یك روش آماری برای کاوش در داده‌های جریان دنباله‌ای

حیدر قاسم‌زاده* محمد قاسم‌زاده
دانشگاه یزد دانشگاه یزد

چکیده

در این مقاله یك روش آماری مبتنی بر حد هُفدینگ برای ایجاد درخت طبقه‌بندی موسوم به «درخت تصمیم بسیار سریع» در داده‌های جریان دنباله‌ای ارائه می‌گردد. در این روش از یك معیار پیشنهادی براي بخش‌بندی داده‌ها بر اساس صفات بهره می‌بریم. این روش مبتنی بر ترکیب حد هفدینگ و خطای طبقه‌بندی است. در سال‌های اخیر داده‌های جریان دنباله‌ای یك زمینه تحقیقاتی مهم و رو به افزایش در انجمن‌های آماری و علم کامپیوتر به حساب می‌آید. این نوع داده‌ها به شکل دنباله‌های مرتب و بالقوه نامحدود از نقاط داده‌ای هستند که نوعاً توسط فرآیند مولد داده غیر ایستا ایجاد می‌شوند. چنین داده‌هایی به طور پیوسته و در جریان زمان تولید می‌شوند و باید خیلی سریع و به صورت بلادرنگ پردازش شوند. نتایج آزمایشی و تحلیل‌های آماری نشان می‌دهند که بکارگیری روش یاد شده منجر به رسیدن به دقت بالاتر و همچنین حد هُفدینگ بهتری می‌گردد.

واژه‌های کلیدی: داده‌های جریان دنباله‌ای، درخت تصمیم بسیار سریع، اندازه‌گیری ناخالصی، معیار بخش‌بندی

۱ مقدمه

در سال‌های اخیر، حجم داده‌ها برای تجزیه و تحلیل کردن بسیار سریع در حال رشد هستند. بنابراین یک زمینه تحقیقاتی جدید به نام داده‌های جریانی به وجود آمد. محققان روش‌ها و الگوریتم‌ها را برای استخراج دانش از این نوع داده‌ها مورد بررسی قرار می‌دهند[۱] و [۲]. داده‌های جریان دنباله‌ای به شکل دنباله‌های مرتب و بالقوه نامحدود از نقاط داده‌ای تولید می‌شوند. در نتیجه تمام داده‌ها نمی‌توانند ذخیره گردند و تجزیه‌وتحلیل باید به صورت جریان دنباله‌ای از عناصر داده‌ها، بسیار سریع و با یک مرتبه اسکن داده‌ها اتفاق بیفتد. بنابراین انجام این عملیات، روش‌های استاندارد داده‌کاوی نمی‌تواند به کار برده شود. مسئله دیگر، رویداد تغییر مفهوم است[۳] و [۴]. تغییر مفهوم به تغییر در توزیع داده یا تغییر در ساختار داده گفته می‌شود. در این صورت مدل‌های تجزیه‌وتحلیل باید خودشان را بر طبق تغییرات بروز رسانی کنند. در این مقاله، یك روش آماری برای طبقه‌بندی داده‌های جریان دنباله‌ای بر اساس درخت تصمیم بسیار سریع مبتنی بر ترکیب حد هُفدینگ و خطای طبقه‌بندی پیشنهاد می‌شود. نتایج آزمایشی و تحلیل‌های آماری نشان می‌دهند که بکارگیری روش یاد شده منجر به رسیدن به دقت بالاتر و همچنین حد هفدینگ بهتری می‌گردد.

۲ حد هُفدینگ

هدف از طراحی درخت‌های تصمیم‌گیری بسیار سریع آن است که به هر نمونه آموزشی فقط یکبار دسترسی و زمان کوتاه ثابتی برای پردازش آن صرف گردد. بخش‌بندی داده‌های جریان دنباله‌ای که از یک گره در حال عبور هستند بر اساس بهترین صفت انجام می‌شود. بهترین صفت در هر گره مبتنی بر یک زیرمجموعه کوچک از نمونه داده‌ها تشخیص داده می‌شود. با ورود جریانی از نمونه‌ها، اولین نمونه‌ها برای انتخاب بهترین صفت ریشه استفاده می‌شوند. بعد از تعیین شدن صفت ریشه درخت