روش‌های شناسایی داده‌ی پرت

نویسنده:
بهناز فرضی

سرشناسه	:	فرضی، بهناز، ۱۳۶۹
عنوان و نام پدیدآور	:	روش‌های شناسایی داده‌ی پرت / نویسنده بهناز فرضی
مشخصات نشر	:	تهران، آروَن، ۱۳۹۹
مشخصات ظاهری	:	۹۲ ص.
شابک	:	۹۷۸ - ۹۶۴ - ۲۳۱ - ۹۳۰ - ۵
وضعیت فهرست‌نویسی	:	فیپا
موضوع	:	داده دورافتاده (آمار) – داده کاوی
موضوع	:	Outliers (Statistics) – Data mining
رده‌بندی کنگره	:	QA ۲۷۶
رده‌بندی دیویی	:	۵۱۹/۵
شماره کتابخانه ملی	:	۷۳۶۹۳۱۵

انتشارات آروَن

روش‌های شناسایی داده‌ی پرت

نویسنده: بهناز فرضی

ناشر: انتشارات آروَن

چاپ اول: ۱۳۹۹

چاپ صدف: ۵۰۰ نسخه

۲۷۰۰۰ تومان

نشانی: میدان انقلاب، خیابان ۱۲ فروردین، خیابان وحید نظری
نرسیده به خیابان منیری جاوید، پلاک ۱۰۵، واحد ۳ تلفن: ۶۶۹۶۲۸۵۰ – ۰۲۱
ایمیل: Arvannashr@yahoo.com وبسایت: www.Arvannashr.ir

بسم الله الرحمن الرحیم

فهرست مطالب

چکیده ۷

فصل اول: کلیات کتاب ۹

مقدمه ۹

بیان مسئله ۱۱

پیشینه‌ی کتاب ۱۲

هدف کتاب ۱۴

اهمیت انجام کتاب ۱۴

معیار ارزیابی ۱۶

فرضیه‌های کتاب ۱۶

فرض‌ها و محدودیت‌های کتاب ۱۷

جنبه‌ی نوآوری کتاب ۱۷

ساختار کتاب ۱۸

فصل دوم: پیشینه‌ی کتاب ۱۹

مقدمه ۱۹

تکنیک امتیازدهی ۲۰

فاکتورهای شناسایی داده‌ی پرت ۲۰

طراحی سیستم تشخیص داده‌ی پرت ۲۲

کاربردهای شناسایی داده‌ی پرت ۲۴

تکنیک آماری ۲۵

تکنیک بر مبنای فاصله ۲۵
تکنیک بر مبنای چگالی ۲۶
تکنیک بر مبنای خوشه بندی ۲۷
تکنیک خوشه بندی بر مبنای تفکیک ۲۸
شناسایی داده‌های پرت توسط الگوریتم‌های خوشه بندی ۲۸
روش‌های خوشه بندی سلسله مراتبی ۲۹
روش دسته بندی بر اساس تعداد متغیر ۳۳
تکنیک بر مبنای همسایگی ۳۴
استفاده از روش‌های بهینه سازی ۳۶
الگوریتم [illegible] تکاملی ۳۸
به کارگیری الگوریتم ژنتیک ۴۰
به کارگیری الگوریتم بهینه سازی علف هرز مهاجم ۴۲
الگوریتم بهینه سازی [illegible] ای مورچگان ۴۲
الگوریتم بهینه سازی [illegible] ات ۴۴
مدل بهینه سازی فاری [illegible] ۴۶
الگوریتم تابکاری شبیه سازی [illegible] ۴۷
الگوریتم جستجوی ممنوع ۴۸
الگوریتم اجتماع زنبور عسل ۴۹
تکنیک بر مبنای قانون ۵۱
تکنیک بر مبنای شبکه‌ی عصبی ۵۱
جمع‌بندی ۵۴

فصل سوم: روش پیشنهادی برای شناسایی داده‌های پرت ۵۵
مقدمه ۵۵
رویکرد پیشنهادی این کتاب ۵۵
رویکرد پیشنهادی برای فاز اول کتاب: جستجوی تخمین نزدیک‌ترین همسایه ۵۶
رویکرد پیشنهادی برای فاز دوم کتاب: پیوندزنی روش‌ها ۵۸
الگوریتم تکاملی علف‌های هرز مهاجم ۵۹
روش خوشه بندی k-means ۶۲
روش حذف داده های پرت توسط خوشه بندی ۶۴
رویکرد پیشنهادی برای فاز سوم کتاب: حذف داده‌های پرت ۶۴
معیار ارزیابی الگوریتم ۶۵
عملکرد کلی الگوریتم ۶۶
جمع‌بندی ۶۸

چکیده

در حال حاضر با افزایش روزافزون داده‌ها و حجم اطلاعات در مسائل دنیای پیرامون خود روبرو هستیم که چالش روبروی ما، مدل سازی و تجزیه و تحلیل داده‌هاست و بهره‌گیری از روش‌هایی همچون داده‌کاوی برای استخراج دانش و اطلاعات نهفته در داده‌ها جزء مراحل ضروری شناسایی داده‌ها به شمار می‌آید. داده‌های پرت با عدم تطابق با سایر داده‌ها سبب بروز مشکل در امر تجزیه و تحلیل داده‌ها می‌گردند. بنابراین لزوم شناسایی داده‌های پرت امری اجتناب ناپذیر است. شناسایی داده‌های پرت و یا خطاها نقش مهمی در کاهش، محدود کردن حجم محاسبات دارند. داده‌های پرت در بسیاری از علوم کامپیوتری، پزشکی و تجارت کاربرد دارد. مسئله‌ی دیگری که امروزه در بحث داده‌کاوی وجود دارد، بحث کاهش خطا در شناسایی داده‌ها است. نقش شناسایی داده‌های پرت و کاهش خطاها، عامل اصلی مطالعه‌ی تکنیک‌های شناسایی داده‌های پرت و بهبود این تکنیک‌ها است.

در این کتاب سعی شده است بسیاری از تکنیک‌های شناسایی داده‌های پرت و معیارهای رده بندی این روش‌ها بیان گردد. 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 پایان نامه، ما قصد داریم تا ضمن مطالعه‌ی کارهای موجود در زمینه‌ی روش‌های شناسایی داده‌های پرت و کاربرد این روش‌ها در داده‌کاوی، روش جدیدی را با هدف کاهش خطا برای شناسایی این دادها ارائه دهیم. ما مسئله خود را در سه فاز مورد مطالعه قرار خواهیم داد. در فاز اول، روش k نزدیک‌ترین همسایه مورد استفاده قرارمی‌گیرد. در فاز دوم، به کمک الگوریتم علف‌های هرز، الگوریتم k-means اجرا می‌گردد. در فاز سوم، پس از تشکیل

خوشه‌ها با استفاده از روش k-means داده‌های پرت شناسایی شده، حذف می‌گردند. بطور کلی دستاوردهای اصلی این کتاب عبارتند از: ۱) ارائه‌ی روش اکتشافی نزدیک‌ترین همسایه به منظور دست یابی به بهترین جواب در مسئله. ۲) ارائه‌ی یک روش پیوندی k-means و علف هرز به کمک روش نزدیک‌ترین همسایه با هدف کاهش خطا در یافتن داده‌های پرت. نتایج حاصل از آزمایشات انجام شده در این پایان‌نامه، نشان‌دهنده برتری چشمگیری روش پیشنهادی با کاهش میانگین مربعات خطا در مجموعه‌ی داده‌ها می‌باشد.