



عنوان

انتخاب ویژگی

(feature selection)

فهرست مطالب

چکیده ۱

فصل اول: مقدمه

۱-۱- مقدمه ۳

فصل دوم: بستر تحقیق

۱-۲- مقدمه ۵

۲-۲- کشف دانش ۶

۳-۲- داده کاوی ۸

۱-۳-۲- داده و اطلاع ۹

۲-۳-۲- استخراج دانش و داده کاوی ۹

۳-۳-۲- پیش پردازش ها ۱۲

۴-۲- انواع روش های داده کاوی ۱۲

۱-۴-۲- خوشه بندی ۱۳

۲-۴-۲- دسته بندی ۱۴

۳-۴-۲- دسته بندی قوانین "اگر-آنگاه" فازی ۱۴

۵-۲- تعریف ویژگی ۱۶

۶-۲- انتخاب ویژگی ۱۶

۱-۶-۲- وظیفه انتخاب ویژگی ۱۷

۷-۲- انتخاب ویژگی در داده های با ابعاد بالا ۱۸

۸-۲- اهداف انتخاب ویژگی ۱۴

۹-۲- مزیت انتخاب ویژگی ۱۹

۱۰-۲- انواع ویژگی ها ۲۰

۱۱-۲- طبقه بندی الگوریتم های انتخاب ویژگی ۲۰

۱۲-۲- جمع بندی ۲۱

فصل سوم: روش های انتخاب ویژگی

۲۳ ۱-۳-۱ مقدمه
۲۴ ۲-۳-۲ روش های کلی انتخاب ویژگی
۲۵ ۳-۳-۳ مروری جامع بر انتخاب ویژگی برای مجموعه داده های عظیم
۲۷ ۱-۳-۳-۱ انتخاب ویژگی با الگوریتم کلونی مورچه
۲۷ ۲-۳-۳-۲ انتخاب ویژگی با الگوریتم ژنتیک
۲۸ ۳-۳-۳-۳ انتخاب ویژگی با الگوریتم جستجو هارمونی
۲۹ ۴-۳-۳-۴ انتخاب ویژگی با الگوریتم فاخته
۳۰ ۵-۳-۳-۵ انتخاب ویژگی با الگوریتم رقابت استعماری
۳۰ ۶-۳-۳-۶ انتخاب ویژگی با الگوریتم کلونی زنبور مصنوعی
۳۳ ۴-۳-۴ جمع بندی
۳۴	مراجع

فهرست اشکال

- شکل ۱-۲- پروسه کشف دانش ۷
- شکل ۲-۲- فرآیند داده کاوی ۱۰
- شکل ۳-۲- مثال جدول اطلاعاتی برای ویژگی ها ۱۶
- شکل ۱-۳- آرایه ای یک بعدی نشان دهنده ی انتخاب ویژگی ۳۰
- شکل ۲-۳- انتخاب ویژگی با الگوریتم کلونی زنبور مصنوعی ۳۱

فهرست جدول‌ها

جدول ۱-۳- نگاشت مفاهیم الگوریتم هارمونی روی مسالهی انتخاب ویژگی ۲۹

چکیده

مساله انتخاب ویژگی، یکی از مسائلی است که در مبحث یادگیری ماشین و همچنین شناسائی آماری الگو مطرح است. این مساله در بسیاری از کاربردها (مانند طبقه‌بندی) اهمیت به سزائی دارد، زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد، که بسیاری از آنها یا بلااستفاده هستند و یا اینکه بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد. و علاوه بر این باعث می‌شود که اطلاعات غیر مفید زیادی را به همراه داده‌های مفید ذخیره کنیم. برای مساله انتخاب ویژگی، راه حل‌ها و الگوریتم‌های فراوانی ارائه شده است که بعضی از آنها قدمت سی یا چهل ساله دارند. مشکل بعضی از الگوریتم‌ها در زمانی که ارائه شده بودند، بار محاسباتی زیاد آنها بود، اگر چه امروزه با ظهور کامپیوترهای سریع و منابع ذخیره سازی بزرگ این مشکل، به چشم نمی‌آید ولی از طرف دیگر، مجموعه‌های داده‌ای بسیار بزرگ برای مسائل جدید باعث شده است که همچنان پیدا کردن یک الگوریتم سریع برای این کار مهم باشد. در این سمینار به مطالعه و بررسی روش‌های انتخاب ویژگی در سال‌های اخیر ارائه شده اند، پرداخته می‌شود.

کلمات کلیدی: انتخاب ویژگی، اطلاعات، روش پوششی، فیلتر، پیش پردازش.

فصل اول

مقدمه

متناسب با پیشرفت سریع تکنولوژی کامپیوترها و پایگاه داده‌ها، داده‌ها با سرعتی بسیار بیشتر از ظرفیت پردازشی انسان در حال انباشته‌سازی هستند. این مجموعه داده‌ها با ابعاد گسترده شامل ویژگی‌های بسیاری می‌باشند که کارایی ابزارهای داده کای را بشدت کاهش می‌دهند. داده کاوی تلاش چندجانبه‌ی منظمی برای بیرون کشیدن قطعه‌های دانش از داده‌هاست. تکثیر مجموعه داده‌های عظیم از دامنه‌های کاری متفاوت، چالش‌های اساسی و مهمی را در روند داده کاوی مطرح می‌کند. نه تنها مجموعه‌ی داده‌ها در حال گسترش هستند بلکه انواع جدیدی از داده‌ها مثل جریان داده‌های دریافتی از وب، میکروآرایه‌ها در ساختار ژنوم‌ها و پروتئین‌ها و سیستم‌های زیست شناسی نیز در حال رایج شدن هستند [1]. بر این اساس محققان دریافته‌اند که داشتن روش‌هایی جهت کاهش ویژگی‌ها و انتخاب تنها تعدادی از آنها به عنوان ویژگی‌های برتر و برجسته‌تر و حذف مابقی آنها، به عنوان رویه‌ای پیش پردازشی برای اعمال الگوریتم‌های داده کاوی بسیاری ضروری است. در این خصوص مسئله‌ی انتخاب ویژگی از جمله مسائل بسیار مهم و مطرح است. انتخاب ویژگی فرآیند انتخاب زیرمجموعه‌ای از ویژگی‌های اصلی با توجه به ضوابطی خاص بوده و تکنیکی مهم و پرکاربرد برای کاهش بعد در داده کاوی است. از جمله‌ی این تاثیرات می‌توان زیر را نام برد. افزایش سرعت الگوریتم‌های داده کاوی، بهبود صحت یادگیری، بهبود کارایی پیش بینی و درک بهتر از داده‌ها و مدل‌های یادگیری. بنابراین، پیدا کردن زیر مجموعه‌ای از ویژگی‌ها از یک مجموعه داده عظیم، مسئله‌ای است که در بسیاری از زمینه‌های مطالعاتی پیش می‌آید. از آنجایی که افزایش تعداد ویژگی‌ها هزینه محاسباتی یک سیستم را افزایش می‌دهد، طراحی و پیاده‌سازی سیستم‌ها با کمترین تعداد ویژگی ضروری به نظر می‌رسد. از طرف دیگر توجه به این موضوع بسیار مهم است که، باید زیر مجموعه موثری از ویژگی‌ها انتخاب شود که کارایی قابل قبولی برای سیستم ایجاد کند [2][3]. برای تشخیص اینکه کدام زیر مجموعه ویژگی موثرتر است، یک راه حل بررسی تمام زیر مجموعه‌های ممکن است که بررسی همه زیر مجموعه‌ها جزء مسائل سخت و دارای پیچیدگی محاسباتی بالاست. این موضوع ما را به سمتی هدایت می‌کند که از الگوریتم‌های فراابتکاری، برای پیدا کردن زیرمجموعه‌ای بهینه از ویژگی‌ها استفاده کنیم. در تحقیقات دهه‌ی اخیر، انتخاب ویژگی به یک حوزه‌ی فعال در بحث داده کاوی تبدیل شده است و کاربردهای وسیعی در بسیاری از شاخه‌ها مثل تشخیص بیماری داشته است.