پیشگویی مقادیر مفقود شده با استفاده از بهینه سازی گروه ذرات مشارکتی

محمد حسین نوروزی بیرامی
دانشکده مهندسی کامپیوتر
دانشگاه آزاد اسلامی واحد اسکو، ایران
mh.noroozi@gmail.com

محمد رضا میبدی
دانشکده مهندسی کامپیوتر و فناوری اطلاعات
دانشگاه صنعتی امیرکبیر، تهران، ایران
mmeybodi@aut.ac.ir

محمد حسین نژاد قویفکر
دانشکده مهندسی کامپیوتر
دانشگاه آزاد اسلامی واحد اسکو، ایران
m.hoseinnejad@yahoo.com

چکیده: در سال های اخیر، مسئله پیشگویی مقادیر صفات مفقود شده[1] در داده کاوی و کشف دانش از پایگاه داده ها مورد توجه محققان قرار گرفته است. ساده ترین روش برخورد با مقادیر مفقود شده، نادیده گرفتن آنها می باشد که در این صورت اطلاعات با ارزشی از دست خواهد رفت. روش های مختلفی برای برخورد با این مسئله پیشنهاد شده است. بیشتر این روش ها از قوانین تصمیم برای پیشگویی مقادیر مفقود شده استفاده می کنند. در این مقاله یک روش جدید با استفاده از بهینه سازی گروه ذرات مشارکتی برای پیشگویی مقادیر مفقود شده پیشنهاد می گردد. این روش بدون استخراج قوانین موجود بین داده، از رکورد های داده ای برای همگرا شدن به مقدار مفقود شده استفاده می کند. در این روش نیازی به دانش افراد خبره برای مشخص کردن ارتباط بین داده ها وجود ندارد. الگوریتم پیشنهادی بر روی داده های ۵۰ سال اخیر هواشناسی شهر تبریز مورد آزمایش قرار گرفته که توانسته با صحت ۹۸/۰۵ درصد داده های مفقود شده را پیشگویی کند.

واژه‌های کلیدی: مقادیر مفقود شده، نظریه مجموعه های نادقیق، قوانین انجمنی، قوانین تصمیم، بهینه سازی گروه ذرات مشارکتی.

۱- مقدمه

یک از مشکلات در تحلیل اطلاعات و استخراج دانش از میان حجم وسیعی از اطلاعات مشکل مقادیر مفقود شده می باشد[1]. دلایل مختلفی برای مفقود شده مقادیر صفات وجود دارد، مثل پاک شدن اطلاعات، پیش بینی نشدن دقیق فیلد ها در مجموعه داده های قدیمی، از بین رفتن داده ها در اثر انتقال از حسگر ها و ...[23]. برای حل مشکل مقادیر مفقود شده روش های مختلفی از جمله انتخاب مقدار صفت بسیار معمول، مفهوم مقدار صفت معمول، انتصاب همه مقادیر ممکن که یک مفهوم را ارائه کنند، چشم پوشی از مقدار صفت مفقود شده، با مقدار مفقود شده به صورت یک مقدار خاص برخورد شود، روش های پوشش وقایع و تعدادی روش های نحوی برای پردازش مستقیم مقادیر صفات مفقود شده گزارش شده است[23,6].

اما با همه اینها اختلال باز هم در داده ها وجود دارد. حالا روش انتصاب مقادیر بسیار معمول را برای مقادیر صفات مفقود شده امتحان می کنیم. این روش مقادیر بسیار تکرار شده را برای مقدار مفقود شده انتصاب می کند که در مثال زیر توضیح داده می شود.

مثال ۱: جدول ۱ را در نظر بگیرید که ۴ مثال داده در مجموعه داده T(C,D) وجود دارد که C مجموعه صفات شرایط و D مجموعه صفات تصمیم می باشد، U نیز مجموعه ای از داده های مثال می باشد. $C=(c_1,c_2,c_3,c_4)$ و $D=(0,1)$ و $U=(u_1,u_2,u_3,u_4)$ که c_3 در u_2 مقدار مفقود شده می باشد و با علامت "؟" نشان داده شده است. طبق این روش مقدار بسیار معمول برای این صفت انتصاب خواهد شد. اما با انتصاب مقدار معمول در این صورت مجموعه شرایط u_1 و u_2 یکی شده ولی صفات تصمیم آنها متفاوت خواهد بود.

جدول ۱: مجموعه داده ها با مقادیر صفات مفقود شده

	Condition				Decision
U	C1	C2	C3	C4	D
1	1	2	2	1	1
2	1	2	?	1	0
3	1	1	3	1	0
4	1	0	2	0	1

یک روش دیگری هم اینکه با مقدار مفقود شده به صورت یک مقدار خاص رفتار شود که این نیز در داده های اصلی اختلال ایجاد می کند. مثلا داده مفقود را به عنوان یک داده ناشناخته در نظر بگیریم.

مثال ۲: جنسیت بیمار را به عنوان صفت گم شده در نظر بگیریم که می تواند دو مقدار مرد و زن داشته باشد، که در این مثال نمی توان مقدار ناشناخته در نظر گرفت[23].

در این مقاله یک روش جدیدی برای برخورد با این مسئله ارائه خواهیم کرد. این روش از بهینه سازی گروه ذرات مشارکتی، برای پیشگویی مقادیر مفقود شده استفاده می کند. الگوریتم پیشنهادی یک روش جدیدی در این حوزه می باشد، که از رکورد های موجود در مجموعه داده ها، برای پیشگویی مقادیر مفقود شده استفاده می کند و نیازی به دانش انسان خبره ندارد.

در ادامه این مقاله، قسمت های مختلف به این صورت سازماندهی شده است. در بخش دوم، روش های موجود را برای پیشگویی مقادیر مفقود شده ارائه می کنیم. در بخش سوم، بهینه سازی گروه ذرات و انواع آن را معرفی می کنیم. در بخش چهارم، الگوریتم پیشنهادی را برای پیشگویی مقادیر مفقود شده ارائه می شود. در بخش پنجم، ارزیابی های انجام شده را توضیح می دهیم و در نهایت در بخش ششم نتیجه گیری و کار های آتی را ارائه خواهیم کرد.

۲- الگوریتم های پیشگویی مقادیر مفقود شده

۲-۱- نظریه مجموعه های نادقیق

نظریه مجموعه های نادقیق[۲] نخستین بار توسط پاولاک در سال ۱۹۸۲ ارائه گردید و از آن زمان به بعد شاهد گسترش کاربردی و نظری این موضوع در دنیا هستیم. این نظریه نگاه ریاضی جدیدی به مفاهیم "نادقیق" و "مبهم" دارد. مبنای این نظریه، دانش و اطلاعاتی است که هر یک از مشاهدات مجموعه مورد مطالعه که در اصطلاح مجموعه مرجع نامیده می شود، در اختیار دارد.

در این روش ها، هر مفهوم نادقیق توسط یک جفت از مفاهیم دقیق مشخص می شود، که به آنها تقریب پایین[۳] و تقریب بالای[۴] مفهوم نادقیق گفته می شود. تقریب پایین از تمام مشاهداتی تشکیل شده است که "قطعا" متعلق به آن مفهوم نادقیق می باشد و تقریب بالا از تمامی مشاهداتی تشکیل شده است که "احتمالا"متعلق به آن مفهوم می باشد. اختلاف بین تقریب بالا و پایین، ناحیه مرزی آن مفهوم را تشکیل می دهد. ایجاد تقریب های بالا و پایین در مجموعه های نادقیق، اصول تحلیل اطلاعات را در مراحل بعدی پایه گذاری می نماید[1].

دو دسته قانون برای پیشگویی استخراج می شود. قوانین معین[۵] از مفهوم تقریب پایین بدست می آید. در واقع تقریب پایین حالت قطعی و معین برای تصمیم گیری می باشد. قوانین ممکن[۶] نیز از مفهوم تقریب بالا بدست می آید و تقریب بالا حالت احتمالی در تصمیم گیری دارد[6,7,8].

۲-۲- تکرار مجموعه اقلام

در این روش از تکنیک کاوش قوانین انجمنی استفاده می کنیم که در داده کاوی برای کشف رابطه بین اقلام در مجموعه داده های با تراکنش های زیاد به کار می رود. در تولید قوانین انجمنی، تکرار مجموعه داده ها بر اساس ارتباط بین قلم با قلم داده ای دیگر با support[۷] معین مشخص می شود. با ارائه تکرار مجموعه اقلام می خواهیم بگوییم که احتمال اینکه یک یا چند قلم باهم در تراکنش جاری حضور داشته باشد چقدر است. به همین دلیل تکرار مجموعه اقلام می تواند برای پشگویی مقادیر صفات مفقود شده به کار گرفته می شود، که به روش مجموعه اقلام[۸] معروف است[23]. برای تولید مجموعه اقلام تکراری در الگوریتم قوانین انجمنی، ابتدا تکرار هر یک از عناصر اولیه را در تراکنش ها می شمارد. وقتی که مجموعه اقلام اولیه بدست آمده support بیشتری از یک حد آستانه مشخص داشته باشد به مجموعه اقلام تولید کننده قوانین انجمنی اضافه می شود. این فرایند تا زمانی ادامه پیدا می کند که دیگر مجموعه اقلام جدیدی یافت نشود. [9,10].

۲-۳- تولید قوانین تصمیم

روش های کلاسیک برای استنتاج قوانین تصمیم یک فرایند دو مرحله ای می باشد. در مرحله اول یک راه حل برای پوشش تمامی مثال های موجود پیدا می کنیم. مجموعه قوانین پوششی یا به صورت مستقیم از استنتاج قوانین عطفی بدست می آید و یا به صورت غیر مستقیم با استفاده از درخت تصمیم استخراج می شود[11]. در راه حل مستقیم معمولا در هر زمان یک قانون استنتاج می شود و آن موردی که توسط قانون پوشش داده شده، حذف می شود. این فرایند تکرار می شود. در مرحله دوم مجموعه قوانین پوششی به یک ساختار کوچکتر تبدیل شود و بهتر است با استفاده از تست های آماری مجموعه قوانین مستقل انتخاب شود. حال قوانینی که از این

روش بدست می آید چارچوبی برای مجموعه داده های خواهد بود که با استفاده از آنها می توان مقادیر صفات مفقود شده را حدس زد[12,13].

۳- بهینه سازی گروه ذرات

۳-۱- بهینه سازی گروه ذرات استاندارد

بهینه سازی گروه ذرات از جمله الگوریتم های جستجوی موازی مبتنی بر جمعیت است که با یک گروه از جواب های تصادفی (ذره ها) شروع به کار می کند، سپس برای یافتن جواب بهینه در فضای مسئله با به روز کردن مکان ذره ها به جستجو ادامه می دهد. هر ذره به صورت چند بعدی (بسته به نوع مسئله) با دو بردار X_{id} و V_{id} که به ترتیب معرف موقعیت مکانی و سرعت بعد d ام از i امین ذره هستند، مشخص می شود. در هر مرحله از حرکت جمعیت، مکان هر ذره با دو مقدار بهترین به روز می شود. اولین مقدار، بهترین تجربه ای است که خود ذره تا به حال بدست آورده است و با p_best نشان داده می شود. دومین مقدار، بهترین تجربه ای است که در بین تمامی ذره ها بدست آمده است و با g_best نشان داده می شود. در هر تکرار، الگوریتم بعد از یافتن دو مقدار بالا، سرعت و موقعیت جدید ذره را بر اساس معادلات (۱) و (۲) بروز رسانی می کند[14].

$$\begin{aligned} v_{id}(t+1) &= w.v_{id}(t) \\ &+c_1.rand_1(p\_best_{id}(t) - x_{id}(t)) \\ &+c_2.rand_2(g\_best_{id}(t) - x_{id}(t)) \end{aligned} \qquad (۱)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \qquad (۲)$$

در معادله (۱) ، w ضریب اینرسی است که به صورت خطی کاهش می یابد که معمولا در بازه [۰-۱] می باشد. c_1 و c_2 ضرایب شتاب می باشد که در بازه [۰-۲] انتخاب می شود که در بیشتر موارد برای هر دوی آنها از مقدار ۱,۴۹ استفاده شده است[2,6,7]. دو عدد$rand_1$ و $rand_2$ نیز اعداد تصادفی در بازه [۰-۱] می باشد. همچنین مقدار نهایی سرعت هر ذره برای جلوگیری از واگرایی الگوریتم به یک بازه محدود می شود. $v_{id} \in [-v_{\max}, v_{\max}]$. شرط خاتمه الگوریتم همگرایی تا حد معین و یا توقف بعد از تعداد معینی تکرار می باشد[3]. معادله (۲) نیز بردار موقعیت جدید ذره را با توجه به سرعت جدید و موقعیت فعلی آن بروز رسانی می کند.

۳-۲- بهینه سازی گروه ذرات مشارکتی

بهینه سازی گروه ذرات مشارکتی تعمیمی بر PSO استاندارد می باشد، که چندین دسته ذره برای پیدا کردن جواب بهینه به صورت مشارکتی با هم کار می کنند، روش های مختلفی برای این کار پیشنهاد شده است، در [15,16] از چند دسته ذره برای جستجو استفاده می شود که یکی از دسته ها دسته اصلی و بقیه دسته ها، دسته های پیرو می باشد که برای بدست آوردن بهترین جواب، دسته اصلی را کمک می کنند. در [17,18] به جای اینکه یک دسته ذره بدنبال حل یک مسئله n بعدی باشند، n دسته ذره برای حل n مسئله یک بعدی تلاش می کنند.

۳-۳- بهینه سازی گروه ذرات مشارکتی تعمیم یافته

در [5] تعمیم جدیدی برای الگوریتم PSO ارائه نموده ایم که اولا مشکلات افتادن در دام بهینگی محلی را به مراتب کاهش می دهد و ثانیا سرعت همگرایی را افزایش می دهد، همچنین مقدار دهی تصادفی اولیه ذرات را از بین می برد و از طرف دیگر به هدف اصلی روش های PSO مشارکتی که امکان موازی سازی است، نائل می شود. شکل ۱ شبه کدی برای الگوریتم CPSO [۹] می باشد.

الگوریتم ارائه شده شامل دو قسمت اصلی می باشد. قسمت اول شامل n دسته ذرات مستقل می باشد. در مرحله اول فضای جستجو را به n قسمت مستقل تقسیم می کنیم و هر دسته ذرات مسئول جستجو در یک قسمت می باشد. بعد از این تقسیم بندی، هر دسته، الگوریتم PSO را بر روی ناحیه مورد جستجو اعمال می کند و بعد از اینکه ۲۰٪ از کل جستجو ها را انجام دادند، کار دسته های اولیه تمام می شود و بهترین جوابی که برای هر دسته بدست آمده است به عنوان جواب های اولیه دسته اصلی محسوب می شود. بعد از اتمام این مرحله، دسته اصلی را با n ذره مقدار دهی می کنیم که مقدار اولیه آنها، بهترین جواب های بدست آمده از تک تک دسته های مرحله اول می باشد. به بیان دیگر بهترین تجربه گروهی تک تک دسته ها، بهینه های محلی هر ناحیه مستقل می باشد و دسته اصلی با استفاده از بهینه های محلی فضای

CPSO Algorithm

```
define
gbest: array [1..n] k-dimension vector // for all swarms
Split search space to n independent area
      // n is numbers of swarms and each swarm to one area P_j   j ∈ [1..n] initialize n k-dimensions PSOs:
for all swarms
   for i=1:t                  // t is 1/5of all iteration in simulation
      apply PSO            // each swarm are limited in own area
      update gbest array for each swarm
   end
end
initialize k-dimension PSO with gbest values  // swarm have n particle
select the best area with minimum fitness for next search
for i=t+1:end of simulation
   apply PSO for search area or selected area
end
```

شکل ۱ : شبه کد الگوریتم CPSO

جستجو مقدار دهی می شود. بعد از این کار ۸۰٪ تکرار های باقیمانده را با الگوریتم PSO بر روی کل فضای جستجو و یا فقط بر روی ناحیه ای که بهترین جواب متعلق به آن است، اعمال می کند.

۴- الگوریتم پیشنهادی برای پیشگویی مقادیر مفقود شده با استفاده از CPSO

با توجه به مطالبی که در بخش های قبل اشاره شد، می توان گفت که اکثر روش های موجود برای پیشگویی مقادیر مفقود شده، مبتنی بر قوانین می باشد که باید قوانینی برای مجموعه داده های موجود پیدا کرد و بر اساس این قوانین مقادیر مفقود شده را پیشگویی کرد. مسئله اصلی این است که برای بدست آوردن قوانینی با درجه اطمینان بالا، نیاز به داده هایی با صحت و دقت و حجم بالا می باشد.

در این بخش می خواهیم روش متفاوتی برای پیشگویی مقادیر مفقود شده ارائه نماییم که مبتنی بر الگوریتم PSO می باشد. با توجه به این که در بخش قبل الگوریتم CPSO را ارائه کردیم، این الگوریتم را برای پیشگویی مقادیر مفقود شده به کار خواهیم گرفت. شکل ۲ الگوریتم پیشگویی مقادیر مفقود شده با استفاده از CPSO را نمایش می دهد.

این الگوریتم با استفاده از PSO برای اولین بار در این حوزه به کار گرفته می شود و مهمترین مرحله کار انتخاب تابع شایستگی است. جدول ۲ نحوه انتخاب این تابع را بر روی داده های هواشناسی نشان می دهد. در واقع اگر فیلد مفقود شده متوسط میزان دما باشد و بازه تغییرات آن [20.5,30.1-] باشد. سایر فیلدهای رکورد با مقدار مفقود شده را عینا کپی می کنیم و مقادیر ممکنی که دمای متوسط می تواند در بازه مورد نظر به خود بگیرد را به طور مساوی تقسیم کرده و به جای مقدار مفقود شده قرار می دهیم و هدف ما همگرا شدن به یکی از این رکورد های تابع شایستگی است. هر چه قدر تعداد این رکورد ها بیشتر باشد، دقت کار افزایش پیدا کرده و از طرف دیگر سرعت کار کاهش می یابد. بنابراین باید حد متوسطی برای تعداد رکورد ها پیدا نماییم. اما با توجه به این روند ممکن است در طی تکرار های مختلف فاصله از تابع شایستگی کم شود ولی فیلد مفقود شده از داده واقعی دور گردد. برای رفع این مشکل بعد از چند تکرار، تعدادی رکورد به صورت تصادفی از خوشه مورد نظر انتخاب کرده و نزدیکترین آنها را با جواب هر ذره جایگزین می کنیم. تعداد بعد های هر ذره بستگی به تعداد صفات رکوردها در مجموعه داده ها دارد.

۵- ارزیابی روش پیشنهادی

با توجه به مطالب ارائه شده در این مقاله می توان گفت که نقطه اشتراک تمامی روش های موجود، مبتنی بر قانون بودن آنها است و باید قوانینی برای هر مجموعه داده استخراج شود و همچنین مقدار مفقود شده باید نوع مشابهی در مجموعه داده داشته باشد تا بتواند به مقدار مفقود شده انتساب شود. ولی در الگوریتم ارائه شده هیچ یک از این دو نقطه مورد استفاده قرار نگرفته است. در واقع سعی می شود یکی سری رکورد تصادفی را به رکوردی با مقدار مفقود شده همگرا نماییم. در صورتی که ممکن است مقدار پیشگویی شده اصلا در مجموعه داده ها وجود نداشته باشد و جواب بدست آمده به مقدار مفقود شده همگرا می شود[4].

۱. ابتدا رکورد های مفقود شده را از مجموعه داده ها جدا کرده و مجموعه داده ای بدون مقدار مفقود شده به وجود می آوریم.
۲. الگوریتم خوشه بندی k-means را بر روی مجموعه داده های بدون مقدار مفقود شده اجرا می کنیم و مجموعه داده ها را به c خوشه، خوشه بندی می کنیم[19,20,21,22].

۳. یکی از رکورد ها با مقدار مفقود شده را برای پیشگویی انتخاب می کنیم.
۴. مینیمم و ماکزیمم مقدار موجود در مجموعه داده ها را برای فیلد مفقود شده پیدا می کنیم.
۵. آرایه ای با h عضو بین مینیمم و ماکزیمم مقدار مفقود شده ایجاد کرده و a می نامیم.
۶. تابع شایستگی را با h عضو ایجاد می کنیم (این تابع دقیقا رکورد مفقود شده می باشد و به جای فیلد مفقود شده عناصر آرایه a را کپی می کنیم)
۷. c دسته ذره ایجاد می کنیم و هر یک از دسته ها را به صورت تصادفی از خوشه ها مقدار دهی اولیه می کنیم (هر دسته دقیقا از یک خوشه مقدار دهی می شود)
۸. الگوریتم PSO را به تعداد تکرار معین (۲۰٪ کل تکرار ها) و به طور جداگانه بر روی خوشه ها اجرا می کنیم. (برای بدست آوردن p_best و g_mest، فاصله تک تک ذرات را از رکورد های تابع شایستگی محاسبه کرده و کمترین فاصله را به عنوان بهترین جواب استفاده می کنیم و برای اینکه از جواب های واقعی فاصله نگیریم بعد از تعداد معینی تکرار، تعدادی رکورد از خوشه مربوطه انتخاب می کنیم و نزدیکترین آنها به جواب ذره را جایگزین مقدار ذره می کنیم) .
۹. یک دسته ذرات اصلی ایجاد می کنیم و ذرات آن را با بهترین جواب هایی که از دسته های مختلف در مرحله قبل بدست آمده است، مقدار دهی می کنیم و خوشه ای که بهترین جواب متعلق به آن است برای جستجوی اصلی انتخاب می کنیم.
۱۰. الگوریتم PSO را به تعداد تکرار معین (۸۰٪ کل تکرارها) بر روی خوشه انتخاب اجرا می کنیم(تابع شایستگی دقیقا مثل مرحله ۸ بر روی جواب های ذرات اعمال می شود و بهترین جواب را انتخاب می کنیم).
۱۱. مقداری که برای فیلد مفقود شده در بهترین جواب نهایی بدست آمده، مقداری است که برای فیلد مفقود شده بدست آمده است

شکل ۲ : الگوریتم CPSO برای پیشگویی مقدار مفقود شده

جدول ۲ : چگونگی انتخاب تابع شایستگی برای داده های هواشناسی

ردیف	متوسط دما (سلسیوس)	کمترین دما (سلسیوس)	بیشترین دما (سلسیوس)	میزان بارش (میلی متر)	متوسط فشار (سانتیمتر جیوه)	...
1	-20.5	15	26	0.2	50	...
2	-17.6789	15	26	0.2	50	...
3	-14.8579	15	26	0.2	50	...
⋮	⋮					
19	30.27895	15	26	0.2	50	...
20	33.1	15	26	0.2	50	...

با توجه به اینکه حوزه پیاده سازی روش ها عنوان شده بسیار متنوع و خارج از حد این مقاله می باشد و با توجه به اینکه میزان موفقیت این الگوریتم ها بستگی به صحت پیشگویی مقادیر مفقود شده می باشد و ارتباطی به روش مورد استفاده ندارد، روشی بهینه تر است که بتواند داده ها را با درصد صحت بالا تری پیشگویی نماید. مثلا در [6] برای قوانین انجمنی با support ۱۰ در صد، صحت پیشگویی ۹۶ در صد بدست آمده است، در صورتی که support ۱۰ درصد برای قوانین، جامعیت کمتری را برای داده های موجود عنوان می کند. در [12] روش های مختلف مبتنی بر قوانین تصمیم و انجمنی مورد مقایسه قرار گرفته است که برای مجموعه داده های مربوط به خودرو های مختلف انجام گرفته است و در نهایت به صحت پیشگویی ۷۳ درصد نائل آمده است. در [13] الگوریتم مبتنی بر قوانین تصمیم بر روی داده های استاندارد UCI مورد آزمایش قرار گرفته است و با صحت حدود ۶۵ در صد داده ها مورد پیشگویی قرار گرفته است.

۵-۱- نتایج حاصل از پیشگویی روش پیشنهادی

برای آزمایش روش پیشنهادی از داده های ۵۰ سال اخیر سازمان هواشناسی آذربایجان شرقی، ایستگاه فرودگاه تبریز استفاده کرده ایم و در کل ۱۷۴۳۴ رکورد معتبر بدست آمده است، فیلد های مربوط به این داده ها و بازه تغییرات هر یک از فیلد ها، در جدول ۳ نشان داده شده است.

با توجه به اینکه تعداد فیلد های این جدول ۱۰ عدد می باشد، بنابراین بعد ذرات را نیز ۱۰در نظر می گیریم و مجموعه داده ها را به ۶ خوشه مجزا تقسیم می کنیم و تعداد رکورد های تابع شایستگی ۴۰ و تعداد رکورد های تصادفی برای جایگزینی با مقدار ذره را نیز ۴۰ در نظر می گیریم و برای دسته های اولیه ۲۰ ذره و برای دسته نهایی ۶ ذره در نظر می گیریم(برای هر خوشه یک ذره). الگوریتم را ۴۰۰ بار تکرار می کنیم که ۸۰ تکرار برای دسته های اولیه و ۳۲۰ تکرار برای جستجو در خوشه ای که بهترین جواب مرحله اول متعلق به آن است، در نظر می گیریم.

جدول ۳ : فیلد های مربوط به داده های هواشناسی شهر تبریز

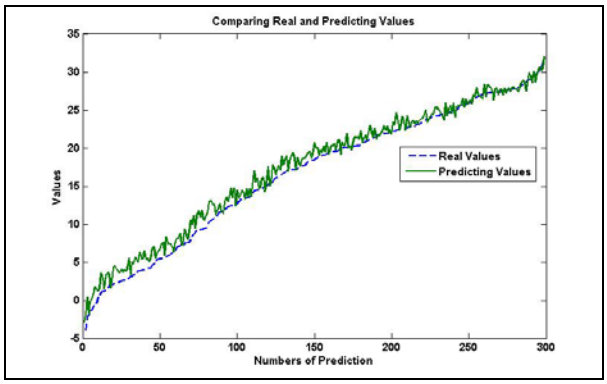
ردیف	نام فیلد	بازه تغییرات
۱	دمای متوسط (سلسیوس)	[-20.5,33.1]
۲	کمترین دما	[-25,33.5]
۳	بیشترین دما	[-14,42]
۴	میزان بارش (میلی متر)	[0,63]
۵	فشار متوسط (میلی متر جیوه)	[843,879]
۶	کمترین فشار	[840,878]
۷	بیشترین فشار	[846,888]
۸	رطوبت متوسط (درصد)	[9.8,98.2]
۹	کمترین رطوبت	[1,96]
۱۰	بیشترین رطوبت	[16,100]

اساس کار به این صورت است که ابتدا به تصادف رکوردی را انتخاب می کنیم و فیلد مفقود شده آن را نیز یکی از سه دمای موجود در جدول در نظر می گیریم. بعد الگوریتم را برای پیشگویی مقدار مفقود شده اجرا می کنیم و نتیجه بدست آمده را با مقدار واقعی مقایسه می کنیم. لازم به توضیح است که رکوردی که برای پیشگویی انتخاب شده، از جدول به طور موقت حذف می کنیم و آن را در محاسبات در نظر نمی گیریم.

جدول ۴ نتایج حاصل از اجرای مستقل بیش از ۱۰۰۰ بار پیشگویی می باشد که متوسط اختلاف مقدار پیشگویی شده را با داده واقعی نشان می دهد. شکل ۳ نیز مقایسه بین مقادیر واقعی و مقادیر پیشگویی شده را برای ۳۰۰ پیشگویی نشان می دهد.

جدول ۴: نتایج حاصل از پیشگویی روش CPSO

بازه میزان اختلاف	درصد پیشگویی ها	متوسط اختلاف	درصد صحت
[0 0.5]	35.89%	± 0.25	99.52%
[0.5 1]	17.62%	± 0.77	98.55%
[1 2]	31.73%	± 1.40	97.38%
بیش از 2	14.76%	± 2.51	95.31%
تمامی داده ها	100%	± 1.04	98.05%



شکل ۳: مقایسه مقادیر واقعی با مقادیر پیشگویی شده برای الگوریتم CPSO

با توجه به نتایج جدول ۴ می توان گفت که روش پیشنهادی برای پیشگویی مقادیر مفقود شده، موفق عمل کرده است. در واقع در این آزمایش، داده های مربوط به کمترین، بیشترین و متوسط دمای موجود در مجموعه داده ها مورد پیشگویی قرار گرفته اند. این داده های در بازه [42.3 -25] درجه سانتیگراد قرار داشتند. در این ارزیابی که بر روی ۱۰۰۰ داده مفقود شده انجام گرفته است، توانسته ایم در ۳۵/۸۹ ٪ پیشگویی ها اختلاف کمتر از ۰/۵ واحد با داده های واقعی داشته باشیم. برای ۱۷/۶۲ ٪ پیشگویی ها اختلافی بین ۰/۵ و ۱ واحد داشته ایم. برای ۳۱/۷۳ ٪ پیشگویی ها اختلافی بین ۱ تا ۲ واحد از داده های واقعی بدست آمده است و در نهایت برای ۱۴/۷۶ ٪ از پیشگویی های اختلافی بیش از ۲ واحد بدست آمده است. در کل ما توانسته ایم داده های مفقود شده را با صحت ۹۸،۰۵٪ پیشگویی نماییم و کل داده های پیشگویی شده با داده های واقعی به طور متوسط ۱،۰۴ $\pm$ واحد اختلاف دارند.

از مقایسه این نتایج با نتایج روش های موجود که در ابتدای این بخش ارائه شده است، به این نتیجه می رسیم که روش پیشنهادی به صورت فوق العاده عمل کرده است و داده های مفقود شده را با صحت بسیار بالایی پیشگویی کرده است. همچنین در این روش دقیقا مقدار مفقود شده جستجو می شود، نه بازه ای که مقدار متعلق به آن می باشد.

۶- نتیجه گیری

در پایان می توان گفت، روش های موجود غالبا مبتنی بر قوانین هستند و باید قوانینی از کل داده های موجود استخراج شود و کیفیت این قوانین و فراگیری آنها تاثیر مستقیمی بر روی پیشگویی مقادیر مفقود شده دارد. اما الگوریتم پیشنهادی سعی می کند مجموعه ای از داده های تصادفی را به رکوردی با مقدار مفقود شده همگرا نماید. در ضمن باید به این نکته تاکید کرد که روش پیشنهادی برای پیشگویی مقادیر مفقود شده نیازی به انسان خبره مسلئه ندارد.

تنها اشکالی که روش پیشنهادی در مقابل روش های موجود دارد، زمان اجرای الگوریتم می باشد. چون این روش مبتنی بر جمعیت می باشد و باید چندین نسل تکرار شود تا به نتیجه برسد، بنابراین نسبت به روش های مبتنی بر قوانین زمانبر خواهد بود.

بنابراین می توان گفت که روش پیشنهادی یک حوزه تحقیقاتی جدیدی در این زمینه می باشد که جای کار بسیار زیادی وجود دارد. نکته ای که در این روش بسیار تعیین کننده است، چگونگی انتخاب تابع شایستگی برای سنجش شایستگی مقادیر پیدا شده می باشد.

با توجه به نتایج شبیه سازی می توان گفت: درست است که این روش در ابتدای راه قرار دارد و جای تکامل بسیار زیادی دارد، ولی در عین حال نتایج بهتری نسبت به روش های موجود کسب کرده است و صحت پیشگویی ۹۸/۰۵ درصد بر روی داده های واقعی بسیار امیدوار کننده است.

مراجع

[۱] علی حیدرنژاد، "پیش بینی شاخص سهام با استفاده از نظریه مجموعه های نادقیق"، پایان نامه کارشناسی ارشد، دانشگاه شهید بهشتی، تهران، ۱۳۸۱.

[۲] حسین نظام آبادی پور و مجید رستمی شهر بابکی، "تعمیمی بر الگوریتم GCBPSO"، دوازدهمین کنفرانس مهندسی کامپیوتر ایران، تهران، ۱۳۸۵، صفحات ۲۹-۳۵.

[۳] محمد شیبانی و محمد رضا میبدی، "PSO-LA : یک مدل جدید برای بهینه سازی"، دوازدهمین کنفرانس مهندسی کامپیوتر ایران، تهران، ۱۳۸۵، صفحات ۱۱۶۲- ۱۱۶۹.

[۴] محمد حسین نورزوی، فریبرز محمودی، "پیشگویی مقادیر مفقود شده"، دهمین کنفرانس دانشجویی مهندسی برق، دانشگاه صنعتی اصفهان، ۱۳۸۶.

[۵] محمد حسین نوروزی، محمد رضا میبدی، "بهینه سازی گروه ذرات فازی مشارکتی"، دومین کنفرانس مشترک سیستم های فازی و هوشمند، دانشگاه مالک اشتر، تهران، ۱۳۸۷.

[6] W. Jerzy and G. Busse, “Rough set strategies to data with missing attribute values”, USA, IEEE, 2003.

[7] W. Jerzy and G. Busse and A. Y. Wang, “Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values”, the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences, 1997.

[8] M. Kryszkiewicz, “Rough set approach to incomplete information systems”, Proceedings of the Second Annual Joint Conference on Information Sciences, 1995.

[9] W. Jerzy and G. Busse and M. Hu. “A comparison of several approaches to missing attribute values in data mining”, In Rough Sets and Current Trends in Computing, 2000.

[10] M. Kryszkiewicz, “Association rules in incomplete databases”, Springer, 1999.

[11] M. Sholom and L. Nitin, “Decision rule solution for data mining with missing values”, USA, 2000.

[12] L. Jiye and C. Nick, “Comparisons on different approaches to assign missing attribute values”, 2006.

[13] C. Blake and E. Keogh and C. Merz, “Uci repository of machine learning database”, University of California Irvine, 1999.

[14] J.Kennedy and R. Eberhart, “Particle swarm optimization”, Proceedings of IEEE International Conference on Neural Networks, vol. 4, 1995.

[15] B. Niu and Y. Zhu, X. He, H. Wu, "MCPSO: A multi-swarm cooperative particle swarm optimizer", Applied Mathematics and Computation 185, Elsevier, 2007, pp. 1050–1062.

[16] B. Niu and Y. Zhu and X. Xian and H. Shen, "A multi-swarm optimizer based fuzzy modeling approach for dynamic systems processing ", Elsevier, 2007.

[17] F. V. D. Bergh and A. Engelbrecht, "A cooperative approach to particle swarm optimization", IEEE Transactions On Evolutionary Computation, Vol. 8, No. 3, 2004.

[18] F. V. D. Bergh and A. Engelbrecht, “Cooperative learning in neural networks using particle swarm optimizers”, South African Comput., vol. 26, 2000, pp. 84–90.

[19] B. Jiang, “Spatial clustering for mining knowledge in support of generalization processes in GIS”, ICA Workshop on Generalisation and Multiple representation, 2004, Leicester.

[20] J. Vesanto and Esa Alhoniemi, “Clustering of the self-organizing map”, IEEE Transactions on Neural Networks, Vol. 11, No. 3, pp. 586-600, 2000.
[21] X.Cui and E. Thomas , “Document clustering using particle swarm optimization”, Oak Ridge, IEEE, 2005.
[22] S. Selim and M. Ismail, “K-means type algorithms: A generalized convergence theorem and characterization of local optimality”, IEEE, 1984.
[23] C. Nick and L. Jiye, “Predicting missing attribute values based on frequent item set and RSFit”, ACM, 2006.

---

[1] - Predicting Missing Attribute Values
[2] - Rough Set Theory
[3] - lower approximation
[4] - upper approximation
[5] - certain rule
[6] - possible rule
[7] - پارامتر support یعنی تعداد تراکنش های شرکت کننده در تولید یک قانون تقسیم بر کل تراکنش های موجود در مجموعه داده ها
[8] - Item Sets approach
[9] - Cooperative Particle Swarm Optimization - CPSO