ارتقاي كيفيت دسته‌بندي متون با استفاده از كميته دسته‌بند دو سطحي

سعيد جليلي[*1] و علي‌اكبر صدري[2]

[1]استاديار گروه مهندسي كامپيوتر - دانشكده فني و مهندسي ـ دانشگاه تربيت مدرس

[2]دانش آموخته كارشناسي ارشد گروه مهندسي كامپيوتر - دانشكده فني و مهندسي ـ دانشگاه تربيت مدرس

(تاريخ دريافت ، تاريخ دريافت روايت اصلاح شده ، تاريخ تصويب)

چكيده

امروزه با توجه به رشد روز افزون دسترسي به اسناد الكترونيكي، دسته‌بندي خودكار اهميت ويژه‌اي يافته است. روش‌هاي معمول در اين زمينه، روش‌هاي يادگيري ماشين هستند. روش‌هاي بر اساس كميته كارايي بهتري نسبت به ساير روش‌ها از خود نشان داده‌اند. در اين مقاله، دو ايده در زمينه كميته‌هاي دسته‌بند ارائه شده است. ايده اول برمبناي كميته bagging كه در آن هركدام از اعضاي كميته روي زيرمجموعه‌اي از مجموعه سندهاي آموزشي، ساخته مي‌شوند، ارائه شده است. بر اساس اين ايده، ابتدا مجموعه آموزش با يک روش خوشه‌بندي به چند خوشه، بر اساس مشابهت يا عدم مشابهت كلاس‌هاي اسناد آموزشي، تقسيم مي‌شود و سپس از روي هر خوشه از اسناد، يک دسته‌بند ساخته شده و به عنوان عضو كميته قرار داده مي‌شود (روش يادگيري اعضاي كميته يكسان است). در ايده دوم ساخت كميته‌اي دو سطحي مطرح شده است. در اين ايده اعضاي كميته‌هاي سطح دوم خود كميته‌هايي هستند كه بر مبناي ايده اول ساخته مي‌شوند. در ارزيابي ايده اول مشخص شد كه افزايش كارايي بر مبناي معيار دقت و معيار بازخواني در كميته بيزين ساده نوع اول، بيشتر از ساير كميته‌هاي نوع اول بوده و به ترتيب ۱۲ درصد و ۵/۱ درصد مي‌باشد. ارزيابي‌ها در كميته نوع دوم (برمبناي ايده دوم)، افزايش كارايي را نسبت به كميته نوع اول نشان مي‌دهند كه بيشترين آن نسبت به كميته Rocchio نوع اول بوده، كه معيار دقت را در حدود ۳/۸ درصد و معيار بازخواني را در حدود ۱۸/۸ درصد افزايش داده است.

واژه هاي كليدي: دسته‌بندي متون – Rocchio - بيزين ساده – SVM - خوشه‌بندي متن - كميته دسته‌بندي - كميته bagging

مقدمه

دسته‌بندي خودكار متون به معني انتساب اسناد به كلاس‌هاي از پيش تعيين شده، در ده سال اخير بسيار جلب توجه كرده است. اين مسئله به خاطر رشد دسترسي به اسناد الكترونيكي و نياز مبرم به سازماندهي آنهاست. جالب توجه است در حالي كه دسته‌بندي متون در حوزه بازيابي اطلاعات[1] قرار دارد، اما اغلب روش‌هاي استفاده شده در اين زمينه، روش‌هاي يادگيري ماشين هستند[۱].

بررسي‌ها و مطالعات نشان مي‌دهند كه روش‌هاي دسته‌بندي متون مانند Rocchio ، بيزين ساده و حتي SVM نمي‌توانند كيفيت دسته‌بندي را از يک حدي بيشتر افزايش دهند اما با بهره‌گيري از كميته‌اي متشكل از تعدادي دسته‌بند، مي‌توان كيفيت دسته‌بندي را ارتقا داد [۱]. در همين راستا در اين مقاله دو ايده جديد در تشكيل كميته‌هاي دسته‌بند ارائه شده است.

ايده اول بر مبناي كميته bagging [۲] كه در آن، هركدام از اعضاي كميته روي زير مجموعه‌اي از يک مجموعه آموزش، ساخته مي‌شوند، ارائه شده است. بر اساس اين ايده، مجموعه اسناد آموزشي اوليه با يكي از روش‌هاي خوشه‌بندي به چند خوشه تقسيم مي‌شود. عمل خوشه‌بندي مي‌تواند بر اساس مشابهت يا عدم مشابهت كلاس‌هاي اسناد آموزشي انجام شود. در اين رويكرد از روي هر كدام از خوشه‌هاي ايجاد شده، يک دسته‌بند ساخته شده و به عنوان عضو كميته قرار داده مي‌شود. همه دسته‌بندها با يک روش يادگيري يكسان از روي خوشه‌ها ساخته مي‌شوند. در ادامه، به كميته‌هاي ساخته شده بر مبناي ايده اول، "كميته نوع اول" گفته مي‌شود.

در ايده دوم، هدف اين است كه كميته‌اي تشكيل شود كه هر عضو آن خود يك كميتة دسته‌بند باشد. بنابراين بر اساس ايده اول مي‌توان كميته بيزين نوع اول، كميته Rocchio نوع اول و كميته SVM نوع اول را بوجود آورد و آنها را به عنوان اعضاي يک كميته سطح بالاتر قرار داد. اعضاي كميته سطح دوم كه خود كميته نوع اول هستند،

* نويسنده مسئول : تلفن :

بر مبناي ايده اول ساخته مي‌شوند. در كميته دو سطحي، اعضاي هر كميته نوع اول با استفاده از يک روش يادگيري يكسان (بر مبناي ايده اول) و از روي خوشه‌هاي مجموعه آموزش ساخته مي‌شوند. در ادامه، به كميته‌هاي ساخته شده بر مبناي ايده دوم،"كميته نوع دوم" گفته مي‌شود. دو ايدة ارائه شدة فوق تاكنون در زمينه متن به كار گرفته نشده‌اند. جهت ارزيابي ايده‌هاي پيشنهادي از مجموعه داده‌هاي رويترز ۲۱-۵۷۸ [۴،۳] استفاده شده است.

علاوه بر دو ايدة فوق، در اين مقاله كميته‌اي كه اعضاي آن با روش‌هاي يادگيري متفاوت روي يک مجموعه آموزش يكسان ساخته مي‌شوند (كميته‌هاي نوع سوم)، با داده‌هاي رويتر ۲۱-۵۷۸ آزمون شده‌اند (تا به عنوان مرجعي براي مقايسه با ايده‌هاي فوق استفاده شود). در [۵]، اين نوع كميته‌ها آزمون شده‌اند ولي به دليل غيراستاندارد و كوچک بودن مجموعه آزمون مورد نظر (تعداد ۸۱۴ سند آموزشي و تعداد ۶۸۰ سند آزمايشي)، نتايج بدست آمده كاملاً قطعي و دقيق نبوده است.

پژوهش‌هاي مرتبط

از دسته‌بندي خودكار در مديريت اطلاعات با حجم زياد استفاده مي‌شود [۶]. دسته‌بندي متون نيز تاكنون در زمينه‌هاي مختلفي به كار گرفته شده است. اين زمينه‌ها عبارتند از بايگاني اسناد و مدارک، دسته‌بندي مقالات جديد، دسته‌بندي صفحات وب، پيدا كردن صفحه خانگي يک شخص، يادگيري خودكار علايق مطالعاتي و پژوهشي كاربران، فيلتر كردن خودكار پست الكترونيكي بر اساس محتوا و غيره. يک مرور كلي بر روش‌ها و كارهاي انجام شده در دسته‌بندي متون در [۷،۸] ارائه شده است.

روش‌هاي دسته‌بندي متون

مهمترين روش‌هاي استفاده شده در دسته‌بندي متون عبارتند از: روش بيزين ساده، روش Rocchio، روش KNN[۲] [۸]، روش‌هاي رگرسيون[۹،۱۰]، درخت‌هاي تصميم‌گيري[۵] ، شبكه‌هاي عصبي [۱۱،۱۰]، روش SVM[۳]، روش‌هاي بر اساس قاعده [۱۲،۱۴،۱۳] و روش‌هاي بر اساس كميته. يكي از روش‌هاي معمول يادگيري ماشين، روش بيزين ساده است [۱۵، ۱۶]. طبيعت ساده و احتمالي اين روش، آن را براي توسعه‌هاي مختلف، مناسب نموده است.

روش Rocchio كه بر اساس مشابهت كار مي‌كند بيشتر به حوزة بازيابي اطلاعات متعلق است و تطبيق آن با يادگيري ماشين براي اولين بار در [۱۷] پيشنهاد شده است. يادگيري در اين روش بسيار سريع انجام مي‌شود[۱۸،۸] . روش SVM اولين بار توسط Joachims [۱۹] در حوزه متن استفاده شد. در [۲۱،۲۰،۱۰] حالت خطي و غير خطي اين روش آزمون شده و در حالت خطي كمي بهتر عمل كرده است. بطور كلي اين روش از لحاظ محاسباتي تا اندازه‌اي پيچيده است.

كميته‌هاي دسته‌بند نيز به اشكال متفاوتي در دسته‌بندي متون مورد استفاده قرار گرفته‌اند. در [۲۲] تركيبي از Rocchio، بيزين ساده و KNN هر سه باهم و به صورت دودويي استفاده شده است و آراء آنها با استفاده از روش تركيب خطي وزن داده شده، تركيب شده است. به دليل كوچک بودن مجموعه اسناد آزمايش(۱۸۷ سند) نتايج بدست آمده، از اطمينان كافي برخوردار نيست.

در [۲۵،۲۴،۲۳]، تركيب هاي مختلفي از دسته‌بندهاي SVM ، KNN ، KNN با رويكرد مبتني بر مدل (KNNM) و Rocchio با مجموعه داده‌هاي 20-NewsGroup آزمون شده است و نتايج نشان مي‌دهد كه بهترين تركيب(كميته)، تركيب دو دسته‌بندي بوده كه از نظر كارايي اولين و دومين بوده‌اند (SVM و KNNM) كه علت آن در [۲۳]، تركيب بهترين دسته‌بند، با دسته‌بندي كه بيشترين نزديكي را به آن دارد، بيان شده است. در [۲۵،۲۴،۲۳]، كه آراي دسته‌بندها در كميته‌ها با روش‌هاي متفاوتي تركيب شده است، ميزان بهبود كارايي در بهترين كميته، به طور متوسط ۲،۶۸ درصد بر مبناي معيار F1، نسبت به بهترين دسته‌بند(SVM) بوده است.

در [۲۶] از تركيب دو دسته‌بند به طور سريالي براي دسته‌بندي استفاده شده است كه در آن، دسته‌بند اول كلاس‌هايي را براي سند جديد كانديد مي‌نمايد سپس دسته‌بند دوم، كلاس نهايي سند جديد را از بين كلاس‌هاي كانديد، انتخاب مي‌نمايد. اين روش با ۲۸۵۰ سند (آموزش و آزمون) ارزيابي شده است. نتايج، بهبود نسبي كارايي را نشان مي‌دهد كه به دليل كم بودن تعداد اسناد، نتايج بدست آمده، از اطمينان كافي برخوردار نيست. روش‌هاي boosting [۲۷] و bagging [۲]، روش‌هاي ديگري براي كميته‌سازي هستند. در اين روش‌ها براي تشكيل اعضاي كميته تنها از يک روش يادگيري استفاده مي‌شود و تفاوت اعضاء فقط در مجموعه

آموزشي آنهاست. در روش boosting، چند دسته‌بند با استفاده از يک الگوريتم يادگيري ساخته شده و به عنوان اعضاي کميته قرار داده مي‌شوند. در اين روش، يادگيري دسته‌بندها به صورت ترتيبي انجام مي‌گيرد زيرا دسته‌بند جديد برمبناي عملکرد دسته‌بندهاي ساخته شده قبلي، آموزش داده مي‌شود. در روش bagging نيز براي تشکيل کميته تنها از يک روش يادگيري استفاده مي‌شود، اما با نمونه‌گيري‌هاي متفاوت همراه با جايگزيني از مجموعه آموزش، چندين مجموعه آموزش جديد بدست مي‌آيد که با هر کدام، مي‌توان يک دسته‌بند جداگانه آموزش داده و کميته را بوجود آورد.

روش‌هاي ترکيب آراي اعضاي کميته

روش‌هاي ترکيب آراي اعضاي کميته با توجه به نوع تشکيل کميته، متفاوت هستند. مهمترين روش‌هاي ترکيب آراء در کميته‌هايي که اعضاي آنها با روش‌هاي يادگيري متفاوت ايجاد مي‌شوند عبارتند از: راي اکثريت[۵] (MV)، ترکيب خطي وزن داده شده[۶] (WLC)، انتخاب پوياي دسته‌بند[۷] (DCS) و ترکيب تطبيقي دسته‌بند[۸] (ACC) که در ادامه معرفي مي‌شوند [۸]. ساده‌ترين راه براي ترکيب آراء استفاده از روش راي اکثريت (MV) مي‌باشد، که در آن خروجي دودويي K دسته‌بند جمع‌آوري مي‌شود و کلاسي که حداقل $\frac{k+1}{2}$ راي آورده باشد، به عنوان کلاس سند جديد d انتخاب مي‌گردد. اين روش بيشتر مناسب دسته‌بندي دودويي است [۸]. در روش ترکيب خطي وزن داده شده (WLC) يک مجموع وزن‌دار از آراء اعضاي کميته براي يک سند جديد d بدست مي‌آيد. وزن Wj ميزان موثر بودن نسبي دسته‌بند Φ_j را نشان مي‌دهد. در اين روش بعد از محاسبه نظر کميته در مورد همه کلاس‌ها، محتمل‌ترين کلاس به عنوان کلاس سند جديد d انتخاب مي‌شود [۲۲].

در روش انتخاب پوياي دسته‌بند(DCS) [۵] در کميته $\{\Phi_1 \dots \Phi_i\}$، دسته‌بند Φ_i، که موثرترين دسته‌بند مي‌باشد، انتخاب مي‌گردد و قضاوت آن دسته‌بند پذيرفته مي‌شود. در اين روش ابتدا دقت محلي[۹] دسته‌بند در k تا همسايگي سند جديد d، با استفاده از روش Leave-one-out محاسبه مي‌گردد[۲۸] و سپس دسته‌بندي که بالاترين دقت محلي را دارد، انتخاب مي‌گردد و راي آن دسته‌بند پذيرفته مي‌شود. وزن هر همسايه d، معمولاً اندازه ميزان مشابهت کسينوسي يا فاصله اقليدسي بين آن همسايه و سند جديد d مي‌باشد. در روش ترکيب تطبيقي دسته‌بند(ACC) به جاي انتخاب يک دسته‌بند با بيشترين دقت محلي براي سند جديد d، کلاسي به d نسبت داده مي‌شود که در ميان همه دسته‌بندها بيشترين دقت محلي را دارد [۵]. اين روش بينابين دو روش قبلي يعني WLC و DCS مي‌باشد. روش‌هاي ديگري براي ترکيب اطلاعات بدست آمده از منابع مختلف (ترکيب آرا) وجود دارد که تحت عنوان روش‌هاي هم‌جوشي تصميم[۱۰] [۳۰و۲۹] مطرح مي‌شوند که مهمترين آنها عبارتند از تئوري شهود، منطق فازي، تئوري اطلاعات و شبکه‌هاي عصبي.

خوشه‌بندي اسناد متني

خوشه‌بندي اسناد، بيشتر در حوزه‌هاي متن‌کاوي[۱۱] و بازيابي اطلاعات کاربرد دارد. در ابتدا از خوشه‌بندي اسناد براي افزايش ميزان معيار دقت و معيار بازخواني در سيستم‌هاي بازيابي اطلاعات استفاده مي‌شد [۳۲و۳۱]. همچنين از خوشه‌بندي براي طبقه‌بندي اسناد وب استفاده مي‌گردد [۱۶]. برخي از رويکردهاي خوشه‌بندي نيز براي پيدا کردن خوشه‌هاي طبيعي در يک طبقه موجود از اسناد و استفاده از اين خوشه‌هاي طبيعي در توليد يک دسته‌بند متن مؤثر، کاربرد دارند [۳۳].

روش‌هاي HAC[۱۲] و K-means، دو روش خوشه‌بندي هستند که در خوشه‌بندي اسناد متني کاربرد زيادي دارند. اگر چه روش HAC کند است ولي عملکرد آن از روش K-means بهتر است[۳۴].

مفاهيم دسته‌بندي متون

در اين بخش شيوه نمايش متن، روش‌هاي وزن‌دهي به خصيصه‌ها، روش‌هاي انتخاب خصيصه و معيارهاي ارزيابي کارآيي دسته‌بند ارائه شده است.

نمايش متون و پردازش‌هاي اوليه

براي استفاده از اسناد متني جهت يادگيري، ابتدا مراحل پيش پردازش، شاخص‌گذاري و کاهش ابعاد روي اسناد متني انجام گرفته و سپس ساخت دسته‌بند و در نهايت دسته‌بندي صورت مي‌گيرد. در مرحله پيش پردازش معمولا سه عمل حذف کلمات زائد[۱۳]، حذف برچسب‌ها و ريشه‌يابي ساده روي کلمات اسناد صورت مي‌گيرد. در مرحله شاخص‌گذاري در واقع شيوه نمايش سند تعيين مي‌گردد. معمول‌ترين شيوه نمايش، استفاده از

مدل فضاي برداري است كه يک روش غير معنايي است. روش‌هاي متعددي [۷،۸،۳۵] براي وزن‌دهي به كلمات يک سند براساس تكرار آنها وجود دارد كه عبارتند از: وزن‌دهي دودويي، وزن‌دهي با تكرار كلمات، وزن‌دهي TFIDF، وزن‌دهي TFC، وزن‌دهي LTC و وزن‌دهي آنتروپي. روش‌هاي وزن‌دهي با تكرار كلمات و TFIDF را كه در اين پژوهش مورد استفاده قرار گرفته‌اند، به طور مختصر شرح مي‌دهيم (با فرض اينكه N تعداد كل اسناد مجموعه آموزش، f_{ik} تعداد دفعات تكرار كلمه i در سند k ، n_i تعداد تكرار كلمه i در كل مجموعه اسناد و a_{ik} وزن كلمه i در سند k باشد). در روش وزن‌دهي با تكرار كلمات، براي وزن‌دهي به كلمات هر سند، از تعداد تكرار كلمه در سند استفاده مي‌شود (يعني $a_{ik} = f_{ik}$). در روش وزن‌دهي TFIDF، وزن كلمه i در سند k با توجه به تعداد دفعات تكرار كلمه در سند و نسبت معكوس تعداد سندهايي از مجموعه، كه كلمه در آنها حداقل يک بار تكرار شده است، محاسبه مي‌شود (يعني $a_{ik} = f_{ik} \times \log(\frac{N}{n_i})$). در مرحله كاهش ابعاد بردارهاي اسناد متني از يكي از روش‌هاي انتخاب خصيصه استفاده مي‌شود[۳۶،۳۷،۳۸]. در روش‌هاي انتخاب خصيصه، سعي بر اين است كه بهترين خصيصه‌ها انتخاب شوند و خصيصه‌هايي كه حاوي اطلاعات نيستند حذف گردند. مهمترين روش‌هاي انتخاب خصيصه [۳۶،۳۷،۳۸] عبارتند از: روش آستانه‌يابي تكرار سند(DF)[۱۴] ، روش بهرة اطلاعاتي(IG)[۱۵] ، روش اطلاعات متقابل (MI)۱۶ ، روش CHI، روش ضريب همبستگي[۱۷] و روش SCHI . در اين پژوهش براي انتخاب خصيصه از روش آستانه‌يابي تكرار سند(DF) استفاده شده است كه در آن تكرار سند (DF) براي هر كلمه، برابر با تعداد اسنادي است كه كلمه در آنها ظاهر شده است. براي هر كلمه واحد در مجموعه سندهاي آموزش اين ضابطه محاسبه شده و در نهايت كلماتي كه بالاترين آستانه را دارند، انتخاب مي‌شوند.

معيارهاي ارزيابي كارايي دسته‌بند

براي ارزيابي كارايي دسته‌بند معمولاً از معيار دقت[۱۸] (P)، معيار بازخواني[۱۹] (R)، معيار F [۷] و معيار نقطه تعادل[۲۰] [۸]، استفاده مي‌شود. وقتي كارايي دسته‌بندها را بررسي مي‌كنيم، كميت‌هاي زير براي هر كلاس مورد توجه قرار مي گيرند: a) تعداد اسنادي كه متعلق به كلاسي تشخيص داده شده‌اند و واقعاً هم متعلق به آن كلاس هستند، b) تعداد اسنادي كه متعلق به كلاسي تشخيص داده شده‌اند اما واقعاً متعلق به آن كلاس نيستند، c) تعداد اسنادي كه متعلق به كلاسي تشخيص داده نشده‌اند اما واقعاً متعلق به آن كلاس هستند. با استفاده از اين كميت‌ها، رابطه‌هاي (۱)، (۲) و (۳) تعريف مي‌شوند:

$$P = a/(a+c) \quad (۱)$$

$$R = a/(a+b) \quad (۲)$$

$$F1 = \frac{2*P*R}{P+R} \quad (۳)$$

رابطه (۱) ميزان دقت در انجام دسته‌بندي را نشان مي‌دهد، در حاليكه رابطه (۲) ميزان كامل بودن مجموعه يافت شده را نشان مي‌دهد. رابطه (۳) تركيبي از دقت و بازخواني را ارائه مي‌دهد. براي بدست آوردن ميانگين كارايي دسته‌بند در كل كلاس‌ها، دو روش استاندارد با نام‌هاي ميكروميانگين[۲۱] و ماكروميانگين[۲۲] وجود دارد.

در روش ميكروميانگين اگر به فرض n كلاس وجود داشته باشد ابتدا a_i ، b_i و c_i براي تمامي كلاس‌ها بدست مي‌آيد. سپس مقادير $a = \sum_{i=1}^{n} a_i$ ، $b = \sum_{i=1}^{n} b_i$ و $c = \sum_{i=1}^{n} c_i$ محاسبه شده و ميزان كارايي دسته‌بند در كل كلاس‌ها محاسبه مي‌گردد. در روش ماكروميانگين، ابتدا معيار كارايي دسته‌بند در هر كلاس محاسبه شده سپس ميانگين اين مقادير محاسبه مي‌شود [۷]. در واقع محاسبه ميكروميانگين و ماكروميانگين نتايج دسته‌بندهاي دودويي خود نوعي تركيب آرا محسوب مي‌شود. محاسبه ميكروميانگين نتايج چند دسته‌بند كه روي زير مجموعه‌هاي يک مجموعه سندهاي آموزش ساخته مي‌شوند، يكي از روش‌هاي ارزيابي دسته‌بندي است.

هنگامي كه معيارهاي دقت، بازخواني و F1 به تنهايي استفاده مي‌شوند، ممكن است گمراه كننده باشند لذا براي ارزيابي دسته‌بندهاي متني معمولاً از معيار نقطه تعادل كه معيار مناسب‌تري براي ارزيابي كارايي دسته‌بندي است، استفاده مي شود. نحوه محاسبه نقطه تعادل در [۸] آمده است. نقطه تعادل مي‌تواند در هر كلاس(T_i) به طور مجزا محاسبه شود يا به طور كلي براي تمام كلاس‌ها (T) يک

نقطه تعادل تعيين گردد. در هر دو حالت اگر احتمال محاسبه شده توسط دسته‌بند، براي تعلق سند جديد d به يک كلاس، كمتر از T_i (يا T) باشد، تعلق سند d به آن كلاس False و در غير اين صورت True مي‌شود. در اين پژوهش نقطه تعادل براي هر كلاس اسناد، به طور مجزا و بر اساس احتمالات توليد شده توسط دسته‌بند براي هر كلاس، محاسبه شده است. يادآوري مي‌گردد دسته‌بند بيزين ساده، دسته‌بند Rocchio به طور كامل و بخش‌هايي از دسته‌بند SVM از طريق برنامه‌نويسي پياده‌سازي شد. براي ارزيابي كارايي دسته‌بندهاي پايه، كميته نوع اول و كميته نوع دوم از معيار دقت (P)، معيار بازخواني (R)، معيار F1 و معيار نقطه تعادل كه در اين بخش تشريح شدند، استفاده كرده‌ايم.

كميته نوع اول (ايده اول)

بر اساس اين ايده مجموعه آموزش با يكي از روش‌هاي خوشه‌بندي به چند خوشه تقسيم مي‌شود. عمل خوشه‌بندي مي‌تواند بر اساس مشابهت يا عدم مشابهت كلاس‌هاي اسناد آموزشي انجام شود. از روي هر كدام از خوشه‌هاي ايجاد شده يك دسته‌بند ساخته شده و به عنوان عضو كميته قرار داده مي‌شود. همه دسته‌بندها با يک روش يادگيري يكسان از روي خوشه‌ها ساخته مي‌شوند. هدف اين ايده ايجاد دسته‌بندهاي تخصصي و قرار دادن آنها در يک كميته، براي افزايش كارايي دسته‌بندي مي‌باشد. در اينجا براي تشكيل اين نوع كميته از روش‌هاي يادگيري بيزين ساده، Rocchio و SVM استفاده شده است يعني كميته بيزين، كميته Rocchio يا كميته SVM ايجاد شده است [٣٩،٤٠]. لازم به ذكر است كه ايجاد كميته دسته‌بند يک سطحي (كميته نوع اول) با اعضايي كه از روي خوشه‌هاي اسناد آموزشي بر اساس معيار عدم مشابهت بين كلاس‌ها يادگيري شده‌اند براي اولين بار در اين مقاله مطرح، استفاده و اثر آن روي دسته‌بندي متون ارزيابي شده است. در ادامه اين بخش، ساخت دسته‌بندهاي پايه، خوشه‌بندي مجموعه آموزش، يادگيري دسته‌بندها از روي خوشه‌ها و تشكيل كميته نوع اول و به روز رساني اعضاي كميته ارائه شده است.

ساخت دسته‌بند بيزين ساده

دسته‌بند بيزين ساده با استفاده از داده‌هاي مجموعه آموزش و دريافت مقادير خصيصه‌هاي يک سند جديد d به تخمين احتمال تعلق آن سند به هر كلاس c_j، بر اساس رابطه (۴) با فرض استقلال كلمات مي‌پردازد [۱۵]:

$$p(c_j \mid d) = p(c_j) \prod_{k=1}^{M} p(w_k \mid c_j) \qquad (۴)$$

در رابطه (۴)، M طول بردار(تعداد كلمات) سند جديد d است و با داشتن مقادير $p(c_j)$ و $p(w_k \mid c_j)$ مي‌توان احتمال $p(c_j \mid d)$ را محاسبه كرد. پس از محاسبه احتمال كلاس‌ها يعني $p(c_j \mid d)$، محتمل‌ترين كلاس، به عنوان كلاس سند d انتخاب مي‌شود. به عبارت ديگر كلاس سند d يعني $C_{\max}$، با استفاده از رابطه (۵) بدست مي‌آيد:

$$C_{\max} = \arg\max_{c_j} p(c_j) \prod_{k=1}^{M} p(w_k \mid c_j) \qquad (۵)$$

از آنجا كه تعداد كلمات سند جديد d ممكن است خيلي زياد باشد و احتمالات فوق‌الذكر نيز همه اعدادي كوچكتر از يک هستند، در نتيجه حاصلضرب اين احتمالات، عددي بسيار كوچک خواهد شد. با توجه به اين كه اكثر زبان‌هاي برنامه‌نويسي قادر به نمايش و پياده‌سازي اين اعداد نيستند، در رابطه (۴) روي دو طرف آن، تابع ln اعمال شده و در نتيجه در پياده‌سازي اين روش از رابطه (۶) استفاده شده است:

$$\ln p(c_j \mid d) = \ln(p(c_j)) + \sum_{k=1}^{M} \ln p(w_k \mid c_j) \qquad (۶)$$

در رابطه (۶) اعداد كوچک اولاً به اعداد منفي تبديل شده‌اند و ثانياً به جاي ضرب باهم جمع شده‌اند كه در نتيجه اعداد قابل فهم‌تر و قابل قياس‌تري بدست مي‌آيد.

در ساخت دسته‌بند بيزين ساده براي وزن‌دهي به خصيصه‌ها از روش وزن‌دهي با تكرار كلمات و براي انتخاب خصيصه از روش آستانه‌يابي تكرار سند (DF) استفاده شده است.

ساخت دسته‌بند Rocchio

اين روش بر اساس مشابهت كار مي‌كند. براي هر كلاس با استفاده از ميانگين‌گيري روي همه بردارهاي اسناد آموزشي متعلق به آن كلاس، يک بردار نمونه محاسبه مي‌شود. به طور رسمي بردار ميانگين c_j، در مجموعه اسناد S متعلق به كلاس c_j، بر اساس رابطه

(۷) محاسبه مي‌شود [۴۱،۸] :

$$C_j = \frac{1}{|S|}\sum_{i=1}^{|S|} doc_i \quad , \quad doc_i \in c_j$$

(۷)

كه در آن doc_i بردار سند آموزشي متعلق به كلاس c_j است. براي دسته‌بندي يک سند جديد d مشابهت آن سند با بردارهاي نمونه متناظر با هر كلاس اندازگيري مي‌شود. ميزان مشابهت [۴۱،۸] را مي‌توان با اندازه‌گيري مشابهت كسينوسي و ضرب داخلي بردار يا مشابهت ژاكارد بدست آورد و سپس شبيه‌ترين كلاس را به عنوان كلاس سند d اعلام نمود. به عبارت ديگر كلاس سند d يعني $C_{\max}$، با استفاده از رابطه (۸) بدست مي‌آيد:

$$C_{\max} = \arg\max_{c_j} Sim(C_j, d)$$

(۸)

كه در آن $Sim(C_j, d)$، مشابهت بين بردار سند جديد d و بردار نمونه متناظر با كلاس c_j را نشان مي‌دهد.

در ساخت اين دسته‌بند براي وزن‌دهي به كلمات از روش وزن‌دهي TFIDF استفاده شد تا اهميت كمتري به كلماتي كه در بيشتر اسناد تكرار شده‌اند، داده شود و براي انتخاب خصيصه از روش آستانه‌يابي تكرار سند(DF) استفاده شد.

ساخت دسته‌بند SVM

ماشين‌هاي SVM جزء روش‌هاي مجزاساز[۲۴] هستند و با استفاده از داده‌هاي آموزشي سعي مي‌كنند بهترين n بعدي را كه داده‌هاي مثبت در يک طرف آن و داده‌هاي منفي در طرف ديگر آن قرار مي‌گيرند، بدست آورند. اين سطح تصميم‌گيري با استفاده از تعداد كمي نمونه آموزشي كه به آنها بردارهاي پشتيبان[۲۵] مي‌گويند قابل انجام است. حسن SVM اين است كه احتياج به انتخاب خصيصه ندارد و با استفاده از يک سري پارامتر پيش فرض، رسيدن به بهترين كارايي را ميسر مي‌سازد[۱۹]. در اين پژوهش براي دسته‌بند SVM، از نسخه تحت ويندوز ابزار آمادة SVM_Light [۱۹] استفاده شد. سندهاي آموزش از نوع مثبت (متعلق به كلاس) يا منفي (غير متعلق به كلاس) هستند. معمولاً به تعداد نمونه‌هاي مثبت، از ساير كلاس‌ها نمونه منفي انتخاب مي‌شود. در اين پژوهش، انتخاب نمونه‌هاي منفي به صورت تصادفي انجام شده است.

سندهاي آزمون نيز از نوع مثبت يا منفي هستند. البته مثبت يا منفي بودن سندهاي آزمون فقط در ارزيابي كارايي دسته‌بند استفاده مي‌شود. تعداد سندهاي آزمون منفي نيز مي‌تواند مساوي با تعداد سندهاي آزمون مثبت، باشد كه در اين حالت نتايج دسته‌بند SVM بسيار خوب است. ولي براي قابل مقايسه شدن روش SVM با ساير روش‌ها (بيزين ساده و Rocchio)، براي هر كلاس، سندهاي آزمون آن كلاس به عنوان نمونه مثبت، و سندهاي آزمون همه كلاس‌هاي ديگر به عنوان نمونه منفي در نظر گرفته شد. يعني همانند ساير دسته‌بندها، بايد تعلق همه سندهاي مجموعه آزمون را به يک كلاس بررسي كنيم و نه بخشي از آن را. نتايج يكايک دسته‌بندهاي مستقل كلاس‌ها در پايان باهم جمع گرديد تا دسته‌بند به صورت ميكروميانگين [۷] ارزيابي شود. يادآوري مي‌گردد تركيب نتايج دسته‌بندهاي مستقل كلاس‌ها از طريق برنامه‌نويسي پياده‌سازي شد. در اين دسته‌بند براي وزن‌دهي به خصيصه‌ها، از روش وزن‌دهي با تكرار كلمات استفاده كرديم و همانطور كه اشاره شد روش SVM نياز به انتخاب خصيصه ندارد.

خوشه‌بندي مجموعه آموزش

براي خوشه‌بندي مجموعه سندهاي آموزش ابتدا با استفاده از روش Rocchio [۴۱،۸] ، بردارهاي نمونه كلاس‌هاي اسناد آموزشي را بدست مي‌آوريم آنگاه پس از بدست آوردن مشابهت دو به دو كلاس‌ها و تشكيل ماتريس مشابهت، با استفاده از الگوريتم خوشه‌بندي HAC [۲۸،۴۲]، خوشه‌بندي را بر اساس مشابهت (يا عدم مشابهت) انجام مي‌دهيم. يعني اسناد كلاس‌هايي در يک خوشه قرار مي‌گيرند كه بيشترين شباهت (يا عدم شباهت) را با همديگر دارند. در الگوريتم HAC ميزان مشابهت بين دو خوشه كه هر كدام از آنها فقط حاوي يک كلاس است، برابر با ميزان مشابهت بين دو كلاس خواهد بود، در غير اين صورت مشابهت بين دو خوشه بر اساس رابطه (۹) برابر با ميزان مشابهت زوجي از اعضاي دو خوشه خواهد بود كه نسبت به مشابهت ساير زوج‌هاي ممكن بين دو خوشه، كمترين مقدار را داشته باشد.

$$sim(cluster_m, cluster_n) = \min(sim(c_i, c_j))$$
$$c_i \in cluster_m \;\&\; cj \in cluster_n$$

(۹)

براي خوشه‌بندي مجموعه سندهاي آموزش بر اساس عدم مشابهت، از عكس مشابهت دو كلاس يعني رابطة (۱۰) استفاده شده است:

$$dissim\ (c_i, c_j) = 1 - sim\ (c_i, c_j)$$

(١٠)

هدف اين است كه كلاس‌هايي در يک خوشه قرار گيرند كه بيشترين افتراق را باهم دارند. بنابراين براي انجام اين نوع خوشه‌بندي، در الگوريتم HAC از عدم مشابهت($dissim$)به جاي مشابهت(sim)استفاده شده است. لازم به ذكر است كه رويكرد عدم مشابهت در خوشه‌بندي اسناد متني، براي اولين بار در اين مقاله مطرح، استفاده و اثر آن روي دسته‌بندي متون ارزيابي شده است. علاوه بر رويكرد مشابهت و عدم مشابهت در مرحلة خوشه‌بندي، براي يادگيري يک دسته‌بند موثر و كارا از روي اسناد آموزشي، عمل فيلتر كردن سندهاي آموزشي نامرتبط با كلاس‌هاي اسناد آموزشي انجام شد. در عمل فيلتر كردن، سندهايي را كه برچسب يک كلاس را داشتند ولي ميزان شباهت (تعلق) آنها به آن كلاس از يک حد آستانه‌اي كمتر بود از زير مجموعه اسناد آن كلاس حذف و در كلاس جداگانه‌اي قرار داده شدند. البته آن دسته از سندهايي كه متعلق به يک كلاس بودند ولي به اشتباه برچسب كلاس ديگري را داشتند، شناسايي و برچسب كلاس آنها اصلاح گرديد.

يادگيري دسته‌بندها و تشكيل كميته نوع اول

با يكي از روش‌هاي يادگيري بيزين ساده، Rocchio و SVM از روي هر خوشه، يك دسته‌بند ساخته شده و به عنوان عضو كميته قرار داده مي‌شود. الگوريتم تشكيل كميته نوع اول در شكل (١) ارائه شده است. در اين الگوريتم، اعضاي كميته كه با يک روش يادگيري يكسان ساخته مي‌شوند، هر كدام روي تعدادي كلاس مشخص از كلاس‌ها با اتكا بر زيرمجموعه‌اي از سندهاي آموزش تخصص دارند. بنابراين به جاي داشتن يک متخصص براي كل مجموعه سندهاي آموزش، تعداد K متخصص وجود خواهد داشت كه هر كدام روي بخشي از مجموعه سندهاي آموزش مهارت دارند و در نتيجه دسته‌بندي يك سند جديد d با داشتن متخصص‌هاي $\Phi_1,...,\Phi_k$، انجام مي‌شود. هر كدام از دسته‌بندهاي $\Phi_1,...,\Phi_k$ را مي‌توان قبل از تشكيل كميته با مجموعه آزمون مربوط به كلاس‌هايي كه دسته‌بند روي آنها ساخته شده است، ارزيابي كرد. يادآوري مي‌گردد در اين پژوهش، كميته نوع اول بر اساس خوشه‌هاي حاوي بيشترين شباهت بين اسناد خوشه و يا بر اساس خوشه‌هاي حاوي كمترين شباهت بين اسناد خوشه ساخته شد.

تركيب آراي اعضاي كميته نوع اول

در كميته نوع اول چون هر كدام از اعضاء، بر مبناي زيرمجموعه‌اي (به عبارت ديگر خوشه‌اي) از مجموعه آموزش ساخته شده‌اند، هر عضو كميته، نظرش را در مورد تعلق سند جديد d به تمامي كلاس‌هايي كه روي آنها تخصص دارد اعلام مي‌كند و در نتيجه در مورد تعلق سند d به كلاس‌هاي ديگر نظري نخواهد داشت. در اين پژوهش براي تصميم‌گيري (تركيب آراء) در كميته بيزين ساده و كميته Rocchio از مقدار آستانه T_i هر كلاس c_i كه به منظور بدست آوردن نقطه تعادل در ارزيابي اعضاي كميته، محاسبه گرديده، استفاده شده است. يعني در دسته‌بندي يک سند جديد اگر احتمال توليد شده براي كلاس c_i، مساوي يا بيشتر از مقدار آستانه T_i آن باشد، كلاس c_i توسط كميته به سند جديد نسبت داده مي‌شود. در كميته SVM ، كلاس c_i كه ماكزيمم احتمال را در بين ساير كلاس‌ها دارد به يک سند جديد نسبت داده مي‌شود[٣٩].

الگوريتم تشكيل كميته نوع اول:

ورودي: ١- يک مجموعه اسناد آموزشي با كلاس‌هاي مشخص.
٢- k : تعداد خوشه‌ها.
٣- L : روش يادگيري

خروجي: كميته‌اي از دسته‌بندها كه با روش يادگيري L از روي خوشه‌ها ساخته شده‌اند.

- بردار كلمات (vector space) مجموعه آموزش را تشكيل بده.
- اسناد آموزشي را به برداري از خصيصه‌ها تبديل كن.
- انتخاب خصيصه را بر اساس يكي از روش‌هاي موجود انجام بده.
- ماتريس مثلثي مشابهت كلاس‌هاي اسناد آموزشي را تشكيل بده.
- با استفاده از الگوريتم خوشه‌بندي HAC، مجموعه آموزش را بر اساس مشابهت (يا عدم مشابهت) كلاس‌ها به k خوشه تبديل كن.
- $i \leftarrow 0$
- تا وقتي كه $i < k$ است مراحل زير را تكرار كن:
  - $i \leftarrow i + 1$
  - دسته‌بند Φ_i را روي خوشه i ام با روش يادگيري L بساز.
  - دسته‌بند Φ_i را به عنوان عضوي از كميته قرار بده.
- كميته‌اي را با دسته‌بندهاي $\Phi_1,...,\Phi_k$ برگردان.

شكل١ : الگوريتم تشكيل كميته نوع اول.

به روز رساني اعضاي كميته

استفاده از دسته‌بندهايي كه خاصيت افزايشي[٢٦] دارند، در كميته مي‌تواند بستر مناسبي براي يادگيري فعال باشد

چرا كه در يادگيري فعال با اضافه شدن هر سند يا مجموعه سند جديد بايد دسته‌بندها را به روز رساني كرد.

با ورود مجموعه سندهاي جديد، دسته‌بندهاي بيزين ساده و Rocchio با هزينه اندكي به روز رساني مي‌شوند [۳۹]. در گذشته دسته‌بند SVM براي ديدن تاثير يک سند جديد بايد از اول آموزش مي‌يافت، اما در [۴۳] روشي ارائه شده كه اين دسته‌بند نيز با هزينه كمتري به روز رساني مي‌شود.

كميته نوع دوم (ايده دوم)

روش‌هاي دسته‌بندي پايه مانند بيزين ساده، Rocchio ، SVM و... هر كدام كارايي خاصي در دسته‌بندي متن دارند و هيچكدام از اين روش‌هاي پايه، بطور كامل روش ديگري را پوشش نمي‌دهد(به عبارت ديگر يک روش نسبت به روش ديگر در دسته‌بندي متن برتري كامل ندارد). در نتيجه نمي‌توان از بين اين روش‌ها فقط بهترين روش را انتخاب كرد. از آنجا كه در كميته نوع اول، براي يادگيري اعضاي كميته فقط از يک روش دسته‌بندي استفاده مي‌گردد، در نتيجه به منظور ارتقاي كارايي كميته و بهره‌گيري از توان تمامي روش‌هاي دسته‌بندي پايه ايدة كميتة دوسطحي (كميته نوع دوم) و چگونگي ايجاد آن در اين مقاله براي اولين مرتبه در حوزة متن مطرح شد. اعضاي كميته نوع دوم را كميته‌هاي نوع اول كه هر كدام بر اساس يک روش دسته‌بندي پايه جداگانه ايجاد مي‌شود، تشكيل مي‌دهند. ابتدا براساس ايده اول مي‌توان كميته بيزين ساده، كميته Rocchio و كميته SVM را بوجود آورد و آنها را به عنوان اعضاي يک كميته سطح بالاتر قرار داد. ايده كميتة نوع دوم بيشتر بر اين مبنا استوار است كه اگر بر مبناي يك مجموعه آموزش، كميته بيزين ساده بهتر از دسته‌بند بيزين ساده و كميته Rocchio بهتر از دسته‌بند Rocchio و كميته SVM بهتر از دسته‌بند SVM عمل مي‌كند، بنابراين كميته‌اي از اين كميته‌ها، كارايي دسته‌بندي را افزايش خواهد داد. الگوريتم تشكيل كميته نوع دوم در شكل (۲) ارائه شده است[۳۹]. از اينجا به بعد، اعضاي كميته نوع دوم را كه خود كميته نوع اول هستند، كميته‌هاي فرعي مي‌ناميم.

تركيب آرا در كميته نوع دوم

با توجه به اينكه در كميته نوع دوم، روش يادگيري اعضاي كميته متفاوت است لذا براي تركيب آراء در اين نوع كميته، مي‌توان از يكي (f) از روش‌هاي تركيب آراء در كميته‌اي كه دسته‌بندهاي آن با روش‌هاي يادگيري متفاوت ساخته مي‌شوند، استفاده كرد. در اين پژوهش براي تركيب آراي اعضاي كميته نوع دوم از روش راي اكثريت استفاده شده است بنابراين براي هر سند جديد d هيچ يا فقط يك كلاس مشخص اعلام مي‌گردد.

الگوريتم كميته نوع دوم:

ورودي: ۱- k : تعداد اعضاي كميته.

۲- $l_1,\ldots,l_k$:روش‌هاي يادگيري مورد نظر براي كميته‌هاي فرعي.

خروجي: كميته‌اي از كميته‌ها

- $i \leftarrow 0$
- تا وقتي كه $i < k$ است مراحل زير را تكرار كن:
  - $i \leftarrow i+1$
  - كميته فرعي $comm_i$ را با استفاده از الگوريتم كميته نوع اول و با روش يادگيري l_i بساز.
  - كميته $comm_i$ را به عنوان عضوي از كميته قرار بده.
- كميته‌اي را با كميته‌هاي فرعي $comm_1,\ldots,comm_k$ برگردان.

شكل۲ : الگوريتم تشكيل كميته‌اي از كميته‌ها (كميته نوع دوم).

ارزيابي كميته‌هاي نوع اول و نوع دوم

همه مراحل روش پيشنهادي (مشتمل بر روش‌هاي وزن‌دهي به خصيصه‌ها، انتخاب خصيصه، خوشه‌بندي، دسته‌بندهاي پايه به غير از دسته‌بند SVM، محاسبه معيارهاي ارزيابي و تصميم‌گيري در كميته‌ها) براي انعطاف‌پذيري بيشتر از طريق برنامه‌نويسي پياده‌سازي شده‌اند. همچنين كميته‌اي كه با روش‌هاي يادگيري متفاوت اما از روي يک مجموعه آموزش واحد ساخته مي‌شود (كه در اين مقاله، كميته نوع سوم نامگذاري مي‌شود) پياده‌سازي و آزمون شده است تا نتايج اين نوع كميته با نتايج روش‌هاي پيشنهادي مقايسه شود.

مراحل مختلف ايجاد كميته نوع اول و كميته نوع دوم روي مجموعه داده‌هاي استاندارد رويترز ۲۱-۵۷۸ آزمون شده است. در اين ارزيابي از گونه ModApte مجموعه رويترز ۲۱-۵۷۸ استفاده شده است كه بيشتر از ساير تقسيم‌بندي‌هاي مجموعه، در كارهاي ديگران استفاده شده است. تعداد اسناد داراي برچسب و غير خالي اين مجموعه حدود ۱۲۳۵۰ سند مي‌باشد كه بر اساس

تقسيم‌بندي انجام شده در خود اين مجموعه، تقريباً ٧٠% اسناد براي آموزش و ٣٠% اسناد براي آزمون در نظر گرفته شده است.

ارزيابي دسته‌بندهاي پايه

نتايج حاصل از ارزيابي دسته‌بندهاي بيزين ساده و Rocchio با ده كلاس اول رويترز كه داراي بيشترين تعداد سند مي‌باشند در جدول (١) و با همه كلاس‌ها در جدول (٢) به طور خلاصه ارائه شده است. در ارزيابي اين دو نوع دسته‌بند از دو رويكرد استفاده شده است. در رويكرد اول از ماكزيمم احتمال تعلق يا شباهت (Max Approach) و در رويكرد دوم از نقطه تعادل در هر كلاس براي تعيين تعلق سند به كلاس استفاده شده است.

نتايج دسته‌بند SVM در حالتي كه در آزمون دسته‌بند هر كلاس، فقط اسناد آزمون آن كلاس به علاوه همان تعداد سند از كلاس‌هاي ديگر استفاده مي‌شود، بسيار خوب است. نتايج اين حالت در جدول (٣) ارائه شده است. اما همانطور كه گفته شد براي قابل مقايسه شدن نتايج روش دسته‌بند SVM با دو روش ديگر دسته‌بندي، دسته‌بند هر كلاس، با همه اسناد مجموعه آزمون، ارزيابي شده و سپس نتايج باهم تركيب شده‌اند.

جدول ١ : نتايج آزمون بيزين ساده و Rocchio روي ده كلاس اول رويتر.

	Use Max Approach		Use Break Even Point Per Class	
	R	P	R	P
Rocchio	0.86	0.95	0.8	0.8
Naïve Bayesian	0.85	0.93	0.87	0.87

جدول ٢ : نتايج آزمون بيزين ساده و Rocchio روي همه كلاس‌هاي رويتر.

	Use Max Approach		Use Breakeven Point Per Classes	
	R	P	R	P
Rocchio	0.7	0.86	0.7	0.7
Naïve Bayesian	0.65	0.80	0.66	0.64

جدول ٣ : نتايج آزمون SVM با مجموعه داده‌هاي رويتر. دسته‌بند هر كلاس با اسناد آزمون آن كلاس به علاوه همان تعداد اسناد از كلاس‌هاي ديگر، آزمون شده است.

	For first 10 Classes		For All Classes	
	R	P	R	P
SVM	0.96	0.97	0.92	0.97

در تركيب نتايج آزمون دسته‌بندها در SVM ، از دو رويكرد استفاده شده است. در رويكرد اول كلاسي كه بالاترين مقدار را به يک سند آزمون نسبت داده است، انتخاب مي‌گردد. در رويكرد دوم همه كلاس‌هاي اختصاص يافته به يک سند آزمون پذيرفته مي‌شوند. نتايج حاصل(ميكروميانگين) از ارزيابي SVM با ده كلاس اول رويتر كه داراي بيشترين تعداد سند مي‌باشند در جدول (۴) و با همه كلاس‌ها در جدول (۵) ارائه شده است. ستون رويكرد اول و دوم در جدول‌هاي (۴) و (۵) به ترتيب با عنوان Max Approach و Multi Class Approach مشخص شده‌اند.

جدول ۴ : نتايج آزمون SVM روي ده كلاس اول رويتر. دسته‌بند هر كلاس با همه اسناد مجموعه آزمون، ارزيابي شده است.

	With Max Approach		With Multi Class Approach	
	R	P	R	P
SVM	0.85	0.95	0.93	0.85

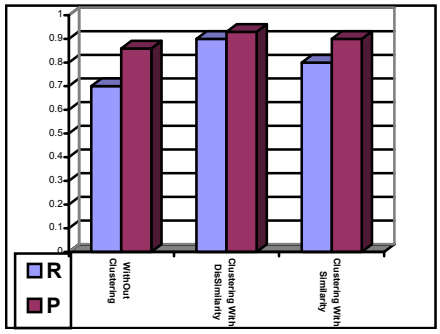
جدول ۵ : نتايج آزمون SVM روي همه كلاس‌هاي رويتر. دسته‌بند هر كلاس با همه اسناد مجموعه آزمون، ارزيابي شده است.

	With Max Approach		With Multi Class Approach	
	R	P	R	P
SVM	0.68	0.90	0.81	0.64

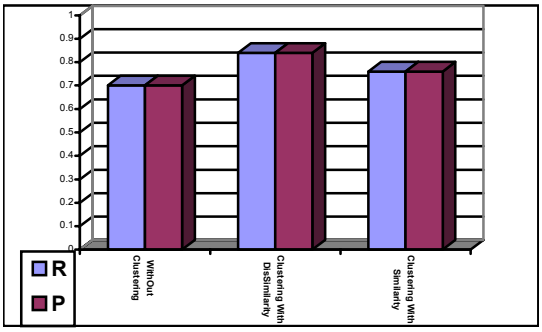
ارزيابي دسته‌بندهاي پايه ايجاد شده از روي خوشه‌ها

مجموعه سندهاي آموزش گونه ModApte به پنج خوشه بر مبناي مشابهت و به پنج خوشه ديگر بر مبناي عدم مشابهت كلاس‌هاي اسناد، تقسيم گرديد. از روي هر خوشه يک دسته‌بند ساخته شد و با مجموعه آزمون مربوط به همان خوشه ارزيابي گرديد. ارزيابي‌ها نشان داد كه كارايي دسته‌بندهاي يادگيري شده از روي خوشه‌هاي بر اساس معيار عدم مشابهت نسبت به دسته‌بندهاي يادگيري شده از روي خوشه‌هاي بر اساس معيار مشابهت براي تمامي روش‌هاي دسته‌بندي پايه بيشتر است. يعني در خوشه‌هايي كه كلاس‌هاي موجود در آنها بيشترين افتراق را باهم دارند نتايج كارايي دسته‌بندهاي اين خوشه‌ها بهتر است. نمودارهاي (١) الي (۶)، نتايج حاصل از ميكروميانگين دسته‌بندهاي پنج خوشه را با دسته‌بندي كه از روي كل مجموعه سندهاي آموزش ساخته شده

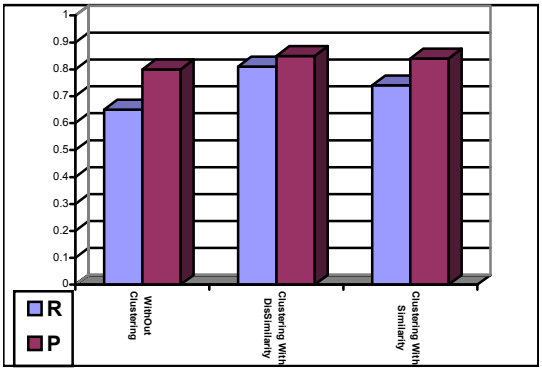
است به تفکیک روش مقایسه مي‌کنند. بر مبناي اين نمودارها ميکرومیانگین نتايج دسته‌بندهاي خوشه‌ها، هم در خوشه‌هايي که بر اساس مشابهت و هم در خوشه‌هايي که بر اساس عدم مشابهت ايجاد شده‌اند بهتر از نتايج دسته‌بند ساخته شده از روي کل مجموعه آموزش است.



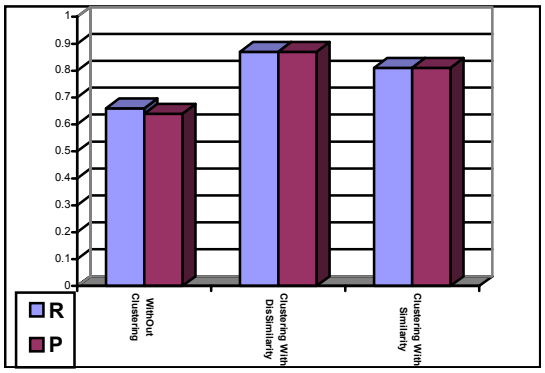
نمودار ۱: مقايسه ميکرومیانگین نتايج دسته بندهاي پنج خوشه، با دسته بندي که از کل مجموعه آموزش ساخته شده است (با روش يادگيري Rocchio و روش دسته‌بندي Max Approach).



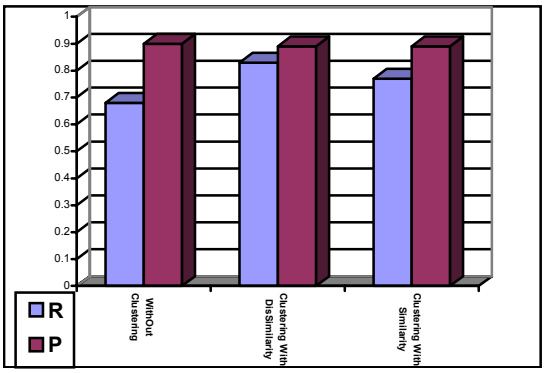
نمودار ۲: مقايسه ميکرومیانگین نتايج دسته‌بندهاي پنج خوشه، با دسته‌بندي که از کل مجموعه آموزش ساخته شده است (با روش يادگيري Rocchio و روش دسته‌بندي Breakeven per Classes).



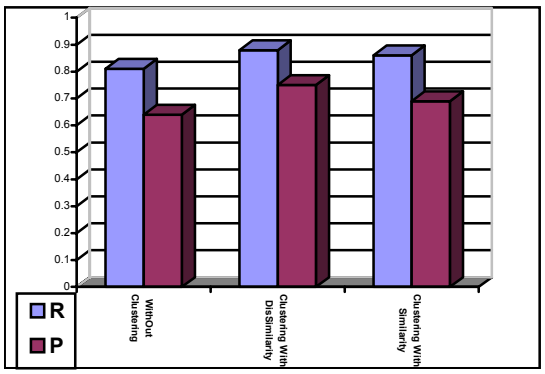
نمودار ۳ : مقايسه ميکرومیانگین نتايج دسته‌بندهاي پنج خوشه، با دسته‌بندي که از کل مجموعه آموزش ساخته شده است (با روش يادگيري بيزين ساده و روش دسته‌بندي Max Approach).



نمودار ۴ : مقايسه ميکرومیانگین نتايج دسته‌بندهاي پنج خوشه، با دسته‌بندي که از کل مجموعه آموزش ساخته شده است (با روش يادگيري بيزين ساده و روش دسته‌بندي Breakeven per Classes).



نمودار ۵: مقايسه ميکرومیانگین نتايج دسته‌بندهاي پنج خوشه، با دسته‌بندي که از کل مجموعه آموزش ساخته شده است (با روش يادگيري SVM و روش دسته‌بندي Max Approach).



نمودار ۶: مقايسه ميکرومیانگین نتايج دسته‌بندهاي پنج خوشه، با دسته‌بندي که از کل مجموعه آموزش ساخته شده است (با روش يادگيري SVM و روش دسته‌بندي Multi Class Approach).

ارزيابي انواع کميته‌هاي نوع اول

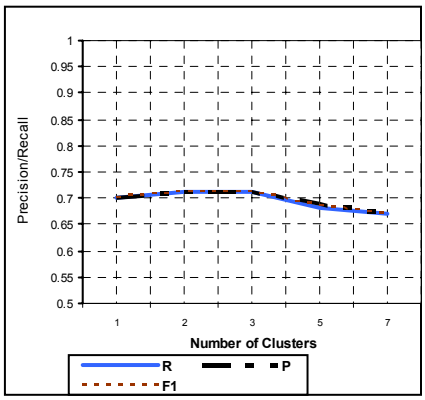
در اين مرحله مجموعه آموزش گونه ModApte به دو، سه، پنج و هفت خوشه بر اساس مشابهت و عدم

مشابهت تقسيم گرديد و بر مبناي خوشه‌هاي ايجاد شده كارايي (بر مبناي معيار بازخواني و معيار دقت) دسته‌بندي به ترتيب در كميته‌هاي دو عضوي، سه عضوي، پنج عضوي و هفت عضوي بررسي گرديد. روند تغيير معيارهاي كارايي (R و P) در اين چهار كميته براي سه روش دسته‌بندي پايه بيزين ساده، Rocchio و SVM در نمودارهاي (٧) الي (١٢) نشان داده شده است. در اين نمودارها، محور Y نشان دهنده تغييرات معيارهاي كارايي(R و P) و محور X نشان دهندة تعداد خوشه‌هاي ايجاد شده (تعداد اعضاي كميته) از روي مجموعه سندهاي آموزش، مي‌باشد. نتايج دسته‌بندي كه از روي تمامي اسناد آموزش ساخته مي‌شود (كميته با يک عضو)، با تعداد يک خوشه در نمودارهاي (٧) الي (١٢) مشخص شده است. براي مشخص كردن كلاس (يا كلاس‌هاي) سند جديد توسط هر عضو كميته نوع اول، در هر سه كميته Rocchio نوع اول، بيزين ساده نوع اول و SVM نوع اول، از رويكردي استفاده شده است كه با توجه به جداول ١ الي ٥ بالاترين كارايي را داشته است.

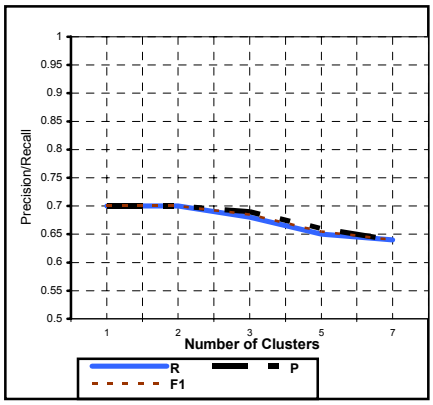
نمودارهاي (٧) الي (١٠) نشان مي‌دهند كه در كميته‌هاي بيزين ساده و Rocchio، اولاً كارايي در كميته‌هاي بر مبناي مشابهت كلاس‌ها، نسبت به كميته‌هاي بر مبناي عدم مشابهت كلاس‌ها، بهتر است. يعني اگر چه تک‌تک دسته‌بندهاي ايجاد شده در خوشه‌هاي بر مبناي عدم مشابهت، نسبت به دسته‌بندهاي ايجاد شده در خوشه‌هاي بر مبناي مشابهت، كارايي بهتري دارند ولي در اين دو كميته نتيجه عكس مي‌باشد و اين مسئله به خاطر اين است كه اولاً در بين اسناد خوشه‌هاي ايجاد شده بر مبناي مشابهت، رويهم افتادگي كمتري نسبت به خوشه‌هاي ايجاد شده بر مبناي عدم مشابهت وجود دارد ثانياً وقتي در اين دو كميته، تعداد خوشه‌ها (يا تعداد اعضاي كميته) از يك حدي بيشتر مي‌شود، كارايي (بر مبناي معيار F1) كاهش مي‌يابد و اين مسئله در كميته‌هاي بر مبناي عدم مشابهت شديدتر است.

در كميته SVM برخلاف دو كميته ديگر وضعيت متفاوت است. با توجه به نمودارهاي (١١) و (١٢) كارايي (بر مبناي معيار F1) در كميته‌هاي بر مبناي عدم مشابهت كلاس‌ها، نسبت به كميته‌هاي بر مبناي مشابهت كلاس‌ها، بهتر است و با افزايش تعداد خوشه‌هاي بر مبناي عدم مشابهت، كارايي نيز افزايش محسوسي پيدا كرده و سپس كاهش مي‌يابد در حاليكه كارايي(بر مبناي معيار F1) در كميته SVM برمبناي مشابهت از ابتدا سير نزولي دارد. به نظر مي‌رسد قدرت جداكنندگي دسته‌بند SVM وقتي كه با خوشه‌هاي برمبناي عدم تشابه آموزش داده مي‌شود بيشتر است و رويهم افتادگي در خوشه‌ها، تاثير چنداني در نتيجه ندارد. همان طور كه در نمودار (٩) مشخص است كميته بيزين ساده در مقايسه با دسته‌بند بيزين ساده، كارايي بهتري را در مقايسه نسبت فوق‌الذكر در ساير كميته‌ها نشان مي‌دهد. علاوه بر آن تغيير تعداد خوشه‌هاي بر مبناي مشابهت، روي معيار دقت در مقايسه با معيار بازخواني تاثير زيادتري دارد. كميته‌هاي بيزين ساده با پنج عضو بر اساس مشابهت، Rocchio با سه(يا دو) عضو بر اساس مشابهت و SVM با سه عضو بر اساس عدم مشابهت، بهترين كارايي را از خود نشان داده‌اند. مقايسه نتيجه ارزيابي كارايي در اين كميته‌ها با دسته‌بندي كه از روي كل مجموعه آموزش ساخته مي‌شود، در نمودارهاي (١٣) الي (١٥)، نشان داده شده است. نمودار(١٣) افزايش مقدار معيار بازخواني و مقدار معيار دقت را ١/٢ درصد نشان مي‌دهد و اين مبين آن است كه كارايي در كميته Rocchio نوع اول نسبت به دسته‌بند Rocchio بهبود چنداني ندارد. نمودار (١۴)، در كميته بيزين ساده نوع اول، افزايش مقدار معيار بازخواني را ۵/١ درصد و افزايش مقدار معيار دقت را ١٢ درصد در مقايسه با دسته‌بند بيزين ساده نشان مي‌دهد. بر مبناي نمودار (١۵)، در كميته SVM نوع اول، مقدار معيار بازخواني ۶ درصد افزايش، مقدار معيار دقت ٢ درصد كاهش و مقدار معيار F1، ٣ درصد افزايش را، نسبت به دسته‌بند SVM داشته است. بنابراين كارايي در كميته بيزين ساده نوع اول نسبت به دو كميته نوع اول ديگر بهبود نسبتاً خوبي دارد اگر چه هنوز دسته‌بند SVM و كميته SVM از روش‌هاي ديگر كارايي بهتري دارند. بر مبناي ارزيابي‌هاي انجام شده و نتايج بدست آمده رفتار كميته دسته‌بند نوع اول (به ويژه كميته بيزين و Rocchio) نسبت به رفتار دسته‌بندهاي پايه با استفاده از خوشه‌بندي، كاملاً متفاوت است. در كميته نوع اول ميزان بهبود كارايي در بهترين كميته (كميته SVM) روي داده‌هاي رويترز ٢١-۵٧٨ از گونه ModApte، به طور متوسط ٣ درصد بر مبناي معيار F1، نسبت به بهترين دسته‌بند(SVM) مي‌باشد در حالي كه در [٢۵،٢۴،٢٣]، ميزان بهبود كارايي روي داده‌هاي 20-

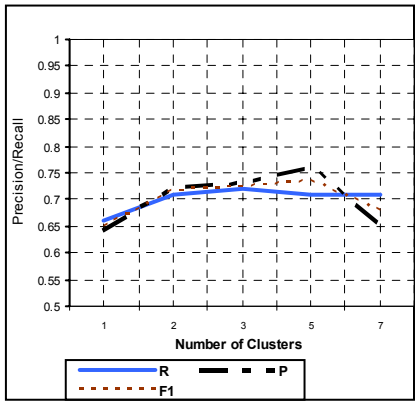
NewsGroup ، به طور متوسط ۲,۶۸ درصد بر مبناي معيار F1، بوده است.



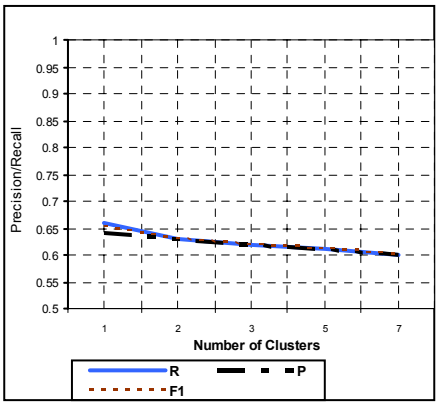
نمودار ۷ : روند تغيير معيارهاي كارايي در كميته هاي Rocchio نوع اول با يک، دو، سه، پنج و هفت عضو. اعضاي هر كميته از روي خوشه‌هاي ايجاد شده بر مبناي مشابهت، ساخته شده‌اند.



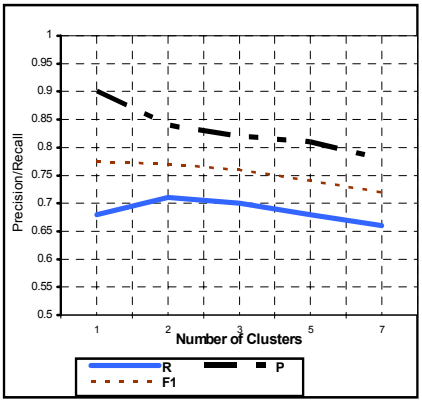
نمودار ۸ : روند تغيير معيارهاي كارايي در كميته هاي Rocchio نوع اول با يک، دو، سه، پنج و هفت عضو. اعضاي هر كميته از روي خوشه‌هاي ايجاد شده بر مبناي عدم مشابهت، ساخته شده اند.



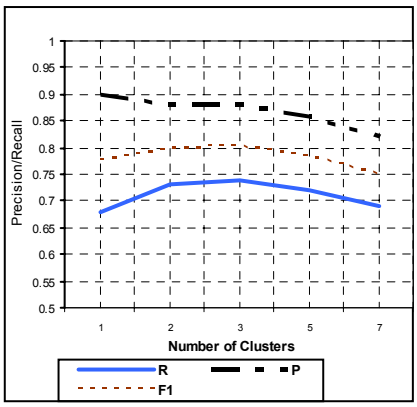
نمودار ۹ : روند تغيير معيارهاي كارايي در كميته‌هاي بيزين ساده نوع اول با يک، دو، سه، پنج و هفت عضو. اعضاي هر كميته از روي خوشه‌هاي ايجاد شده بر مبناي مشابهت، ساخت شده‌اند.



نمودار ۱۰ : روند تغيير معيارهاي كارايي در كميته‌هاي بيزين ساده نوع اول با يک، دو، سه، پنج و هفت عضو. اعضاي هر كميته، از روي خوشه‌هاي ايجاد شده بر مبناي عدم مشابهت، ساخته شده‌اند.



نمودار ۱۱: روند تغيير معيارهاي كارايي در كميته‌هاي SVM نوع اول با يک، دو، سه، پنج و هفت عضو. اعضاي هر كميته از روي خوشه‌هاي ايجاد شده بر مبناي مشابهت، ساخته شده‌اند.

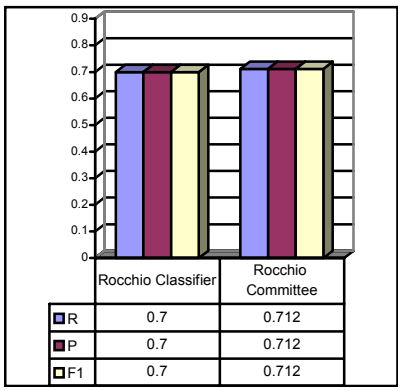


نمودار ۱۲ : روند تغيير معيارهاي كارايي در كميته‌هاي SVM نوع اول با يک، دو، سه، پنج و هفت عضو. اعضاي هر كميته، از روي خوشه‌هاي ايجاد شده بر مبناي عدم مشابهت، ساخته شده‌اند.

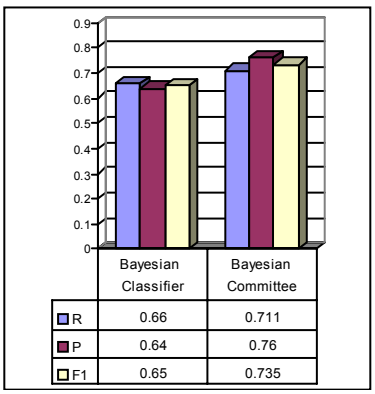
ارزيابي كميته نوع دوم

كميته نوع دوم با سه كميته فرعي (كميته نوع اول)

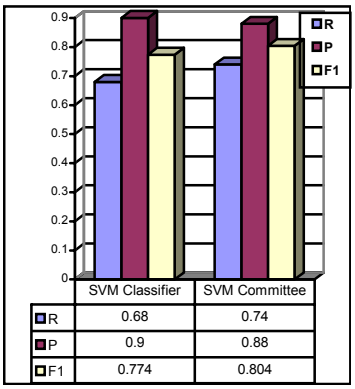
بيزين ساده، Rocchio و SVM تشكيل شد. در ساخت كميته نوع دوم، از كميته‌هاي فرعي كه كارايي بهتري داشتند استفاده گرديد. در اين پژوهش براي تركيب آراء در كميته از روش راي اكثريت استفاده شده است. نتايج حاصل از ارزيابي كميته نوع دوم با استفاده از نمودارهاي (١٦) الي (١٨) ارائه شده است. اين نمودارها نشان مي‌دهند كه كارايي در كميته نوع دوم، از كارايي تك‌تك اعضاي آن (كميته‌هاي فرعي) بهتر است. نمودار (١٦) افزايش مقدار معيار بازخواني را ٣/٨ درصد و افزايش مقدار معيار دقت را ١٨/٨ درصد نسبت به كميته Rocchio نوع اول نشان مي‌دهد. نمودار (١٧) افزايش مقدار معيار بازخواني را ٣/٩ درصد و افزايش مقدار معيار دقت را ١٤ درصد نسبت به كميته بيزين ساده نوع اول نشان مي‌دهد. بر مبناي نمودار (١٨)، مقدار معيار بازخواني فقط يک درصد و مقدار معيار دقت فقط ٢ درصد نسبت به كميته SVM نوع اول، افزايش يافته است و اين نشان مي‌دهد كه كارايي در كميته نوع دوم نسبت به كميته SVM نوع اول به مقدار خيلي كمي بهبود يافته است ولي نسبت به دو كميته نوع اول ديگر بهبود نسبتاً خوبي داشته است. نتايج بدست آمده از ارزيابي كميتة نوع دوم و مقايسه آن با نتايج كميته‌هاي نوع اول مجدداً اين بحث را كه هيچكدام از روش‌هاي دسته‌بندي پايه و كميته‌هاي نوع اول برتري صددرصد نسبت به ديگر روش‌ها ندارند را تاييد مي‌كند و در نتيجه ايجاد كميته نوع دوم (يا كميته دسته‌بند دو سطحي) را به منظور افزايش كارايي دسته‌بندي توجيه مي‌كند.



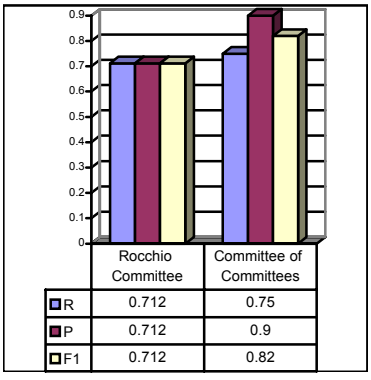
نمودار١٣ : مقايسه كارايي كميته Rocchio نوع اول با دسته‌بند Rocchio ، بر اساس مجموعه رويتر. كميته با سه(يا دو) عضو از روي خوشه‌هاي ايجاد شده بر مبناي مشابهت، ساخته شده است.



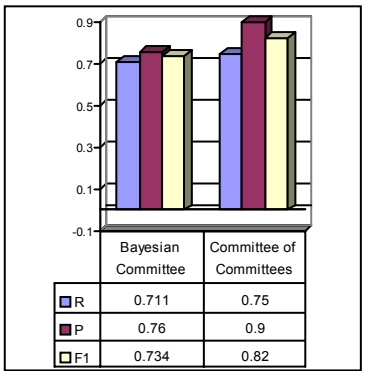
نمودار١٤ : مقايسه كارايي كميته بيزين ساده نوع اول با دسته‌بند بيزين ساده، بر اساس مجموعه رويتر. كميته با پنج عضو از روي خوشه‌هاي ايجاد شده بر مبناي مشابهت، ساخته شده است.



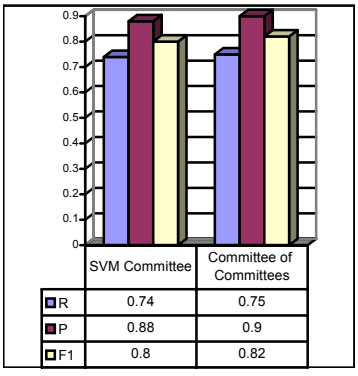
نمودار١٥ : مقايسه كارايي كميته SVM نوع اول با دسته‌بند SVM ، بر اساس مجموعه رويتر. كميته با سه عضو از روي خوشه‌هاي ايجاد شده بر مبناي عدم مشابهت، ساخته شده است.



نمودار١٦ : مقايسه كارايي كميته نوع دوم با كميته Rocchio نوع اول، بر اساس مجموعه داده رويتر.



نمودار١٧: مقايسه كارايي كميته نوع دوم با كميته بيزين ساده نوع اول، بر اساس مجموعه داده رويتر.



نمودار ١٨ : مقايسه كارايي كميته نوع دوم با كميته SVM نوع اول، بر اساس مجموعه داده رويتر.

در کمیته نوع دوم میزان بهبود کارایي روي داده‌هاي رویتر۲۱-۵۷۸ از گونه ModApte، به طور متوسط ۴,۶ درصد بر مبناي معیار F1، نسبت به بهترین دسته‌بند (SVM) مي‌باشد در حالي که در [۲۵،۲۴،۲۳]، میزان بهبود کارایي روي داده‌هاي 20-NewsGroup، به طور متوسط ۲,۶۸ درصد بر مبناي معیار F1، بوده است.

ارزیابي کمیته نوع سوم

کمیته نوع سوم (یعني کمیته یک سطحي که اعضاي آن با روش‌هاي یادگیري متفاوت ولي از روي یک مجموعه سندهاي آموزش ساخته مي‌شوند) براي مقایسه با روش‌هاي پیشنهادي و آزمون با مجموعه داده‌هاي استاندارد پیاده‌سازي و آزمون شد. در واقع این کمیته به منظور بررسي جایگاه کمیته نوع سوم در مقایسه با کمیته‌هاي نوع اول و نوع دوم یادگیري شده است.

این کمیته با سه عضو یعني دسته‌بندهاي بیزین ساده، Rocchio وSVM تشکیل شد و از روش راي اکثریت (MV)، براي ترکیب آراي آن استفاده گردید. نتایج حاصل براي ده کلاس اول رویتر و براي همه کلاس‌هاي رویتر با دو رویکرد مختلف در جدول‌هاي (۶) و (۷) ارائه شده است. این جدول‌ها نشان مي‌دهند که نتیجه حاصل از کمیته نوع سوم بر مبناي معیار F1، در همه حالات از نتیجه تک‌تک اعضاي آن بهتر است.

بهبود نتایج نسبت به دسته‌بند SVM چندان زیاد نیست ولي نسبت به دو دسته‌بند دیگر خوب است. این کمیته در حالتي که دسته‌بندهاي آن، از رویکرد ماکزیمم احتمال تعلق یا شباهت (Max Approach) براي دسته‌بندي استفاده مي‌کنند، نتیجه بهتري را بر اساس معیارF1 از خود نشان مي‌دهد.

مقایسه نتایج کمیته‌هاي نوع اول، نوع دوم و نوع سوم

در نمودار (۱۹) نتایج ارزیابي کمیته‌هاي نوع اول، نوع دوم و نوع سوم روي مجموعه رویتر ۲۱-۵۷۸ با گونه ModApte و با همه کلاس‌هاي این مجموعه، باهم مقایسه شده‌اند. براي مقایسه کمیته نوع سوم با سایر کمیته‌ها، نتایج آن هم بر مبناي جدول (۶) و هم بر مبناي جدول (۷) که در نمودار (۱۹) به ترتیب با عنوان "Committee 3-1 " و "Committee 3-2" مشخص شده‌اند، مورد استفاده قرار گرفته است. کمیته نوع دوم که در نمودار با عنوان "Committee 2" مشخص شده است نسبت به کمیته نوع سوم (Committee 3-1) بر اساس معیار F1، ۲ درصد افزایش کارایي را نشان مي‌دهد. در ضمن کمیته SVM نوع اول، در مقایسه با دو کمیته نوع اول دیگر کارایي بهتري از خود نشان مي‌دهد و با کمیته‌هاي نوع دوم و سوم قابل مقایسه است. براساس معیار F1، کارایي کمیته SVM نوع اول با کمیته نوع سوم یکسان است ولي در مقایسه با کمیته نوع دوم، ۲ درصد کارایي کمتري دارد. بنابراین کمیته نوع دوم بر مبناي مجموعه داده‌هاي رویتر نسبتاً از کمیته‌هاي دیگر بهتر عمل مي‌کند.

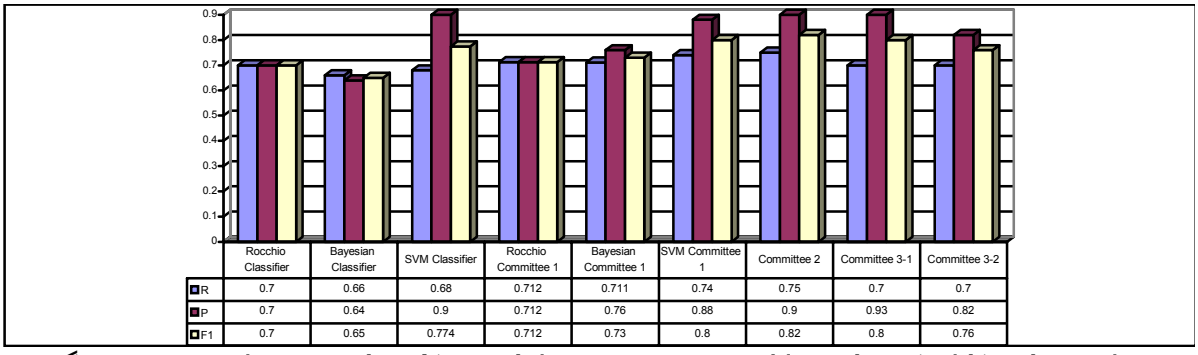
جدول۶ : مقایسه نتیجه کمیته نوع سوم با نتایج سه عضو آن. هر کدام از اعضاي کمیته از رویکرد ماکزیمم احتمال تعلق یا شباهت براي دسته‌بندي استفاده مي‌کنند .

		10 First Classes			All Classes		
		R	P	F1	R	P	F1
Use Max Approach	Rocchio	0.86	0.95	0.90	0.7	0.86	0.77
	Naïve Bayesian	0.85	0.93	0.89	0.65	0.80	0.72
	SVM	0.85	0.95	0.90	0.70	0.90	0.79
Committee Use MV Method		0.87	0.97	0.92	0.70	0.93	0.80

جدول۷ : مقایسه نتیجه کمیته نوع سوم با نتایج سه عضو آن. دسته‌بندهاي بیزین ساده و Rocchio از رویکرد نقطه breakeven در هر کلاس و دسته‌بند SVM از رویکرد چند کلاسي براي دسته‌بندي استفاده مي‌کنند.

		First 10 Classes			All Classes		
		R	P	F1	R	P	F1
Use Breakeven Point Per Classes	Rocchio	0.80	0.80	0.80	0.70	0.70	0.70
	Naïve Bayesian	0.87	0.87	0.87	0.66	0.64	0.65
SVM Use Multi Class Approach		0.93	0.85	0.89	0.81	0.65	0.72
Committee Use MV Method		0.90	0.89	0.90	0.70	0.82	0.76

کل ارزیابي‌ها با توجه به نمودار(۱۹) نشان مي‌دهد که رویکرد کمیته(حتي کمیته نوع سوم) براي افزایش کارایي دسته‌بندي، یک رویکرد کاملاً مؤثر است. در مجموع کمیته نوع دوم با توجه به نمودار (۱۹) نسبت به دسته‌بندهاي پایه، کمیته نوع اول و کمیته نوع سوم کارایي بهتري دارد.



	Rocchio Classifier	Bayesian Classifier	SVM Classifier	Rocchio Committee 1	Bayesian Committee 1	SVM Committee 1	Committee 2	Committee 3-1	Committee 3-2
R	0.7	0.66	0.68	0.712	0.711	0.74	0.75	0.7	0.7
P	0.7	0.64	0.9	0.712	0.76	0.88	0.9	0.93	0.82
F1	0.7	0.65	0.774	0.712	0.73	0.8	0.82	0.8	0.76

نمودار ۱۹: مقايسه كارايي كميته‌هاي نوع اول، نوع دوم و نوع سوم بر اساس همه كلاس‌هاي مجموعه داده رويتر ۵۷۸-۲۱ گونه ModeApte.

نتيجه‌گيري و كارهاي آينده

در اين مقاله، مفاهيم، روش‌ها و كارهاي انجام شده در زمينه دسته‌بندي متون بررسي و دو ايده جديد در اين زمينه ارائه شد. در ايدة اول، تشكيل كميته‌هاي يك سطحي (كميته‌هاي نوع اول) از طريق تقسيم مجموعه اسناد آموزشي به چندين خوشه و يادگيري يك دسته‌بند از روي هر خوشه (با يك روش يادگيري يكسان) و قرار دادن دسته‌بندهاي ايجاد شده در يك كميته، مطرح شد. در ايده دوم تشكيل كميته‌هاي دو سطحي(كميته‌هاي نوع دوم)كه اعضاي آن كميته‌هاي ايجاد شده بر اساس ايدة اول هستند، مطرح شد. با توجه به مراجعي كه در آنها روش‌هاي گوناگون مطرح يا مقايسه شده‌اند، به طور كلي نمي‌توان گفت روشي هميشه برتر از ديگر روش‌ها عمل مي‌كند، زيرا بستگي به بستر داده‌ها نيز دارد و تغييرات مختلفي كه روي روش‌ها مي‌تواند اعمال شود آنها را از نظر كارايي بيشتر به هم نزديک مي‌كند. البته روش‌هاي بر اساس كميته كه از چند دسته‌بند به جاي يک دسته‌بند استفاده مي‌كنند و سپس روش SVM اكثراً بهتر از بقيه عمل كرده‌اند. ارزيابي كميته نوع اول نشان داد كه كارايي در كميته‌هاي بيزين ساده نوع اول و Rocchio نوع اول بر مبناي مشابهت كلاس‌ها، نسبت به كميته‌هاي بر مبناي عدم مشابهت كلاس‌ها، بهتر است زيرا رويهم افتادگي اسناد در بين خوشه‌هاي بر مبناي مشابهت، نسبت به خوشه‌هاي بر مبناي عدم مشابهت، كمتر است. در كميته SVM نوع اول برخلاف دو كميته ديگر وضعيت متفاوت بوده و نشان مي‌دهد كه رويهم افتادگي خوشه‌ها در قدرت جداكنندگي دسته‌بند SVM، تاثير چنداني ندارد. كميته بيزين ساده نوع اول در مقايسه با دسته‌بند بيزين ساده، كارايي نسبتاً خوبي دارد (معيار بازخواني ۵/۱ درصد و معيار دقت ۱۲ درصد بهبود يافته است) در حالي كه اين بهبود در دو نوع كميته ديگر ملاحظه نمي شود.

نتايج حاصل از ارزيابي كميته نوع دوم نشان مي‌دهد كه كارايي اين نوع كميته از كارايي تک‌تک اعضاي آن، بهتر است. بيشترين افزايش در كميته نوع دوم، نسبت به كميته Rocchio نوع اول بوده، كه مقدار معيار بازخواني را ۳/۸ درصد و مقدار معيار دقت را ۱۸/۸ درصد بهبود داده است. كارايي در كميته نوع دوم نسبت به كميته SVM نوع اول، مقدار بهبود خيلي كمي را نشان مي‌دهد كه در آن مقدار معيار بازخواني فقط يک درصد و مقدار معيار دقت فقط ۲ درصد نسبت به كميته SVM نوع اول، افزايش يافته است. در مقايسه انواع كميته‌ها مشخص گرديد كه كميته نوع دوم نسبت به كميته نوع سوم بر اساس معيار F1، ۲ درصد افزايش كارايي را نشان مي‌دهد. در ضمن كميته SVM نوع اول، در مقايسه با دو كميته نوع اول ديگر كارايي بهتري از خود نشان مي‌دهد و با كميته‌هاي نوع دوم و سوم قابل مقايسه است. كميته‌هاي دسته‌بندي معرفي شده در [۲۳،۲۴،۲۵،۲۶،۲۲] (كه از نظر نوع تشكيل، مشابه با كميته‌هاي اين مقاله مي‌باشند) و كميته‌هاي دسته‌بندي ارائه شده در اين مقاله، نشان مي‌دهند كه به طور كلي كميته‌هاي دسته‌بندي متون، كارايي دسته‌بندي را بر اساس معيارهاي كارايي بهبود مي‌دهند. هر چند كه كميته‌هاي ارائه شده در [۲۶،۲۲] روي مجموعه داده‌هاي استاندارد آزمون نشده‌اند و كميته‌هاي ارائه شده در [۲۴،۲۵،۲۳] فقط روي داده‌هاي 20-NewsGroup آزمون شده‌اند، آزمون كميته‌هاي ارائه شده در اين مقاله با داده‌هاي رويترز ۵۷۸-۲۱ از گونه ModApte (كه استانداردترين مجموعه داده در حوزه دسته‌بندي متون است) بهبود كارايي دسته‌بندي (توسط

کمیته‌هاي دسته‌بندي) را بر اساس معیارهاي کارایي، مطمئن‌تر و قابل اعتمادتر مي‌نماید. اگر کمیته‌هاي معرفي شده در [۲۳،۲۴،۲۵،۲۶،۲۲] با مجموعه داده‌هاي استاندارد (به ویژه داده‌هاي رویتر ۵۷۸-۲۱ از گونه ModApte) آزمون مي‌شدند، نتایج آزمون آنها با نتایج آزمون کمیته‌هاي این مقاله، قابل قیاس‌تر مي‌شد. البته به نظر مي‌رسد نتایج به دست آمده از آزمون کمیته‌هاي این مقاله، با توجه به روش‌هاي پیشنهاد شده و مجموعه داده‌هاي در نظر گرفته شده براي آزمون آنها، وضعیت بهتري نسبت به نتایج کمیته‌هاي معرفي شده در [۲۶،۲۲،۲۳،۲۴،۲۵] دارد. به طور کلي چون مرحله یادگیري کمیته‌هاي معرفي شده در این مقاله، فقط یک بار و مراحل بروز رساني آنها به تعداد خیلي کم انجام مي‌شود از این جهت مي‌توان بحث هزینه را در مراحل یادگیري و بروز رساني رها کرد. اما هزینه مرحله دسته‌بندي مي‌تواند قابل بحث باشد. تخصیص کلاس به یک سند جدید d (دسته‌بندي سند جدید) به صورت برخط خیلي مهم است که بایستي هم سریع و هم دقیق باشد. در کمیته نوع اول ترکیب آرا، یک مسئله مهم و قابل پژوهش است، چرا که وجود یک روش مناسب در این زمینه، مي‌تواند کارایي کمیته را به میزان قابل ملاحظه‌اي افزایش دهد. چگونگي تقسیم مجموعه آموزش به زیرمجموعه‌هاي کوچکتر و موثرتر به نحوي که بهترین کارایي را در دسته‌بندي ایجاد نماید موضوع دیگري است که مي‌تواند ما را به سمت تخصصي کردن دسته‌بندها رهنمون سازد.

مراجع

1 - Giorgetti, D. and Sebastiani, F. (2003). "Multiclass text categorization for automated survey coding." *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, Melbourne, US, PP. 798-802.

2 - Breiman, L. (1996). "Bagging predictors." *Machine Learning,* Vol. 24, PP. 123-140.

3 - Debole, F. and Sebastiani, F. (2004). "An analysis of the relative difficulty of Reuters-21578 subsets." *In* Proceedings of LREC-04, 4th Int. Conf. on Language Resources and Evaluation*,* Lisbon, PT, PP. 971-974.

4 - Lewis, D. D., Yang, Y., Rose, T. and Li, F. (2004). "RCV1: A New Benchmark Collection for Text Categorization Research." *Journal of Machine Learning Research,* Vol. 5, PP. 361-397.

5 - Li, Y. H. and Jain, A. K. (1998). "Classification of text documents." *The Computer Journal*, Vol. 41, No. 8, PP.537-546.

6 - Joachims, T., Freitag, T. and Mitchell, D. (1997). "WebWatcher: A tour guide for the world wide web." *In proceedings of 15th Int. joint Conf. on artificial intelligence.*

7 - Aas, K. and Eikvil, L. (1999). "Text categorization: A survey." *Int. Conf. on machine learning*, PP.128-156.

8 - Sebastiani, F. (2002). "Machine learning in automated text categorization." *Journal of ACM Computing Surveys,* Vol. 34, No. 1, PP. 1-47.

9 - Yang, Y. and Chute, C. G. (1994). "An example-based mapping method for text categorization and retrieval." *ACM Transactions on Information System*, Vol. 12, No. 3, PP.252-277.

10 - Yang, Y. and Liu, X. (1999). "A re-examination of text categorization methods." *In proceedings of SIGIR-99, 22nd ACM Int. Coference on Research and Development in Information Retrieval* (Berkeley, US), PP. 42-49.

11 - Dagan, I., Karov, Y. and Roth, D. (1997). "Mistakedriven learning in text categorization." *In Proceedings of 2nd Conf. on Empirical Methods in Natural Language Processing* (Providence, RI), PP.55-63.

12 - Li, H. and Yamanishi, K. (1999). "Text classification using ESC-based stochastic decision lists." *In Proceedings of 8th ACM Int. Conf. on Information and Knowledge Management,* PP.122-130.

13 - Cohen, W. and Singer, Y. (1999). "Context sensitive learning methods for text categorization." *ACM Trans. Inform. Syst.*, Vol. 17, No. 2, PP.141-173.

14 - Apte, C., Damerau, F. J. and Weiss, S. M. (1999), "Automated learning of decision rules for text categorization." *ACM Trans. On inform. Syst.*, Vol. 12, No. 3, PP.233- 251.

15 - Lewis, D. D. (1998). "Naive(Bayes) at Forty: The independence assumption in information retrival." *ECML*.

16 - Koller, D. and Sahami, M. (1997). "Hierarchically classifying documents using very few words." *Proceedings of the 14th Int. Conf. on Machine Learning (ML),* Nashville, Tennessee, PP. 170-178, July.

17 - Hull, D. A. (1994). "Improving text retrieval for routing problem using latent semantic indexing." *In Proceedings of 17th ACM Int. Conf. on Research and Development in Information Retrieval*, PP. 282-289.

18 - Jalili, S. and Aghaee A. (2003). *Text Classification by class centroid method*. 11th conference on electrical engineering (ICEE 2003).

19 - Joachims, T. (1998). "Text categorization with support vector machines: Learning with many relevant features." *ECML,* PP.137-142.

20 - Klinkenberg, R. and Joachims, T. (2000). "Detecting concept drift With support vector machines." *In Proceedings of 17th Int. Conf. on Machine Learning,* PP.487-494.

21 - Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M. (1998). "Inductive learning algorithms and representations for text categorization." *In Proceedings of CIKM-98, 7th ACM Int. Conf. on Information and Knowledge Management (Bethesda, MD),* PP.148-155.

22 - Larkey, L. S. and Croft, W. B. (1996). "Combining classifiers in text categorization." *In Proceedings of SIGIR-96, 19th ACM Int. Conf. on Research and Development in Information Retrieval* (Switzerland), PP.289-297.

23 - Bi, Y. and Be, D. (2004). *Classification Decision Combination for Text Categorization: An Experimental Study*, DEXA 2004, LNCS 3180, PP. 222 - 231.

24 - Bell, D. A. and Guan, J. W. and Bi, Y. X. (2005). "An evidential approach to classification combination for text categorisation." Studies in Fuzziness and Soft Computing, Vol. 185, PP. 13-22.

25 - Bi, Y. and Be, D. (2004). *Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization*, LNAI 3131, PP. 127-138.

26 - Zheng, Z. and Zhou, S. and Zhou, A. (2004). *Sequential Classifiers Combination for Text Categorization: An Experimental Study*, WAIM 2004, LNCS 3129, PP. 509–518.

27 - Nardiello, P., Sebastiani, F. and Sperduti, A. (2003). "Discretizing continuous attributes in AdaBoost for text categorization." *Proceedings of ECIR-03, 25th European Conf. on Information Retrieval*, PP. 320—334.

28 - Steinbach, M., Karypis, G. and Kumar, V. (2000). "A comparison of document clustering techniques." *Proc. TextMining Workshop, KDD*.

29 - Tsikrika, T. and Lalmas, M. (2001). "Merging techniques for performing data fusion on the web." Proceedings of the 10th Int. Conf. on information and Knowledge management, Atlanta, Georgia, USA, October.

30 - Bordogna, G. (2002). "Soft fusion of information accesses." *Proceedings of IEEE Int. Conf. on Fuzzy Systems,* Vol 2, PP.1466-1471.

31 - Beil, F. Ester, M. and Xu, X. (2002). "Frequent term-based text clustering." *Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD'02).*

32 - Zamir, O., Etzioni, O., Madani, O., Karp, R. M. (1997). "Fast and intuitive clustering of web documents." KDD ’97, PP. 287-290.

33 - Aggarwal, C. C., Gates, S. C. and Yu, P. S. (1999). "On the merits of building categorization systems by supervised clustering." *Proc.of the 5th ACM Int. onf. on Knowledge Discovery and Data Mining,* PP. 352 – 356.

34 - Cutting, D. R., Karger, D. R., Pedersen, J. O. and Tukey, J. W. (1992). "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections." SIGIR ‘92, PP. 318 – 329.

35 - Debole, F. and Sebastiani, F. (2003). "Supervised term weighting for automated text categorization." *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, Melbourne, US, PP. 784-788.

36 - Galavotti, L., Sebastiani, F. and Simi, M. (2000). "Experiments on the use of feature selection and negative evidence in automated text categorization." *In Proceedings of 4th European Conf. on Research and Advanced Technology for Digital Libraries,* PP. 59-68.

37 - Yang, Y. and Pedersen, J. O. (1997). "A comparative study on feature selection in text categorization." *In Proce. of 14th Int. Conf. on Machine Learning*, PP.412-420.

38 - Jalili, S. and Bitarafan, M. (2006). "Increasing performace of text classification based on improving feature selection." *Journal of faculty of engineering (Tehran university), Special Issue:Electrical Engineering*, Vol 40, No.3.

39 - Sadri, A.A. (2005). *Text Classification Based on Committee and Feature Selection*. Thesis of Master of Science (M.S.) in Computer Engineering (Software)" , Tarbiat Modares University, Faculty of Engineering.

40 - Sadri, A.A. and Jalili, S. (2006). *Quality Enhancement in Text Classification with classification committees.* 14th conference on electrical engineering(ICEE 2006) , AmirKabir University.

41 - Moens, M. and Dumortier, J. (2000). "Text categorization: the assignment of subject descriptors to magazine articles." *Information Processing and Management,* Vol. 36, Elsevier Science, PP.841-861.

42 - El-Hamdouchi, A. and Willet, P. (1989). "Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval." *The Computer Journal,* Vol. 32, No.3.

43 - Cauwenberghs, G. and Poggio, T. (2000). "Incremental and Decremental support vector machine learning." *Aadvances In Neural Information Processing*, Vol. 13.

واژه هاي انگليسي به ترتيب استفاده در متن

1 - Information Retrieval
2 - k Nearest Neighbor
3 - Support Vector Machines
4 - Pair Wise Coupling
5 - Majority voting
6 - Weighted linear Combination
7 - Dynamic classifier selection
8 - Adaptive classifier combination
9 - Local accuracy
10 - Decision Fusion
11 - Text Mining
12 - Hierarchic Agglomerative Clustering
13 - Stop words
14 - Document Frequency
15 - Information Gain
16 - Mutual Information
17 - Correlation Coefficient
18 - Precision
19 - Recall
20 - Breakeven
21 - Micro Average
22 - Macro Average
23 - Monotone
24 - Discriminative
25 - Supprot Vectors
26 - Incremental