کاربرد الگوریتم های تکاملی در داده کاوی

مصطفی تقوی، نازک نوبری
گروه فلسفه علم دانشگاه صنعتی‌شریف
taghavi11@yahoo.com

چکیده

داده‌کاوی عبارت است از فرایند کشف و تحلیل حجم بزرگی از داده‌ها با استفاده از روش‌های آماری و ریاضی. این مقاله توصیفی است از کاربرد الگوریتم‌های تکاملی(EAs)[1] در داده‌کاوی و کشف دانش. هدف، بیان شیوه انطباق افراد، اپراتورهای ژنتیک و توابع هماهنگی با استخراج دانشی سطح بالا از داده است. در بیشتر موارد، داده‌های حاصل از سیستم‌های بزرگ و پیچیده، الگوی مشخصی ندارد و در طی زمان و مکان تغییر می‌کند. بنابراین برای تحلیل این نوع داده‌ها باید به دنبال روش‌هایی بود که بتواند روش‌های آماری کلاسیک را کامل کند. داده‌کاوی بر اساس روش‌های آماری کلاسیک، هم زمان قابل توجّهی را صرف می‌کند و هم مسبوق به نظریّه است.

در این مقاله به منظور بررسی نقش الگوریتم اکتشافی، با عنوان الگوریتم تکاملی، در گام‌های متوالی فرایند داده‌کاوی به تعریف و توصیف فرایند داده‌کاوی و الگوریتم‌های تکاملی پرداخته و با نظر به مفاهیم هر یک از آنها، رابطه الگوریتم‌های تکاملی و فرایند داده‌کاوی را بررسی کرده ایم. در نهایت معلوم گردیده است که تنها فرایندی مانند فرایند داده‌کاوی با قابلیت تحلیلی زیاد می‌تواند اطلاعاتی قابل درک برای انسان ارائه نماید و نیز معلوم گردیده است که چگونه دو فرایند داده-کاوی و الگوریتم تکاملی یکدیگر را پشتیبانی می‌نمایند.

کلید واژه‌ها: داده، استخراج، انتخاب، طبقه‌بندی، دسته‌بندی، داده‌کاوی، الگوریتم تکاملی

۱ـ مقدمه

با نظر به این امر که حجم دادهٔ ذخیره شده در پایگاه‌های داده[۲] به سرعت زیاد می‌شود و این حجم عظیم دادهٔ ذخیره شده شامل دانشی با ارزش، اما پنهان است و هر داده می‌تواند نقشی مؤثر در فرایند تصمیم‌گیری داشته باشد و از طرفی دیگر، تعداد تحلیل‌گران داده با نرخی کمتر از حجم داده‌های ذخیره شده افزایش میابد؛ این خود دلیلی برای استفاده از روش‌های خودکار در استخراج دانش از داده است. داده‌کاوی قدم اصلی فرایند گسترده‌ای است که کشف دانش[۳] از پایگاه داده نام دارد..

1. Evolutionary Algorithm

۲ .Data base . پایگاه داده: مجموعه ای از داده است که به روشی نظام‌مند و معمولا در رایانه، ذخیره می‌شوند. مانند برنامه‌ای رایانه‌ای که برای پاسخ به پرسش‌ها استفاده می‌شود و منبعی است برای تصمیم‌گیری.

3. Knowledge discovery

این فرایند شامل کاربرد چندین روش پیش پردازشی[1] با هدف تسهیل کاربرد الگوریتم داده‌کاوی و روش‌های فرا پردازشی[2] با هدف پردازش مجدد و ارتقای دانش بدست آمده است.

ماهیت داده‌کاوی، استخراج دانش از داده با استفاده از ساز و کارهایی نسبتاً خودکار است. اما این عبارت سوالی را پیش می‌آورد که چه نوع دانشی باید کشف شود و این کشف چگونه محقق می‌شود. در داده‌کاوی به‌دنبال کشف دانشی با قابلیت پیش‌بینی‌کنندگی هستیم. در حقیقت انسان به دنبال کشف دانش با درجه بالایی از صحت است. بعلاوه این دانش کشف شده باید برای کاربر نیز قابل درک باشد تا در نهایت بتواند وی را در تصمیمات پشتیبانی نماید. در غالب موارد، اگر دانش کشف شده صرفاً به مثابه جعبه‌ای سیاه باشد و پیش‌بینی را بدون بیان هر گونه توضیحی انجام دهد، کاربر به صحت آن اعتماد نخواهد کرد.(D. Michael *et al*, 1994)

۲ـ داده‌کاوی

با تعریف داده‌کاوی به صورت فرایند کشف و تحلیل دادهٔ متعلق به مجموعه‌های بزرگ با استفاده از ابزار خودکار و بر اساس الگوها و قوانین معنی‌دار (Michael Berry *et al*, 2000)، متوجه می‌شویم که داده‌کاوی استفاده از ابزار تحلیل داده برای کشف ناشناخته‌ها، الگوهای معتبر و روابط مجموعه‌های بزرگ داده است.(Pieter Adriaans *et al*, 1996) این ابزار شامل مدل‌های آماری، الگوریتم‌های ریاضی و روش‌های یادگیری ماشینی[3](الگوریتم هایی که عملکردشان را به صورت خودکار و با نظر به تجربه‌ها بهبود می‌دهند. مانند شبکه‌های عصبی[4] یا درخت‌های تصمیم‌گیری[5])است. پس داده‌کاوی چیزی بیش از جمع‌آوری و اداره داده است و تحلیل و پیش‌بینی داده را نیز شامل می‌شود.

معمولاً در تحلیل داده از روش‌هایی مانند پرسشنامه ساختاریافته(متداول در اکثر پایگاه های تجاری داده) یا نرم افزارهای تحلیل آماری استفاده می‌شود. اکثر ابزارهای ساده تحلیلی از روشی مبتنی بر تحقیق استفاده می‌کنند. نقطه آغازین فرایند این ابزارهای ساده، ارائهٔ فرضیّه است و سپس با استفاده از داده، درستی یا نادرستی آن فرضیّه آزمون می‌شود (Seifert. Jeffrey W, 2004). در این روش‌ها بحث مسبوقیت مشاهده بر نظریه مطرح می‌شود که خود دلیلی برای محدودیت کاربرد این نوع ابزارهای ساده آماری است.

باید توجه داشت که اثربخشی این روش با توجه به قوه ابتکار کاربر در ارائهٔ فرضیّه‌های گوناگون متناسب با ساختار نرم افزار مورد استفاده محدود می‌شود. در مقابل، داده‌کاوی نیز با به‌کارگیری روشی اکتشافی و کاونده به آزمون هم‌زمان روابط بین داده می‌پردازد و درجه یگانگی و تکراری بودن داده را بررسی می‌کند. روش‌های داده‌کاوی هیچ فرض یا ساختاری را بر داده تحمیل نمی‌کنند و به داده اجازه می‌دهند تا خود، مراحل بعدی را تعیین کند. البته روش‌های داده‌کاوی، جانشین روش‌های آماری سنّتی نیستند، بلکه آنها را کامل می‌کنند. پس می‌توان گفت داده‌کاوی فرایندی مربوط به الگوهای اکتشافی، پیوسته، مدل‌های آماری و نامتعارف در داده است.(Fayyad et al., 1996). این تعریف به مواردی اشاره دارد که داده از نظر حجم، بسیار بزرگ و یا انجام تحلیل دو جانبه بیش از حد بغرنج است.

---

1. Preprocessing
2. Post processing
3. Machine learning . یادگیری ماشینی از مباحث هوش مصنوعی است و روشی است برای ایجاد برنامه‌های رایانه ای و از تحلیل مجموعه داده استفاده می‌کند. با استفاده از این برنامه‌ها رایانه قابلیت یادگیری پیدا می‌کند. (Michie D., D.J. Spiegelhalter and C.C. Taylor. 1994)
4. Neural network. شبکه عصبی شامل گروهی از یاخته‌های عصبی زیستی(مغز انسان) و یا مصنوعی (شبیه‌سازی‌های مکانیکی و یا الکتریکی شبکه‌های عصبی زیستی) است. شبکه‌های عصبی در این مقاله، همان شبکه‌های عصبی مصنوعی است. (Michie D., D.J. Spiegelhalter and C.C. Taylor. 1994)
5. Decision tree

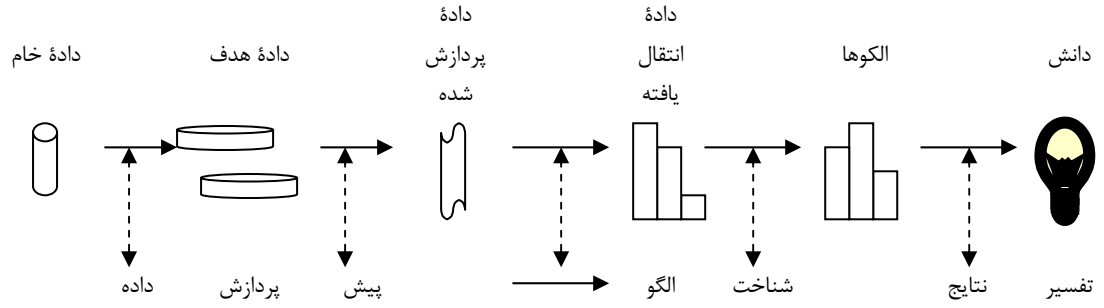
۱ـ ۲ـ ضرورت داده‌کاوی

روش‌های بهبود مدیریت داده، نقشی مهم در افزایش قابلیت استفاده از اطلاعات و کاهش هزینه‌های ذخیره سازی دارند. زیرا طی دهه‌های گذشته شاهد افزایشی سریع در حجم اطلاعات جمع آوری و ذخیره شده(با نظر گرفتن این فرض که هر سال تقریبا حجم داده جهان دوبرابر می‌شود) هستیم (Patrick Dillon, 1998).

با نظر به افزایش سریع قدرت رایانه‌ها در دهه‌های گذشته و افزایش تعداد مجموعه‌های بزرگ داده و شناخت ارزش تغییرات ملایم[1] دیگر، روش‌های سنتی به تنهایی قادر به ارائه تحلیل‌هایی قدرتمند از داده نیستند و این خود بر ضرورت استفاده از قابلیت بالای روش‌شناسی تحلیل رایانه[2] تأکید می‌کند.

بسیاری از سازمان‌ها چه در بخش خصوصی و چه در بخش دولتی، با استفاده از فناوری‌ها و فرایندهای کسب و کار پیشرو، جذب روش‌های داده‌کاوی شده‌اند. بعضی از این تغییرات عبارتند از: رشد شبکه‌های رایانه‌ای[3](اتصال به پایگاه‌های داده)، توسعه فنون جستجو مانند شبکه‌های عصبی و الگوریتم‌های پیشرفته، گسترش مدل‌های مشتری - کارگر[4] ، افزایش میزان دسترسی کاربران به منابع مرکزی داده و افزایش توانایی ترکیب داده از منابع گوناگون و تبدیل آن به یک منبع واحد قابل جستجو.(Pieter Adriaans *et al*, 1996). به‌علاوه، سازمان‌ها از داده‌کاوی به عنوان ابزاری در تحقیقات مشتری، صرفه جویی و تحقیقات پزشکی نیز استفاده می‌کنند.

داده‌کاوی از دو گام اصلی پیش پردازش داده و شناخت الگو تشکیل می‌شود. پردازش داده شامل مواردی است که یا تعداد ترکیب‌های عناصر داده زیاد باشد و یا نشانه‌ها، برگرفته از چندین دادهٔ ساده باشند. معمولا فرایند پیش پردازش زمان‌بر است. شناخت الگو نیز در جایی‌که الگوی داده ترکیبی باشد، استفاده می‌شود.



شکل شماره ۱- داده‌کاوی به عنوان فرایندی تکراری و تعاملی

از فنون داده‌کاوی برای تحلیل داده در زمینه‌های گوناگون استفاده می‌شود و زمانی‌که بیشترین توجه داده‌کاوی بر الگوریتم‌های شناخت الگو باشد، گام‌های پیش پردازش نقشی مهم در موفقیت فرایند خواهند داشت.(Langley and Simon, 1995; Burl et al., 1998).

۲ـ ۲ـ محدودیت‌های داده‌کاوی

همان اندازه که نتایج داده‌کاوی می‌توانند بیانگر ابزاری قدرتمند برای پاسخ به پرسش‌ها باشند، این فرایند در اجرا بی‌نیاز از غیر نیست. داده‌کاویِ موفق، نیازمند مهارت‌های فنی و متخصصان تحلیلگری است که قادر به تحلیل و تفسیر نتایج بدست

1. Untapped
2. Computer- intense
3. Computer networks
4. Client - server

آمده باشند. پس می‌توان گفت، محدویت‌های داده‌کاوی بیشتر فردگرا و مربوط به داده‌های اولیه است تا فن‌گرا. زیرا فرد در دام مسبوقیت مشاهده بر نظریه می‌افتد و در تحلیل‌های خود دست به تحریف می‌زد. این تحریف می‌تواند به صورت خودآگاه و یا ناخودآگاه صورت گیرد.

اگرچه داده‌کاوی راهی برای تعیین الگوها و روابط میان آنها است؛ هرگز ارزش و اهمیت این الگوها را به کاربران نشان نمی‌دهد. به طور مشابه، اعتبار الگوهای کشف شده نیز وابسته به مقایسه آنها با شرایط دنیای واقعی است.

محدودیت دیگر داده‌کاوی، در شیوه شناسایی روابط بین رفتارها و متغیرها است که ضرورتاً این شناسایی نتیجه فرایندی اتفاقی نخواهد بود.

۳ـ ۲ـ مباحث داده‌کاوی

علاوه بر اهمیت قابلیت‌های مربوط به شیوه این فرایند در کشف و تحلیل داده، عوامل دیگری نیز مانند کیفیت داده، توانایی عملکرد[1]، هدفمندی[2]، محدودیت دسترسی عمومی و ... بر موفقیت نتایج طرح تأثیر می‌گذارند. Seifert. Jeffrey W, (2004)

کیفیت داده

کیفیت داده مبحثی چالش برانگیز در فرایند داده‌کاوی است و به صحت و کامل بودن داده بر می‌گردد. بعلاوه، داده‌کاوی می‌تواند بر ساختار و درجه سازگاری داده تحلیل شده نیز اثر بگذارد. تگرار و نسخه‌برداری از داده‌های ضبط شده، عدم وجود استانداردها، شرایط زمانی در به‌روز آوری و خطاهای انسانی از جمله عوامل اثرگذار بر اثربخشی فنون پیچیده داده‌کاوی هستند(تفاوت‌های ظریف بین داده و حساسیت آن‌ها).

توانایی عملکرد

با توجه به کیفیت داده، مبحث توانایی عملکرد به پایگاه‌های متفاوت داده و نرم افزارهای داده کاوی[3] بر می‌گردد. این مفهوم توانایی سامانه رایانه‌ای و یا داده را برای کار با سایر سامانه‌ها یا استفاده از داده در فرایندهای جاری و مرسوم، تعیین می‌کند.

توانایی عملکرد پایگاه‌های داده و نرم افزار مورد استفاده، در فرایند داده‌کاوی نقشی مهم دارند. زیرا در قابلیت جستجو و تحلیل هم‌زمان پایگاه‌های چندگانه داده و تضمین سازش‌پذیری فعالیت‌های فرایند داده‌کاوی مؤثر هستند.

هدفمندی

هدفمندی، یکی از مخاطره‌های مشخص داده‌کاوی است و بیانگر نحوه کنترل اطلاعات فرد است. این مفهوم توجه کاربر را به هدف داده‌کاوی و اولویت جمع‌آوری داده جلب می‌کند.

یکی از دلایل اولیه اشتباه در نتایج، وجود داده‌های نادرست[4] است. تمام مجموعه‌های داده به دنبال صحت داده هستند. البته اگر خود داده ارزش اقتصادی بالایی نداشته باشد، هزینه کسب اطمینان از صحت داده هرگز توجیه‌پذیر نخواهد بود.

محدودیت دسترسی عمومی[5]

عامل اثر گذار دیگر، امنیت داده و محدودیت دسترسی به آن است. این مفهوم نقشی مهم در میزان تسهیم اطلاعات و شروع فرایند داده‌کاوی خواهد داشت.(Seifert. Jeffrey W, 2004)

---

1. Interoperability
2. Mission creep
3. Soft wares of Data Mining. مانند نرم افزار Orange که شامل فرایند مدل‌سازی و فنون کشف داده است.
4. Inaccurate
5. Privacy

۴ـ ۲ـ وظایف داده‌کاوی

به طور کلی اگر وظایف اصلی داده‌کاوی را در طبقه‌بندی، انتخاب و استخراج بدانیم هر کدام از این وظایف می‌توانند به منزله مشکلی در نظر گرفته شوند که الگوریتم داده‌کاوی به رفع آن می‌پردازد.

۱ـ ۴ـ ۲ـ طبقه بندی[1]

طبقه‌بندی وظیفه‌ای قابل توجه است. این وظیفه طی دهه‌های گذشته از طریق یادگیری ماشینی و جوامع آماری مطالعه شده است. هدف این وظیفه، پیش‌بینی ارزش صفت مورد نظر کاربر، مبتنی بر ارزش سایر صفت‌هایی(خصایصی) است که صفات پیش‌بینی کننده نام دارند.

قوانین طبقه‌بندی که به صورت اگر- آنگاه هستند، نوع خاصی از قوانین پیش‌بینی در نظر گرفته می‌شوند. قسمت اگر در این قوانین، شامل ترکیبی از موقعیت‌های ارزش صفت پیش‌بینی است و قانون نتیجه در قسمت آنگاه، شامل ارزش پیش بینی شده برای صفت هدف است.

در وظیفه طبقه‌بندی، دادهٔ کاویده شده به دو مجموعه آموزشی[2] و آزمون تقسیم می‌شود. الگوریتم داده‌کاوی تنها با دسترسی به مجموعه آموزش به کشف قوانین می‌پردازد. برای این منظور، الگوریتم مورد نظر باید هم به ارزش صفت‌های پیش-بینی و هم به صفت هدف هر نگاشته[3](مدرک) در مجموعه آموزش دسترسی داشته باشد.

زمانی‌که فرایند طبقه‌بندی به پایان می‌رسد و الگوریتم، مجموعه‌ای از قوانین را پیدا می‌کند؛ قابلیت پیش‌بینی این قوانین در مجموعه آزمون ارزیابی می‌شود.

۲ـ ۴ـ ۲ـ مدل سازی وابسته[4]

این وظیفه در واقع تعمیم وظیفه طبقه‌بندی است. در طبق‌بندی، پیش از آنکه بخواهیم ارزش چندین صفت را برای هدف پیش‌بینی کنیم، مجددا به کشف قوانین (اگر- آنگاه) پیش‌بینی، برای دستیابی به دانشی سطح بالاتر می‌پردازیم.

البته این شکلی عمومی است و هر صفتی می‌تواند هم در قانون مقدم(قسمت اگر) واقع شود و هم در قانون نتیجه(قسمت آنگاه). اما باید توجه داشت که صفت مور نظر نمی‌تواند هم‌زمان در هر دو قسمت قرار گیرد.

۳ـ ۴ـ ۲ـ دسته بندی[5]

همانطور که قبلا ملاحظه شد، در وظیفه طبقه‌بندی، طبقه آموزشی در حکم ورودی الگوریتم داده‌کاوی در نظر گرفته می‌شود و تعیین کننده شکلی از یادگیری نظارتی است. اما در وظیفه دسته بندی، الگوریتم داده‌کاوی باید با جداکردن موارد به دسته‌هایی که هر کدام شکلی از یادگیری نظارت نشده هستند، خودش به کشف روابط و دانش بپردازد.(Fisher DH, 1987 و Fisher D and Hapanyengwi G, 1993). البته تنها یک‌بار که دسته‌ها مشخص شدند، هر دسته می‌تواند به عنوان یک طبقه در نظر گرفته شود. بنابراین می‌توان یک الگوریتم طبقه بندی را بر اساس دادهٔ دسته بندی شده اجرا نمود. الگوریتم ژنتیک در دسته‌بندی داده کاربردی فراوان دارد.

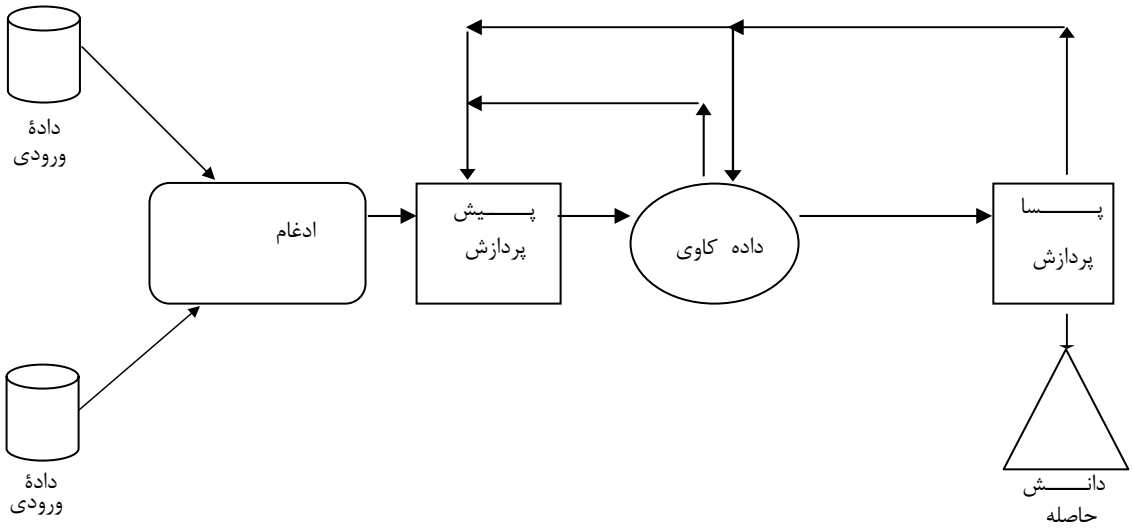
(Falkenauer E. 1998 وFreitas AA. 1999 و Park Y and Song M. 1998).

۴ـ ۴ـ ۲ـ کشف قوانین پیوسته [6]

در شکل استاندارد این وظیفه، هر دادهٔ مشاهده شده یا ضبط شده شامل مجموعه‌ای از صفت‌های دوتایی است که آیتم[7] نام می‌گیرند. گرچه هم طبقه‌بندی و هم قوانین پیوسته، ساختاری(اگر- آنگاه) دارند، از هم متفاوت هستند. باید توجه که در

1. Classification
2. Training test
3. Record
4. Dependence modeling
5. Clustering
6. Discovery of Association Rules
7. Item

وحله نخست قوانین پیوسته می‌توانند بیشتر از یک آیتم را در قانون نتیجه شامل شوند و این درحالی است که قوانین طبقه-بندی همیشه دربرگیرنده یک صفت(هدف) هستند. بعلاوه، برعکس وظیفه پیوستگی، وظیفه طبقه‌بندی، قائل به تناسب بین رابطه صفت‌های پیش‌بینی و صفت هدف نیست و صفت‌های پیش‌بینی فقط می‌توانند در قانون مقدم واقع شوند ور این در حالی است که صفت هدف صرفاً در قانون نتیجه قرار می‌گیرد. (Freitas, AA, 2000)



شکل شماره ۲- مروری بر فرایند کشف دانش. (Freitas , Alex A.2000)

۵ـ ۲ـ روش‌های داده‌کاوی

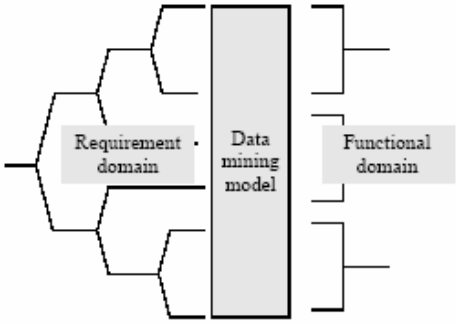
معمولاً در داده‌کاوی از روش‌هایی مانند روش‌های زیر استفاده می‌کنند:

- روش‌های مبتنی بر درخت[1]
- خط رگرسیون چندگانه
- شبکه‌های عصبی
- دسته‌بندی

در این میان مناسب‌ترین روش، روشی است که خصوصیات زیر را داشته باشد:

- انتخاب خودکار پیش‌بینی کننده با هدف استفاده در مدل.
- قابلیت برخورد با آشوب‌های ناگهانی[2] ، اختلالات در داده و داده ناقص.
- ارائه تصویری واضح از ترکیب نتایج و بازخورهای قابل استفاده برای تحلیل.
- وجود مدل‌هایی با ساختاری قوی.
- تأکید مدل بر یادگیری، آزمون و اعتبار داده.

1. Tree based methods
2. Flawed



شکل شماره ۳- نقش داده‌کاوی در تابع انتقال (Fayyad, U. 1996.)

۳ـ محاسبه تکاملی

محاسبه تکاملی عبارت است از سامانه‌های حل مساله‌ای که مدل‌های محاسباتی فرایندهای حل مساله را به عنوان عنصر اصلی در طراحی و اجرا به کار می‌برند. تعدادی از مدل‌های محاسباتی تکاملی توسعه یافته عبارتند از: الگوریتم‌های تکاملی، الگوریتم‌های ژنتیک، استراتژی تکامل[1]، برنامه‌ریزی تکاملی[2] و حیات مصنوعی[3] (Kusiak , Andrew, 2000)

۱ـ ۳ـ الگوریتم تکاملی چیست؟

الگوریتم تکاملی(EA) : الگوریتمی است که جنبه‌های انتخاب طبیعی و تداوم هماهنگی را ترکیب می‌کند. الگوریتم تکاملی از جمعیت ساختارهای قوانین انتخاب، ترکیب‌بندی مجدد، تغییر و بقا، حفاظت می‌کند. این ساختارها مبتنی بر اپراتورهای ژنتیکی هستند. در این روش، محیط خود تعیین‌کننده هماهنگی یا عملکرد هر یک از افراد جمعیت است و از افراد هماهنگ‌تر برای تولید مجدد استفاده می‌کند.(Gopalakrishnan and A.Gunasekaran, 2000)

پس می‌توان گفت: الگوریتم‌های تکاملی رویه‌های جستجو تصادفی با استفاده از ساز و کار ژنتیکی و انتخاب طبیعی هستند. با مروری بر تاریخچه الگوریتم تکاملی مشخص می‌شود ایده اصلی تمام این نوع الگوریتم‌ها یکی است و به فرض جمعیتی از افراد برمی‌گردد که فشار محیطی آنها را وادار به انتخاب طبیعی می‌کند و تابعی کیفی برای حداکثر کردن آن چیزی خواهند بود که به صورت تصادفی موجب پدید آمدن مجموعه‌ای از راه حل‌ها و عناصر می‌شود. به‌کارگیری تابع کیفی به عنوان معیاری برای هماهنگی در نظر گرفته می‌شود. بر اساس این هماهنگی، بعضی از کاندیداهای بهتر برای ایجاد نسل بعدی انتخاب می‌شوند و این فرایند تا رسیدن به جواب بهینه یا پایان یافتن زمان اجرا ادامه میابد. الگوریتم‌های تکاملی کاربردهای مختلفی دارند.

(Fonseca and Fleming, 1995. Bäck, 1996. Coello, 1999. and Van Veldhuizen and Lamont,2000)

معمولاً از این الگوریتم‌ها در توابع بهینه‌سازی استفاده می‌شود. زیرا الگوریتم‌های تکاملی مناسب بهینه‌سازی، تابع ترکیبی هستند و این نقش را با توجه به کروموزمهایشان در ارائه راه‌حل‌ها ایفا می‌کنند. هر فرد می‌تواند رشته‌ای ساده از صفر و یک و یا به پیچیدگی یک برنامه رایانه‌ای باشد. جمعیت اولیه به صورت تصادفی انتخاب می‌شود. سپس الگوریتم به منظور دستیابی به راه‌حل بهینه به ارزیابی افراد می‌پردازد. معمولا هر راه حل منحصر به همان مسأله است و باید توسط کاربر حمایت و اداره شود. فرایند تکاملی باعث می‌شود تا جمعیت با محیط تطابق بهتر و بهتری پیدا کند.

اجرای فرایند تکاملی در بسیاری از موارد اتفاقی است. هنگام انتخاب، افراد مناسب‌تر شانس بیشتری نسبت به سایرین دارند. اما معمولاً حتی نامناسب ترین فرد هم شانسی برای والد و انتخاب شدن دارد. در ترکیب افراد، انتخاب اینکه کدام

1 . Evolution strategy
2. Evolutionary programming
3. Artificial life

قسمت ترکیب شود نیز تصادفی است. البته الگوریتم تکاملی متفاوت از الگوریتم‌های ژنتیک است. الگوریتم ژنتیک هر یک از افراد را بر اساس کدهای شناخته شده‌ای به نام کروموزوم تولید می‌کند و کروموزوم‌ها برای تولید افراد جدید یا ترکیب می‌شوند و یا جهش می‌یابند .(Goldberg, D.E., 1989)

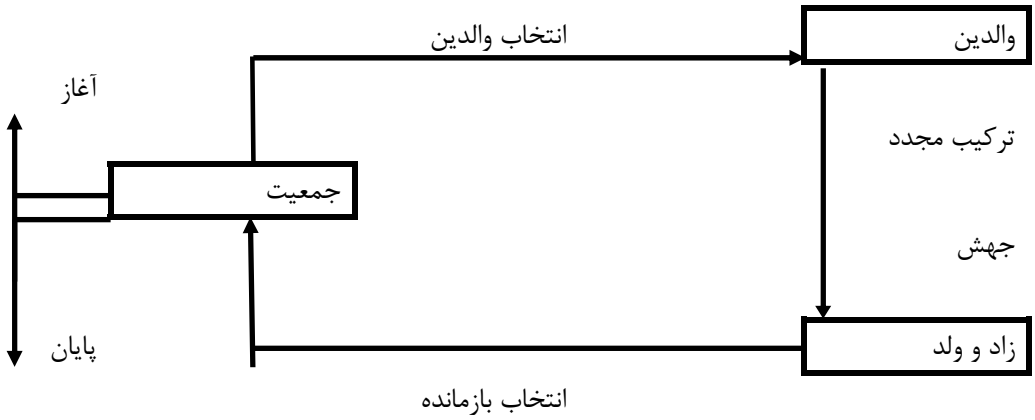
اجزای الگوریتم‌های تکاملی را می‌توان به‌ترتیب زیر در نظر گرفت.

- تعریف افراد
- ارزیابی تابع(تابع هاهنگی)
- انتخاب جمعیت
- مکانیزم انتخاب والد
- اپراتورهای متغیّر ، ترکیب مجدّد و جهش
- ساز و کار انتخاب بازمانده (جانشینی)

هر کدام از این اجزا باید با توجه به یک الگوریتم تکاملی تعیین شوند.

طرح عمومی یک الگوریتم تکاملی را نیز می‌توان به صورت زیر در نظر گرفت.

Begin
INITIALISE population with random candidate solutions;
EVALUATE each candidate;
REPEAT UNTIL (TERMINATION CONDITION is satisfied)
1-SELECT parents;
2-RECOMBINE pairs of parents;
3-MUTATE the resulting offspring
4-EVALUATE new candidate
5-SELECT individuals for the next generation;
END



شکل شماره ۴- فلوچارت طرح عمومی الگریتم تکاملی (Smith, S. F. 1983)

۲ـ۳ـ راهبرد تکامل[1] (ES)

راهبرد تکامل که برای اولین بار توسط اینگو ریچن‌برگ و هانس ـ پائول شوفول[2] در سال ۱۹۶۳ در دانشگاه برلین ارائه شد، الگوریتمی است که افراد (راه حل های بالقوه) با توجه به مجموعه‌ای از متغیرهای مشاهده‌ای با ارزش واقعی کدگزاری

1. Evolution Strategy
2. Ingo Rechenberg and Hans- Paul Schwefel

می‌شوند. برای هر متغیر هدف، هر فرد راهبردی متغیر دارد که این راهبردها درجه تغییر متغیرهای مربوط راتعیین می کنند. متغیرهای راهبردی با توجه به نرخ تغییر متغیرهای هدف، خود تغییر می‌کنند. ویژگی‌های یک راهبرد تکامل به‌وسیله اندازه جمعیت و تعداد زاد و ولد در هر مرحله تعیین می‌شود.

۳ـ ۳ـ برنامه‌ریزی ژنتیک[1] (GP)

تفاوت اصلی الگوریتم ژنتیک و برنامه‌ریزی ژنتیک در شیوه ارائه راه حل است. برنامه‌ریزی ژنتیک برنامه‌های رایانه‌ای را به زبان خود رایانه به عنوان راه حل تولید می‌کند. اما الگوریتم ژنتیک با ایجاد رشته‌ای از اعداد که آن اعداد خود معرف راه‌حل‌ها هستند، ویژگی می‌یابد.

برنامه‌ریزی ژنتیک شامل چهار گام زیر است:

۱- تولید جمعیت اولیه، ترکیبی تصادفی از توابع.(برنامه های رایانه ای)

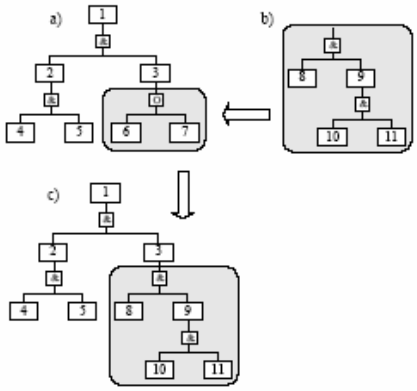
۲- اجرای هر برنامه رایانه‌ای در جمعیت و تعیین ارزش هماهنگی آن در حل مساله.

۳- ایجاد جمعیتی جدید با استفاده از برنامهٔ رایانه‌ای

۳-۱ نسخه‌برداری از بهترین برنامه‌های موجود

۳-۲ ایجاد برنامه‌های رایانه‌ای بوسیله تغییر

۳-۳ ایجاد برنامه‌های جدید از طریق تقاطع برنامه‌ها



۴- بهترین برنامهٔ رایانه‌ای در هر بار تولید و ارایهٔ راه‌حل، به‌عنوان نتیجه برنامه‌ریزی ژنتیک در نظر گرفته می‌شود (Koza 1992, Benzhaf et al, 1998).

وجه تمایز برنامه‌ریزی ژنتیک از سایر الگوریتم‌های تکاملی در استفاده این برنامه‌ریزی از مدل‌های سلسله مراتبی(درختی) است. انعطاف‌پذیری در این برنامه‌ریزی بسیار مهم است. زیرا زیرساخت داده به صورت خودکار کشف می‌شود. اما یکی از مشکلات اولیه این برنامه‌ریزی، وجود تعداد راه‌حل‌هایی است که بدون هیچ بهبودی مرتباًتکرار می‌شوند. (Kusiak, Andrew, 2000)

۴ـ نقش الگوریتم‌های تکاملی در داده‌کاوی

از آنجا که در این مقاله داده‌کاوی فرایندی چند مرحله‌ای در نظر گرفته شده است، بررسی نقش الگوریتم‌های تکاملی در هر یک از مراحل این فرایند ضروری است. (H. A. Abbass et al., 2001)

1. Genetic Programming

۱ـ ۴ـ الگوریتم تکاملی در استخراج ترکیب[1]

فرایند استخراج ترکیب داده که مربوط به مسأله است در داده‌کاوی امری است بسیار مشکل و وابسته به داده. در بعضی از انواع داده، ترکیب به آسانی قابل تشخیص است. از جمله در دادهٔ مربوط به متن که داده، همان کلمات متن هستند. اما زمانیکه بعضی از انواع داده توجه خود را صرفا معطوف به استخراج می‌کنند این وظیفه دشوار خواهدبود. به‌ویژه در جایی‌که استخراج ترکیب داده بیش از اندازه چالش برانگیز باشد. داده تصویر[2] مثال روشنی از این نوع است. در گذشته، دادهٔ تصویر تنها محدود به رشته‌هایی مانند ستاره‌شناسی بود. با این حال امروزه این نوع داده در زمینه‌های مختلفی به کار می‌رود. مانند تصویربرداری پزشکی[3]، تصاویر چند رسانه‌ای در شبکه و تصویرهای ویدیوئی.(Michalewicz. Z, 1996)

علاوه بر الگوریتم‌های ژنتیک، محققانی چند نیز به کاربرد برنامه‌ریزی ژنتیک در امر پردازش تصویر اشاره نموده‌اند. پلی[4] در ۱۹۹۶ نشان داد که چگونه برنامه‌ریزی ژنتیک فیلتر های مؤثری را برای تصاویر پزشکی پیدا می‌کند.

۲ـ ۴ـ الگوریتم تکاملی در انتخاب ترکیب[5]

معمولاً به دنبال استخراج داده، مجموعه‌ای از ترکیبات نیز تعیین می‌شوند. در بسیاری از موقعیت‌ها، شناخت اولویت مربوط بودن شکل استخراج شده از داده به مسأله، امری دشوار است. ترکیب‌های نامربوط نه تنها پیچیدگی زمانی الگوریتم‌ها را زیاد می‌کنند، زمان مورد نیاز استخراج خود ترکیب را نیز افزایش می‌دهند. اغلب روش تکاملی به‌کار رفته برای انتخاب ترکیب‌ها توسط الگوریتم یادگیری صورت می‌گیرد. در این روش هنگام ارزیابی الگوریتم یادگیری با توجه به محاسبات الگوریتمی، هماهنگی ترکیب زیرمجموعه‌ها نیز بدست می‌آید.

۳ـ ۴ـ الگوریتم تکاملی در طبقه‌بندی[6]

الگوریتم‌های تکاملی با الگوریتم‌های طبقه‌بندی مانند سامانه‌های مبتنی بر قانون[7] و درخت تصمیم‌گیری ارتباط مستقیم دارند.

۱ـ ۳ـ ۴ـ سامانه‌های مبتنی بر قانون

در یادگیری ماشینی مفاهیمی مانند مجموعه قوانین بسیار متداول است. زیرا در میان سایر روش ها، قوانین به سادگی ارائه می شوند و انسان نیز می تواند به آسانی آنها را تفسیر کند. در الگوریتم‌های تکاملی دو شیوه اصلی برای ارائه مجموعه‌های قوانین وجود دارد. در روش میشیگان[8] (Holland, 1975) , (Goldberg and Holland, 1989) (Booker, هر فرد جمعیت ارائه کننده قانونی ثابت است و جمعیت کل، دربردارنده مفهوم هدف است. در مقابل در روش پیتسبرگ[9] (Smith, 1980, 1983; DeJong, Spears, Gordon, 1993) هر فرد با متغیر رده بندی شده اش، ارائه کننده مجموعه‌ای کامل از قوانین خواهد بود.

حلقه اصلی در یک سامانه طبقه‌بندی، سامانه‌ای است که با ورودی‌هایی از محیط آغاز به کار می‌کند و خروجی‌ها به صورت پیام‌هایی به لیست پیام‌های قبلی اضافه و شروع به‌کار می‌کنند و قوی‌ترین قانونِ هماهنگ با پیام ها استخراج می‌شود.

سامانه‌های طبقه‌بندی در جایی‌که دانش کافی یا تخصصی برای کنترل مرسوم وجود نداشته باشد، به مثابه سامانه‌های کنترل برای محیط‌های نامطمئن استفاده می‌شوند. (Goldberg, 1989)

---

1. Feature extraction
2. Image data
3. Medical imaging
4. Poly
5. Feature selection
6. Classification
7. Rule-based systems
8. Michigan Approach
9. Pittsburgh” approach

۲ـ ۳ـ ۴ـ درخت تصمیم گیری و الگوریتم‌های تکاملی

درخت تصمیم‌گیری روشی مرسوم در طبقه بندی است. زیرا به آسانی توسط چندین گره، ساخته می‌شود و متخصصان به راحتی می‌توانند آنها را تفسیر کنند. گره‌های داخلی بیانگر آزمون‌هایی بر اساس ترکیب‌ها هستند و به توصیف داده می‌پردازند. گره‌های برگ‌ها بیانگر عنوان طبقه‌ها هستند.

مسیر گره ریشه به یکی از برگها بیانگر پیوستگی آزمون‌ها است. برنامه‌ریزی ژنتیک در ارائه راه حل از درخت تصمیم استفاده می‌کند. کزا[1] (۱۹۹۲)، اولین مثال کاربرد برنامه‌ریزی ژنتیک را در طبقه‌بندی ارائه نمود. بدین‌صورت که هماهنگی هر درخت تصمیم مبتنی بر صحت یک مجموعه آموزش بود.(H. A. Abbass et al., 2001)

۴ـ ۴ـ الگوریتم تکاملی در دسته‌بندی

در این رابطه می‌توان بین دو روش اصلی استفاده از الگوریتم‌های تکاملی در مسائل دسته‌بندی وجه تمایز قایل شد. در روش اول، هر موقعیت در کروموزوم‌ها، آیتمی را در مجموعه آموزش ارائه می‌کند. اگر شماره دسته‌ها (K) اولویت در نظر گرفته شوند، هر موقعیت در کروموزوم‌ها می‌تواند ارزشی به صورت [K,1] داشته باشد. این روش مشابه کدگذاری مستقیم شبکه‌های طبیعی است. این روش در اجرا ساده است زیرا نیازی به اپراتور تکاملی با تخصص ویژه نیست. اما مشکلاتی نیز دارد. از جمله این‌که اندازه افراد دقیقا اندازه مجموعه آموزش است و برای مشکلات بزرگ، این انتخاب کاربردی نیست. کاربرد دیگر الگوریتم‌های تکاملی در دسته‌بندی ، شناسایی مراکز ثقل هر یک از دسته ها است. (H. A. Abbass et al., 2001)

۵ـ نتایج

اگر داده‌کاوی را فرایند کشف و تحلیل حجمی بزرگ از داده‌ها با استفاده از روش‌های آماری و ریاضی بدانیم، از آنجا که در بیشتر موارد، داده‌های حاصل از سیستم‌های بزرگ و پیچیده، الگوی مشخصی ندارند و در طی زمان و مکان تغییر می‌کند و نیز با نظر به سرعت افزایش حجم داده و تنوع پایگاه‌های داده و نیز تغییرات سریع حاصل از عوامل اثر گذار بر ماهیت پدیده، بیش از پیش ضرورت روشی تحلیلی را در شناسایی پدیده مشخص می‌شود. زیرا با فرض اینکه معنای هر نشانه واقعیتی پنهان است و تنها با تفسیر برملا و آشکار می‌گردد؛ می‌توان گفت نشانه به‌وسیله تفسیر به سخن می‌آید و آنچه در درون خود دارد بیرون می‌ریزد و هیچ تفسیری فارغ از داده، صورت نخواهد پذیرفت.

از طرفی، داده‌کاوی بر اساس روش‌های آماری کلاسیک، هم زمان قابل توجّهی را صرف می‌کند و هم مسبوق به نظریّه است. برای یافتن روشی که با عدم توانایی روش‌های آماری ساده در تشخیص روابط چندجانبه و غیر خطی میان پدیده‌ها و عدم هماهنگی میان تحلیل‌های حاصل از روش‌های آماری ساده و ادارک انسان مواجهه کند، می‌توان به نتایج زیر توجه نمود.
اگر قدرت تشخیص و مؤثر بودن تصمیم انسان را تابعی از فرایندی معناشناسیک بدانیم که مبتنی بر شیوه دیدن و برداشت وی از پدیده‌ها است، نوعی تنوع و انحصاری در برداشت بوجودخواهد آمد که تنها فرایندی مانند فرایند داده‌کاوی با قابلیت تحلیلی زیاد می‌تواند اطلاعاتی قابل درک برای انسان ارائه نماید.

بعلاوه، در این مقاله با بررسی الگوریتم‌های تکاملی و توجه به ویژگی‌های ساز و کار ژنتیک و مقایسه هم‌گونی فرایند این الگوریتم‌ها با شیوه تصمیم‌گیری انسان، قانون حاکم اگر(موقعیتی با وضعیتی راضی کننده باشد) آنگاه(ارزشهایی برای صفتی خاص پیش بینی می شود)، شناسایی شد و معلوم گردید این دو فرایند داده‌کاوی و الگوریتم تکاملی یکدیگر را پشتیبانی می‌نمایند.

در پایان، اگرچه لزوم استفاده از داده‌کاوی و تناسب الگوریتم تکاملی به تفصیل بررسی شد، کابرد الگوریتم تکاملی در پردازش مجدد داده‌های حاصل از تصمیم‌های مبتنی بر دانش حاصل شده و تعیین نقش و میزان اثر شیوه تصمیم‌گیری انسان در تغییر دانش و تبدیل آن به داده جدید توصیه می‌شود.

1. Koza

منابع

1. Abbass H. A., R.A. Sarker and C. S Newton, 2001.ON THE USE OF EVOLUTIONARY ALGORITHMS IN DATA MINING. “Data Mining: A Heuristic Approach".
2. Banzhaf W., P. Nordin , RE. Keller and FD. Francone. 1998. Genetic Programming ~ anIntroduction: On the Automatic Evolution of Computer Programs and ItsApplications. Morgan Kaufmann,.
3. Burl, M., L. Asker, P. Smyth, U. Fayyad, P. Perona,L. Crumpler and J. Aubele. 1998. Learning to recognize volcanoes on Venus. Machine Learning, 30, 165-195.

Fayyad, U., G. Piatetsky-Shapiro, P. Adriaans, P. Smyth and R. Uthurusamy. 1996. Advances inknowledge discovery and data mining. Menlo Park, CA: AAAI Press/ the MIT Press.

4. Fisher DH. 1987. Knowledge acquisition via incremental conceptual clustering. Machine Learning. 2. 139-172.
5. Fisher, D. ,G. Hapanyengwi. 1993. Database management and analysis tools of machine induction. Journal of Intelligent Information Systems, 2(1), Mar., 5-38.
6. Falkenauer E. 1998. Genetic Algorithms and Grouping Problems. John Wiley &Sons,
7. Freitas, AA. 1999. On Rule Interestingness Measures. Knowledge-Based Systems 12(5-6), 309-315. Oct.
8. Freitas, AA. 2000. Understanding the crucial differences between classification and discovery of association rules - a position paper. To appear in ACM SIGKDD Explorations, 2
9. Freitas, AA. 2000. A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery. Postgraduate. Brazil. http://www.ppgia.pucpr.br/~alex.
10. Goldberg, D. E. 1983. Computer-aided gas pipeline operation using genetic algorithms and rule learning. Dissertation Abstracts International. 44 (10). 3174B. Doctoral dissertation, University of Michigan.
11. Goldberg, D.E. 1989. Genetic Algorithms in Search, Optimization and MachineLearning. Addison-Wesley,

Gopalakrishnan and A. Gunasekaran (Eds). 2000. Proceedings of the SPIE Conference on Intelligent Systems and Advanced Manufacturing SPIE, Vol. 4192, Boston, MA. pp. 1-10.

12. Koza, JR. 1992. Genetic Programming: on the Programming of Computers by Means of Natural Selection. MIT Press.
13. Kusiak **,** Andrew. 2000 Evolutionary Computation and Data Mining of the SPIE Conference on Intelligent Systems and Advanced Manufacturing, B.Gopalakrishnan and A. Gunasekaran (Eds), SPIE, Vol. 4192, Boston, MA, November, pp. 1-10.
14. Langley, P. and H. A. Simon. 1995. Applications of machine learning and rule induction.Communications of the ACM. 38 (11). 55-64.
15. Michalewicz Z. 1996. Genetic Algorithms + Data Structures = Evolution Programs. 3rd Ed. Springer-Verlag.
16. Michie D., D.J. Spiegelhalter and C.C. Taylor. 1994. Machine Learning, Neural and Statistical Classification. New York: Ellis Horwood.
17. Park, Y. and M. Song. 1998. A genetic algorithm for clustering problems. Genetic Programming 1998: Proc. 3rd Annual Conf. 568-575.
18. Patrick Dillon, 1998 Data Mining: Transforming Business Data Into Competitive Advantage and Intellectual Capital. AtlantaGA: The Information Management Forum. pp. 5-6.
19. Pieter, Adriaans. Dolf Zantinge, 1996. Data Mining (New York: Addison Wesley)
20. Seifert. Jeffrey W. 2004 .Data mining: An Overview. Congressional Research Service.
21. Smith, S. F. 1983. Flexible learning of problem solving heuristics through adaptive search. In Proceedings of the 8th International Joint Conference on Artificial Intelligence. pp. 422-425.