بسم الله الرحمن الرحيم

دانشکده فنی و مهندسی

گروه مهندسی برق

پایان‌نامه تحصیلی برای دریافت درجه کارشناسی ارشد

رشته مهندسی برق گرایش مخابرات سیستم

بهینه سازی موقعیت نماینده‌ها در طبقه‌بند نزدیکترین همسایه با استفاده از الگوریتم‌های ابتکاری

مؤلف:

محدثه رضایی

استاد راهنما:

دکتر حسین نظام آبادی پور

استاد مشاور:

دکتر داریوش عباسی مقدم

شهریورماه ۱۳۹۳

تقدیم به :

پدر بزرگوار و مادر مهربانم که هر لحظه وجودم را از چشمه‌سار پر از عشق چشمانشان سیراب می‌کنند،

و به همسرم، به پاس قدردانی از قلبی آکنده از عشق و معرفت که محیطی سرشار از امنیت و آرامش برای من فراهم آورده است،

و به فرزند دلبندم، که کودکی گمشده‌ام را در چهره معصومش یافته‌ام.

تشکر و قدردانی:

سپاس بی‌کران پروردگار یکتا را که به ما هستی بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه‌چینی از علم و معرفت را روزیمان ساخت.

نمی‌توانم معنایی بالاتر از تقدیر و تشکر بر زبانم جاری سازم و سپاس خود را در وصف استادان خویش آشکار نمایم که هر چه گویم، کم گفته‌ام. از استاد اندیشمند و شایسته جناب آقای دکتر نظام آبادی پور که در کمال سعه صدر، با حسن خلق و فروتنی از هیچ کمکی دریغ ننمودند و زحمت راهنمایی این پایان‌نامه را بر عهده گرفتند و از استادان گرامی، آقای دکتر سریزدی و آقای دکتر افتخاری که زحمت داوری این پایان‌نامه را متقبل شدند، کمال تشکر و قدردانی را دارم. همچنین از پدر و مادر عزیز و همسر مهربانم که آرامش روحی و آسایش فکری فراهم نمودند تا با حمایت‌های همه جانبه در محیطی مطلوب، مراتب تحصیلی و نیز پایان‌نامه درسی را به نحو احسن به اتمام برسانم، سپاسگذاری می‌نمایم. باشد که این خردترین، بخشی از زحمات آنان را سپاس گوید.

چکیده:

در سال‌های اخیر تلاش‌های زیادی در زمینه‌ی استفاده از الگوریتم‌های ابتکاری در حل مسائل طبقه‌بندی و خوشه‌بندی صورت گرفته است. در این پایان‌نامه برای نخستین بار از الگوریتم جستجوی گرانشی که یکی از جدیدترین الگوریتم‌های ابتکاری مبتنی بر جمعیت است، در حل مسأله تولید نماینده برای طبقه‌بند k-نزدیک‌ترین همسایه استفاده شده است. به منظور بهبود عملکرد، راهکارهایی چون استفاده از «قانون رأی‌گیری اکثریت» و اصلاح جمعیت اولیه‌ی الگوریتم ارائه شده است. همچنین دو روش ترکیبی که مسائل تولید و انتخاب نماینده را با هم ترکیب می‌کنند، پیشنهاد شده‌اند. روش‌های پیشنهادی روی پانزده مجموعه داده ارزیابی شده و نتایج آنها با ۹ روش روزآمد مقایسه شده است. نتایج بدست آمده نشان می‌دهند که روش‌های پیشنهادی دقت طبقه‌بندی بالاتری نسبت به روش‌های مورد مقایسه داشته و عملکرد مناسبی در زمینه تولید نماینده دارند. بالاترین میانگین دقت طبقه‌بندی بدست آمده در میان روش‌های پیشنهادی روی مجموعه آزمون، تقریباً ۱/۵٪ بیشتر از بهترین روش مورد مقایسه است.

واژه‌های کلیدی: طبقه‌بندی، طبقه‌بند k-نزدیکترین همسایه ، تولید نماینده، الگوریتم جستجوی گرانشی، الگوریتم‌های ابتکاری

فهرست مطالب

فصل سوم: پیش‌زمینه : کاهش داده و الگوریتم جستجوی گرانشی

فصل چهارم: تولید نماینده با استفاده از الگوریتم جستجوی گرانشی

فهرست جدول‌ها

فهرست شکل‌ها

فصل اول

مقدمه

۱-۱ مقدمه

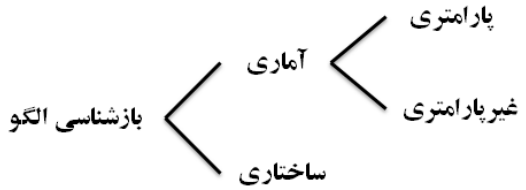
بازشناسی الگو یک زمینه علمی است که هدف آن طبقه‌بندی الگوها در تعدادی دسته[۱] یا گروه[۲] است. بسته به نوع کاربرد، این الگوها می‌توانند تصاویر، شکل موج سیگنال‌ها یا هر سنجش دیگری باشند که نیاز به طبقه‌بندی دارد. در بازشناسی الگو، بر اساس فضای نمایش مورد استفاده می توان دو شاخه را از یکدیگر متمایز کرد (شکل ۱-۱). بازشناسی الگوی ساختاری[۳] مبتنی بر پردازش زبان‌های طبیعی است [۱]. بر این اساس، همانطور که قواعد دستور زبان میان جمله‌ها ارتباط برقرار می‌کنند، روابط بین الگوها نیز توسط یک دستور زبان تعیین می‌شود. بنابراین درطبقه‌بندی بوسیله‌ی بازشناسی الگوی ساختاری، باید مشخص کرد که آیا یک زنجیره می‌تواند با یک دستور زبان خاص تولید شود یا نه.

شاخه دوم، بازشناسی الگوی آماری[۴] [۲] است که مبتنی بر تئوری تصمیم است. بر این اساس، فضای نمایش یک فضای برداری است و در این فضا فرض می‌شود که هیچ ارتباط ساختاری میان ویژگی‌های مختلف وجود ندارد. یک الگو نقطه ای در این فضای نمایش است و مختصات آن با یک بردار ویژگی تعیین می‌شود. در مرحله‌ی طراحی یک طبقه‌بند آماری، هدف تعیین مرزهایی برای جداسازی الگوهای مربوط به گروه‌های مختلف، در این فضای نمایش است که مرزهای تصمیم نام دارند. بنابراین در طبقه‌بندی بوسیله‌ی بازشناسی الگوی آماری، برچسب گروه یک داده‌ی جدید براساس ناحیه‌ای که درآن قرار می‌گیرد، تعیین می‌شود.

۱-۲ طبقه‌بند K-نزدیکترین همسایه[۵]

بازشناسی الگوی آماری را می‌توان به دو دسته روش‌های پارامتری و روش های غیرپارامتری تقسیم کرد (شکل ۱-۱). در تقریب پارامتری، دانش قبلی راجع به توزیع احتمال هر گروه در فضای نمایش وجود دارد که با تعدادی پارامتر مشخص می‌شود. با استفاده از این توزیع‌ها مرزهای تصمیم تعریف می‌شوند. در مقابل روش‌های پارامتری، روش‌های غیرپارامتری وجود دارند که هیچ دانش قبلی راجع به توزیع احتمال گروه‌ها در فضای نمایش ندارند. تنها اطلاعات موجود برای

[1] Category
[2] Class
[3] Syntactic Pattern Recognition
[4] Statistical Pattern Recognition
[5] k-Nearest Neighbor



شکل ۱- ۱ شاخه‌های بازشناسی الگو

این روش‌ها، یک مجموعه داده به نام مجموعه آموزش[۶] (TR) است که برچسب گروه داده‌ها در آن معلوم است. طبقه‌بندهای غیرپارامتری با استفاده از دانش استخراج شده از این مجموعه آموزش، برچسب نمونه‌های مجموعه آزمون[۷] (TS)را پیش بینی می‌کنند.

در میان طبقه‌بندهای غیرپارامتری، می‌توان روش‌های مبتنی بر همسایگی را به دلیل سادگی مفهومی، از دیگر روش‌ها متمایز کرد. یکی از توابع تصمیم مبتنی بر همسایگی پرکاربرد در طبقه‌بندی، قانون k نزدیکترین همسایه (K-NN) است [۳]. برای طبقه‌بندی یک داده جدید، ابتدا براساس معیار فاصله مورد استفاده، k همسایه نزدیکتر آن از میان نمونه‌های مجموعه آموزش مشخص می‌شوند، سپس برچسب گروه اکثریت به این داده جدید اختصاص می‌یابد. طبقه‌بند K-NN بدون ایجاد یک مدل در مرحله یادگیری، برای تعیین برچسب یک داده جدید تنها از داده‌های آموزش خام استفاده می‌کند و به همین دلیل پیاده‌سازی آن به سادگی صورت می‌گیرد. علاوه بر این سادگی مفهومی، رفتار طبقه‌بند K-NN نزدیک به بهینه بیز بوده و نرخ خطای این طبقه‌بند حداکثر دو برابر خطای طبقه بند بیز است [۴].

با وجود سادگی و دقت طبقه‌بندی بالا، چند ضعف در ساختار طبقه‌بند K-NN وجود دارد. نخستین مشکل این طبقه‌بند، نیاز به حافظه‌ی زیاد برای ذخیره‌ی کل مجموعه آموزش است. علاوه بر این، محاسبه میزان فاصله‌ی نمونه‌های ناشناخته با همه‌ی نمونه‌های مجموعه آموزش، منجر به افزایش هزینه محاسباتی خواهد شد. سومین اشکال این طبقه‌بند، حساسیت زیاد آن به نویز است. چون همه داده‌ها حتی داده‌های نویزی به عنوان مجموعه آموزش ذخیره می‌شوند، ممکن است نتایج حاصل از طبقه‌بندی تفاوت زیادی با نتایج بهینه داشته باشند.

[6] Training set
[7] Test set

تاکنون روش‌های متعددی برای بهبود عملکرد طبقه‌بند K-NN پیشنهاد شده‌اند. روش‌های مبتنی بر کاهش داده[8] می‌توانند به طور همزمان بر هرسه ضعف طبقه‌بند K-NN غلبه کنند [۵] و [۶]. در کاهش داده، هدف بدست آوردن یک مجموعه داده با اندازه‌ای کوچکتر نسبت به مجموعه آموزش اصلی است که به کمک آن بتوان یک داده‌ی جدید را با دقت مشابه یا حتی بالاتر طبقه‌بندی کرد. از دید فضای ویژگی می‌توان انتخاب ویژگی[9] [۷] و استخراج ویژگی[10] [۸] را به عنوان روش‌های اصلی کاهش داده در نظر گرفت. انتخاب ویژگی، انتخاب یک زیرمجموعه ویژگی نماینده از فضای ویژگی است، در حالی که استخراج ویژگی، تولید ویژگی‌های جدید برای نمایش داده‌هاست.

کاهش داده را می‌توان متناظر با کاهش تعداد نمونه‌ها دانست. از این نقطه نظر، روش‌های کاهش داده را می‌توان به دو دسته انتخاب نماینده[11] [۹] و تولید نماینده[12] [۱۰] تقسیم کرد. در روش اول یک زیرمجموعه از نمونه‌های آموزش انتخاب شده و به عنوان نماینده کل داده‌ها معرفی می‌شود. این زیرمجموعه با حذف داده‌های نامناسب و نویزی بدست می‌آید. در تولید نماینده، در کنار انتخاب داده‌ها می‌توان نماینده‌های جدیدی نیز تولید کرد و مجموعه داده اصلی را با نماینده‌های مصنوعی تولید شده، جایگزین نمود. این فرآیند، تعریف نماینده برای نواحی فاقد نماینده را ممکن می‌سازد. در انتخاب نماینده فرض اولیه این است که بهترین نماینده‌ها را می‌توان در میان نمونه‌های مجموعه داده اصلی یافت، در حالی که روش‌های تولید نماینده، در صورت لزوم نماینده‌های جدیدی را برای نمایش داده‌ها تولید می‌کنند.

به کمک روش‌های وزن‌دهی نیز می‌توان عملکرد طبقه‌بند K-NN را بهبود بخشید. وزن‌دهی ویژگی[13] [۱۱] روش شناخته شده‌ای است که در آن به هر یک از ویژگی‌های موجود یک وزن اختصاص می‌یابد. به این ترتیب محاسبه فاصله میان نمونه‌ها اصلاح می‌گردد. وزن‌دهی نمونه‌ها[14] روش دیگری است که در آن به هر نمونه مجموعه آموزش یک وزن اختصاص می‌یابد. این روش نیز محاسبه فاصله بین نمونه‌ها را اصلاح می‌کند [۱۲].

---

[8] Data reduction
[9] Feature selection
[10] Feature extraction
[11] Prototype selection
[12] Prototype generation
[13] Feature weighting
[14] Instance weighting

استفاده از روش‌های ترکیبی که به طور همزمان از چند روش کاهش داده استفاده می‌کنند یا روش‌های کاهش داده را با روش‌های وزن‌دهی ترکیب می‌کنند نیز بسیار مورد توجه قرار گرفته است [۱۳]. استفاده همزمان از این روش‌ها می‌تواند دقت طبقه‌بندی را افزایش داده و باعث دستیابی به نرخ کاهش بالاتری گردد. در این روش‌ها تولید نماینده یا انتخاب نماینده با روش‌های انتخاب ویژگی یا وزن‌دهی ترکیب می‌شود [۱۴] و [۱۵] و [۱۶].

در میان پیشنهادهایی که تاکنون برای حل مسأله تولید نماینده صورت گرفته است، روش‌های تولید نماینده مبتنی بر الگوریتم‌های ابتکاری، توانسته‌اند بهترین نتایج را بدست آورند. در این روش‌ها، در یک روند بهینه‌سازی مجموعه‌ای از نماینده‌ها به عنوان جواب مسأله، در موقعیت بهینه قرار می‌گیرد. چون الگوریتم‌های ابتکاری از روش‌های تقریبی هستند و رسیدن به جواب بهینه به تعادل میان کاوش[۱۵] و بهره‌گیری[۱۶] آنها بستگی دارد، آزمودن روش‌های جدید در حل این مسأله می‌تواند در رسیدن به جواب بهینه مؤثر باشد. هنگامی که از الگوریتم‌های ابتکاری در مسأله تولید نماینده استفاده می‌شود، باید راهکارهای مناسبی برای الف)انتخاب جمعیت اولیه، ب)روند بهینه-سازی موقعیت نماینده‌ها و ج)حذف نماینده‌های نامناسب ارائه شود. اگر یک مرحله انتخاب نماینده پیش از آغاز مرحله تولید صورت گیرد، می‌تواند به انتخاب مناسب جمعیت اولیه کمک کند و زمانی که پس از آن انجام شود باعث حذف نماینده‌های نامناسب می‌شود. بنابراین با ترکیب روش‌های تولید و انتخاب نماینده می‌توان نتایج خوبی بدست‌آورد. در فصل دوم الگوریتم‌هایی که تاکنون برای تولید نماینده ارائه شده‌اند، مرور و بررسی می‌شوند.

۱-۳ اهداف پایان‌نامه

در این پایان‌نامه، یک رویکرد جدید برای تولید نماینده، با استفاده از الگوریتم جستجوی گرانشی ارائه می‌شود. همچنین ترکیب روش‌های تولید و انتخاب نماینده با استفاده از الگوریتم جستجوی گرانشی، به صورت توأم و متوالی و تأثیر آن بر بهبود نتایج مورد بررسی قرار می‌گیرد. علاوه بر این، دو راهکار برای انتخاب جمعیت اولیه پیشنهاد می‌شود. از دیگر اهداف پایان‌نامه مقایسه توان الگوریتم‌های ابتکاری جمعیت ذرات، تکامل تفاضلی و جستجوی گرانشی و تعدادی از روش‌های روزآمد، در مسأله تولید نماینده است.

[15] Exploration
[16] Exploitation

۱-۴ سازمان پایان‌نامه

ساختار پایان‌نامه به این صورت است که ابتدا در فصل دوم، مروری بر الگوریتم‌هایی که تاکنون برای تولید نماینده ارائه شده‌اند، صورت می‌گیرد. پس از آن در فصل سوم پیش‌زمینه‌ای از کاهش داده و الگوریتم جستجوی گرانشی مطرح می‌شود. در فصل چهارم روش‌های پیشنهادی برای تولید نماینده با استفاده از الگوریتم جستجوی گرانشی معرفی می‌شوند. در فصل پنجم مجموعه داده‌هایی که از آنها برای مقایسه روش ها استفاده شده، معرفی شده و نتایج آزمایش‌ها ارائه می‌شود. در پایان نیز در فصل ششم پایان‌نامه جمع‌بندی شده و پیشنهادهایی برای ادامه کار معرفی می‌شود.

فصل دوم

مروری بر کارهای انجام شده در زمینه تولید نماینده

۲-۱ مقدمه

همانگونه که در فصل نخست بیان شد، روش‌های تولید نماینده برای بهبود عملکرد طبقه‌بند K-NN معرفی شدند. این روش‌ها را می توان بر اساس معیارهای مختلفی دسته‌بندی کرد. ما بر اساس مرجع [۱۰] چهار معیار الف) نوع کاهش، ب) نوع مجموعه نهایی، ج) مکانیزم‌های تولید و د) ارزیابی جستجو را برای دسته‌بندی این الگوریتم ها در نظر می‌گیریم. اگرچه برای همه الگوریتم‌ها نمی‌توان این چهار معیار را به طور کامل مشخص نمود.

روشی که هر الگوریتم برای تعیین تعداد مناسب نماینده‌ها بکار می‌برد، با معیار نوع کاهش مشخص می‌شود. معیار دیگری که الگوریتم‌های تولید نماینده را از یکدیگر متمایز می‌کند، نوع مجموعه‌ی نهایی تولید شده توسط الگوریتم است که این مجموعه می‌تواند شامل داده‌های مرزی، داده‌های داخلی یا هر دو نوع باشد. معیار سوم بیانگر مکانیزم‌های مختلفی است که در الگوریتم‌ها برای تولید نماینده بکار رفته است و در ارزیابی جستجو، روش ارزیابی الگوریتم بررسی می‌شود. در ادامه فصل، ابتدا به بیان جزئیات چهار معیار فوق پرداخته، سپس الگوریتم‌ها و روش‌هایی که در مقالات برای حل مساله تولید نماینده پیشنهاد شده‌اند را معرفی خواهیم کرد.

۲-۲ مشخصه‌های اصلی روش‌های تولید نماینده

۲-۲-۱ نوع کاهش

روش‌های تولید نماینده به دنبال یافتن یک مجموعه کاهش یافته از نماینده‌ها برای نمایش مجموعه آموزش می‌باشند و اندازه این مجموعه کاهش یافته را می‌توان با روش‌های افزاینده، کاهنده، ثابت و مختلط تعیین کرد که در ادامه به بیان این روش‌ها می‌پردازیم.

الف) افزاینده[17]

در این روش مجموعه نماینده‌های اولیه تهی در نظر گرفته شده یا شامل تعداد کمی نماینده از هر گروه می‌باشد. طی اجرای الگوریتم، نماینده‌های جدیدی به این مجموعه اضافه شده یا موقعیت نماینده‌های موجود اصلاح می‌شود. مهم‌ترین مزیت این روش‌ها

[17] Incremental

این است که در مرحله یادگیری سریعتر بوده و نیازمند حافظه کمتری هستند. عیب اصلی این روش‌ها حساسیت به ترتیب ورود نمونه‌هاست.

ب)کاهنده[18]

در این روش کل مجموعه آموزش به عنوان مجموعه نماینده اولیه در نظر گرفته شده و الگوریتم شروع به کاهش این مجموعه می‌کند. این کاهش می‌تواند با ادغام، حرکت، حذف یا تغییر برچسب نماینده‌ها صورت گیرد. مزیت این روش این است که در تولید نماینده‌ها همه‌ی نمونه‌های آموزش نقش دارند، اما نسبت به دیگر روش‌ها هزینه محاسباتی بالاتری دارد.

ج) ثابت[19]

این روش‌ها تعداد نهایی نماینده‌ها را با استفاده از پارامتری که توسط کاربر تعریف می‌شود، مشخص می‌کنند و این مهم‌ترین ضعف آنهاست. زیرا تعداد نماینده‌ها بسیار وابسته به مجموعه داده است و مجموعه داده‌های مختلف به تعداد نماینده‌های متفاوتی نیاز دارند. به دلیل اینکه تعداد نماینده‌ها از قبل مشخص است، این روش‌ها تنها بر روی افزایش دقت طبقه‌بندی تمرکز می‌کنند.

د) مختلط[20]

در یک کاهش مختلط مجموعه اولیه نماینده‌ها با انتخاب تصادفی تعداد ثابتی نماینده یا با اجرای یک الگوریتم انتخاب نماینده مشخص می‌شود. سپس با روش‌های خاصی، تعدادی نماینده اضافه و حذف شده یا نماینده‌های موجود اصلاح می‌شوند. به این ترتیب این روش‌ها می‌توانند از مزایای روش‌های قبل بهره‌مند شوند، اما برای بیش‌برازش داده‌ها بسیار مستعد بوده و معمولاً هزینه محاسباتی بالایی دارند.

[18] Decremantal
[19] Fixed
[20] Mixed

۲-۲-۲ نوع مجموعه نهایی

این معیار مربوط به نوع مجموعه‌ای است که الگوریتم در نهایت تولید کرده است. این مجموعه می‌تواند شامل داده‌های مرزی، داده‌های داخلی یا هر دو باشد. مجموعه نهایی می‌تواند با یکی از روش‌های ذیل تولید شود.

الف) متراکم‌سازی[۲۱]

الگوریتم‌هایی که از روش متراکم‌سازی استفاده می‌کنند، مجموعه نماینده نهایی آنها شامل نماینده‌هایی نزدیک به مرزهای تصمیم است. دلیل نگه داشتن نماینده‌های مرزی این است که نقاط داخلی به اندازه نقاط مرزی بر مرزهای تصمیم اثر ندارند و با تأثیر منفی اندکی بر دقت طبقه‌بندی می‌توانند حذف شوند. این روش‌ها سعی می‌کنند دقت طبقه‌بندی روی مجموعه آموزش را حفظ کنند که این کار می‌تواند برقدرت تعمیم روش اثر منفی داشته باشد. توانایی کاهش روش‌های متراکم‌سازی به طور ذاتی بالاست چون تعداد داده‌های مرزی بسیار کمتر از داده‌های داخلی است.

ب)ویرایش[۲۲]

در این روش، داده‌های نویزی یا داده‌هایی که برچسبی متفاوت با نزدیک‌ترین همسایه های خود دارند، حذف می‌شوند. چون عموماً نمونه‌های مرزی این شرایط را دارند، نمونه‌های داخلی باقی مانده و نمونه‌های مرزی حذف می‌شوند. به این ترتیب، به دلیل بوجود آمدن مرزهای تصمیم نرم‌تر، دقت تعمیم روی داده‌های آموزش بهبود می‌یابد، اما الگوریتم نرخ کاهش پایین‌تری حاصل می‌کند.

ج)ترکیبی[۲۳]

برای بدست آوردن کوچک‌ترین مجموعه‌ای که قادر باشد دقت تعمیم روی داده‌های آزمون را حفظ کند یا حتی افزایش دهد، از روش‌های ترکیبی استفاده می‌شود. به این منظور در این الگوریتم‌ها داده‌های مرزی و داخلی بر اساس معیارهای مشخصی اصلاح می‌شوند. طبقه‌بند نزدیک‌ترین همسایه، انطباق زیادی با روش‌های ترکیبی دارد.

---

[21] Condensation
[22] Edition
[23] Hybrid