چکیده:

پاکسازی داده ها از خطاها و نویز یکی از بخش‌های اصلی فرایند نگهداری داده محسوب می‌گردد. پاکسازی به توالی عملیات‌های انجام شـــده با هدف بهبود کیفیت کلی مجموعه های داده‌ها اتلاق می‌شـــود. در واقع پاکســـازی داده کاهش و حذف نســـخه‌های کپی در مجموعه داده می‌باشـــد. مســـئله‌ای در برنامه‌های کاربردی پایگاه داده رخ می‌دهد و زمانی بدتر می‌شود که داده‌های منابع بخواهند با هم تلفیق شوند. از این رو ، پاکسازی داده به عنوان بخش اصلی فرایند یکپارچگی داده شناخته شده است.

هنوز مسائل و چالش‌های زیادی از جمله تصحیح خطا و حل تعارض، نگهداری داده‌های پاکسازی شده و پاکسازی داده در محیط‌های مجازی یکپارچه در زمینه پاکسازی داده وجود دارد. برای حل چالش‌های پیش رو پاکسازی داده توسعه یک چارچوب مناسب نیز احساس می‌شود.

از این رو در این تحقیق یک روش آماده‌سازی و پاکسازی داده مبتنی بر خوشه‌بندی ارایه شده است که در آن ابتدا داده‌های خام را به دلیل اینکه به سرعت به پاک‌سازی و نرمال‌سازی‌های آتی داده کمک میکند، خوشه‌بندی نموده سپس عملیات پاک‌سازی داده‌ها و در مرحله آخر عمل تبدیل داده انجام می‌شود. در این تحقیق، آزمایش‌ها در بدترین حالت میزان نویز، در نظر گرفته شده و با این وضعیت آزمایش پیش برده شده، تا بهترین نتیجه برای روش پیشنهادی بدست آید. به طور کلی روش پیشنهادی یک چارچوب پاکسازی داده جهت بهبود صحت روش‌های داده‌کاوی برروی داده‌های پاکسازی شده را نشان می‌دهد و مزیت اصلی روش ارایه شده این است که نیاز به بروزرسانی در محیط‌های پویا را نداشته و با قراردادن هر داده جدید در دسته مناسب همیشه منبع را به‌روز نگه می‌دارد. در روش پیشنهادی از مجموعه داده تشخیص حروف با 20000داده استفاده شده است که بعد از اعمال گام های روش پیشنهادی بر روی مجموعه داده، دو مجموعه پاکسازی شده با روش پیشنهادی مورد مقایسه قرار گرفته و طبق نتایج بدست آمده، روش پیشنهادی الگوریتم طبقه بندی Kstar را 25 درصد بهبود داده است.

واژه‌های کلیدی: پاکسازی، خوشه بندی، داده، نویز و خطا