



عنوان مقاله ۳۱: کاربرد نمونه‌گیری مجموعه-رتبه‌دار در تحلیل داده‌های بزرگ

نویسنده/نویسندگان: محمد حسین زرین کلاه، هادی جبّاری نوقابی و محمد مهدی صابر

چکیده:

یکی از اساسی‌ترین چالش‌ها در مدل‌سازی و تحلیل داده، برآورد دقیق پارامترهای مدل است. در تحلیل داده‌های بزرگ، به دلیل حجم بالا و پیچیدگی، پردازش کل مجموعه داده‌ها از نظر محاسباتی و زمانی غیرعملی یا بسیار پرهزینه است. در چنین شرایطی، استفاده از نمونه‌های معرف به‌عنوان جایگزینی برای کل داده‌ها، علاوه بر کاهش هزینه‌های محاسباتی و افزایش سرعت آموزش مدل‌ها، به مدیریت محدودیت‌های حافظه و منابع نیز کمک می‌کند. در این مقاله، از روش نمونه‌گیری مجموعه-رتبه‌دار میانه‌ای (Median Ranked-Set Sampling) برای انتخاب نمونه‌های نماینده و برآورد پارامترهای مدل استفاده شده است. برای برآورد پارامترهای مدل در داده‌های درآمدی مرکز آمار ایران، روش‌های نمونه‌گیری تصادفی ساده (Simple Random Sampling) و نمونه‌گیری مجموعه-رتبه‌دار میانه‌ای (MRSS) به کار گرفته شد. نتایج نشان داد که در مواردی که داده‌ها نامتقارن هستند، استفاده از MRSS منجر به برآوردهایی با دقت بیشتر نسبت به SRS می‌شود.

کلمات کلیدی: نمونه‌گیری مجموعه-رتبه‌دار، داده‌های بزرگ، بررسی‌های نمونه‌ای