



دانشگاه صنعتی امیرکبیر

(پلی تکنیک 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

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

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

صفحه

عنوان

۲	مفاهیم تئوری	۲
۳	۱-۲ روش‌های انتخاب ویژگی	۳
۳	۱-۱-۲ بهره اطلاعاتی	۳
۳	۲-۱-۲ شاخص جینی	۳
۴	۳-۱-۲ نسبت نابرابری	۴
۴	۴-۱-۲ معیار زائدی کمینه شباهت بیشینه	۴
۴	۵-۱-۲ معیار تمایزگر نسبی	۴
۵	۲-۲ دسته‌بندی روش‌های انتخاب ویژگی	۵
۵	۳-۲ الگوریتم ژنتیک	۵
۶	۳ روش‌های ارائه شده	۶
۷	۱-۳ روش IGFSS	۷
۷	۱-۱-۳ مراحل الگوریتم	۷
۷	۲-۱-۳ مثال و تحلیل	۷
۸	۲-۳ روش MRDC	۸
۹	۱-۲-۳ مراحل الگوریتم	۹
۹	۲-۲-۳ مثال و تحلیل	۹
۹	۳-۳ روش برپایه الگوریتم ژنتیک	۹
۱۰	۱-۳-۳ شناسنامه الگوریتم ژنتیک	۱۰
۱۱	۲-۳-۳ مراحل الگوریتم	۱۱
۱۲	۴ ارزیابی و مقایسه	۱۲
۱۳	۱-۴ مقایسه تئوری	۱۳
۱۴	۵ جمع‌بندی و نتیجه‌گیری	۱۴
۱۵	منابع و مراجع	۱۵

صفحه

فهرست اشکال

شکل

صفحه	فهرست جداول	جدول
۸	مجموعه داده نمونه برای روش IGFSS	۱-۳
۸	امتیاز معیارهای انتخاب ویژگی برای روش IGFSS	۲-۳
۸	تفاوت روش سنتی با روش IGFSS برای مثال ارائه شده	۳-۳
۱۰	مجموعه داده نمونه برای روش MRDC	۴-۳

فصل اول

مقدمه

فصل دوم

مفاهیم تئوری

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

۱-۲ روش‌های انتخاب ویژگی

۱-۱-۲ بهره اطلاعاتی

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

$$(1-2) \quad IG(t) = - \sum_{i=1}^M P(C_i) \log P(C_i) + P(t) \sum_{i=1}^M P(C_i|t) \log P(C_i|t) + P(\bar{t}) \sum_{i=1}^M P(C_i|\bar{t}) \log P(C_i|\bar{t})$$

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 $P(\bar{t})$ به 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.

۲-۱-۲ شاخص جینی

شاخص جینی^۲ معیاری دیگر برای انتخاب ویژگی است که در مقالاتی مورد استفاده قرار گرفته است [۱][۲]. نحوه محاسبه این معیار در رابطه ۲-۲ آورده شده است.

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

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.

¹Information Gain

²Gini index

۳-۱-۲ نسبت نابرابری

نسبت نابرابری^۳ معیاری است که برای انتخاب ویژگی در مقاله اویسال^۴ استفاده شده است [۲]. نحوه محاسبه این معیار در رابطه ۳-۲ آورده شده است.

$$OR(t, C_i) = \log \frac{P(t|C_i)[1 - P(t|\bar{C}_i)]}{[1 - P(t|C_i)]P(t|\bar{C}_i)} \quad (3-2)$$

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 $0/01$ به صورت و مخرج افزوده شده است [۲].

۴-۱-۲ معیار زائدی کمینه شباهت بیشینه

معیار زائدی کمینه شباهت بیشینه^۵ که با نماد $mRMR$ یک روش انتخاب ویژگی چند متغیره است که در مقاله لبنی و همکاران مورد استفاده قرار گرفته است [۱]. نحوه محاسبه این معیار در رابطه ۴-۲ آمده است.

$$mRMR(f_j) = I(f_j, C_k) - \frac{1}{|S| - 1} \sum_{f_i \in S} I(f_i, f_j) \quad (4-2)$$

در این رابطه مجموعه S به معنی مجموعه ویژگی‌های انتخابی است. $I(a, b)$ به معنای اطلاعات متقابل^۶ a و b است.

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.

۵-۱-۲ معیار تمایزگر نسبی

معیار تمایزگر نسبی^۷ یک روش انتخاب ویژگی برای مسائل دسته‌بندی دودویی است که در مقاله لبنی و همکاران [۱] مورد استفاده بدو است. نحوه محاسبه این معیار در رابطه ۵-۲ آمده است.

³Odds Ration

⁴Uysal

⁵Minimal redundancy maximal relevance

⁶Mutual information

⁷Relative discriminative criterion

$$RDC(t, tc_i(t)) = \frac{|df_{pos}(t) - df_{neg}(t)|}{\min(df_{pos}(t), df_{neg}(t)) \cdot tc_i(t)} \quad (5-2)$$

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^۸ مطابق رابطه ۶-۲ حاصل می‌شود. نهایتاً $AUC(t, tc_i)$ به ازای آخرین سند به عنوان امتیاز نهایی اعلام خواهد شد.

$$\begin{cases} AUC(t, tc_1) = 0 \\ AUC(t, tc_i) = AUC(t, tc_{i-1}) + \frac{RDC(t, tc_i) + RDC(t, tc_{i+1})}{2} \end{cases} \quad (6-2)$$

۲-۲ دسته‌بندی روش‌های انتخاب ویژگی

۳-۲ الگوریتم ژنتیک

^۸Area Under the Curve(AUC)

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 عیناً 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 کاملاً منطبق بر چیزی است که در مقالات بیان شده است.

۱-۳ روش IGFSS

روش IGFSS توسط اويسال [۲] 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.

۱-۱-۳ 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 f_s باشد. همچنین فرض کنید که نسبت تعداد ویژگی‌های منفی به کل ویژگی‌ها برابر با n_{frs} 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 را 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 f_s باشد، لازم است تا تعدادی ویژگی به مجموعه اضافه شود. این ویژگی‌ها را بر اساس معیار انتخاب ویژگی جهانی انتخاب می‌شوند. یعنی ویژگی‌ها را با بیشترین امتیاز که تا به الان انتخاب نشده‌اند به مجموعه ویژگی‌های انتخاب‌شده افزوده می‌شوند تا به اندازه مورد نظر برسیم.

۲-۱-۳ مثال و تحلیل

برای درک بهتر از نحوه اجرای الگوریتم بهتر است تا یک مثال را مورد بررسی قرار دهیم. [۲] در جدول ۱-۳ یک مجموعه داده کوچک شامل محتوا و کلاس اسناد آورده شده است.

جدول ۱-۳: مجموعه داده 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 ۲-۳: امتیاز معیارهای انتخاب ویژگی برای روش IGFSS

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

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

۲-۳ روش MRDC

روش توسط لبنی و همکاران [۱] 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

۱-۲-۳ مراحل الگوریتم

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

- حذف ایست‌واژه‌ها^۱: 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^۲: خیلی از کلمات هستند که به طریق مختلف نوشته می‌شوند ولی به یک کلمه مرتبط هستند؛ به عنوان مثال کلمات «می‌روم»، «رفت»، «بروید» تماماً 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 ۱-۳ آورده شده است.

$$MRDC(f_i) = RDC(f_i) - \sum_{f_i \neq f_j, f_j \in S} correlation(f_i, f_j) \quad (1-3)$$

۲-۲-۳ مثال و تحلیل

۳-۳ روش برپایه الگوریتم ژنتیک

در کار تحقیقاتی غارب و همکاران[۹] برای انتخاب ویژگی‌های مسائل دسته‌بندی از روشی مبتنی بر الگوریتم ژنتیک بهره گرفتند. این بخش این روش را تشریح می‌کند. ه

^۱Stop word

^۲Stemming

جدول ۳-۴: مجموعه داده 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

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 کروموزوم 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 کاملاً 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 $0/8$ برای آن استفاده کرده‌اند. $c(s_i)$ امتیاز مجموعه ویژگی را مشخص می‌کند.

$$fitness(s_i) = z.c(s_i) + (1 - z) \cdot \frac{1}{|s_i|} \quad (2-3)$$

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

$$p(s_i) = \frac{fitness(s_i)}{\sum_{i=1}^n fitness(s_i)} \quad (3-3)$$

۵. باز ترکیبی: برای باز ترکیبی، هر کروموزوم والد به دو بخش کاملاً مساوی تقسیم می‌شود. سپس بر اساس وزن‌های TF-IDF مشخص می‌شود که هر بخش از هر کروموزوم 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 کروموزوم 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. نهایتاً در خروجی این گام یک مجموعه ویژگی نهایی حاصل می‌گردد.

فصل چهارم

ارزیابی و مقایسه

۱-۴ مقایسه تئوری

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, Mahdiah, 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.