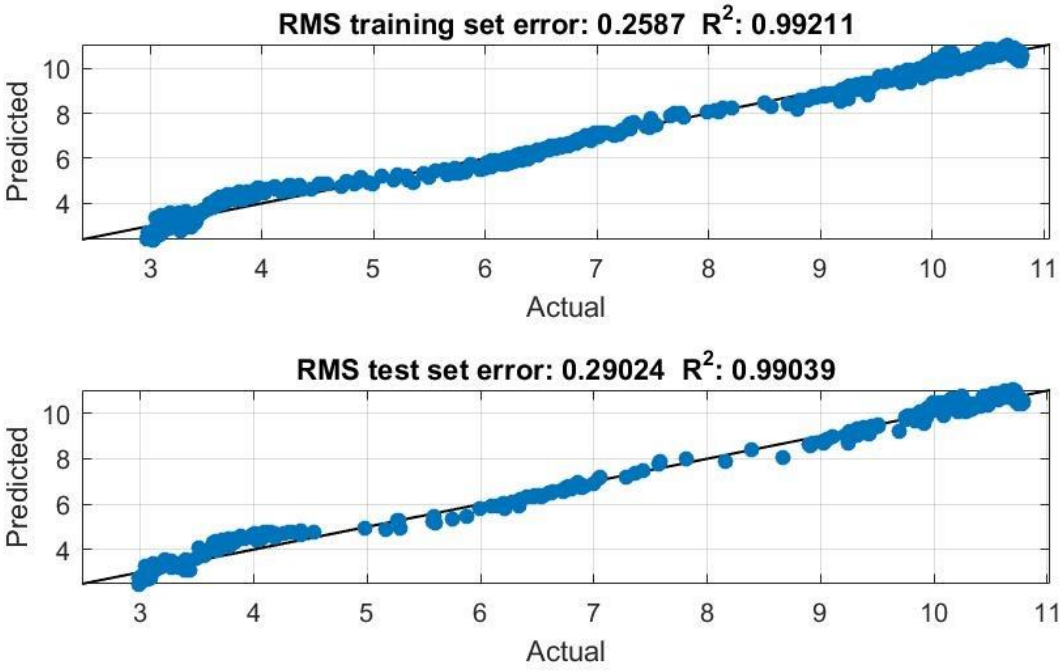
شکل ۱۴: ژن ایستگاه شماره ۶

ارزیابی کارایی و همگرایی سیستم

در اجرای الگوریتم های مکاشفه ای همواره باید همگرایی سیستم چک شود. شکل‌های این بخش همگرایی سیستم های را برای تشخیص درست دست نوشته ها نشان می‌دهد.



شکل ۱۵: خطای rms برای داده‌های تست و آموزش در طول اجرای برنامه حاصل از میانگین اجراها



شکل ۱۶: خطای رگرسیون rms برای داده‌های تست و آموزش در طول اجرای برنامه حاصل از میانگین گیری شده

جدول زیر پیچیدگی هر یک از اجراها را در محاسبه مدل نهایی به همراه مدل نشان می‌دهد. این جدول نشان می‌دهد با وجـود بهینه سازی هیچ یک از معیارهای کارایی افت نکرده اند.

جدول -۶: پیچیدگی کدل هر سرور به همراه مدل به دست آمده از آن و شاخص R2

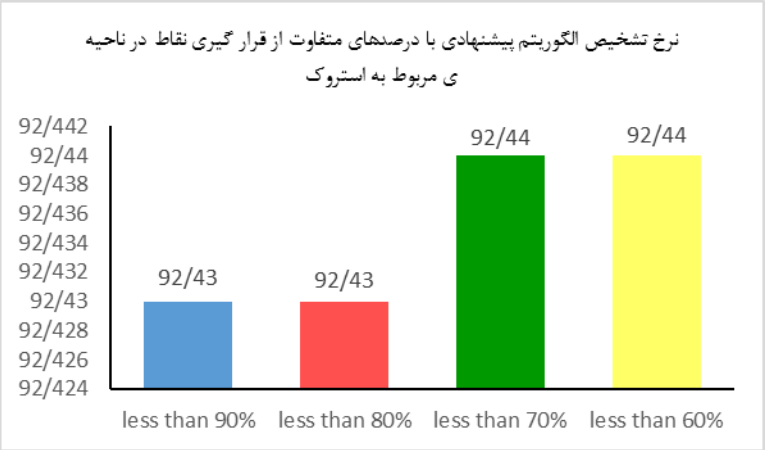
Model ID	Goodness of fit (R^2)	Model complexity	Model
463	0.98	16	0.00431 x1 x32 - 0.148 x3 - 1.24 x1 - 0.00121 x32 x4 + 22.6
507	0.987	24	4.93 x2 - 1.14 x1 + 0.536 x4 - 1.24 tanh(x1 - 1.0 x3) - 5.61e-4 x1 x32 + 8.61
511	0.935	3	13.5 x2 - 1.96 x1 + 0.334 x4 + 21.6
546	0.97	11	12.1 x2 + 0.177 x3 - 1.59 tanh(x1 - 1.0 x3) - 1.61
562	0.983	20	0.163 x3 - 0.707 x1 + 0.65 x4 - 1.19 tanh(x1 - 1.0 x3) - 6.85e-4 x1 x3 x4 + 1.38
564	0.992	48	6.57 x2 - 1.35 x1 - 0.258 x4 - 0.97 tanh(x1 - 1.0 x3) - 4.91e-4 tanh(x3) - 4.91e-4 x2 x4 - 4.91e-4 x32 x4 + 0.00153 x1 x3 x4 + 25.9
568	0.991	28	5.41 x2 - 4.2 x1 - 1.11 x4 - 1.16 tanh(x1 - 1.0 x3) + 0.0983 x1 x4 - 1.08e-4 x3 x42 + 59.4
576	0.992	60	6.58 x2 - 1.35 x1 - 0.258 x4 - 0.97 tanh(x1 - 1.0 x3) - 4.91e-4 tanh(x3) - 4.91e-4 x32 x4 - 4.91e-4 x2 (2.0 x3 + x4) + 0.00153 x1 x3 x4 + 25.9
591	0.992	42	6.55 x2 - 1.35 x1 - 4.91e-4 x3 - 0.257 x4 - 0.969 tanh(x1 - 1.0 x3) - 4.91e-4 tanh(x3) - 4.91e-4 x32 x4 + 0.00153 x1 x3 x4 + 25.9

نتیجه گیری

هدف پژوهش حاضر ارائه‌ی سیستم جدیدی بر پایه‌ی نمایش جدید ژن ها در الگوریتم برنامه نویسی ژنتیک جهت بهبود عملکرد آنها بودد. ویژگی هایی مستخرج از سیگنال های حرکتی جهت تشخیص برخط دست نوشته‌ی پیوسته‌ی فارسی ورودی اصلی الگوریتم بوده است. روال تشخیص در سیستم پیشنهادی با استفاده از آتاماتای متناهی انجام شده است که از ترکیب تشخیص بدنه‌ی اصلی و نوع استروک براساس یک الگوریتم یادگیری مبتنی بر ژنتیک و بررسی موقعیت مکانی delayed strokeها تشکیل شده است. در این کار ما داده ها را بر روی سیستم های موازی پخش کرده و بصورت موازی هر سیستم یک بخش فضای جستجو را حل کرده است. نتایج آزمایشات مختلف نشان دهنده ی سومند بودن روش تعریف شده و روش پیشنهادی جهت بهبود تشخیص حروف و اعداد فارسی می‌باشد.

همانگونه که از نتایج آزمایشات مشخص است، برای تابع ارزیابی الگوریتم یادگیری سراسری، استفاده از طبقه‌بندهای KNN و DT نسبت به SVM سرعت بالاتری از الگوریتم پیشنهادی را به دست می‌دهد. شاید بتوان علت این موضوع را در این دانست که برخلاف SVM که برای طبقه‌بندی مدل می‌سازد، طبقه‌بند DT در ساختار خود یک feature selection دارد و منجر به حذف ویژگی‌های غیرمتمایزکننده می‌شود.

بیشترین خطا در سیستم پیشنهادی به دلیل قرار گرفتن تعداد کمتر از ۹۰٪ ژن های مشترک در ناحیه‌ی تعیین شده است. خطای حاصل از این موضوع را می‌توان با کاهش درصد مربوط به تعداد نقاط قرارگیری ژن های تکراری در ناحیه‌ی مربوطه بهبود بخشید. سیستم پیشنهادی با خطای قرار گرفتن تعداد کمتر از ۸۰٪ ، ۷۰٪ و ۶۰٪ از نقاط قرارگیری ژن های تکراری در ناحیه‌ی مذکور تست شده است. نتیجه در شکل ۱-۵ آورده شده است. همانگونه که از نمودار مشخص است، با در نظر گرفتن سرعت حاصل از قرار گرفتن تعداد کمتر از ۷۰٪ و ۶۰٪ ، می‌توان ۰٫۰۱٪ درصد خطای حاصل را بهبود بخشید.



شکل۹: نرخ تشخیص الگوریتم پیشنهادی با درصدهای متفاوت از قرار گیری نقاط در ناحیه ی تکراری ژن ها

منابع:

1. Aburas, A.A. and S.M. Rehiel, Off-line Omni-style handwriting Arabic character recognition system based on wavelet compression. Arab Research Institute in Sciences & Engineering, 2007. 3(4): p. 123-135.
2. AlKhateeb, J.H., et al. Knowledge-based baseline detection and optimal thresholding for words segmentation in efficient pre-processing of handwritten Arabic text. in Information Technology: New Generations. ITNG 2008. Fifth International Conference on. IEEE.
3. Al-Rashaideh, H., Preprocessing phase for Arabic word handwritten recognition. information transmisssion in computer networks, 2006. 6(1).
4. Ghods, V., E. Kabir, and F. Razzazi, Effect of delayed strokes on the recognition of online Farsi handwriting. Pattern Recognition Letters, 2013. 34(5): p. 486-491.
5. Alabau, V., A. Sanchis, and F. Casacuberta, Improving on-line handwritten recognition in interactive machine translation. Pattern Recognition, 2014. 47(3): p. 1217-1228.
6. Das, N., et al., Handwritten Bangla character recognition using a soft computing paradigm embedded in two pass approach. Pattern Recognition, 2015. 48(6): p. 2054-2071.
7. Mozaffari, S., et al., Lexicon reduction using dots for off-line Farsi/Arabic handwritten word recognition. Pattern Recognition Letters, 2008. 29(6): p. 724-734.
8. Al Abodi, J. and X. Li, An effective approach to offline Arabic handwriting recognition. Computers & Electrical Engineering, 2014. 40(6): p. 1883-1901.
9. Kamble, P.M. and R.S. Hegadi, Handwritten Marathi character recognition using R-HOG Feature. Procedia Computer Science, 2015. 45: p. 266-274.
10. Parvez, M.T. and S.A. Mahmoud, Arabic handwriting recognition using structural and syntactic pattern attributes. Pattern Recognition, 2013. 46(1): p. 141-154.
11. Elarian, Y., et al., An Arabic handwriting synthesis system. Pattern Recognition, 2015. 48(3): p. 849-861.
12. Parodi, M. and J.C. Gómez, Legendre polynomials based feature extraction for online signature verification. Consistency analysis of feature combinations. Pattern Recognition, 2014. 47(1): p. 128-140.
13. Fierrez, J., et al., HMM-based on-line signature verification: Feature extraction and signature modeling. Pattern recognition letters, 2007. 28(16): p. 2325-2334.
14. Kholmatov, A. and B. Yanikoglu, Identity authentication using improved online signature verification method. Pattern recognition letters, 2005. 26(15): p. 2400-2408.
15. Rashidi, S., A. Fallah, and F. Towhidkhah, Feature extraction based DCT on dynamic signature verification. Scientia Iranica, 2012. 19(6): p. 1810-1819.
16. Narima, Z., R. Messaoud, and B. Mouldi. Neuro-Markovian hybrid system for handwritten Arabic word recognition. in Electronics, Circuits and Systems. ICECS 2003. Proceedings of the 2003 10th IEEE International Conference on. IEEE.
17. Razzak, M.I., et al., HMM and fuzzy logic: a hybrid approach for online Urdu script-based languages' character recognition. Knowledge-Based Systems, 2010. 23(8): p. 914-923.
18. Natarajan, P., et al., Multi-lingual offline handwriting recognition using hidden Markov models: A script-independent approach, in Arabic and Chinese Handwriting Recognition. 2008, Springer. p. 231-250.
19. Lee, H. and B. Verma, Binary segmentation algorithm for English cursive handwriting recognition. Pattern Recognition, 2012. 45(4): p. ۱۳۰۶-۱۳۱۷.
20. Zhang, D., et al., Chinese comments sentiment classification based on word2vec and SVM perf. Expert Systems with Applications, 2015. 42(4): p. 1857-1863.
21. Robertson, J. and R. Guest, A feature based comparison of pen and swipe based signature characteristics. Human movement science, 2015. 43: p. 169-182.
22. Mohamed, A., et al., Baseline extraction algorithm for online signature recognition. WSEAS Transactions on Systems, 2009. 8(4): p. 491-500.

23. Fahmy, M.M., Online handwritten signature verification system based on DWT features extraction and neural network classification. Ain Shams Engineering Journal, 2010. 1(1): p. 59-70.
24. Lee, J., et al., Using geometric extrema for segment-to-segment characteristics comparison in online signature verification. Pattern Recognition, 2004. 37(1): p. 93-103.
25. Guru, D. and H. Prakash, Online signature verification and recognition: An approach based on symbolic representation. IEEE transactions on pattern analysis and machine intelligence, 2009. ۳۱(٦): p. 1059-1073.
26. Ghods, V. and E. Kabir. Feature extraction for online Farsi characters. in 2010 12th International Conference on Frontiers in Handwriting Recognition. 2010.
27. Ziaratban, M., K. Faez, and F. Allahveiradi, Novel Statistical Description for the Structure of Isolated Farsi/Arabic Handwritten Characters. ICFHR, Canada, 2008.
28. Ahmed, H. and S.A. Azeem. On-line Arabic handwriting recognition system based on HMM. in 2011 International Conference on Document Analysis and Recognition. IEEE.
29. Abdelazeem, S. and H.M. Eraqi. On-line Arabic handwritten personal names recognition system based on HMM. in 2011 International Conference on Document Analysis and Recognition. IEEE.
30. Ghods, V., E. Kabir, and F. Razzazi, Decision fusion of horizontal and vertical trajectories for recognition of online Farsi subwords. Engineering Applications of Artificial Intelligence, 2013. 26(1): p. 544-550.
.31 Baghshah, M.S., S.B. Shouraki, and S. Kasaei. A novel fuzzy classifier using fuzzy LVQ to recognize online persian handwriting. in 2006 2nd International Conference on Information & Communication Technologies. 2006. IEEE.
32. Baghshah, M.S. A novel fuzzy approach to recognition of online Persian handwriting. in 5th International Conference on Intelligent Systems Design and Applications (ISDA'05). 2005. IEEE.
33. Potrus, M.Y., U.K. Ngah, and B.S. Ahmed, An evolutionary harmony search algorithm with dominant point detection for recognition-based segmentation of online Arabic text recognition. Ain Shams Engineering Journal, 2014. 5(4): p. 1129-1139.
34. De Stefano, C., A. Della Cioppa, and A. Marcelli, Character preclassification based on genetic programming. Pattern Recognition Letters, 2002. 23(12): p. 1439-1448.
35. De Stefano, C., et al., A GA-based feature selection approach with an application to handwritten character recognition. Pattern Recognition Letters, 2014. 35: p. 130-141.
36. Cha, S.-H., C.C. Tappert, and S.N. Srihari. Optimizing Binary Feature Vector Similarity Measure using Genetic Algorithm and Handwritten Character Recognition. in ICDAR. 2003. Citeseer.
37. Bui, T., et al. Development of algorithms for face and character recognition based on wavelet transforms, PCA and neural networks. in Control and Communications (SIBCON), 2015 International Siberian Conference on. IEEE.
38. Sun, Y., S. Wei, and J. Chen. SMT product character recognition based on Principal Component analysis. in Electronic Packaging Technology (ICEPT), 2015 16th International Conference on. IEEE.
39. Shao, L., L. Liu, and X. Li, Feature learning for image classification via multiobjective genetic programming. IEEE Transactions on Neural Networks and Learning Systems, 2014. 25(7): p. 1359-1371.
40. Parkins, A. and A.K. Nandi, Genetic programming techniques for hand written digit recognition. Signal Processing, 2004. 84(12): p. 2345-2365.
41. Searson, D.P., GPTIPS 2: an open-source software platform for symbolic data mining, in Handbook of genetic programming applications. 2015, Springer. p. ۵۵۱-۵۷۳.
42. Razavi, S. and E. Kabir. A dataset for online Farsi handwriting. in 6th national conference on intelligent systems (in Farsi). 2004.

43. Kholmatov, A. and B. Yanikoglu, SUSIG: an on-line signature database, associated protocols and benchmark results. Pattern Analysis and Applications, 2009. 12(3): p. 227-236.

44. Preece, S.J., et al., A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. IEEE Transactions on Biomedical Engineering, ۲۰۰۹. ٥٦(۳): p. 871-879.

.45 Kherallah, M., F. Bouri, and A.M. Alimi, On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm. Engineering Applications of Artificial Intelligence, 2009. 22(1): p. 153-170.

46. Boubaker, H., et al., Online Arabic databases and applications, in Guide to OCR for Arabic Scripts. 2012, Springer. p. 541-557.

47. Eraqi, H.M. and S.A. Azeem. An on-line arabic handwriting recognition system: Based on a new on-line graphemes segmentation technique. in 2011 International Conference on Document Analysis and Recognition. IEEE.

48. Liwicki, M. and H. Bunke. IAM-OnDB-an on-line English sentence database acquired from handwritten text on a whiteboard. in Eighth International Conference on Document Analysis and Recognition (ICDAR'05). 2005. IEEE.

49. Liwicki, M. and H. Bunke. HMM-based on-line recognition of handwritten whiteboard notes. in Tenth international workshop on frontiers in handwriting recognition. 2006. Suvisoft.

مطالعات علوم کاربردی در مهندسی

دوره ۹، شماره ۱، بهار ۱۴۰۲، صفحات ۲۰-۱

Improving Genetic programming Algorithms considering multi-parameter functions in gene formation

Abstract

In recent years, due to the ever-increasing communication between machines and humans, much attention has been paid to digital pen technology in mobile phones and tablets.

The increase in the use of this technology has led to the need to create virtual keyboards based on the pen.

Despite the efforts made in the English language to create this virtual keyboard, the defects of such systems in Persian and Arabic languages are obvious.

The aim of the current research is to improve the genetic algorithm for the performance of a letter recognition algorithm as a case study.

The study of the methods used is the recognition of Persian handwriting with the help of a global learning algorithm based on genetic programming with the definition of distinguishing features of Persian and Arabic characters. The main body of letters and the type of strokes of each letter are recognized by a global learning algorithm based on genetic programming. After identifying the main body of the letters, according to the type of stroke placed after the main body and according to the location of the stroke part of the letter, the final detection of the letter is done by a DFA. The proposed algorithm recognizes individual Persian letters and numbers with 97.52% and a sequence of consecutive Persian letters and numbers with 92.43%.

Keyword: digital pen, genetic programming, gene display, algorithm improvement, finite automata machine