روش یادگیری دسته بند فازی Emmff و بکارگیری آن در داده های پزشکی

سعید جلیلی، فاطمه فرجی دانشگر
f_daneshgar@modares.ac.ir, sjalili@modares.ac.ir
دانشگاه تربیت مدرس، دانشکده فنی، بخش برق، آزمایشگاه یادگیری نمادین ماشین

چکیده مبسوط

سیستم های خبره مبتنی بر قانون اغلب برای پشتیبانی تصمیم گیری در دامنه های مختلفی مانند تشخیص خطا[1]، زیست شناسی[2]، و پزشکی به کار گرفته می شوند. در بعضی از دامنه ها مانند پزشکی، ترجیح داده می شود از روش های دسته بندی به صورت جعبه سیاه(مانند شبکه عصبی) استفاده نشود تا کاربر قادر به درک دانش دسته بند[۱]. دسته بندهای مبتنی بر قانون فازی برای این منظور بسیار مناسب هستند. زیرا آنها حاوی قوانین تفسیرپذیر زبانی ساده ای هستند و بعضی محدودیت های دسته بندهای قطعی[3] را ندارند. دسته بند ها اغلب باید طی یک فرآیند یادگیری از روی داده ها ایجاد شوند، زیرا دانش متخصص در بسیاری از دامنه ها محدود است و نمی توان تمام پارامترهای سیستم را بوسیله دانش متخصص مشخص کرد.

یکی از کاربردهای داده کاوی در پزشکی، مسئله تشخیص سرطان است. تا کنون از روش های مختلفی برای این مسئله استفاده شده است که به طور کلی می توان به SVM [۲] ، دسته بندی فازی[۱] و [۳-۵] ، شبکه عصبی[۶-۸] و الگوریتم ژنتیک[۹] اشاره کرد.

MMFF[4] یکی از روش های دسته بندی فازی است که طی سه مرحله: ۱) انتخاب خصیصه ۲) تولید توابع عضویت و ۳) تولید جدول تصمیم گیری، قوانین فازی را از نمونه های آموزشی استخراج می کند. این روش ازیک سو الگوریتم بسیار ساده ای دارد که می تواند برای کاربردهای پزشکی مناسب باشد و از سوی دیگر نتایج خوبی از لحاظ دقت دسته بندی بین روش های مشابه دارد.

در روش MMFF ابتدا تعدادی از خصیصه ها به عنوان خصیصه های مرتبط انتخاب می شود. سپس برای خصیصه های مرتبط، توابع عضویت تولید می شود. در مرحله بعد با استفاده از توابع عضویت تولید شده، بردار تصمیم گیری ساخته می شود. بردار تصمیم گیری، یک بردار n بعدی است که n تعداد خصیصه های مرتبط است و اندازه هر بعد نیز متناظر با تعداد مجموعه های فازی خصیصه انتخاب شده است. شکل(۱) یک بردار تصمیم گیری دو بعدی مربوط به دو خصیصه x و y را نشان می دهد که خصیصه x دارای چهار مجموعه فازی $\{x_1,\ldots,x_4\}$ و خصیصه y نیز دارای چهار مجموعه فازی $\{y_1,\ldots,y_4\}$ می باشد. هر سلول بردار تصمیم گیری، برچسب نمونه های موجود در سلول را نشان می دهد. هر سلول بردار تصمیم گیری معادل با یک قانون فازی است.

y_1	O1		O2	O3
y_2		O2		
y_3		O2	O2,O3	O3
y_4	O1		O3	O3
	x_1	x_2	x_3	x_4

شکل۱ـ بردار تصمیم گیری دو بعدی

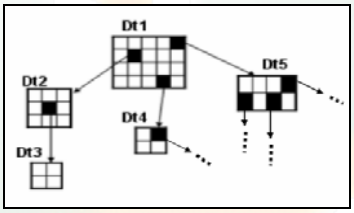
در روش MMFF، در سلول هایی که بیش از یک برچسب کلاس وجود دارد(مانند سلول ردیف سوم . ستون سوم در شکل(۱))، یکی از کلاس ها به عنوان برچسب نهایی سلول انتخاب می شود و قانون مربوط به سلول با این برچسب کلاس تولید می شود. ما در این

[1] Fault Detection
[2] Biology
[3] Crisp
[4] Merging Membership Function First

مقاله، روش MMFF را برای تشخیص سرطان با استفاده از داده های Wisconsin Breast Cancer(WBC) [۱۰] به کار بردیم. مجموعه داده WBC شامل ۶۹۹ نمونه است که هر نمونه با ۹ خصیصه توصیف شده است. پس از آزمایشات مختلف روی این مجموعه داده با روش MMFF، مشاهده شد که نتایج پایین تر از حد انتظار است. پس از بررسی روش MMFF دریافتیم که انتخاب یک برچسب کلاس در سلول هایی که بیش از یک برچسب کلاس دارند، در بسیاری موارد منجر به از دست دادن اطلاعات و دسته بندی نادقیق می شود. بنابراین ما در این مقاله روش MMFF را توسعه دادیم، به نحوی که در سلول هایی که بیش از یک برچسب کلاس وجود دارد، دسته بندی نمونه ها را با استفاده از خصیصه های باقیمانده (خصیصه هایی که تا کنون برای دسته بندی نمونه های سلول استفاده نشده اند) ادامه می دهیم. ما این روش را Emmff نامگذاری کردیم.

الگوریتم توسعه یافته Emmff ، یک الگوریتم دسته بندی فازی بازگشتی است که در فرآیند یادگیری دسته بند، با استفاده از نمونه های آموزشی، یک درخت تصمیم گیری فازی یادگیری می کند. در درخت تصمیم گیری روش Emmff، بر خلاف درخت های تصمیم گیری فازی معمول که هر گره آن یک خصیصه فازی است، هر گره، یک بردار تصمیم گیری است.

شکل(۲) ، ساختار درخت یادگیری شده در روش Emmff را نشان می دهد. در این شکل، سلول های سیاه، سلول هایی هستند که نمونه های آموزشی آنها دارای یک نوع برچسب نیستند. در نتیجه فرآیند یادگیری برای این سلول ها بایستی با استفاده از سایر خصیصه ها ادامه یابد. در حالی که نمونه های آموزشی درون سلول های سفید، همه دارای فقط یک برچسب هستند که به معنی پایان یادگیری در آن سلول ها می باشد. نقاط برگ این درخت، بردار هایی است که همه سلول های آن سفید است.



شکل۲ـ ساختار درختی یادگیری شده توسط روش Emmff

آزمایش روش Emmff با مجموعه داده WBC نشان داد که دسته بند فازی تولید شده به روش Emmff، نسبت به دسته بند روش MMFF و همچنین نسبت به سایر روش های موجود در تشخیص سرطان، قادر به دسته بندی و تشخیص دقیق تری است.

منابع

[1] D. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data", Artificial Intelligence in Medicine 16 (1999) 149–169
[2] W. Wolberg and O. Mangasarian, "Multi surface method of pattern separation for medical diagnosis, applied to breast cytology". Proceedings of National Academy of Sciences (1990) 9193–6
[3] F. Schleif , T. Villmann and B. Hammer, "Prototype based fuzzy classification in clinical proteomics", In Proceedings of International Journal of Approximate Reasoning (2007)
[4] T. Nakashima, Y. Yokota , H. Ishibuchi and A. Bargiela, "Effect of Data Weighting Methods on the Performance of Fuzzy Classification Systems", IEEE Annual Meeting of the North American Fuzzy Information Processing Society (2005) 216-221
[5] E. G. Mansoori, M. J. Zolghadri and S. D. Katebi, "A weighting function for improving fuzzy classification systems performance", Fuzzy Sets and Systems 158 (2007) 583 – 591
[6] H. Yi-Chung , "Fuzzy integral-based perceptron for two-class pattern classification problems", Information Sciences 177 (2007) 1673–1686
[7] Y. Wu, M. Giger, K. Doi, C. Vyborny, R. Schmidt and C. Metz," Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer", Radiology 187 (1993) 81-87
[8] J. Baker, P. Kornguth, J. Lo, M. Williford and C. Floyd, "Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon", Radiology 196(1995) 817-822
[9] C. A. Peña-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis", Artificial Intelligence in Medicine17 (1999) 131-155
[10] http://www.ics.uci.edu: mlearn:MLRepository