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

۲ مفاهیم تئوری ۲
۲-۱ روش‌های انتخاب ویژگی ۳
۲-۱-۱ بهره اطلاعاتی ۳
۲-۱-۲ شاخص جینی ۳
۲-۱-۳ نسبت نابرابری ۴
۳ روش‌های ارائه‌شده ۵
۴ ارزیابی و مقایسه ۶
۵ جمع‌بندی و نتیجه‌گیری ۷
منابع و مراجع ۸

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.

[3] Odds Ration

[4] Uysal

[5] Minimal redundancy maximal relevance

[6] Mutual information

فصل سوم

روش‌های ارائه‌شده

فصل چهارم

ارزیابی و مقایسه

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.