ارائه یک روش جدید انتخاب ورودی بر اساس دسته‌بندی نیمه نظارتی برای تخمین سری‌های زمانی

راضیه محمدی[۱،*] و فرشید کی نیا[۲]

اطلاعات مقاله

دریافت مقاله: ۱۳۹۷/۰۷/۲۱

پذیرش مقاله: ۱۳۹۸/۰۴/۰۵

واژگان کلیدی:

انتخاب ویژگی،

یادگیری ماشین،

الگوریتم فرا ابتکاری،

سری زمانی.

چکیده

انتخاب ویژگی نقش مهمی در بهبود عملکرد الگوریتم‌های یادگیری ماشین ایفا می‌کند. بر این اساس در این تحقیق با اشاره به برخی روش‌های یادگیری ماشین و انتخاب ویژگی باهدف ارائه یک روش جدید انتخاب ورودی بر اساس دسته‌بندی نیمه نظارتی برای تخمین سری‌های زمانی، یک دیدگاه کلی بر قابلیت‌ها و موارد کاربرد این روش‌ها ارائه گردید. در این تحقیق ۱۲ رویکرد در ایجاد یک طبقه‌بندی کننده بهینه مبتنی بر ماشین بردار پشتیبان (SVM) و شبکه‌های عصبی (MLP و RBF) بر مبنای الگوریتم ژنتیک (GA)، فاخته (Cuckoo) و ازدحام ذرات (PSO) ارائه گردید. در این راستا سعی شده سیستمی طراحی شود که منجر به کاهش هزینه در جمع‌آوری داده‌ها شود. برای این منظور در تحقیق حاضر از سه مجموعه داده با قابلیت سری زمانی از داده‌های استاندارد UCI، استفاده گردید. نتایج حاصل از رویکردهای استفاده شده در این تحقیق بیانگر عملکرد خوب تمامی الگوریتم‌های استفاده شده دارد. با این‌حال، توانایی و عملکرد هر کدام از رویکردها با توجه به نوع و ماهیت داده‌ها متفاوت است. همین امر باعث شده است که گاهاً رویکرد شبکه عصبی MLP و الگوریتم GA یا Cuckoo نتایج بهتری داشته باشند و در برخی موارد نیز رویکرد ماشین بردار پشتیبان با الگوریتم PSO نتایج بهتری ارائه می‌دهد. با توجه به نتایج حاصل می‌توان بیان کرد که استفاده از انتخاب ویژگی بر اساس دسته‌بندی نیمه نظارتی باعث کاهش خطای سیستم، افزایش دقت و افزایش سرعت تخمین سری‌های زمانی می‌گردد. از این‌رو با استفاده از طبقه‌بندی کننده‌ی کارا و قدرتمند شبکه عصبی MLP و ماشین بردار پشتیبان در کنار الگوریتم بهینه‌سازی و فرا ابتکاری، می‌توان یک سیستم طبقه‌بندی ترکیبی بهینه برای تخمین سری‌های زمانی طراحی نمود.

۱-مقدمه

داده‌های سری زمانی مجموعه‌ای از مشاهدات است که به صورت پیوسته با زمان و مقادیر گذشته آن‌ها برای پیش-بینی سری‌های زمانی استفاده می‌شوند [۱]. از این‌رو انتخاب متغیرهای ورودی[۲] یک گام مهم در توسعه مدل‌های پیش‌بینی سری زمانی است. در این فصل ابتدا در مورد مسئله پژوهشی، اهداف و اهمیت تحقیق صحبت خواهد شد و مشکلات موجود در این زمینه به صورت مختصر بیان

* پست الکترونیک نویسنده مسئول: mohammadiraziyeh@yahoo.com

۱. کارشناسی ارشد پژوهشکده انرژی، پژوهشگاه علوم و تکنولوژی پیشرفته و علوم محیطی، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته، کرمان، ایران

۲. دانشیار پژوهشکده انرژی، پژوهشگاه علوم و تکنولوژی پیشرفته و علوم محیطی، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته، کرمان، ایران

[2] Input Variables

خواهند شد. در انتها نیز چارچوب انجام تحقیق ارائه خواهد شد.

پیشرفت‌های به وجود آمده در جمع‌آوری داده سری زمانی و قابلیت‌های ذخیره‌سازی در طی دهه‌های اخیر باعث شده در بسیاری از علوم با حجم بزرگی از اطلاعات روبرو شویم. در حال حاضر با رشد فناوری و تراکم انبوه داده‌های سری زمانی، جمع‌آوری داده‌های مناسب سری زمانی توسط نیروی انسانی زمان‌بر و پرهزینه و همراه با محدودیت کاربری است. محققان در زمینه‌های مختلف هر روز با حجم مشاهدات بیشتری روبرو می‌شوند. از طرف دیگر در مقایسه با بسترهای داده‌های قدیمی و کوچک‌تر، بسترهای داده‌ای امروزی چالش‌های جدیدی در تحلیل داده‌های سری زمانی به وجود آورده‌اند. روش‌های آماری سنتی به دو دلیل، امروزه کارایی خود را از دست داده‌اند. علت اول افزایش تعداد مشاهدات است و علت دوم که از اهمیت بالاتری برخوردار است افزایش تعداد متغیرهای مربوط به یک مشاهده است. همچنین داده‌های خام معمولاً دچار مشکلاتی مانند نویز، بایاس، تغییرات شدید در بازه دینامیکی و نمونه‌برداری هستند و استفاده از آن‌ها به فرم اولیه و بدون اعمال پیش‌پردازش موجب تضعیف طراحی-های بعدی خواهد شد. تمرکز مسئله انتخاب ویژگی[1] بر روی یافتن یک زیرمجموعه از ویژگی‌های ارزنده بر اساس ارزش آن‌ها در طبقه‌بندی داده‌های سری زمانی است که این مهم با حذف ویژگی‌های افزونه، بی‌ارزش و نویزی صورت می‌گیرد. در حقیقت این عمل منجر به بهبود عملکرد یادگیری، کاهش پیچیدگی محاسباتی، کاهش فضای ذخیره‌سازی و ساخت مدلی بهتر با قابلیت تعمیم می‌شود. از طرف دیگر به علت وقت‌گیر بودن طبقه‌بندی داده‌های سری زمانی به صورت دستی از طبقه‌بندی‌های نیمه نظارتی، بدون دانش و اطلاع پیشین در مورد داده‌های سری زمانی جهت طبقه‌بندی استفاده می‌شود.

تاکنون تحقیقاتی در زمینه انتخاب ویژگی صورت گرفته است. یک روش انتخاب ویژگی بر اساس کلونی مورچگان ارائه شده است. انواع روش‌های انتخاب ویژگی اعم از روش‌های پایه فیلتر و بسته‌بندی را مورد بررسی قرار گرفت [۲]. با استفاده از الگوریتم ژنتیک و ترکیب آن با شبکه‌ی عصبی مصنوعی بهترین ورودی‌ها را از بین ورودی‌های کاندید را تعیین کردند [۳]. دو روش انتخاب ویژگی مبتنی بر مدل شبکه عصبی مصنوعی و الگوریتم ژنتیک مورد مقایسه قرار گرفته است [۴]. در تحقیقی از الگوریتم جستجوی فاخته درزمینه انتخاب ویژگی استفاده کردند [۵]. همچنین یک چارچوب ارزیابی عملکرد را که شامل یک معیار کمی و چند معیار کیفی برای تکنیک‌های انتخاب متغیر ورودی بود ارائه شده است [۶].

نوآوری این تحقیق استفاده از انتخاب ویژگی بر اساس دسته‌بندی نیمه نظارتی است. فرضیه اصلی این تحقیق این است که استفاده از دسته‌بندی نیمه نظارتی باعث کاهش خطای سیستم، افزایش دقت و افزایش سرعت تخمین سری‌های زمانی می‌گردد.

۲-جمع‌آوری داده‌ها

جمع‌آوری و ترکیب داده‌ها از چندین منبع به درون یک محل منسجم به نام انبار داده‌ها، یکی از وظایف اولیه در فرایند داده‌کاوی محسوب می‌شود. این منابع می‌تواند شامل چند پایگاه داده‌ی ناهمگن باشد. افزونگی و در نتیجه ناسازگاری داده‌ها از مسائل مهمی هستند که در جمع‌آوری داده‌ها باید به آن توجه کرد. صفات خاصه‌ی اطلاعاتی که از منابع مختلف جمع‌آوری می‌شوند را می‌توان توسط داده-های مِتا[2] که معمولاً هر پایگاه داده یا انبار داده شامل آن است، تشخیص داد و بدین ترتیب از افزونگی جلوگیری نمود [۷].

نگهداری و یا حذف صفات خاصه‌ای که مقادیر آن‌ها از مقدار دیگر صفات مشتق می‌شوند، به دلیل استفاده از فرایند داده‌کاوی[3] برمی‌گردد. چنانچه به دست آوردن این صفات مشتق شده از پیچیدگی زمانی بالایی برخوردار باشد، بهتر است که این صفات نگهداری شوند. در غیر اینصورت با تشخیص مناسب حذف می‌شوند. با محاسبه‌ی همبستگی میان صفات خاصه می‌توان این نوع افزونگی را کنترل نمود. در هر حال تشخیص افزونگی بین صفات خاصه می‌تواند به تشخیص در سطح نمونه‌ها نیز منجر شود.

۲-۱-تغییر شکل داده‌ها

شکل مناسب داده‌ها به عنوان ورودی الگوریتم‌های داده‌کاوی نقش بسزایی در این فرایند بازی می‌کند و در

[1] Datat Mining

[2] Features Selection
[3] Metadata

مرحله‌ی آماده‌سازی داده‌ها این نقش پررنگ است.
استفاده از توابع تجمعی[1]، نرمال‌سازی[2]، خوشه‌بندی[3] و رگرسیون[4] روش‌های مرسومی هستند که برای تغیر و تبدیل شکل داده‌ها می‌توان از آن‌ها استفاده نمود. انتخاب روش مناسب به ماهیت داده‌ها، مقدار آن و اهداف داده-کاوی بستگی دارد. در جمع‌آوری داده‌ها و ساخت انبار داده‌ها از توابع تجمعی نظیر ماکزیمم، مینیمم، میانگین و مجموع استفاده می‌شود. این توابع علاوه بر تغییر شکل داده‌ها می‌توانند از حجم آن نیز بکاهند [۸].
روش‌های نرمال‌سازی معمولاً برای مواردی که محاسبه‌ی فواصل بین نمونه‌ها در داده‌کاوی مطرح است، مفید هستند. در این روش مقادیر اندازه‌گیری شده به محدوده‌ی جدیدی نگاشت می‌شوند. این محدوده‌ی جدید با توجه به الگوریتم‌های موجود بیشتر بین ۰ تا ۱ یا بین ۱ - تا ۱ انتخاب می‌شود .سه روش مرسوم برای نرمال‌سازی در ادامه توضیح داده شده است.
* روش اول به وسیله‌ی حرکت نقطه‌ی اعشار در داده‌ها، محدوده جدید را درست می‌کند. برای تبدیل مقدار یک داده در این روش فقط کافی است آن را بر عددی که توانی از ده است، تقسیم کنیم. این عدد با توجه به حداکثر قدر مطلق مقادیر موجود تعیین می‌شود [۹].
* با داشتن حداکثر و حداقل مقادیر موجود می‌توانیم مطابق با رابطه (۱) آن‌ها را به هر محدوده‌ی جدید دلخواهی نرمال‌سازی کنیم.

$$v = \frac{v - Min}{Max - Min} \qquad (۱)$$

در رابطه (۱) مقادیر به دست آمده دربازه بین ۰ و ۱ قرار می‌گیرند.
* نرمال‌سازی نوع سوم برای مواقعی مناسب است که در نمونه‌های صفت خاصه مقدار خارج از محدوده وجود داشته باشد و همین‌طور مقدار حداکثر و حداقل داده‌ها مشخص نیستند. این روش از میانگین و انحراف استاندارد صفات خاصه استفاده می‌کند. برای تبدیل مقدار V_{old} به مقدار V_{new} از رابطه (۲) استفاده می‌کنیم.

$$V_{new} = \frac{V_{old} - Mean}{St_Dev} \qquad (۲)$$

در رابطه (۲)، Mean میانگین مقادیر صفت خاصه‌ی موردنظر و St_Dev انحراف استاندارد همان مقادیر است [۱۰].

۲-۲- روش‌های کاهش ابعاد داده و انتخاب ویژگی

رویکردهای استفاده شده در تحقیق حاضر به منظور کاهش ابعاد داده‌ها و انتخاب بهینه ویژگی‌ها از ۱۲ رویکرد استفاده شده است که در جدول (۱) بیان شده‌اند [۱۱].

جدول ۱- انواع رویکردهای به کار گرفته‌شده در تحقیق حاضر

رویکردها		
RBF	MLP	SVM
RBF+GA	MLP+GA	SVM+GA
RBF+PSO	MLP+PSO	SVM+PSO
RBF+Cuckoo	MLP+Cuckoo	SVM+Cuckoo

۲-۳- ارزیابی مدل‌های انتخاب‌شده

برای سنجش انطباق یک پیش‌بینی با یک الگوی داده سری زمانی، از معیارهای سنجش خطای پیش‌بینی استفاده شده است، اگر Z و *Z نشانگر مقدار واقعی و پیش‌بینی شده متغیر در زمان t باشد خطای پیش‌بینی عبارت است از e=Z-Z* بنابراین برای یک دوره زمانی و برای n مقدار پیش‌بینی‌شده، معیارهای سنجش پیش‌بینی مطابق با روابط (۳، ۴ و ۵) عبارت‌اند از:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(e)^2 \qquad (۳)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e)^2} \qquad (۴)$$

$$\mathrm{R}^2 = 1 - \sum_{i=1}^{N}(\mathrm{Z} - \hat{\mathrm{Z}}) \Big/ \sum_{i=1}^{N}(e)^2 \qquad (۵)$$

از لحاظ دقت بین معیارهای فوق، ملاک برای مقایسه دقت مدل‌ها معیار RMSE است. RMSE یک شاخص کمّی مناسب است که از آن می‌توان در برآورد میزان دقت مدل استفاده کرد. ضریب تبیین (R^2) بیان‌گر میزان همبستگی بین مقادیر پیش‌بینی شده و داده‌های اندازه‌گیری شده است.
در تحقیق حاضر به‌منظور ارائه یک روش جدید انتخاب ورودی بر اساس دسته‌بندی نیمه نظارتی برای تخمین

[1] Cumulative functions
[2] Normalization
[3] Clustering
[4] Regression

سری‌های زمانی، از سه مجموعه داده باقابلیت سری زمانی از دادهای استاندارد UCI، استفاده گردید. مجموعه داده-های استفاده‌شده در این تحقیق شامل:

داده‌های کیفیت هوا که این مجموعه داده شامل ۱۵ ویژگی است که در تحقیق حاضر ویژگی‌های اول و دوم (ستون‌های زمان و تاریخ) را حذف و درنهایت از ۱۳ ویژگی آن استفاده شد. همچنین به علت زمان‌بر بودن پردازش و محاسبات در رویکردهای استفاده‌شده (جدول ۳-۱) در تحقیق حاضر و محدودیت سخت‌افزاری سیستم کامپیوتر قابل‌دسترس دانشجو، تعداد ۹۹ نمونه از کل نمونه‌های مجموعه داده کیفیت هوا استفاده شد (ماتریس ۱۳×۹۹).

۱- داده‌های بازار بورس استانبول

این مجموعه داده شامل ۸ ویژگی ورودی، یک ویژگی به‌عنوان هدف و تعداد ۵۳۶ نمونه است. در تحقیق حاضر تمامی نمونه‌ها را به خاطر در دست داشتن خروجی از یک مجموعه داده با حجم بالاتر استفاده نمودیم (ماتریس ۹×۵۳۶).

۲- مجموعه داده‌های SML ۲۰۱۰

این مجموعه داده شامل ۲۴ ویژگی و تعداد ۴۱۳۷ نمونه است. در تحقیق حاضر به علت حجم بالای این مجموعه داده و محدودیت سخت‌افزاری، درنهایت از ۱۸ ویژگی و تعداد ۱۷۰ نمونه این مجموعه داده استفاده شد (ماتریس ۱۸×۱۷۰).

۳- نتایج به دست آمده

۳-۱- نتایج پردازش مجموعه داده‌های کیفیت هوا

۳-۱-۱- نتایج به دست آمده از ماشین بردار پشتیبان و رویکردهای آن

همان‌طور که مشاهده می‌شود عملکرد رویکرد SVM+PSO و SVM+GA با توجه به مقادیر R2 و RMSE که هر دو به ترتیب برابر با ۰/۹۰ و ۰/۰۷ هستند نسبت به ماشین بردار پشتیبان و رویکردSVM+ Cuckoo بهتر بوده است. تعداد کاهش بعد در رویکرد SVM+PSO، ۳ می باشد این در صورتی است که کاهش بعد در رویکرد SVM+Cuckoo، ۴ است.

جدول ۲- نتایج به دست آمده برای SVM و رویکردهای آن بر روی داده‌های کیفیت هوا

data name: a part of AirQualityUCI.xlsx				
data size: ۹۹*۱۳				
State	SVM	SVM+GA	SVM+Pso	SVM+Cuckoo
Dim	۱۳	۱۰	۱۰	۹
R	۰.۹۶	۰.۹۷	۰.۹۷	۰.۹۶
R2	۰.۸۷	۰.۹۰	۰.۹۰	۰.۸۸
MSE	۰.۰۰۶۱	۰.۰۰۴۸	۰.۰۰۴۸	۰.۰۰۵۵
RMSE	۰.۰۷۸	۰.۰۷۰	۰.۰۷۰	۰.۰۷۴
feature selected by GA:		[۱ ۲ ۳ ۴ ۵ ۷ ۹ ۱۰ ۱۱ ۱۲]		
feature selected by Pso:		[۱ ۲ ۳ ۴ ۵ ۷ ۹ ۱۰ ۱۱ ۱۲]		
feature selected by cuckoo:		[۲ ۳ ۴ ۵ ۶ ۹ ۱۰ ۱۱ ۱۲]		

۳-۱-۲- نتایج به‌دست‌آمده از شبکه عصبی MLP و رویکردهای آن

همان‌طور که در جدول ۴ مشاهده می‌شود، عملکرد رویکرد MLP+Cuckoo با توجه به مقادیر R^2 و RMSE که این مقادیر به ترتیب برابر با ۰/۹۹ و ۰/۰۰۷۲ است، نسبت به شبکه عصبی MLP که برابر با ۰/۹۹ و ۰/۰۱۶ است و دیگر رویکردها از دقت بالاتری برخوردار است. اگرچه تعداد کاهش بعد در رویکردهای MLP+Cuckoo و MLP+GA برابر با ۸ است، ولی با این حال الگوریتم فاخته نسبت به الگوریتم ژنتیک عملکرد بهتری در ترکیب با شبکه عصبی MLP داشته است.

۳-۱-۳- نتایج به دست آمده از شبکه عصبی RBF و رویکردهای آن

با توجه به نتایج به دست آمده در جدول ۵، عملکرد تمامی رویکردها نسبت به شبکه عصبی RBF تفاوت چندانی ندارد و تنها نکته قابل توجه کاهش بعد در رویکردها ترکیبی بوده است که در همه‌ی آن‌ها یکسان و مشابه است.

جدول ۴- نتایج به دست آمده برای MLP و رویکردهای آن بر روی داده‌های کیفیت هوا

Report: MLP + (GA,Pso,Cuckoo)				
data name: a part of AirQualityUCI.xlsx				
data size: ۹۹*۱۳				
State	MLP	MLP+GA	MLP+Pso	MLP+Cuckoo
Dim	۱۳	۵	۷	۵
R	۰.۹۹	۰.۹۸	۰.۹۹	۰.۹۹
R2	۰.۹۹	۰.۹۶	۰.۹۹	۰.۹۹
MSE	۰.۰۰۰۲۷	۰.۰۰۱۹	۰.۰۰۰۲۵	۰/۰۰۰۰۵۲
RMSE	۰.۰۱۶	۰.۰۴۳	۰.۰۱۶	۰.۰۰۷۲
feature selected by GA:	[۳ ۵ ۷ ۱۱ ۱۲]			
feature selected by Pso:	[۲ ۵ ۷ ۸ ۱۰ ۱۱ ۱۲]			
feature selected by cuckoo:	[۲ ۴ ۷ ۱۱ ۱۲]			

جدول ۵- نتایج به دست آمده برای RBF و رویکردهای آن بر روی داده‌های کیفیت هوا

Report: RBF + (GA, Pso,Cuckoo)				
data name: a part of AirQualityUCI.xlsx				
data size: ۹۹*۱۳				
State	RBF	RBF+GA	RBF+Pso	RBF+Cuckoo
Dim	۱۳	۹	۹	۹
R	۰.۹۹	۰.۹۹	۰.۹۹	۰.۹۹
R2	۰.۹۹	۰.۹۹	۰.۹۹	۰.۹۹
MSE	۰.۰۰۰۳	۰.۰۰۲۹	۰.۰۰۰۲۹	۰.۰۰۰۲۹
RMSE	۰.۰۱۷	۰.۰۱۷	۰.۰۱۷	۰.۰۰۱۷
feature selected by GA:	[۱ ۲ ۴ ۵ ۶ ۸ ۹ ۱۱ ۱۲]			
feature selected by Pso:	[۱ ۲ ۴ ۵ ۶ ۸ ۹ ۱۱ ۱۲]			
feature selected by cuckoo:	[۱ ۲ ۴ ۵ ۶ ۸ ۹ ۱۱ ۱۲]			

جدول ۶- نتایج به دست آمده برای SVM و رویکردهای آن بر روی داده‌های بورس استانبول

Report: SVM + (GA, Pso,Cuckoo)				
data name: ISTANBUL STOCK EXCHANGE.xlsx				
data size: ۵۳۶*۹				
State	SVM	SVM+GA	SVM+Pso	SVM+Cuckoo
Dim	۹	۷	۶	۷
R	۰.۹۰	۰.۹۰	۰.۹۰	۰.۹۰
R2	۰.۸۰	۰.۸۰	۰.۷۹	۰.۸۰
MSE	۰.۰۰۱۸	۰.۰۰۱۸	۰.۰۰۱۹	۰.۰۰۱۸
RMSE	۰.۰۴۲	۰.۰۴۲	۰.۰۴۳	۰.۰۴۲
feature selected by GA:	[۱ ۲ ۳ ۵ ۶ ۷ ۸]			
feature selected by Pso:	[۱ ۲ ۳ ۶ ۷ ۸]			
feature selected by cuckoo:	[۱ ۲ ۳ ۵ ۶ ۷ ۸]			

۳-۲- نتایج پردازش مجموعه داده‌های بازار بورس استانبول

۳-۲-۱- نتایج به دست آمده از ماشین بردار پشتیبان و رویکردهای آن

در برآورد شاخص بازار بورس استانبول همان‌طور که در جدول ۶ مشاهده می‌گردد رویکردهای ماشین بردار پشتیبان با الگوریتم‌های ژنتیک، ازدحام ذرات و فاخته تغییر قابل‌توجهی ارائه نداده‌اند و از لحاظ کاهش بعد نیز چندان خوب عمل ننموده‌اند کمترین مقدار R^2 برابر با ۰/۷۹ در حالت SVM+PSO به دست آمده است. این در حالی است که، الگوریتم ازدحام ذرات در ترکیب با ماشین بردار پشتیبان با انتخاب ۶ ویژگی از نظر کاهش بعد موفق‌تر عمل کرده است هر چند که نتایج مقداری ضعیف‌تر به دست آمده است. نتایج و عملکرد دیگر رویکردها نیز مشابه هم است.

جدول ۷- نتایج به دست آمده برای MLP و رویکردهای آن بر روی داده‌های بورس استانبول

Report: MLP + (GA, Pso,Cuckoo)				
data name: ISTANBUL STOCK EXCHANGE.xlsx				
data size: ۵۳۶*۹				
State	MLP	MLP+GA	MLP+Pso	MLP+Cuckoo
Dim	۹	۷	۷	۷
R	۰.۹۰	۰.۸۶	۰.۸۸	۰.۸۹
R2	۰.۷۸	۰.۶۴	۰.۷۵	۰.۷۹
MSE	۰.۰۰۱۹	۰.۰۰۳۲	۰.۰۰۲۲	۰.۰۰۱۹
RMSE	۰.۰۴۴	۰.۰۵۶	۰.۰۴۷	۰.۰۴۴
feature selected by GA: [۱ ۲ ۳ ۵ ۶ ۷ ۸]				
feature selected by PSO: [۱ ۲ ۳ ۵ ۶ ۷ ۸]				
feature selected by cuckoo : [۱ ۲ ۳ ۵ ۶ ۷ ۸]				

جدول ۸: نتایج به دست آمده برای RBF و رویکردهای آن بر روی داده‌های بورس استانبول

Report: RBF + (GA, Pso,Cuckoo)				
data name: ISTANBUL STOCK EXCHANGE.xlsx				
data size: ۵۳۶*۹				
State	RBF	RBF+GA	RBF+Pso	RBF+Cuckoo
Dim	۹	۳	۳	۳
R	-۰.۲۵	۰.۲۲	۰.۱۶	۰.۱۶
R2	-۳۵	-۶۲۱.۷۰	-۹.۵۸	-۹.۵۸
MSE	۳۱	۵.۵۹	۰.۰۹۵	۰.۰۹۵
RMSE	۱۷.۰۶	۳.۳۶	۰.۳۱	۰.۳۱
feature selected by GA: [۶ ۷ ۸]				
feature selected by PSO: [۲ ۶ ۷]				
feature selected by Cuckoo: [۲ ۶ ۷]				

۳-۲-۲- نتایج به دست آمده از شبکه عصبی MLP و رویکردهای آن

با توجه به جدول ۷، در برآورد شاخص بازار بورس استانبول رویکردهای شبکه عصبی MLP نسبت به خود شبکه ازنظر نتایج نهایی عملکردی ضعیف داشته‌اند. در حالی که، با کاهش بعد ویژگی‌های ورودی در رویکردها مواجه هستیم. مقدار کاهش بعد در شبکه عصبی MLP برابر با ۴ و در ترکیب با الگوریتم‌های فرا ابتکاری برابر با ۶ به دست آمده

است. عملاً ترکیب شبکه عصبی MLP با الگوریتم فاخته بهترین عملکرد را داشته است چرا که مقدار خطا را نیز تا مقدار ۰/۰۴۴ کاهش داده است. مقدار RMSE به ترتیب برای الگوریتم ازدحام ذرات و الگوریتم ژنتیک در ترکیب با شبکه عصبی MLP برابر با ۰/۰۴۷ و ۰/۰۵۶ به دست آمده است.

۳-۲-۳- نتایج به دست آمده از شبکه عصبی RBF و رویکردهای آن

با توجه به نتایج به دست آمده برای رویکردهای شبکه عصبی با تابع پایه شعاعی(RBF) در برآورد شاخص بازار بورس استانبول (جدول ۸)، همان‌طور که مشاهده می‌شود، دقت الگوریتم‌های مورد استفاده بسیار پایین است. میزان کاهش بعد برای شبکه عصبی برابر با ۴ و در ترکیب با الگوریتم‌های ژنتیک، ازدحام ذرات و فاخته برابر با ۱۰ به دست آمده است. اگرچه الگوریتم‌های استفاده شده از نظر کاهش بعد بسیار خوب عمل نموده‌اند اما از دقت و اعتبار کافی برخوردار نیستند. بنابراین نمی‌توان بیان کرد که کاهش ابعاد در این حالت از اعتبار لازم کافی برخوردار است.

۳-۳- نتایج پردازش مجموعه داده‌های SML ۲۰۱۰

۳-۳-۱- نتایج به دست آمده از ماشین بردار پشتیبان و رویکردهای آن

جدول ۹- نتایج به‌دست‌آمده برای SVM و رویکردهای آن بر روی داده‌های SML ۲۰۱۰

Report: SVM + (GA, Pso,Cuckoo)				
data name part of SML۲۰۱۰ Data Set.xlsx				
data size: ۱۷۰*۱۸				
SVM+Cuckoo	SVM+Cuckoo	SVM+Cuckoo	SVM+Cuckoo	SVM+Cuckoo
۱۳	۱۳	۱۳	۱۳	۱۳
۰.۷۵	۰.۷۵	۰.۷۵	۰.۷۵	۰.۷۵
۰.۵۰	۰.۵۰	۰.۵۰	۰.۵۰	۰.۵۰
۰.۰۲	۰.۰۲	۰.۰۲	۰.۰۲	۰.۰۲
۰.۱	۰.۱	۰.۱	۰.۱	۰.۱
feature selected by GA: [۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۱۰ ۱۱ ۱۲ ۱۳ ۱۴ ۱۵ ۱۶ ۱۷]				
feature selected by Pso: [۱ ۳ ۴ ۷ ۸ ۹ ۱۰ ۱۱ ۱۲ ۱۳ ۱۴ ۱۵ ۱۶ ۱۷]				
feature selected by Cuckoo: [۱ ۲ ۴ ۵ ۶ ۷ ۸ ۱۰ ۱۱ ۱۲ ۱۳ ۱۴ ۱۵ ۱۶ ۱۷]				

در خصوص مدل‌سازی مجموعه داده‌های SML ۲۰۱۰ با استفاده از رویکردهای ذکرشده در جدول ۹، از ۱۸ ویژگی ورودی استفاده شده است. نتایج به دست آمده حاکی از آن هست که میزان کاهش بعد در الگوریتم SVM و در ترکیب با الگوریتم ژنتیک، الگوریتم ازدحام ذرات و الگوریتم فاخته به ترتیب برابر با، ۱، ۱، ۴، ۵ است. بیشترین کاهش بعد در ترکیب ماشین بردار پشتیبان در ترکیب با الگوریتم فاخته است. بنابراین، الگوریتم‌های ازدحام ذرات و فاخته به ترتیب اولویت، کاهش بعد بیشتری داشته‌اند و نتایج به نسبت بهتری در بر داشته‌اند. مقدار R^2 در الگوریتم فاخته برابر با ۰/۵ و مقدار RMSE برابر با ۰/۱ به دست آمده است. با توجه به مقادیر به دست آمده برای ضریب تبیین و خطای الگوریتم، نتایج حاصل از دقت کافی برخوردار نیست.

۳-۳-۲- نتایج به دست آمده از شبکه عصبی MLP و رویکردهای آن

در جدول ۱۰ رویکردهای شبکه عصبی MLP در رابطه با مدل‌سازی مجموعه داده‌های SML ۲۰۱۰ نتایج جالبی نشان داده‌اند (جدول ۱۰). با استفاده از شبکه عصبی MLP هیچ گونه کاهش بعدی مشاهده نشده است. ولی در ترکیب شبکه عصبی با الگوریتم ژنتیک، الگوریتم ازدحام ذرات و الگوریتم فاخته، میزان کاهش بعد به ترتیب ذکر شده برابر با ۷،۶ و ۱۰ بوده است.. بیشترین کاهش بعد مربوط به رویکرد MLP+Cuckoo است. در رتبه بعدی رویکرد MLP+GA قرار می‌گیرد. این در حالی است که در رویکرد MLP+GA مقادیر RMSE کمتر و R^2 بالاتری مشاهده شود.

جدول ۱۰- نتایج به دست آمده برای MLP و رویکردهای آن بر روی داده‌های SML ۲۰۱۰

Report: MLP + (GA, Pso,Cuckoo)				
data name: part of SML ۲۰۱۰ Data Set.xlsx				
data size: ۱۷۰*۱۸				
MLP+Cuckoo	MLP+Cuckoo	MLP+Cuckoo	MLP+Cuckoo	MLP+Cuckoo
۸	۸	۸	۸	۸
۰.۸۶	۰.۸۶	۰.۸۶	۰.۸۶	۰.۸۶
۰.۶۹	۰.۶۹	۰.۶۹	۰.۶۹	۰.۶۹
۰.۰۱۴	۰.۰۱۴	۰.۰۱۴	۰.۰۱۴	۰.۰۱۴
۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲	۰.۱۲
feature selected by GA: [۱ ۲ ۳ ۴ ۵ ۷ ۹ ۱۲ ۱۳ ۱۴ ۱۶ ۱۷]				
feature selected by Pso: [۱ ۲ ۳ ۴ ۵ ۷ ۹ ۱۲ ۱۳ ۱۴ ۱۶ ۱۷]				
feature selected by GA: [۲ ۳ ۵ ۶ ۱۰ ۱۳ ۱۴ ۱۱۶]				

جدول ۱۱- نتایج به دست آمده برای RBF و رویکردهای آن بر روی داده‌های SML۲۰۱۰

Report: RBF + (GA, Pso,Cuckoo)				
Data Set.xlsx۲۰۱۰data name: part of SML				
data size: ۱۷۰*۱۸				
State	RBF	RBF+GA	RBF+Pso	RBF+Cuckoo
Dim	۱۸	۱۴	۱۵	۱۶
R	۰.۸۸	۰.۶۲	۰.۷۹	۰.۹
R2	۰.۶۱	-۲.۸۲	۰.۳۳	۰.۷۰
MSE	۰.۰۱	۰.۱۹	۰.۰۳	۰.۰۱
RMSE	۰.۱	۰.۴	۰.۱۸	۰.۱۱
feature selected by [۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۱۰ ۱۲ ۱۳ ۱۵ ۱۷]				
feature selected by Pso: [۱ ۲ ۳ ۴ ۵ ۷ ۸ ۹ ۱۰ ۱۲ ۱۳ ۱۴ ۱۵ ۱۶ ۱۷]				
feature selected by Cuckoo:: [۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ ۱۰ ۱۱ ۱۲ ۱۳ ۱۵ ۱۶ ۱۷]				

۳-۳-۳- نتایج به دست آمده از شبکه عصبی RBF و رویکردهای آن

با توجه به نتایج به دست آمده در جدول ۱۱ عملکرد کلی رویکردهای شبکه عصبی RBF چندان مناسب نبوده است به نحوی که در ترکیب با الگوریتم ژنتیک نتایج مناسبی ارائه نشده است. اگرچه بیشترین میزان کاهش بعد در ترکیب شبکه عصبی RBF با الگوریتم ژنتیک اتفاق افتاده است ولی مقدار ضریب تبیین مدل حاکی از دقت پایین آن است. مقدار ضریب تبیین مدل در الگوریتم RBF+Cuckoo نسبت به دیگر الگوریتم‌ها بیشتر است که نشان دهنده دقت بیشتر در این حالت در مقابل کاهش بعد کمتر آن است.

۴- نتیجه‌گیری

در تحقیق حاضر جهت انتخاب ورودی مناسب جهت تخمین سری‌های زمانی از تکنیک‌های انتخاب ویژگی از الگوریتم ژنتیک باینری (GA)، فاخته (Cuckoo) و ازدحام ذرات (PSO) و برای دسته‌بندی الگوریتم از طبقه‌بندی

نیمه نظارتی ماشین بردار پشتیبان (SVM) و شبکه‌های عصبی مصنوعی (MLP و RBF) استفاده شده است.
در پژوش حاضر مشخص گردید، انتخاب ویژگی‌ها بر اساس دسته‌بندی نیمه نظارتی باعث کاهش خطای سیستم در تخمین سری‌های زمانی شد.
در این راستا سعی شد سیستمی طراحی شود که منجر به کاهش هزینه در جمع‌آوری داده‌ها شود. در تحقیق حاضر از سه مجموعه داده با قابلیت سری زمانی از دادهای استاندارد UCI، استفاده گردید.
در این مطالعه، به طور کلی ۱۲ رویکرد (رجوع به جدول ۱) از ترکیب‌های الگوریتم و ماشین یادگیری تعریف گردید و به کار گرفته شد. رویکردهای مذکور در ۳ سناریوی متفاوت با استفاده از داده‌های استاندارد UCI، مورد آزمایش قرار گرفتند که در ادامه به بررسی عملکرد تمامی رویکردها در سناریوهای مورد مطالعه خواهیم پرداخت.

۴-۱- سناریوی اول؛ داده‌های کیفیت هوا

در این سناریو بعد از اعمال الگوریتم‌ها بر روی داده‌های کیفیت هوا و انجام مبحث انتخاب ویژگی، با استفاده از ماشین‌های یادگیری مدل‌سازی و برآورد داده‌های سری زمانی کیفیت هوا انجام گرفت. بر اساس مقادیر R^2 و RMSE و همچنین از نظر توانایی در کاهش بُعد، بهترین رویکردها با توجه به نوع ماشین‌های یادگیری (SVM، MLP و RBF) در این سناریو برگزیده شدند که در جدول ۱۲ قابل مشاهده است.
با توجه به جدول ۱۲ مشخص است که عملکرد رویکردهای شبکه‌های عصبی (MLP و RBF) نسبت به ماشین بردار پشتیبان با در نظر گرفتن، میزان کاهش بعد، خطای مدل و ضریب تبیین، بهتر بوده است. نکته دیگری که به خوبی محرز است این است که رویکردهای شبکه عصبی تابع پایه شعاعی (RBF) عملکرد کاملاً مشابهی داشته و الگوریتم‌های استفاده شده در ترکیب با این شبکه عصبی به جز در کاهش بُعد آن هم به طور کاملاً مشابه دخالت چندانی نداشته‌اند. با این حال، درخصوص سناریوی اول بهترین عملکرد مختص ترکیب شبکه عصبی پرسپترون چند لایه با الگوریتم فاخته (MLP+Cuckoo) است که از نظر کاهش بعد بسیار عالی عمل نموده و از لحاظ دقت و صحت برآورد نیز با اختلاف کمی نسبت به شبکه عصبی RBF، عملکرد بهتری دارد.

جدول ۱۲- نتایج عملکرد بر روی داده‌های کیفیت هوا

رویکردهای منتخب	بُعد اولیه	بُعد انتخاب شده	RMSE	R2
SVM+PSO	۱۳	۱۰	۰/۰۷	۰/۹۰
MLP+Cuckoo	۱۳	۵	۰	۰/۹۹
RBF +GA	۱۳	۹	۰/۰۲	۰/۹۹
RBF +PSO	۱۳	۹	۰/۰۲	۰/۹۹
RBF+Cuckoo	۱۳	۹	۰/۰۲	۰/۹۹

۴-۲- سناریوی دوم؛ برآورد داده‌های بازار بورس استانبول

در این سناریو نیز بر اساس مقادیر R^2 و RMSE و همچنین از نظر توانایی در کاهش بُعد، بهترین رویکردها با توجه به نوع ماشین‌های یادگیری (SVM، MLP و RBF) برگزیده شدند که در جدول ۱۳ مشاهده می‌گردد. الگوریتم ازدحام ذرات در انتخاب ویژگی موفق‌تر عمل نموده است. بنابراین، با توجه به نتایج به دست آمده درخصوص سناریوی دوم، رویکرد ماشین بردار پشتیبان برگزیده می‌شود.

جدول ۱۳- نتایج عملکرد بر روی داده‌های بازار بورس استانبول

رویکردهای منتخب	بُعد اولیه	بُعد انتخاب شده	RMSE	R^2
SVM+PSO	۹	۶	۰/۰۴	۰/۷۹
MLP+Cuckoo	۹	۷	۰/۰۴	۰/۷۹

با توجه به نتایج به دست آمده برای رویکردهای شبکه عصبی RBF در برآورد شاخص بازار بورس استانبول همان‌طور که مشاهده گردید رویکردهای شبکه عصبی RBF بسیار نتایج ضعیفی داشته و کاملاً نامناسب عمل کرده‌اند. الگوریتم‌های استفاده شده از نظر کاهش بعد بسیار خوب عمل نموده‌اند اما از آنجا که شبکه عصبی RBF در انجام مدل مناسب موفق نبوده است در نهایت کاهش بعد نیز تأثیری در نتیجه نهایی نداشته است. با این حال، با توجه به جدول ۱۳ می‌توان بیان نمود رویکردهای شبکه عصبی پرسپترون و ماشین بردار پشتیبان بسیار به هم

نزدیک می‌باشند و از لحاظ دقت و صحت برآورد نیز با اختلاف کمی نسبت به شبکه عصبی RBF، عملکرد بهتری داشته است.

۴-۳- سناریوی سوم؛ تخمین داده‌های SML ۲۰۱۰

همچنین در خصوص سناریوی سوم می‌توان چنین بیان داشت که نتایج حاصل آزمون رویکردها بر اساس مقادیر R^2 و RMSE و همچنین از نظر توانایی در کاهش بُعد (انتخاب ویژگی)، بهترین رویکردها با توجه به نوع ماشین‌های یادگیری (SVM، MLP و RBF) برگزیده شدند که در جدول ۱۴ قابل مشاهده است.

جدول ۱۴- نتایج عملکرد بر روی داده‌های SML ۲۰۱۰

رویکردهای منتخب	بُعد اولیه	بُعد انتخاب شده	RMSE	R2
SVM+PSO	۱۸	۱۵	۰/۱۵	۰/۵۱
MLP+Cuckoo	۱۸	۸	۰/۱۲	۰/۶۹
MLP +GA	۱۸	۱۱	۰/۰۷۸	۰/۸۷
RBF+Cuckoo	۱۸	۱۶	۰/۱۱	۰/۷۰

رویکردهای شبکه عصبی MLP در رابطه با مدل‌سازی مجموعه داده‌های SML ۲۰۱۰ نتایج جالبی نشان داده‌اند (جدول ۱۴). به طور کلی تمامی آن‌ها کاهش بعد مناسبی داشته‌اند. بیشترین کاهش بعد مربوط به رویکرد MLP+Cuckoo است و در رتبه بعدی رویکرد رویکردهای ماشین بردار پشتیبان در این سناریو کاهش بُعد بیشتری نسبت به شبکه عصبی RBF داشته است.

MLP+GA عملکرد پایین‌تری داشته است. به هر حال، درخصوص سناریوی سوم می‌توان بیان داشت که به طور کلی رویکردهای شبکه عصبی MLP دقت و عملکرد مناسب و بهتری داشته‌اند. هرچند رویکرد MLP+Cuckoo کاهش بعد بسیار قابل توجهی داشته است اما از آنجا که معیار مقایسه R^2 و RMSE می‌باشد رویکرد برگزیده ما MLP +GA است.

با توجه به نتایج حاصل از این تحقیق می‌توان گفت که استفاده از انتخاب ویژگی بر اساس دسته‌بندی نیمه نظارتی باعث کاهش خطای سیستم، افزایش دقت و افزایش سرعت تخمین سری‌های زمانی می‌گردد.

به طور کلی با توجه به بررسی‌های صورت گرفته و اینکه هر روش یادگیری ماشین و انتخاب ویژگی دارای مزایا و محدودیت‌هایی است؛ روش مناسب برای یک کاربرد خاص بر اساس اهداف پژوهش و معیارهای مـوردنظر از جمله سادگی، ثبات، تعداد ویژگی‌های کاهش‌یافته، دقت طبقه-بندی، حافظـه و زمـان محاسـبات مورد نیاز مـی‌توانـد انتخاب گردد.

در این تحقیق ۱۲ رویکرد در ایجاد یک طبقه‌بندی کننده بهینه مبتنی بر ماشین بردار پشتیبان، شبکه عصبی MLP و RBF بر مبنای الگوریتم ژنتیک، فاخته و ازدحام ذرات ارائه گردید. نتایج حاصل در مقایسه با نوع داده‌ها، بیانگر برتری الگوریتم ژنتیک در بیشتر موارد، به‌خصوص با افزایش ابعاد فضای جستجو است. همچنین نتایج بهینه با به‌کارگیری الگوریتم ژنتیک در حل هم‌زمان تعیین پارامتر و انتخاب ویژگی حاصل می‌شود که در آن با انتخاب ویژگی-های کمتر، دقت بالاتر حاصل شد. از این‌رو با استفاده از طبقه‌بندی کننده‌ی کارا و قدرتمند شبکه عصبی و ماشین بردار پشتیبان در کنار الگوریتم بهینه‌سازی و فرا ابتکاری، می‌توان یک سیستم طبقه‌بندی ترکیبی بهینه برای تخمین سری‌های زمانی طراحی کرد.

به طور کلی با توجه به نتایج حاصله از این تحقیق مشخص گردید که رویکردهای مختلف ترکیب ماشین‌های یادگیری (SVM,MLP,RBF) و الگوریتم‌های بهینه‌سازی (PSO,Cuckoo,GA) به طور کلی باعث بهبود نتایج و خروجی‌ها می‌شود. البته توانایی و عملکرد هر کدام از رویکردها با توجه به نوع و ماهیت داده‌ها متفاوت است. همین امر باعث شده است که گاهاً رویکرد شبکه عصبی MLP و الگوریتم GA یا Cuckoo نتایج بهتری داشته باشد و در برخی موارد نیز رویکرد ماشین بردار پشتیبان با الگوریتم PSO نتایج بهتری داشته است. بنابراین، آن‌چه مهم است این نکته است که انتخاب ویژگی‌ها بر اساس دسته‌بندی نیمه نظارت شده در رویکردهای استفاده شده در این تحقیق باعث افزایش سرعت و دقت تخمین سری‌های زمانی گردیده است.

از میان روش‌های مورد بررسی روش‌های یادگیری گروهی به دلیل بکارگیری چندین یادگیرنده در یک روند تکراری

قادر به انتخاب ویژگی‌های بهینه در طول آموزش می‌باشند، که سبب بهبود عملکرد طبقه‌بندی می‌شوند. با این حال این رویکردها به دلیل ماهیت تکرار شونده، بار محاسباتی سنگینی دارند. همچنین از میان روش‌های منفرد، روش ماشین بردار پشتیبان به دلیل استفاده از ویژگی‌های هندسی داده‌ها و تولید دقت‌های بالای طبقه‌بندی یک روش مناسب و متداول در تخمین سری‌های زمانی است اما این روش از پیچیدگی محاسباتی رنج می‌برد. علاوه بر این، الگوریتم‌های یادگیری نیمه نظارت شده را به دلیل دارا بودن مزیت روش‌های یادگیری نظارت شده و نظارت نشده می‌توان به عنوان یک روش مؤثر برای انتخاب ویژگی و طبقه‌بندی به کار برد. پیشنهاد می‌گردد در تحقیقات آینده با توجه به اهمیت مسئله سری‌های زمانی، از الگوریتم‌های فرا ابتکاری دیگر از قبیل روش درخت تصمیم‌گیری و روش مبنی بر خرد جمعی به منظور بهینه‌سازی شبکه عصبی و ماشین‌های بردار پشتیبان در تخمین سری‌های زمانی استفاده گردد. همچنین بهینه‌سازی شبکه عصبی و ماشین‌های بردار پشتیبان در تخمین سری‌ها زمانی استفاده گردد. همچنین بهینه‌سازی اتوماتیک پارامترهای الگوریتم ژنتیک که در این تحقیق با ازمون و خطا به دست آمد از دیگر موضوعات پیشنهادی است.

مراجع

[1] حسینعلی نیرومند، ابوالقاسم بزرگ نیا، مقدمه‌ای بر تحلیل سری‌های زمانی (ترجمه) (The Analysis Of Time Series An Introduction, C. Chatfield) / تألیف سی چتفیلد؛ (انتشارات دانشگاه مشهد؛ ISBN 964-5782-88-0132)، ۱۳۸۹.

[2] محمد ناظری تهرودی، "مقایسه مدل‌های PAR و MPAR در مدل‌سازی سری زمانی ماهانه دبی رودخانه تحت تأثیر عوامل هواشناسی"، مجله مدل‌سازی در مهندسی، دوره ۱۶، شماره ۵۲، ۱۳۹۷، صفحه ۱۸۳-۱۹۱.

[3] شیرین پیران، علیرضا برهانی داریان، "انتخاب هوشمند ورودی‌ها با استفاده از الگوریتم ژنتیک باینری در شبکه‌ی عصبی تکاملی"، چهارمین کنگره بین المللی عمران، معماری و توسعه شهری، ۱۳۹۵.

[4] N. Muttil, & K.W. Chau, "Machine-learning paradigms for selecting ecologically significant input variables", Engineering Applications of Artificial Intelligence, Vol. 20, No. 6, 2007, pp. 735–744.

[5] A.R. Yildiz, "Cuckoo search algorithm for the selection of optimal machining parameters in milling operations", The International Journal of Advanced Manufacturing Technology, Vol. 64, Nos. 1–4, 2013, pp. 55–61.

[6] S. Galelli, G.B. Humphrey, H.R. Maier, A. Castelletti, G.C. Dandy, & M.S. Gibbs, "An evaluation framework for input variable selection algorithms for environmental data-driven models", Environmental Modelling & Software, Vol. 62, 2014, pp. 33–51.

[7] M. Dash, H. Liu, "Feature selection for classification", Intelligent Data Analysis, Vol. 1, No. 3, 1997, pp. 131–156.

[8] P.J. Brockwell, R.A. Davis, Introduction to Time Series and Forecasting, 2nd edition, Springer, New York, 2002.

[9] H. Liu, H. Motoda, Feature Extraction, Construction, and Selection: A Data Mining Perspective, Kluwer

[10] J. Neter, M.H. Kutner, C.J. Nachtsheim, L. Wasserman, Applied Linear Statistical Models, 4th edition, Irwin, 1996.

[11] R. Rajabioun, "Cuckoo Optimization Algorithm", Applied Soft Computing, Vol.11, No.8, 2011, pp. 5508–5518.