باسمه تعالی

دانشکده مهندسی کامپیوتر

دسته‌بندی اسناد با استفاده از یادگیری ماشین

گزارش پروژه درس هوش مصنوعی

نویسنده :

محسن رحیمی (۸۹۵۲۱۱۶۹)

استاد:

دکتر بهروز مینایی

نیمسال اول ۹۲-۹۱

چکیده

دسته‌بندی اسناد امروزه به طور گسترده در بازیابی اطلاعات به منظور سازماندهی اسناد به کار گرفته می‌شود. در روش دسته‌بندی نظارت شده اسناد، اطلاعاتی صحیح در مورد اسنادی که قبلا دسته‌بندی شده اند با روشهایی در اختیار ما قرار می گیرد و ما بر اساس این اطلاعات به دسته‌بندی اسناد جدید می پردازیم. روش هایی که به بررسی آنها خواهیم پرداخت عبارتند از: روش نزدیک ترین همسایه، روش ماشین بردار پشتیبان و روش ساده بیزی. الگوریتم پیاده سازی بر اساس روش ساده بیزی خواهد بود. در این پروژه به دسته‌بندی موضوعی اسناد موجود در پیکره همشهری می پردازیم و سپس با استفاده از الگوریتم های موجود برای یک سند جدید موضوع را پیش بینی می کنیم. پیکره همشهری مجموعه ای از مقالات خبری همشهری اولین روزنامه آنلاین در ایران است که بیش از ۲۰ سال است که منتشر می‌شود و آرشیو آن در اختیار عموم قرار دارد. این پیکره شامل ۳۴۵ مگابایت متن برچسب‌دار اخبار با ساختار مناسب است.

کلمات کلیدی

یادگیری ماشین، رده بندی، دسته بندی، دسته بندی کننده بیزین، بیزین ساده، پیکره همشهری

۱- مقدمه

امروزه دسته‌بندی به عنوان مهمترین مسئله‌ی یادگیری با ناظر[1] در بسیاری از حوزه‌ها و بخصوص تحلیل داده‌های آماری و بازیابی اطلاعات مورد توجه بسیاری قرار گرفته است.دسته‌بندی خودکار متون به معنی انتساب اسناد متنی به دسته های از پیش تعیین شده در ده سال اخیر تمام توجهات را به خود جلب کرده است. این مسئله به خاطر دسترسی به اسناد الکترونیکی فراوان و نیاز مبرم به سازماندهی آنهاست. در جامعه تحقیقاتی روش اصلی در این زمینه، روشهای بر اساس یادگیری ماشین هستند. در این پروژه ضمن آشنایی با روش های دسته‌بندی اسناد پیاده سازی یکی از الگوریتم ها را نیز خواهیم دید.

برای دسته بندی نظارت شده اسناد ما باید ابتدا روشی را برای یادگیری از روی مجموعه آموزشی داشته باشیم.برای ذخیره اطلاعات بدست آمده از یادگیری می توانیم آنها را در یک فضای n بعدی نگاشت کرده و اطلاعات مربوط به اسناد هر دسته را نگهداری کنیم الگوریتم نزدیکترین همسایه از این روش استفاده می‌کند.

روش دیگر برای نگهداری از اطلاعات بدست آمده از یادگیری این است که آنها را به صورت مجموعه احتمالات مربوط به اسناد و موضوعات ذخیره کنیم و در هنگام نیاز به آنها از این احتمالات برای دسته بندی اسناد جدید استفاده کنیم.

یکی دیگر از مشکلات موجود در دسته بندی اسناد این است که تا چه اندازه دسته بندی اسناد جدید درست و قابل اطمینان خواهد بود. همچنین گاهی محاسبات زیاد ما را بر آن خواهد داشت که در بعضی موارد از تقریب برای بدست آوردن دسته‌بندی یک سند استفاده کنیم؛ آیا تقریب های در نظر گرفته شده در الگوریتم ها نتایج درستی را در پی خواهد داشت؟

۲- یادگیری ماشین[]

یادگیری ماشین یکی از شاخه های وسیع و پر کاربرد هوش مصنوعی می باشد که به معنی طراحی و توسعه الگوریتم هایی است که بر اساس آنها رایانه ها و یا دیگر ماشین ها توانایی تعلم و یادگیری را پیدا می کنند.

الگوریتم های یادگیری ماشین بر اساس نوع داده های در اختیار به انواع زیر تقسیم می شوند:

الف) **یادگیری با نظارت**[]: یکی از روشهای یادگیری است که مجموعه ای از جفت‌های ورودی – خروجی ارائه شده و سیستم تلاش می‌کند تا تابعی از ورودی به خروجی را فرا گیرد

ب) **یاد گیری بدون نظارت**[]: یکی از روشهای یادگیری است که یک مجموعه از مثالهای یادگیری وجود دارد که در آن فقط مقدار ورودی‌ها مشخص است و اطلاعاتی در مورد خروجی صحیح در دست نیست.

ج) **یاد گیری نیمه نظارت شده**[]: ترکیبی از دو روش یادگیری قبلی که عامل هوش‌مند هم از داده‌های بدون برچسب[1] وهم از داده‌های با برچسب استفاده می‌کند را یادگیری نیمه نظارتی می‌گویند

---

[1] Supervised learning

[] Machine Learning

[] Supervised learning

[] Unsupervised learning

[] Semi-supervised learning

د) **یادگیری تقویتی**[]: در یادگیری تقویتی هیچ نوع زوج ورودی- خروجی ارائه نمی‌شود. به جای آن، پس از اتخاذ یک عمل، حالت بعدی و پاداش بلافصل به عامل ارائه می‌شود.

۲-۱- یادگیری با نظارت (نظارت شده)

یادگیری تحت نظارت، یک روش عمومی در یادگیری ماشین است. یک مجموعه از مثالهای یادگیری وجود دارد بازای هر ورودی، مقدار خروجی و یا تابع مربوطه نیز مشخص است. هدف سیستم یادگیر بدست آوردن فرضیه‌ای است که تابع و یا رابطه بین ورودی و یا خروجی را حدس بزند به این روش یادگیری با نظارت گفته می‌شود. به این معنی که یک شخص ناظری وجود دارد که برچسب‌گذاری برای تمایز دسته‌های مختلف را بر روی اسناد اعمال می‌کند.

۳- دسته‌بندی اسناد[]

دسته‌بندی اسناد یا رده بندی اسناد یک مسئله مهم در علوم مرتبط با کتابخانه ها و همچنین علوم کامپیوتر است. کار اصلی دسته‌بندی این است که اسناد را به یک یا چند طبقه یا رده تقسیم کند. این کار ممکن است به صورت دستی یا به صورت الگوریتمی انجام شود. رده بندی به صورت دستی مانند کاری است که در یک کتابخانه انجام می شود و کتابها به دسته های مختلف تقسیم می شوند. روش الگوریتمی رده بندی اصولا در علم کامپیوتر و در مبحث داده کاوی جای می گیرد و به صورت یادگیری ماشین انجام می شود.

اسناد ممکن است بر اساس موضوع و یا براساس یک ویژگی خاص (مانند نام نویسنده) رده بندی شوند. در رده بندی موضوعی اسناد ابتدا اسنادی با موضوع معین به عنوان نمونه برای یادگیری به ماشین داده می شود و ماشین با توجه به کلمات داخل هر سند به یادگیری می پردازد و با در نظر گرفتن احتمال وجود کلمات به پیش بینی موضوعی سندهای دیگر می پردازد.

۳-۱- روش‌های دسته‌بندی

به منظور اعمال روشهای دسته‌بندی، بایستی سندهای متنی به روشی که نشان دهنده محتوای آنها باشد، نمایش داده شوند. معمولا سندها را با استفاده از تک کلمات درون سندها، با در نظر نگرفتن کلمات بی ارزش[] (مانند از ، را ، چرا و...) نمایش می دهند .

در روش های دسته‌بندی، به منظور یادگیری این روش ها، سندهای متنی به دو دسته سندهای آموزشی و سندهای آزمون تقسیم می شوند. از سندهای آموزشی برای دسته‌بندی و از سندهای آزمون برای ارزیابی میزان دقت و کامل بودن روش دسته‌بندی استفاده می شود.

---

[1] Label

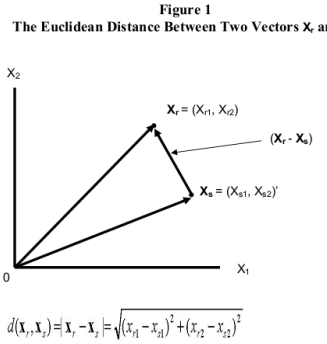
[] Reinforcement learning

[] Document Classification

[] Stop Words

تاکنون روشهای زیادی از جمله روشهای آماری برای دسته‌بندی خودکار سندهای متنی تدوین شده اند. از جمله روش های دسته‌بندی می‌توان به روش دسته‌بندی نزدیکترین همسایه، روش بیزین، روش SVM، درخت تصمیم و شبکه های عصبی اشاره نمود.

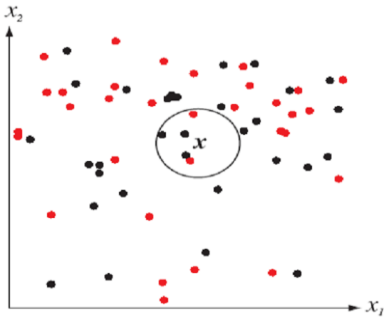
۳-۱-۱- روش نزدیکترین همسایه[1]



الگوریتم نزدیکترین همسایه یک الگوریتم یادگیری با نظارت است. در حالت کلی از این الگوریتم به دو منظور استفاده می شود؛ برای تخمین تابع چگالی توزیع داده‌های تعلیم و برای طبقه‌بندی داده‌های تست بر اساس الگوهای تعلیم.

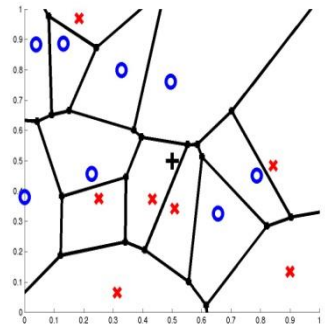
در این روش فرض میشود که تمام نمونه ها نقاطی در فضای n بعدی حقیقی هستند و همسایه ها برمبنای فواصل اقلیدسی استاندارد تعیین میشوند.

فرض می کنیم $D^n = \{x_1,...,x_n\}$ مجموعه‌ای از n الگوی ورودی باشد و $x' \in D^n$ نزدیکترین ورودی به نقطه‌ی تست x باشد. قانون نزدیک ترین همسایه برای طبقه‌بندی x آنرا در کلاسی مشابه با کلاس x' قرار میدهد.



قانون k نزدیک‌ترین همسایه گسترشی از قانون نزدیک ترین همسایه است به این صورت که این قانون x را در دسته ای قرار می‌دهد که بیشترین تکرار را در بین k نزدیک ترین همسایه x دارد. به عنوان مثال در شکل زیر x توسط این قانون در کلاس سیاه قرار می‌گیرد.

اگرچه این الگوریتم هرگز فرضیه عمومی مشخصی ایجاد نمیکند، با این وجود ممکن است سطح تصمیم القا شده توسط الگوریتم برای یک فضای دو بعدی را بصورت ترکیبی از چندوجهی ها نشان داد که هر چند وجهی مجموعه ای از نقاطی را که توسط آن دسته‌بندی خواهند شد مشخص مینماید. نقاط خارج چندوجهی نقاطی خواهند بود که توسط سایر چندوجهی ها دسته بندی خواهند شد.



این نوع نمودار Voronoi diagram خوانده میشوند.

الگوریتم k-NN را میتواند بسادگی برای توابع هدف پیوسته نیز استفاده نمود.

در این حالت بجای انتخاب متداولترین مقدار موجود در همسایگی مقدار میانگین k مثال همسایه محاسبه میشود.

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} f(x_i)}{k}$$

[1] Nearest Neighbor

میتوان عملکرد این الگوریتم را با در نظر گرفتن وزنی برای هر یک از k مثال همسایگی بهتر نمود. این وزن بر اساس فاصله نمونه ها تا نمونه مورد بررسی اعمال میشود و معمولا با فاصله نمونه ها رابطه معکوس دارد.

در حالت گسسته:

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{\operatorname{argmax}} \sum_{i=1}^{k} w_i \delta(v, f(x_i)) \qquad \text{where } w_i = \frac{1}{d(x_q, x_i)^2}$$

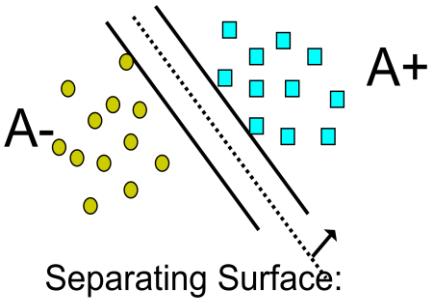
در حالت پیوسته

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} w_i f(x_i)}{\sum_{i=1}^{k} w_i} \qquad \text{where } w_i = \frac{1}{d(x_q, x_i)^2}$$

در صورت اعمال وزن این امکان وجود خواهد داشت که به جای k نمونه همسایه از تمامی نمونه ها برای دسته بندی استفاده کنیم. اما این انتخاب باعث کند شدن عمل دسته بندی خواهد شد.

۳-۱-۲- روش ماشین بردار پشتیبان[1]

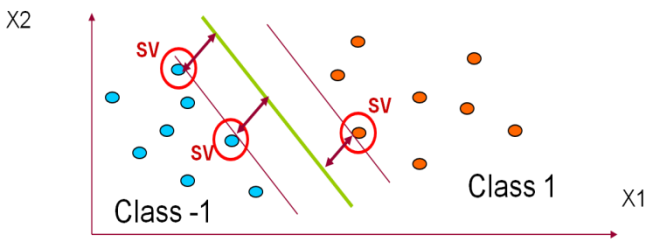
هدف این دسته الگوریتم ها تشخیص و متمایز کردن الگوهای پیچیده در داده هاست (از طریق کلاسترینگ، دسته بندی، رنکینگ، پاکسازی و غیره).



مبنای کاری دسته‌بندی کنندهٔ SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد.

با فرض اینکه دسته ها بصورت خطی جداپذیر باشند، ابرصفحه هائی با حداکثر حاشیه (maximum margin) را بدست می آورد که دسته ها را جدا کنند. اگر دو دسته وجود داشته باشند که بصورت خطی از هم جداپذیر باشند، بهترین جدا کننده این دو دسته چیست؟

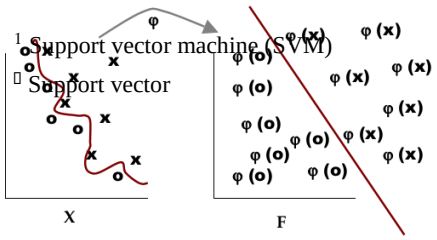
ایده SVM برای جداسازی دسته ها به این صورت است که ابتدا یک صفحه دسته‌بندی بسازید. دو صفحه مرزی موازی با صفحه دسته بندی رسم کرده و آندو را آنقدر از هم دور میکنیم که به داده ها برخورد کنند.صفحه دسته بندی که بیشترین فاصله را از صفحات مرزی داشته باشد، بهترین جدا کننده خواهد بود.



نزدیکترین داده های آموزشی به ابر صفحه های جدا کننده بردار پشتیبان[2] نامیده میشوند.

فشرده سازی اطلاعات در الگوریتم SVM به این صورت است که بجای داده های آموزشی از بردارهای پشتیبان استفاده میکند.

Transformation to separate



[1] Support vector machine (SVM)

[2] Support vector

در مسایلی که داده ها بصورت خطی جداپذیر نباشند داده ها به فضای با ابعاد بیشتر نگاشت پیدا میکنند تا بتوان آنها را در این فضای جدید بصورت خطی جدا نمود.

۳-۱-۳- روش ساده بیزی[1]

یک روش بسیار کاربردی یادگیری روش یادگیرنده ساده بیزی می‌باشد که عموماً روش طبقه‌بندی ساده بیزی نامیده می‌شود . در برخی زمینه‌ها نشان داده شده است که کارایی آن قابل قیاس با کارایی روش‌هایی مانند شبکه عصبی و درخت تصمیم می‌باشد .

طبقه‌بندی ساده بیزی برای مسائلی که هر نمونه x درآن توسط مجموعه‌ای از مقادیر صفات و تابع هدف $f(x)$ از مجموعه‌ای مانند V انتخاب می‌گردد کاربرد دارد . روش بیزی برای طبقه‌بندی نمونه جدید این است که محتمل‌ترین طبقه یا مقدار هدف[2] v_{MAP} را با داشتن مقادیر صفات $< a_1, a_2, ..., a_n >$ که توصیف کننده نمونه جدید است شناسایی کند،

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, ..., a_n)$$

با استفاده از قضیه بیز می‌توان عبارت بالا را به صورت زیر بازنویسی کرد،

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1, a_2, ..., a_n, \mid v_j) P(v_j)}{P(a_1, a_2, ..., a_n)}$$
$$= \arg\max_{v_j \in V} P(a_1, a_2, ..., a_n, \mid v_j) P(v_j)$$

محاسبه $P(v_j)$ از روی داده‌های آموزشی به این صورت که میزان تکرار v_j در داده‌ها چقدر است، آسان می‌باشد . اما محاسبه جملات مختلف $P(a_1, a_2, ..., a_n \mid v_j)$ به این صورت قابل قبول نخواهد بود مگر اینکه حجم بسیار بسیار زیادی از داده‌های آموزشی در اختیار داشته باشیم . مشکل اینجاست که تعداد این جملات برابر تعداد نمونه‌های ممکن ضرب در تعداد مقادیر تابع هدف می‌باشد . بنابراین باید هر نمونه را چندین بار مشاهده کنیم تا تخمین مناسبی از آن بدست آید .

فرض روش طبقه‌بندی ساده بیزی بر اساس این ساده سازی است **که مقادیر صفات با داشتن مقادیر تابع هدف از یکدیگر مستقل شرطی[3] می‌باشند** . به عبارت دیگر ، این فرض بیانگر این است که به شرط مشاهده خروجی تابع هدف احتمال مشاهده صفات $a_1, a_2, ..., a_n$ برابر ضرب احتمالات هر صفت به طور جداگانه می‌باشد . اگر این را جایگزین معادله بالا کنیم روش طبقه‌بندی ساده بیزی را نتیجه می‌دهد،

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

---

[1] Naïve Bayes

[2] Maximum A Posteriori (MAP)

[3] conditional independence

که v_{NB} خروجی طبقه‌بندی ساده بیزی برای تابع هدف می‌باشد . توجه کنید که تعداد جملات $P(a_i \mid v_j)$ که در این روش باید محاسبه شوند برابر تعداد صفات ضرب در تعداد دسته‌های خروجی برای تابع هدف می‌باشد که این مقدار از تعداد جملات $P(a_1, a_2, ..., a_n \mid v_j)$ بسیار کمتر است .

نتیجه اینکه یادگیری ساده بیزی سعی در تخمین مقادیر مختلف $P(v_j)$ و $P(a_i \mid v_j)$ با استفاده از میزان تکرار آنها در داده‌های آموزشی دارد . این مجموعه تخمین‌ها متناظر با فرض یاد گرفته شده است . سپس از این فرض برای طبقه بندی نمونه‌های جدید استفاده می‌شود که این کار با استفاده از فرمول بالا صورت می‌گیرد . هر گاه فرض مستقل شرطی بودن روش طبقه بندی ساده بیزی بر آورده شود طبقه ساده بیزی معادل طبقه MAP خواهد بود .

۴- پیاده سازی پروژه

برای دسته بندی اسناد از روش ساده بیزی که بر اساس تئوری بیز استوار است استفاده می کنیم.

پیاده سازی پروژه از دو قسمت تشکیل شده است؛ این دو قسمت پس از پیاده سازی برای برچسب گذاری یک سند جدید و تعیین موضوع آن استفاده خواهند شد.

قسمت اول مربوط به یادگیری می باشد. ماشین در این قسمت باید بتواند از یک مجموعه آموزشی مقادیر مربوط به احتملات مختلف را بدست بیاورد. از جمله احتملاتی که در این مرحله باید محاسبه شود احتمال مربوط به رخداد هر یک از موضوعات در دادگان آموزشی می باشد به عبارت دیگر باید مقدار $P(v_j)$ را به ازای هر یک از موضوعات بدست آوریم. برای این کار از فرمول زیر استفاده خواهیم کرد:

$$P(v_j) = \frac{N_{Vj}}{N}$$

که در آن N_{Vj} تعداد اسناد با موضوع v_j در دادگان آموزشی و N تعداد کل اسناد دادگان آموزشی است.

احتمال دیگری که در این قسمت باید محاسبه گردد احتمال شرطی $P(a_i \mid v_j)$ است.این احتمال بیانگر رخداد هر یک از صفات به شرط وقوع هر یک از موضوعات می باشد. برای محاسبه این احتمال از فرمول زیر استفاده می کنیم

$$P(a \mid v) = \frac{N_{av}}{\sum_{a' \in a_i} N_{a'v}}$$

که در آن N_{av} تعداد تکرار صفت a در موضوع v می باشد همچنین عبارت مخرج کسر برابر با مجموع تعداد کل صفات در موضوع v می باشد.

به دلیل اینکه ممکن است یک صفت در یک موضوع اصلا تکراری نداشته باشد و احتمال بالا برابر با صفر شود و در نتیجه احتمال مربوط به قانون بیز را که برابر حاصلضرب همه احتمالات فرمول بالا است صفر کند از یک مرحله نرم سازی استفاده می شود و فرمول بالا را به فرمول زیر تبدیل می کند:

$$P(a \mid v) = \frac{N_{av} + 1}{\sum_{a' \in a_i} (N_{a'v} + 1)}$$

قسمت دوم پیاده سازی پروژه مربوط به تعیین موضوع یک سند جدید است. در این قسمت باید با توجه به یادگیری های مرحله قبل بتوانیم برای یک سند که موضوع آن مشخص نیست با توجه به محتوای داخلی آن موضوعی را برچسب بزنیم.

برای این کار باید به محاسبه احتمال زیر بپردازیم:

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, ..., a_n)$$

همانطور که گفته شد با استفاده از تئوری و قوانین بیز و همچنین فرض ساده بیزی به فرمول زیر خواهیم رسید:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

احتمالات مربوط به $P(v_j)$ و $P(a_i \mid v_j)$ برای هر موضوع و صفت در مرحله قبل محاسبه شده است در نتیجه کافی است با استفاده از فرمول بالا برای هر موضوع این احتمال را محاسبه کنیم و موضوعی که بیشترین مقدار احتمال را دارد به عنوان دسته پیشنهادی برای سند تست برگزینینم.

نکته ای که در پیاده سازی این قسمت مورد توجه باید قرار بگیرد این است که حاصلضرب مقادیر احتمال $P(a_i \mid v_j)$ برای هر صفت در یک موضوع منجر به تولید عدد بسیار کوچکی خواهد شد که خود موجب بروز زیرریز[1] در محاسبه خواهد شد. برای جلوگیری از بروز چنین اتفاقی از لگاریتم گیری استفاده می کنیم؛ با توجه به اینکه داریم: $\log(xy) = \log(x) + \log(y)$ می توانیم بجای ضرب کردن مقادیر احتمال از حاصلجمع مقادیر لگاریتمی آنها استفاده کنیم در نتیجه فرمول قانون ساده بیزی به صورت زیر در خواهد آمد:

$$v_{NB} = \arg\max_{v_j \in V} [\log P(v_j) + \sum_i \log P(a_i \mid v_j)]$$

۵- دادگان مورد استفاده

با توجه به این که تاکنون تلاش‌های بسیار محدودی در زمینه کاربرد روش‌های دسته‌بندی و داده‌کاوی بر روی متون فارسی انجام شده است، تصمیم گرفته شد تا از پیکره[2] همشهری برای آزمون الگوریتم‌های دسته‌بندی استفاده کنیم. مجموعه آزمایش همشهری یکی از معتبرترین منابع در زبان فارسی است. نسخه ۱ همشهری شامل بیش از ۱۶۰۰۰۰سند، ۶۵ درخواست و قضاوت‌های مرتبط است که از سال ۱۳۷۵ تا ۱۳۸۱ توسط افراد مختلف با موضوعات مختلف نوشته شده است. نسخه ۲ همشهری نسبت به نسخه قبل بزرگ‌تر و جامع‌تر است که تصاویر مقالات را نیز در بر دارد. مولفان روزنامه همشهری بصورت دستی مقالات خود را به دسته‌های مختلفی تقسیم کردند و آن را در سایت پیکره همشهری[3] قرار دادند. تمام اسناد در این مجموعه به ۸۲ موضوع مختلف بر اساس دسته اخبار موجود در سایت روزنامه دسته‌بندی شده‌اند. بعنوان مثال siasi به معنی سیاسی است در فارسی و political در انگلیسی است. نام‌های انگلیسی و فارسی دسته‌های اصلی (۱۶ دسته مهم از بین ۸۲ دسته موجود که بیشترین اسناد را دارند) پیکره در جدول زیر آمده است.

لیست 16 دسته مهم پیکره همشهری

تگ دسته	نام دسته به فارسی	نام دسته به انگلیسی
Adabh	هنری-ادبی	Art-Literature
Akhar	اخبار کوتاه	Shoet news

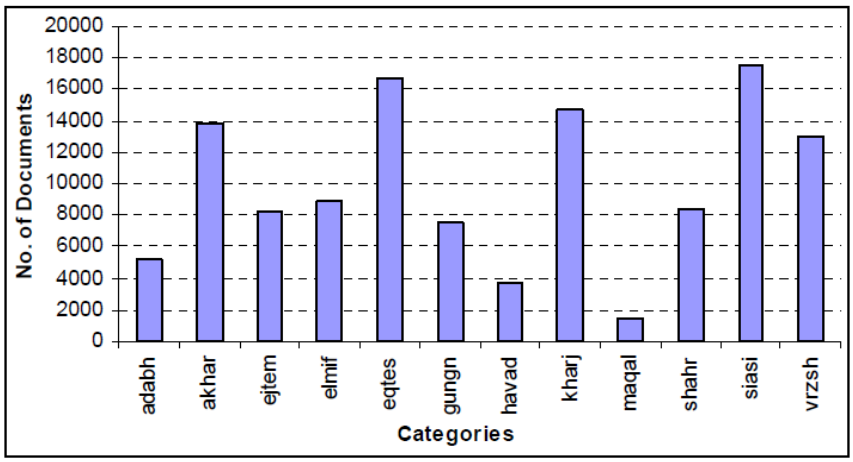
[1] underflow

[2] Hamshahri corpus

[3] آدرس سایت: http://ece.ut.ac.ir/dbrg/Hamshahri/faindex.html

Stock market & Banking	بورس و بانک	Bankb
World’s Economy	اقتصاد جهانی	Econw
Society	اجتماعی	Ejtem
Science & Culture	علمی و فرهنگی	Elmif
Economy (in Iran)	اقتصادی	Eqtes
Tourism	گردش گری	Gards
Miscellaneous	گوناگون	Gungn
Social event	حوادث	Havad
Information Technology	فناوری اطلاعات	Ikaba
Foreign news	اخبار خارجی	Kharj
Tehran & Municipal affairs	شهر تهران	Shahr
Iran cities (except Tehran)	شهرستان‌ها	Shrst
Polotic	سیاسی	Siasi
Sport	ورزشی	Vrzsh

از ۸۲ دسته موجود تنها ۱۲ دسته بیش از ۱۰۰۰ سند دارند که ۷۲ درصد مجموعه را شکل می‌دهند.
شکل زیر توزیع اسناد در ۱۲ دسته اصلی را نشان می‌دهد.



توزیع اسناد پیکره همشهری در ۱۲ دسته اصلی

۶- ارزیابی پروژه و نتیجه گیری

پس از پیاده سازی الگوریتم بیزین ساده ابتدا پیکره آموزشی همشهری به الگوریتم داده خواهد شد. سپس برای تست و ارزیابی آن می توان چند سند دیگر به الگوریتم دارد تا الگوریتم بر اساس یادگیری های گذشته به دسته بندی آنها بپردازد.

برای تست برنامه مجموعه ۶ سند که با موضوعات مختلف به الگوریتم داده شد که الگوریتم توانست دسته یا موضوع ۴ سند را درست پیش بینی کند. در کل با بالا رفتن مقدار سندهای مورد بررسی در مجموعه آموزشی میزان دقت و صحت پیش بینی الگوریتم نیز بالا خواهد رفت.

البته باید در نظر داشت که تقریب مورد استفاده در روش بیزین ساده برای همه داده ها نتایج قابل قبولی نخواهد داشت. زیرا واضح است که فرض استقلال برای کلمات داخل سند درست نخواهد بود به طور مثال احتمال اینکه در یک سند سیاسی کلمه انقلاب به تنهایی بیاید با احتمال اینکه به صورت انقلاب اسلامی بیاید با هم متفاوت خواهد بود. البته می توان برای بهبود این الگوریتم از محاسبه احتمالات مختلف برای عبارات یک، دو و سه کلمه به صورت جداگانه نیز بهره جست.

منابع و مآخذ:[1]

[1] Sahami, M. (1999). Using Machine Learning to Improve Information Access.Ph.D. Thesis,

[2] Lewis, D. (1994). "An introduction to information retrieval." In Proceedings of 17 Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval.

[3] A Yang, Y. (1999). "An evaluation of statistical approaches to text categorization." Journal of Information Retrieval,Vol. 1, No. 1/2, PP.67-88.

[4] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, 2002.

[5] Hinrich Sch¨utze. Introduction to Information Retrieval, Text Classification & Naive Bayes. Institute for Natural Language Processing, Universit¨at Stuttgart, 2009.06.09

[6] Prof. Thomas B. Fomby. K-Nearest Neighbors Algorithm: Prediction and Classification ,Department of Economics Southern Methodist University, February 2008

[7] Tristan Fletcher.Support Vector Machines Explained, March 1, 2009

[8] CHRISTOPHER J.C. BURGES. A Tutorial on Support Vector Machines for Pattern Recognition

[9] Tom M. Mitchell. Machine Learning, GENERATIVE AND DISCRIMINATIVE CLASSIFIERS:NAIVE BAYES AND LOGISTIC REGRESSION

[1] تمامی منابع و ماخذ ذکر شده به صورت ضمیمه پروژه قرار گرفته است