هوش محاسباتی

پروژه سوم

یکی از مسائل مهم در ارتباطات پیامکی، شناسایی پیامک‌های جعلی[1] از واقعی است. برای توسعه سامانه‌ای که بتواند صحت پیامک‌ها را شناسایی کند (spam filtering)، یک مجموعه داده از پیامک‌های جعلی و واقعی می‌تواند برای آموزش سامانه در یک فرآیند بهینه‌سازی مورد استفاده قرار گیرد. با توجه به حساسیت موجود در شناسایی برخی پیامک‌ها به عنوان جعلی که ممکن است منجر به عدم ارسال آنها برای گیرنده شود، لازم است روند شناسایی پیامک‌های جعلی به صورت شفاف و مشخص قابل توضیح باشد. از طرف دیگر تصمیم‌گیری در مورد جعلی بودن یک پیامک به دلیل ابهام در معانی و مفاهیم بکار رفته در آن دارای عدم قطعیت زیادی است. با در نظر گرفتن این محدودیت‌ها، در توسعه این سامانه تصمیم گرفته شده از یک پایگاه قوانین[2] دسته‌بندی[3] مبتنی بر منطق فازی استفاده شود. چنین رویکردی به دسته‌بندی معمولاً به عنوان سامانه‌های دسته‌بند یادگیر[4] شناخته می‌شود.

وظیفه اصلی دانشجویان در این پروژه توسعه چنین پایگاه قوانینی برای دسته‌بندی پیامک‌ها به جعلی و واقعی با توجه به خصوصیات توضیحات داده شده در بخش‌های بعدی است. در بخش ۱ مسأله و جزئیات آن به صورت دقیق‌تر تشریح شده است. در بخش ۲ ملاحظات مدنظر در انجام این پروژه آورده شده است. بخش ۳ موارد نهایی که باید تحویل داده شود را توضیح داده است. مهلت تحویل این پروژه تا پایان روز جمعه ۱۲ خرداد است.

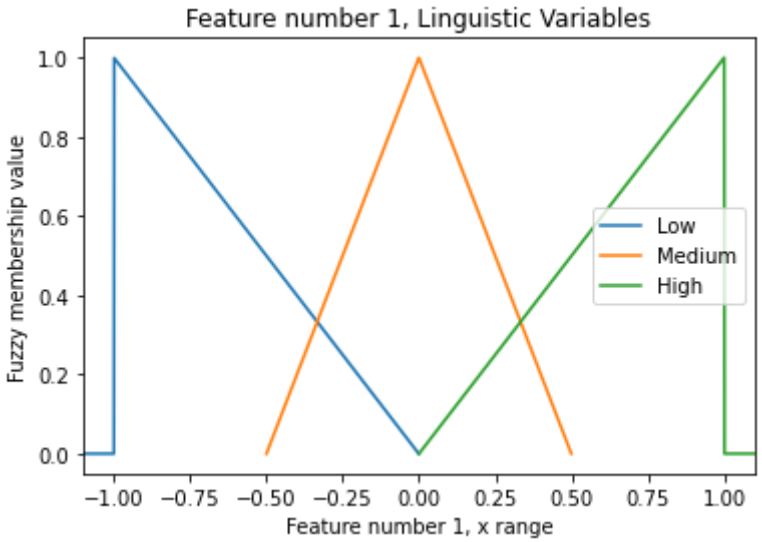
۱ – تشریح مسئله

هدف از این پروژه توسعه یک پایگاه قوانین فازی و استفاده از استدلال تقریبی[5] برای دسته‌بندی پیامک‌ها است. هر قانون در این پایگاه برای نگاشت نمونه‌های منطبق با شرایط توصیف شده توسط یک مقداردهی برای متغیرهای زبانی[6] به یک دسته به کار گرفته می‌شود. نمونه‌ای از چنین قانونی به صورت زیر است:

$$\text{If } X_1 \text{ is } A_{1i} \text{ and } X_2 \text{ is } A_{2j} \text{ and } \dots \text{ and } X_n \text{ is } A_{nk} \text{ Then } Y = 0 \qquad (1)$$

که در آن X_1، ...، X_n متغیرهای زبانی مسأله بوده و هر کدام مرتبط با یکی از ویژگی‌هایی است که برای توصیف پیامک در نظر گرفته شده است. هر یک از این متغیرها دارای مجموعه‌ای از مقادیر زبانی است که با استفاده از مجموعه‌های فازی تعریف می‌شوند: $T(X_i) = \{A_{i1}, \dots, A_{im(i)}\}$. متغیر Y نشان دهنده دسته بوده و دارای یکی از دو مقدار 0 (واقعی) و 1 (جعلی) است. شکل ۱ مثالی از یک متغیر زبانی با مقادیر Low، Medium و High که بر روی مجموعه جهانی اعداد حقیقی در بازه $[-1, 1]$ تعریف شده‌اند را نشان می‌دهد.

[1] Spam
[2] Rule Base
[3] Classification
[4] Learning Classifier Systems
[5] Approximate Reasoning
[6] Linguistic Variable



شکل ۱. مقادیر زبانی در نظر گرفته شده برای یک ویژگی

پس از استخراج ویژگی‌های هر پیامک به صورت ترد[۷]، در مرحله فازی‌سازی[۸] میزان تطابق مقدار هر ویژگی با مقادیر زبانی مختلف متغیر مربوط به آن ویژگی بدست می‌آید (از محاسبه درجه عضویت مقدار آن ویژگی در مجموعه‌های فازی هر یک از مقادیر). مثلاً اگر مقدار مشاهده شده برای ویژگی ۱ (در مثال شکل ۱) 0.25- باشد، میزان تطابق آن با مقادیر Low، Medium و High از متغیر زبانی مربوط به ویژگی ۱ به ترتیب برابر با 0.25، 0.5 و 0 است. بر این اساس می‌توان با تجمیع[۹] میزان تطابق شرط‌های مختلف یک قانون میزان تطابق کلی آن قانون با یک ورودی را تعیین کرد. رابطه زیر میزان تطابق کلی قانون نشان داده شده در رابطه ۱ را با استفاده از عملگر ضرب جبری برای تجمیع نشان می‌دهد.

$$g_R\left(\boldsymbol{x}^{(p)}\right) = \mu_{A_{1i}}\left(x_1^{(p)}\right) \times \mu_{A_{2j}}\left(x_2^{(p)}\right) \times \ldots \times \mu_{A_{nk}}\left(x_n^{(p)}\right) \qquad (2)$$

این روند برای هر یک از قوانین موجود در پایگاه انجام شده و میزان تطابق کلی هر یک از آنها با ورودی محاسبه می‌شود. در این صورت می‌توان با تجمیع میزان تطابق قوانین مرتبط با هر یک از دسته‌ها (در این مسأله فقط دسته 0 و 1)، دسته‌ای که دارای تطابق بیشتری با ورودی است را مطابق روابط زیر برای آن ورودی در نظر گرفت.

$$g_c\left(\boldsymbol{x}^{(p)}\right) = \sum_{R_j \in\, class(c)} g_{R_j}\left(\boldsymbol{x}^{(p)}\right) \qquad (3)$$

$$\hat{y}\left(\boldsymbol{x}^{(p)}\right) = \arg \max_{c \in \{0,1,\cdots\}} g_c\left(\boldsymbol{x}^{(p)}\right) \qquad (4)$$

برای توسعه پایگاه قوانینی که به این شکل به کار گرفته می‌شود، در این پروژه از مجموعه داده SMS spam collection که از پایگاه UCI[۱۰] قابل دسترسی است استفاده می‌شود. این مجموعه داده دارای ۵۵۷۴ داده متنی است که به یکی از دو دسته خروجی

[7] Crisp
[8] Fuzzification
[9] Aggregation
[10] https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

متعلق هستند. پایگاه قوانین باید به شکلی طراحی شود که با داده‌های موجود در این مجموعه داده تطبیق پیدا کند. پارامترهای قابل تطبیق برای آموزش پایگاه قوانین در بخش ۲ مشخص شده‌اند.

با توجه به فضای پیچیده حاصل از مقادیر مختلف برای پارامترهای چنین پایگاه قوانینی، از *الگوریتم‌های تکاملی* برای بهینه‌سازی در روند آموزش سامانه استفاده می‌شود. در این رویکرد، به علت غیرقطعی بودن قوانین، کیفیت (برازندگی) هر قانون ایجاد شده در پایگاه قوانین را می‌توان با توجه به ضریب اطمینان[11] (CF) آن قانون در هنگام دسته‌بندی مشخص کرد که مرتبط با درجه تطابق کلی قانون برای نمونه‌های هر دسته است:

$$f_c(R_j) = \sum_{x^{(p)}:y^{(p)}=c} g_{R_j}(x^{(p)}) \quad (5)$$

$$CF(R_j) = \frac{f_{y_j}(R_j) - f^{neg}(R_j)}{\sum_{c\in\{0,1,\cdots\}} f_c(R_j)} \quad (6)$$

$$f^{neg}(R_j) = \frac{1}{r-1}\sum_{c\neq y_j} f_c(R_j) \quad (7)$$

در صورت کسر داده شده در رابطه ۶، $f_{y_j}(R_j)$ نشان دهنده درجه تطابق کلی نمونه‌های آموزشی دسته تعیین شده در خروجی قانون R_j است و $f^{neg}(R_j)$ میانگین تطابق کلی قانون R_j با هر یک از کلاس‌های دیگر است. مقدار r در رابطه ۷ نشان دهنده تعداد کل دسته‌ها است.

در این پروژه ویژگی‌های TF-IDF[12] از متن پیامک‌های موجود در مجموعه داده در روند پیش‌پردازش استخراج می‌شود (به فایل کد پیوست مراجعه کنید)، هر چند انواع دیگری از ویژگی‌ها به این منظور قابل استفاده است. با توجه به تعداد زیاد این ویژگی‌ها، در روند پیش‌پردازش تعداد آنها کاهش می‌یابد تا ایجاد پایگاه قوانین ساده‌تر شود. برای کاهش ابعاد ویژگی‌ها دو رویکرد در نظر گرفته شده است: ۱) انتخاب ویژگی[13] که به انتخاب زیرمجموعه‌ای از ویژگی‌های مهمتر می‌پردازد و در این پروژه از معیار اطلاعات متقابل[14] برای شناسایی چنین ویژگی‌هایی استفاده شده است؛ ۲) استخراج ویژگی که ویژگی‌های جدیدی از روی ویژگی‌های اولیه ایجاد می‌کند و در این پروژه از روش تحلیل مؤلفه‌های اصلی[15] به این منظور استفاده شده است. پیاده‌سازی هر دو روش در فایل کد پیوست موجود است. دانشجویان باید پایگاه قوانین فازی را با توجه به ویژگی‌های بدست آمده پس از پیش‌پردازش متون توسعه دهند.

۲ – ملاحظاتی که در حل مسئله باید در نظر گرفته شوند

الف) مدل‌سازی زبانی مناسب برای ۵ ویژگی مهمتر بدست آمده از پیش‌پردازش متون را با شرایط زیر بدست آورید:

- هر متغیر زبانی (متناظر با هر یک از ویژگی‌ها) می‌تواند بین ۳ تا ۵ مقدار زبانی داشته باشد

[11] Certainty Factor
[12] Term Frequency – Inverse Document Frequency
[13] Feature Selection
[14] Mutual Information
[15] Principal Components Analysis

- هر مقدار زبانی می‌تواند با یکی از چهار مجموعه فازی زیر نشان داده شود:
  - مجموعه فازی مثلثی متساوی‌الساقین ($s > 0$)

$$\mu_{iso-tri}(x) = \max\left(\min\left(\frac{x-m}{s}, \frac{m-x}{s}\right), 0\right)$$

  - مجموعه فازی ذوزنقه‌ای قائم‌الزاویه ($|s| > 0$)

$$\mu_{rect-trap}(x) = \max\left(\min\left(\frac{x-m}{s}, 1\right), 0\right)$$

  - مجموعه فازی گاوسی

$$\mu_{gaussian}(x) = e^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2}$$

  - مجموعه فازی سیگموئید

$$\mu_{sigmoid}(x) = \frac{1}{1 + e^{-\frac{x-m}{s}}}$$

- هر مجموعه فازی دارای دو پارامتر m و s است

ب) بر اساس مدل‌سازی زبانی انجام شده قوانین فازی دسته‌بندی را با توجه به شرایط زیر بدست آورید:

- برای هر قانون باید قسمت شرط آن با تعیین یکی از مقادیر برای هر متغیر زبانی تعیین شود
  - هر یک از مقادیر به صورت مستقیم یا نفی شده (negated) می‌تواند در قانون بکار گرفته شود
  - ممکن است برخی متغیرها در یک قانون بکار نروند
- برای هر قانون یکی از دسته‌ها به عنوان خروجی تعیین شود
- تعداد قوانین پایگاه محدود است (اما هدف انتخاب زیرمجموعه بهینه از قوانین نیست بلکه دستیابی به قوانینی که بتواند عملیات دسته‌بندی را با عملکرد مناسب انجام دهد کافی است)

پ) مؤلفه‌های اصلی الگوریتم تکاملی مورد استفاده برای تطبیق پارامترهای این پایگاه قوانین فازی شامل روش نمایش راه‌حل‌ها، تابع هدف، روش انتخاب، عملگرهای تغییر، روش مقداردهی اولیه جمعیت، اندازه جمعیت و شرایط توقف باید به صورت کامل و صریح مشخص شوند.

ت) دلیل انتخاب هر یک از مؤلفه‌ها در الگوریتم تکاملی را بیان کنید.

ث) برای سنجش عملکرد سامانه دسته‌بندی مجموعه داده‌ها را به دو قسمت آموزشی و آزمایشی تقسیم کرده و علاوه بر نحوه بهبود عملکرد روی داده‌های آموزشی در حین بهینه‌سازی، نتیجه عملکرد مجموعه قوانین نهایی را روی داده‌های آزمایشی گزارش کنید. برای تقسیم مجموعه داده به دو قسمت می‌توانید از تابع آماده train_test_split در کتابخانه sklearn استفاده کنید[16].

ج) نتیجه نهایی مدل‌سازی زبانی و مجموعه قوانین فازی بدست آمده را مشخص کنید.

چ) نحوه دسته‌بندی یکی از داده‌ها را با مشخص کردن قوانین بکار رفته در روند استنتاج آن و میزان تطابق آنها توضیح دهید.

[16] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

ح) تأثیر تعداد قوانین موجود در پایگاه را با بررسی حداقل ۳ اندازه مختلف در بازه [۵۰، ۵۰۰] برای اندازه پایگاه بررسی کنید.

خ) تأثیر استفاده از عملگر ضرب جبری بجای عملگر استاندارد min برای تجمیع قسمت‌های مختلف شرط قوانین را در عملکرد کلی سامانه دسته‌بندی بررسی کنید.

د) تأثیر استفاده از دو روش کاهش بعد ارائه شده (انتخاب ویژگی و استخراج ویژگی) را در عملکرد کلی سامانه دسته‌بندی بررسی کنید.

۳ – مواردی که باید تحویل داده شود

- فایل(های) کد برنامه مورد استفاده برای پیاده‌سازی پروژه در یک پوشه به نام Code
  - استفاده از کتابخانه آماده برای بخش الگوریتم تکاملی تمرین مجاز بوده اما قسمت فازی باید از ابتدا[17] پیاده‌سازی شود و استفاده از کتابخانه آماده نمره‌ای ندارد.
  - هرگونه کپی کد یا گزارش چه از فضای اینترنت و چه از گروه‌های دیگر نمره منفی خواهد داشت.
  - میزان مشارکت هر فرد گروه در پروژه باید به صورت شفاف مشخص باشد. هر یک از اعضای گروه باید به کلیات روش حل مسئله و نیز جزییات آن بخشی از پروژه که مسئولیتش را به عهده داشته‌اند تسلط کافی داشته باشد. این مورد به صورت ارائه حضوری مورد ارزیابی قرار خواهد گرفت.
- فایل گزارش با نام Doc.pdf شامل موارد زیر:
  - نتایج حل مسئله به همراه ملاحظات مشخص شده در بخش ۲
  - تشریح و تحلیل نتایج به دست آمده از نظر شما
  - هر گونه توضیح اضافی در مورد نحوه انجام پروژه

* دقت کنید که گزارش شما حتما باید به صورت یک گزارش فنی باشد.

- فایل‌های کد و گزارش را به صورت یک فایل فشرده در قالب ZIP و با نام CI _PR3_Names تحویل دهید (به جای Names نام خانوادگی اعضای گروه را قرار دهید).
- پاسخ‌ها باید از طریق سایت درس در کوئرا ارسال شوند.

مهلت تحویل این پروژه تا پایان روز جمعه ۱۲ خرداد خواهد بود.

موفق باشید

[17] From Scratch