عنوان مقاله:
روشی برای بهبود شناسایی داده پرت با استفاده از تکنیک نزدیک ترین همسایه

محل انتشار:
چهارمین کنفرانس بین المللی مهندسی دانش بنیان و نوآوری در حوزه مهندسی کامپیوتر و برق (سال: 1396)

تعداد صفحات اصل مقاله: 10

نویسندگان:
بهناز فرضی - *گروه کامپیوتر، واحد بوئین زهرا، دانشگاه آزاد اسلامی بوئین زهرا، ایران*

حسن نادری - *گروه کامپیوتر، واحد بوئین زهرا، دانشگاه آزاد اسلامی بوئین زهرا، ایران*

خلاصه مقاله:
در حال حاضر با افزایش روزافزون داده ها و حجم اطلاعات در مسائل دنیای پیرامون خود روبرو هستیم که چالش روبروی ما، مدل سازی و تجزیه و تحلیل داده هاست و بهره گیری از روشهایی همچون داده کاوی برای استخراج دانش و اطلاعات نهفته در داده ها، جزء مراحل ضروری شناسایی داده ها به شمار می آید. داده های پرت با عدم تطابق با سایر داده ها سبب بروز مشکل در امر تجزیه و تحلیل داده ها میگردند . بنابراین لزوم شناسایی داده های پرت امری اجتناب ناپذیر است. شناسایی داده های پرت و یا خطاها نقش مهمی در کاهش، محدود کردن حجم محاسبات دارند. داده های پرت در بسیاری از علوم کامپیوتری، پزشکی و تجارت کاربرد دارد. مسئله ی دیگری که امروزه در بحث دادهکاوی وجود دارد، بحث کاهش خطا در شناسایی داده ها است. نقش شناسایی داده ی پرت و کاهش خطاها، عامل اصلی مطالعهی تکنیکهای شناسایی داده های پرت و بهبود این تکنیکها است.در این مقاله تکنیکهای شناسایی داده های پرت و معیارهای رده بندی این روشها بیان شدند. از جمله این تکنیکها میتوان به تکنیک مبتنی بر خوشه بندی، تکنیک مبتنی بر همسایگی، تکنیک مبتنی بر چگالی و تکنیک امتیازدهی اشاره کرد. در این مقاله، ما مسئله خود را در سه فاز مورد مطالعه قرار خواهیم داد. در فاز اول، روش k نزدیک ترین همسایه مورد استفاده قرارمیگیرد. در فاز دوم، به کمک الگوریتم علفهای هرز، الگوریتم k-means اجرا میگردد. در فاز سوم، پس از تشکیل خوشه ها با استفاده از روش k-means داده های پرت شناسایی شده، حذف میگردند. بطور کلی دستاوردهای اصلی این تحقیق عبارتند از: (1 ارائه ی روش اکتشافی نزدیک ترین همسایه به منظور دست یابی به بهترین جواب در مسئله. (2 ارائه ی یک روش پیوندی k-means و علف هرز به کمک روش نزدیک ترین همسایه با هدف کاهش خطا در یافتن داده های پرت. نتایج حاصل از آزمایشات انجام شده در این مقاله، نشاندهنده برتری چشمگیری روش پیشنهادی با کاهش میانگین مربعات خطا در مجموعهی داده ها میباشد.

کلمات کلیدی:
شناسایی داده ی پرت مبتنی بر فاصله، شناسایی داده ی پرت مبتنی بر چگالی، شناسایی داده ی پرت مبتنی بر خوشه بندی، خوشه بندی بر مبنای درخت پوشای کمینه، تخمین جستجوی k نزدیک ترین همسایه