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

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.

۲-۱ روش‌های انتخاب ویژگی

۲-۱-۱ بهره اطلاعاتی

بهره اطلاعاتی[1] یکی از معیارهای محبوب برای انتخاب ویژگی در مقالات است [۱][۲]. نحوه محاسبه این معیار برای یک کلمه در رابطه ۲-۱ آمده است.

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

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.

۲-۱-۲ شاخص جینی

شاخص جینی[2] 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.

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

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.

[1] Information Gain

[2] Gini index

۲-۱-۳ نسبت نابرابری

نسبت نابرابری [3] معیاری است که برای انتخاب ویژگی در مقاله اویسال [4] استفاده شده است [۲] . نحوه محاسبه این معیار در رابطه ۲-۳ آورده شده است.

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

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 [۲] .

۲-۱-۴ معیار زائدی کمینه شباهت بیشینه

معیار زائدی کمینه شباهت بیشینه [5] 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 ۲-۴ آمده است.

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

در این رابطه مجموعه S به معنی مجموعه ویژگی‌های انتخابی است. $I(a, b)$ به معنای اطلاعات متقابل [6] 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.

۲-۱-۵ معیار تمایزگر نسبی

معیار تمایزگر نسبی [7] یک روش انتخاب ویژگی برای مسائل دسته‌بندی دودویی است که در مقاله لبنی و همکاران [۱] مورد استفاده بدوه است. نحوه محاسبه این معیار در رابطه ۲-۵ آمده است.

[3] Odds Ration
[4] Uysal
[5] Minimal redundancy maximal relevance
[6] Mutual information
[7] Relative discriminative criterion

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

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

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

[8] Area 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 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[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.

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.[۲] 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 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

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

۳-۳ روش برپایه الگوریتم ژنتیک

[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

فصل چهارم

ارزیابی و مقایسه

۴-۱ مقایسه تئوری

فصل پنجم

جمع‌بندی و نتیجه‌گیری

منابع و مراجع

[1] 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.

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