بازبینی

Two scalable algorithms for associative text classification

داریوش حسن‌پور آده

۹۳۰۸۱۶۴

این مقاله در مورد طبقه‌بندی انجمنی متون[1] می‌باشد، مساله‌ای که در مقاله بیان شده و سعی در حل این مساله داشته این است که آورده در برخی الگوریتم‌های از این نوع برای بالا بردن کارایی و دقت خود تعداد بسیار زیادی قوانین انجمنی زیادی را تولید می‌کنند، بنابراین زمان زیادی را این الگوریتم ها نیاز دارند و همچنین در برخی از موارد ممکن است به علت تولید بیش از حد نیاز قوانین کارایی پایینی داشته باشند. بنابراین این مقاله آمده است ۲ عدد الگوریتم همکار جهت افزایش کارایی و کاهش پیچیدگی محاسباتی برای این هدف معرفی کرده است که اولی قوانین استخراج شده را به صورت موثر ذخیره می‌کند و دیگری سرعت تطبیق قوانین[2] بالا می‌برد.
به طور کلی هدف یک دسته‌بند متون، دسته‌بندي اسناد در قالب تعداد معینی دسته‌هاي از پیش تعیین شده می باشد. هر سند می‌تواند در یک، چند و یا هیچ دسته‌اي قرار بگیرد. این موضوع می‌تواند در قالب یک یادگیري خودکار قرار گیرد تا با استفاده از آن بتوان هر سند را به طور خودکار به دسته‌اي نسبت داد.
در این مقاله، از روش دسته‌بندي بر مبناي قواعد انجمنی که از روي فرایند کاوش الگوهاي مکرر مجموعه داده هاي آموزشی تولید شده استفاده کرده‌اند، می شود. این فرآیند با فرآیندي که در داده کاوي داده‌هاي بزرگ پایگاه داده‌ها استفاده می شود یکسان می‌باشد. استفاده از قواعد انجمنی و ترکیب آن با قواعد دسته‌بندي و ایجاد مدل جدیدي با عنوان دسته‌بندي انجمنی و استفاده از آن براي دسته‌بندي متون می باشد.

در مورد نقات قوت روش ارائه شده می‌توان گفت که به علت اینکه روش ارائه شده به صورت ماژولار[3] ارائه شده اند – یعنی به دوقسمت ذخیره‌سازی و تطبیق قوانین تقسیم شده است، کلیه فواید حاکم به سیستم های ماژولار را دارا می‌باشد و اینکه با توجه به اینکه تطبیق قوانین یکی از وقت‌گیرترین فاز اجرایی الگوریتم‌های طبقه‌بندی انجمنی می‌باشد بنابراین با پرداختن به این فاز و تلاش برای افزایش سرعت این فاز از امتیازات این روش می‌توان به‌حساب آورد. از دیگر مزایای این روش می‌توان به قابلیت تفسیر ساده، درک آسان قواعد توسط انسان اشاره کرد. حذف قواعد ضعیف می تواند تا حد فوق‌العاده دقت دسته‌بندي را افزایش دهد. و همچنین ویژگی‌ها هم می‌توانند منفرد باشند وهم چندگانه، یعنی می‌توان از اطلاعات ترکیبی ویژگی‌هاي چندگانه استفاده کرد.

دسته‌بندي بر مبناي قواعد انجمنی داراي معایبی هم هست که از آن جمله افزایش بعد فضاي برداري ویژگی‌ها می‌باشد که براي رفع این مشکل از تکنیک‌هاي کاهش بعد فضاي ویژگی‌ها استفاده می شود، و همچنین افزایش تعداد قواعدي که در فاز آموزش تولید شده‌اند و باعث افزایش بیهوده زمان محاسبات و کاهش تاثیر در دسته‌بندي انجمنی می شوند. براي رفع این مشکل هم از تکنیک هرس کردن قواعد استفاده می‌شود. در این تکنیک فقط قواعدي که داراي کیفیت و تأثیر بالایی هستند انتخاب می شوند.

---
[1] Associative Text Classification
[2] Rule Matching
[3] Modular