

عنوان مقاله:

ارائه مدلی ترکیبی جهت خوشه بندی و تشخیص داده های پرت در داده های جریانی

محل انتشار:

ششمین کنفرانس ملی پژوهش های کاربردی در مهندسی برق، مکانیک و مکاترونیک (سال: 1399)

تعداد صفحات اصل مقاله: 14

نویسندگان:

میثم خادم - دانشجوی Ph.D. مهندسی کامپیوتر گرایش نرم افزار، دانشگاه آزاد اسلامی واحد بیرجند، بیرجند، ایران

حمید سعادتفر - استادیار گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه بیرجند، بیرجند، ایران

خلاصه مقاله:

امروزه با توجه به افزایش روزافزون داده های جریانی، خوشه بندی این نوع از داده ها جهت استخراج دانش و تصمیم گیری به موقع تبدیل به یکی از مباحث داغ تحقیقاتی شده است. در این راستا خوشه بندی به عنوان یک رویکرد یادگیری بدون ناظر، میتواند در شناسایی الگو و استخراج دانش از چنین داده هایی موثر واقع شود. خوشه بندی به خودی به خود یک وظیفه چالش برانگیز است، حال با توجه به ماهیت در حال گذر بودن داده های جریانی و محدودیت دسترسی به آنها و همچنین لزوم واکنش سریع و به موقع، این وظیفه با چالش های جدیدی روبروست. از یک سو با توجه به این حقیقت که این نوع از داده ها نامتناهی بوده و به سرعت در حال تغییرند، می بایست خوشه بندی آنها نیز در گذر زمان و به صورت تکاملی انجام شود. از سوی دیگر وجود داده های پرتی که میتوانند دقت یادگیری را تحت تاثیر خود قرار دهند و منجر به تصمیم گیری های نادرست شوند، تشخیص خوشه های درست را قدری دشوار می سازد. در این راستا، الگوریتم های متنوعی در حوزه ی خوشه بندی و همچنین تشخیص داده های پرت در داده های جریانی مطرح شده است. اغلب این الگوریتم ها به صورت آنلاین (برخط) و با رویکردهای مبتنی بر فاصله به تشخیص داده های پرت می پردازند. با توجه به اهمیت موضوع در مقاله جاری سعی شده است تا با رویکردی ترکیبی ضمن خوشه بندی داده ها توسط الگوریتم k-means و تشخیص داده های پرت کاندید به صورت برخط، بار دیگر داده های پرت به صورت غیربرخط (آفلاین) غربالگری شده و در نهایت داده های پرت باقیمانده به عنوان داده های پرت واقعی گزارش شوند. الگوریتم پیشنهادی بر روی مجموعه داده های استاندارد KDD-CUP 99 که دربردارنده ی اطلاعات برچسب خوردهای در ارتباط با وضعیت عادی و وضعیت حمله در شبکه هستند (اعمال شده است). نتایج تحقیقات حاکی از این است که رویکرد پیشنهادی نسبت به رویکردهایی که اخیرا در این حوزه مطرح شده است دارای دقت تشخیص بالاتر بوده و خوشه های بدست آمده دارای خلوص بالاتری هستند. علاوه براین، تحلیلها نشان میدهند روش پیشنهادی از سرعت بالاتری نیز نسبت به رویکردهای مشابه برخوردار است.

کلمات کلیدی:

داده های جریانی، خوشه بندی، k-means، داده های پرت.

لینک ثابت مقاله در پایگاه سیویلیکا:

<https://civilica.com/doc/1160758>

