افزایش دقت کلاسه بندی در داده کاوی با استفاده از ترکیب کلاسه بندها

حمیدرضا طهماسبی[۱] ، حسن احمدی[۲]

چکیده

اگر چه بعضی از کلاسه بندها در برخی موارد نسبت به بقیه نتایج بهتری تولید می کنند ولی هیچ یک از آنها بر سایرین برتری نداشته و نمی تواند تمام داده ها را بدون هیچ خطایی کلاسه بندی کند. هر کلاسه بند قوت ها و ضعف های خاص خود را دارد. ترکیب مناسب کلاسه بندها، می تواند نتایج کلاسه بندی بهتری نسبت به هر کلاسه بند و حتی بهترین آنها تولید کند.

در این مقاله، روشی برای ترکیب کلاسه بندها پیشنهاد می شود که نتایج کلاسه بندهای نزدیکترین k- همسایه، درخت تصمیم و بیز ساده را با استفاده از تئوری ترکیب باورها ترکیب می کند. این روش بهمراه سایر روشهای ترکیبی معروف بر روی دو مجموعه داده با کاربردهای مختلف مورد ارزیابی قرار گرفته و نشان داده می شود که علاوه بر بیشتر بودن دقت روش پیشنهادی نسبت به کلاسه بندهای بکار رفته در ترکیب، نسبت به سایر روش های ترکیبی نیز از دقت بیشتری برخوردار است. بعلاوه با توجه به این آزمایشات، تاثیر تعداد و نوع کلاسه بندها و همچنین ترتیب ترکیب آنها نیز مورد بررسی و تحلیل قرار می گیرند.

کلمات کلیدی

داده کاوی، کلاسه بندی ترکیبی، ترکیب باورها ، میزان دقت کلاسه بندی

Increasing Classification Accuracy in Data Mining Using Multiple Classifiers Combination

HamidReza Tahmasebi[1], Hasan Ahmadi[2]

Abstract

Although some classifiers may succeed better than others in some cases, none is perfect and none can classify any data without mistakes. Each classifier has its own strengths and weaknesses. Combining classifiers, in an efficient way, can achieve better classification results than any single classifier (even the best one).

In this paper, we propose a classifiers combination approach using theory of beliefs combination that combines the results of k Nearest Neighbors, Decision tree and Naive Bayes classifiers. This approach is evaluated through several experiments that are performed using two dataset with different applications. Classification results producted by the proposed shows improvement compared to the classification results produced by the individual classifiers that are used in the combination and also another common classifiers combination methods. Also important conclusions about the effects of the type, number and the order of the combined classifiers in the classifiers combination process are extracted from these experiments.

Keywords

Data Mining, Multiple Classification, Beliefs Combination, Classification Accuracy

---

۱. عضو هیئت علمی گروه کامپیوتر دانشگاه آزاد اسلامی واحد کاشمر htahmasebi2002@yahoo.com

۲. استادیار گروه کامپیوتر دانشکده فنی و مهندسی دانشگاه آزاد اسلامی واحد مشهد hahmadi@mshdiau.ac.ir

۱- مقدمه

در فرآیند کلاسه بندی، بر اساس داده های توزیع شده مدل اولیه ای ایجاد می گردد و سپس این مدل برای طبقه بندی داده های جدید و تخصیص آنها به کلاس های مجزا مورد استفاده قرار می گیرد. الگوریتم های کلاسه بندی در معماری، روش یادگیری یا نحوه آموزش و نمایش ویژگی ها با یکدیگر متفاوتند [۱]. اگر چه بعضی از کلاسه بندها نسبت به بقیه بهتر عمل می کنند ولی هیچ کلاسه بندی نمی تواند نسبت به سایرین برتری داشته و یا داده ها را بدون هیچگونه خطایی طبقه بندی کند [۲]. مطالعات انجام شده نشان می دهند که روش های ترکیب کلاسه بندها به یک ابزار مؤثر بمنظور افزایش کارایی و میزان دقت کلاسه بندی تبدیل شده اند [۷-۱].

یک سیستم کلاسه بندی ترکیبی، از ترکیب دو یا چند کلاسه بند معمولی ساخته می شود. این روش اطلاعات خروجی کلاسه بندهای مختلف را ترکیب کرده و تصمیم نهایی کلاسه بندی بر اساس اطلاعات ترکیب شده، گرفته می شود. هر سیستم کلاسه بندی ترکیبی از دو مؤلفه اساسی تشکیل شده است [۱و۸]. اولین مؤلفه بستگی به کاربردی دارد که کلاسه بندی داده ها به آن منظور باید انجام بگیرد. مواردی از قبیل تعداد و نوع کلاسه بندهایی که باید انتخاب شوند، نوع ویژگی هایی که توسط هر کلاسه بند باید استفاده شود و مسائل مربوط به ساختار هر کلاسه بند جزء این مؤلفه محسوب می شوند. مؤلفه دوم که اغلب در همه کاربردها یکسان است، مسائل مرتبط با این پرسش است که چگونه نتایج حاصل از کلاسه بندهای مختلف ترکیب شوند تا بهترین نتیجه بدست آید [۶].

در این مقاله، از سه الگوریتم کلاسه بندی k- نزدیکترین همسایه[۱] (KNN)، بیز ساده[۲] (NB) و درخت تصمیم[۳] (DT) استفاده شده و یک سیستم کلاسه بندی ترکیبی پیشنهاد می گردد که خروجی های سه کلاسه بند مذکور را با استفاده از تئوری ترکیب شواهد دمستر - شافر[۴] (DS) با هم ترکیب کرده و با آزمایش آن روی دو مجموعه داده ی سرطان سینه Wisconsin و Iris نشان داده می شود که نتایج بهتری نسبت به سه کلاسه بند به کار رفته در ترکیب و همچنین یک روش ترکیبی دیگر مبتنی بر DS و روش های کلاسه بندی ترکیبی مشهورِ رأی اکثریت[۵] (MV) [۴]، ترکیب خطی وزن دار[۶] (WLC) [۴]، ترکیب ماکزیمم[۷] (MX) [۴]، ترکیب میانگین[۸] (AV) [۴] و ترکیب میانه[۹] (MD) [۴] تولید کرده و از دقت بهتری برخوردار است. تئوری شواهد دمستر - شافر تعمیم یافته احتمال بیزین می باشد که خصوصیات بازیابی صریح عدم قطعیت و قاعده ترکیب شواهد آن [۹] موجب استفاده از آن در این روش شده است.

با توجه به خروجی نامناسب الگوریتم های k- نزدیکترین همسایه و درخت تصمیم برای استفاده در تئوری دمستر - شافر ، ابتدا خروجی این دو الگوریتم به شکل قابل پذیرش توسط تابع ترکیب تبدیل شده و سپس عمل ترکیب انجام می گیرد.

در بخش دوم این مقاله، تئوری ترکیب شواهد دمستر - شافر به اختصار معرفی می گردد. بخش سوم به معرفی روش های کلاسه بندی ترکیبی موجود مبتنی بر تئوری دمستر - شافر می پردازد. روش پیشنهادی در بخش چهارم معرفی و نتایج حاصل از به کارگیری آن روی دو مجموعه داده ی استاندارد و مقایسه دقت آن با سایر روش های ترکیبی متداول در بخش پنجم مطرح می گردد. نتیجه گیری و کارهای آتی نیز بخش ششم را تشکیل می دهند.

۲- تئوری ترکیب شواهد دمستر – شافر

فرض کنید Θ مجموعه ی همه خروجی های ممکن در یک آزمایش باشد. به مجموعه Θ چارچوب مشاهدات گفته می شود [۹]. در تئوری شواهد دمستر - شافر ، باور[۱۰] مقداری است که برای بیان قطعیت یک گزاره به کار می رود. یک باور معمولاً براساس یک تابع $m : 2^{\Theta} \to [0,1]$ به نام تخصیص احتمال پایه[۱۱] (BPA) محاسبه می شود. BPA دو شرط دارد:

$$m(\phi) = 0 \qquad (۱)$$

$$\sum_{A \subseteq \Theta} m(A) = 1 \qquad (۲)$$

m(A) بیانگر میزان باور جزئی است که دقیقاً به A نسبت داده می شود نه باور کل به A [۳و۹]. برای اندازه گیری باور کامل اختصاص یافته به A باید همه کمیت های *m(B)* که B زیرمجموعه A است با هم جمع شوند:

$$Bel(A) = \sum_{\forall B \subseteq A} m(B) \qquad (۳)$$

از آنجایی که تابع باور نمی تواند بیانگر نقیض A (A') باشد شافر میزان تردید در A را بصورت باور کامل به A' تعریف کرده است. تابع محتمل بودن[۱۲]، میزان باور کاملی که می توان به A اختصاص داد را محاسبه می نماید و برابر است با :

$$Pl(A) = 1 - Bel(A') = \sum_{B \subseteq \Theta} m(B) - \sum_{B \subseteq A'} m(B) = \sum_{B \cap A \neq \phi} m(B) \qquad (۴)$$

در مقابل *Bel(A)* که خلاصه همه دلایل برای باور به A می باشد، *Pl(A)* بیان می کند که اگر همه آنچه کـه فعـلاً مـا از آن آگـاه نیـستیم A را حمایت کنند، چقدر باید به A باور داشت. بنابراین میزان باور صحیح به A در فاصله [*Bel(A),Pl(A)*] قرار دارد.

قاعده ترکیب شواهد دمستر، دو بدنه شواهد مستقل تعریف شده در یک چارچوب مشاهدات را با هم ترکیب کرده و به یک بدنـه شـواهد تبـدیل می کند. در حقیقت این قاعده، روشی برای ترکیب شواهد از منابع مختلف می باشد. به طوری که شواهد را روی هم گذاری کرده و یک تابع باور بـه دست می آورد که تناظر این شواهد است. فرض کنید دو بدنه شواهد به صورت شکل (۱) موجود باشند:

$$\{A_1, A_2, ..., A_l\}, \{m_1(A_1), m_1(A_2), ..., m_1(A_l)\}$$
$$\{B_1, B_2, ..., B_m\}, \{m_2(B_1), m_2(B_2), ..., m_2(B_m)\}$$

شکل (۱) - دو بدنه شواهد A و B

برای بدنه شواهد جدید حاصل از ترکیب این دو بدنه شواهد، مقدار BPA برابر است با :

$$m(C) = (m_1 \oplus m_2)(C) = \frac{\sum_{A_i \cap B_j = C} m_1(A_i) m_2(B_j)}{1 - N} \qquad (۵)$$

بنابراین قواعد ترکیب شواهد دمستر میزان تطابق بین دو بدنه شواهد تعریف شده در یک چارچوب مشاهدات را محاسبه می نماید. در این رابطه که قاعده ترکیب دمستر نام دارد، مخرج کسر یک عامل نرمال سازی است که باعث می شود m(C) نیز یک BPA باشد. N فـاکتور مغـایرت اسـت و مقدار آن برابر است با:

$$N = \sum_{A_i \cap B_j = \phi} m_1(A_i) m_2(B_j) \qquad (۶)$$

تئوری دمستر - شافر به طور کامل تر در [۶و۹] تشریح شده است.

۳- روش های ترکیبی موجود مبتنی بر تئوری دمستر – شافر

تفاوت اصلی هر یک از روش هایی که برای ایجاد یک سیستم کلاسه بندی ترکیبی با استفاده از تئوری DS پیشنهاد شده است، در نحوه بدسـت آوردن مقادیر باور کلاس ها در هر کلاسه بند می باشد. مهمترین این روش ها در ادامه به اختصار معرفی می گردند.

Xu [۱۰]، مدلی برای ترکیب با استفاده از تئوری DS پیشنهاد نمود که مقادیر بـاور بـا اسـتفاده از نـرخ هـای تـشخیص درسـت[۱۳] و تـشخیص اشتباه[۱۴] محاسبه می شوند. این روش، این واقعیت را که یک کلاسه بند کارایی یکسانی روی کلاس های مختلف ندارد را نادیده می گیرد و در نتیجه ممکن است کارایی خوبی نداشته باشد [۵].

Rogova [۱۱]، یک مدل برای ترکیب نتایج کلاسه بندی شبکه های عصبی با استفاده از تئوری DS پیشنهاد نمود که مقـادیر بـاور بـه کمـک رابطه میان بردارهای خروجی کلاسه بندها و یک سری بردارهای مرجع که از میانگین بردارهای خروجی بدست می آیند، محاسبه مـی شـوند. عیـب اصلی این روش در نحوه محاسبه بردارهای مرجع می باشد به طوری که میانگین بردارهای خروجی ممکـن اسـت بهتـرین انتخـاب نباشـد [۱۲]. بـه همین دلیل، Al-Ani [۷]، روش مشابه ای ارائه نمود که بردارهای مرجع با استفاده از مینیمم کردن میانگین مربع خطاها بین نتـایج کلاسـه بنـدی ترکیبی و کلاس های واقعی نمونه های آموزشی طی یک فرایند تکراری بدست می آیند و منجر بـه کـارایی بهتـری نـسبت بـه روش Rogova مـی گردد. ولی هزینه محاسبات در این روش زیاد است. به طوری که برای هر کلاس یک بردار مرجع در نظر می گیرد که در صورت افزایش تعداد کلاس ها و یا تعداد کلاسه بندها، مدت زمان آموزش نیز افزایش می یابد [۱۲].

روشی نیز توسط Mahajani و Aslandogan [۱۳] پیشنهاد گردید که خروجـی هـای سـه الگـوریتم کلاسـه بنـدی KNN، NB و DT را بـا استفاده از تئوری DS ترکیب می نماید تا نتایج کلاسه بندی بهتری تولید کند. این روش را به اختصار DS0 می نامیم.

خروجی های کلاسه بندهای KNN و DT بصورت برچسب کلاس[۱۵] می باشد. در این روش برای محاسبه مقادیر باور در KNN، k- نزدیکتـرین همسایه ی نمونه جدید x_s، به مجموعه هایی که نمونه های هر دسته دارای کلاس یکسانی هستند تقسیم می شوند. سپس مقـدار فاصـله D_m بـرای هر نمونه ی x_i در هر مجموعه با استفاده از رابطه زیر بدست می آید:

$$D_m = e^{-\frac{d_s}{d_{mean}}} \qquad (۷)$$

d_s، فاصله بین نمونه x_s و x_i و d_{mean} میانگین فاصله بین نمونه های آموزشی هم مجموعه با x_i می باشد. مقادیر d_s و d_{mean} به مقادیری در بازه [0,1] نرمال شده اند. مقدار باور برای هر کلاس برابر با میانگین D_m های نمونه های متعلق به مجموعه آن کلاس می باشد.

در DT ، مقادیر باور از محاسبه درجه اطمینان[۱۶] (CF) برای هر کلاس بدست می آیند:

$$CF = \frac{P(\ feature\ \ set \mid P(class)}{P(\ feature\ \ set)} \qquad (۸)$$

در واقع CF، همان احتمال پسین یک کلاس به ازای مجموعه ویژگی های داده شده می باشد.

۴- روش پیشنهادی

در اینجا روشی پیشنهاد می شود که در آن نتایج حاصل از سه الگوریتم کلاسه بندی متفاوت KNN، DT و NB با استفاده از تئوری شواهد دمستر- شافر با هم ترکیب می شوند تا به دقت بهتری دست یافت.

مراحل کار این روش بصورت زیر است:

۱- کلاسه بندی نمونه ورودی توسط سه کلاسه بند KNN، DT و NB و تبدیل خروجی کلاسه بندهای KNN، DT به فرمت مناسب

۲- استخراج مقادیر باور از نتایج خروجی کلاسه بندها

۳- ترکیب مقادیر باور با استفاده از تئوری دمستر - شافر

۴- تصمیم گیری و انتخاب برچسب کلاس مناسب برای نمونه ورودی

۴-۱ کلاسه بندی نمونه ورودی و تبدیل خروجی ها به فرمت مناسب

ابتدا داده نرمال شده ی ورودی توسط سه کلاسه بند KNN، DT و NB کلاسه بندی می گردد. خروجی دو الگوریتم نزدیکترین k- همسایه و درخت تصمیم بصورت برچسب کلاس می باشد، جهت استخراج مقادیر باور از خروجی کلاسه بندها، خروجی آنها می بایست مقادیر حقیقی در بازه [0,1] باشد. بدین منظور برای کلاسه بند KNN از الگوریتم پیشنهادی توسط Denoex [۱۴] استفاده می کنیم. این الگوریتم یک روش KNN ترکیبی مبتنی بر دمستر - شافر می باشد که از این تئوری به عنوان یک ساختار ترکیبی در داخل الگوریتم کلاسه بندی استفاده می کند.

برای کلاسه بندی نمونه X، اگر بخواهیم حتماً X به یکی از کلاس ها تعلق یابد، خروجی الگوریتم Denoex، درجه محتمل بودن برای هر کلاس C_i می باشد:

$$Pl(\{C_i\}) = m(\{C_i\}) + m(C) \qquad i = 1,...,M \qquad (۹)$$

هر یک از مقادیر $Pl(C_i)$ را نرخ اطمینان کلاس C_i نامیده و با نشان $Cr(i)$ می دهیم. بنابراین خروجی کلاسه بند KNN بصورت زیر می باشد:

$$C_{KNN} = \{Pl(C_1),...,Pl(C_M)\} = \{Cr(1),...,Cr(M)\} \qquad (۱۰)$$

برای استخراج خروجی های مناسب از کلاسه بند DT از ماتریس تداخل استفاده می کنیم.

ماتریس تداخل برای کلاسه بند e_t در شکل (۲) مشاهده می شود. در این ماتریس، تعداد کلاس ها برابر M می باشد. $n_{ij}^{(t)}$ بیانگر تعداد نمونه هایی است که کلاس واقعی آنها C_i است و توسط کلاسه بند e_t کلاس C_j به آنها انتساب شده است (i=1,…,M و j=1,…,M+1). $n^{(t)}_{i(M+1)}$ تعداد نمونه هایی است که متعلق به کلاس C_i می باشند ولی کلاسه بند هیچ کلاسی برای آنها تعیین نمی کند.

$$PT_t = \begin{bmatrix} n_{11}^{(t)} & \cdots & n_{1j}^{(t)} & \cdots & n_{1(M+1)}^{(t)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{i1}^{(t)} & \cdots & n_{ij}^{(t)} & \cdots & n_{i(M+1)}^{(t)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{M1}^{(t)} & \cdots & n_{Mj}^{(t)} & \cdots & n_{M(M+1)}^{(t)} \end{bmatrix}$$

شکل (۲) - ماتریس تداخل

با توجه به این ماتریس، مقادیر BPA با استفاده از رابطه زیر به دست می آیند [۱۰].

$$Bel(C_i) = P(X \in C_i \mid e_t(X) = j) = \frac{n_{ij}^t}{\sum_{i=1}^{M} n_{ij}^t} \qquad i = 1,...,M \qquad (۱۱)$$

بنابراین:

$$C_{DT} = \{Bel(C_1),...,Bel(C_M)\} = \{Cr(1),...,Cr(M)\} \qquad (۱۲)$$

در کلاسه بند NB نیز $P(C_i / X)$ را نرخ اطمینان کلاس C_i در نظر می گیریم. خروجی این کلاسه بند به صورت زیر است:

$$C_{NB} = \{P(C_1 \mid X),..., P(C_M \mid X)\} = \{Cr(1),...,Cr(M)\} \qquad (۱۳)$$

۴-۲ استخراج مقادیر باور

به منظور استخراج مقادیر باور یا BPA برای هر کلاسه بند، کلاس هایی که نرخ اطمینان نزدیک به هم دارند در یک گزاره قرار می گیرند و سپس برای هر گزاره یک BPA بدست می آید. نزدیکی نرخ اطمینان دو کلاس با استفاده از مقدار آستانه $\lambda_t \in [0,1]$ تعیین می گردد. بدین منظور الگوریتم زیر برای هر کلاسه بند بطور جداگانه اجرا می گردد:

۱- مقادیر $Cr(i)$ را بصورت نزولی در یک بردار V بصورت $V = \{V(1), V(2),..., V(M)\}$ مرتب کن بطوری که $V(1)$ متناظر با کلاسی است که بزرگترین نرخ اطمینان را دارد.

۲- دو متغیر a و j را مقدار اولیه یک بده. $(a \leftarrow 1, j \leftarrow 1)$

۳- یک گزاره جدید تهی بنام A_j بساز. $(A_j = \{\phi\})$

۴- $V(a)$ را در A_j قرار بده.

۵- مقدار a را یک واحد افزایش بده. $(a \leftarrow a+1)$

۶- اگر $a = M+1$ می باشد به مرحله ۸ برو.

۷- اگر $Cr(V(a-1)) - Cr(V(a)) < \lambda_t$ است به مرحله ۴ برو. در غیر اینصورت مقدار j را یک واحد افزایش بده و به مرحله ۳ برو.

۸- BPA برای هر گزاره ی A_j برابر است با میانگین نرمال شده ی نرخ های اطمینان کلاس های متعلق به آن گزاره. یعنی:

$$m(A_j) = \frac{\mu_{A_j}}{S} \qquad (۱۴)$$

که مقدار μ_{A_j} در این رابطه برابر است با:

$$\mu_{A_j} = \frac{\sum_{i \in A_j} Cr(i)}{|A_j|} \qquad (۱۵)$$

و S پارامتر نرمال سازی است که مقدار آن برابر است با :

$$S = \sum_{j \in A} \mu_{A_j} \qquad (۱۶)$$

۹- پایان.

این الگوریتم برای هر کلاسه بند تعدادی گزاره تولید می کند. تعداد این گزاره ها بین یک تا M گزاره می باشد. در حالتی که $\lambda_t = 0$ انتخاب شود، M گزاره تولید می شود که هر گزاره شامل یک کلاس خواهد بود.

انتخاب مقدار مناسب برای λ_t با تست روش پیشنهادی به ازای مقادیر مختلف λ_t روی مجموعه داده های آموزشی مختلف بدست می آید.

۴-۳ ترکیب مقادیر باور

مقادیر باور یا BPA استخراج شده از سه کلاسه بند با استفاده از تئوری دمستر - شافر با هم ترکیب می شوند. این قاعده ترکیب، چند گزاره جدید $A'_{j'}$ تولید می کند که هر یک دارای یک مقدار BPA می باشد.

۴-۴ انتخاب برچسب کلاس مناسب

در انتخاب کلاس مناسب برای نمونه ورودی X به صورت زیر عمل می شود:

۱- اگر $N \geq \lambda_{rej}$ باشد، یا حداقل دو کلاس C_I و C_J وجود داشته باشند به طوری که :

$$P(I) = P(J) = \max\{P(1),..., P(M)\} \qquad (۱۷)$$

آنگاه نمونه X به هیچ کلاسی تعلق نمی گیرد.

۲- اگر $N < \lambda_{rej}$ باشد، نمونه ورودی X به کلاس C_J تعلق می گیرد به طوری که :

$$P(I) = \max\{P(1),..., P(M)\} \qquad (۱۸)$$

$P(l)$ ، ($l \in A'_j$) از رابطه زیر بدست می آید:

$$P(l) = \frac{m(A'_{j'})}{|A'_{j'}|} \qquad (۱۹)$$

پارامتر N، عامل مغایرت در رابطه (۵) است و $\lambda_{rej} \in [0,1]$ مقدار آستانه ای است که برای رد کردن کلاس یک نمونه تعیین شده است. اگر $\lambda_{rej} = 1$ انتخاب شود، مطمئناً یکی از کلاس ها بعنوان کلاس نمونه ورودی تعیین می گردد. هر چه مقدار λ_{rej} کاهش یابد، نرخ این که نمونه ورودی رد شود، افزایش می یابد. در صورتی که نمونه ورودی X به کلاس C_J تعلق یابد، خروجی نهایی، برچسب J و در غیر اینصورت برچسب M+1 می باشد.

۵- نتایج تجربی

در این بخش روش پیشنهادی به همراه کلاسه بندهای شرکت کننده در ترکیب و روش های کلاسه بندی ترکیبی مشهور رأی اکثریت (MV)، ترکیب ماکزیمم (MX)، ترکیب میانگین (AV)، ترکیب میانه (MD)، ترکیب خطی وزن دار (WLC) و روش ترکیبی مبتنی بر تئوری دمستر - شافر DS0 روی مجموعه داده های Iris و سرطان سینه Wisconsin آزمایش و نتایج حاصل مورد مقایسه و بررسی قرار می گیرند. روش پیشنهادی را به اختصار، DS می نامیم.

۵-۱ مجموعه داده ها

مجموعه داده سرطان سینه Wisconsin [۱۵] شامل ۶۹۹ نمونه داده از دو کلاس خوش خیم و بدخیم می باشد. ۴۵۸ نمونه متعلق به کلاس خوش خیم و ۲۴۱ نمونه متعلق به کلاس بدخیم می باشد. این مجموعه داده، بجز صفت معرف برچسب کلاس، ۱۰ صفت دارد.

مجموعه داده ی Iris [۱۵] با تعداد ۱۵۰ نمونه، از سه کلاس تشکیل شده است که هر کلاس حاوی ۵۰ نمونه با ۴ صفت کمی می باشد. داده ها مربوط به کلاسه بندی گل هایی از دسته Iris می باشد. صفاتی از گل ها که موجود است به ترتیب عبارتند از: طول کاسبرگ، عرض کاسبرگ، طول گلبرگ و عرض گلبرگ. کلاس ها شامل سه تیره گل با نام های Iris-setosa، Iris-versicolor و Iris-virginica می باشد.

در بررسی کارایی هر یک از روش های کلاسه بندی از تکنیک ارزیابی ۱۰ تکه برابر[۱۷] استفاده می کنیم.

۵-۲ نتایج

۵-۲-۱ آزمایش سه کلاسه بند معمولی

میانگین دقت حاصل از ده بار اجرای الگوریتم KNN معرفی شده به ازای تعداد همسایه های متفاوت روی مجموعه داده های تست در جدول (۱) درج شده است. برای مجموعه داده سرطان سینه Wisconsin به ازای k=9 و برای مجموعه داده Iris به ازای k=5 و k=7، میانگین دقت بهتری نسبت به سایر مقادیر k حاصل شده است. بنابراین مقدار k برای داده های سرطان برابر ۹ و برای داده های Iris برابر ۵ انتخاب گردید. برای آموزش و تست الگوریتم های NB و DT از نسخه 3.4 نرم افزار Weka [۱۶] استفاده شده است که پارامترهای استفاده شده برای هر الگوریتم همان مقادیر پیش فرض تعیین شده در این نرم افزار می باشد.

جدول (۱) - میانگین دقت روش KNN به ازای مقادیر مختلف k

مقدار k	1	3	5	7	9
میانگین دقت Wisconsin (%)	0	86.71	89.57	89.43	**93.14**
میانگین دقت Iris (%)	83.33	88	**89.33**	**89.33**	88.67

میانگین دقت ده بار آزمایش هر یک از این سه الگوریتم بر روی دو مجموعه داده ی مذکور در جداول (۲) و (۳) نشان داده شده است. همان طور که مشاهده می شود میانگین دقت روش KNN تا حدودی نسبت به دو روش دیگر بیشتر بوده و نتایج دقیق تری تولید کرده است.

جدول (۲) - میانگین دقت اجرای ۱۰ بار سه الگوریتم کلاسه بندی روی مجموعه داده سرطان سینه Wisconsin

الگوریتم	NB	DT	KNN
میانگین دقت (%)	93.86	89.71	**95.71**

جدول (۳) - میانگین دقت اجرای ۱۰ بار سه الگوریتم کلاسه بندی روی مجموعه داده Iris

الگوریتم	NB	DT	KNN
میانگین دقت (%)	88.67	88.67	**89.33**

۵-۲-۲ آزمایش روش پیشنهادی و روش های ترکیبی متداول

در روش پیشنهادی برای بدست آوردن مقدار مناسب برای λ_t و λ_{rej}، در ابتدا برای λ_t مقادیر 0، 0.1، 0.2، 0.3، 0.4، 0.5، 0.6، 0.7، 0.8، 0.9 و 1 انتخاب گردید. سپس به ازای هر یک از این مقادیر، مقادیر 0.1، 0.2، 0.3، 0.4، 0.5، 0.6، 0.7، 0.8، 0.9 و 1 برای λ_{rej} انتخاب شده و به ازای آنها روش پیشنهادی روی مجموعه داده ها مورد آزمایش قرار گرفت. برای مجموعه داده سرطان سینه Wisconsin در حالتی که $\lambda_t = 0.4$ و $\lambda_{rej} = 1$ انتخاب گردیده بود، بیشترین دقت حاصل گردید و لذا این مقادیر برای دو پارامتر λ_t و λ_{rej} برگزیده شد. همچنین برای مجموعه داده Iris در حالتی که $\lambda_t = 0.7$ و $\lambda_{rej} = 0.9$ انتخاب گردیده بود، بیشترین دقت حاصل شد و در نتیجه این مقادیر برای دو پارامتر مذکور در مجموعه داده Iris برگزیده شد.

میانگین دقت ده بار آزمایش روش پیشنهادی و روش های ترکیبی برای دو مجموعه داده مذکور در جدول های (۴) و (۵) مشاهده می شود.

برای مجموعه داده سرطان سینه با توجه به جداول (۲) و (۴) مشاهده می شود میانگین دقت کلیه روش های ترکیبی آزمایش شده نسبت به سه الگوریتم کلاسه بندی KNN، NB و DT بیشتر می باشد. همچنین جدول (۴) نشان می دهد که روش ترکیبی پیشنهادی نسبت به سایر روش های ترکیبی آزمایش شده از دقت بهتری برخوردار است.

با دقت در جداول (۳) و (۵) برای مجموعه داده Iris نیز مشاهده می شود که برای این مجموعه داده نیز میانگین دقت روش های ترکیبی آزمایش شده نسبت به سه کلاسه بند شرکت کننده در ترکیب بیشتر می باشد. بعلاوه میانگین دقت روش ترکیبی پیشنهادی از سایر روش های ترکیبی آزمایش شده بیشتر می باشد.

در کل مشاهده می شود که روش پیشنهادی در مقایسه با سه الگوریتم کلاسه بندی بکار رفته در ترکیب و روش های کلاسه بندی ترکیبی مذکور از دقت بهتری برخوردار می باشد.

جدول (۴) - میانگین دقت کلاسه بندهای ترکیبی آزمایش شده و روش پیشنهادی برای مجموعه داده سرطان سینه Wisconsin

الگوریتم	AV	MX	MV	MD	WLC	DS0	DS
میانگین دقت(%)	95.86	95.97	95.86	95.86	96	96.71	97.57

جدول (۵) - میانگین دقت کلاسه بندهای ترکیبی آزمایش شده و روش پیشنهادی برای مجموعه داده Iris

الگوریتم	AV	MX	MV	MD	WLC	DS0	DS
میانگین دقت(%)	90	90.67	90	90	90.67	90.67	91.33

۶- نتیجه گیری و پیشنهاد

در این مقاله روشی برای کلاسه بندی پیشنهاد گردید که به کمک تئوری ترکیب شواهد دمستر - شافر، نتایج حاصل از سه الگوریتم کلاسه بندی k- نزدیکترین همسایه، بیز و درخت تصمیم را با هم ترکیب می کند تا نتایج کلاسه بندی دقیق تری تولید گردد. روش پیشنهادی به همراه سه الگوریتم کلاسه بندی مذکور و همچنین روش های کلاسه بندی ترکیبی متداول، برای تشخیص سرطان سینه روی مجموعه داده سرطان سینه Wisconsin و همچنین کلاسه بندی گل هایی از دسته Iris مورد تست و ارزیابی قرار گرفت و مشاهده گردید که در مقایسه با سه الگوریتم کلاسه بندی معمولی بکار رفته و همچنین روش های ترکیبی متداول، از دقت بهتری بر خوردار می باشد.

هر چند که استفاده از سیستم های کلاسه بندی ترکیبی، بویژه روش ترکیب شواهد دمستر - شافر، موجب افزایش زمان اجرا و پیچیدگی زمانی می شود، ولی در کاربردهایی مثل کاربردهای پزشکی که دقت برای ما از اهمیت و اولویت بیشتری برخوردار است، سیستم های کلاسه بندی ترکیبی بسیار سودمند خواهند بود.

در کارهای آتی، بکارگیری روش پیشنهادی بر روی سایر مجموعه داده ها با ویژگی ها، ابعاد و کلاس های مختلف و بررسی نتایج حاصله مورد توجه خواهد بود. علاوه بر این بررسی چگونگی انتخاب تعداد و نوع کلاسه بندهای شرکت کننده در ترکیب با توجه به کاربرد مورد نظر نیز مد نظر قرار دارد.

۷- مراجع

[۱] احمدی، حسن، طهماسبی، حمیدرضا، **"بررسی تکنیک های کلاسه بندی ترکیبی در داده کاوی"**، دومین کنفرانس داده کاوی ایران، دانشگاه صنعتی امیر کبیر، آبان ۱۳۸۷.

[۲] احمدی، حسن، طهماسبی، حمیدرضا، **"کلاسه بندی مبتنی بر شواهد در داده کاوی پزشکی"**، دومین کنفرانس داده کاوی ایران، دانشگاه صنعتی امیر کبیر، آبان ۱۳۸۷.

[3] Bi, Yaxin, Bell, David, Wang, Hui, Guo, Gongde, Guan. Jiwen, *"Combining Multiple Classifiers Using Dempster's rule for text Caregorization"*, Applied Artificial Intelligence, 21:3, 211- 239, 2007.

[4] Kuncheva, L,I, "*Combining Pattern Classifiers, Methods and Algorithms"*,New York, NY: Wiley Interscience, 2005.

[5] Tulyakov, S, Jaeger, S, Govindaraju, V, Doermann, D, "*Review of Classifier combination Methods"*, Studies in Computational Inteligence(SCI) 90,361-386 , 2008.

[6] Polikar, R, *"Ensemble Based Systems in Decision Making*", IEEE circuits & systems magazine, Third Quarter, 21-45, 2006.

[7] Al-Ani, A, Deriche, M, *"A New Technique for Combining Multiple Classifiers using The Dempster-Shafer Theory of Evidence"*, Journal of Artificial Intelligence Research 17:333-361, 2002.

[8] Ruta, D, Gabrys, B, "*An Overview of classifier fusion Methods"*, University of Paisley, Computing & Information Systems, 1-10, 2000.

[9] Sentz, K, Ferson, S, *"Combination of Evidence in Dempster-Shafer Theory*", Binghamton University, 3-94, 2002.

[10] Xu, L, Kryzak, A, Suen, C, "*Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition"*, IEEE Transactions on Sytems, Man, and Cybernetics, 22(3), 418-435, 1992.

[11] Rogova, G, *"Combining the results of several neural network classifiers*", Neural Networks 7(5):777-781, 1994.

[12] Bi, Y, Guan, J, Bell, D, "*The Combination of Multiple Classifiers Using an Evidential Reasoning Approach*", Artificial Intelligence, 2008.

[13] Aslandogan ,Y, Alp, Mahajani, Gauri A, *"Evidence Combination in Medical Data Mining*", *itcc*, p. 465, International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 2, 2004.

[14] Denoeux, T, *"A k-nearest neighbor classification rule based on Dempster-Shafer Theory*", IEEE Transactions on Systems, Man and Cybernetics, 25 (5):804-813, 1995.

[15] www.ics.uci.edu/~mlearn/MLRepository.html.

[16] www.cs.waikato.ac.nz/me/weka.

زیرنویس ها

[1] K-Nearest Neighbor
[2] Naive Baysian
[3] Decision Tree
[4] Dempster-Shafer
[5] Majority Vote
[6] Weighted Linear Combination
[7] Max Rule
[8] Average Rule
[9] Median Rule
[10] Belief
[11] Basic Probability Assignment
[12] Plausibility Function
[13] Recognation Rate
[14] Substitution Rate
[15] Abstract Level
[16] Confidence
[17] 10- Fold Cross Validation