7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

# A feature-based Method by butterfly Optimization Algorithm to Predict the Execution Time of a Job in Hadoop

Hadi Mollanoori, Abolfazl Gandomi*

[1]Department of Computer Engineering, Yazd Branch, Islamic Azad University, Yazd, Iran,
Mollanoori.hadi@gmail.com, gandomi@iauyazd.ac.ir

**Abstract**

Hadoop is among important platforms used for big data processing whereby MapReduce operations can be used to process big data in real time. One important challenge in big data processing using Hadoop is proper scheduling of jobs executed in the framework, because correct and optimal execution of jobs hinges upon predicting their execution time. This study aimed at estimating the execution time of jobs. To this end, butterfly optimization algorithm was used to select important job features and the artificial neural network was used for learning. Analyses showed lower error rates for butterfly optimization algorithm compared to Particle Swarm Optimization algorithm, the Spotted Hyena Optimization algorithm, and the Firefly algorithm. Results showed that the value of the objective function for feature selection decreased in the proposed iteration-based method. In order to predict job execution time, the initial population in this method increased, which in turn, reduced the Root Mean Square Error by about 25.85%. The proposed method showed a lower execution time estimation error in comparison to other methods like Multilayer Neural Network, Recursive Neural Network, Decision Tree, and Random Forest.

**KEYWORDS:**Hadoop, MapReduce, Butterfly Optimization Algorithm, Neural Network, Estimation of Job Execution Time

7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

# یک روش مبتنی بر انتخاب ویژگی با الگوریتم بهینه سازی پروانه به منظور پیش بینی زمان اجرای Job های مبتنی بر نگاشت-کاهش

هادی ملانوری، ابوالفضل گندمی

[1]گروه مهندسی کامپیوتر، واحد یزد، دانشگاه آزاد اسلامی، یزد، ایران،
Mollanoori.hadi@gmail.com , ،Gandomi@iauyazd.ac.ir

**چکیده**

**یکی از بسترهای مهم برای پردازش کلان داده استفاده از معماری هدوپ است که می‌توان با استفاده از عملیات نگاشت-کاهش، داده‌های بزرگ را در زمان واقعی مورد پردازش قرار داد. یکی از چالش‌های مهم پردازش کلان داده در هدوپ زمانبندی دقیق jobهایی است که در این بستر اجراء می‌شوند و نیاز است که قبل از اجرای jobها زمان اجرای آن را پیش‌بینی نمود تا به درستی و به صورت بهینه زمانبندی شوند. در این مقاله برای تخمین زمان اجرای دقیق jobها از الگوریتم بهینه‌سازی پروانه برای انتخاب ویژگی-های مهم jobها و از شبکه عصبی مصنوعی برای یادگیری استفاده شده است. تجزیه و تحلیل ها نشان می‌دهد که این الگوریتم، از الگوریتم بهینه‌سازی ذرات، الگوریتم بهینه‌سازی کفتار و الگوریتم کرم شب‌تاب خطای کمتری دارد. آزمایشات نشان می‌دهد مقدار تابع هدف انتخاب ویژگی، در روش پیشنهادی بر حسب تکرار، یک روند نزولی و کاهشی است. به منظور پیش بینی زمان اجرای jobها، افزایش جمعیت اولیه در این روش، متوسط مجذور خطا را در حدود ۲۵.۸۵٪ کاهش می‌دهد. مقایسه روش پیشنهادی با روش‌های دیگر نشان می‌دهد خطای تخمین زمان اجرا در روش پیشنهادی از شبکه عصبی چند لایه، شبکه عصبی بازگشتی، درخت تصمیم‌گیری و جنگل تصادفی کمتر است.**

**کلمات کلیدی**

**نگاشت - کاهش، الگوریتم بهینه‌سازی پروانه، شبکه عصبی، زمان اجرای job**

## ۱- مقدمه

داده‌های که پیرامون ما ایجاد می‌شوند دارای دانش نهفته بوده و ارزش اطلاعاتی بالایی دارند و اگر به کمک روش‌های کشف دانش بتوان این الگوها را استخراج نمود می‌توان بخوبی به تجزیه و تحلیل داده‌ها پرداخت و آنها را تفسیر نمود. در واقع تجزیه و تحلیل داده‌های تولید شده در بیشتر نرم-افزارهای کاربردی در جهت استخراج دانش یکی از وظایف کشف دانش و الگوریتم‌های آن در تکنیک‌های یادگیری ماشین و داده‌کاوی است. امروزه روش‌های کشف دانش به طور گسترده برای بکارگیری در انواع داده‌ها استفاده می‌شود تا الگوی پنهان درون آنها کشف شود اما استفاده از این روش‌ها برای کشف دانش در داده‌های که حجم و ابعاد زیادی دارند یک فرآیند چالش‌برانگیز است. به عبارت بهتر روش‌های کشف دانش و یادگیری ماشین برای اجراء بر روی داده‌های بزرگ نیاز به راه‌کارهای جدیدی دارند زیرا زمان پردازش این داده‌ها در روش‌های سنتی داده‌کاوی زیاد بوده و استفاده کاربردی از این روش‌ها را برای تحلیل داده‌ها با چالش مواجه می-سازد[۱]. امروزه تعداد زیادی نرم‌افزار در حالت آفلاین یا آنلاین وجود دارند که حجم زیادی از داده را ایجاد نموده که در حوزه کلان داده یا داده بزرگ قرار دارند که نمونه این بسترها را می‌توان در شبکه‌های اجتماعی نام برد که تعدادی زیادی کاربر وجود دارد و هر کاربر نیز حجم زیادی از داده را تولید می‌نماید و یک مثال دیگر از این نوع داده‌ها را می‌توان داده‌های ایجاد شده در سرویس پست الکترونیک را مثال زد[۲]. افزایش حجم داده‌ها در اینترنت

و پیرامون ما باعث شده حجم زیادی از داده وجود داشته باشد که نیاز است به سرعت مورد پردازش و تحلیل قرار گرفته شوند تا بتوان به اطلاعات مهم آن دسترسی پیدا نمود و از طرفی اگر این دانش در زمان واقعی کشف نشود ارزش این اطلاعات از بین می‌رود که نمونه آن تشخیص پولشویی و تقلب در حجم بالایی از داده‌ها است که در اینترنت و سیستم بانکی وجود دارد و نیاز است که این تراکنش‌ها با تحلیل و پردازش در بسترهای مناسب مانند پردازش ابری مورد تجزیه و تحلیل قرار گرفته شوند[۳]. ظهور بانکداری الکترونیک باعث شده است بسیاری از بانکها خدمات خود را تحت وب ارایه دهنده و در روز میلیون‌ها تراکنش در فضای مجازی انجام می‌شود و به نوعی کلان داده محسوب می‌شوند. اطلاعات مالی کاربران و تراکنش‌های آن یک نمونه واقعی از کلان داده است که دارای ارزش اطلاعات بالایی است و این موضوع زمانی اهمیت پیدا می‌کنند که در صدد باشیم به کمک تجزیه و تحلیل تراکنش‌ها رفتار مشتریان بانک را تحلیل نموده و اطلاعات باارزشی را از داده‌های خام و البته کلان داده استخراج نمایم[۴]. مطالعات نشان می‌دهد که حجم داده‌ها در حال افزایش بود و بخش زیادی از این داده‌ها مرتبط با فضای وب و نرم‌افزارهای کاربردی در این حوزه است و افزایش تعداد کاربران اینترنت و تلفن‌های همراه که می‌توانند به اینترنت متصل شوند و داده را تولید نمایند بر شتاب این داده‌ها و حجم آنها افزوده است. یکی دیگر از بسترهای که کلان داده تولید می‌نماید مرتبط با شبکه اینترنت اشیاء است و هر کدام از اشیاء هوشمندی که در این شبکه قرار دارند می‌توانند داده زیادی را تولید نمایند[۵]. امروزه داده‌های بزرگ در حوزه‌های مختلفی تولید می‌شوند که می‌توان این داده‌ها را در حوزه تجاری، پزشکی، سلامت، هواشناسی، ستاره-شناسی، نجوم و امنیتی مشاهده نمود و مسلماً برای پردازش این داده‌ها نیاز است که ابزارها و فناوری‌های استفاده شود این حجم داده را مورد پردازش قرار داده و آن را تحلیل نماید. یکی از بسترهای مناسب برای پردازش کلان داده استفاده از شیوه‌های محاسبات ابری است و این شیوه خود نیز دارای فناوری‌های مختلفی است که از جمله آنها می‌توان به هدوپ[۶] اشاره نمود. هدوپ یک بستر پردازش کلان داده است که با استفاده از تعدادی کلاستر موازی عمل پردازش کلان داده را انجام دهد و برای این منظور هر کدام از کلاسترها با تکنیک نگاشت و کاهش پردازش‌ها را موازی بر روی تعدادی کلاستر پیاده‌سازی و اجراء نموده اما این روش نیز چالش خاص خود را دارد. یکی از چالش‌های هدوپ آن است که حافظه اصلی استفاده نمی‌نماید و سرعت آن کاهش می‌یابد و برای رفع این چالش از فناوری آپاچی اسپارکبه جای آن استفاده می‌شود. یکی از چالش‌های مهم در پردازش job ها در بسترهایی مانند هدوپ و اسپارک آن است که نیاز است به درستی زمانبندی و اجراء شوند که در این مقاله یک راه‌کار پیشنهادی برای آن ارائه می‌گردد. چالش مهم پردازش اطلاعات و داده‌ها در سیستم‌های ابری فقط حجم داده‌ها نبوده بلکه برای پردازش در سیستم ابری چالش‌های فراوان دیگری وجود دارد که می‌توان به مدیریت منابع و زمانبندی و مجازی‌سازی اشاره نمود. هر کدام از این چالش‌ها به نوعی نیاز به این مسئله دارد که یک پروسه یا job در سیستم پردازش ابری چقدر زمان نیاز دارد و به عبارت بهتر تخمین زمان پروسه و job در سیستم پردازش ابری در بسیاری از کاربردهای مرتبط با پردازش ابری مهم و حیاتی است و باعث می‌شود منابع بهتر بین jobها تخصیص داده شود. یکی از روش‌های پیش بینی پروسه‌ها در سیستم پردازش ابری استفاده از روش‌های داده‌کاوی و یادگیری ماشین است که در این پژوهش از شبکه عصبی مصنوعی چند لایه برای این منظور استفاده می-شود[۷] . چالش مهم پیش بینی دقیق زمان اجرای یک job در چارچوب هدوپ دارای ویژگی‌های زیادی است که یک پروسه دارد و یادگیری در بین این همه ویژگی دقت پیش بینی را کاهش می‌دهد از این جهت برای بهبود پیش بینی شبکه عصبی مصنوعی می‌توان فاز انتخاب ویژگی را اعمال نمود تا یادگیری فقط بر روی ویژگی‌های مهم انجام شود. مکانیزم انتخاب ویژگی در واقع انتخاب نمودن تعدادی از ویژگی‌ها است که می‌تواند خطای طبقه‌بندی را کاهش دهد و برای این منظور نیاز است که با رویکرد یک مسئله بهینه‌سازی با آن رفتار شود زیرا نیاز است که ویژگی‌های انتخاب شود که خطا را کاهش دهد. برای انتخاب بهینه ویژگی‌ها می‌توان از روش‌های فرابتکاری استفاده نمود تا بهترین ویژگی‌ها برای تشخیص زمان اجرای jobها استخراج شود. یکی از الگوریتم‌های فرابتکاری که دارای رویکرد هوش گروهی است الگوریتم بهینه‌سازی پروانه[۱۰] است که از رفتار انتشار فرمون توسط این حشرات و جذب سایر پروانه‌ها به سمت جواب بهینه الگوبرداری شده است. این مقاله دارای چند بخش است و در ابتدا مفاهیم مرتبط با مقاله نظیر هدوپ ارایه شده و سپس مروری بر الگوریتم بهینه‌سازی پروانه شده و مطالعات مرتبط با زمانبندی jobها در هدوپ بررسی شده و سپس روش پیشنهادی برای پیش بینی زمان اجرای jobها با استفاده از الگوریتم پروانه و شبکه عصبی مصنوعی چند لایه ارایه شده و در ادامه روش پیشنهادی پیاده‌سازی و مورد تحلیل قرار گرفته شده و نتایج آن بیان می‌شود.

## ۲- مروری بر مطالعات انجام شده

در[۱۳] بیان گردیده علاوه بر اشیاء هوشمند نرم‌افزارهای کاربردی نیز داده-های زیادی تولید می‌نماید و نیاز است که داده‌ها به درستی مورد پردازش قرار گرفته شود و مهمترین بستر برای این داده‌ها و پردازش آنها استفاده از چارچوب‌های ابری مانند هدوپ است که می‌توانند این داده‌ها را در زمان نزدیک به زمان واقعی پردازش نمایند. با توجه به اینکه در حوزه تحلیل داده-های کلان، عملکرد بیدرنگ یا تقریباً زمان حقیقی بسیار مورد توجه پژوهشگران است، تلاش می‌شود تا حد امکان از بروز تأخیر در پردازش داده-ها در سیستم جلوگیری شود. با وجود آنکه پردازش در قالب هدوپ به عنوان بستر مناسب برای پردازش کلان داده در نظر گرفته می‌شود اما منابع بکار رفته در این سیستم نیز برای پردازش کلان داده محدود بوده و برای رفع این چالش ضرورت دارد مواردی رعایت شود

در[۱۴] برای پیش بینی و مدیریت منابع در ماشین مجازی، مناطق جذاب در محیط ابر یک روش مبتنی بر بهینه‌سازی ترتیبی ارایه شده است. در واقع

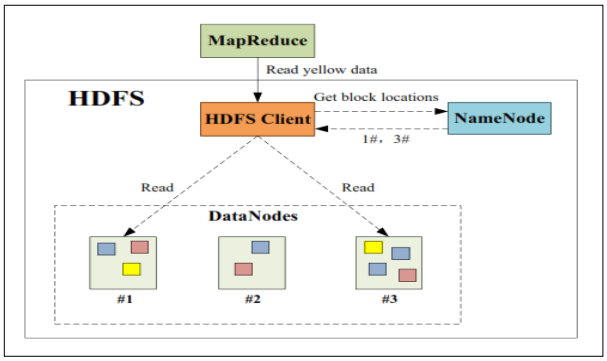
پیش بینی زمانبندی در ماشین‌های مجازی یک وظیفه مهم برای اجرای jobها برای به حداقل رساندن تاخیر و اجتناب از حالت‌های غیر ضروری در سیستم ابری است. نتایج تحلیل روش پیشنهادی آنها نشان می‌دهد این روش می‌تواند در زمان مناسب پروسه‌ها و jobها را تشخیص داده و با توجه به زمان اجرای پیش بینی آنها منابع بهینه در ماشین‌های مجازی را برای آنها فراهم نماید. در[۱۵]، برای زمانبندی پروسه‌ها در سیستم پردازش ابری یک روش مبتنی بر الگوریتم جستجوی ممنوعه و الگوریتم جستجوی هارمونی ارایه و معرفی شد. نتایج آزمایشات آنها نشان می‌دهد الگوریتم پیشنهادی از لحاظ هزینه تولید و زمانبندی در مقایسه با جستجوی ممنوعه، جستجو هارمونی اثرگذاری بیشتری دارد. در[۱۶] یک روش پیش بینی و تخمین زمان اجرای پروسه در فضای ابری ارایه شد که بر اساس یادگیری ماشین و الگوریتم شبکه عصبی LSTM عمل می‌نماید. نتایج آزمایشات نشان می‌دهد که دقت روش پیشنهادی آنها در مقایسه با روش های فعلی بهبود یافته است. علاوه بر این، برای درک بهتر زمان اجرای یک پروسه، آنها از یک الگوریتم خوشه بندی سلسله مراتبی نظارت نشده، BIRCH، برای طبقه بندی و تفسیر نتایج استفاده نمودند. در[۱۷] برای تخمین زمان اجرای jobها در پردازش ابری از روش‌های مبتنی بر تحلیل سری زمانی استفاده شده است تا بتوانند زمان لازم برای اجرای یک پروسه که در پردازش داده‌های کلان استفاده می‌شود تخمین زده شود و تحلیل و ارزیابی آنها نشان می‌دهد روش پیشنهادی در صورتی که بار متوازنی بر روی سیستم باشد توانایی خوبی برای تخمین زمان اجراء دارد. در[۱۸] یک روش جدید برای زمانبندی jobها در سیستم هدوپ برای پردازش کلان داده ارائه نمودند. نتایج تجربی و شبیه سازی آنها اثربخشی طراحی را به شدت تایید کرده است و زمانبندی job جدید ما می تواند به طور متوسط زمان پردازش jobها را تا ۴۵٪ کاهش دهد. در[۱۹]، یک روش جدید زمانبندی jobها بر اساس نگاشت و کاهش در سیستم هدوپ ارایه شد. هدوپ یک پلت فرم تجزیه و تحلیل داده بزرگ برای منبع باز است که به طور گسترده در هر دانشگاه و صنعت استفاده می-شود. جداسازی چارچوب مدیریت منابع و برنامه‌ریزی، نسل بعدی هدوپ، یعنی Hadoop YARN، به چارچوب برنامه نویسی مختلف متصل شده و قادر است انواع مختلفی از حجم jobها مانند تجزیه و تحلیل تعاملی و پردازش جریان داده را مدیریت نمایند. با این حال، بسیاری از الگوریتم‌های زمانبندی در YARN برای پردازش دسته‌ای طراحی شده‌اند. این پژوهش یک FSPY (پروتکل ملاقات منصفانه در YARN) را پیشنهاد می‌کند تا پاسخگویی را با تضمین عدالت بین jobها بهبود بخشد. نتایج تجربی نشان می‌دهد که زمان‌بندی آنها تا ۱۰ برابر با توجه به پاسخگویی تحت بار کاری سنگین بهتر عمل می‌نمایند. در[۲۰] یک روش جدید زمانبندی برای jobها در هدوپ یکی از برجسته ترین و پیشگام‌ترین فناوری‌ها برای مدیریت داده-های بزرگ است معرفی نمودند. الگوریتم‌های مختلف برنامه‌ریزی jobها در چارچوب هدوپ در دهه گذشته بخوبی توسعه یافته است و تلاش شده تا یک مرور کلی از روش‌های زمانبندی jobها در هدوپ ارائه گردد. رویکردهای مورد بررسی تحت دسته‌های مختلف قرار می‌گیرند که از جمله آنها اولویت بندی وظایف، اولویت بندی منابع، اولویت بندی اندازه job، روش های ترکیبی را نام برد.

## ۳-روش پیشنهادی

### ۳-۱- هدوپ

یکی از فناوری‌های جدید در حوزه پردازش کلان داده سیستم توزیع شده هدوپ می‌باشد. هدوپ یک پروژه متن باز از بنیاد آپاچی بوده و برای پردازش کلان داده در نظر گرفته شده است. هدوپ برای پردازش کلان داده به دو جزء مهم سیستم فایل توزیع شده و عملیات نگاشت و کاهش متکی است. سیستم فایل توزیع شده هدوپ باعث می‌شود که کل فضای ذخیره‌سازی کلاسترها به عنوان یک سیستم فایل واحد در نظر گرفته شود و هر داده بر روی یکی از کلاسترها قرار گرفته شود و از طرفی مجموعه‌ای از توابع نگاشت و کاهش نیز باعث می‌شود که پردازش‌ها به قطعات کوچک تقسیم شده و به سمت داده‌ها که در فضای توزیع شده قرار دارند هدایت شوند. مزیت مهم عملیات نگاشت و کاهش در آن است که توسط مجموعه‌ای از توابع نگاشت و کاهش کد پردازش به صورت توزیع شده بر روی داده‌ها اجراء می‌شود بدون آنکه داده‌ها بین گره‌ها و کلاسترها جابجا شوند و این حالت سرعت عملیات پردازش کلان داده را افزایش می‌دهد. یک معماری از سیستم پردازش هدوپ که در آن دو جزء نگاشت و کاهش به همراه سیستم فایل توزیع شده نقش کلیدی دارند در شکل(۱)، نمایش داده شده است:

مطابق شکل فوق، داده‌ها در سیستم فایل توزیع شده قرار داده می‌شوند و عملیات موازی و پردازش توزیع شده در قالب دستورات نگاشت و کاهش به سمت داده‌های که بر روی سیستم فایل قرار دارند ارسال می‌شوند و داده‌ها به صورت همزمان در کلاسترهای مختلف مورد پردازش توسط عملیات نگاشت و کاهش قرار گرفته می‌شوند. منظور از عملیات نگاشت و کاهش تبدیل دستورات پردازش بر روی داده‌های است که به صورت توضیح شده بر روی مجموعه‌ای از کلاسترها قرار دارد. مکانیزم نگاشت و کاهش که در سیستم



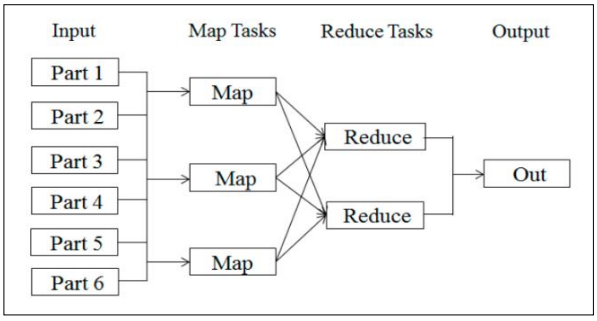

**شکل ۱: یک معماری از چارچوب پردازش هدوپ [۱۱].**

هدوپ بکار گرفته می‌شود در شکل(۲)، نمایش داده شده است و طبق این چارچوب نگاشت کدهای اصلی را به ریزکدهای تقسیم می‌نماید که بر روی داده‌های توزیع شده اجراء می‌شوند و کاهش نتایج میانی را برای محاسبه نهایی استفاده می‌نماید. مطابق شکل فوق می‌توان دریافت که در مدل محاسباتی نگاشت و کاهش مراحل و مکانیزم‌های ذیل انجام می‌شود:

- در نگاشت و کاهش دو وظیفه نگاشت و کاهش وجود دارد که هر کدام از آنها توسط توابعی در هدوپ ارایه می‌گردد.
- در هر مرحله و بعد از فرآیند نگاشت مرحله کاهش بر روی داده‌ها اجراء می‌شود.
- در مرحله نگاشت مقادیر جفت و کلید متناظر با آنها ایجاد می‌شود.
- خروجی مرحله نگاشت یا همان جفت کلید و مقدار به عنوان ورودی مرحله کاهش در نظر گرفته می‌شود.

هدف مرحله کاهش تبدیل داده‌ها به زیر مجموعه‌های کوچکتر از داده‌ها می‌باشد .

پردازش در بسترهای محاسبات توزیع شده مانند هدوپ با چالش‌های زیادی برای داده‌های بزرگ مواجه است که از جمله آنها می‌توان زمانبندی jobها اشاره نمود. یک jobیه پروسه‌ای گفته می‌شود که برروی داده‌های بزرگ یا داده‌های بارگذاری شده در هدوپ پیاده‌سازی می‌شود. یک job برای اجراء می‌تواند از منابع هدوپ استفاده نماید اما این منابع چون محدود است و تعداد jobها از این منابع بیشتر است نیاز به زمانبندی jobها وجود دارد. در زمانبندی jobها به عوامل مختلفی می‌توان توجه نمود که از جمله آنها می‌توان به زمان پردازش jobها اشاره نمود. برای بهبود زمانبندی jobها می‌توان jobها را دسته‌بندی نمود و بر اساس زمان اجرای آنها یا زمانی که نیاز است پروسه آنها تکمیل شود طبقه‌بندی شوند و سپس jobها را اولویت داد و به عنوان مثال jobهایی که زمان اندکی نیاز دارند اولویت بالا دارد و jobهایی که زمان زیادی نیاز دارند با اولویت کمتر اجراء نمود. چالش مهم این روش در آن است که نیاز است زمان اجراء در ابتدا پیش بینی شود و سپس از این زمان استفاده شود. در روش پیشنهادی برای پیش بینی زمان اجرای هر job برای زمانبندی در هدوپ موارد ذیل در نظر گرفته می‌شود:




**شکل ۲: عملیات نگاشت و کاهش در هدوپ [۱۲]**

- ویژگی‌های مرتبط با jobها که بر روی زمان اجراء تاثیر دارند به عنوان ورودی یک روش پیش‌بینی در داده‌کاوی مانند شبکه عصبی مصنوعی چند لایه در نظر گرفته می‌شوند.
- jobها بر اساس ویژگی‌ها به دسته‌های مختلفی از نظر زمان اجراء تقسیم می‌شوند که در حالت استاندارد و ساده می‌توان دو دسته زمان اجرای زیاد و کم را برای هر job در نظر گرفت.
- jobها به دسته‌های مختلف از نظر زمانی تقسیم شده و برای کاهش دادن خطای طبقه‌بندی jobها به دسته‌های مختلف از فاز انتخاب ویژگی استفاده می‌شود.
- در روش پیشنهادی از الگوریتم بهینه‌سازی پروانه که در سال ۲۰۱۹ ارائه شده به عنوان یک روش انتخاب کننده ویژگی استفاده می‌شود تا ویژگی‌های مهم jobها برای زمانبندی استفاده گردد.

در مورد ویژگیهای job نیز میتوان ویژگیهای دیگری مانند : تعداد map ها، تعداد Reduce ها، تعدادcore ها، حجم حافظه، زمان اجرای map task، زمان اجرای Reduce task و غیره را نام برد. یکی از این ویژگیها زمان اجرای job هست که ما می خواهیم این ویژگی را از طریق پروفایل با استفاده از زبان متلب تخمین بزنیم و ورودی برنامه نیز همان پروفایل job می باشد. job های بیان شده جزئی از پروفایل job هستند.

## ۳-۲- الگوریتم بهینه سازی پروانه

پروانه‌ها ازجمله حشراتی می‌باشند که می‌توانند باهم از طریق موارد شیمیایی که به آن‌ها فرومون گفته می‌شود ارتباط برقرار نمایند. مشاهده می‌شود که پروانه‌ها به انگیزه‌های مختلف نظیر یافتن جفت یا غذا در هوا فرومون آزاد نموده تا سایر پروانه‌ها را به سمت خود جذب نمایند و به‌طورمعمول پروانه‌ها سعی می‌کنند به سمت منبعی از فرومن‌ها حرکت نمایند که بیشترین مقدار فرومون در آن ناحیه انتشاریافته است

الگوریتم بهینه‌سازی پروانه یک الگوریتم فراابتکاری گروهی است که از رفتار پروانه‌ها در انتشار فرومون شیمیایی و جذب آن‌ها توسط سایر پروانه‌ها الگوبرداری شده است. در این الگوریتم هر پروانه یک راه‌حل مساله است و بر اساس شایستگی خود می‌تواند مقداری فرومون در هوا منتشر نماید. در الگوریتم بهینه‌سازی پروانه یک پروانه می‌تواند دو نوع حرکت ذیل را انجام دهد[ ۱۰]:

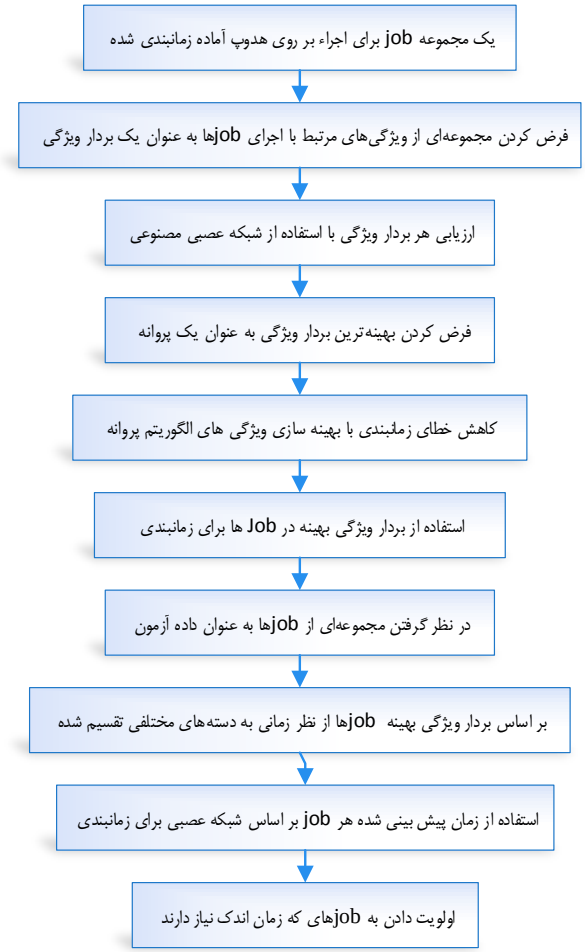
- با یک احتمال تصادفی به سمت بیشترین مقدار فرومون موجود در فضای مساله حرکت نماید که توسط شایسته‌ترین پروانه منتشر می‌شود و درواقع نوعی حرکت جستجوی محلی پیرامون جواب بهینه در این حالت انجام می‌شود.

- یک پروانه در جمعیت دو پروانه را تصادفی انتخاب نموده و به سمت فرومون آن پروانه‌ها جذب می‌شود و درواقع نوعی حرکت جستجوی سراسری در این حالت انجام می‌شود.

یک سلسه مراحل در روش پیشنهادی وجود دارد که در شکل(۳)، این مراحل برای انتخاب ویژگی jobهای قابل اجراء در هدوپ در زمانبندی نمایش داده شده است:

در روش پیشنهادی برای انتخاب ویژگی و زمابندی jobها در هدوپ راه حل ذیل وجود دارد:

- استفاده از زمان اجرای تعدادی از jobها برای آموزش و انتخاب ویژگی
- انتخاب ویژگی‌های بهینه مرتبط با زمان اجرای jobها با الگوریتم بهینه‌سازی پروانه




**شکل ۳: مکانیزم پیشنهادی برای انتخاب ویژگی توسط الگوریتم پروانه در زمانبندی**

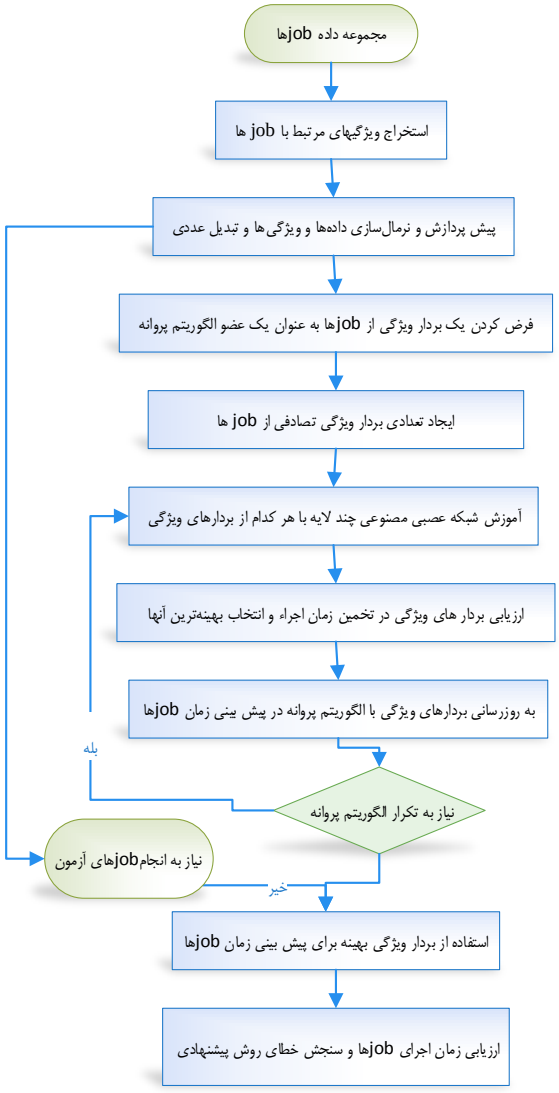
- ورود تعدادی jobبرای زمانبندی به هدوپ و استفاده از ویژگی‌های انتخاب شده
- تخمین نوع کلاس زمانبندی jobبا شبکه عصبی مصنوعی
- اولویت دادن به زمانبندی با توجه به کلاس و دسته زمانی نوع job

در شکل(۴)، چارچوب روش پیشنهادی برای انتخاب ویژگی jobها در هدوپ نمایش داده شده است. در این چارچوب تعدادی job که برای اجراء در هدوپ در نظر گرفته شده است از مجموعه داده مورد نظر استخراج شده و سپس در ادامه این مجموعه jobها که دارای تعدادی ویژگی است مورد فاز استخراج ویژگی قرار گرفته شده و یک بردار ویژگی از jobها به عنوان یک پروانه یا یک عضو الگوریتم بهینه‌سازی پروانه تعریف می‌شود:

هر بردار پروانه دارای تعدادی ویژگی است که می‌تواند آنها را انتخاب نشده یا انتخاب شده در نظر گرفت که در اینجا آنها با صفر و یک نشان داده می‌شود. هر پروانه یا بردار ویژگی به عنوان ورودی شبکه عصبی مصنوعی استفاده شده و برای آموزش آن استفاده می‌شود و با داده‌های آموزشی فرآیند آموزش انجام می‌شود و میزان شایستگی هر بردار وبژگی برای زمانبندی jobها به دو عامل مهم ذیل بستگی دارد:

- خطای تشخیص زمان اجرای jobها و نوع زمان از نظر اندک یا زیاد یا متوسط
- تعداد ویژگی انتخاب شده در هر بردار ویژگی برای زمانبندی jobها

هر بردار ویژگی که این دو شاخص را کاهش دهد یک بردار ویژگی بهتری است که می‌تواند برای زمانبندی jobها نیز از آن استفاده نمود. با اجرای الگوریتم بهینه‌سازی پروانه می‌توان بر روی بردارهای ویژگی اثر گذاشته و آنها را به روز نموده و سپس آنها را با شبکه عصبی مصنوعی مورد ارزیابی مجدد قرار داد و در نهایت در تکرار آخر می‌توان از بردار ویژگی بهینه که دارای حداقل خطای ممکن در تخمین زمان jobها است برای زمانبندی استفاده نمود. با استخراج بردار ویژگی بهینه می‌توان از آن به عنوان ورودی شبکه عصبی مصنوعی در انتخاب نوع job و زمان و دسته آن قضاوت نمود و به کمک داده‌های آزمون که نمونه‌های jobهای ورودی برای ارزیابی می‌باشند می‌توان این jobها را در دسته زمانی خود برای زمانبندی قرار داد و الگوریتم زمانبندی دلخواه را بر روی آنها اجراء نمود. فلوچارت روش پیشنهادی برای تشخیص دسته و کلاس نوع jobها از نظر زمانی در شکل(۵)، نمایش داده شده است و در این فلوچارت مراحلی مانند کدینگ راه‌حل، ایجاد جمعیت اولیه، جستجوی محلی و سراسری و از طرفی تبدیل بردار ویژگی‌ها از حالت پیوسته به باینری با استفاده از توابع تبدیل نمایش داده شده است. در فلوچارت روش پیشنهادی در ابتدا نیاز است که یک بردار ویژگی اولیه از jobها در نظر گرفته شود و این بردار ویژگی یک پروانه مانند فرمول (۱)، است و در فاز اول تعدادی از این بردارهای ویژگی تصادفی ایجاد شده و در قالب یک جمعیت اولیه آنها را ایجاد نموده تا هر کدام از آنها برای آموزش شبکه عصبی مصنوعی استفاده



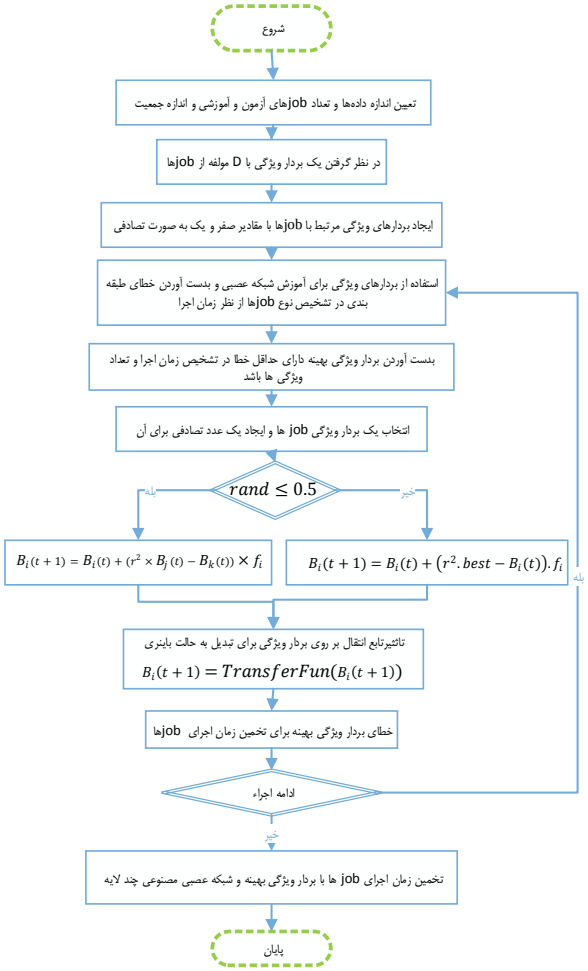

**شکل ٤: چارچوب روش پیشنهادی برای انتخاب ویژگی**

شود و در طبقه‌بندی jobها از نظر زمانی استفاده شود و این جمعیت را می‌توان به مانند فرمول (۲)، نمایش داد:

$$B_i = \{B_i^1, B_i^2, \dots, B_i^D\} \qquad (۱)$$

$$BOA = \ll B_1, B_2, \dots, B_N \gg \qquad (۲)$$

در اینجا هر راه‌حل یا پروانه $B_i$ یک بردار ویژگی است که مقدار آن دارای صفر و یک است و به ترتیب اگر یک مولفه بردار ویژگی صفر باشد نشان می‌دهد. در این رابطه، N اندازه جمعیت اولیه بردارهای ویژگی و $BOA$ نیز جمعیت اولیه از راه‌حل‌ها یا پروانه‌ها یه به عبارت بهتر بردارهای ویژگی در ارتباط با زمانبندی jobها است. در روش پیشنهادی هر بردار ویژگی بر روی مجموعه داده مرتبط با jobها نگاشت داده می‌شود و از این داده‌ها برای آموزش و یادگیری شبکه عصبی مصنوعی استفاده می‌شود.




**شکل ٥: فلوچارت روش پیشنهادی در انتخاب ویژگی و کاهش ابعاد**

یک شبکه عصبی مصنوعی با ضرب ورودی‌ها در اوزان لایه‌های پنهان و جمع آنها با مقادیر بایاس می‌تواند خروجی مورد نظر خود را تولید نماید و این خروجی را می‌توان به مانند تابع فرمول (۳)، مدلسازی نمود:

$$\bar{Y}_i = \sum w.x + b \qquad (۳)$$

در این رابطه، w ،x و b به ترتیب ورودی شبکه عصبی مصنوعی، اوزان لایه پنهان و بایاس لایه پنهان است و $\bar{Y}_i$ مقدار پیش بینی شبکه عصبی مصنوعی از زمان اجرای یک job است. اگر تعداد نمونه‌های مورد ارزیابی برابر m باشد آنگاه می‌توان متوسط خطا را مانند فرمول(۴)، برای ارزیابی هر بردار ویژگی job ها در هدوپ استفاده نمود:

$$Error = \frac{1}{m}.\sum_{i=1}^{m}(\bar{Y}_i - Y_i)^2 \qquad (۴)$$

در این رابطه، $Y_i$ برابر دسته واقعی یک job نظیر i از نظر زمان اجراء بوده و $\bar{Y}_i$ مقدار پیش بینی شده کلاس توسط شبکه عصبی مصنوعی آموزش داده شده توسط یک بردار ویژگی است و m تعداد نمونه‌های است که برای ارزیابی مورد استفاده قرار گرفته می‌شود. این تابع به تنهایی برای ارزیابیس بردارهای ویژگی می‌تواند استفاده شود اما یک تابع بهتر را می‌توان به صورت تابع فرمول (۵)، بیان نمود که متوسط خطای تشخیص نوع job و متوسط تعداد ویژگی انتخاب شده در آن را می‌توان در نظر گرفت:

$$CostFun_{Job} = \begin{cases} \min Error = \frac{1}{m}.\sum_{i=1}^{m}(\bar{Y}_i - Y_i)^2 \\ \min FS = \frac{SelectFeatures}{AllFeatures} \end{cases} \quad (۵)$$

تابع مورد نظر را می‌توان به صورت یک تابع واحد بازنویسی نمود و آن را در ضرایبی ضرب نمود که مجموعه آنها برابر ۱ است و این تابع را می‌توان به صورت فرمول (۶)، ارائه نمود:

$$CostFun_{Job} = 0.5 \times \frac{1}{m}\sum_{i=1}^{m}(\bar{Y}_i - \quad Y_i)^2\ 0.5 \times FS \quad (۶)$$

هدف نهایی کمینه نمودن این تابع با استفاده از بردار ویژگی بهینه job ها در هدوپ است و برای این منظور از الگوریتم بهینه‌سازی پروانه استفاده می‌شود. در ابتدا نیاز است که هر بردار ویژگی از جمعیت اولیه را در این تابع هدف قرار داد و آن را مورد ارزیابی قرار داده و بردار ویژگی بهینه‌ای که دارای کمترین خطا است به مانند فرمول (۷)، به عنوان پروانه بهینه انتخاب نمود:

$$best = min\{CostFun_{Job(B_1)}, CostFun_{Job(B_2)} \dots, CostFun\_Job(B_N)\} \quad (۷)$$

در این رابطه، $CostFun\_Job$ تابع هدف انتخاب ویژگی بوده و $N$ اندازه جمعیت پروانه‌ها، $B_i$ نیز یک پروانه یا بردار ویژگی بوده و $best$ نیز بردار ویژگی بهینه است که دارای حداقل خطای ممکن برای تشخیص نوع زمان یک job است و می‌توان به کمک این بردار ویژگی بهینه و به مانند الگوریتم بهینه‌سازی پروانه هر بردار ویژگی را به روزرسانی نمود. برای به روزرسانی بردارهای ویژگی مرتبط با job های قابل اجراء می‌توان در تکرار رابطه(۸) یا فرمول(۹)، را بر روی هر کدام از آنها اجرا نمود تا این مولفه‌ها به روزرسانی شوند:

$$B_i(t+1) = B_i(t) + \left(rand(0,1) \times best - \quad B_i(t)\right) \times f_i \quad (۸)$$

$$B_i(t+1) = B_i(t) + (rand(0,1) \times B_j(t) - \quad B_k(t)) \times f_i \quad (۹)$$

$B_i(t+1)$ موقعیت جدید بردار ویژگی $B_i(t)$ در تکرار t است و $B_k(t)$ و $B_j(t)$ موقعیت فعلی دو پروانه دلخواه جمعیت مانند j و k است و از طرفی $f_i$ میزان جذابیت یک پروانه است. می‌توان این رابطه را به صورت یک معادله واحد و مانند فرمول (۱۰)، بازنویسی نمود:

$$B_i(t+1) = \begin{cases} B_i(t) + (rand(0,1) \times best - B_i(t)) \times f_i & rand < 0.5 \\ B_i(t) + (rand(0,1) \times B_j(t) - B_k(t)) \times f_i & rand > 0.5 \end{cases} \quad (۱۰)$$

$f_i$ میزان جذابیت یک پروانه است و بر اساس فرمول (۱۱)، می‌توان آن را محاسبه نمود. در این رابطه، $S$ ضریب عددی جذب که برابر ۰.۰۱ فرض شده و p هم توان جذب نام دارد که برابر ۰.۱ در نظر گرفته می‌شود. می‌توان هر کدام از این رابطه‌ها را در یک تابع واحد قرار داد و مانند فرمول (۱۲) و (۱۳) آنها را فرموله نمود:

$$f_i = S \times CostFun\_Job(B_i)^p \quad (۱۱)$$

$$B_i(t+1) = B_i(t) + (rand(0,1) \times best - B_i(t)) \times S \times CostFun\_Job(B_i)^p \quad (۱۲)$$

$$B_i(t+1) = B_i(t) + (rand(0,1) \times B_j(t) - B_k(t)) \times S \times CostFun\_Job(B_i)^p \quad (۱۳)$$

با انجام روابط، یک بردار ویژگی مانند $B_i(t)$ به روزرسانی شده و به بردار ویژگی $B_i(t+1)$ تبدیل شده و این بردار ویزگی می‌تواند از حالت باینری خارج شده باشد و نیاز دارد که مقدار فیلدهای آن به صورت صفر و یک تبدیل شود. هر کدام از بردارهای ویژگی مرتبط با job ها که تحت تاثیر الگوریتم بهینه‌سازی پروانه قرار گرفته می‌شود به علت وجود روابط غیرباینری می‌توانند بردارهای ویژگی را از حالت باینری خارج نموده و آنها را پیوسته نمایند که در این حالت بردارهای مورد نظر برای انتخاب ویژگی کارایی نداشته و نیاز است که با استفاده از یک مجموعه توابع که تابع انتقال نامیده می‌شوند بتوان آنها را باینری نمود. دو تابع انتقال کاربردی برای انتخاب ویژگی وجود دارد که S-Shape و V-Shape نامیده می‌شوند و معادله این توابع تبدیل یا انتقال به ترتیب در فرمول (۱۴) و (۱۵)، تعریف شده است و می‌توان با استفاده از فرمول (۱۶) و (۱۷) به ترتیب و بر اساس توابع انتقال S-Shape و V-Shape هر کدام از بردارهای ویژگی را انتخاب نمود:

$$T(B_i) = \left|\frac{B_i}{\sqrt{a+B_i^2}}\right| \quad (۱۴)$$

$$T(B_i) = \frac{1}{1+e^{-a.B_i}} \quad (۱۵)$$

$$B_i = \begin{cases} \neg B_i & rand(0,1) < T(B_i) \\ B_i & rand(0,1) > T(B_i) \end{cases} \quad (۱۶)$$

$$B_i = \begin{cases} 0 & rand(0,1) < T(B_i) \\ 1 & rand(0,1) > T(B_i) \end{cases} \quad (۱۷)$$

در واقع با استفاده از توابع انتقال S-Shape و V-Shape می‌توان هر مولفه بردار ویژگی را مجدد صفر یا یک تنظیم نمود و به عنوان نمونه در تابع S-Shape اگر مقدار عدد تصادفی بین صفر و یک کمتر از نگاشت این توابع باشد مولفه ویژگی برابر یک می‌شود و در غیر اینصورت مقدار آن برابر یک می‌شود. شبکه کد روش پیشنهادی برای پیش بینی زمان jobها در هدوپ را می‌توان در شکل(۶)، مشاهده نمود. در فناوری هدوپ از تکنیک نگاشت و کاهش برای موازی‌سازی و اجرای jobها در سیستم‌های مختلف استفاده می‌شود در نگاشت و کاهش jobها ابتدا تقسیم شده و هر سیستم بخشی از jobها را انجام می‌دهد و سپس نتایج محاسبات در فرآیند کاهش با هم تلفیق شده و به خروجی ارسال می‌شود. در روش پیشنهادی برای انتخاب ویژگی با استفاده از الگوریتم بهینه‌سازی پروانه ارائه شد و بردارها ویژگی بهینه برای انتخاب ویژگی‌های موثر در پیش بینی زمان jobها در فناوری هدوپ مورد استفاده قرار گرفته شد و در مکانیزم پیشنهادی با انتخاب بردار ویژگی بهینه می‌توان از آن برای یادگیری شبکه عصبی مصنوعی در طبقه‌بندی jobها از نظر زمانی استفاده نمود. در روش پیشنهادی تعدادی از ویژگی‌های مهم مرتبط با jobهای که قرار است در هدوپ اجراء و زمانبندی شوند در قالب یک بردار ویژگی در نظر گرفته شده شده و هر بردار ویژگی مرتبط با jobها یا پروسه‌ها به صورت یک پروانه و یک عضو الگوریتم بهینه‌سازی پروانه فرض شده سپس به کمک الگوریتم بهینه‌سازی پروانه بردارهای ویژگی بهینه انتخاب شده تا خطای پیش بینی و طبقه‌بندی jobها از نظر زمانی کاهش داده شود. در روش پیشنهادی برای ارزیابی هر بردار ویژگی در پیش بینی نوع زمان و کلاس زمانی هر داده از شبکه عصبی مصنوعی چند لایه استفاده می‌شود. شبکه عصبی مصنوعی چند لایه در روش پیشنهادی به عنوان یک مکانیزم یادگیری استفاده شده و بردارهای ویژگی با شبکه عصبی مصنوعی مورد ارزیابی قرار گرفته شده تا مشخص شود کدام بردار ویژگی دارای کارایی بیشتری است.

## ٤-پیاده سازی و تحلیل

برای ارزیابی روش پیشنهادی از مجموعه داده Aloja [21] استفاده شده است و تعدادی از رکوردها و نمونه‌های آن مورد پیش پردازش قرار گرفته شده است و هر سطر آن یک نمونه است که دارای تعدادی ویژگی اولیه است و این ویژگی‌ها برابر ۵۴ عدد بوده و یک ویژگی خروجی نیز برای آن در نظر گرفته شده است که می‌تواند حالت زمانی را نشان دهد که دو حالت در اینجا در نظر گرفته شده است و حالتها با صفر و یک کدگذاری شده که به ترتیب زمان اجرای اندک و زیاد را نشان می‌دهد و هر ردیف و نمونه بکار رفته در این مجموعه داده ویژگی‌های یک job است که در هدوپ اجراء می‌شود.

برای تجزیه و تحلیل روش پیشنهادی می‌توان از دو شاخص MSE و RMSE در کنار شاخص متوسط تابع هدف انتخاب ویژگی استفاده نمود. هر یک از این شاخص‌ها به ترتیب در فرمول (۱۸) و (۱۹) فرموله و مدلسازی شده‌اند و رابطه تابع هدف انتخاب ویژگی نیز در فصل سوم ارائه شده و از ذکر مجدد آن خودداری می‌شود:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n} \quad (۱۸)$$

**Butterfly Optimization algorithm for predict the execution time of a job**

*Define feature vector* $B_i = \{B_i^1, B_i^2, \dots, B_i^D\}$ *for predict the execution time of a job*
*Set objective function :* $CostFun\_Job = 0.5 \times \frac{1}{m}\sum_{i=1}^{m}(\bar{Y}_i - Y_i)^2 + 5 \times FS$
*Create initial population of Job feature vector* $BOA = \ll B_1, B_2, \dots, B_N \gg$
*Normalization dataset:*
$X' = \frac{X - min}{max - min}(b - a) + a$
*Divide data into training and test data*
*Creating initial population of feature vectors*
**for** *i=1 to n*
  **if** *rand<0.5*
    $B_i$=1
  **else**
    $B_i$=0
  **end if**
  *Classification Model Training by ANN*
  *Evaluation* $B_i$ *by* $CostFun_{Job} = 0.5 \times \frac{1}{m}\sum_{i=1}^{m}(\bar{Y}_i - Y_i)^2\ 0.5 \times FS$
**Update** *best feature vectors*
**end for**
**while** *(t <= MaxIt)*
**for** *i=1 to n*
 **if** *rand<0.*
$$B_i(t+1) = B_i(t) + (rand(0,1) \times best - B_i(t)) \times S \times CostFun\_Job(B_i)^p$$
 **else**
$$B_i(t+1) = B_i(t) + (rand(0,1) \times B_j(t) - B_k(t)) \times S \times CostFun\_Job(B_i)^p$$
 **end if**
*Use conversion function:*
 **if** $rand(0,1) < T(B_i)$
$$B_i(t+1) = \neg B_i(t);$$
 **else**
$$B_i(t+1) = B_i(t)$$
**end if**
**Update** *best feature vectors*
**end for**
 **end while**
*BestSol vector for Classification and predict the execution time of a job*

**شکل (٦): شبه کد روش پیشنهادی**

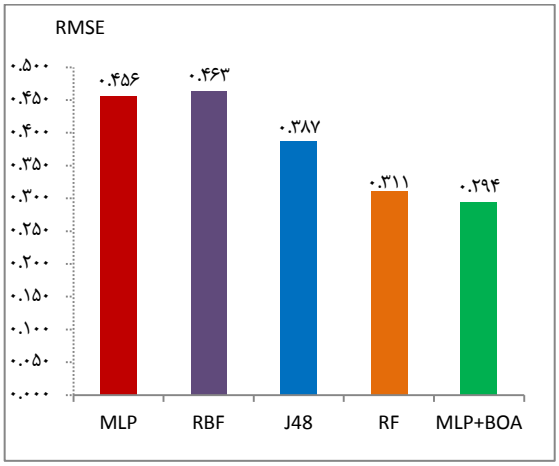
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}} \quad (۱۹)$$

در این رابطه، $y_i$ شماره واقعی کلاس یک نمونه job از نظر زمان اجراء و $\bar{y}_i$ شماره کلاس تخمینی نمونه مورد نظر در روش پیشنهادی است و n تعداد نمونه‌ها یا job های بکار رفته برای ارزیابی است که برای ارزیابی شبکه عصبی مصنوعی چند لایه برای طبقه‌بندی نوع job ها استفاده می‌شود. برای تجزیه و تحلیل الگوریتم پیشنهادی از محیط برنامه‌نویسی متلب ۲۰۱۸ استفاده شده است و از ۷۰٪ نمونه‌ها به عنوان نمونه آموزشی و از ۳۰٪ دیگر به عنوان نمونه آزمون استفاده می‌شود. در پیاده‌سازی‌ها از یک شبکه عصبی مصنوعی ۲ لایه استفاده شده که هر لایه آن ۴ نورون دارد. برای ارزیابی روش پیشنهادی در تشخیص نوع کلاس زمانبندی job ها در هدوپ می‌توان از شاخص خطای RMSE که یک شاخص مهم در ارزیابی‌ها است استفاده نمود. در اینجا خطای RMSE نشان می‌دهد که در پیش بینی نوع زمان job چقدر خطا انجام شده است و این خطا چه ارتباطی به اندازه جمعیت دارد زیرا اندازه جمعیت یک فاکتور مهم در الگوریتم‌های بهینه‌سازی و فرابتکاری است و افزایش آن می‌تواند در دقت آنها موثر باشد زیرا شانس یافتن و جستجوی بهتر با افزایش اندازه جمعیت افزایش خواهد یافت. با افزایش جمعیت این خطا در حال کاهش است زیرا شانس یافتن بردار ویژگی بهینه افزایش خواهد یافت و این موضوع نیز باعث کاهش یافتن خطای تشخیص و پیش بینی job های قابل اجراء می‌گردد. تحلیل و ارزیابی نمودار نشان می‌دهد که در جمعیت ۵، ۱۰، ۱۵ و ۲۰ به ترتیب متوسط مجذور خطا برابر ۰.۲۹۴، ۰.۲۷۲، ۰.۲۵۴ و ۰.۲۱۸ است و با افزایش جمعیت از ۵ به ۲۰ خطای متوسط در تشخیص و پیش بینی زمان اجرای job ها در هدوپ در حدود ۲۵.۸۵٪ کاهش پیدا می‌نماید. با توجه به خروجی‌های وکا و روش پیشنهادی می‌توان متوسط خطای هر روش را مطابق جدول(۱) و نمودار شکل(۷)، با هم مقایسه و ارزیابی نمود و بر حسب شاخص خطا به ارزیابی هر یک از روش‌ها پرداخت. با توجه به آزمایشات و خروجی‌ها می‌توان نتیجه گرفت که خطای پیش بینی شبکه عصبی چند لایه، شبکه عصبی بازگشتی، درخت تصمیم-گیری و جنگل تصادفی و الگوریتم بهینه سازی پروانه به ترتیب برابر ۰.۴۵۶، ۰.۴۶۳، ۰.۳۸۷، ۰.۳۱۱ و ۰.۲۹۴ است و در بین این روش‌ها خطای روش پیشنهادی از آنها کمتر است و در مرتبه دوم نیز جنگل تصادفی قرار دارد و سپس درخت تصمیم‌گیری، شبکه عصبی مصنوعی چند لایه و شبکه عصبی بازگشتی قرار دارد.

تجزیه و تحلیل روش پیشنهادی نشان می‌دهد خطای روش پیشنهادی نسبت به شبکه عصبی مصنوعی چند لایه در حالتی که جمعیت برابر ۵ و تعداد تکرار برابر ۳۰ است در حدود ۱.۵۵ برابر کمتر شده است و این کاهش به آن دلیل است که الگوریتم پروانه بردارهای ویژگی را بهینه انتخاب می-نماید. آزمایشات نشان می‌دهد که روش پیشنهادی نسبت به الگوریتم جنگل تصادفی نیز در حدود ۵.۴۶٪ نیز خطای کمتری دارد و اگر اندازه جمعیت برابر

**جدول(۱): مقایسه خطای الگوریتم پیشنهادی و سایر الگوریتم ها**

| روش | متوسط RMSE |
|---|---|
| MLP | ۰.۴۵۶ |
| RBF | ۰.۴۶۳ |
| J48 | ۰.۳۸۷ |
| RF | ۰.۳۱۱ |
| MLP+BOA | ۰.۲۹۴ |




**شکل(۷): مقایسه متوسط مجذور خطای تشخیص زمان jobها در روش پیشنهادی و الگوریتم های دیگر**

۲۰ در نظر گرفته شود خطای روش پیشنهادی در حدود ۰.۲۱۸ است و خطای جنگل تصادفی در حدود ۰.۳۱۱ است و این کاهش خطا نسبت به جنگل تصادفی در این حالت برابر ۲۹.۹۰٪ کاهش خواهد داشت.

## ۵-نتیجه گیری

در پیش بینی زمان اجرای یک job می‌توان از روش‌های یادگیری ماشین مانند شبکه عصبی مصنوعی استفاده نمود و ویژگی‌های هر job را به عنوان ورودی شبکه عصبی مصنوعی در نظر گرفت تا بتوان زمان اجرای یک job را پیش بینی نمود. در اینجا می‌توان دسته‌ها یا کلاسهایی را ایجاد نمود و نوع job را بر اساس زمان اجرایش در این کلاسها قرار داد و در اینجا هر job برای اجراء و پیش بینی زمان اجرای آن نیاز دارد ویژگی‌های مناسب و دقیقی را به شبکه عصبی مصنوعی ارائه دهد تا به عنوان ورودی دقیق بتوان از آنها استفاده نمود. چالش مهم شبکه عصبی مصنوعی در پیش بینی زمان اجرای job ها توسط شبکه عصبی مصنوعی در آن است که نیاز می‌باشد یادگیری و پیش بینی فقط بروی ویژگی‌های مهم انجام شود. در روش پیشنهادی برای افزایش دقت شبکه عصبی مصنوعی در پیش بینی نوع job از نظر زمان اجراء از مکانیزم انتخاب ویژگی توسط الگوریتم‌های فرابتکاری استفاده می-

شود و در اینجا از الگوریتم بهینه‌سازی پروانه به عنوان الگوریتم انتخاب کننده ویژگی در کنار شبکه عصبی مصنوعی استفاده می‌شود. تجزیه و تحلیل الگوریتم پیشنهادی برای کاهش دادن خطای تشخیص زمان اجرای job ها در هدوپ توسط روش پیشنهادی نتایج ذیل را نشان می‌دهد مقدار تابع هدف انتخاب ویژگی در الگوریتم بهینه‌سازی پروانه بر حسب تکرار یک روند نزولی و کاهشی است. نتایج نشان می دهد انتخاب ویژگی با الگوریتم بهینه‌سازی پروانه باعث کاهش متوسط خطا شده است از طرفی با افزایش اندازه جمعیت در الگوریتم بهینه‌سازی پروانه و تابع انتخاب ویژگی، خطای پیش بینی زمان اجرای job ها به دلیل جستجوی بردار ویژگی بهینه کاهش یافته است. همانطور که نتایج نشان می‌دهد با تکرار الگوریتم بهینه‌سازی پروانه شانس یافتن بردار ویژگی بهتر افزایش یافته و احتمال کاهش خطای پیش بینی زمان اجرای job ها بیشتر می‌شود. آزمایشات مختلفی که انجام شده نشان می‌دهد در بین روش‌های یادگیری و طبقه‌بندی، خطای روش پیشنهادی کمتر است و روش های دیگر همانند جنگل تصادفی، درخت تصمیم‌گیری، شبکه عصبی مصنوعی چند لایه و شبکه عصبی بازگشتی در جایگاه بعدی قرار دارد. از پژوهش‌های آتی می‌توان به استفاده از الگوریتم‌های فراابتکاری دیگر نظیر الگوریتم بهینه‌سازی شاهین برای انتخاب ویژگی، ترکیب الگوریتم پیشنهادی با روشهای زمانبندی در هدوپ برای زمانبندی job ها اشاره نمود.

## مراجع


[1] Basha, S. A. K., Basha, S. M., Vincent, D. R., & Rajput, D. S. (2019). Challenges in Storing and Processing Big Data Using Hadoop and Spark. In *Deep Learning and Parallel Computing Environment for Bioengineering Systems* (pp. 179-187). Academic Press.

[2] Goyal, M. (2019, February). Demonetization-Twitter Data Analysis using Big Data & Hadoop. In *2019 Amity International Conference on Artificial Intelligence (AICAI)* (pp. 156-158). IEEE.

[3] Amalina, F., Hashem, I. A. T., Azizul, Z. H., Fong, A. T., Firdaus, A., Imran, M., & Anuar, N. B. (2019). Blending Big Data Analytics: Review on Challenges and a Recent Study. *IEEE Access*.

[4] Wang, H., Ma, S., & Dai, H. N. (2019). A rhombic dodecahedron topology for human-centric banking big data. *IEEE Transactions on Computational Social Systems*, *6*(5), 1095-1105.

[5] Mo, Y. (2019). A Data Security Storage Method for IoT Under Hadoop Cloud Computing Platform. *International Journal of Wireless Information Networks*, 1-6.

[6] Ghani, N. A., Hamid, S., Hashem, I. A. T., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior*, *101*, 417-428.

[7] Thakor, D., & Patel, B. (2019). PDLB: An Effective Prediction-Based Dynamic Load Balancing Algorithm for Clustered Heterogeneous Computational Environment. In Recent Findings in Intelligent Computing Techniques (pp. 593-603). Springer, Singapore.

[8] Devi, M. A., Ravi, S., Vaishnavi, J., & Punitha, S. (2016). Classification of cervical cancer using artificial neural networks. *Procedia Computer Science*, *89*, 465-472.

[9] Moridnejad, A., Abdollahi, H., Alavipanah, S. K., Samani, J. M. V., Moridnejad, O., & Karimi, N. (2015). Applying artificial neural networks to estimate suspended sediment concentrations along the southern coast of the Caspian Sea using MODIS images. Arabian Journal of Geosciences, 8(2), 891-901.

[10] Arora, S., & Anand, P. (2019). Binary butterfly optimization approaches for feature selection. *Expert Systems with Applications*, *116*, 147-160.

[11] Li, Z., Su, D., Zhu, H., Li, W., Zhang, F., & Li, R. (2017). A Fast Synthetic Aperture Radar Raw Data Simulation Using Cloud Computing. *Sensors*, *17*(1), 113.

[12] Wang, L., Wang, Y., & Xie, Y. (2015). Implementation of a parallel algorithm based on a spark cloud computing platform. *Algorithms*, *8*(3), 407-414.

[13] Capra, M., Peloso, R., Masera, G., Ruo Roch, M., & Martina, M. (2019). Edge computing: A survey on the hardware requirements in the internet of things world. Future Internet, 11(4), 100

[14] Sonkar, S. K., & Kharat, M. U. (2019). Load prediction analysis based on virtual machine execution time using optimal sequencing algorithm in cloud federated environment. International Journal of Information Technology, 11(2), 265-275.

[15] Alazzam, H., Alhenawi, E., & Al-Sayyed, R. (2019). A hybrid job scheduling algorithm based on Tabu and Harmony search algorithms. The Journal of Supercomputing, 1-18.

[16] Zhu, Z., & Fan, P. (2019). Machine Learning Based Prediction and Classification of Computational Jobs in Cloud Computing Centers. arXiv preprint arXiv:1903.03759.

[17] Chen, J., Li, K., Rong, H., Bilal, K., Li, K., & Philip, S. Y. (2019). A periodicity-based parallel time series prediction algorithm in cloud computing environments. Information Sciences, 496, 506-537.

[18] Hu, Z., Li, B., Qin, Z., & Goh, R. S. M. (2017, June). Job scheduling without prior information in big data processing systems. In 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS) (pp. 572-582). IEEE.

[19] Liu, Y., Zeng, Y., & Piao, X. (2016, August). High-responsive scheduling with mapreduce performance prediction on hadoop yarn. In 2016 IEEE 22nd International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA) (pp. 238-247). IEEE.

[20] Johannessen, R., Yazidi, A., & Feng, B. (2017, April). Hadoop MapReduce scheduling paradigms. In 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 175-179). IEEE.

[21] Gautam, J. V., Prajapati, H. B., Dabhi, V. K., & Chaudhary, S. (2015, March). A survey on job scheduling algorithms in Big data processing. In 2015 IEEE International Conference on Electrical, Computer

and Communication Technologies (ICECCT) (pp. 1-11). IEEE.
[22] https://aloja.bsc.es/
[23] https://github.com/hadi-operator/Butter-y-Optimization-algorithm-for-predict-the-execution-time-of-a-job/blob/main/CODE