دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

پروژه نهایی درس شناسایی آماری الگو

روش‌های انتخاب ویژگی برای مسائل دسته‌بندی متن

نگارش

علیرضا مازوچی

استاد درس

دکتر محمد رحمتی

بهمن ۱۴۰۰


بسم الله الرحمن الرحيم

چکیده

در این قسمت چکیده پایان نامه نوشته می‌شود. چکیده باید جامع و بیان‌کننده خلاصه‌ای از اقدامات انجام‌شده باشد. در چکیده باید از ارجاع به مرجع و ذکر روابط ریاضی، بیان تاریخچه و تعریف مسئله خودداری شود.

واژه‌های کلیدی:

کلیدواژه اول، ...، کلیدواژه پنجم (نوشتن سه تا پنج واژه کلیدی ضروری است)

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

فهرست نمادها

نماد	مفهوم
$\mathbb{R}^n$	فضای اقلیدسی با بعد n
$\mathbb{S}^n$	کره یکه n بعدی
M^m	خمینه m-بعدی M
$\mathfrak{X}(M)$	جبر میدان‌های برداری هموار روی M
$\mathfrak{X}^1(M)$	مجموعه میدان‌های برداری هموار یکه روی (M,g)
$\Omega^p(M)$	مجموعه p-فرمی‌های روی خمینه M
Q	اپراتور ریچی
$\mathcal{R}$	تانسور انحنای ریمان
ric	تانسور ریچی
L	مشتق لی
Φ	۲-فرم اساسی خمینه تماسی
∇	التصاق لوی-چویتای
Δ	لاپلاسین ناهموار
∇^*	عملگر خودالحاق صوری القا شده از التصاق لوی-چویتای
g_s	متر ساساکی
∇	التصاق لوی-چویتای وابسته به متر ساساکی
Δ	عملگر لاپلاس-بلترامی روی p-فرم‌ها

فصل اول

مقدمه

بالای ۸۰ درصد از اطلاعات موجود در قالب داده‌های متنی ذخیره شده‌اند[۱]. پردازش این داده‌ها در حوزه پردازش زبان طبیعی[۱] است. یکی از کاربردهای این حوزه دسته‌بندی متون[۲] به تعدادی دسته از پیش تعیین شده است؛ به عنوان مثال ایمل‌های دریافتی یک فرد را در نظر بگیرید. تعدادی از ایمیل‌ها، ایمیل‌هایی هستند که کاربر مایل به دریافت آن است و تعدادی دیگر هرزنامه[۳] هستند. طبیعی است که کاربران دوست نداشته باشند که صندوق دریافتی آن‌ها شامل هرزنامه‌ها شود؛ پس در این شرایط نیاز به سیستمی است که پیام‌های متنی ورودی را به دو کلاس تقسیم کند. تشخیص هرزنامه‌ها شاید یکی از معروف‌ترین کاربردهای دسته‌بندی متن باشد اما قطعا تنها کاربرد آن نیستند!

هر متن دارای ویژگی‌هایی است؛ این ویژگی‌ها در روش‌های دسته‌بندی مختلف استفاده می‌شوند و به سبب آن‌ها امکان توسعه یک مدل دسته‌بند متن وجود خواهد داشت. برای بدست آوردن ویژگی‌های یک متن روش‌های گوناگونی وجود دارد. یکی از روش‌ها تهیه بردارهایی از متون است که هر بعد آن متناسب با یکی از کلمات موجود در دیکشنری باشد. بدین شکل که اگر متنی تعداد زیادی از کلمه اول را در خود داشته باشد، مقدار بعد اولش زیاد خواهد بود و بالعکس. این روش اگرچه در مقایسه با روش‌های عصبی روش جدیدی محسوب نمی‌شود ولی با این حال چندان قدیمی هم نیست و همچنان در شرایطی که داده‌ی کافی وجود نداشته باشد قابل استفاده هستند. همانطور که گفته شد در این روش‌ها به ازای هر کلمه در لغتنامه، یک ویژگی[۴] درنظر گرفته می‌شود و بدین ترتیب ابعاد فضای مسئله بسیار بالا خواهد بود. ابعاد بالای مسئله باعث خواهد شد که روش‌های مرسوم برای دسته‌بندی دچار مشکل جدی شوند.

برای حل مشکل ابعاد بالا یک راه حل استفاده از روش‌های انتخاب ویژگی[۵] است. در روش‌های انتخاب ویژگی متناسب با شرایط مسئله تعدادی از ویژگی‌ها انتخاب می‌شوند و مابقی ویژگی‌ها حذف می‌شوند. بدین ترتیب در یک فضای با ابعاد کمتر و ویژگی کمتر با سهولت بیشتر می‌تواند روش‌های دسته‌بندی متن را استفاده کرد. سوالی که باید به آن جواب داد این است که «چگونه می‌توان ویژگی‌های یک مسئله دسته‌بندی متن را انتخاب کرد؟»

در این پروژه تحقیقاتی، من سه روش جدید و معتبر [۳] [۲] [۱]که برای انتخاب ویژگی در مسائل دسته‌بندی معرفی شده‌اند را تبیین می‌کنم و تفاوت میان آن‌ها را مورد بررسی قرار خواهم داد. بدین ترتیب ابتدا در فصل دوم مفاهیم تئوری که برای درک روش‌های مذکور مورد نیاز است بیان خواهد شد. در فصل سوم و با تکیه به مفاهیم تئوری هر یک از سه روش در یک بخش مجزا تشریح می‌شود. در فصل چهارم ارزیابی و مقایسه‌ای میان سه روش صورت می‌گیرد. نهایتا در فصل پنجم جمع‌بندی و نتیجه‌گیری مطالب گفته‌شده در مقاله ارائه خواهد شد.

[1] Natural language processing(NLP)
[2] Text Classification
[3] Spam
[4] Feature
[5] Feature Selection

فصل دوم

مفاهیم تئوری

در این بخش قصد داریم در مورد مفاهیم تئوری که در روش‌های مورد بررسی این پروژه استفاده شده‌اند بپردازیم.

۲-۱ دسته‌بندی روش‌های انتخاب ویژگی

روش‌های انتخاب ویژگی به چندین دسته تقسیم می‌شوند. دو روش متداول و شناخته‌شده‌تر آن روش‌های فیلتر[۱] و پوشاننده[۲] هستند. در روش‌های پوشاننده مستقیما ویژگی‌های انتخاب‌شده را در یک مسئله واقعی که در اینجا یک مسئله دسته‌بندی متن است استفاده می‌کنند و لذا امتیازی که برای یک مجموعه ویژگی انتخاب‌شده بدست می‌آید امتیاز دقت واقعی برای مسئله دسته‌بندی است. در مقابل و در روش‌های فیلتر، با اعمال روش‌های آماری سعی می‌شود که یک امتیاز برای یک مجموعه ویژگی انتخاب‌شده حاصل گردد.

روش‌های پوشاننده چون به صورت مستقیم مجموعه ویژگی را بررسی می‌کند منجر به خروجی دقیق‌تری می‌شود؛ اما باید توجه داشت که روش‌های فیلتر زمان اجرای به مراتب بهتری دارند و بر روی مسئله دسته‌بندی بایاس نخواهند شد[۲]. در مسائل دسته‌بندی چون تعداد ویژگی‌ها بسیار بالاست نمی‌توان از روش‌های پوشاننده مستقیم استفاده کرد و لذا یا باید از روش‌های فیلتر استفاده کرد و یا به صورت ترکیبی از این دو شیوه بهره گرفت.

روش‌های فیلتر خود به چندین دسته قابل تقسیم هستند؛ نخست آنکه می‌توان این روش‌ها را به روش‌های محلی[۳] و روش‌های جهانی[۴] تقسیم کرد. در روش‌های جهانی به ویژگی یک امتیاز مطلق داده می‌شود اما در روش‌های محلی به هر ویژگی متناسب با هر کلاس یک امتیاز داده می‌شود؛ بعنی در روش‌های محلی مشخص است که یک ویژگی برای هر کلاس تا چه میزان خاصیت متمایزکننده دارد. در حالتی که از یک معیار محلی استفاده می‌شود می‌توان مشخص کرد که یک ویژگی می‌توان عضویت یک متن به یک کلاس را نشان دهد یا آنکه عدم عضویت را می‌تواند به خوبی نشان دهد. اگر عضویت را بتواند بهتر نشان دهد آن را یک ویژگی مثبت[۵] و در غیر این صورت آن را یک ویژگی منفی[۶] برای آن کلاس به حساب می‌آورند.[۳].

[1] Filter
[2] Wrapper
[3] Local
[4] Global
[5] Positive
[6] Negative

روش‌های فیلتر را می‌توان به دو دسته تک متغیره[۷] و چندمتغیره [۸] هم تقسیم کرد. در روش‌های تک متغیره هر ویژگی به صورت مستقل از سایر ویژگی‌ها امتیاز دریافت می‌کند، ولی در روش‌های چند متغیره در کنار آن که به شباهت ویژگی به هدف نگاه می‌شود، به زائد نبودن ویژگی‌ها نسبت به یکدیگر هم توجه می‌شود.[۲]

۲-۲ محاسبه روش‌های انتخاب ویژگی

در این بخش مهم‌ترین معیارهای انتخاب ویژگی معرفی می‌شوند و نحوه محاسبه آن‌ها ارائه می‌شود. این معیارها تماما جز روش‌های انتخاب ویژگی فیلتر هستند.

۲-۲-۱ بهره اطلاعاتی

بهره اطلاعاتی[۹] یکی از معیارهای محبوب برای انتخاب ویژگی در مقالات است [۲][۳]. نحوه محاسبه این معیار برای یک کلمه در رابطه ۲-۱ آمده است.

(۱-۲)
$$IG(t) = -\sum_{i=1}^{M} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{M} P(C_i|t)\log P(C_i|t) + P(\bar{t})\sum_{i=1}^{M} P(C_i|\bar{t})\log P(C_i|\bar{t})$$

در این رابطه $IG(t)$ به معنای مقدار بهره اطلاعاتی برای کلمه t است. M برابر با تعداد کلاس‌ها است. $P(C_i)$ احتمال کلاس C_i است؛ یعنی چه تعدادی از اسناد به این کلاس تعلق دارند. $P(t)$ احتمال مربوط به کلمه t است؛ یعنی آنکه چه تعدادی از اسناد شامل این کلمه هستند. به طور مشابه $P(\bar{t})$ معنای احتمال عدم این کلمه است؛ یعنی آنکه چه تعدادی از اسناد شامل این کلمه نیستند. $P(C_i|t)$ احتمال کلاس C_i به شرط کلمه t است؛ بدین معنا که چه تعدادی از اسناد شامل کلمه t به کلاس C_i تعلق دارند. به طور مشابه $P(C_i|\bar{t})$ هم تعریف می‌شود.

۲-۲-۲ شاخص جینی

شاخص جینی[۱۰] معیاری دیگر برای انتخاب ویژگی است که در مقالاتی مورد استفاده قرار گرفته است [۲][۳]. نحوه محاسبه این معیار در رابطه ۲-۲ آورده شده است.

[7] Univariate
[8] Multivariate
[9] Information Gain
[10] Gini index

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 P(C_i|t)^2 \qquad (۲-۲)$$

در این رابطه $GI(t)$ به معنای مقدار شاخص جینی برای کلمه t است. $P(t|C_i)$ احتمال شرطی کلمه t نسبت به کلاس C_i است؛ بدین تعریف که بررسی می‌کند که چه تعداد از اسناد متعلق به کلاس C_i دارای کلمه t هستند. سایر نمادهای این رابطه در بخش قبل تعریف شده است.

۲-۲-۳ نسبت نابرابری

نسبت نابرابری [11] معیاری است که برای انتخاب ویژگی در مقاله اویسال [12] استفاده شده است [۳] . نحوه محاسبه این معیار در رابطه ۲-۳ آورده شده است.

$$OR(t, C_i) = \log \frac{P(t|C_i)[1 - P(t|\bar{C}_i)]}{[1 - P(t|C_i)]P(t|\bar{C}_i)} \qquad (۳-۲)$$

در این رابطه $OR(t, C_i)$ نسبت نابرابری به ازای کلمه t و کلاس C_i محاسبه شده است. در کار تحقیقاتی اویسال برای جلوگیری از صفر شدن مخرج مقدار ۰/۰۱ به صورت و مخرج افزوده شده است [۳] .

۲-۲-۴ معیار زائدی کمینه شباهت بیشینه

معیار زائدی کمینه شباهت بیشینه [13] که با نماد $mRMR$ یک روش انتخاب ویژگی چند متغیره است که در مقاله لبنی و همکاران مورد استفاده قرار گرفته است [۲] . نحوه محاسبه این معیار در رابطه ۲-۴ آمده است.

$$mRMR(f_j) = I(f_j, C_k) - \frac{1}{|S| - 1} \sum_{f_i \in S} I(f_i, f_j) \qquad (۴-۲)$$

در این رابطه مجموعه S به معنی مجموعه ویژگی‌های انتخابی است. $I(a, b)$ به معنای اطلاعات متقابل [14] a و b است.

[11] Odds Ration

[12] Uysal

[13] Minimal redundancy maximal relevance

[14] Mutual information

اگر به منطق این رابطه نگاه کنیم، در می‌یابیم با این معیار به دنبال ویژگی‌های هستیم که با داده‌های یک کلاس ارتباط بالایی داشته باشند و با ویژگی‌هایی که در حال حاضر انتخاب شده‌اند شباهت پایین.

۲-۲-۵ معیار تمایزگر نسبی

معیار تمایزگر نسبی[15] یک روش انتخاب ویژگی برای مسائل دسته‌بندی دودویی است که در مقاله لبنی و همکاران[۲] مورد استفاده بدوه است. نحوه محاسبه این معیار در رابطه ۲-۵ آمده است.

$$RDC(t, tc_i(t)) = \frac{|df_{pos}(t) - df_{neg}(t)|}{\min(df_{pos}(t), df_{neg}(t)).tc_i(t)} \quad (۵-۲)$$

در این رابطه $RDC(t, tc_i(t))$ به معنای امتیاز تمایزگر نسبی یک کلمه t و سند i-ام است. $df_{pos}(t)$ و $df_{neg}(t)$ به ترتیب به معنای تعداد اسناد کلاس مثبت و کلاس منفی که شامل کلمه t هستند می‌شود. منظور از $tc_i(t)$ تعداد دفعات تکرار کلمه t در سند i-ام است. برای آنکه بتوان یک امتیاز نهایی به کلمه t نسبت داد باید تمام این امتیازها را باهم به نوعی جمع زد. مساحت زیر منحنی[16] مطابق رابطه ۲-۶ حاصل می‌شود. نهایتا $AUC(t, tc_i)$ به ازای آخرین سند به عنوان امتیاز نهایی اعلام خواهد شد.

$$\begin{cases} AUC(t, tc_1) = 0 \\ AUC(t, tc_i) = AUC(t, tc_{i-1}) + \frac{RDC(t,tc_i)+RDC(t,tc_{i+1})}{2} \end{cases} \quad (۶-۲)$$

۲-۳ الگوریتم ژنتیک

الگوریتم ژنتیک[17] یک الگوریتم تکاملی[18] است که با اقتباس از فرآیند تکامل موجودات زنده ارائه شده است. از آنجایی که این الگوریتم قسمت اصلی مقاله غارب و همکاران[۱] را تشکیل می‌دهد، در این قسمت به صورت مختصر توضیح داده می‌شود.

در الگوریتم ژنتیک ابتدا باید هر جوابی که برای مسئله وجود دارد را در قالب یک وضعیت بازنمایی[19] کرد. در این حالت هر وضعیت نقش کرومزوم یک شخص را خواهد داشت و ژن‌های این کرومزوم مرتبط با جزئیات آن وضعیت است. سپس باید یک تعداد زیادی فرد با کرومزوم اولیه ایجاد کرد؛ چیزی که به

[15] Relative discriminative criterion
[16] Area Under the Curve(AUC)
[17] Genetic algorithm
[18] Evolutionary algorithm
[19] Representation

آن جمعیت اولیه[۲۰] گفته می‌شود. در الگوریتم‌های ژنتیک لازم است تا یک تابع شایستگی[۲۱] تعریف شود. فردی که شایستگی بیشتری دارد باید مطابق با قانون تکامل شانس بیشتری برای زنده ماندن و تکثیر نسل داشته باشد. این چیزی است که در گام انتخاب والدین رخ می‌دهد. در گام انتخاب والدین، افراد با شایستگی بیشتر انتخاب خواهند شد. سپس هر دو والد دو فرزند را ایجاد می‌کنند که ژن این دو حاصل ترکیب ژن والدین است. نحوه ترکیب ژن والدین و ایجاد ژن فرزندان را بازترکیب[۲۲] گویند. نهایتا باید عمل جهش[۲۳] هم تعریف شود. در جهش برخی از ژن‌ها یک فرد تغییر می‌کند. پس از آنکه نسل جدید به وجود آمدند، نسل پیشین از بین می‌رود و الگوریتم ژنتیک با نسل جدید ادامه پیدا می‌کند تا جایی که یک شرط خاتمه برقرار شود. این شرط خاتمه می‌تواند تعداد نسل مشخص و یا همگرایی نسل‌ها باشد.

در ابتدای یک الگوریتم ژنتیک عملا تعدادی جواب اولیه برای مسئله داریم و در حین الگوریتم با نسل‌های جدید، جواب‌های موجود هم بهتر می‌شود؛ چراکه یک جواب مناسب در صورتی که تابع شایستگی به خوبی تعریف شده باشد، منجر به ایجاد جواب‌های بیشتری مبتنی بر خود می‌شود و جواب‌ها نامناسب کنار گذاشته می‌شوند. نهایتا آنکه عمل بازترکیب و جهش می‌توانند تنوع جواب‌ها را حفظ کنند و به وضعیت‌هایی برسیم که در ابتدا قابل ساختن نبوده است. برای آنکه یک الگوریتم برپایه ژنتیک معرفی شود لازم است تا گام‌های گفته شده طراحی شوند؛ یعنی به عنوان مثال مشخص باشد که عمل بازترکیب چگونه رخ می‌دهد.

[20] Initial population
[21] Fitness function
[22] Crossover
[23] Mutation

فصل سوم

روش‌های ارائه‌شده

در این فصل قرار است سه روش انتخاب ویژگی برای مسائل دسته‌بندی بررسی شود. لازم به ذکر است که در این فصل روش‌ها عینا مطابق با چیزی که در متن مقاله گفته شده است بیان نشده است؛ یعنی آنکه برخی از جزئیات حذف شده است و ممکن است نحوه بیان برخی از قسمت‌های روش تغییر یافته باشد. با تمام این‌ها ایده و خروجی روش‌ها کاملا منطبق بر چیزی است که در مقالات بیان شده است.

۳-۱ روش IGFSS

روش IGFSS توسط اویسال[۳] معرفی شده است و این بخش بر اساس مقاله وی تبیین شده است. ابتدا این روش را معرفی می‌کنیم و سپس مثالی برای اجرای این الگوریتم در ادامه خواهیم آورد.

۳-۱-۱ مراحل الگوریتم

این الگوریتم از چهار گام تشکیل شده است:

۱. برچسب‌گذاری ویژگی‌ها: در این گام برای هر ویژگی یک امتیاز انتخاب ویژگی محلی نسبت به هر کلاس محاسبه می‌شود. هر کدام از این ویژگی‌ها عضویت یا عدم عضویت یک کلاس نسبت به سایر کلاس‌ها را بهتر نمایش می‌دهد. در این مرحله با یک برچسب شماره کلاس و عضویت یا عدم عضویت یک ویژگی را مشخص می‌کنیم.

۲. انتخاب ویژگی جهانی: این بار با یک شاخص انتخاب ویژگی جهانی برای هر ویژگی امتیاز آن را محاسبه می‌کنیم و لیست را بر اساس این امتیاز مرتب می‌کنیم.

۳. ساخت مجموعه ویژگی: فرض کنید که اندازه مجموعه ویژگی‌های انتخاب شده برابر با fs باشد. همچنین فرض کنید که نسبت تعداد ویژگی‌های منفی به کل ویژگی‌ها برابر با $nfrs$ باشد. در این مرحله از ابتدای لیستی که در گام قبل ساخته شده است به سمت انتهای لیست حرکت می‌کنیم. برای هر کلاس و با توجه به برچسب‌هایی که در گام اول مشخص کردیم ویژگی‌ها با بیشترین امتیاز جهانی را انتخاب می‌کنیم و در عین حال باید نسبت ویژگی‌های منفی و مثبت رعایت شود.

۴. بخش شرطی: چنانچه اندازه مجموعه ویژگی‌های انتخاب شده کمتر از fs باشد، لازم است تا تعدادی ویژگی به مجموعه اضافه شود. این ویژگی‌ها را بر اساس معیار انتخاب ویژگی جهانی انتخاب می‌شوند. یعنی ویژگی‌ها با بیشترین امتیاز که تا به الان انتخاب نشده‌اند به مجموعه ویژگی‌های انتخاب‌شده افزوده می‌شوند تا به اندازه مورد نظر برسیم.

۳-۱-۲ مثال و تحلیل

برای درک بهتر از نحوه اجرای الگوریتم بهتر است تا یک مثال را مورد بررسی قرار دهیم.[۳] در جدول ۳-۱ یک مجموعه‌داده کوچک شامل محتوا و کلاس اسناد آورده شده است.

جدول ۳-۱: مجموعه‌داده نمونه برای روش IGFSS

شماره سند	محتوای سند	کلاس
۱	موش گربه گرگ	C_1
۲	موش گربه اسب سگ	C_2
۳	موش گربه سگ مرغ اسب	C_2
۴	خفاش گاو اردک اسب پلیکان	C_3
۵	خفاش گاو اسب پلیکان	C_3
۶	خفاش گاو شتر اسب مرغ	C_3

جدول ۳-۲: امتیاز معیارهای انتخاب ویژگی برای روش IGFSS

ویژگی	امتیاز شاخص جینی	امتیاز نسبت نابرابری کلاس‌ها	برچسب ویژگی
خفاش	۱	۴/۶۱۵۱ ، -۴/۳۳۰۷ ، -۴/۱۱۰۹	مثبت C_3
گاو	۱	۴/۶۱۵۱ ، -۴/۳۳۰۷ ، -۴/۱۱۰۹	مثبت C_3
سگ	۱	-۴/۲۱۴۶ ، ۴/۶۱۵۱ ، -۳/۷۱۳۶	مثبت C_2
گرگ	۱	-۳/۵۳۶۱ ، -۳/۲۵۸۱ ، ۴/۶۱۵۱	مثبت C_1
گربه	۰/۵۵۵۶	-۴/۶۱۵۱ ، ۴/۳۳۰۷ ، ۴/۱۱۰۹	منفی C_3
موش	۰/۵۵۵۶	-۴/۶۱۵۱ ، ۴/۳۳۰۷ ، ۴/۱۱۰۹	منفی C_3
اسب	۰/۵۲۰۰	۳/۵۳۶۱ ، ۳/۲۵۸۱ ، -۴/۶۱۵۱	منفی C_1
پلیکان	۰/۴۴۴۴	۳/۸۱۶۵ ، -۳/۹۳۱۸ ، -۳/۷۱۳۶	منفی C_2
اردک	۰/۱۱۱۱	۲/۴۹۴۱ ، -۳/۲۵۸۱ ، -۳/۰۴۴۵	منفی C_2
شتر	۰/۱۱۱۱	۲/۴۹۴۱ ، -۳/۲۵۸۱ ، -۳/۰۴۴۵	منفی C_2
مرغ	۰/۰۹۰۳	-۱/۲۹۲۹ ، ۰ ، -۳/۷۱۳۶	منفی C_1

بر اساس مجموعه‌داده معرفی‌شده می‌توان معیارهای انتخاب ویژگی مرتبط را بدست آورد و برچسب‌گذاری پیشنهادی در گام اول الگوریتم را انجام داد. خروجی این موارد در جدول ۳-۲ آورده شده است.

۳-۲ روش MRDC

روش MDRC توسط لبنی و همکاران [۲] ارائه شده است. مانند قسمت قبل ابتدا روش را تشریح می‌کنیم و سپس سعی می‌کنیم در قالب یک مثال تحلیل اولیه از آن داشته باشیم.

جدول ۳-۳: تفاوت روش سنتی با روش IGFSS برای مثال ارائه‌شده

روش	مجموعه ویژگی‌های انتخاب‌شده	C_1	C_2	C_3
روش سنتی برپایه شاخص جینی	خفاش، گاو، سگ، گرگ، گربه و موش	۱	۱	۴
روش IGFSS	خفاش، سگ، گرگ، گربه، اسب و پلیکان	۲	۲	۲

۳-۲-۱ مراحل الگوریتم

۱. پیش‌پردازش: به طور خلاصه پردازش‌های زیر بر روی داده‌ها انجام می‌شود:

- حذف ایست‌واژه‌ها[1]: برخی از کلمات نظیر حروف اضافه در غالب اسناد به تعداد بالا یافت می‌شود و لذا دانش مفیدی برای دسته‌بندی متون نداند که بهتر است حذف شوند.
- حذف کلمات نادر: تعدادی از کلمات هستند که تنها در تعداد بسیار کمی از اسناد ظاهر می‌شوند. مطابق با قانون Zipf تعداد این کلمات بسیار زیاد است و حذف آن باعث کاهش چشمگیر تعداد ویژگی‌ها می‌شود. در روش مقاله کلماتی که در کمتر از چهار سند آمده‌اند را حذف کرده‌اند.
- ریشه‌یابی[2]: خیلی از کلمات هستند که به طریق مختلف نوشته می‌شوند ولی به یک کلمه مرتبط هستند؛ به عنوان مثال کلمات «می‌روم»، «رفت»، «بروید» تماما ریشه یکسانی دارند. در روش پیشنهادی نیز از ریشه‌یابی استفاده شده است.

۲. محاسبه امتیاز جهانی: در گام بعد برای تمام ویژگی‌های باقی مانده امتیاز ویژگی مطابق با معیار تمایزگر نسبی محاسبه می‌شود.

۳. انتخاب ویژگی‌ها: در این گام سعی در انتخاب ویژگی‌هایی است که هم امتیاز جهانی بالایی داشته باشند و هم آنکه همبستگی کمی با یکدیگر داشته باشند. مجموعه S مجموعه ویژگی‌های انتخاب‌شده نهایی است. در ابتدا این مجموعه با ویژگی‌ای که بیشترین امتیاز جهانی را داشته باشد تشکیل می‌شود. سپس به صورت تکرارشونده ویژگی که دارای بالاترین امتیاز $MRDC$ باشد به مجموعه S افزوده می‌شود تا مجموعه S به اندازه مدنظر برسد. نحوه محاسبه معیار $MRDC$ به ازای ویژگی f_i در رابطه ۳-۱ آورده شده است.

$$MRDC(f_i) = RDC(f_i) - \sum_{f_i \neq j_j, f_j \in S} correlation(f_i, f_j) \qquad (۱-۳)$$

۳-۲-۲ مثال و تحلیل

برای درک بهتر این الگوریتم یک مثال از نحوه اجرای الگوریتم بررسی می‌شود.[۲] در جدول ۳-۴ یک مجموعه‌داده نمونه برای این روش ارائه شده است. در این مجموعه‌داده چهار کلمه و طبیعتا چهار ویژگی وجود دارد. فرض کنید ویژگی‌های «گربه»، «ماهی»، «موش» و «سگ» به ترتیب ویژگی‌های f_1، f_2، f_3، f_4 باشند. همچنین برای نشان دادن کارایی الگوریتم فرض کنید ویژگی f_5 نیز به ویژگی «ماهی»

[1] Stop word
[2] Stemming

جدول ۳-۴: مجموعه‌داده نمونه برای روش MRDC

شماره سند	محتوای سند	کلاس
۱	گربه ماهی	C_1
۲	گربه موش ماهی	C_1
۳	موش ماهی	C_1
۴	موش گربه ماهی موش ماهی	C_1
۵	ماهی گربه ماهی گربه	C_1
۶	ماهی موش	C_1
۷	سگ موش	C_2
۸	سگ سگ	C_2
۹	ماهی ماهی موش	C_2
۱۰	موش	C_2
۱۱	گربه ماهی	C_2
۱۲	سگ ماهی	C_2

جدول ۳-۵: مقایسه دو معیار تمایزگر نسبی و MDRC برای مجموعه داده نمونه

روش	f_1	f_2	f_3	f_4	f_5
تمایزگر نسبی	۶	۱	۵	۱۵	۱
MDRC	۵/۹۰۲	-۰/۱۹۳	۴/۶۳	۱۵	-۰/۸۴

یعنی همان ویژگی ۲ اشاره داشته باشد. چنانچه برای هر ویژگی مقدار معیار تمایزگر نسبی و مقدار معیار MRDC را بدست بیاوریم به اعداد موجود در جدول ۳-۵ می‌رسیم. همانطور که در جدول ۳-۵ هم مشخص است دو ویژگی کاملا یکسان f_2 و f_5 امتیاز $MDRC$ یکسانی نخواهند داشت و نسبت به یکدیگر زائد محسوب می‌شوند.

۳-۳ روش برپایه الگوریتم ژنتیک

در کار تحقیقاتی غارب و همکاران[۱] برای انتخاب ویژگی‌های مسائل دسته‌بندی از روشی مبتنی بر الگوریتم ژنتیک بهره گرفتند. این بخش این روش را تشریح می‌کند.

۳-۳-۱ شناسنامه الگوریتم ژنتیک

مانطور که در فصل قبل در مورد الگوریتم‌های ژنتیک توضیح دادیم، برای ارائه یک الگوریتم بر پایه ژنتیک باید گام‌ها و توابع موجود در آن را به طور دقیق تعریف کرد. توابع و جزئیاتی پیشنهادی آنان به شرح زیر است:

۱. بازنمایی: هر ژن در یک کرومزوم متناسب با ویژگی است. در صورتی که مقدار آن صفر باشد

یعنی آن ویژگی انتخاب نشده است و اگر مقدار آن یک باشد یعنی ویژگی انتخاب شده است.

۲. جمعیت اولیه: برای ساخت جمعیت اولیه به صورت کاملا تصادفی کروموزم‌ها ساخته می‌شود.

۳. تابع شایستگی: تابع شایستگی در این مقاله به دو هدف اهمیت می‌دهد؛ اول آنکه مجموعه ویژگی انتخاب شده باید برای دسته‌بندی مناسب باشد و دوم آنکه باید حتی الامکان اندازه آن کوچک باشد. در رابطه ۳-۲ تابع شایستگی آورده شده است. پارامتر z برای تنظیم نسبت اهمیت دو مولفه گفته شده است. در مقاله از عدد ۰/۸ برای آن استفاده کرده‌اند. $c(s_i)$ امتیاز مجموعه ویژگی را مشخص می‌کند.

$$fitness(s_i) = z.c(s_i) + (1 - z).\frac{1}{|s_i|} \qquad (۳-۲)$$

۴. انتخاب: انتخاب افراد برتر باتوجه به امتیاز شایستگی تعیین می‌شود. مطابق با رابطه ۳-۳ احتمال انتخاب هر فرد تعیین می‌شود.

$$p(s_i) = \frac{fitness(s_i)}{\sum_{i=1}^{n} fitness(s_i)} \qquad (۳-۳)$$

۵. باز ترکیبی: برای بازترکیبی، هر کرومزوم والد به دو بخش کاملا مساوی تقسیم می‌شود. سپس بر اساس وزن‌های TF-IDF مشخص می‌شود که هر بخش از هر کرومزوم والد دارای چه مجموع وزنی است. سپس یک فرزند را از دو قسمتی می‌سازند که بیشترین وزن ممکن به وجود آید و یک فرزند را از دو بخش باقی‌مانده.

۶. جهش: برای جهش در روش پیشنهادی مقاله، ابتدا بررسی می‌شود که آیا امتیاز والدین یک فرزند از یک حد آستانه‌ای پایین‌تر است یا خیر. اگر پایین‌تر بود ژن‌های فرزند باید تغییر کند. برای جهش، تعدادی از ویژگی‌ها با پایین‌ترین وزن حذف می‌شود و به جای آن ویژگی‌ها با اهمیت بالا در بهترین کرومزوم نسل قبل جایگزین می‌شود.

۳-۳-۲ مراحل الگوریتم

روش پیشنهادی غارب و همکاران در دو گام اصلی انجام می‌گیرد:

۱. انتخاب ویژگی‌های برتر: در این گام و با کمک معیارهای انتخاب ویژگی با نگرش روش‌های فیلتر، بهترین ویژگی‌ها انتخاب می‌شود. این ویژگی‌ها ویژگی نهایی نیست؛ بلکه در این گام سعی شده

است تا تعداد ویژگی‌های اصلی که بسیار زیاد است را به تعداد معقولی کاهش دهد تا اجرای یک الگوریتم ژنتیک امکان‌پذیر باشد.

۲. اجرای الگوریتم ژنتیک: در این گام مطابق با توضیحات بخش قبل الگوریتم ژنتیک اجرا می‌شود. در اینجا لازم به ذکر است که برای محاسبه مناسب بودن یک مجموعه ویژگی از روش‌های پوشاننده استفاده می‌شود. نهایتا در خروجی این گام یک مجموعه ویژگی نهایی حاصل می‌گردد.

فصل چهارم

ارزیابی و مقایسه

در این فصل قصد داریم ارزیابی از دقت‌های گزارش‌شده در مقالات روش‌های ارائه‌شده را بیاوریم. همچنین سعی می‌کنیم مقایسه‌ای میان روش‌ها داشته باشیم.

۴-۱ مقایسه پیچیدگی زمانی

روش بر پایه ژنتیک غارب و همکاران [۱] پیچدگی زمانی بیشتری نسبت به دو روش دیگر دارد. در این روش در مرحله اول با کمک شش معیار انتخاب ویژگی فیلتر تعدادی از ویژگی‌های مناسب‌تر انتخاب می‌شود و سپس در مرحله بعد یک الگوریتم ژنتیک آن هم با معیار انتخاب ویژگی پوشاننده استفاده می‌شود. دو روش دیگر تنها از یک معیار انتخاب ویژگی فیلتر استفاده کرده‌اند که حتی می‌توان گفت پیچیدگی کمتری نسبت به زمان مرحله اول روش بر پایه ژنتیک دارد. به علاوه در روش برپایه ژنتیک مرحله دوم پیچیدگی زمانی زیادی را دارد؛ چراکه روش‌های ژنتیک و روش‌های پوشاننده روش‌های کندی هستند.

حال باید دو روش دیگر را مقایسه کرد. در روش IGFSS آیسال یک بار امتیاز یک معیار جهانی و یک بار امتیاز یک معیار محلی حساب می‌شود. سپس در بدترین حالت دو بار باید لیست ویژگی‌ها را پیمایش کرد؛ یک بار هنگام تشکیل مجموعه ویژگی‌های انتخاب‌شده اولیه و بار دیگر در مرحله بخش شرطی و رساندن تعداد ویژگی‌ها به یک اندازه خاص.[۳] در روش MRDC لبنی و همکاران یک بار برای تمام ویژگی‌ها معیار تمایزگر نسبی را حساب می‌کنند و سپس نیاز است تا مقدار MDRC حساب شود که محاسبه Correlation اصلی‌ترین قسمت آن است.[۲] در این شرایط به نظر می‌رسد که روش IGFSS روش سریع‌تری است چرا که لازم نیست تا دو ویژگی نسبت به هم سنجیده شوند و در نتیجه پیچیدگی آن در شرایطی که ابعاد مسئله بسیار بالاست به $O(|F|^2)$ نمی‌رسد ولی پیچیدگی زمانی MDRC از $O(|F|^2)$ بیشتر است.

۴-۲ مقایسه پیچیدگی حافظه

از منظر حافظه‌ی مورد نیاز الگوریتم هم باز روش برپایه ژنتیک به حافظه بیشتری نیاز دارد؛ چراکه در مرحله دوم که قرار است الگوریتم ژنتیک اجرا شود به تعداد اعضای هر نسل باید مجموعه‌ای از ویژگی‌ها نگهداری شود. دو الگوریتم دیگر از نظر حافظه تفاوت چندانی بای یکدیگر ندارند.

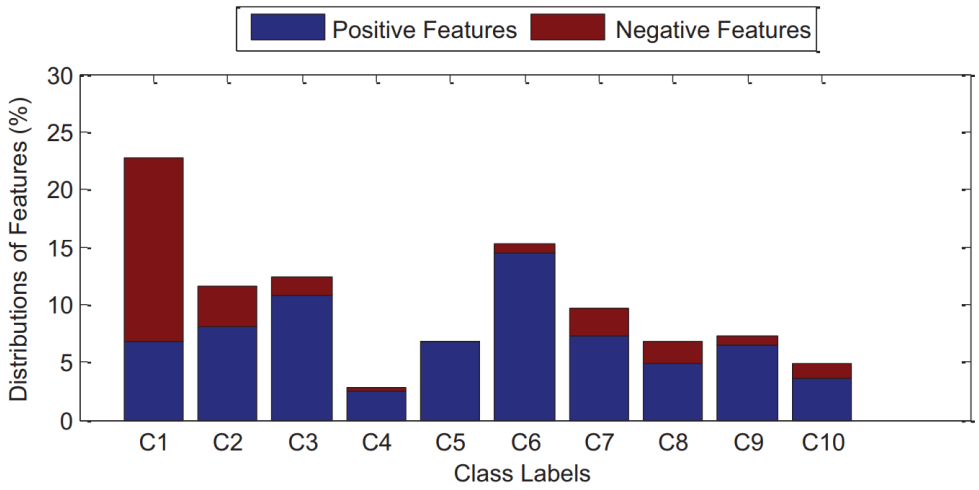
۴-۳ مقایسه دقت

در این پروژه پیاده‌سازی‌ای از الگوریتم‌ها تهیه نشده است و در عین حال پیاده‌سازی آماده‌ای هم برای این‌ها در دسترس نبوده است؛ لذا برای مقایسه دقت مستقیما به اعداد مقاله مراجعه کردم. اما اعداد در

مقاله امکان مقایسه دقیق و عادلانه را ندارد. چراکه غارب و همکاران از پیکره‌های عربی استفاده کرده‌اند. آیسول و لبنی و همکاران از تعدادی پیکره استفاده کرده‌اند که برخی از آن‌ها مشترک است ولی با این حال تنظیمات متفاوت که اعمال کرده‌اند باعث می‌شود که همچنان مقایسه عادلانه‌ای را نتوان انجام داد. برای این دو روش نتایج بر روی مجموعه‌داده رویترز را گزارش خواهیم کرد. این مجموعه‌داده هم در در روش مشترک است و هم آنکه قسمت آموزش و ارزیابی آن توسط خود مجموعه‌داده تعیین شده است. در اینجا بنا به محدودیت فقط همین مورد بررسی می‌شود و برای دیدن سایر نتایج می‌توانید به خود مقالات مراجعه کنید. مجموعه‌داده رویترز شامل ده کلاس است.

۴-۳-۱ دقت روش IGFSS

یکی از مشکلاتی که در کار تحقیقاتی اویسال به آن اشاره شده است این است که معیارهای سنتی به تعداد ویژگی هر کلاس و نسبت ویژگی‌های منفی اهمیت نمی‌دهند. این مورد در تصویر ۱ به خوبی پیدا است. در این تصویر توزیع ویژگی‌ها برای معیار شاخص جینی آورده شده است.



شکل ۴-۱: فراوانی ویژگی‌های انتخاب‌شده نسبت به هر کلاس برای شاخص جینی در روش IGFSS [۳]

در جدول ۴-۱ و ۴-۲ به ترتیب دقت مربوط به روش‌های مختلف انتخاب ویژگی برای دسته‌بند SVM و Naive bayes بدون استفاده از روش IGFSS و با استفاده از آن آورده شده است. با بررسی کلی در می‌یابیم که استفاده از روش پیشنهادی در مقاله منجر به بهبود روش پایه می‌شود اما این بهبود چندان موثر نیست و در هیچ یک از موارد شاهد بیش از ۲ درصد بهبود نیستیم.

جدول ۴-۱: معیار F_1 برای روش‌های پایه و IGFSS برای دسته‌بند SVM [۳]

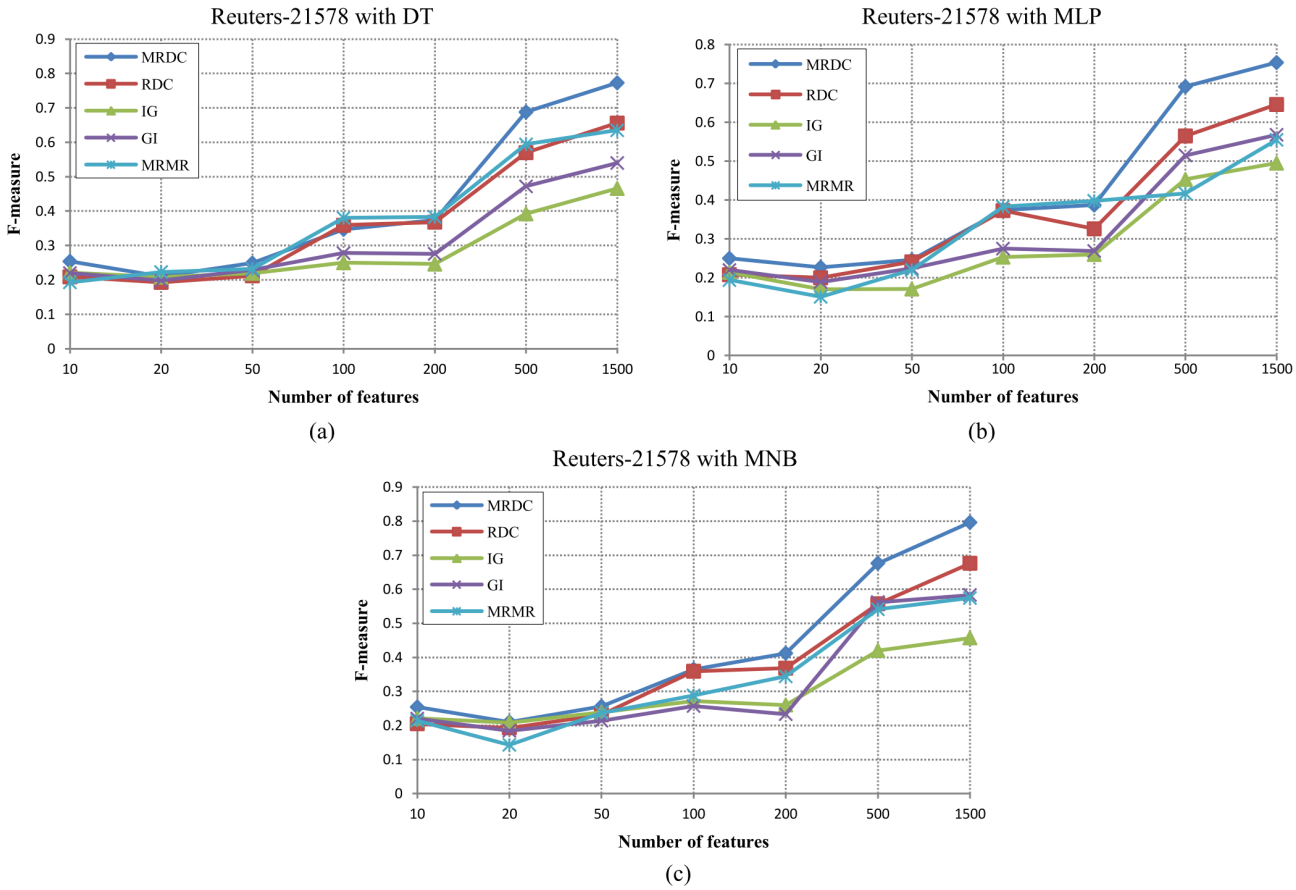
روش	nfr	۲۵۰	۳۰۰	۳۵۰	۴۰۰	۴۵۰	۵۰۰
IG	–	۸۵/۷۵۵	۸۶/۰۰۶	۸۶/۰۰۶	۸۵/۸۶۳	۸۶/۰۰۶	۸۵/۸۲۷
IG+IGFSS	۰/۶	۸۵/۳۶۱	۸۶/۴۷۳	۸۶/۱۵۰	۸۶/۲۹۴	۸۶/۱۱۴	۸۶/۰۰۶
GI	–	۸۵/۹۳۵	۸۵/۹۷۱	۸۶/۰۰۶	۸۶/۴۰۱	۸۶/۰۷۸	۸۶/۴۳۷
GI+IGFSS	۰/۳	۸۵/۶۴۸	۸۵/۷۹۱	۸۶/۳۲۹	۸۶/۴۳۷	۸۶/۷۶۰	۸۵/۹۳۵
DFS	–	۸۵/۸۹۹	۸۵/۸۹۹	۸۵/۹۷۱	۸۵/۷۹۱	۸۵/۸۹۹	۸۵/۷۹۱
DFS+IGFSS	۰/۸	۸۵/۰۰۲	۸۶/۲۵۸	۸۶/۴۷۳	۸۶/۲۵۸	۸۶/۱۱۴	۸۵/۸۶۳

جدول ۴-۲: معیار F_1 برای روش‌های پایه و IGFSS برای دسته‌بند NB [۳]

روش	nfr	۲۵۰	۳۰۰	۳۵۰	۴۰۰	۴۵۰	۵۰۰
IG	–	۸۳/۵۳۱	۸۲/۳۸۲	۸۲/۳۸۲	۸۲/۵۶۲	۸۱/۹۱۶	۸۱/۷۳۷
IG+IGFSS	۰/۶	۸۴/۱۰۵	۸۴/۲۸۴	۸۴/۳۲۰	۸۴/۲۱۲	۸۴/۵۳۵	۸۴/۰۳۳
GI	–	۸۴/۵۳۵	۸۴/۲۱۲	۸۳/۹۶۱	۸۴/۱۴۱	۸۳/۶۷۴	۸۳/۴۲۳
GI+IGFSS	۰/۳	۸۵/۱۰۹	۸۵/۴۶۸	۸۴/۸۲۲	۸۴/۹۶۶	۸۴/۳۵۶	۸۴/۵۷۱
DFS	–	۸۴/۹۳۰	۸۴/۲۸۴	۸۴/۰۳۳	۸۳/۸۸۹	۸۳/۶۰۲	۸۳/۱۰۰
DFS+IGFSS	۰/۸	۸۴/۶۰۷	۸۵/۱۸۱	۸۵/۲۸۹	۸۴/۶۷۹	۸۴/۷۸۷	۸۴/۷۵۱

۴-۳-۲ دقت روش MDRC

در تصویر ۴-۲ دقت متناسب با معیار F_1 برای روش‌های مختلف پایه به همراه روش MDRC برای سه روش دسته‌بندی آورده شده است. از این نمودارها می‌توان دریافت که به ازای تعداد ویژگی کم این روش برتری جدی‌ای نسبت به روش‌های پیشین ندارد اما وقتی تعداد ویژگی‌ها بیشتر می‌شود برتری آن نسبت به سایر روش‌ها کاملا حس می‌شود. همچنین می‌توان دید که روش MDRC نسبت به سایر روش‌ها برای حالات بیش‌تر از ۵۰۰ ویژگی حداقل ۱۰ درصد بهبود دارد. این بهبود واقعا قابل ملاحظه است و چیزی است که در روش IGFSS مشاهده نشده بود؛ لذا می‌توان گفت که به نظر می‌رسد روش MDRC دقت بهتری نسبت به روش IGFSS دارد.



شکل ۴-۲: امتیاز معیار F_1 برای روش‌های مختلف انتخاب ویژگی و روش MDRC و روش‌های دسته‌بندی (a) درخت تصمیم (b) روش MLP (c) روش MNB [۲]

فصل پنجم

جمع‌بندی و نتیجه‌گیری

منابع و مراجع

[1] Ghareb, Abdullah Saeed, Bakar, Azuraliza Abu, and Hamdan, Abdul Razak. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications, 49:31–47, 2016.

[2] Labani, Mahdieh, Moradi, Parham, Ahmadizar, Fardin, and Jalili, Mahdi. A novel multivariate filter method for feature selection in text classification problems. Engineering Applications of Artificial Intelligence, 70:25–37, 2018.

[3] Uysal, Alper Kursat. An improved global feature selection scheme for text classification. Expert systems with Applications, 43:82–92, 2016.