سمینار هفتگی گروه آمار

طراحی و ارزیابی یک مدل دسته بندی با مدیریت داده های گمشده پزشکی

سخنران:

جناب آقای گلاب پور

دانشجوی دکتری انفورماتیک پزشکی، دانشگاه علوم پزشکی مشهد

زمان:

۱۱ اسفند ماه ۱۳۹۵ ساعت ۱۰ صبح

مکان:

سالن دکتر بزرگ نیا

چکیده:

بیماری و درمان با مشاهده و تفسیر داده‌های پزشکی ارتباط تنگاتنگی دارد. جمع‌آوری و تفسیر معانی داده‌های پزشکی، محور مراقبت‌های بهداشتی است. اهمیت داده‌های پزشکی در مراقبت بهداشتی ناشی از بحرانی بودن نقش آن‌ها در فرایند تصمیم‌گیری است. در واقع همه فعالیت‌های مراقبتی با جمع‌آوری، تحلیل و استفاده از داده‌های پزشکی مرتبط هستند.

داده‌های گمشده یکی از چالش‌های موجود در پیش‌پردازش داده‌ها در علوم پزشکی است. منظور از داده گمشده، داده‌ای است که مقدار آن برای یک متغیر ثبت نشده است. طی سال‌های اخیر روش‌های بسیاری برای غلبه بر این مشکلات ارائه‌شده است. با این وجود، به دلایلی ازجمله عدم آشنایی پژوهشگران با این روش‌ها و عدم دانش چگونگی استفاده از آن‌ها، بسیاری از پژوهشگران همواره از روش‌های کلاسیک اولیه استفاده می‌کنند. حال‌آنکه استفاده از این روش‌ها در بسیاری از موارد به دلیل انحرافیکه وارد مسئله می‌نمایند، بیش از آنکه کیفیت داده‌های پزشکی را افزایش دهند، باعث کاهش کیفیت داده‌های پزشکی می‌شوند. در اکثر مدل‌های ارائه‌شده، بیشتر به دنبال افزایش یکی از ویژگی‌های داده شامل صحت، ویژگی و غیره هستند و معمولاً این

مدل‌ها قادر به بهبود چندین ویژگی از داده‌ها به‌صورت همزمان نیستند.

داده‌های پزشکی با مقادیر گمشده، منجر به کشف دانش داده‌کاوی استخراج‌شده باکیفیت پایین می‌شوند. لذا تکنیک‌های پیش‌پردازش و پاک‌سازی داده‌ها با هدف ارتقاء کیفیت داده‌ها به کار می‌روند. یکی از مشکلات رایج برای پژوهشگران در هنگام کار با داده‌ها وجود مقادیر گمشده است، که به دلایل گوناگونی رخ می‌دهند. در این موارد، پیش از مدل‌سازی نیاز به پر کردن مقادیر داده‌ها یا حذف رکوردهایی با فیلد گمشده است. حذف رکوردها باعث می‌شود قسمتی از اطلاعات داده حذف شود، به‌عنوان‌مثال با حذف یک رکورد داده که دارای تعدادی فیلد گمشده است کلیه اطلاعات مرتبط با آن رکورد حذف می‌شود، لذا کیفیت دانش استخراج‌شده از آن داده کاهش می‌یابد درصورتی‌که با پر کردن مقادیر گمشده با مقدار مناسب کیفیت دانش استخراج‌شده افزایش می‌یابد.

در تمام روش‌های بکار رفته در حوزه پزشکی از مدل‌های آماده موجود استفاده می‌شود و معمولاً هیچ پژوهشگری مدل ویژه‌ی کار خود را استفاده نمی‌نماید. ما در این پژوهش به کمک الگوریتم‌های بهینه‌سازی به دنبال ارائه یک مدل هستیم که بتواند چندین معیار از داده‌ها را به‌صورت همزمان بهبود دهد. این معیارها شامل حساسیت، دقت و ویژگی داده‌های پزشکی است. این مدل به‌صورت ویژه‌ای برای داده‌های پیوسته‌ی پزشکی مناسب است هرچند که برای داده‌های گسسته هم قابل‌استفاده خواهد بود.