تشخیص ناهنجاری در فضاهای با مقیاس بزرگ و ابعاد بالا با استفاده از مدل ترکیبی ماشین بردار پشتیبان تک کلاسه و یادگیری عمیق

احمد اسدی - ۹۴۱۳۱۰۹۱

آبان ماه ۱۳۹۵

چکیده

تشخیص ناهنجاری در مسائلی که با داده‌های با ابعاد بالا روبرو هستند با چالش‌های مختلفی روبرو است. یکی از مهم‌ترین این چالش‌ها، مشکل «نفرین ابعاد» است. با افزایش تعداد ابعاد مورد استفاده در مساله، تعداد ویژگی‌های استخراج شده که ارتباط معناداری با برچسب داده‌ها ندارند، افزایش خواهد یافت. این مساله باعث ایجاد مشکلات متعددی در مسیر تشخیص ناهنجاری در فضاهای با بعد بالا می‌شود. برای حل این مشکل، می‌توان با استفاده از روش‌های مبتنی بر خوشه‌بندی، ابتدا ویژگی‌های مناسبی را از بین تمام ویژگی‌های موجود انتخاب کرد به طوری که علاوه بر کاهش ابعاد مساله، توانایی مناسبی در ایجاد توزیع‌های خوش‌فرم دربرچسب‌های مختلف را نیز داشته باشند. سپس با استفاده از این ویژگی‌ها و الگوریتم‌های معمول در حوزه تشخیص ناهنجاری، عملیات مورد نظر را اجرا نمود. در این پژوهش، در مرحله اول با استفاده از یک شبکه عصبی باور عمیق، ویژگی‌های مناسب استخراج شده و سپس با به‌کارگیری یک مدل ماشین بردار پشتیبان تک کلاسه در مرحله دوم، عملیات دسته‌بندی انجام می‌شود.

۱ مقدمه

استخراج ویژگی برای داده‌ها از جمله مهم‌ترین چالش‌ها در فرآیند حل مساله است. علاوه بر این، در شرایطی که تعداد داده‌ها بسیار زیاد باشد، عدم وجود داده‌های برچسب خورده به اندازه کافی، لزوم استفاده از یادگیری بدون نظارت را بیش از پیش جلوه می‌دهد. هدف اصلی در پژوهش‌های مربوط به تشخیص ناهنجاری این است که داده‌هایی را که رفتاری غیر عادی نسبت به داده‌های دیگر از خود نشان می‌دهند، شناسایی شوند. چالش اصلی در تشخیص ناهنجاری، توانایی کنترل مناسب مجموعه‌دادگان بزرگ و نویزی است. در پژوهش مورد مطالعه، با ارائه یک روش ترکیبی بدون نظارت، سعی می‌شود تا حد ممکن بر این مشکلات غلبه شود.

مجموعه‌های دادگان ابعاد بالا، مشکلاتی برای تشخیص ناهنجاری ایجاد می‌کنند که از جمله مهم‌ترین آن‌ها می‌توان به ۱) افزایش نمایی فضای جستجو، ۲) وجود ویژگی‌های نامربوط به برچسب‌ها. مدل‌های مختلفی برای حل این مساله ارائه شده است. یکی از مشهورترین این مدل‌ها، دسته مدل‌های موسوم به ماشین‌های بردار پشتیبان تک کلاسه[1] هستند. این دسته از مدل‌ها سعی در مدل‌سازی توزیع داده‌های عادی موجود در مجموعه داده‌ها دارند و به طور همزمان سعی می‌کنند تا حد ممکن، مدل ارائه شده را نسبت به داده‌های نویزی یا ناهنجاری‌های موجود، غیرحساس کنند. به همین منظور، با استفاده از یک تابع هسته[2] داده‌های موجود را به یک فضای با ابعاد بالاتر نگاشت می‌کنند به طوری‌که بتوان در فضای جدید، داده‌های عادی را به راحتی از ناهنجاری‌ها جدا نمود.

از جمله مزایای این دسته از مدل‌ها می‌توان به سه مورد زیر اشاره کرد:

۱. قدرت تعمیم‌پذیری بسیار بالا

[1] One Class SVMs
[2] Kernel Function

۲. عدم وجود مشکل اکسترمم‌های محلی

۳. قدرت مدل‌سازی هر مجموعه‌داده‌ای بسته به نوع تابع هسته تعریف شده

با وجود این‌که ماشین‌های بردار پشتیبان مزایای زیادی دارند، محدودیت‌هایی که در مسائل ایجاد می‌کنند باعث می‌شود نتوان از آن‌ها مستقیما در مسائل با مقیاس بزرگ و ابعاد بالا به خوبی استفاده کرد. از جمله این محدودیت‌ها، رابطه نمایی زمان اجرای این الگوریتم‌ها با تعداد رکوردهای موجود در مجموعه‌داده است. از طرفی با توجه به مشکل «نفرین ابعاد» با افزایش بعد مساله باید تعداد داده‌های آموزشی به طور نمایی افزایش یابد که باعث می‌شود نتوان از ماشین‌های بردار پشتیبان در مسائل با ابعاد بالا استفاده نمود.
یکی از مدل‌های مطرح در زمینه دسته‌بندی و کاهش بعد، شبکه‌های باور عمیق هستند. این شبکه‌ها با استفاده از یک الگوریتم حریصانه، به شکل لایه به لایه آموزش داده می‌شوند و قادرند در مسائل دسته‌بندی چندکلاسه عملکردهای بسیار مناسبی از خود نشان بدهند.
از مزایای شبکه باور عمیق می‌توان به قابلیت بالای این شبکه در مدل‌سازی داده‌ها با ابعاد بالا و مقیاس بزرگ اشاره کرد. همین‌طور این شبکه‌ها قادرند داده‌های پیچیده و با ابعاد بالا را تحت یک روش بدون نظارت، در یک فضای با ابعاد کوچکتر (یا بزرگتر) بازتولید کنند.