



بهبود ترکیب نمونه انتخاب برای انتساب داده های گمشده با کمک الگوریتم های ماشین بردار
پشتیبان وجستجوی گرانشی

بتول نخعی^۱ . روح الله ذوالفقاری^۲

۱- هنرآموز رایانه استان کرمان ناحیه ۲ دانشگاه غیرانتفاعی بهمنیار کرمان، batolnakhae1355@gmail.com

۲- هنرآموز رایانه استان سمنان شهرستان دامغان دانشگاه آزاد اسلامی علوم و تحقیقات واحد سمنان
Rohollah.zolfaghary@gmail.com

خلاصه

گمشده گی داده در تمامی پژوهش های علوم اجتماعی، رفتاری، پزشکی وجود دارد. در آمار، گم شدن داده به وضعیتی گفته می شود که تعدادی از اطلاعات داده ها گزارش نشده باشند. گمشده گی داده باعث کاهش تطابق جامعه نمونه با جامعه کل شده و می تواند منجر به نتیجه گیری اشتباه در مورد جمعیت اصلی شود. گمشده گی داده یک اتفاق معمول بوده و بسته به میزان آن، می تواند اثر قابل توجهی در نتیجه گیری به دست آمده از داده ها داشته باشد. تمامی روش های برآورد پارامترها بر پایه فرض کامل بودن مجموعه داده ها استوار است و تحت برقراری این شرایط منجر به برآوردهایی نارایب می شوند؛ و البته با افزایش نسبت گمشدگی، مقدار آریبی نیز افزایش خواهد یافت. در این پژوهش بدین صورت عمل گردیده است که یک سری مجموعه داده واقعی بر حسب درصد گمشده گی مقادیری را از دست می دهند سپس بر اساس الگوریتم جستجوی گرانشی مقادیر گمشده تخمین زده می شوند و بر اساس معیار میانگین خطای مطلق میزان درصد خطای بدست آمده از روش جستجوی گرانشی نسبت به مقادیر واقعی محاسبه می شود که این معیار بر اساس نتایج ارزیابی، خطای کمتری را نشان می دهد و هر چه داده ها بیشتر باشند این معیار نیز به تناسب بیشتر می شود. سپس با استفاده از الگوریتم طبقه بندی SVM برای طبقه بندی داده ها بر اساس مقادیر تعیین شده توسط الگوریتم جستجوی گرانشی دقت حاصل با دقت طبقه بندی بر روی داده های حاصل از پر کردن با میانگین ویژگی و همچنین پر کردن داده ها با الگوریتم نزدیکترین همسایگی ($K-NN$) مقایسه می گردد. که در بخش ارزیابی نتایج نشان داده می شود که دقت طبقه بندی بر روی داده های بدست آمده از روش جستجوی گرانشی نسبت به روشهای دیگر بالاتر است و هرچه درصد گمشدگی بیشتر شود این دقت کمتر می شود.

واژه های کلیدی

داده های گمشده، جستجوی گرانشی، داده کاوی، الگوریتم طبقه بندی SVM - دقت طبقه بندی

۱. مقدمه