



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

صفحه

فهرست اشکال

شکل

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

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.

³Odds Ration

⁴Uysal

⁵Minimal redundancy maximal relevance

⁶Mutual information

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

این روش 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.

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

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 ۳-۱: 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 ۳-۲: 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	-۴/۷۱۳۶ ، ۴/۶۱۵۱ ، -۴/۲۱۴۶	C_2 مثبت
گرگ	۱	۴/۶۱۵۱ ، -۳/۲۵۸۱ ، -۳/۵۳۶۱	C_1 مثبت
گربه	۰/۵۵۵۶	۴/۱۱۰۹ ، ۴/۳۳۰۷ ، -۴/۶۱۵۱	C_3 منفی
موش	۰/۵۵۵۶	۴/۱۱۰۹ ، ۴/۳۳۰۷ ، -۴/۶۱۵۱	C_3 منفی
اسب	۰/۵۲۰۰	۴/۶۱۵۱ ، -۳/۲۵۸۱ ، ۳/۵۳۶۱	C_1 منفی
پلیکان	۰/۴۴۴۴	-۳/۷۱۳۶ ، -۳/۹۳۱۸ ، ۳/۸۱۶۵	C_2 منفی
اردک	۰/۱۱۱۱	-۳/۰۴۴۵ ، -۳/۲۵۸۱ ، ۲/۴۹۴۱	C_2 منفی
شتر	۰/۱۱۱۱	-۳/۰۴۴۵ ، -۳/۲۵۸۱ ، ۲/۴۹۴۱	C_2 منفی
مرغ	۰/۰۹۰۳	-۳/۷۱۳۶ ، ۰ ، -۱/۲۹۲۹	C_1 منفی

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

۲-۳ روش MRDC

۳-۳ روش برپایه الگوریتم ژنتیک

فصل چهارم

ارزیابی و مقایسه

۱-۴ مقایسه تئوری

فصل پنجم

جمع‌بندی و نتیجه‌گیری

منابع و مراجع

- [1] 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.
- [2] Uysal, Alper Kursat. An improved global feature selection scheme for text classification. *Expert systems with Applications*, 43:82–92, 2016.