Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 8, Number 1, Spring 2019, pp. 31-40
Received: 01/11/2017, Revised: 08/04/2018, Accepted: 02/03/2019

# Instance Reduction Using Multi-Objective Chaotic Particle Swarm Optimization Algorithm

Javad Hamidzadeh[1*], Nilufar Kashefi[2*] and Mona Moradi[3]
1*- Faculty of Computer Engineering and Information Technology, Sadjad University Of Technology, Mashhad, Iran.
2- Faculty of Computer Engineering and Information Technology, Sadjad University Of Technology, Mashhad, Iran.
3- Faculty of Computer Engineering and Information Technology, Sadjad University Of Technology, Mashhad, Iran.
[1*]J_hamidzadeh@sadjad.ac.ir, [2]N.kashefi111@sadjad.ac.ir, and [3]Mmoradi@semnan.ac.ir

Corresponding author address: Javad Hamidzadeh, Faculty of Computer Engineering and Information Technology, Sadjad University Of Technology, Mashhad, Iran, Post Code : 9188148848.

**Abstract-** Today, it is important to reduce the original huge data set to a manageable volume. Also, unbalanced data distribution between different classes is a serious challenge in data mining. In the proposed method, the instance reduction problem is considered as a multi-objective problem, which can perform well by considering the two contradict criteria, classification accuracy and reduction rate of instances. The multi-objective problem is solved using the chaotic particle swarm optimization algorithm. The distance-based decision classifier has the task of distinguishing the maintenance or deletion of test instances. Creating and maintaining balances for different types of data distribution is the main goal of the proposed method. The results of the experiments have been compared with the state-of-the-art methods, which show superiority of the proposed method in terms of classification accuracy and reduction percentage.

**Keywords-** Instance Reduction, Multi Objective Particle Swarm Optimization Algorithm, Unbalanced Data, Reduction Rate, Accuracy Rate, Chaotic Functions.

مجله علمی پژوهشی رایانش نرم و فن‌آوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۸، شماره ۱، بهار ۱۳۹۸ ، صفحه ۴۰-۳۱
دریافت: ۹۵/۱۰/۲۲، بازنگری: ۹۷/۰۵/۱۳، پذیرش: ۹۷/۱۱/۱۴

# کاهش نمونه در داده‌ها به کمک الگوریتم بهینه‌سازی چندهدفه ازدحام ذرات آشوبی

جواد حمیدزاده*[۱]، نیلوفر کاشفی[۲]، منا مرادی[۳]
*۱- دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی سجاد، مشهد، ایران.
۲- دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی سجاد، مشهد، ایران.
۳- دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی سجاد، مشهد، ایران.
[1*]J_hamidzadeh@sadjad.ac.ir, [2]N.kashefi111@sadjad.ac.ir, and [3]Mmoradi@semnan.ac.ir

* نشانی نویسنده مسئول: جواد حمیدزاده، مشهد، بلوار جلال آل احمد، جلال آل احمد ۶۴، دانشگاه صنعتی سجاد، کد پستی: ۹۱۸۸۱۴۸۸۴۸

**چکیده- امروزه با توجه به حجم وسیع داده‌ها، مسئله کاهش نمونه حائز اهمیت است. همچنین عدم وجود توازن در توزیع داده‌ها بین کلاس‌های مختلف یک چالش جدی در داده‌کاوی است. در روش پیشنهادی، مسئله کاهش نمونه به‌عنوان مسئله‌ی چندهدفه در نظر گرفته شده است که توانسته است با درنظر گرفتن دو معیار متضاد صحت طبقه‌بندی و نرخ کاهش نمونه‌ها و همچنین توجه به معیارهای مربوط به داده‌های نامتوازن عملکرد خوبی داشته باشد. ایجاد و حفظ توازن در انواع مختلف توزیع داده مهمترین هدف روش پیشنهادی است. مسئله چندهدفه طراحی شده با استفاده از الگوریتم بهینه‌سازی ازدحام ذرات آشوبی حل شده است. سطح تصمیم مبتنی بر فاصله در روش پیشنهادی، وظیفه تشخیص حفظ و یا حذف نمونه‌های آزمایشی را دارد. نتایج آزمایشات نشان‌دهنده‌ی برتری روش پیشنهادی از نظر دقت و صحت طبقه‌بندی و نرخ کاهش داده‌ها نسبت به روش‌های مرز دانش است.**

**واژه‌های کلیدی:** کاهش نمونه، الگوریتم بهینه سازی چندهدفه ازدحام ذرات، داده‌های نامتوازن، نرخ صحت، نرخ کاهش نمونه‌ها، توابع آشوب.

## ۱- مقدمه

داده‌کاوی، فرایند طبقه‌بندی داده‌ها و آشکارسازی اطلاعات مرتبط با هم می‌باشد [۱]. اغلب، حجم داده‌های مورد بررسی بسیار زیاد است و معمولا شامل نویز، مقادیر اضافی و یا تکراری می‌باشند. پردازش این قبیل داده‌ها مشکلاتی مانند کمبود حافظه، دقت پایین طبقه‌بندی و کاهش سرعت را به‌همراه دارد. یکی از روش‌های مقابله با این قبیل مشکلات، کاهش نمونه است که در آن زیرمجموعه‌ای از داده‌های اصلی انتخاب می‌شود که فاقد نویز، مقادیر اضافی و یا تکراری است و ما را به دقتی از طبقه‌بندی می‌رساند که با مجموعه داده‌های اصلی می‌رسیدیم، ضمن این که حافظه‌ی مورد نیاز و زمان اجرای الگوریتم کاهش می‌یابد.

در عمل کاهش نمونه معمولاً با مسئله توازن بین اندازه مجموعه آموزشی و کیفیت طبقه‌بندی مواجه می‌شویم. الگوریتم موفق معمولاً اندازه مجموعه آموزشی را کاهش می‌دهد بدون این‌که صحت کلی طبقه‌بند کاهش چشم‌گیری داشته باشد. چالش اصلی فرآیند کاهش نمونه، انتخاب معیار مناسب برای حذف نمونه‌هاست.

در بسیاری از مسائل داده‌کاوی توزیع غیریکنواخت نمونه‌ها در کلاس‌های مختلف منجر به تشکیل کلاس‌های نامتوازن می‌شود. به کلاس‌هایی که تعداد داده‌های بسیار زیادی دارند کلاس اکثریت و به کلاس‌هایی که داده‌های بسیار کمی را شامل می‌شوند کلاس اقلیت می‌گویند. معمولا کلاس اقلیت برای ما از اهمیت بالایی برخوردار است و در بیشتر مسائل شناسایی داده‌های این کلاس‌ها هدف اصلی است. مواجهه با مسائل کلاس‌های نامتوازن کار بسیار

مشکلی است و از طرفی بیشتر مسائل دنیای واقعی کلاس‌هایی با توزیع غیریکنواخت دارند. لذا در ارائه یک روش کاهش نمونه، نحوه توزیع داده‌ها در بین کلاس‌ها چالشی است که باید در نظر گرفته شود [۲].

موضوع مهم در کلاس‌های نامتوازن این است که در نظر گرفتن معیار صحت طبقه‌بندی به‌تنهایی کافی نیست، زیرا طبقه‌بندی نادرست نمونه‌های کلاس اقلیت تأثیر کمی در نرخ صحت دارد. به این معنی که نرخ صحت بالا می‌تواند تنها از نمونه-های کلاس اکثریت ناشی شود. لذا در مواجهه با مسائلی که کلاس-های نامتوازن دارند، اغلب معیارهای دیگری نیز در نظر گرفته می-شود. روش کاهش نمونه قابل‌قبول باید چند معیار نظیر دقت طبقه‌بندی و میزان کاهش نمونه‌ها را درنظر بگیرد. اگر روشی میزان قابل توجهی کاهش در تعداد داده اما دقت کم طبقه‌بندی و یا بالعکس را داشته باشد، روشی کاربردی تلقی نمی‌شود. از این‌رو، مسئله‌ی کاهش نمونه را می‌توان به‌عنوان یک مسئله‌ی بهینه‌سازی چندهدفه با دو معیار متضاد در نظر گرفت.

مسئله‌ی کاهش نمونه، مسئله‌ای چندهدفه با فضای راه‌حل بسیار بزرگ است. در این مسئله، مجموعه‌ی بسیار بزرگی از راه‌حل‌ها وجود دارد که هر راه‌حل زیرمجموعه‌ای از نمونه‌ها می‌باشد. اگر فرض کنیم تعداد نمونه‌ها $n$ باشد، $2^n$ زیرمجموعه‌ی ممکن از نمونه‌ها وجود دارد پس می‌توان نتیجه‌گیری کرد که مسئله‌ی کاهش نمونه به جستجوی سراسری فضای مسئله نیاز دارد. الگوریتم‌های تکاملی از جمله روش‌هایی هستند که قابلیت بالایی در جستجوی سراسری با سرعت و دقت قابل قبول دارند و به نتایج بسیار خوبی در حوزه‌ی کاهش نمونه دست یافته‌اند [۳] . الگوریتم بهینه‌سازی ازدحام ذرات یکی از روش‌های تکاملی کاربردی می‌باشد که نسبت به برخی روش‌های معروف تکاملی دیگر از قبیل الگوریتم ژنتیک، هزینه محاسباتی کمتر و سرعت همگرایی بالاتری دارد [۴-۵]. این الگوریتم و توسعه‌های آن به‌دلیل قابلیت جستجوی سراسری فضای مسئله با سرعت مناسب می‌توانند به‌عنوان یک روش کابردی در حوزه‌ی کاهش نمونه به‌کار روند [۶-۸].

در مقاله حاضر روشی برای کاهش نمونه پیشنهاد شده است که می‌تواند با ایجاد و حفظ توازن در مجموعه داده‌های نامتوازن و متوازن عملکرد قابل قبولی داشته باشد. برای حل مشکل موجود در کلاس‌های نامتوازن، علاوه بر معیارهای نرخ کاهش و نرخ صحت، معیارهای دیگری نظیر G-mean [۹]، که از جمله مهمترین معیارهای داده‌های نامتوازن است استفاده شده است. الگوریتم پیشنهادی طوری طراحی شده است که بتواند داده‌های زیادی را از مجموعه داده اصلی با حفظ دقت، صحت و توازن در طبقه‌بند حذف نماید اما این اهداف با یدیگر در تضاد هستند زیرا کاهش شدید نمونه تاثیر نامطلوبی بر دقت و صحت طبقه‌بند می‌گذارد بنابراین ارائه راه‌کاری برای تعدیل اهداف متضاد ضروری به نظر می‌رسد.

در روش پیشنهادی، از روش بهینه‌سازی چندهدفه ازدحام ذرات به‌منظور جستجوی سراسری فضای مسئله با در نظر گرفتن اهداف متضاد استفاده گردیده و از جواب‌های نامغلوب بدست آمده از بهینه‌ی پارتو به‌عنوان مجموعه داده کاهش یافته استفاده شده است. به‌منظور کنترل سرعت همگرایی الگوریتم و جلوگیری از به دام افتادن در بهینه‌های محلی، اعداد تصادفی به کمک توابع آشوب تولید می‌شوند. تشخیص حفظ/حذف نمونه توسط سطح تصمیم مبتنی بر فاصله انجام می‌شود. برقراری توازن در توزیع داده‌ها پس از اجرای کاهش نمونه در مجموعه داده‌هایی با توزیع اولیه متوازن/نامتوازن هدف روش پیشنهادی است. تاکنون از این ایده به منظور کاهش نمونه استفاده نشده است.

در ادامه‌ی مقاله، در بخش دوم، مروری بر کارهای پیشین در حوزه‌ی کاهش نمونه انجام شده است. در بخش سوم، مفاهیم مقدماتی مرتبط با روش پیشنهادی توضیح داده شده‌اند. روش پیشنهادی در بخش چهارم معرفی می‌گردد. نتایج آزمایش‌ها و تفسیر آنها در بخش پنجم نشان داده شده است و سپس نتیجه‌گیری و کارهای آینده در بخش ششم بیان شده‌اند.

## ۲- مروری بر کارهای پیشین

تاکنون تحقیقات زیادی در زمینه‌ی کاهش نمونه انجام شده است. روش‌های مبتنی بر قوانین ویرایش نزدیک‌ترین همسایه نظیر قانون نزدیک‌ترین همسایه خلاصه شده [۱۰]، قانون نزدیک‌ترین همسایه کاهش یافته [۱۱] و قانون نزدیک‌ترین همسایه [۱۲] از رویکردهای مطرح در زمینه کاهش نمونه هستند. گرچه قانون نزدیک‌ترین همسایه خلاصه شده طبقه بندی صحیح نمونه‌های مجموعه را تضمین می‌کند اما ایراد آن حساسیت نسبت به نویز است. قانون نزدیک‌ترین همسایه کاهش‌یافته توانایی حذف نمونه‌های نویزی و داخلی و حفظ نقاط مرزی را دارد. قوانین ویرایش نزدیک‌ترین همسایه علاوه بر حذف نمونه‌های نویزی، مرزهای تصمیم‌گیری یکنواخت‌تری به‌جا می‌گذارد، همچنین تمام نقاط داخلی را حفظ می‌کند. ایراد این روش این است که حافظه را بهینه نمی‌کند.

در روش نمونه تعمیم‌یافته تودرتو (NGE) [۱۳] نمونه‌های آموزشی به درون ابرمستطیل‌ها تعمیم یافته‌اند و نمونه‌ی ورودی با توجه به کلاس نزدیک‌ترین ابرمستطیل آن طبقه‌بندی می‌شود.

روش LMIRA [۱۴]، نمونه‌ها را با حذف نمونه‌های غیرمرزی کاهش می‌دهد و مرزی بزرگ برای جداسازی کلاس‌ها ایجاد می‌کند. روش IRAHC [۱۵]، روشی مبتنی بر خوشه‌بندی ابرمستطیل می‌باشد. ابرمستطیل، مستطیلی است $n$-بعدی که توسط نقاط بیشینه و کمینه و تابع فاصله‌ی متناظر تعریف می‌شود.

روش‌های [۱۷-۱۶-۱] از الگوریتم ژنتیک برای کاهش نمونه استفاده کرده‌اند و نشان داده‌اند که الگوریتم‌های تکاملی می‌توانند نتایج بهتری نسبت به روش‌های سنتی یا غیرتکاملی کسب کنند. روش [۱۸] از الگوریتم بهینه‌سازی گروه مورچگان برای کاهش نمونه استفاده کرده است. روش‌های [۲۱-۱۹] از ترکیب‌های مختلف انتخاب نمونه و ویژگی و وزن‌دهی نمونه و ویژگی برای کاهش نمونه استفاده کرده‌اند. در [۲۲] روشی به کمک الگوریتم‌های تلفیقی جهت کاهش داده‌ها به کار گرفته شده است که در آن دو روش ENN و CNN به وسیله‌ی الگوریتم تلفیقی، ترکیب شده‌اند. در چارچوب روش‌های تلفیقی، مجموعه‌ای از طبقه‌بندها داریم که هر طبقه‌بند یک نگاشت یک بردار نمونه به مجموعه‌ی برچسب‌هاست. روش ENN تمرکزش بر دقت کلی طبقه‌بندی و روش CNN تمرکزش بر روی نرخ کاهش داده‌ها می‌باشد. لذا استفاده‌ی همزمان دو روش، باعث برآورده ساختن هر دو معیار می‌شود. در [۲۳] نیز عملیات کاهش نمونه به‌کمک روش‌های تلفیقی صورت گرفته شده است و نتایج آن بهتر از استفاده از یک طبقه‌بند به‌تنهایی می‌باشد. همچنین در این پژوهش ثابت شده که روش‌های تلفیقی در حضور نویز نیز نتایج قابل قبولی دارند. روش [۲۴] به کاهش نمونه در داده‌های حجیم می پردازد. ایده‌ی آن، استفاده از نگاشت کاهش است که به‌دلیل عملکرد موازی باعث افزایش سرعت آنالیز داده‌ها می‌شود. برخی از پژوهش‌ها نظیر [۲۵]، از روش وزن‌دهی به نمونه‌ها استفاده کرده‌اند.

## ۳- مفاهیم مقدماتی

برای درک بهتر روش پیشنهادی، در این بخش الگوریتم بهینه سازی ازدحام ذرات و بهینه پارتو معرفی می‌گردند.

### ۳-۱- الگوریتم بهینه سازی ازدحام ذرات

الگوریتم بهینه‌سازی ازدحام ذرات[۱] PSO [۴] الهام گرفته از رفتارهای اجتماعی حیوانات می‌باشد. این الگوریتم با جمعیتی از جواب‌های تصادفی (ذرات) آغاز می‌شود. در طول اجرا، ذرات با استفاده از تجربیات قبلی خود و سایر ذرات از بهترین مکان‌هایی که تاکنون در آن‌ها قرار داشته‌اند به سمت بهینه محلی خود یا به سمت بهترین ذره گروه حرکت می‌کنند [۴]. هر ذره شامل دو فاکتور سرعت $v$ و موقعیت $x$ در فضای $d$-بعدی است و از روابط زیر محاسبه می‌گردد:

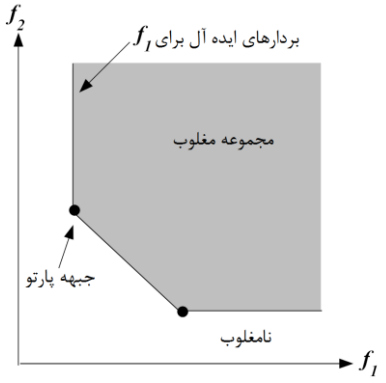
$$v_{id}(t+1) = wv_{id}(t) + c_1 r_1\left(p_{id}(t) - x_{id}(t)\right) + c_2 r_2\left(p_g(t) - x_{id}(t)\right) \quad (۱)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad i = 1 \dots N \quad (۲)$$

در روابط بالا، $N$ تعداد ذرات می‌باشد. $w$ فاکتور اینرسی وزن و همچنین اعداد تصادفی $r_1$ و $r_2$ از اعداد بازه [۱-۰] هستند. $d$ تعداد ابعاد فضای جستجو است. مقادیر $v$ و $x$ به‌ترتیب بیانگر سرعت و مکان ذرات می‌باشند. $p_i$ بهترین موقعیت ذره $i$ و $p_g$ بهترین موقعیت پیدا شده در کل ذرات می‌باشد. $c_1$ و $c_2$ ثابت‌هایی هستند که سرعت ذره به سوی $p_i$ و $p_g$ را تغییر می‌دهند و از بازه [۱-۰] انتخاب می‌شود.

### ۳-۲- بهینه پارتو

در بسیاری از مسائل بهینه‌سازی، درصدد بهینه کردن چند هدف (معمولا در متضاد) به‌طور همزمان می‌باشیم، لذا تابع بهینه تعریف شده باید تمام معیارها را مورد توجه قرار دهد. ماهیت مسائل چندهدفه به‌گونه‌ای است که مقایسه‌ی جواب‌ها به‌سادگی انجام نمی‌شود. جبهه پارتو تکنیکی است که با دسته‌بندی جواب‌ها به دو دسته مغلوب و نامغلوب جواب بهینه را به‌دست می‌آورد. جبهه پارتو می‌تواند به عنوان مجموعه‌ای از جواب‌های نامغلوب تعریف شود. یک نقطه پاسخ بهینه پارتو یا پاسخ نامغلوب است اگر هیچ بردار شدنی (از متغیرهای تصمیم در فضای جستجو) وجود نداشته باشد به طوری که برخی از اهداف را کاهش دهد بدون اینکه حداقل باعث کاهش همزمان یکی دیگر از اهداف نیز شود (شکل ۱). با در نظر گرفتن میزان آگاهی ما از مسئله و شرایط مرزی پاسخ بهینه نهایی از مجموعه پاسخ‌های جبهه پارتو انتخاب می‌شود [۲۶].




**شکل ۱: مجموعه جواب‌های مغلوب و جبهه پارتو در فضای دو هدفه.**

## ۴- روش پیشنهادی

روش پیشنهادی با هدف کاهش نمونه و با تاکید بر ایجاد و حفظ توازن در مجموعه داده‌های با توزیع نا متوازن و متوازن طراحی شده است. جزئیات روش پیشنهادی بدین شرح است:

### ۴-۱- کاهش نمونه با بهره‌گیری از الگوریتم بهینه‌سازی چندهدفه ازدحام ذرات آشوبی

الگوریتم بهینه‌سازی چندهدفه ازدحام ذرات نسخه‌ی تغییریافته‌ی PSO است [۲۷]. این الگوریتم دارای یک مخزن جهت نگهداری جواب‌های نامغلوب است. تفاوت الگوریتم بهینه‌سازی چندهدفه ازدحام ذرات با PSO در داشتن همین مخزن است که در آن $p_g$ به‌صورت مخزنی از جواب‌هاست که آن را $rep_i$ برای ذره $i$ می‌نامیم. در روش پیشنهادی، از این الگوریتم جهت کاهش نمونه استفاده شده است؛ اما روش PSO ذاتاً پیوسته بوده و برای استفاده در مسئله‌ی کاهش نمونه که ماهیتی گسسته دارد باید تغییر یابد. با باز-تعریف مفهوم سرعت در PSO به نسخه‌ی دودوئی این الگوریتم می‌رسیم. در نسخه دودوئی، موقعیت هر ذره در هر بعد به دو مقدار صفر و یک محدود می‌شود. در این نسخه مفهوم سرعت به مفهوم احتمال تغییر یافته و $v_{id}$ احتمال یک بودن $x_{id}$ را بیان می‌کند. بدین معنی که $v_{id}$ مقداری بین [۰-۱] می‌گیرد که این مقدار[۱] بیان‌کننده احتمال یک بودن $x_{id}$ است. بدین منظور، ابتدا سرعت ذره در هر بعد با استفاده از رابطه (۱) محاسبه می‌شود. سپس، این مقدار با استفاده از تابع محدود کننده سیگموئید در رابطه (۳)، به مقداری بین صفر و یک نگاشت می‌شود و در نهایت موقعیت ذره $i$ام در بعد $d$ام با رابطه (۴) به روز می‌شود:

$$S(v_{id}) = Sigmoid(v_{id}) = \frac{1}{1+e^{-v_{id}}} \quad (۳)$$

$$\begin{matrix} if\ S(v_{id}(t+1) = max\{S(v_{id}(t+1))\} \\ then\ x_{id}(t+1) = 1\ else\ x_{id}(t+1) = 0 \end{matrix} \quad (۴)$$

با استفاده از ایده دودوئی، در روش پیشنهادی، هر ذره از جمعیت به‌صورت یک آرایه به طول کل تعداد نمونه‌های آموزش در نظر گرفته شده است. این آرایه به‌منظور تولید جمعیت مورد نیاز الگوریتم بهینه‌سازی می‌باشد. چنانچه سرعت ذره به مقدار بیشینه خود تا تکرار فعلی برسد، نمونه متناظر با آن حفظ می‌شود (مقدار درایه مربوطه برابر با یک می‌گردد)؛ در غیر این‌صورت، آن نمونه حذف می‌شود (مقدار درایه مربوطه صفر می‌گردد). با انجام این روش تعداد بسیار زیادی آرایه با ترکیب‌های مختلفی از صفر و یک به دست می‌آید که هر کدام از این آرایه‌ها می‌تواند جواب مسئله باشد. به منظور جستجو در آرایه‌ها و انتخاب بهترین آرایه (بهترین زیرمجموعه از نمونه‌ها) به‌عنوان راه‌حل مسئله از الگوریتم بهینه-سازی چندهدفه ازدحام ذرات استفاده می‌نمائیم. لازم به ذکر است که شکل دودویی آرایه‌ها برای بیان مسئله به شکل نمادین می‌باشد و در حل مسئله، مقادیر اکتسابی آرایه در طول حرکت ذرات بدون نگاشت حفظ می‌شوند. در این الگوریتم هر آرایه را به‌عنوان یک ذره و مجموع تمام آرایه‌ها را جمعیت می‌نامیم. به‌عنوان مثال در شکل ۲ یک آرایه (ذره) به‌عنوان یکی از جواب‌های مسئله به‌صورت نگاشت شده به شکل دودویی آورده شده است. در این شکل فرض شده است مجموعه داده موردنظر شامل ۱۰ نمونه است.

| #نمونه | ۱ | ۲ | ۳ | ۴ | ۵ | ۶ | ۷ | ۸ | ۹ | ۱۰ |
|---|---|---|---|---|---|---|---|---|---|---|
| | ۰ | ۰ | ۱ | ۱ | ۱ | ۰ | ۱ | ۱ | ۰ | ۰ |

**شکل ۲: مثال از یکی از جواب‌های مسأله به‌صورت نگاشت شده به حالت دودویی.**

براساس شکل ۲، راه‌حل ارائه شده بیان‌گر انتخاب نمونه‌های سه، چهار، پنج، هفت و هشت و حذف نمونه‌های یک، دو، شش، نه و ده می‌باشد.

مراحل کاهش نمونه در روش پیشنهادی به شرح زیر است:

به منظور شناسایی جواب‌های نامغلوب و ذخیره آن‌ها در مخزن، مجموعه‌ای از ذرات با خصوصیات گفته شده طبق روابط (۴-۱) ایجاد می‌شوند. به منظور ارزیابی میزان کارایی ذره، مجموعه داده انتخابی (ذره) به سطح تصمیم طراحی شده داده می‌شود. مقداری که سطح تصمیم برمی‌گرداند، مقدار برازندگی برای هر جواب است. به دلیل آن‌که به‌دنبال ذراتی هستیم که رابطه (۵) را کمینه کنند، این رابطه را به‌عنوان تابع برازندگی و مسئله را به‌صورت مسئله کمینه‌سازی تابع چندهدفه درنظر می‌گیریم:

$$\begin{matrix} \min \omega_1(100 - Gmean) \\ + \omega_2(100 - Reduction_{Rate}) \\ + \omega_3(100 - Fmeasure) \\ + \omega_4(100 - Accuracy) \end{matrix} \quad (۵)$$

تابع برازندگی شامل چهار معیار G-mean (رابطه (۶))، نرخ کاهش (رابطه (۷))، دقت طبقه‌بندی (رابطه (۸)) و صحت طبقه‌بندی (رابطه (۹)) می‌باشد.

$$Gmean = \sqrt{TPR \times TNR} \quad (۶)$$

$$ReductionRate = \frac{T-R}{T} \quad (۷)$$

$$presicion = \frac{TP}{TP+FP}; recall = \frac{TP}{TP+FN}$$
$$Fmeasure = 2 \cdot \frac{precision \cdot recall}{presicion + recall} \quad (۸)$$

$$Acuuracy = (TP + TN)/(TP + TN + FP + FN) \quad (۹)$$

در رابطه (۷)، $T$ تعداد کل داده‌های آموزش و $R$ تعداد نمونه‌های باقی‌مانده است.

لازم به ذکر است در حالت مطلوب نمونه‌هایی انتخاب می‌شوند که بتوانند مقادیر معیارهای G-mean، F-measure و صحت طبقه-بندی را افزایش دهند. علاوه‌براین، اگر چه هدف آن است که مجموعه داده‌ای با کمینه اندازه حاصل گردد (بیشترین نرخ کاهش) اما کاهش تعداد نمونه‌ها بر روی عملکرد روش پیشنهادی تاثیر منفی دارند. لذا، با مسئله چندهدفه روبه رو هستیم و باید بین معیارهای متضاد مصالحه برقرار کنیم. اگر مقدار پارامترهای موجود در رابطه (۵) به صورت درصد محاسبه شوند و بهترین مقادیر برای این معیارها ۱۰۰٪ باشد، این رابطه به دنبال جواب-هایی است که کمترین اختلاف را با مقدار ایده‌آل ۱۰۰٪ داشته باشند. از طرف دیگر، با استفاده از وزن‌های $\omega_i$، $\sum_{i=1}^{4} \omega_i = 1$، میزان اهمیت معیارهای متفاوت تمیز داده شده است. در این مسئله، ابتدا ده درصد از جواب‌هایی که کمترین برازندگی را دارند به‌عنوان جواب‌های مغلوب در نظر گرفته می‌شوند.

۲. در این مرحله، هر ذره یکی از جواب‌های نامغلوب موجود در مخزن را به‌طور تصادفی به‌عنوان رهبر درنظر می‌گیرد و موقعیت خود را به کمک روابط (۱۰) و (۱۱) به‌روز می‌کند.

$$v_i(t+1) = wv_i(t) + c_1r_1(p_i(t) - x_i(t) + c_2r_2(rep_i(t) - x_i(t)) \quad (۱۰)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (۱۱)$$

بهترین خاطره شخصی هر آرایه به‌روز رسانی می‌شود.

۳. جواب‌های نامغلوب جدید به مخزن افزوده می‌شوند.

۴. اعضای مغلوب ایجاد شده داخل مخزن حذف می‌شوند.

۵. بررسی شروط توقف: ۱۰۰۰رسیدن به تکرار و یا عدم تغییر جواب بهینه در ۱۰ تکرار اخیر. اگر شروط توقف برقرار نبود، مراحل الگوریتم از گام ۲ ادامه پیدا می‌کند.

۶. یکی از جواب‌ها از مجموعه جواب نهایی به‌عنوان مجموعه کاهش یافته به‌طور تصادفی انتخاب می‌شود. از آنجا که نسخه‌ی دودویی الگوریتم PSO همگرایی مناسبی ندارد بنابراین از مفهوم آشوب برای تسریع همگرایی استفاده می‌کنیم. از توابع آشوب برای محاسبه مقادیر تصادفی $r_1$ و $r_2$ موجود در رابطه (۱۰) استفاده شده است. این مقادیر در هر به‌روزرسانی حرکت ذره محاسبه می-شود. توابع آشوب روش‌های جستجوی تصادفی هستند که می-توانند به‌عنوان تولیدکننده‌ی متغیرهای تصادفی ایفای نقش کنند. اعداد تولید شده توسط الگوریتم‌هایی که قابلیت تولید اعداد تصادفی را دارند به‌طور تناوبی تکرار می‌شوند. الگوریتم‌های مبتنی بر آشوب به‌علت عدم تکرار در مقایسه با الگوریتم‌های مبتنی بر احتمال سرعت بالاتری دارند. این افزایش سرعت می‌تواند با اعمال توابع آشوب مختلفی در رابطه (۱۰)، محقق گردد. توابع آشوب علاوه بر افزایش کیفیت جستجو، گاها با فرار از بهینه محلی قابلیت جستجوی سراسری را افزایش می‌دهند. این الگوریتم تصادفی به‌نظر می‌رسد اما در واقع تحت شرایط مشخص در یک سیستم غیرخطی مشخص اتفاق می‌افتد. تابع آشوب به‌کار گرفته شده در روش پیشنهادی، نگاشت لجستیک می‌باشد. رابطه این الگو به‌صورت یک دینامیک غیرخطی از جمعیت زیستی با رفتار آشوبی به‌دست می‌آید [۲۸]:

$$x_{k+1} = ax_k(1 - x_k) \quad (۱۲)$$

در رابطه (۱۲) $x_k$، $k$امین عدد آشوبی است. بر اساس تجربه مقدار $a$ برابر با ۴ درنظر گرفته می‌شود [۲۸].

**۴-۲- سطح تصمیم مبتنی بر فاصله**

در روش پیشنهادی به‌منظور تعیین مرز بین دو کلاس هدف (حاوی نمونه‌هایی که در مجموعه داده باقی می‌مانند) و کلاس غیرهدف (حاوی نمونه‌هایی که از مجموعه داده حذف می‌شوند)، از ایده سطح تصمیم‌گیری مبتنی بر فاصله [۲۹] استفاده شده است. فرض می‌کنیم. $d(x, x_i)$ مشخص‌کننده‌ی فاصله‌ی نمونه‌ی $x$ از مجموعه آموزش $x_i$ (کلاس هدف) و $d(x, x_j)$ مشخص‌کننده‌ی فاصله‌ی نمونه‌ی $x$ از مجموعه آموزش $x_j$ (کلاس غیرهدف) است. در این روش به‌دنبال نمونه‌های $x$ای هستیم که میانگین مجذور فاصله‌ی آن از هر دو کلاس برابر باشد. طبق رابطه‌ی (۱۳) داریم:

$$\frac{1}{n_1}\sum_{i=1}^{n_1} \mu_i d_1(x, x_i) = \frac{1}{n_2}\sum_{j=1}^{n_2} \mu_j\, d_2(x, x_j) \quad (۱۳)$$

در این رابطه، $\mu_i$ مشخص کننده تعداد همسایه‌های نمونه $x_i$ می-باشد که روش محاسبه آن در [۲۹] بیان گردیده شده است، $n_1$ و $n_2$ به‌ترتیب بیانگر تعداد نمونه‌های کلاس اول و کلاس دوم می-باشند. با توجه به این‌که در روش پیشنهادی تمرکز زیادی روی داده‌های نامتوازن وجود دارد و از آن‌جایی که رابطه (۱۳) در مواجهه با کلاس‌های نامتوازن نتایج مطلوبی به‌همراه ندارد، اصلاح این رابطه ضروری به‌نظر می‌رسد. یک راه حل برای رفع این نقص، نرمال کردن حجم داده‌های دو کلاس می‌باشد. به دلیل آن‌که هدف، محاسبه میانگین فاصله است و معیار فاصله در صورت کسر رابطه (۱۳) به صورت مجذور محاسبه شده، تقسیم طرفین رابطه به مجذور تعداد نمونه‌های آموزشی هر کلاس، راه‌کار ارائه شده

برای حل این مشکل می‌باشد که در رابطه (۱۴) نمایش داده شده است [۲۹]:

$$\frac{1}{n_1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\mu_i d_1(x,x_i)\right) = \frac{1}{n_2}\left(\frac{1}{n_2}\sum_{j=1}^{n_2}\mu_j d_2(x,x_j)\right) \quad (۱۴)$$

برای محاسبه‌ی فاصله‌ی بین نمونه $x$ و نمونه‌های دو کلاس از مجذور فاصله‌ی اقلیدسی (2-norm) استفاده گردید. با جایگذاری این فاصله در رابطه‌ی (۱۴)، رابطه‌ی (۱۵) را داریم:

$$n_2^2 \times \sum_{i=1}^{n_1}\mu_i \|x-x_i\|^2 = n_1^2 \times \sum_{j=1}^{n_2}\mu_j \|x-x_j\|^2 \quad (۱۵)$$

حال با فرض $\|x-x_i\|^2 = (x-x_i)^T(x-x_i)$ و ساده‌سازی روابط داریم:

$$x^T x\left(n_2^2\sum_{i=1}^{n_1}\mu_i - n_1^2\sum_{j=1}^{n_2}\mu_j\right) - 2x^T \times \left(n_2^2\sum_{i=1}^{n_1}\mu_i x_i - n_1^2\sum_{j=1}^{n_2}\mu_j x_i\right) + n_2^2\sum_{i=1}^{n_1}\mu_i x_i^T x_i - n_1^2\sum_{j=1}^{n_2}\mu_j x_j^T x_j = 0 \quad (۱۶)$$

سطح تصمیم مورد استفاده در رابطه (۱۷) ارائه شده است. مقادیر $a$، $b$ و $c$ طبق روابط (۱۸)، (۱۹) و (۲۰) محاسبه می‌شوند.

$$F(x) = ax^T x + bx + c = 0 \quad (۱۷)$$

$$a = \left(n_2^2\sum_{i=1}^{n_1}\mu_i - n_1^2\sum_{j=1}^{n_2}\mu_j\right) \quad (۱۸)$$

$$b = 2\left(n_2^2\sum_{i=1}^{n_1}\mu_i x_i - n_1^2\sum_{j=1}^{n_2}\mu_j x_j\right) \quad (۱۹)$$

$$c = n_2^2\sum_{i=1}^{n_1}\mu_i x_i^T x_i - n_1^2\sum_{j=1}^{n_2}\mu_j x_j^T x_j \quad (۲۰)$$

در رابطه (۱۷) با اعمال داده‌های نمونه‌ی $x$ و با تعیین علامت آن می‌توان کلاس نمونه‌ی موردنظر را تشخیص داد. هدف حذف نمونه‌هایی است که تاثیر کمتری در تعیین مرز دارند. از طرف دیگر، در این سطح با نگاشت داده‌ها به فضایی با ابعاد بالاتر توسط روش‌های کرنل به فرمول غیرخطی دیگری به‌عنوان سطح تصمیم‌گیرنده رسیدیم که در رابطه (۲۱) آورده شده است. جزئیات محاسبه‌ی این دو سطح تصمیم‌گیرنده در [۲۹] می‌باشد.

$$n_2 \times \sum_{i=1}^{n_1}\mu_i[k(x,x) - 2\times k(x,x_i) + k(x_i,x_i)] = n_1 \times \sum_{j=1}^{n_2}\mu_j[k(x,x) - 2\times k(x,x_j) + k(x_j,x_j)] \quad (۲۱)$$

$k(x,x)$ معرف تابع کرنل است. چنانچه تابع پایه شعاعی[۲] [۳۰] در نظر گرفته شود، سطح تصمیم از رابطه (۲۲) محاسبه می‌شود:

$$n_1\sum_{j=1}^{n_2}\mu_j e^{-\frac{\|x-x_j^2\|}{2\sigma^2}} - n_1\sum_{j=1}^{n_2}\mu_j + n_2\sum_{i=1}^{n_1}\mu_i - n_2\sum_{i=1}^{n_1}\mu_i e^{-\frac{\|x-x_i^2\|}{2\sigma^2}} = 0 \quad (۲۲)$$

الگوریتم روش پیشنهادی در ادامه آورده شده است:

***Algorithm of the proposed method***

***Initial parameter:***
*Size of population=200;*
*Size of repository=20;*

***Input:*** *main Data Set*
***Output:*** *Reduced Data Set*

```
While(stop condition)
    Particles=Create initial population using random value(main Data Set);
    While(i<=(size(Particles)))
        If(Particle(i) =Max(Particle(i)))
            Add Particle(i) to SelectedSet;
    end
    FitnessValue=DecisionSurface(SelectedSet);
    Repository=Select 20 best fitness value as nondominated particles;
    DominatedParticle=(~(Repository));
    The position of Dominated Particles are updated based on nondominated particles;
    Best known position of each particle are updated
    Add new nondominated into repository
    Remove dominated particles from repository
End
ReducedSet=Select one solution from Repository randomly;
```

***Decision Surface algorithm***

***Initial parameter:***
*$w_1$=0.2; $w_2$=0.25; $w_3$=0.25; $w_4$=0.3;*
***Input:*** *SelectedSet*
***Output:*** *FitnessValue*

```
Divide SelectedSet into TrainData and Testdata;
DecisionSurface= Make Decision Surface by TrainData;
While(i<=size(TestData))
    If (DecisionSurface(TestData(i))>=0)
        LableOfTestData(i)=1;
    Else
        LableOfTestData(i)=2;
End
Compute accuracy, precision, G-mean, ReductionRate;
FitnessValue=w1.accuracy+w2.precision
```

$+w_3.G\text{-}mean+w_4.\ ReductionRate;$

## ۵- نتایج آزمایش‌ها و مقایسه‌ها

در این بخش نتایج حاصل از آزمایش‌ها که بر روی کامپیوتری با پردازنده‌ی Intel، core i5، 4GB Ram، 2.7GHz، توسط نرم افزار Matlab2014b پیاده‌سازی شده است بیان می‌شود. جدول ۱ پارامترها و مقادیر استفاده شده در الگوریتم بهینه‌سازی ذرات آشوبی را نشان می‌دهد.

**جدول ۱: تنظیمات پارامترها در بهینه‌سازی ازدحام ذرات آشوبی**

| پارامتر | مقدار |
|---|---|
| $w$ | {۰,۹-۰,۷-۰,۴} |
| $c_1$ | ۰,۵ |
| $c_2$ | ۰,۷ |
| $a$ | ۴ |

مجموعه داده‌های ذکر شده در جدول ۲، از منبع داده‌ی UCI [۳۱] استخراج شده‌اند. همان‌طور که در جدول ۶ نشان داده شده است، سه مجموعه داده heart، glass و Haberman توزیع نامتوازن‌تری دارند.

**جدول ۲: پایگاه داده‌های انتخابی از UCI**

| شماره | مجموعه داده | #نمونه | #ویژگی | #کلاس |
|---|---|---|---|---|
| ۱ | Ionosphere | ۳۵۱ | ۳۴ | ۲ |
| ۲ | Sonar | ۲۰۸ | ۶۰ | ۲ |
| ۳ | Wdbc | ۵۶۹ | ۳۰ | ۲ |
| ۴ | Liver | ۳۴۵ | ۶ | ۲ |
| ۵ | Haberman | ۳۰۶ | ۳ | ۲ |
| ۶ | Heart | ۲۶۷ | ۴۴ | ۲ |
| ۷ | Pima | ۷۶۸ | ۸ | ۲ |
| ۸ | Iris | ۱۵۰ | ۴ | ۳ |
| ۹ | Glass | ۲۱۴ | ۹ | ۶ |
| ۱۰ | Census | ۲۹۹۲۸۵ | ۴۱ | ۲ |

در تمام آزمایش‌ها، از روش اعتبارسنجی متقابل ۱۰تایی[۳] به‌منظور ارزیابی روش پیشنهادی استفاده شده است. در روش پیشنهادی از کرنل تابع پایه‌ای شعاعی و همچنین تابع آشوب لجستیک استفاده شده است. در تخمین مقدار $\delta$ کرنل از مقادیر {۱۶-۱۲-۸-۴-۲-۱-۰/۸-۰/۷۵-۰/۵-۰/۴-۰/۳-۰/۲-۰/۱} استفاده شد. دقت طبقه‌بندی داده‌های کاهش داده شده در روش پیشنهادی، با طبقه‌بند SVM با مقادیر $C=$ {۰/۱-۰/۲-۰/۴-۰/۸-۱-۲-۴-۸-$inf$} ارزیابی شده است. به این‌صورت که این طبقه‌بندها با داده‌های کاهش داده شده آموزش می‌بینند سپس با داده‌های آزمون این طبقه‌بندها ارزیابی می‌شوند. وزن‌های استفاده شده در تابع برازندگی (رابطه (۵)) بنا به کاربرد قابل تغییرند. از آنجا که هدف روش پیشنهادی برقراری و حفظ توازن پس از عمل کاهش نمونه در مجموعه داده است لذا، به معیار G-mean که برای ارزیابی داده‌ها با کلاس‌های نامتوازن به کار می‌رود وزن بیشتری نسبت داده شد. بقیه وزن‌ها به طور مساوی بین سایر معیارها تقسیم گردید. $\omega_1 = 1$ ، $\omega_2 = 0/25$ ، $\omega_3 = 0/25$ ، $\omega_4 = 0/3$.

در آزمایش‌های انجام شده، رتبه هر روش در پرانتز نشان داده شده است. علاوه بر این، مقایسه نتایج به سه روش رتبه‌بندی، مقدار p-value و آزمون رتبه‌بندی علامت‌دار ویلکاکسون[۴] صورت می‌گیرد. در این دو آزمون فرض اولیه‌ای به این‌صورت در نظر گرفته می‌شود: تفاوت چشمگیری بین روش پیشنهادی و سایر روش‌ها وجود ندارد. سپس بررسی می‌شود که آیا این فرضیه صحت دارد یا رد می‌شود. در واقع این آزمون‌ها اتفاقی بودن اختلاف بین روش پیشنهادی و سایر روش‌ها را بررسی می‌کنند. مقدار p-value احتمال برقراری فرض اولیه است که هرچه کمتر باشد بهتر است، که برای آن به‌طور قراردادی مرز ۰/۰۵ در نظر گرفته می‌شود. مقدار p-value اگر از این مرز کمتر باشد به این معناست که تفاوت بین روش پیشنهادی و سایر روش‌ها معنادار است. این مقدار در جداول گزارش شده است. در آزمون رتبه‌بندی علامت‌دار ویلکاکسون که به‌صورت دوطرفه[۵] در نظر گرفته شده است در ابتدا یک فرض تهی به این صورت که تفاوتی بین روش پیشنهادی یا سایر روش‌ها وجود ندارد در نظر گرفته می‌شود. در این آزمون سطح اهمیت ۰/۰۵ و مقدار بحرانی[۶] برای ۱۰ داده به کمک جدول A5 در [۳۲] برابر ۸ می‌باشد. در جداول نتایج آزمایش‌ها، آزمون ویلکاکسون در دو سطر آخر آورده شده است. مقدار یک در ردیف آخر به معنای رد فرض تهی می‌باشد و به این معناست که روش پیشنهادی با سایر روش‌ها اختلاف چشم‌گیری دارد و مقادیر صفر نیز عدم رد فرض تهی را نشان می‌دهد و بیان‌گر این است که روش با سایر روش‌ها اختلاف چندانی ندارد.

برای بررسی قابلیت‌های روش پیشنهادی، این روش با سایر روش‌های مطرح در زمینه کاهش نمونه از قبیل [۱۴-۱۵] و [۱۸-۱۷] مقایسه گردیده است. در تمام این روش‌ها از kNN با k=۳ و فاصله‌ی اقلیدسی جهت انجام محاسبات استفاده شده است.

نرخ صحت در جدول ۳ آورده شده است. همان‌طور که مشاهده می‌شود روش پیشنهادی در همه مجموعه‌های داده به جز Sonar دارای بیشترین نرخ صحت و بیشترین میانگین نرخ صحت در مقایسه با سایر روش‌ها است. مقادیر p-value گزارش شده بیان‌گر وجود اختلاف معنادار بین روش پیشنهادی و سایر روش‌هاست. مقادیر یک حاصل از نتیجه آزمون ویلکاکسون در سطر آخر جدول، بیانگر رد آزمون تهی است، به این معنا که نرخ خطا در

روش پیشنهادی، با سایر روش‌های مطرح شده اختلاف قابل توجهی دارد.

**جدول ۳: نرخ صحت**

| مجموعه داده | روش پیشنهادی | [۱۵] | [۱۷] | [۱۸] | [۱۴] |
|---|---|---|---|---|---|
| Ionosphere | **۸۵/۲۳(۱)** | ۸۵/۱۵(۲) | ۸۱/۷۳(۵) | ۸۲/۵۲(۴) | ۸۴/۸۷(۳) |
| Sonar | ۸۹/۴۱(۲) | **۸۹/۷۷(۱)** | ۷۱/۱۳(۵) | ۷۸/۹۲(۴) | ۸۷/۷۷(۳) |
| Wdbc | **۸۴/۷۵(۱)** | ۸۳/۱۴(۲) | ۸۲/۰۵(۴) | ۷۹/۳۳(۵) | ۸۲/۶۹(۳) |
| Liver | **۶۸/۹۷(۱)** | ۶۷/۵۵(۲) | ۶۵/۱۸(۴) | ۶۴/۲۱(۵) | ۶۷/۲۱(۳) |
| Haberman | **۷۷/۱۲(۱)** | ۷۴/۳۰(۳) | ۶۶/۱۰(۴) | ۶۴/۱۹(۵) | ۷۶/۴۶(۲) |
| Heart | **۶۸/۸۸(۱)** | ۶۸/۲۸(۲) | ۶۳/۸۷(۵) | ۶۶/۵۶(۴) | ۶۷/۸۷(۳) |
| Pima | **۷۱/۸۹(۱)** | ۷۰/۹۰(۲) | ۶۸/۴۷(۴) | ۷۰/۰۸(۳) | ۶۷/۵۵(۵) |
| Iris | **۹۶/۹۱(۱)** | ۹۶/۱۳(۳) | ۹۵/۰۸(۴) | ۹۴/۱۵(۵) | ۹۶/۵۳(۲) |
| Glass | **۸۰/۷۳(۱)** | ۸۰/۱۵(۲) | ۷۲/۴۹(۴) | ۷۰/۷۸(۵) | ۷۹/۱۵(۳) |
| Census | **۸۰/۶۱ (۱)** | ۷۹/۸۲(۲) | ۷۷/۵۲(۳) | ۷۵/۱۶(۵) | ۷۶/۳۳(۴) |
| میانگین رتبه | **۱/۱(۱)** | ۲/۱(۲) | ۴/۲(۴) | ۴/۵(۵) | ۳/۱(۳) |
| p-value | | ۰/۰۰۸۵ | ۰/۰۰۴۵ | ۰/۰۰۱۱ | ۰/۰۰۳۳ |
| آماره T | | ۳/۳۵۳۱ | ۳/۷۵۶۰ | ۴/۷۲۹۲ | ۳/۹۵۸۱ |
| آزمون ویلکاکسون | | ۱ | ۱ | ۱ | ۱ |

در جدول ۴ نرخ کاهش بررسی شده است. روش پیشنهادی در بیشتر مجموعه داده‌ها به جز Ionosphere، Iris و Glass دارای نرخ کاهش بیشتری نسبت به سایر روش‌ها است و به‌جز روش [۱۵] با سایر روش‌ها اختلاف قابل توجهی دارد. مقدار p-value نیز این مسئله را تایید می‌کند.

**جدول ۴: نرخ کاهش**

| مجموعه داده | روش پیشنهادی | [۱۵] | [۱۷] | [۱۸] | [۱۴] |
|---|---|---|---|---|---|
| Ionosphere | ۷۸/۰۲(۲) | **۷۹/۱۳(۱)** | ۶۲/۱۹(۴) | ۵۹/۰۷(۵) | ۷۶/۶۲(۳) |
| Sonar | **۹۱/۹۲(۱)** | ۹۰/۰۱(۳) | ۷۳/۰۳(۴) | ۶۹/۱۱(۵) | ۹۰/۳۵(۲) |
| Wdbc | **۷۲/۱۴(۱)** | ۶۶/۷۲(۴) | ۵۹/۰۷(۵) | ۶۷/۳۹(۲) | ۶۶/۹۳(۳) |
| Liver | **۷۸/۱۳(۱)** | ۷۶/۳۶(۳) | ۶۷/۸۱(۵) | ۶۸/۳۹(۴) | ۷۷/۹۲(۲) |
| Haberman | **۸۱/۹۵(۱)** | ۷۹/۴۵(۳) | ۸۰/۴۵(۲) | ۷۹/۱۵(۴) | ۷۳/۵۱(۵) |
| Heart | **۹۰/۰۴(۱)** | ۸۸/۱۵(۴) | ۸۹/۰۲(۲) | ۷۹/۲۱(۵) | ۸۸/۹۵(۳) |
| Pima | **۹۳/۶۳(۱)** | ۹۲/۸۰(۳) | ۹۳/۱۷(۲) | ۶۲/۳۴(۵) | ۸۵/۳۱(۴) |
| Iris | ۷۸/۱۹(۲) | ۷۵/۶۸(۳) | ۷۰/۰۹(۵) | **۷۹/۲۹(۱)** | ۷۳/۵۸(۴) |
| Glass | ۸۹/۰۲(۳) | **۹۲/۸۰(۱)** | ۷۸/۱۸(۴) | ۷۷/۴۰(۵) | ۸۹/۶۳(۲) |
| Census | **۸۰/۲۲(۱)** | ۷۷/۱۸(۲) | ۷۳/۸۸(۴) | ۷۳/۳۱(۵) | ۷۴/۱۴(۳) |
| میانگین رتبه | **۱/۴(۱)** | ۲/۷(۲) | ۳/۷(۴) | ۴/۱(۵) | ۳/۰(۳) |
| p-value | | ۰/۰۸۸۳ | ۰/۰۰۲۰ | ۰/۰۰۴۲ | ۰/۰۰۷۳ |
| آماره T | | ۱/۹۱۱۲ | ۴/۲۸۵۴ | ۳/۷۹۶۳ | ۳/۴۴۶۸ |
| آزمون ویلکاکسون | | ۰ | ۱ | ۱ | ۱ |

در روش پیشنهادی جهت کنترل توازن داده‌های متوازن بعد از عملیات کاهش و همچنین کنترل دقت طبقه‌بندی کلاس اقلیت در مجموعه داده‌های نامتوازن، از معیار G-mean استفاده شده است. معیار G-mean از جمله معیارهای مهم در برخورد با مجموعه داده‌های نامتوازن می‌باشد زیرا پایین بودن دقت یک کلاس موجب کاهش این معیار می‌شود، پس به‌کمک این معیار می‌توان میزان توجه به کلاس اقلیت را افزایش داد. در جدول ۵ مقادیر میانگین G-mean آورده شده است. روش پیشنهادی دارای نرخ G-mean بالایی است، لذا در مواجهه با مجموعه داده‌های نامتوازن عملکرد خوبی دارد. از طرف دیگر در مقایسه با سایر روش‌ها به‌جز روش [۱۵] مقدار G-mean اختلاف قابل قبول دارد.

**جدول ۵: میانگین G-mean**

| مجموعه داده | روش پیشنهادی | [۱۵] | [۱۷] | [۱۸] | [۱۴] |
|---|---|---|---|---|---|
| Ionosphere | **۷۵/۶۹(۱)** | ۷۲/۸۵(۲) | ۶۵/۱۹(۵) | ۷۱/۲۲(۳) | ۷۰/۸۵(۴) |
| Sonar | **۸۷/۱۳(۱)** | ۸۱/۲۸(۴) | ۷۹/۸۴(۵) | ۸۱/۲۹(۳) | ۸۳/۳۸(۲) |
| Wdbc | ۷۰/۰۹(۳) | ۵۱/۰۴(۵) | ۷۲/۱۹(۲) | **۷۲/۳۳(۱)** | ۶۵/۶۲(۴) |
| Liver | ۸۴/۰۹(۲) | **۸۸/۳۶(۱)** | ۷۸/۴۴(۵) | ۸۲/۹۲(۳) | ۸۲/۹۱(۴) |
| Haberman | **۹۱/۸۲(۱)** | ۹۱/۷۰(۲) | ۸۳/۲۹(۵) | ۸۳/۴۸(۴) | ۸۹/۳۹(۳) |
| Heart | **۹۶/۸۸(۱)** | ۹۲/۸۱(۳) | ۸۸/۶۱(۵) | ۸۹/۲۱(۴) | ۹۱/۸۲(۲) |
| Pima | ۸۴/۳۹(۳) | ۸۵/۰۹(۲) | ۷۹/۸۹(۵) | **۸۵/۷۰(۱)** | ۸۲/۹۲(۴) |
| Iris | **۹۳/۹۱(۱)** | ۹۰/۲۹(۳) | ۸۲/۰۹(۵) | ۹۲/۰۲(۲) | ۹۲/۲۵(۴) |
| Glass | **۹۷/۱۸(۱)** | ۹۶/۶۳(۳) | ۹۷/۰۱(۲) | ۹۳/۵۲(۴) | ۹۳/۴۹(۵) |
| Census | ۸۹/۹۱(۲) | ۸۶/۳۴(۳) | ۸۰/۸۶(۴) | ۹۰/۲۷(۱) | ۷۹/۲۰(۵) |
| میانگین رتبه | **۱/۶(۱)** | ۲/۸(۳) | ۴/۳(۵) | ۲/۶(۲) | ۳/۷(۴) |
| p-value | | ۰/۱۱۰۷ | ۰/۰۰۱۴ | ۰/۰۳۴۲ | ۰/۰۰۱۶ |
| آماره T | | ۱/۷۶۹۲ | ۴/۵۳۷۵ | ۲/۴۹۳۷ | ۴/۴۶۸۶ |
| آزمون ویلکاکسون | | ۰ | ۱ | ۱ | ۱ |

در جدول ۶ چگونگی توزیع داده‌ها بین کلاس‌ها در مجموعه داده‌های انتخابی آورده شده است. در جدول ۷ درصد توزیع داده‌ها را بعد از اعمال روش پیشنهادی نشان می‌دهد. خط تیره‌های موجود در جدول‌های ۶ و ۷ بیانگر عدم وجود داده‌ای در کلاس مذبور در مجموعه داده‌ی مورد نظر می‌باشد. همان‌طورکه جدول ۷ نشان می‌دهد، کاهش داده‌ها به‌صورت متوازن صورت گرفته است، به‌طوری‌که مجموعه داده‌های متوازن بعد از اعمال روش پیشنهادی همچنان به‌صورت متوازن باقی مانده‌اند و از طرفی عدم توازن داده‌های نامتوازن بر اثر عملیات کاهش نمونه، بیشتر نشده است. دلیل این امر وجود معیار G-mean است که به هر دو کلاس اکثریت و اقلیت اهمیت می‌دهد.

**جدول ۶: درصد توزیع داده‌ها بین کلاس‌ها**

| مجموعه داده | #class6 | #class5 | #class4 | #class3 | #class2 | #class1 |
|---|---|---|---|---|---|---|
| Ionosphere | - | - | - | - | ۳۶ | ۶۴ |
| Sonar | - | - | - | - | ۵۳ | ۴۷ |
| Wdbc | - | - | - | - | ۶۳ | ۳۷ |
| Liver | - | - | - | - | ۵۸ | ۴۲ |
| Haberman | - | - | - | - | ۲۶ | ۷۴ |
| Heart | - | - | - | - | ۲۱ | ۷۹ |
| Pima | - | - | - | - | ۶۵ | ۳۵ |
| Iris | - | - | - | ۳۳ | ۳۳ | ۳۴ |
| Glass | ۴ | ۶ | ۸ | ۳۶ | ۳۳ | ۷ |
| Census | - | - | - | - | ۲۸ | ۷۲ |

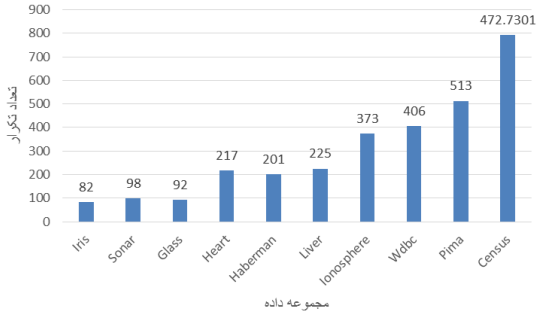
**جدول ۷: درصد توزیع داده‌ها بعد از اعمال روش پیشنهادی**

| مجموعه داده | #class6 | #class5 | #class4 | #class3 | #class2 | #class1 |
|---|---|---|---|---|---|---|
| Ionosphere | - | - | - | - | ۴۱/۰۹ | ۵۸/۹۱ |
| Sonar | - | - | - | - | ۵۱/۶۸ | ۴۸/۳۲ |
| Wdbc | - | - | - | - | ۵۸/۷۶ | ۴۱/۲۴ |
| Liver | - | - | - | - | ۶۰/۷۹ | ۳۹/۲۱ |
| Haberman | - | - | - | - | ۲۹/۴۴ | ۷۰/۵۶ |
| Heart | - | - | - | - | ۲۷/۲۹ | ۷۲/۷۱ |
| Pima | - | - | - | - | ۶۱/۵۸ | ۳۸/۴۲ |
| Iris | - | - | - | ۳۲/۴۲ | ۳۳/۰۹ | ۳۳/۴۹ |
| Glass | ۱۳/۷۹ | ۵/۳۱ | ۶/۸۳ | ۷/۱۲ | ۳۴/۲۹ | ۳۲/۶۶ |
| Census | - | - | - | - | ۳۰/۱۸ | ۶۹/۸۲ |

زمان اجرا بر حسب ثانیه در جدول ۸ و نمودار همگرایی روش پیشنهادی بر حسب تعداد تکرار در شکل ۳ نشان داده شده‌اند. مشاهده می‌گردد به دلیل استفاده از تابع آشوب، روش پیشنهادی سریع‌تر از سایرین به جواب نهایی و همگرایی می‌رسد.

**جدول ۸: زمان اجرا**

| مجموعه داده | روش پیشنهادی | [۱۵] | [۱۷] | [۱۸] | [۱۴] |
|---|---|---|---|---|---|
| Ionosphere | ۱۴/۱۳ | ۱۴/۷۷ | ۱/۵۲ | ۱۶۸/۴۵ | ۱۵۶/۵۴ |
| Sonar | ۴۱/۴۰ | ۴۴/۴۸ | ۵۴/۳۰ | ۵۹/۵۸ | ۵۱/۶۵ |
| Wdbc | ۱۸/۲۷ | ۱۹/۵۹ | ۲۱/۵۲ | ۲۲۰/۶۳ | ۲۰۶/۹۶ |
| Liver | ۹۸/۳۷ | ۱۰/۶۱ | ۱۱۶/۱۲ | ۱۲۳/۳۸ | ۱۱۲/۴۸ |
| Haberman | ۶۵/۹۰ | ۶۸/۵۳ | ۷۶/۹۰ | ۸۱/۴۰ | ۷۴/۶۴ |
| Heart | ۷۶/۰۳۷ | ۸۰/۶۰ | ۹۵/۱۳ | ۱۰۲/۹۵ | ۹۱/۲۲ |
| Pima | ۲۶۷/۸۹۱۶ | ۲۷۰/۶۴ | ۲۷۹/۴۵ | ۲۸۴/۱۸ | ۲۷۷/۰۸ |
| Iris | ۱۹/۱۰۱۹ | ۲۳/۷۱ | ۳۸/۴۰ | ۴۶/۲۹ | ۳۴/۴۴ |
| Glass | ۳۷/۸۴۰۲ | ۳۹/۷۴ | ۴۵/۸۲ | ۴۹/۰۸ | ۴۴/۱۸ |
| Census | ۴۷۲/۷۳۰۱ | ۴۷۵/۵۴ | ۴۸۴/۵۰ | ۴۸۹/۳۲ | ۴۸۲/۰۹ |




**شکل ۳: نمودار همگرایی الگوریتم پیشنهادی**

همان‌طور که مشاهده می‌گردد زمان اجرای روش پیشنهادی در مجموعه داده Census نسبت به زمان اجرای سایر مجموعه داده‌ها بیشتر است و دیرتر به همگرایی می‌رسد. البته این مسئله با توجه به حجم زیاد نمونه‌ها توجیه‌پذیر است.

## ۶- نتیجه‌گیری و کارهای آینده

در این مقاله، روشی برای کاهش داده‌ها با استفاده از الگوریتم بهینه‌سازی چندهدفه ازدحام ذرات آشوبی ارائه گردید. همچنین سطح تصمیم مبتنی بر فاصله جهت تعیین برچسب کلاس نمونه‌ها به‌کار گرفته شده است. در روش پیشنهادی، نوآوری‌های متفاوتی وجود دارد که عبارتند از: (۱) حل مسئله‌ی کاهش داده به‌صورت یک مسئله‌ی چند هدفه؛ (۲) حفظ توازن داده‌های متوازن بعد از عملیات کاهش و (۳) وجود دقت بالای طبقه‌بندی کلاس اقلیت در داده‌هایی با توزیع نامتوازن. استفاده از معیار G-mean در روش پیشنهادی، تاثیر به‌سزایی در محقق ساختن اهداف دوم و سوم داشته است. بر اساس نتایج حاصل از آزمایش‌ها، روش پیشنهادی در معیارهای نرخ صحت، نرخ کاهش و G-mean عملکرد قابل قبولی را از خود نشان داده است. در روش پیشنهادی، با به‌کار بردن الگوریتم تکاملی، فضای مسئله به‌خوبی جستجو می‌شود. از طرفی، استفاده از توابع آشوب، سرعت همگرایی الگوریتم را تسریع می‌بخشد و منجر به جستجوی بهینه سراسری می‌شود. استفاده از بهینه‌سازی چندهدفه تکاملی علاوه‌بر ایجاد مصالحه بین معیارهای مورد نظر موجب انتخاب مجموعه کوچکی از نمونه‌ها می‌شود که توانایی تضمین دقت مطلوب طبقه‌بندی را دارند. به دلیل آن‌که پایین بودن دقت یک کلاس موجب کاهش مقدار G-mean می‌شود، توجه به حفظ دقت کلاس اقلیت می‌تواند موجب کارائی قابل قبول روش پیشنهادی بر روی داده‌های با توزیع نامتوازن شود.

از آنجا که روش پیشنهادی یک روش برون‌خط[۷] است، نویسندگان توسعه کار در کاربردهای برخط[۸] را به‌عنوان کارهای آتی، پیشنهاد می‌دهند.

## مراجع


[1]. L. Nanni and A. Lumini, "Prototype Reduction Technique: A Comparison among Different Approaches," *Expert System with Application*, 2011.

[2]. D. J. Dittman, T. M. Khoshgoftaar and A. Napolitano, "Selecting the Appropriate Data Sampling Approach for Imbalanced and High-Dimensional Bioinformatics Datasets," IEEE 14th International Conference on Bioinformatics and Bioengineering, 2014.

[3]. J. Derrac, S. Garcia and F. Herrera, "A Survey on Evolutionary Instance Selection and Generation," International Journal of Applied Metaheuristic Computing, vol. 1, no. 1, pp. 60-92, January-March 2010.

[4]. J. Kennedy and R. Eberhart, "Particle Swarm Optimization," IEEE International Conference on Neural Network, vol. 4. Pp. 1942-1948, 1995.

[5]. Y. Shi and R. Eberhart, "A Modified Particle Swarm Optimizer", IEEE International CEC, pp. 69-73, 1998.

[6]. T. Zhai and Zh. He, "Instance Selection for Time Series Classification Based on Immune Binary Particle Swarm Optimization," Knowledge-Based Systems, vol. 49, pp. 106-115, 2013.

[7]. S. Sakinah S. Ahmad and W. Pedrycz, "Feature and Instance Selection Via Cooperative PSO," Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, pp. 2127-2132. IEEE, 2011.

[8]. C. Alejandro, I. Galván and P. Isasi "Michigan particle swarm optimization for prototype reduction in classification problems," New Generation Computing, vol. 27, no. 3, pp. 239-257, 2009.

[9]. N. Garcia-Pedrajas, J. Perez-Rodriguez and A. Haro-Garcia, "OligoIS: Scalable Instance Selection for Class-Imbalanced Data Sets," IEEE Transactions on Cybernetics, vol. 43, no. 1, February 2013.

[10]. P.E. Hart, "The Condensed Nearest Neighbor Rule," IEEE Transaction on Information Theory, vol. 14, no. 3, pp. 515-516, May 1968.
[11]. G.W. Gates, "The Reduced Nearest Neighbor Rule," IEEE Transaction Information Theory, vol. 18, no. 3, pp. 431-433, May 1972.
[12]. D.L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data", IEEE Transaction on Systems, Man, and Cybernetics, vol. 2, no. 3, pp. 408-421, July 1972.
[13]. S. Alzberg, "A Nearest Hyperrectangle Learning Method," Machine Learning, vol. 6, pp. 251-276, 1991.
[14]. J. Hamidzadeh, R. Monsefi and H. Sadoghi Yazdi, "LMIRA: Large Margin Instance Reduction Algorithm," Neurocomputing, vol. 145, pp. 477-487, 2014.
[15]. J. Hamidzadeh, R. Monsefi and H. Sadoghi Yazdi, "IRAHC: Instance Reduction Algorithm using Hyperrectangle Clustering," Pattern Recognition, vol. 48, pp. 1878-1889, 2015.
[16]. J.R. Cano, F. Herrera, M. Lozano, "Using evolutionary Algorithms as Instance Selection for Data Reduction: an experimental study," IEEE Transactions on Evolutionary Computation, vol. 7, no. 6, pp. 561–575, 2003.
[17]. Ch. Tsai, Z. Chen and Sh. Ke, "Evolutionary Instance Selection for Text Classification," The Journal of Systems and Softwares, vol. 90, 2014.
[18]. I. M. Anwar, Kh. M. Salama and A. M. Abdelbar, "Instance Selection with Ant Colony Optimization," Procedia Computer Science, vol. 53, 2015.
[19]. J. Perez, A. German and N. Garcia, "Simultaneous instance and feature selection and weighting using evolutionary Computation: Proposal and Study," Applied Soft Computing, vol. 37, 2015.
[20]. J. Derrac, S. Garcia, F. Herrera and IFS-CoCo, "Instance and Feature Selection Based on Cooperative Coevolution with Nearest Neighbor Rule," Pattern Recognition, vol. 43, 2010.
[21]. N. Garcia and C. Garcia, "Boosting for Class-Imbalanced Datasets Using Genetically Evolved Supervised Non-Linear Projections," Artificial Intelligence, vol. 2, 2013.
[22]. M. Blachnik, "Ensemble of Instance Selection Methods based on Feature Subset," 18th International Conference on Knowledge-based and Intelligent Information & Engineering Systems, 2014.
[23]. N. Garcia and A. de haro, "Boosting Instance Selection Algorithms," Knowledge-based systems, no. 19, 2014.
[24]. I. Trieguero, D. Peralta, J. Bakardit, S. Garcia and F. Herrera, "A MapReduce Solution for Prototype Reduction in Big Data Classification," Neurocomputing, vol. 150, 2015.
[25]. F. Dornaika and I. Kamal Aldine, "Detrimental Sparse Modeling Representative selection for Prototype Selection," Pattern Recognition, vol. 48, 2015.
[26]. M. Reyes-Sierra and C. A. Coello Coello, "Multi-Objective Particle Swarm Optimizers: A Survey of the State-of-the-Art," International Journal of Computational Intelligence Research, vol. 2, no. 3, pp. 287-308, 2006.
[27]. A. Charlos and C. Coello, "Handling Multiple Objective with Particle Swarm Optimization," IEEE transactions on Evolutionary Computation, vol. 8, no. 3, June 2004.
[28]. E. Ott, Chaos in Dynamic System, Cambridge UK: Cambridge University Press, 2002.
[29]. J. Hamidzadeh, R. Monsefi and H. Sadoghi Yazdi, "DDC: Distance-based Decision Classifier," Nerual Comput & Applic, pp. 1697-1707, vol. 21, 2012.
[30]. V. Hooshmand Moghadam and J. Hamidzadeh, "New Hermite Orthogonal Polynomial Kernel and Combined Kernels in Support Vector Machine classifier," Pattern Recognition, vol. 60, pp. 921-935, 2016.
[31]. A. Asuncion and D.J. Newman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA.
[32]. Sheskin D., Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, 2003.

## زیرنویس‌ها:

[1] Particle Swarm Optimization (PSO)

[2] Radial basis function

[3] 10-fold

[4] Wilcoxon signed-rank test

[5] Two-side

[6] Critical value

[7] Offline

[8] Online