انتخاب ویژگی ها از صفحات وب مبتنی بر تئوری مجموعه ناهموار و اتوماتای یادگیر توزیع شده

بابک نصیری[۱]؛ فریبرز محمودی[۲]؛ محمدرضا میبدی[۳]

چکیده

صفحات وب حجم انبوهی از اطلاعات را در خود جای داده اند، اما برای وب کاوی بسیاری از آنها زائد و اضافی می باشند. از این رو لازم است بعنوان پیش پردازش، ویژگی های مناسب از بین این حجم اطلاعات استخراج شود. از آنجا کـه انتخـاب ویژگـی هـای مناسب یک مسئله NP-hard بشمار می رود، جستجو برای الگوریتم های تقریبی سریع و کارا همچنان ادامه دارد. در این مقاله یک روش ترکیبی جدید مبتنی بر تئوری مجموعه ناهموار[۱] و اتوماتای یادگیر توزیع شده برای انتخاب ویژگی های مناسب در صفحات وب ارائه شده است. نتایج حاصل از پیاده سازی روش پیشنهادی بر روی چندین مجموعه داده از جمله یـک مجموعـه مبتنـی بـر وب، حکایت از کارایی روش پیشنهادی در مقایسه با سایر روشهای شناخته شده دارد.

کلمات کلیدی

انتخاب ویژگی ها، وب کاوی، اتوماتای یادگیر توزیع شده، تئوری مجموعه ناهموار، محتوی کاوی وب.

Feature Selection from web pages based Rough set Theory and Distributed Learning Automata

Babak Nasiri[۱]; Fariborz Mahmoudi[۲]; Mohammad reza Meybodi[۳]

[۱،۲] Computer Engineering and Information Technology Department, Azad Islamic University, Qazvin, Iran

[۳] Computer Engineering and Information Technology Department, Amirkabir University of Technology ,Tehran, Iran

Abstract

Web pages have massive amounts of information in itself, but for web mining many of them are unneeded and extra. Therefore, as a pre-processing is required to extract appropriate features from huge amount of information. Since feature selection is a NP-Hard Problem, searching for fast and efficient algorithm continues. In this paper, a new hybrid approach based on rough set theory and distributed learning automata for feature selection from web pages are presented. The results of the implementation of the proposed Approach on several data set, including a Web dataset, shows that the proposed approach is comparable with other known methods.

۱. مقدمه

صفحات وب حجم انبوهی از اطلاعات را در خود جای داده اند، اما برای وب کاوی بسیاری از آنها زائد و اضافی می باشند. چنانچه ایـن حجـم زائد از اطلاعات مورد پاکسازی قرار نگرفته و حذف نشوند، الگوریتمهای وب کاوی را دچار مشکل کرده و در سرعت و دقت آنها بسـیار تـاثیر گـذار خواهند بود. انتخاب ویژگی های مناسب در بین اطلاعات یک روش خوب برای حذف اطلاعات زائد مـی باشـد. از آنجـا کـه انتخـاب ویژگـی هـای مناسب، یک مسئله NP-hard بشمار می رود، جستجو برای الگوریتم های تقریبی سریع و کارا همچنـان ادامـه دارد. در ایـن مقالـه مجموعـه داده هایی از وب با یک خصیصه تصمیم، بعنوان ورودی در نظر گرفته شده اند و هدف، انتخاب ویژگی های مناسب از بین کلیه ویژگی های موجود مـی باشد. این کار باید بنحوی صورت گیرد که قدرت طبقه بندی را کاهش ندهد.

در دهه گذشته روشهای مختلفی برای انتخاب ویژگی ها پیشنهاد شده است که به طور کلی می توان آنها را به دو گروه تقسیم نمود . الگوریتم های رتبه بندی ویژگیها۲ و الگوریتم های یافتن زیرمجموعه مینیمم۳ . در روش رتبه بندی ویژگیها ، به هر ویژگی یک رتبه نسـبت مـی دهنـد و

۱ دانشکده برق، کامپیوتر و فناوری اطلاعات،دانشگاه آزاد اسلامی، قزوین، ایران، nasiri_babak@yahoo.com

۲ دانشکده برق، کامپیوتر و فناوری اطلاعات،دانشگاه آزاد اسلامی، قزوین، ایران، mahmoudi@qazviniau.ac.ir

۳ دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه امیر کبیر، تهران، ایران، mmeybodi@aut.ac.ir

سپس آنها را بر اساس رتبه شان مرتب می کنند .این روش، زیرمجموعه مینیمم از ویژگیها (مجموعه کاهش یافته) را برای آنالیز های بعدی مشخص نمی کند و فقط فاصله موجود بین ویژگیها را بدست می آورد. اما در روش یافتن زیر مجموعه مینیمم، این موضوع بصورت معکوس وجود دارد [۸]. روش ارائه شده در این مقاله جزو گروه دوم بشمار می آید .

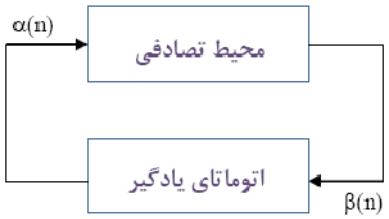
ادامه مقاله بصورت زیر سازماندهی شده است. در بخش دوم مفاهیم مورد نیاز از تئوری مجموعه ناهموار و اتوماتای یادگیر ارائه شده است. در بخش سوم، روش پیشنهادی برای انتخاب ویژگی ها، مبتنی بر تئوری مجموعه های ناهموار و اتوماتای یادگیر ارائه شده است و روش پیاده سازی آن تشریح شده است . در بخش چهارم نتایج حاصل از ارزیابی این روش در مقایسه با روشهای دیگر بر روی مجموعه داده های مختلف ارائه شده است و بخش آخر به نتیجه گیری و پیشنهادهایی برای بهبود تخصیص دارد.

۲. مفاهیم مورد نیاز

در این قسمت به شرح مفاهیم مورد نیاز برای ارائه روش پیشنهادی می پردازیم.

۱.۲. اتوماتای یادگیر

یک اتوماتای یادگیر یک مدل انتزاعی است که بطور تصادفی یک عمل از مجموعه متناهی اعمال خود را انتخاب کرده و بر محیط اعمال می کند. محیط، عمل انتخاب شده توسط اتوماتای یادگیر را ارزیابی کرده و نتیجه ارزیابی خود را توسط یک سیگنال تقویتی به اتوماتای یادگیر اطلاع می دهد. سپس اتوماتای یادگیر با اطلاع از عمل انتخاب شده و سیگنال تقویتی ، وضعیت داخلی خود را بروز کرده و عمل بعدی خود را انتخاب می کند. هدف نهایی این است که اتوماتا یاد بگیرد تا از بین اعمال خود بهترین عمل را انتخاب کند. بهترین عمل، عملی است که احتمال دریافت پاداش از محیط را به حداکثر برساند. کارکرد اتوماتای یادگیر در تعامل با محیط، درشکل (۱) مشاهده میشود[۳].



شکل (۱): ارتباط بین اتوماتای یادگیر با محیط

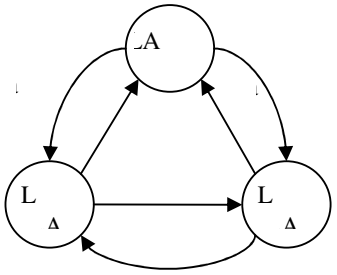
محیط را می توان توسط سه تایی $E=\{\alpha,\beta,C\}$ نشان داد که در آن $\alpha=\{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ مجموعه ورودیها، $\beta=\{\beta_1, \beta_2, \ldots, \beta_m\}$ مجموعه خروجیها و $C=\{C_1, C_2, \ldots, C_r\}$ مجموعه احتمالات جریمه می باشد. هرگاه β مجموعه دو عضوی باشد محیط از نوع P می باشد. در چنین محیطی $\beta_1 = 1$ بعنوان پاسخ نامطلوب یا شکست، $\beta_2 = 0$ بعنوان پاسخ مطلوب یا موفقیت در نظر گرفته می شوند. در محیط از نوع Q، مجموعه β دارای تعداد متناهی عضو می باشد و در محیط از نوع S، تعداد اعضاء مجموعه β نامتناهی است. C_i نشان دهنده احتمال نامطلوب بودن سیگنال تقویتی محیط در پاسخ به عمل α_i می باشد. در یک محیط ایستا مقادیر C_i ها ثابت هستند، حال آنکه در یک محیط غیر ایستا این مقادیر در طی زمان تغییر می کنند. بر اساس اینکه تابع بروزرسانی وضعیت اتوماتای یادگیر (که با اطلاع از عمل انتخاب شده و سیگنال تقویت β، وضعیت بعدی اتوماتای یادگیر را محاسبه کند) ثابت یا متغیر باشد، اتوماتای یادگیر به دو دسته اتوماتای یادگیر با ساختار ثابت و اتوماتای یادگیر با ساختار متغیر تقسیم می گردند.

اتوماتای یادگیر با ساختار متغیر را میتوان توسط یک چهارتایی بصورت $\{\alpha,\beta,P,T\}$ نشان داد که در آن $\alpha=\{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ مجموعه اعمال اتوماتای یادگیر، $\beta=\{\beta_1, \beta_2, \ldots, \beta_m\}$ مجموعه ورودیهای اتوماتای یادگیر، $P = \{P_1, P_2, \ldots, P_r\}$ بردار احتمال انتخاب هریک از عملها و T، $P(k+1) = T[\alpha(k), \beta(k), P(k)]$ الگوریتم یادگیری اتوماتای یادگیر میباشد. الگوریتم های یادگیری متنوعی برای اتوماتای یادگیر ارائه شده است که در ادامه یک الگوریتم یادگیری خطی برای اتوماتای یادگیر بیان می گردد. فرض کنید اتوماتای یادگیر در مرحله nام اقدام α_i خود را انتخاب نموده و محیط ارزیابی خود را توسط سیگنال تقویتی $\beta(n)$ به اتوماتای یادگیر اعلام می کند. با استفاده از الگوریتم یادگیری خطی، اتوماتای یادگیر، بردار احتمال انتخاب اقدام های خود را مطابق رابطه (۱) تنظیم می کند[۵].

$$P_i(n+1) = P_i(n) + a.(1 - \beta(n)).(1 - P_i(n)) - b.\beta(n).P_i(n)$$

$$P_j(n+1) = P_j(n) + a.(1-\beta(n)).P_j(n) + \frac{b.\beta(n)}{r-1} - b.\beta(n).P_i(n) \qquad \text{if } j \neq i \quad (1)$$

که a پارامتر پاداش و b پارامتر جریمه می باشد. اگر a و b با هم برابر باشند، الگوریتم L_{R-P}[۴]، اگر b از a خیلی کوچکتر باشد، الگوریتم $L_{R\in P}$[۵] و اگر b صفر باشد، الگوریتم L_{R-I} [۶] نام دارد[۳,۵].



شکل (۲) :اتوماتای یادگیر توزیع شده

۱.۱.۲. اتوماتای یادگیر توزیع شده(DLA)[۷] :

این اتوماتا شبکه ای از اتوماتاهای یادگیر است که برای حل یک مسئله با یکدیگر همکاری میکنند. تعداد اقدامهای یک اتوماتای یادگیر در DLA برابر با تعداد اتوماتاهای متصل به اتوماتای مورد نظر میباشد. انتخاب یک اقدام توسط اتوماتا در شبکه، اتوماتای متناظر با این اقدام را فعال میسازد . به عنوان مثال در شکل (۲) هر اتوماتا دارای دو اقدام میباشد. انتخاب اقدام a_1 توسط LA_1، اتوماتای یادگیر LA_2 را فعال می کند. LA_2 به نوبه خود یکی از اقدامهای خود را انتخاب می کند و سبب فعال شدن اتوماتای یادگیر دیگری خواهد شد. در هر زمان فقط یک اتوماتای یادگیر در شبکه فعال می باشد.

۲.۲. تئوری مجموعه ناهموار

جهان پیرامون ما سرشار از مسائلی است که با ابهام، عدم قطعیت و گهگاه تناقض در داده ها همراه است. در اقتصاد مفاهیمی نظیر مناسب بودن، کارآمد بودن، به صرفه بودن و هزاران مورد دیگر با آنها بطور روزمره مواجه هستیم، ذاتا با ابهام آمیخته هستند. امروزه با آمیزش حوزه های مختلف علوم، اینگونه مفاهیم را به وفور میتوان در مسائل مهندسی نیز مشاهده نمود. از این رو وجود ابزاری برای کار با داده های مبهم و غیرقطعی ضروری بنظر می رسد. در دهه های اخیر تلاش های بسیاری برای ایجاد و توسعه ابزارهایی بدین منظور صورت گرفته است که تئوری مجموعه ناهموار را می توان مکمل تئوری های سنتی قبلی در این زمینه، نظیر تئوری احتمالات، تئوری شواهد و تئوری مجموعه های فازی در نظر گرفت[۱,۲].

تئوری مجموعه ناهموار اولین بار توسط Pawlak در سال ۱۹۸۲ معرفی شد. این تئوری برپایه رابطه تشخیص ناپذیری و عدم توانایی در تمایز بین اشیاء استوار بوده و یک تقریب از مفاهیم و مجموعه ها را بر پایه روابط دودویی از روی داده های تجربی ایجاد می کند. این تقریب ها مدلی از هدف ما را تشکیل می دهند.

در ادامه به معرفی یک سری از تعاریف پایه و مورد نیاز از این تئوری پرداخته شده است.

- **سیستم اطلاعاتی:** جدولی از داده ها است که هر سطر آن بیانگر یک مثال، نمونه یا حالت از سیستم است. ستونهای این جدول نیز مقدار و یا وضعیت یک متغیر، صفت و یا خصیصه را برای هر نمونه بیان می کند. یک سیستم اطلاعاتی بصورت I=(U,A) بیان می شود که در آن، U یک مجموعه متناهی و ناتهی به نام مجموعه مرجع و A یک مجموعه متناهی و ناتهی از خصیصه ها می باشد. *[۴,۵]*.

- **سیستم تصمیم گیری**: به سیستم اطلاعاتی گفته می شود که یکی از خصیصه ها به نام خصیصه تصمیم، به عنوان نتیجه ای برای دسته بندی نمونه های جدول مطرح می شود. سیستم تصمیم گیری بصورت $I=(U,A\cup\{d\})$ بیان می شود که $d\notin A$ خصیصه تصمیم و اعضای مجموعه A، خصیصه های شرطی یا شروط نامیده می شوند.

- **رابطه تشخیص ناپذیری:** به ازای هر $P\subseteq A$، یک رابطه هم ارزی $IND(P)$ به نام رابطه تشخیص ناپذیری وجود دارد که بصورت زیر تعریف می گردد :

$$IND(P)=[x]_P=\{(x,y)\in U\times U : a(x)=a(y) \quad for\ every\ a\in P\} \qquad (۲)$$

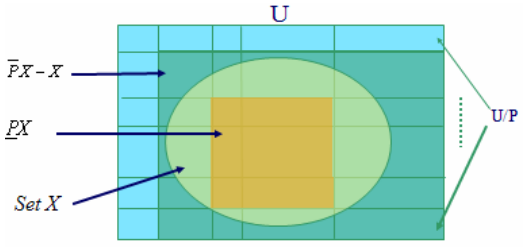
که IND(P)، رابطه تشخیص ناپذیری روی P خوانده می شود. اگر $(x,\ y)\in IND(P)$ باشد. دو نمونه x و y بوسیله خصیصه های P نسبت به هم تشخیص ناپذیرند و در دسته هم ارزی $[x]_p$ قرار می گیرند.

- **مجموعه ناهموار:** گاهی نمی توان به کمک استقرا توصیفی قطعی یا دقیق از نمونه های موجود در جدول یک سیستم اطلاعاتی بدست آورد .در این حالت می توان گفت که خصیصه های یک نمونه از این سیستم می توانند بیانگر سه حالت ممکن باشند : متعلق بودن نمونه به یک پدیدهٔ خاص، عدم تعلق نمونه به آن پدیده و در نهایت قرارگرفتن نمونه در ناحیه مرزی. در صورت قرار گرفتن نمونه ای در ناحیه مرزی، آن مجموعه را یک مجموعهٔ ناهموار می گویند.

- **تقریب مجموعه:** اگر $I=(U,A)$ یک سیستم اطلاعاتی باشد و $P\subseteq A$ و $X\subseteq U$ باشند، آنگاه می توان به کمک اطلاعات موجود در P مجموعه X را تقریب زد (شکل ۳). این تقریب با بیان "تقریب بالا" و "تقریب پائین" از X امکان پذیر است. تقریب بالای X ($\overline{P}X$) با توجه به P شامل اعضایی است که با توجه به خصیصه های P می توان آنها را بعنوان اعضای احتمالی در X دسته بندی کرد(رابطه ۳). همچنین تقریب پائین X ($\underline{P}X$) با توجه به P شامل اعضایی است که با توجه به خصیصه های P می توان آنها را بعنوان اعضای قطعی X دسته بندی کرد (رابطه ۴).

$$\underline{P}X=\{X\,|\,[X]_P\subseteq X\} \qquad (۳)$$

$$\overline{P}X=\{X\,|\,[X]_P\cap X\neq\phi\} \qquad (۴)$$



شکل (۳) : تقریب مجموعه X با توجه به خصیصه های P

- **ناحیه مثبت :** چنانچه $P,Q\subseteq A$ باشد، ناحیه مثبت P نسبت به Q طبق رابطه زیر قابل محاسبه می باشد.

$$POS_P(Q)=\bigcup_{X\in U/Q}\underline{P}X \qquad (۵)$$

- **درجه وابستگی:** چنانچه $P,Q\subseteq A$ باشد، درجه وابستگی بین Q و P بصورت رابطه (۵) قابل تعریف می باشد.

$$K=r_P(Q)=|POS_P(Q)|/|U| \qquad (۶)$$

که |Y| تعداد عناصر موجود در مجموعه Y می باشد. چناچه K برابر با ۱ باشد، Q بطور کامل به P وابسته می باشد که آن بدین معنی است که کلیه مقادیر خصیصه های موجود در Q بصورت یکتا با مقادیر خصیصه های موجود در P قابل شناسایی می باشد.. اگر ۰<K<۱ باشد بدین معنی است که Q به P با درجه K وابسته می باشد. اگر K برابر با ۰ باشد، Q به P به هیچ وجه وابسته نمی باشد.

- **مجموعه کاهش یافته**[8] : اگر R یک زیرمجموعه از C باشد، به R یک مجموعه کاهش یافته می گویند اگر رابطه زیر برقرار باشد.

$$r_R(D) = r_C(D) \wedge \forall R' \subset R, r_{R'}(D) < r_C(D) \quad (۷)$$

همچنین یک مجموعه کاهش یافته با حداقل تعداد عناصر، مجموعه کاهش یافته مینیمم[9] نامیده می شود. هدف از انتخاب خصیصه ها، پیدا کردن یک مجموعه کاهش یافته مینیمم می باشد.

۳. روش پیشنهادی

روش پیشنهادی برای انتخاب ویژگی های مناسب از یک روش ترکیبی مبتنی بر اتوماتای یادگیر توزیع شده و مجموعه ناهموار برای انتخاب ویژگی های مهم استفاده می کند. برای این منظور از یک DLA، بعنوان گراف ورودی بهره می جوید. این گراف از روی نمونه های موجود در مجموعه، بصورت زیر ساخته می شود.

ابتدا نمونه ها به تفکیک کلاس از هم جدا می شوند. سپس به هر خصیصه، یک اتوماتای یادگیر (نود) تخصیص داده می شود. حال به ازای هر نمونه، از هر یک از خصیصه ها(نودها) که دارای مقدار می باشد، به کلیه خصیصه های دیگر که دارای مقدار می باشند، ارتباط برقرار می گردد(اقدام). بدین ترتیب یک گراف (شبکه) از LA ها تولید می گردد.

روش ساده تر برای تولید گراف ذکر شده، ایجاد یک گراف کامل می باشد. اما از آنجا که در وب، ویژگی های هر صفحه با بقیه صفحات بسیار متفاوت می باشد. لذا در هر نمونه، خصیصه های بسیاری دارای مقدار نمی باشند. از این رو ایجاد گراف کامل، به کند شدن روش پیشنهادی و طولانی تر شدن زمان برای رسیدن به جواب منتهی می شود.

با تشکیل این گراف، هدف مسئله به پیدا کردن مسیری در گراف که خصیصه های انتخاب شده در آن مسیر تشکیل یک مجموعه کاهش یافته را بدهند و از cardinality پائینی نیز برخوردار باشند، تبدیل می شود..

الگوریتم پیشنهادی بصورت زیر می باشد.

الگوریتم ۱: الگوریتم پیشنهادی برای انتخاب ویژگی ها

Algorithm۱: *feature selection based DLA & RS*

initialize C = { All nodes } , D = { decision }, R={}, best_R={All nodes} , set ε

Step ۱: Create a DLA Graph from Data

Step ۲: - Select a LA Randomly as LA_k

- Add LA_k to R

Step ۳: while $r_R(D) < r_C(D) - \varepsilon$ and $|R| < |best\_R|$

- Activate an action of LA_k based probability as $Action_{kj}$.
- Add LA_j to R

Step ۴: If $|R| < |best\_R|$

- Give Reward to each LA_k and $Action_{kj}$ in R
- Give reward
- best_R = R

Else

Penalize each LA_k and $Action_{kj}$ in R

در الگوریتم بالا، C خصیصه های شرطی، D خصیصه تصمیم، R خصیصه های انتخاب شده تاکنون، best_R بهترین راه حل پیدا شده تاکنون و ε میزان خطای قابل اغماض می باشد.

در ابتدای الگوریتم، مجموعه C، R و مقدار ε (معمولا ۰.۱) مقداردهی اولیه می گردند. یک گراف DLA از روی داده ها ساخته شده و به هر خصیصه یک اتوماتای یادگیر تخصیص داده می شود. سپس DLA شروع بکار نموده و در هر مرحله، اتوماتای فعال، یکی از اقدام های خود را انتخاب و به مجموعه R اضافه می کند. میزان وابستگی بین مجموعه خصیصه های موجود در R با خصیصه تصمیم محاسبه شده و با میزان وابستگی بین مجموعه خصیصه های موجود در C مقایسه می گردد. در صورتیکه راه حل یافت شده بهتر از راه حل های قبلی بود و تفاوت میزان وابستگی ها کمتر از ε بود، این حلقه تکرار می گردد. پس از خروج از حلقه، چناچه راه حل مناسبی پیدا شده بود، کلیه اقدامهای انتخاب شده از اتوماتاهای یادگیر موجود در R پاداش می گیرند و در غیر این صورت کلیه اقدامهای انتخاب شده از اتوماتاهای یادگیر موجود در R جریمه می شوند.

۴. ارزیابی روش پیشنهادی

در این بخش به بررسی و ارزیابی روش پیشنهادی در مقایسه با سایر الگوریتم ها می پردازیم. برای اینکار از شش مجموعه داده استفاده شده است. چهار مجموعه داده Vote، Cancer، Cpu و Flag جزو مجموعه داده های استانداردی می باشند که در UCI Repository قابل دسترس می باشند. مجموعه داده Adeater برگرفته از ویژگی های بکار گرفته شده در حذف عکس های تبلیغاتی در صفحات اینترنت می باشد. این مجموعه داده دارای ۳ خصیصه عددی و ۱۵۵۵ خصیصه دودویی است . مجموعه داده Insurance نیز دارای ۸۵ خصیصه عددی می باشد ، که این داده ها برگرفته شده از اطلاعات بیمه گذاران می باشد و برای پیش بینی سرمایه گذاری در یک نوع بیمه خاص مورد استفاده قرار گرفته است . هر دو مجموعه همچنین در [۶] به مسابقه گذاشته شده است.. در جدول زیر مشخصات مربوط به مجموعه داده های استفاده شده در این کار نشان داده شده است.

جدول (۱): مجموعه داده های مورد استفاده.

نام مجموعه داده	تعداد نمونه ها	تعداد خصیصه ها	درصد مقادیر گمشده	تعداد کلاس ها
Vote	۴۳۵	۱۷	٪۴	۲
Cancer	۶۹۹	۱۰	٪۱۰	۲
CPU	۲۰۹	۸	٪۲	۲
Flag	۲۴	۲۷	٪۰	۲
[۶]Insurance	۵۸۲۲	۸۶	٪۰	۲
[۴]ADeater	۳۲۷۹	۱۵۵۹	٪۳	۲

در مجموعه داده های اول تا پنجم می توان از گراف کامل نیز بهره برد (چون تعداد خصیصه ها کم می باشند) اما در مجموعه داده adeater به شیوه مطرح شده، گراف DLA ساخته می شود.

همچنین روش پیشنهادی برای انتخاب ویژگی ها با دو الگوریتم دیگر به نام های روش GA و Best First برای انتخاب ویژگی ها مورد مقایسه و ارزیابی قرار گرفته است.

نتایج حاصل از اعمال روش پیشنهادی بر روی مجموعه داده ها بصورت جدول (۲) می باشد. نتایج بدست آمده، حاصل از ۲۰ بار اجرای الگوریتم و میانگین گیری می باشد.

جدول (۲) : نتایج حاصل از اجرای الگوریتم پیشنهادی بر روی مجموعه داده ها.

نام مجموعه داده	تعداد خصیصه های اولیه	تعداد مشخصه ها		
		روش الگوریتم ژنتیک	روش Best First	روش پیشنهادی
Vote	۱۷	۵ (٪۸۵)	۷ (٪۸۵)	۵ (٪۸۵)
Cancer	۱۰	۷ (٪۱۰۰)	۷ (٪۱۰۰)	۷ (٪۱۰۰)
CPU	۸	۳ (٪۱۰۰)	۳ (٪۱۰۰)	۳ (٪۱۰۰)
Flag	۱۰	۷ (٪۱۰۰)	۸ (٪۱۰۰)	۷ (٪۹۸)
Insurance	۸۶	۳۴ (٪۱۰۰)	۲۵ (٪۹۵)	۲۶ (٪۹۷)
Adeater	۱۵۵۹	۳۸۰ (٪۹۲)	۴۲۰ (٪۹۴)	۴۵۰ (٪۹۸)

در جدول بالا، درصد موجود در هر سلول میزان پوشش رکوردها را مشخص می کند. برای مثال در روش الگوریتم ژنتیک، روی مجموعه داده Vote، ۸۵ درصد از رکوردها با استفاده از ۵ خصیصه انتخاب شده بصورت یکتا قابل شناسایی می باشند. همانطور که مشاهده می شود، نتایج بر روی اکثر مجموعه داده ها حکایت از کارایی روش پیشنهادی نسبت به سایر روشها دارد.

۵. نتیجه گیری

در این مقاله یک روش جدید برای انتخاب ویژگی های مناسب از صفحات وب مبتنی بر تئوری مجموعه ناهموار و اتوماتای یادگیر توزیع شده ارائه شد. تئوری مجموعه ناهموار و اتوماتای یادگیر توزیع شده مورد معرفی قرار گرفت و بخوبی از قابلیت های تئوری مجموعه ناهموار در کار بر روی مجموعه های غیردقیق و دارای ابهام و از اتوماتای یادگیر توزیع شده برای جستجو در فضاهای بزرگ بهره گرفته شد. بمنظور ارزیابی، روش پیشنهادی بر روی تعدادی مجموعه داده استاندارد آزمایش شد و با نتایج حاصل از دو الگوریتم GA و Best First مورد مقایسه قرار گرفت. نتایج حاکی از کارایی این الگوریتم در مقابل روشهای دیگر بود. از مزایای این روش می توان به پیدا کردن یک مجموعه از reduct ها بجای پیدا کردن یک reduct، در مقایسه با الگوریتم های دیگر ارزیابی شده، اشاره نمود. از معایب آن نیز می توان به مشکل ذاتی اتوماتای یادگیر در همگرایی دیررس اشاره کرد که در این مقاله با تشکیل گراف غیر کامل تا حدی این مشکل برطرف شده است. همچنین برای بهبود سرعت بر روی مجموعه داده های بزرگ، می توان ابتدا مجموعه داده را به چند قسمت تقسیم کرده و الگوریتم را روی هر قسمت اعمال نمود و سپس از اشتراک راه حل های بدست آمده، پاسخ نهایی را بدست آورد.

مراجع

[۱] Lian-Yin Zhai*, Li-Pheng Khoo, Sai-Cheong Fok , Feature extraction using rough set theory and genetic algorithms an application for the simplification of product quality evaluation , Computers & Industrial Engineering ۴۳ (۲۰۰۲) ۶۶۱–۶۷۶.

[۲] Komorowski, J., Pawlak, Z., Polkowski, L. & Skowron, A. (۱۹۹۹). Rough sets: A tutorial ۹۸.

[۳] Narendra, K. S., and Thathachar, M. A. L., Learning Automata: An Introduction, Printice-Hall Inc, ۱۹۸۹.

[۴] Nicholas kushmerick , Learning to remove Internet advertisements ,۳rd Int. Conf. on Autonomous Agents ۱۹۹۹.

[۵] Thathachar, M. A. L., Sastry, P. S., "Varieties of Learning Automata: An Overview", IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. ۳۲, No. ۶, PP. ۷۱۱-۷۲۲, ۲۰۰۲.

[۶] Coil Challenge ۲۰۰۰ - http://www.dcs.napier.ac.uk/coil/challenge/.

[۷] Nicholas kushmerick , Learning to remove Internet advertisements ,۳rd Int. Conf. on Autonomous Agents ۱۹۹۹.

[۸] M. Kantardzic , Data Mining Concepts , Models , Method and Algorithms , Wiley-InterScience , ۲۰۰۳ (Chapter ۲)

[۹] R.J. Roiger , M.W.Geatz , Data Mining A Tutorial – Based Primer , Addison Wesley ۲۰۰۳.

زیرنویس

[۱] Rough Set Theory
[۲] Feature Ranking
[۳] Minimum Subset
[۴] Linear Reward-Penalty
[۵] Linear Reward Epsilon Penalty
[۶] Linear Reward Inaction
[۷] Distributed Learning Automata
[۸] Reduct Set
[۹] Minimal Reduct