

بررسی ساختارهای داده متفاوت جهت انتخاب روش خوشه بندی مناسب با مطالعه موردی بر روی پایگاه داده سرطان Wisconsin سینه

محمود شاکری¹، محمد عبدالهی²

¹محمود شاکری، مرکز آموزش عالی علمی کاربردی جهاد دانشگاهی مشهد، mah_shakeri@yahoo.com
²محمد عبدالهی، مرکز آموزش عالی علمی کاربردی جهاد دانشگاهی مشهد، mabdolahi512@yahoo.com

چکیده

الگوریتم‌های خوشه‌بندی داده‌ها را به گروه‌هایی تقسیم می‌کنند؛ به صورتی که داده‌های درون هر خوشه بیش‌ترین شباهت و داده‌های خوشه‌های متفاوت کم‌ترین شباهت را به یکدیگر داشته باشند. انتخاب روش خوشه‌بندی مناسب و معیار فاصله یا شباهت مورد استفاده برای هر پایگاه داده، از اهمیت ویژه‌ای برخوردار است. به منظور تعیین روش مناسب، اولین قدم شناخت کافی از ساختار داده‌های موجود در پایگاه داده است. به عنوان مثال، توزیع داده‌های هر خوشه، تعداد خوشه‌ها و ارتباطات درون خوشه‌ای و بین خوشه‌ای از جمله مواردی بوده که بر انتخاب معیار فاصله یا شباهت و روش خوشه‌بندی تاثیر زیادی داشته و در بهینه سازی عمل خوشه‌بندی اثر مستقیم دارند. پس از تحلیل ساختار داده‌ها، می‌توان با توجه به ویژگی‌های معیارهای فاصله یا شباهت و روش‌های خوشه‌بندی، تصمیم‌گیری نموده و بهترین و مناسب‌ترین الگوریتم خوشه‌بندی را برگزید. در این مقاله پس از تحلیل ساختار داده‌های پایگاه داده سرطان سینه Wisconsin، روش‌های مختلف خوشه‌بندی و معیارهای شباهت مناسب مورد بررسی قرار گرفته است. نتایج آزمایشات با تحلیل‌های صورت گرفته تطابق کامل دارند.

کلمات کلیدی: خوشه بندی، یادگیری بدون نظارت، شبکه‌های عصبی، سرطان سینه