الگوریتم RESSEL

پروژه پایان ترم کلاس یادگیری ماشین: توصیف مدل و شیوه پیاده سازی

استاد: جناب آقای دکتر شاقوزی

محقق: سجاد صابری

معرفی روشRESSEL

روش (Reliable Semi-Supervised Ensemble Learning) **RESSEL** ، یک نوع روش یادگیری نیمه‌نظارتی است که ترکیبی از یادگیری نیمه‌نظارتی و یادگیری شورایی (Ensemble Learning) را به‌کار می بندد. در این روش، از داده‌های برچسب‌دار و بدون‌برچسب استفاده می‌شود تا با بهره‌گیری از داده‌های بدون‌برچسب، مدل یادگیری نظارتی بهبود یابد. هدف اصلی RESSEL این است که با استفاده از مجموعه‌ای از طبقه‌بندها، از نمونه‌های بدون‌برچسب به شکل موثرتری برای بهبود دقت پیش‌بینی‌ها استفاده شود.

هدف اصلیRESSEL

در این روش، مجموعه کوچکی از داده‌های برچسب‌دار و یک مجموعه بزرگ از داده‌های بدون‌برچسب انتخاب می شوند. این کار با خودآموزی (Self-Training) انجام می‌شود؛ به این معنا که مدل پس از یادگیری اولیه، نمونه‌هایی از داده‌های بدون‌برچسب که با اعتماد بیشتری طبقه‌بندی می‌شوند را به عنوان داده‌های برچسب‌دار به مجموعه آموزشی اضافه می‌کند. هدف این مسیر، بهبود پیشبینی و دسته بندی صحیح داده ها با بازده بالاتر در محیط واقعی است.

پارامترهای در نظر گرفته شده

در کد پیاده‌سازی RESSEL به زبان پایتون، چندین پارامتر کلیدی در نظر گرفته شده است که برای کنترل فرایند یادگیری مهم هستند:

1. **ensemble_size تعداد مدل‌های جمعی :** تعداد مدل‌های جمعی یا طبقه‌بندهایی که قرار است در الگوریتم جمعی استفاده شوند. تعداد بیشتر باعث ایجاد تنوع بیشتر و بهبود دقت کلی می‌شود.
2. **Iterations تعداد تکرارهای خودآموزی:** تعداد تکرارهایی که در آن داده‌های بدون‌برچسب با اطمینان بالا به مجموعه داده‌های برچسب‌دار اضافه می‌شوند.
3. **batch_size اندازه دسته داده‌های بدون‌برچسب:** تعداد نمونه‌هایی که در هر تکرار از داده‌های بدون‌برچسب با اعتماد بالا انتخاب شده و به داده‌های برچسب‌دار اضافه می‌شوند.
4. **unlabeled_fraction نسبت داده‌های بدون‌برچسب:** نسبت داده‌های بدون‌برچسب که در هر تکرار برای خودآموزی در نظر گرفته می‌شود.
5. **oob_threshold آستانه خطای(OOB) :** آستانه‌ای که برای توقف زودهنگام (Early Stopping) براساس خطای OOB (خطای خارج از کیسه) استفاده می‌شود. اگر خطای OOB به کمتر از این مقدار برسد، مدل متوقف می‌شود تا از بیش‌برازش جلوگیری شود.

مراحل پیاده‌سازی در پایتون

1. **بارگذاری داده‌ها** :ابتدا داده‌های **Spambase** که شامل ویژگی‌های ایمیل‌ها برای تشخیص هرزنامه است، بارگذاری می‌شود. داده‌ها به دو مجموعه **داده‌های برچسب‌دار** (برای آموزش اولیه) و **داده‌های بدون‌برچسب** (برای خودآموزی) تقسیم می‌شوند.
2. **پیش‌پردازش داده‌ها**: داده ها با استفاده از ابزار **MinMaxScaler**مقیاس‌بندی شده و از **SelectFromModel**برای انتخاب ویژگی‌های مهم استفاده می‌شود. مدل جنگل تصادفی (RandomForest)برای انتخاب ویژگی‌ها به‌کار می‌رود.

3. **تعریف مدل پایه: DecisionTreeClassifier**به عنوان مدل پایه استفاده می‌شود. این طبقه‌بند به عنوان واحد اصلی در روش جمعی به کار می‌رود.
4. **استفاده از BaggingClassifier :** برای ایجاد مجموعه ای از مدل ها، از **BaggingClassifier** استفاده می‌شود که چندین طبقه‌بند تصمیم‌گیری را روی داده‌های آموزشی بوت‌استرپ شده آموزش می‌دهد و از OOBبرای ارزیابی مدل استفاده می‌کند.
5. **خودآموزی:** در هر تکرار، مدل با استفاده از داده‌های برچسب‌دار اولیه آموزش می‌بیند و سپس از داده‌های بدون‌برچسب استفاده می‌کند تا نمونه‌های با ضریب اعتماد بالا را به عنوان داده‌های برچسب‌دار به مجموعه آموزشی اضافه کند.
6. **محاسبه خطای OOB** : پس از هر تکرار، خطای OOB محاسبه می‌شود. این خطا برای ارزیابی عملکرد مدل روی داده‌هایی که در نمونه‌برداری بوت‌استرپ حضور نداشته‌اند به کار می‌رود.
7. **توقف زودهنگام**: اگر خطای OOB به زیر مقدار تعیین‌شده برسد، فرایند خودآموزی متوقف می‌شود.

در نهایت، این کد به داده‌های Spambase اعمال می‌شود که شامل ویژگی‌های ایمیل‌هاست. این کد از روش خودآموزی برای افزایش دقت طبقه‌بندهای جمعی استفاده می‌کند و از داده‌های بدون‌برچسب بهره می‌برد تا مدل بهتری ایجاد کند. در قسمت انتهایی کد، فرآیند ارزیابی مدل انجام می‌شود تا عملکرد نهایی مدل بر روی داده‌های تست شده بررسی گردد.