

بررسی داده‌های دانشجویی با استفاده از روش‌های داده‌کاوی

سید مجتبی سجادی^۱ و ابراهیم رضانی^۲

^۱ دانشجوی دانشگاه آزاد اسلامی واحد مشهد مقطع کارشناسی ارشد، s_mjtb@yahoo.com

^۲ دانشجوی دانشگاه پیام نور مرکز ری مقطع کارشناسی ارشد، r_ebi2008@yahoo.com

چکیده: روش‌های داده‌کاوی کمک فوق‌العاده‌ای به محققان برای استخراج دانش و اطلاعات پنهانی که در داده‌های استفاده شده توسط محققان به ارث برده شده است، دارد. این یک روش پردازش استخراج الگوهای معتبر، نوین، موثر و قابل فهم از پایگاه داده است. در این مقاله عملیات مختلف داده‌کاوی از قبیل خوشه‌بندی، کشف قوانین وابستگی، تشخیص ناهمگونی و رده‌بندی را بر روی یک مجموعه داده آموزشی بکار می‌بریم و نشان می‌دهیم چگونه می‌توان از نتایج این عملیات در جهت شناخت بهتر وضعیت تحصیلی دانشجویان و پیش‌بینی نمرات نهایی آنها استفاده کرد. همچنین می‌توان اطلاعات بدست آمده از عملیات داده‌کاوی را برای افزایش کارایی دانشجویان و نیز ایجاد بهبود و تغییرات در برنامه‌های آموزشی آنها، ارزیابی و اصلاح سیستم‌های آموزش الکترونیکی و موارد از این قبیل استفاده کرد.

واژگان کلیدی: تحلیل رفتار آموزشی، داده‌کاوی آموزشی، خوشه‌بندی، طبقه‌بندی.

تکنیک نوظهور داده‌کاوی است که می‌تواند اطلاعات مربوط به حوزه تحصیلات را اعمال کند. این روش از تکنیک‌های زیادی مانند درخت‌های تصمیم، شبکه‌های عصبی، K-همسایه نزدیک و بسیاری دیگر استفاده می‌کند. بکارگیری این تکنیک می‌تواند انواع مختلف دانش را با استفاده از قوانین انجمنی، طبقه‌بندی و خوشه‌بندی کشف کند. داده‌های جمع‌آوری شده در مورد دانشجویان می‌تواند شخصی یا آموزشی باشد که از طریق دفاتر و پایگاه داده‌های موجود در مدارس یا دانشکده‌ها جمع‌آوری می‌شوند. این نوع داده‌ها همچنین از طریق سیستم‌های آموزش الکترونیکی قابل دستیابی هستند. با بکارگیری تکنیک‌های داده‌کاوی روی داده‌های آموزشی می‌توان اطلاعات و دانش مفیدی را از آنها استخراج کرد که این دانش نیز به نوبه خود

۱- مقدمه

داده‌کاوی نوعی مرتب‌سازی به روشی است که در واقع برای استخراج الگوهای پنهان از پایگاه داده‌های بزرگ مورد استفاده قرار می‌گیرد. مفاهیم و روش‌های داده‌کاوی را می‌توان در زمینه‌های مختلف مانند بازاریابی، پزشکی، املاک و مستغلات، مدیریت ارتباط با مشتری، مهندسی، وب‌کاوی، و غیره استفاده کرد. در سال‌های اخیر تحقیقات زیادی در زمینه بکارگیری فرآیند داده‌کاوی در امر آموزش صورت می‌گیرد. این زمینه تحقیقاتی جدید، داده‌کاوی آموزشی نامیده می‌شود که به امر توسعه روش‌های کشف دانش از داده‌های محیط‌های آموزشی خصوصاً دانشجویان می‌پردازد. به واقع داده‌کاوی آموزشی یک

گروه‌های متمایز یا کلاس اشیا نیز استفاده شود، ولی آنها را هزینه‌بر می‌کند بنابراین خوشه‌بندی می‌تواند به عنوان روش پیش پردازش برای انتخاب زیر مجموعه ویژگی و طبقه‌بندی مورد استفاده قرار گیرد.

اسنادی: تکنیک‌های رگرسیون می‌تواند برای اسناد تطبیق

داده شود. آنالیز رگرسیون می‌تواند برای مدل کردن ارتباط بین یک یا چند متغیر مستقل و وابسته مورد استفاده قرار گیرد. در داده‌کاوی متغیرهای مستقل ویژگی‌ها شناخته شده هستند و متغیرهای وابسته همان چیزی است که ما می‌خواهیم پیش‌بینی کنیم. متاسفانه بیشتر مشکلات دنیا واقعی به سادگی قابل پیش‌بینی نیستند. بنابراین بیشتر تکنیک‌های پیچیده (رگرسیون منطقی، درختان تصمیم‌گیری یا شبکه‌های عصبی) ممکن است برای پیش‌بینی مقادیر بعدی لازم شود. انواع مدل‌های مشابه اغلب می‌تواند هم برای رگرسیون و هم برای طبقه‌بندی استفاده شود. برای مثال CART، الگوریتم درخت تصمیم‌گیری می‌تواند هم برای ایجاد طبقه‌بندی درختان (به منظور طبقه‌بندی پاسخ متغیرهای قطعی) و هم رگرسیون درختان (برای پیش‌بینی متغیرهای وابسته مداوم) مورد استفاده قرار گیرد. شبکه‌های عصبی همچنین می‌توانند هم طبقه‌بندی و هم رگرسیون مدل‌ها را ایجاد کنند.

قوانین انجمن: انجمن‌ها و ارتباطات معمولا برای پیدا

کردن مجموعه یافته‌های مکرر در میان مجموعه داده‌های بزرگ وجود می‌آیند. این نوع یافته‌ها به کسب و کار برای تصمیم‌گیری‌های خاص مانند طراحی کاتالوگ، عبور بازاریابی و تجزیه و تحلیل رفتار مشتری کمک می‌کند. الگوریتم‌های قوانین انجمن باید قادر به تولید قوانین با ارزش اطمینان کمتر از یک باشد. با این حال تعداد قوانین انجمن حاصل برای یک مجموعه داده‌ی داده شده ممکن است بسیار زیاد بوده و درصد بالایی از آن قوانین کم ارزش باشند.

شبکه‌های عصبی: شبکه‌های عصبی مجموعه‌ای از

واحدهای ورودی، خروجی متصل هستند و هر اتصال دارای وزن موجود با آن است. در طول مرحله یادگیری شبکه یاد می‌گیرد

می‌تواند برای درک و فهم رفتار دانشجویان، کمک در امر آموزش و تدریس، ارزیابی و بهبود سیستم‌های آموزش الکترونیکی، بهبود برنامه آموزشی، افزایش بازدهی و کارایی دانشجویان و اهداف دیگری بکار گرفته شود.

۲- عملیات داده‌کاوی در سیستم‌های آموزشی

تکنیک‌ها و الگوریتم‌های مختلفی از جمله طبقه‌بندی، خوشه‌بندی، رگرسیون، هوش مصنوعی، قوانین انجمن، درخت‌های تصمیم‌گیری، ژنتیک، الگوریتم روش نزدیک‌ترین همسایه و غیره برای کشف دانش از پایگاه داده استفاده می‌کنند. این تکنیک‌ها و روش‌ها در داده‌کاوی برای بهتر درک کردن نیاز به اشاره مختصری دارد.

طبقه‌بندی: طبقه‌بندی رایج‌ترین تکنیک داده‌کاوی مورد

استفاده است که مجموعه‌ای از نمونه‌های پیش طبقه‌بندی شده را بکار می‌گیرد تا مدلی را که می‌تواند سوابق جمعیتی را در حجم بالا طبقه‌بندی کند، توسعه دهد. این روش اغلب از درخت‌های تصمیم‌گیری و یا شبکه عصبی مبتنی بر الگوریتم‌های طبقه‌بندی استفاده می‌کند. فرآیند داده‌های طبقه‌بندی شده شامل یادگیری و طبقه‌بندی است. در یادگیری آموزش داده‌ها به وسیله الگوریتم طبقه‌بندی شده آنالیز می‌شوند. در آزمون طبقه‌بندی داده‌ها به منظور برآورد دقت قوانین طبقه‌بندی استفاده شده است. اگر دقت و صحت قابل قبول باشد، قوانین می‌تواند به داده‌های چند تایی جدید اعمال شوند. الگوریتم طبقه‌بندی آموزشی، هر یک از این نمونه‌های طبقه‌بندی را به منظور تعیین مجموعه پارامترهای مورد نیاز برای تبعیض مناسب استفاده می‌کند. سپس این پارامترها را به یک مدل بنام طبقه‌بندی کدگذاری می‌کند.

خوشه‌بندی: به شناسایی کلاس‌های مشابه از اشیا

خوشه‌بندی گفته می‌شود. با استفاده از تکنیک‌های خوشه‌بندی بیشتر می‌توان مناطق متراکم و پراکنده در فضای جسم را شناسایی کرده و الگوی توزیع کلی و ارتباط بین ویژگی‌های داده‌ها را کشف کنیم. روش طبقه‌بندی می‌تواند در ابزارهای موثر

خصایص ممکن است خود شامل یک یا چند خصیصه دیگر نیز باشند. بخاطر اینکه ما دنبال خصایصی می‌گردیم که نمره نهایی دانشجو را مشخص و تعیین می‌کنند تالی این قوانین $x =$ نمره نهایی می‌باشد که x یک مقدار برای نمره نهایی مثل خوب، مردود و ... می‌باشد.

جدول (۱) قوانین نمونه‌ای را نشان می‌دهد که برای نمره نهایی "خوب" دانشجویان استخراج گردیده‌اند. این قوانین بر اساس مقدار اطمینان مرتب گردیده‌اند. مقدار اطمینان، کسری از موارد است که وقتی مقدم قانون در آنها ظاهر شده است، تالی نیز در آنها وجود دارد. به عنوان مثال اولین قانون جدول (۱) با مقدار اطمینان ۰.۷۲۱ به این معنی است که یک ارتباط مستقیم قوی ما بین مقدم: حضور خوب، انجام تکالیف و نمره میان ترم خوب با تالی نمره نهایی خوب برقرار است. همچنین این نتایج نشان می‌دهند که حضور در کلاس تأثیر مهمی در نمره نهایی دانشجویان دارد.

۳-۲- طبقه‌بندی

طبقه‌بندی یکی از عملیات موجود در داده‌کاوی است که عضویت یک نمونه داده را در یکی از گروه‌های از قبل مشخص شده پیش‌بینی می‌کند. در داده‌کاوی آموزشی ممکن است بخواهیم با در نظر گرفتن مقدار تلاش یک دانشجو نمره نهایی او را پیش‌بینی کنیم. در این تحقیق ما درخت FT را برای طبقه‌بندی استفاده کرده‌ایم. درخت مزبور خیلی بزرگ است. جدول (۲) بعضی از قوانین درخت حاصل را برای داده‌های تحقیق نشان می‌دهد. با استفاده از طبقه‌بندی می‌توان به عنوان مثال دانشجویانی را که احتمال مردود شدن آنها در امتحان پایان ترم وجود دارد قبلاً تشخیص داد و روی آنها کار کرد و جلوی این امر را گرفت. باید توجه داشت که طبقه‌بندی متفاوت از قوانین وابستگی می‌باشد. قوانین وابستگی شرایط موجود را نشان می‌دهند در حالیکه طبقه‌بندی شرایط و وضعیت آینده را پیش‌بینی می‌کند.

که به وسیله تنظیم وزن، قادر به پیش‌بینی برچسب صحیح کلاس از ورودی‌های چندگانه باشد. شبکه‌های عصبی دارای توانایی قابل توجه استنتاج معانی از داده‌های پیچیده و مبهم هستند و می‌توانند برای استخراج الگوها و شناسایی روندهایی که هم برای بشر و هم سایر تکنیک‌های کامپیوتری بسیار پیچیده است، مورد استفاده قرار گیرند. این روش برای ورودی‌ها و خروجی‌هایی که ارزش مداوم دارند، بسیار مناسب است. شبکه‌های عصبی در شناسایی الگوها یا روندهای موجود در داده‌ها بهترین هستند و به خوبی برای پیش‌بینی یا پیش‌بینی نیازها مناسب‌اند.

درختان تصمیم‌گیری: درخت تصمیم‌گیری ساختاری

درخت گونه است که نشان دهنده‌ی مجموعه‌ای از تصمیم‌گیری‌ها است. این قوانین تصمیماتی را برای طبقه‌بندی مجموعه‌ی داده‌ها تولید می‌کنند. روش‌های خاص درخت تصمیم‌گیری شامل طبقه‌بندی و رگرسیون درختان (CART) و تشخیص تعامل خودکار (CHAID) Chi Square است.

روش نزدیک‌ترین همسایه: تکنیکی است که در آن هر

رکورد در یک مجموعه داده بر اساس ترکیبی از کلاس‌های رکوردهای K تایی مشابه به هم در یک مجموعه داده قدیمی، طبقه‌بندی می‌شود (جایی که K بزرگتر یا مساوی ۱ است). گاهی اوقات تکنیک K همسایه‌ی نزدیک نامیده می‌شود.

۳-۲- عملیات داده‌کاوی در سیستم‌های آموزشی

در این بخش نتایج حاصل از بکارگیری برخی تکنیک‌های داده‌کاوی بر روی داده‌های یک تحقیق را نشان می‌دهیم.

۳-۱- قوانین انجمن

این قوانین ارتباطات جالب و پنهان را مابین خصایص یک مجموعه داده نشان می‌دهند. آنها امکان پیدا کردن روابطی شبیه اگر---مقدم---در اینصورت---تالی--- را می‌دهند که مقدم و تالی خصایصی از مجموعه داده هستند. البته هر یک از این

۳-۳ - خوشه‌بندی

خوشه‌بندی یعنی یافتن گروه‌هایی از اشیا بطوریکه اشیا موجود در یک گروه تا حد امکان شبیه هم باشند و با اشیا موجود در سایر گروه‌ها متفاوت باشند. با استفاده از این تکنیک می‌توان دانشجویان را بر اساس ویژگی‌های مختلفی به عنوان مثال میزان فعال بودن آنها گروه‌بندی کرد و با هر گروه از آنها بصورت مناسبی رفتار کرد. در این تحقیق از الگوریتم EM برای خوشه‌بندی دانشجویان استفاده کرده‌ایم بدین صورت که دانشجویان را به ۵ گروه مختلف جهت مدیریت ارتباط و رفتار با آنها تقسیم کرده‌ایم که جدول (۳) مقدار میانگین خوشه‌ها را برای هر یک از خصیصه‌ها نشان می‌دهد.

۳-۴ - تشخیص ناهمگونی

این تکنیک داده‌ها را جستجو می‌کند تا آنهایی را که خیلی متفاوت از بقیه هستند پیدا کند. در داده‌کاوی آموزشی می‌توان از این تکنیک در جهت پیدا کردن دانشجویانی که مشکلات خاصی از قبیل یادگیری دارند استفاده نمود. در مجموعه داده‌های این تحقیق ۳۶ مورد انحراف و ناهمگونی تشخیص داده شده‌اند. جدول (۴) این ناهمگونی‌ها و خصیصی را که انحراف در آنها اتفاق افتاده است را نشان می‌دهد. با تشخیص این موارد مدرس می‌تواند راهکار مناسبی را در جهت رفع آنها بکار گیرد.

جدول (۱) قوانین انجمن داده‌های دانشجویان

اطمینان	مقدم
۰.۷۲۱	حضور = خوب میان ترم = خوب انجام تکالیف = بله
۰.۶۷۳	حضور = خوب میان ترم = خوب پروژه = خوب
۰.۶۷۵	حضور = خوب انجام تکالیف = بله پروژه = خوب
۰.۵۹۴	انجام تکالیف = بله میان ترم = خوب پروژه = خوب
۰.۴۳۹	حضور = خوب انجام تکالیف = بله پروژه = خوب
۰.۴۳۰	پروژه = خوب میان ترم = خوب حضور = خوب
۰.۴۲۵	حضور = خوب میان ترم = خوب پروژه = خوب
۰.۴۱۱	پروژه = خوب حضور = خوب انجام تکالیف = بله

جدول (۲) قوانین درخت تصمیم حاصل از طبقه‌بندی

$If\ midterm = good\ and\ attendance = good\ then\ final\ grade = good$
$If\ midterm = fail\ and\ project = fail\ then\ final\ grade = fail$
$If\ midterm = weak\ and\ homework = weak\ then\ final\ grade = fail$
$If\ attendance = good\ and\ homework = good\ then\ final\ grade = good$
$If\ project = excellent\ and\ midterm = good\ then\ final\ grade = excellent$
$If\ project = good\ and\ homework = good\ then\ final\ grade = good$

جدول (۳) قوانین درخت تصمیم حاصل از طبقه‌بندی

خوشه ۵	خوشه ۴	خوشه ۳	خوشه ۲	خوشه ۱	خصیصه
۱۰/۱۵	۱۱/۱۵	۱۲/۱۵	۱۳/۱۵	۱۵/۱۵	حضور
۱/۳	۱.۲۵/۳	۱.۵/۳	۲/۳	۲.۲۵/۳	تکالیف
۱.۲۵/۳	۱.۵/۳	۱.۵/۳	۲.۵/۳	۳/۳	پروژه
۱.۵/۴	۲/۴	۲.۵/۴	۳/۴	۳.۷۵/۴	میان‌ترم
۸/۲۰	۱۰/۲۰	۱۲/۲۰	۱۶/۲۰	۱۹/۲۰	نمره نهایی

جدول (۴) داده‌های ناهمگون

مقدار	تعداد	متغیر
۲/۲۰	۳	نمره نهایی
۳/۲۰	۴	نمره نهایی
۲/۱۵	۲	حضور
۰/۳	۱۰	پروژه
۰.۵/۳	۹	تکلیف
۰.۲۵/۴	۸	میان‌ترم

اطلاعاتی دانشجویان جهت پیش‌بینی تقسیم دانشجویان بر اساس بانک اطلاعاتی قبلی مورد استفاده قرار می‌گیرد. رویکردهای زیادی برای خوشه‌بندی داده‌ها استفاده می‌شوند، که در اینجا با استخراج اطلاعات و دانش نهفته در این مجموعه داده راهکارهای مناسبی را برای هر یک از موارد در جهت رفع مشکلات فعلی و افزایش بازدهی دانشجویان و همچنین پیش‌بینی مشکلات احتمالی آینده و جلوگیری از وقوع آنها معرفی کردیم.

اطلاعاتی نظیر حضور در کلاس، آزمون کلاسی، سمینار و نمرات تکالیف از بانک اطلاعاتی قبلی دانشجویان و اساتید جهت بهبود تقسیم دانشجویان کمک خواهد کرد. این تحقیق جهت شناخت آن دسته از دانشجویانی که توجه ویژه نیاز دارند تا نرخ شکستشان کاهش یابد و دادن فعالیت‌های مناسب برای آزمون پایان ترم بعدی نیز کار می‌کند.

۴- نتیجه‌گیری

در این تحقیق جنبه‌های مختلف داده‌کاوی آموزشی مورد بررسی قرار گرفتند. همچنین عمل خوشه‌بندی روی بانک

مراجع

- [1] U. K. Pandey and S. Pal, "Data Mining: A prediction of performer or underperformer using classification", (*IJCSIT*) *International Journal of Computer Science and Information Technology*, Vol. 2(2), pp. 686-690, 2011.
- [2] Shaeela Ayesha and Tasleem Mustafa, "Data mining model for higher education system", *European Journal of Scientific Research*, Vol. 43, No. 1, pp. 24-29, 2010.
- [3] Er.Rimmy Chuchra, "Use of Data Mining Techniques for the Evaluation of Student Performance: A case study", *International Journal Of Computer Science and Management Research*, Vol 1, 2012.
- [4] Baker, R.S.J.d., Barnes, T., Beck, J.E. (Eds.) *Proceedings of the 1st International Conference on Educational Data Mining*, 2008.
- [5] Witten, I.H., Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition, San Francisco: Morgan Kaufmann, 2005.
- [6] Z. N. Khan, "Scholastic achievement of higher secondary students in science stream", *Journal of Social Sciences*, Vol. 1, No. 2, pp. 84-87, 2005.
- [7] Romero, C., Ventura. S., *Educational data mining: A survey from 1995 to 2005*, Expert Systems with Applications 135-146, 2007.

