# کاهش ویژگی سیستم های اطلاعاتی ناقص با تئوری مجموعه راف فازی با استفاده از الگوریتم چرخه آب

**وحید نجف پور[۱*]، ناصر سلطانی[۲] و بهروز زادمهر[۳]**

**۱- دانشکده مهندسی برق، دانشگاه پدافند هوایی خاتم الانبیا(ص)، تهران ، ایران**
**۲- دانشکده مهندسی برق، دانشگاه پدافند هوایی خاتم الانبیا(ص)، تهران ، ایران**
**۳- دانشکده مهندسی برق و کامپیوتر، دانشگاه امیر کبیر، تهران ، ایران**

**چکیده**

در سال های اخیر تئوری مجموعه راف به یکی از راه حل های قدرتمند در حل مسئله هوش مصنوعی و داده کاوی تبدیل شده است. کاهش ویژگی در سیستم های اطلاعاتی به معنای حذف ویژگی های غیرضروری و کاهش بعد داده ها است. استفاده از تئوری مجموعه های فازی و الگوریتم چرخ آب می‌تواند در کاهش ویژگی های ناقص سیستم های اطلاعاتی مفید باشد. تئوری مجموعه فازی و نظریه مجموعه های راف دو نظریه متمایز اما مکمل است که با عدم اطمینان در داده ها مواجه می شوند. ویژگی های برجسته هر دو نظریه در محدوده تئوری تنظیم راف فازی قرار گرفته است. هدف از این مقاله ارائه یک رویکرد مجموعه ای فازی بر اساس غلبه بر سیستم های اطلاعاتی با ارزش ناقص است. در این مقاله، یک رویکرد بهینه سازی جدید، که به عنوان الگوریتم چرخه آب شناخته می شود برای حل این مسئلهٔ مورد استفاده قرار گرفته است . برای ارزیابی عملکرد الگوریتم WCA در ۹ مجموعه داده ای شناخته شده از UCI که در جدول نتایج نشان داده شده مورد آزمایش قرار گرفته شده و تست گردید. نتایج آزمایشات، نشان می دهد که راف فازی و الگوریتم پیشنهاد شده نتایج مناسبی ارائه داد که درخـور تأمـل است.

**واژه‌های کلیدی:** الگوریتم چرخه آب، تئوری مجموعهٔ راف ، راف فازی، سیستم اطلاعاتی ناقص،کاهش ویژگی

# Attribute Reduction of Incomplete Information systems based on Fuzzy Rough Set by the Water Cycle Algorithm

**Abstract**

In recent years, rough set theory has become one of the powerful solutions in solving the problem of artificial intelligence and data mining. Feature reduction in information systems means removing unnecessary features and reducing data dimension. Using the theory of fuzzy sets and the water wheel algorithm can be useful in reducing the incomplete features of information systems. Fuzzy set theory and rough set theory are two distinct but complementary theories that deal with uncertainty in data. The salient features of both theories are within the scope of fuzzy rough tuning theory. The purpose of this paper is to present a fuzzy set approach based on overcoming incomplete value information systems. In this article, a new optimization approach, which is known as water cycle algorithm, is used to solve this problem. To evaluate the performance of the WCA algorithm, it was tested on 9 datasets selected from UCI, which are shown in the results table. The results of the experiments show that the fuzzy rough and the proposed algorithm gave appropriate results, which is reasonable.

**Key words:** Water cycle algorithm, rough set theory, fuzzy rough, incomplete information system, feature reduction

ایمیل نویسنده مسئول: v.najafpoor1@gmail.com

## ۱-مقدمه

در اختیار داشتن حجم زیاد داده ها در دنیای واقعی نمایانگر این است که ما از نظر مقدار داده غنی ، و از نظر مقدار دانش، فقیر هستیم و به همین دلیل متدهای تحلیل داده کاوی جهت استخراج دانش از دل داده ها معرفی شده اند [۱]. همچنین با توجه به ذخیرهٔ دهها، صدها یا حتی هزاران ویژگی در پایگاه دادهٔ برنامه‌های کاربردی، بحث کاهش ویژگی در جهت تحلیل این داده‌ها امری ضروری و انکار ناپذیر است[۳-۲]. تئوری مجموعه‌های فازی (Fuzzy Set Theory) یک روش ریاضی است که برای تعامل با عدم قطعیت در داده ها و اطلاعات استفاده می شود. یک مجموعه فازی به دنباله‌ای از اعداد فازی گفته می‌شود که هر عدد فازی نشان دهنده احتمال عضویت عنصر در آن مجموعه است. کاهش ویژگی بر مبنای تئوری مجموعه‌های فازی (Fuzzy Set Reduction)یکی از روش‌های مورد استفاده در پردازش داده‌های فازی است که به منظور کاهش بعد داده‌ها و حذف ویژگی‌های تکراری یا غیرضروری، استفاده می‌شود. در مجموعه‌های فازی، ویژگی‌های تکراری یا غیرضروری ممکن است باعث ایجاد ابهام در تصمیم‌گیری و حفظ سطح دقت مدل‌سازی نشوند. با استفاده از روش کاهش ویژگی، امکان حذف این ویژگی‌ها و در نتیجه سرعت بخشیدن به فرآیند آموزش مدل‌های پیش‌بینی و تصمیم‌گیری فراهم می‌شود. تحقیقات پیشین در زمینه کاهش ویژگی بر مبنای تئوری مجموعه‌های فازی، شامل روش‌های مختلفی مانند الگوریتم جستجوی ژنتیک، الگوریتم هایبریدی، الگوریتم تکاملی، روش های خوشه بندی و غیره می‌شود. هر یک از این روش‌ها با استفاده از مفاهیم تئوری مجموعه‌های فازی و الگوریتم‌های خاص خود، قادر به کاهش بعد داده‌ها و یافتن ویژگی‌های مهم و حذف ویژگی‌های تکراری هستند. در بررسی پیشینهٔ نظری کاهش ویژگی بر مبنای تئوری مجموعهٔ راف، دو رویکـرد اساسـی را در مقابل داریم. ابتدا رویکردهایی که کلاسیک هستند و از همان نسخهٔ اصلی تئوری مجموعهٔ راف استفاده میکنند. رویکرد دوم، توسعه یافتهٔ رویکرد قبلی است که الگوریتم های اکتشـافی کـاهش ویژگی را ارائه میکند. نکتهٔ مهم در رویکردهای مختلف کاهش ویژگی بر مبنای تئوری مجموعهٔ راف، نوع سیستم های اطلاعاتی است. در اغلب پژوهش های اجراشده بر سیستمهـای اطلاعـاتی کامل تمرکز شده است؛ اما در دنیای واقعی، مجموعهٔ داده هـای جمـع آوری شـده همیشـه کامـل نیستند. این مشکل بخصوص در سامانه های اطلاعات نظامی بیشتر به چشم می خورد. پس چالش اساسی کاهش ویژگی مبتنی بر تئوری مجموعهٔ راف، کار روی سیسـتم هـای اطلاعاتی ناقص است.

کاهش ویژگی در سیستم های اطلاعاتی به معنای حذف ویژگی های غیرضروری و کاهش بعد داده ها است. استفاده از تئوری مجموعه های فازی و الگوریتم چرخ آب می‌تواند در کاهش ویژگی های ناقص سیستم های اطلاعاتی مفید باشد . برای این منظور، ابتدا باید داده های مورد نظر را به شکل مجموعه های فازی تبدیل کنیم. سپس با استفاده از الگوریتم چرخ آب، ویژگی های ناقص را شناسایی کرده و حذف می‌کنیم. این الگوریتم با استفاده از خصوصیت های مجموعه های فازی، ورودی هایی را که در محدوده تعریف شده قرار ندارند را شناسایی می کند و آن ها را حذف می کند. به عنوان مثال، در صورتی که داده هایی از سیستم های اطلاعاتی را داریم که شامل ویژگی های نام، سن، آدرس، تحصیلات و شغل هستند، می‌توانیم با استفاده از تئوری مجموعه های فازی و الگوریتم چرخ آب، ویژگی های ناقص را شناسایی کرده و حذف کنیم. به عنوان مثال، اگر ویژگی "آدرس" در بیشتر موارد از داده های ما مقدار نداشته باشد، می‌توانیم این ویژگی را حذف کنیم تا اطلاعات مورد نیاز را با دقت بیشتری به دست آوریم.

کاهش ویژگی به عنوان روش کاهش تعدادی از ویژگی ها در یک سیستم اطلاعات شناخته شده است و عمل آن یک مرحله حیاتی از پردازش داده کاوی است. در فرآیند کاهش ویژگی، حداقل مجموعه ای از صفات انتخاب شده است. افزایش حجم پایگاه داده‌ها مسئله کاهش ویژگی را ایجاد و پراهمیت می‌کند. در بررسی پایگاه‌های داده، گاهی با سیستم‌های اطلاعاتی ناقص مواجه هستیم. یک سیستم اطلاعاتی ناقص به جداولی از داده‌ها اطلاق می‌شود که برخی درایه‌های صفات آن مقداری ندارند. یک رویکرد برای تحلیل این سیستم‌ها این است که سطر مربوط به مقادیر ناقص را حذف کنیم. اما این رویکرد مناسب نیست زیرا اندازهٔ داده‌ها را کاهش می‌دهد که ممکن است اطلاعات مفیدی از بین برود. یک روش دیگر برای تحلیل این نوع سیستم‌های اطلاعاتی این است که داده‌های ناموجود را بصورت‌های مختلفی مقداردهی کنیم. برای مثال از تحلیل‌های آماری استفاده کرده و داده‌های ناموجود را حدس بزنیم. هر دو رویکرد برای کامل شدن سیستم‌های اطلاعاتی کاربرد دارند اما بکارگیری یک روش مناسب برای کامل کردن سیستم‌های اطلاعاتی ناقص خود یک موضوع تحقیق پیچیده است.

یکی از روشهای کنترل و مدیریت داده ها با حجم زیاد تئوری مجموعه راف است. در این تئوری با تفکیک ویژگی ها به ویژگی های شرط و تصمیم گیری،به کمک تئوری نظریه مجموعه ها و رویکردهای ریاضی اقدام به شناسایی رابطه و وابستگی بین این دو گروه ویژگی ها می نماید. تئوری مجموعه راف تلاش می کند،یک زیر مجموعه ای با حداقل ویژگی از مجموعه ویژگی های اصلی با استفاده از روش تولید تصادفی ایجاد کند. اشکال اصلی روش تولید تصادفی، این است که تنها برای مجموعه داده های کوچک مناسب است و روش بسیار وقت گیر است. یک عامل مهم درفرایند انتخاب ویژگی ها، جستجو در تمام فضای زیر مجموعه ویژگی های ممکن است. جستجو جامع در این فضا ، معمولا امکان پذیر نیست. روش کلاسیک معمولا منجر به بهینه

محلی می شود. به منظور این نقطه ضعف ، استراتژی های جستجو مبتنی بر الگوریتم های فرابتکاری پیشنهاد شده است. هدف این الگوریتم ها ، رسیدن به بهترین جواب یافت شده در کوتاهترین زمان ممکن می باشد. امروزه با بزرگ و پیچیده شدن مسائل ، استقبال از روش های فراابتکاری بطور چشم گیری افزایش یافته است. این امر باعث شده بسیاری از محققان به سمت بررسی مشکلات کاهش ویژگی با استفاده از انواع مختلفی از الگوریتم ها ، مانند : الگوریتم ژنتیک[۴] ، بهینه سازی ازدحام ذرات[۵]، جستجو پراکنده[۶] ، بهینه سازی کلونی مورچه ها و زنبور عسل[۹-۷] شبیه سازی حرارتی [۱۰] تشویق گردند.

اغلب روش‌های بهینه‌سازی مبتنی بر هوش مصنوعی همانند الگوریتم ژنتیک، شبیه‌سازی فرآیندهای طبیعی می‌باشند. یکی از دلایل این امر ملموس بودن، سادگی فرموله‌کردن و درک تکامل این الگوریتم‌ها است. الگوریتم چرخه آب، با الهام‌گیری از فرایند چرخه آب در طبیعت نسبت به دیگر روش‌های مطرح شده در این زمینه دارای توانایی بالایی بوده و از سرعت مناسبی نیز برخوردار می‌باشد. در اغلب پژوهش‌های انجام شده، سیستم‌های اطلاعاتی کامل مورد هدف بوده‌اند. اما در دنیای واقعی، مجموعه داده‌های جمع آوری شده همیشه بصورت کامل نیستند. پس یک چالش اساسی کاهش ویژگی ، کار بر روی سیستم‌های اطلاعاتی ناقص می‌باشد. درگذشته، روش‌هایی برای حل مسئله کاهش ویژگی معرفی شده‌اند؛ در این مقاله از تئوری مجموعه راف برای سیستم های اطلاعاتی ناقص که از جمله مسائل NP-hard بشمار می‌آید، استفاده می‌شود. رویکرد مجموعه راف را به سیستم های اطلاعات ناقص، یک سیستم اطلاعاتی ناقص به جداولی از داده‌ها اطلاق می‌شود که برخی درایه‌های صفات آن مقداری ندارند یعنی سیستم هایی که مقادیر مشخصه برای اشیاء ممکن است ناشناخته ( missing، null) باشند، ارائه می دهیم. نگرانی اصلی ما برای یافتن قوانین از چنین سیستم هایی است[۱۱].

درادامه این مقاله در بخش ۲ ، تئوری مجموعه راف ، راف ناقص و راف فازی به طور دقیق تشریح شده است. ساختار الگوریتم چرخه آب نحوه سازگاری آن با مسئله کاهش ویژگی در بخش ۳ و نتایج پیاده سازی الگوریتم مذکور در بخش ۴ ارائه شده است. نهایتا ، خلاصه و نتیجه گیری و رویکرد پیشنهادی برای کاهش ویژگی و دستاوردهای آن در بخش ۵ جمع بندی شده است.

## ۲- تئوری مجموعه راف(RST)

در اختیار داشتن حجم زیاد داده ها در دنیای واقعی نمایانگر این است که ما از نظر مقدار داده غنی، و از نظر مقدار دانش، فقیر هستیم و به همین دلیل متدهای تحلیل داده کاوی جهت استخراج دانش از دل داده ها معرفی شده اند. تئوری مجموعه راف یکی از این روش هاست که توسط محقق لهستانی پاولاک در سال ۱۹۸۰ معرفی گردیده است [۴]. موفقیت RST و توسعه آن را می توان به دلیل قابلیت ها و مزایای این تئوری دانست که مهمترین آن ها به شرح زیر است:این تئوری تنها بر پایه داده های خام بنا شده و نیاز به اطلاعات خارجی ندارد.RST نه تنها برای تحلیل ویژگی های کیفی، بلکه برای مطالعه ویژگی های کمی نیز به کار گرفته می شود.این تئوری امکان کاهش معیارهای اضافی را داشته و الگوریتم هایی ساده برای این کار ارائه می دهد.نتیجه RST یکسری قوانین تصمیم گیری است که از مدل های مختلف آن استخراج می شود.

RST نیازی به اصلاح ناسازگاری ها نداشته و می تواند قوانین متناقض را به قوانین قطعی و غیر قطعی تقسیم کند؛ به طوری که قوانین ناشی از این مدل ها به راحتی قابل درک باشند. نقطه آغاز RST مجموعه داده هاست که معمولا در قالب یک جدول، سازمان داده شده و تحت نام سیستم اطلاعاتی یا پایگاه داده معرفی می شوند. اصلی ترین عملیات در RST تقریب بالا و پایین است و این تقریبها برای تشخیص اجزای کاملا وابسته یا نیمه وابسته در جدول اطلاعات به کار می روند.تقلیل داده نیز یکی از مهمترین مفاهیم این تئوری است، که در ادامه هر یک از این مفاهیم اساسی تشریح می شود[۱۲]. با این حال یکی از محدودیت‌های اصلی تئوری راف این است که مبتنی بر داده های گسسته است. در تئوری راف کالاسیک، رسیدگی به داده های نویز دار و داده های حقیقی و پیوسته ممکن نیست. در نتیجه بسیاری از محققان، به دنبال تعمیم تئوری راف بودند، بطوری که بتوان از آن برای رسیدگی به مجموعه داده های نویز دار وپیوسته استفاده کرد[۱۳]. یکی از ایدوئولوژی‌های بکار رفته در تئوری مجموعه راف این است که فقط از داده‌های تولید شده استفاده می‌کند؛ در حالی که در سایر روش های موجود برای حل مسئله علاوه بر داده های تولید شده، نیاز به یک دانش تکمیلی دیگری هم دارند.

### ۲-۱- فرایند کاهش راف فازی

مجموعه راف فازی که بر پایه انتخاب ویژگی است، براساس مفهوم پایین تقریب فازی، عملیات کاهش مجموعه داده‌ها که شامل ویژگی-های حقیقی می شود، را آغاز می‌کند. عملگر منطقه مثبت در حالت مجموعه راف سنتی تنها بوسیله عملگر پایین تقریب تعریف می‌شود، در حالی که در حالت راف فازی، به همان حالت است، تنها درجه عضویت شی $x \in u$ به آن اضافه شده[۱۴]، در نتیجه منطقه مثبت در حالت فازی به این صورت است[۱۵]:

$$\mu_{POS_P(Q)}(x) = sup_{x \in U/Q} \mu_{\underline{P}X}(x). \qquad (۱)$$

با استفاده از تعریف منطقه مثبت، تابع وابستگی راف فازی به این صورت تعریف می‌شود[۱۱]:

$$\gamma_P(Q) = \frac{\left|\mu_{POS_P(Q)}(x)\right|}{|U|} = \frac{\sum_{x \in U} \mu_{POS_P(Q)}(x)}{|U|}. \qquad (۲)$$

در حالت crisp مجموعه راف، وابستگی Q بر روی P یعنی، نسبت اشیائی که از کل مجموعه داده قابل تشخیص هستند. اما در حالت فازی، وابستگی یعنی تعیین کاردینالیتی فازی $\mu_{POS_P(Q)}(x)$، تقسیم بر تعداد اشیاء موجود در مجموعه مرجع. اگر روند مجموعه راف فازی بدرستی نتیجه دهد، این امکان وجود دارد که بتوان آن را برای ویژگی های با خاصیت های متعدد استفاده کرد و وابستگی بین زیر مجموعه های مختلف از مجموعه ویژگی های اصلی را می‌توان براحتی پیدا کرد. برای مثال، می‌توان درجه وابستگی ویژگی‌های شرطی $P = \{a, b\}$ را با توجه به ویژگی تصمیم گیری مورد نظر بدست آورد. در حالت crisp، عملگر $U/P$ دربردارنده گروهی از اشیاء است که با توجه به هر دو ویژگی a,b نوشته می‌شود. اما در حالت فازی اشیاء ممکن است به بسیاری از کلاس‌های هم‌ارزی تعلق داشته باشند، بنابراین حاصل ضرب دکارتی $U/IND(\{a\})$ و $U/IND(\{b\})$ باید در نظر گرفته شود. در حالت کلی:

$$U/P = \otimes \{a \in P: U/IND(\{a\})\} \qquad (۳)$$

هر مجموعه‌ای که در $U/P$ است، مشخص کننده کلاس هم‌ارزی است. برای مثال اگر

$$P = \{a, b\}, \quad U/IND(\{a\}) = \{N_a, Z_a\},\ U/IND(\{b\}) = \{N_b, Z_b\}, \qquad (۴)$$

آنگاه:

$$U/P = \{N_a \cap N_b, N_a \cap Z_b, Z_a \cap N_b, Z_a \cap Z_b\} \qquad (۵)$$

میزان اینکه یک شیء متعلق به کلاس هم‌ارزی است، با استفاده از ترکیب کلاس‌های هم‌ارزی فازی تشکیل دهنده، محاسبه می شود. مثلا:

$$F_i, i = 1,2, \dots, n$$

$$\mu_{F_1 \cap \dots \cap F_n}(x) = \min\left(\mu_{F_1}(x), \mu_{F_2}(x), \dots, \mu_{F_n}(x)\right) \qquad (۶)$$

### ۲-۲-سیستم‌های اطلاعاتی ناقص

یک سیستم اطلاعاتی معمولاً به فرم زیر ارائه می‌شود:

$$IS = (U, A, \{V_a\}, f_a)_{a \in A} \qquad (۷)$$

که در رابطهٔ ۱، U، تعداد متناهی از اشیای دارای مقدار و A، تعداد محدودی از ویژگی‌های دارای مقدار می‌باشند. $V_a$، مجموعهٔ مقادیر ویژگی‌ها و $f_a$، یک تابع اطلاعاتی از سیستم اطلاعات است که $V_a$ را به U نگاشت می‌کند. اگر در سیستم اطلاعاتی یک شی x و یک ویژگی به نام a وجود داشته باشد که تابع $f_a(x)$ مقداری نداشته باشد، آنگاه به آن سیستم اطلاعاتی ناقص گویند.

مثال: جدول شمارهٔ ۱ نمونه‌ای از سیستم اطلاعاتی ناقص را نشان می‌دهد. این سیستم شامل اطلاعات مربوط به چند فرد است که ویژگی‌های آن‌ها شامل قد، وزن و سن می‌باشد. هر کدام از مقادیر مربوط به این ویژگی‌ها با مقادیر کیفی مشخص شده‌اند.

**جدول۱- نمونه‌ای از یک سیستم اطلاعاتی ناقص**

| افراد | قد | وزن | سن |
|---|---|---|---|
| شمارهٔ ۱ | بلند | زیاد | خیلی بالا |
| شمارهٔ ۲ | کوتاه | زیاد | بالا |
| شمارهٔ ۳ | * | کم | پایین |
| شمارهٔ ۴ | بلند | زیاد | خیلی بالا |
| شمارهٔ ۵ | خیلی بلند | * | خیلی بالا |
| شمارهٔ ۶ | بلند | متوسط | * |
| شمارهٔ ۷ | * | کم | بالا |
| شمارهٔ ۸ | بلند | * | خیلی پایین |

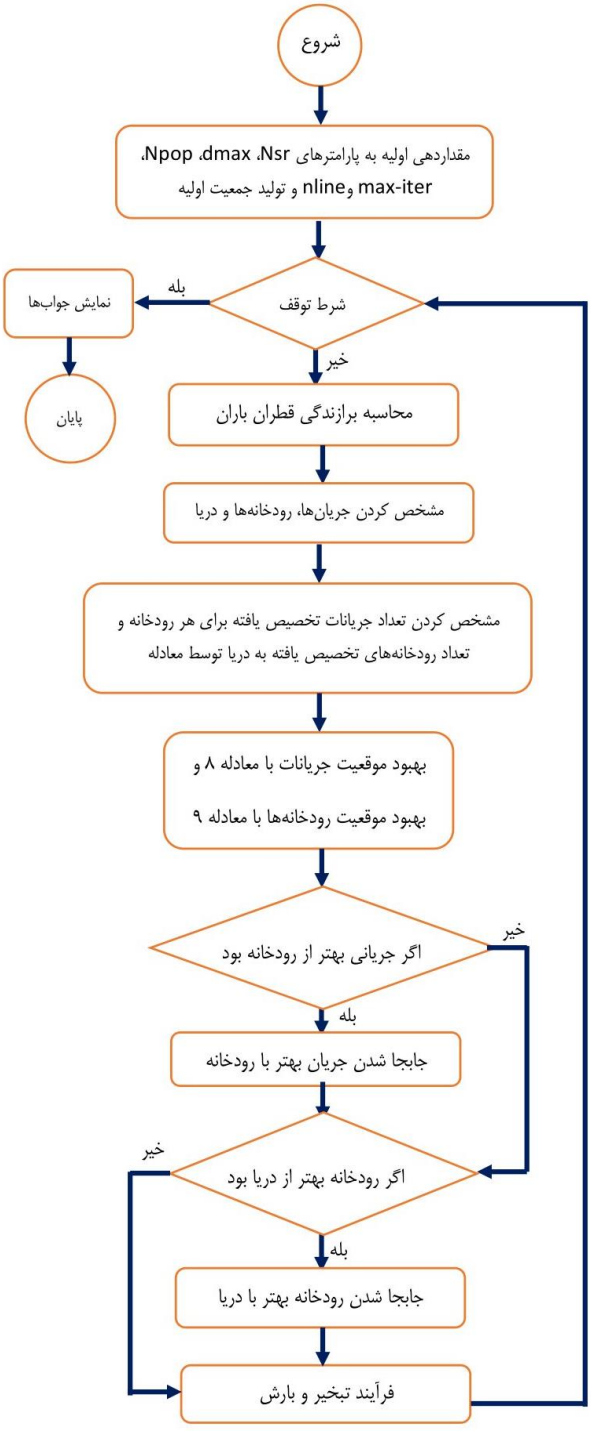
## ۳- الگوریتم چرخه آب برای کاهش ویژگی

مجموعه راف فازی برای انتخاب ویژگی، به منظور کاهش ابعاد داده های حجیم و استفاده از ویژگی های مهم تر برای تحلیل و پردازش داده ها استفاده می شود. در این روش، با استفاده از مفهوم پایین تقریب فازی، ابعاد داده ها با حفظ دقت مورد نیاز، کاهش می یابد. برای این منظور، ابتدا یک مجموعه اولیه از ویژگی ها به عنوان ورودی در نظر گرفته می شود. سپس با استفاده از الگوریتم های مختلفی مانند الگوریتم ژنتیک، الگوریتم جستجوی محلی و غیره، فضای ویژگی ها جستجو شده و ویژگی های بی اهمیت حذف می شوند. در این روش، با توجه به مفهوم پایین تقریب فازی، برای هر ویژگی، یک مقدار عضویت در نظر گرفته می شود. سپس با استفاده از الگوریتم های کاهش ابعاد، ویژگی هایی که دارای مقدار عضویت پایینی هستند، حذف می شوند.

به عنوان مثال، فرض کنید یک مجموعه داده شامل ۱۰ ویژگی باشد. با استفاده از مجموعه راف فازی، می توانیم تعدادی از ویژگی های مهم را انتخاب کنیم و ویژگی های بی اهمیت را حذف کنیم. برای مثال، اگر ویژگی های ۲ و ۶ و ۹ دارای مقدار عضویت پایینی هستند، آن ها حذف می شوند و مجموعه داده به صورت ۷ ویژگی کاهش می یابد. در نهایت، با استفاده از مجموعه راف فازی، می توانیم تنها با استفاده از ویژگی های مهم تر داده ها را تحلیل کنیم و همچنین در زمان اجرای الگوریتم ها صرفه جویی شود.

الگوریتم چرخه آب ، به عنوان یک روش بهینه‌سازی جدید که توسط اسکندر و همکاران در سال ۲۰۱۲ معرفی شده است، این الگوریتم همانند چرخه آب در طبیعت، با الگوبرداری از تبخیر آب از اقیانوس تشکیل ابر، تشکیل رودخانه و همچنین الگوبرداری از سرازیر شدن آب از گودال‌ها در طبیعت ارائه شده است که قابلیت بالایی در فرار از بهینه محلی و همچنین سرعت زیادی در رسیدن به بهینه سراسری دارد [۷]. مشابه سایر الگوریتم‌های متاهیورستیک، این الگوریتم هم با یک جمعیت اولیه آغاز می‌شود که قطرات باران نامیده می‌شوند. در ابتدا ما فرض می‌کنیم باران یا تگرگ داریم و بهترین شخص را به عنوان دریا در نظر می‌گیریم. پس تعدادی از قطرات باران را که موقعیت بهتری دارند، به عنوان رودخانه و بقیه قطرات باران را به عنوان جریان‌هایی که به سمت رودخانه و دریا جاری هستند، در نظر می‌گیریم. مقدار آبی

که از یک جریان به رودخانه یا دریا می‌ریزد، متفاوت است با جریانی دیگر، به علاوه رودخانه‌ها به دریا که در مکان پایین‌تری قرار دارد، سرازیر می‌شوند.در شکل ۱ فلوچارت چرخه آب را آورده ایم.




**شکل ۱- فلوچارت الگوریتم چرخه آب**

### ۳-۱-گام اول: ایجاد جمعیت اولیه و مقداردهی اولیه به پارامترها

ایجاد جمعیت اولیه: برای حل یک مسئله بهینه‌سازی با استفاده از روش‌های متاهیورستیک مبتنی بر جمعیت لازم است که مقادیر متغیرهای مسئله همانند یک آرایه تشکیل شوند. در PSO و GA، این آرایه، موقعیت ذره و کروموزوم نامیده می‌شود. در این روش به آن قطره‌ی باران می‌گویند. در یک مسئله بهینه‌سازی N بعدی، یک قطره-ی باران، آرایه‌ی $1 \times N$تایی است که این آرایه به صورت معادله (۸)تعریف شده است:

$$Raindrop = [x_1, x_2, x_3, \dots, x_n] \qquad (۸)$$

مقداردهی اولیه: به تعداد Npop قطره‌ی باران ایجاد می‌شود. برای این کار، یک ماتریس با سایز $N_{pop} \times N_{var}$ تولید شده که سطرها تعداد اعضای جمعیت و ستون‌ها تعداد متغیرها هستند و درواقع نشان-دهنده جمعیتی از قطرات باران است. مقادیر متغیرها با توجه به مسئله‌ی مورد بررسی می‌تواند عدد حقیقی، اعشاری یا گسسته باشد. از این‌رو، ماتریس X به صورت تصادفی تولید شده و به صورت معادله(۹) است:

$$Population\ of\ raindrops = \begin{bmatrix} Raindrop_1 \\ Raindrop_2 \\ Raindrop_3 \\ \vdots \\ Raindrop_{Npop} \end{bmatrix} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_{Nvar}^1 \\ x_1^2 & x_2^2 & \cdots & x_{Nvar}^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{Npop} & x_2^{Npop} & \cdots & x_{Nvar}^{Npop} \end{bmatrix} \qquad (۹)$$

### ۳-۲-گام دوم: ارزیابی جمعیت

هزینه‌ی یک قطره‌ی باران به‌صورت معادله (۱۰)ارزیابی می‌شود:

$$C_i = Cost_i = f(x_1^i, x_2^i, \dots, x_{Nvar}^i)\ i = 1,2,3, N_{pop} \qquad (۱۰)$$

### ۳-۳-گام سوم: انتخاب دریا و رودخانه

به تعداد Nsr از بهترین افراد (که دارای کمترین مقدار Cost هستند) به عنوان دریا و رودخانه‌ها انتخاب می‌شوند. قطره‌ای که کمترین مقدار را در میان دیگران دارد، به عنوان دریا در نظر گرفته می‌شود. درحقیقت، Nsr مجموعه‌ای از رودخانه‌ها و یک دریا است که این تعداد از معادله‌ی (۱۱) محاسبه می‌شود:

$$N_{sr} = Number\ of\ Rivers + \frac{1}{sea} \qquad (۱۱)$$

باقی جمعیت(قطرات باران)، جریان‌ها را شکل می‌دهند. این جریانات به سمت رودخانه‌ها یا مستقیماً به سمت دریا جاری می‌شوند.

$$N_{Raondrops} = N_{pop} - N_{sr} \qquad (۱۲)$$

برای تخصیص قطرات باران به رودخانه‌ها یا دریا با توجه به شدت جاری شدن، از معادله‌ی (۱۳) استفاده می‌کنیم:

$$NS_n = round\left\{\left|\frac{Cost_n}{\sum_{i=1}^{Nsr} Cost_i}\right| \times N_{Raindrops}\right\}.n = 1,2,\dots,N_{sr} \qquad (۱۳)$$

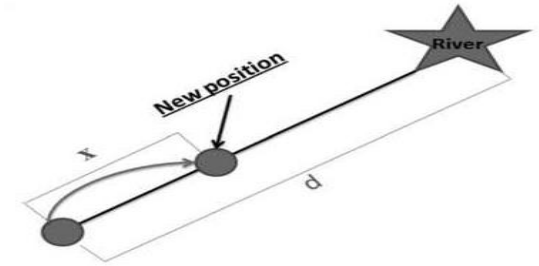
که $NS_n$ تعداد جریاناتی است که به رودخانه‌های خاص یا دریا جاری می‌شوند.

### ۳-٤-گام چهارم:. بهبود موقعیت

همه رودخانه‌ها و جریان‌ها به دریا ختم می‌شوند (بهترین نقطه‌ی بهینه). یک جریان در طول خط اتصالش با رودخانه به سمت رودخانه جاری می‌شود، توسط فاصله‌ای که به صورت تصادفی با معادله (۱۴)انتخاب می‌شود:

$$X\epsilon\ (0\ , C \times d), C > 1 \qquad (۱۴)$$

که C یک مقدار بین ۱ و ۲ است (نزدیک به ۲). بهترین مقدار برای C می‌تواند ۲ باشد. فاصله‌ی بین جریان (نهر) و رودخانه d در نظر گرفته شده است. مقدار X در معادله‌ی (۱۴) با یک عدد تصادفی توزیع شده بین $(0\ , C \times d)$ مرتبط است. مقدار C بیشتر از یک باعث می‌شود که جریان‌ها (نهرها) قادر باشند، در جهت‌های مختلف به سمت رودخانه‌ها جاری شوند.



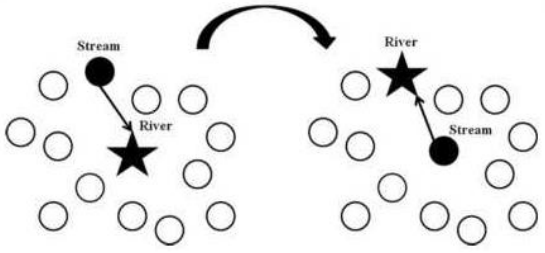

**شکل ۲- دید سمانتیک از جاری شدن نهر ها به یک رودخانه دیگر**

این مفهوم هم‌چنین ممکن است در جریان رودخانه‌ها به سمت دریا هم استفاده شود. بنابراین موقعیت جدید جریان‌ها و رودخانه‌ها به‌صورت معادله (۱۵) و (۱۶) بدست می‌آید:

$$X_{stream}^{i+1} = X_{stream}^{i} + rand \times C\ \times \left(X_{River}^{i} - X_{stream}^{i}\right) \qquad (۱۵)$$

$$X_{River}^{i+1} = X_{River}^{i} + rand \times C\ \times \left(X_{sea}^{i} - X_{River}^{i}\right) \qquad (۱۶)$$

که rand یک عدد تصادفی توزیع شده به صورت یکنواخت است بین [۱و۰]. اگر راه‌حل داده شده توسط یک جریان بهتر از رودخانه‌ی متصل به آن باشد، موقعیت‌های رودخانه و نهر (جریان) جابجا می‌شود (یعنی رودخانه نهر می‌شود و نهر رودخانه می‌شود). چنین جابجایی‌ای می‌تواند به طور مشابه برای رودخانه‌ها و دریا نیز اتفاق بیفتد. شکل۳، جابجایی جریانی را که در میان دیگر جریان‌ها بهترین است را با رودخانه نشان می‌دهد.




**شکل۳-مبادله موقعیت جریان‌ها و رودخانه‌ها که ستاره نماد رودخانه و دایره با رنگ سیاه نماد بهترین جریان در بین دیگر جریان‌ها است.**

### ۳-۵-گام پنجم: تبخیر

تبخیر یکی از فاکتورهای بسیار مهمی است که می‌تواند از همگرایی زودرس الگوریتم جلوگیری کند. آب تبخیر شده در اتمسفر برای تشکیل ابر در اتمسفرِ سردتر، متراکم می‌شود و ابر آب را به شکل باران به زمین برمی‌گرداند. باران جریان‌های (نهرهای) جدید را ایجاد می‌کند و جریان‌های جدید به سمت رودخانه‌ها یا دریا جاری می‌شوند. در روش ارائه شده فرآیند تبخیر، باعث می‌شود که آب دریا مانند رودخانه/نهر که به سمت دریا جاری هستند، تبخیر شود. این فرض به منظور جلوگیری از قرار گرفتن در بهینه محلی در نظر گرفته شده است. شبه کد زیر را در نظر بگیرید:

If $\left|X_{sea}^{i} - X_{River}^{i}\right| < d_{max} i = 1,2,3, \ldots, N_{sr} - 1$

Evaaporation and raining process end

dmax

یک عدد کوچک نزدیک به صفر است. اگر فاصله بین رودخانه و دریا کمتر از dmax باشد، نشان می‌دهد که رودخانه به دریا ملحق شده است. در این وضعیت، فرایند تبخیر انجام می‌شود و همان‌طورکه در طبیعت می‌بینیم بعد از میزان کافی عمل تبخیر، رطوبت آغاز خواهد شد.

یک مقدار بزرگ برای dmax جستجوی نزدیک به دریا را کاهش می‌دهد، در حالی که مقدار کوچک باعث می‌شود، جستجو نزدیک دریا شدت یابد. بنابراین dmax شدت جستجو در نزدیکی دریا را کنترل می‌کند (راه‌حل بهینه). مقدار dmax به صورت تطبیقی به صورت زیر کاهش می‌یابد:

$$d_{max}^{i+1} = d_{max}^{i} - \frac{d_{max}^{i}}{\max iteration} \qquad (۱۷)$$

### ۳-٦-گام ششم: فرایند بارش

بعد از عمل فرآیند تبخیر، بارش انجام می‌شود. در فرآیند بارندگی قطرات باران جدید جریان‌هایی را در مکان‌های مختلف شکل می‌دهند (شبیه عملگر جهش در الگوریتم ژنتیک). برای مشخص کردن مکان-های نهرهای جدید و تولید شده، معادله‌ی(۱۸) استفاده می‌شود:

$$x_{stream}^{new} = LB + rand \times (UB - LB) \qquad (۱۸)$$

LB و UB محدوده‌ی پایین و بالایی هستند که توسط مسئله تعریف شده‌اند. مجدداً، قطره‌ی بارانی که جدیداً تشکیل شده است، به عنوان یک رودخانه که به سمت دریا می‌رود، در نظر گرفته می‌شود. بقیه‌ی قطرات جدید برای تشکیل نهرهای جدیدی که به سمت رودخانه یا دریا جاری هستند، فرض شده‌اند. به‌منظور افزایش نرخ همگرایی و عملکرد محاسباتی الگوریتم معادله‌ی (۱۹) فقط برای جریان‌هایی که به سمت دریا جاری هستند در نظر گرفته شده است. این معادله قصد دارد کاوش در نزدیکی دریا را بهبود بخشد.

$$X_{stream}^{new} = X_{sea} + \sqrt{\mu}\ \times randn(1, N_{var}) \qquad (۱۹)$$

که $\mu$ یک ضریب است که محدوده‌ی ناحیه‌ی جستجو نزدیک دریا را نشان می‌دهد. randn یک عدد تصادفی توزیع شده‌ی یکنواخت است. مقادیر بزرگ‌تر برای $\mu$ امکان خروج از ناحیه‌ی موجه را افزایش می‌دهد. به عبارت دیگر مقادیر کوچک‌تر برای $\mu$ باعث می‌شود الگوریتم در ناحیه‌ی کوچک‌تر نزدیک دریا به جستجو بپردازد. مقدار مناسب برای $\mu$، ۰.۱ است.

راه‌حل الگوریتم چرخه آب مجموعه‌ای از راه‌حل ها (موقعیت‌ها) و یا مجموعه حالاتی از کارها را معرفی راه‌حل می گویند. برای شروع کار چرخه آب تنظیم راه‌حل کار مهم و ضروری است. چگونگی نمایش و ارائه راه‌حل از روی مشکلی که باید حل شود، تعیین می‌گردد. در واقع این امر، این اجازه را می‌دهد که یک راه‌حل به راحتی ساخته و به سرعت مورد ارزیابی قرار گیرد. الگوریتم چرخه آب مبتنی بر جمعیت است، پس با مجموعه ای از راه‌حل‌ها کار خود را شروع می‌کند. که تعداد این راه‌حل‌ها با تعداد ویژگی‌های شرطی $|C|$ برابر است. در این الگوریتم هر راه‌حل معادل یک قطره باران است. که به حالات مختلف تولید می‌شود. تولید راه‌حل در الگوریتم چرخه آب برای حل مسئله کاهش ویژگی با توجه به اینکه از کدام حالت تئوری راف استفاده شود، متفات است. الگوریتم چرخه آب از راه‌حل‌های باینری استفاده می‌کند. ساخت راه‌های اولیه که در این مرحله صورت می‌گیرند، اصولا بصورت تصادفی انجام می‌گیرد. برخی از راه‌حل‌های اولیه‌ی تولید شده، دارای درجه وابستگی یک می‌باشند. ویژگی‌ها که در اینجا همان راه‌حل‌ها هستند، با برداری به نام Y نمایش داده می‌شوند. در میان تمام ویژگی‌ها از بردار Y، اگر خانه $y_I$ از بردار Y که $i=1,\ldots,|C|$، مقدارش برابر با یک باشد، آنگاه ویژگی $i^{th}$ به عنوان بخشی از زیر مجموعه راه‌حل‌ها انتخاب می‌شود؛ و اگر مقدار خانه از بردار Y برابر با صفر باشد، آن ویژگی بخشی از راه‌حل محسوب نمی‌شود. در این روش ویژگی‌ها بصورت تصادفی انتخاب می‌شوند.

### ۳-۷-شرط خاتمه

شرط خاتمه در الگوریتم را می‌توان تعداد دفعات تکرار الگوریتم یا رسیدن به بهترین زیر مجموعه حداقلی تعریف کرد. در این پژوهش ما شرط پایان الگوریتم را مطابق با سایر پژوهش‌های انجام شده در این زمینه [20، 7]، تعداد دفعات تکرار الگوریتم و برابر با ۲۰ بار تعریف می‌کنیم.

## ۴- نتایج پیاده سازی

برای ارزیابی عملکرد الگوریتم WCA در ۹ مجموعه داده ای شناخته شده از UCI که در سایت http://www.ics.uci/mlearn قابل دسترسی است که در جدول (۲) نشان داده شده مورد آزمایش قرار گرفته شده است. در ادامه نتایج بدست آمده کاهش ویژگی با استفاده از الگوریتم چرخه آب (WCA)با نتایج به دست آمده با الگوریتم های معروف ابتکاری ژنتیک(GA) وجستجوی هارمونی(HS) در جدول (۳) آمده است.این نتایج بهترین از ۲۰ بار تکرار هر الگوریتم است که تعداد ویژگی های کاهش یافته را نشان می دهد.برای برنامه نویسی این الگوریتم ها از متلب Matlab R2017a استفاده شده است. الگوریتم ها بر روی یک کامپیوتر با حافظه اصلی ۴ گیگابایت و پردازنده دو هسته‌ای ۲.۲ گیگا هرتز اینتل اجرا شده انددر همه ی این الگوریتم ها مقدار جمعیت اولیه برابر ۱۰۰ وشرط توقف نیز تعداد تکرار ۲۰ در نظر گرفته شده است.

**جدول ۲-داده های پایگاه داده UCI**

| نام دیتاست | تعداد ویژگی ها | تعداد سطرها |
|---|---|---|
| **Cleveland** | 297 | 14 |
| **Glass** | 214 | 10 |
| **Heart** | 270 | 14 |
| **Ionosphere** | 230 | 35 |
| **Olitos** | 120 | 26 |
| **Water 2** | 390 | 39 |
| **Water3** | 390 | 39 |
| **Web** | 149 | 2557 |
| **Wine** | 178 | 14 |

## ۵-نتیجه‌گیری و بحث

نتایج برآمده از متد ارائه شده، بخوبی نمایانگر توانایی های الگوریتم چرخه آب در امر کاهش ویژگی سیستم های اطلاعاتی ناقص میباشد. با بررسی دیدگاه های مُختلف میتوان پی برد که؛ زیرمجموعه های کاهشی قابل قبول، موازی سازی الگوریتم به منظور صرف زمان کمتر در رسیدن به پاسُخ، کاردینالیتی کمتر و کیفیت، دقت و صحت بالا در امر کاهش ویژگی همه و همه دلایلی بر برتری مزایا و راندمان مُناسب این روش پیشنهادی میباشند. بررسی الگوریتم چرخه آب بعنوان روشی مبتنی بر جمعیت، میتواند سبب بهبود تمام راه حل های موجود (درجمعیت) گردد. داده های شناخته شده از UCI که در ۹ مجموعه داده استفاده شده‌اند و الگوریتم WCA بر روی آن‌ها آزمایش شده است، نتایج نشان می‌دهد که الگوریتم پیشنهادی بهترین عملکرد را نسبت به راف فازی و الگوریتم WCA ارائه می‌دهد. این نتایج نشان می‌دهند که الگوریتم پیشنهادی در تحلیل داده ها و استخراج اطلاعات، عملکرد بهتری نسبت به روش های موجود دارد. با این حال، برای ارزیابی بهتر و قطعی تر عملکرد الگوریتم پیشنهادی، می توان بر روی مجموعه داده های دیگری نیز آزمایشات جدیدی انجام داد.

جدول ۳-نتایج بدست آمده توسط هر یک از روش ها

| الگوریتم | WCA | | GA | | HS | |
|---|---|---|---|---|---|---|
| دیتاست | تعداد ویژگی ها | زمان اجرا(ثانیه) | تعداد ویژگی ها | زمان اجرا(ثانیه) | تعداد ویژگی ها | زمان اجرا(ثانیه) |
| **Cleveland** | $9^{12}\ 10^{8}$ | 521 second | $10^{19}\ 11^{1}$ | 490 second | $9^{10}\ 10^{10}$ | 671 second |
| **Glass** | $9^{11}\ 10^{9}$ | 320 second | $9^{10}\ 10^{10}$ | 501 second | $10^{19}\ 11^{1}$ | 452 second |
| **Heart** | $8^{20}$ | 40 second | $8^{18}\ 9^{2}$ | 77second | $8^{17}\ 9^{3}$ | 129 second |
| **Ionosphere** | $8^{19}\ 9^{1}$ | 373 second | $9^{17}\ 10^{3}$ | 800 second | $8^{15}\ 9^{5}$ | 326 second |
| **Olitos** | $6^{18}\ 9^{2}$ | 238 second | $7^{20}$ | 130 second | $6^{16}\ 7^{4}$ | 197 second |
| **Water 2** | $7^{20}$ | 829 second | $7^{19}\ 8^{1}$ | 1003 second | $7^{17}\ 8^{3}$ | 3123 second |
| **Water3** | $6^{13}\ 7^{7}$ | 3618 second | $8^{20}$ | 2321 second | $7^{19}\ 8^{1}$ | 1928 second |
| **Web** | $18^{12}\ 19^{6}\ 20^{2}$ | 2238 second | $20^{20}$ | 2520 second | $20^{18}\ 2^{12}$ | 2639 second |
| **Wine** | $6^{20}$ | 156 second | $6^{19}\ 7^{1}$ | 201 second | $6^{20}$ | 195 second |

## مراجع


[1] Han, J., J. Pei, and M. Kamber, Data mining: concepts and techniques. 2011: Elsevier.

[2] Guyon, I. and A. Elisseeff, An introduction to variable and feature selection. "*Journal of machine learning research*," vol. 3, pp. 1157-1182, 2003.

[3] Yu, J., "General C-means clustering model, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp.1197-1211, 2005.

[4] Al Imran, Abdullah, Md Rifatul Islam Rifat, and Rafeed Mohammad. "Enhancing the classification performance of lower back pain symptoms using genetic algorithm-based feature selection." *Proceedings of International Joint Conference on Computational Intelligence*: IJCCI 2018. Singapore: Springer Singapore, 2019.

[5] Tawhid, Mohamed A., and Abdelmonem M. Ibrahim. "Hybrid binary particle swarm optimization and flower pollination algorithm based on rough set approach for feature selection problem." *Nature-inspired computation in data mining and machine learning*, pp. 249-273, 2020.

[6] Som, Tanmoy, et al. "Fuzzy Rough Set Theory-Based Feature Selection: A Review." *Mathematical Methods in Interdisciplinary Sciences*, pp. 145-166, 2020.

[7] Nasir, Mohammad, et al. "A comprehensive review on water cycle algorithm and its applications." Neural Computing and Applications, vol. 32, pp. 17433-17488, 2020.

[8] Palanisamy, S. and S. Kanmani, "Artificial Bee Colony Approach for Optimizing Feature Selectio," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 3, 2012.

[9] Ke, L., Z. Feng, and Z. Ren, "An efficient ant colony optimization approach to attribute reduction in rough set theory," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1351-1357, 2008.

[10] Abdolrazzagh-Nezhad, M., and Ali Adibiyan. "Attribute reduction based on rough set theory by soccer league competition algorithm," Iranian Journal of Electrical and Computer Engineering, vol. 82, no. 3, 2021.

[11] Pawlak, Z., "Rough sets," *International journal of computer & information sciences*, vol. 11, no. 5, pp. 341-356, 1982.

[12] Jain, Pankhuri, Anoop Tiwari, and Tanmoy Som. "Fuzzy rough assisted missing value imputation and feature selection." *Neural Computing and Applications*, vol. 35. no.3, pp. 2773-2793, 2023.

[13] Al Shalabi, L., Z. Shaaban, and B. Kasasbeh, "Data mining: A preprocessing engine," *Journal of Computer Science*, vol. 2, no. 9, pp. 735-739, 2006.

[14] Zadeh, L.A., "The concept of a linguistic variable and its application to approximate reasoning—I," *Information sciences*, vol. 8, no. 3, pp. 199-249, 1975.

[15] Ayesha, Shaeela, Muhammad Kashif Hanif, and Ramzan Talib. "Overview and comparative study of dimensionality reduction techniques for high dimensional data." *Information Fusion*, vol. 59, pp. 44-58, 2020.