پنجمین همایش ملی بیوتکنولوژی جمهوری اسلامی ایران
3-5 آذر ماه 1386، سالن اجلاس سران
The 5th National Biotechnology Congress of Iran
24-26 Nov, 2007, Summit Meeting Conference Hall, Tehran- Iran

رویکرد فازی به خوشه بندی و انتخاب خصیصه‌ها برای کلاسه بندی داده‌های توصیفی ژن ها

الهام چیت ساز*، محمد طاهری، سراج الدین کاتبی**
chitsaz@cse.shirazu.ac.ir
mtaheri@cse.shirazu.ac.ir
katebi@shirazu.ac.ir
دانشگاه شیراز

چکیده

انتساب برچسب‌های گسسته به نمونه‌های مورد پردازش براساس مقادیر آنها در خصیصه‌ها مختلف را کلاسه‌بندی داده‌ها می‌نامند. در این تحقیق، مجموعه داده‌هایی مورد نظراست که دارای تعداد نمونه‌های کم و حجم عظیمی از خصیصه‌ها هستند. بسیاری از مجموعه داده‌های بیولوژیکی مانند داده‌های میکرو آرایه‌ها دارای چنین ویژگی می‌باشند. مهمترین بحث مورد بررسی این مقاله، رویکرد فازی به روشی در انتخاب خصیصه‌ها جهت کلاسه‌بندی داده‌ها است؛ که توسط Wai-Ho Au ارائه شد. این روش مبتنی بر خوشه‌بندی خصیصه‌ها بنا به وابستگی بین آنها است. پایداری بیشتر ،همگرایی سریعتر و بهبود نتایج حاصل از کلاسه‌بندی، از برتری‌های روش پیشنهادی به نسبت رویکرد پیشین (غیر فازی) هستند. همچنین در این مقاله، روشی نوین جهت گسسته کردن داده‌های پیوسته با استفاده از معیار Fisher ارائه شده است به انضمام این که یک روش انتساب اولیه به مراکز خوشه‌ها نیز پیشنهاد شده است. روش این مقاله بر روی مجموعه داده Leukemia اجرا شده که به نسبت رویکرد پیشین از بهبود قابل ملاحظه‌ای برخوردار است.

کلمات کلیدی: بیوانفورماتیک، تشخیص الگو، انتخاب خصیصه، منطق فازی، خوشه‌بندی

مقدمه

انتخاب خصیصه‌های مناسب، تا کنون در کاربردهای فراوان و متفاوتی از جمله کلاسه‌بندی داده‌ها، به عنوان یک پیش پردازش مطرح شده است. انتخاب خصیصه‌های مناسب با تعداد کمتر منجر به افزایش سرعت یادگیری، کاهش فضای حافظه مورد نیاز و بهبود نتیجه کلاسه‌بندی می‌شود. کاهش هزینه جمع آوری خصیصه‌ها برای داده‌های جدید، از دیگر انگیزه‌های کاهش تعداد خصیصه‌ها است. داده‌های بیولوژیکی (به عنوان مثال، میکرو آرایه‌ها) معمولا عریض و کم عمق هستند بدین معنی که، از تعداد کمی نمونه و حجم عظیمی از خصیصه‌ها برخوردارند. نمونه‌هایی کاربردهایی که در ارتباط با این گونه مجموعه داده‌ها هستند، عبارتند از:

1- کلاس‌بندی نمایه‌های توصیفی ژن‌ها در میکرو آرایه‌ها به توجه به سطح توصیفی تعداد زیاد ژن‌های موجود.
2- پیش‌بینی ساختار پروتئین‌ها مبنی بر دنباله DNA متناظر
3- تشخیص برخی از بیماریهای روانی بر اساس خصیصه‌های به دست آمده از سیگنال‌های EEG.

تعداد کم نمونه‌های داده‌ای، دانش قابل دسترس از داده‌ها را محدود می‌کند و نتیجتاً احتمال استدلال صحیح، در مورد تاثیر مقدار هر یک از خصیصه‌ها بر روی برچسب نمونه‌ها، کاهش می‌یابد. تعداد کمتر خصیصه‌ها، موجب آسان و سریعتر شدن تولید قوانین کلاسه‌بندی در یک سیستم کلاسه بندی می‌شود. از سوی دیگر، تعداد بیشتر خصیصه‌ها که غالباً مناسب نمی‌باشند، می‌تواند سیستم را در لحظه یادگیری دچار ابهام کند چنانکه همگرایی را نیز به مخاطره اندازد. همچنین، تعداد بیشتر خصیصه‌ها، زمان اجرا و حافظه بیشتری را می‌طلبد در حالیکه معمولاً تعداد کمی از خصیصه-های تاثیر گذار برای کلاسه‌بندی داده‌ها کفایت می‌کند. درواقع:

1- مقدار بسیاری از خصیصه‌ها معمولا مستقل از برچسب داده‌ها است.
2- مقادیر برخی از خصیصه‌ها از وابستگی بسیاری برخوردارند به قسمی که تنها انتخاب تعداد کمی از آنها جهت کلاسه‌بندی داده‌ها کافی است.

گزینش خصیصه‌های مناسب در زمینه‌ها و با رویکردهای مختلفی تا کنون مورد بررسی قرار گرفته است. به عنوان مثال،روش های filter ، wrapper وembedded[1]،روش های آماری مانند PCA [2]وLDA[3]، و الگوریتم‌های ژنتیک[4]. انتخاب ژن‌های موثر در کلاسه‌بندی داده‌های میکروآرایه‌ای از زمینه‌های تحقیقاتی است که بررسی‌های بسیاری تا کنون بر روی آن صورت گرفته است [5] و [6]. Wai-Ho Au [7] نیز در سال 2005، یک روش انتخاب خصیصه‌های مناسب را مبتنی بر خوشه‌بندی خصیصه‌های همبسته ارائه کرده است. خوشه‌بندی مربوطه در روش مذکور با استفاده از تکنیک جدیدی به نام k-modes انجام می‌شود که برگرفته از روش خوشه‌بندی k-means می‌باشد. در این مقاله، یک رویکرد نوین پیشنهاد شده است که خصیصه‌های وابسته را با استفاده از نسخه فازی k-modes گروه بندی می‌کند. رویکرد فازی، با دیدگاه عدم قطعیت، به پایداری بیشتر و در نتیجه درجه صحت کلاسه‌بندی بالاتری می-رسد. همچنین تغییرات دیگری در راستای انتخاب خصیصه‌های مناسب بعد از خوشه‌بندی انجام شده است. در این مرحله بر خلاف روش پیشین (غیر فازی) به جای انتخاب تعدادی خصیصه از هر خوشه بنا به وابستگی درون خوشه-ای، وابستگی هر خصیصه با کل خصیصه‌های دیگر را البته با توجه به درجه عضویت آنها در هر خوشه مورد بررسی قرار داده و بهترین را گزینش می‌کند. همچنین یک روش نوین گسسته سازی داده‌های پیوسته بر اساس معیار Fisher[3] به انضمام تکنیک جدیدی جهت مقداردهی اولیه مراکز خوشه‌ها پیشنهاد شده است.

در این مقاله، ارزیابی خصیصه‌های انتخاب شده، با استفاده از درخت تصمیم‌گیری C4.5[8] انجام می‌شود. مجموعه داده Leukemia[9] که داده میکروآرایه‌ای می‌باشدبا 73 نمونه داده‌ای و 7129 ژن، در آزمایشات مورد استفاده قرار گرفت که به دلیل پاره‌ای از پیاده‌سازی‌های سریع و کمبود حافظه، تنها 1000 ژن ابتدایی آن در نظر گرفته شده است.

در ادامه، روش ارائه شده توسط Wai-Ho Au جهت انتخاب خصیصه‌های بهینه ،رویکرد فازی پیشنهادی، معیار جدید گسسته سازی داده‌ها، تکنیک استفاده شده در مقداردهی ابتدایی مراکز خوشه‌ها و نتایج آزمایشات بررسی می‌شود. نهایتاً جمع‌بندی مقاله مد نظر قرار خواهد گرفت.

رویکرد فازی پیشنهادی

Wai-Ho Au روشی به نام k-modes را جهت خوشه‌بندی خصیصه‌ها، مبتنی بر روش k-means با دو تفاوت اصلی مطرح کرد. اول آنکه در k-modes وابستگی بین دو خصیصه به عنوان معیار شباهت مورد استفاده قرار می‌گیرد در حالیکه k-means فاصله بین دو نمونه (معمولاً فاصله اقلیدسی) را به عنوان معیار عدم شباهت استفاده می‌کند. دیگر اینکه مرکز یک خوشه در k-means[10] همواره میانگین نمونه‌های متعلق به آن است در حالیکه k-modes یکی از خصیصه‌های هر خوشه را که با دیگر خصیصه‌های متعلق به آن خوشه، طبق رابطه (1)، بیشترین ارتباط را داشته باشد به عنوان مرکز آن خوشه انتخاب می‌کند.

$$MR(A_i) = \sum_{\substack{A_j \in Cluster(i), \\ j \neq i}} R(A_i : A_j) \tag{1}$$

در جاییکه، A_i نمایانگر iامین خصیصه و *Cluster(i)* مجموعه خصیصه‌های متعلق به خوشه A_i است. $R(A_i:A_j)$ معیار همبستگی بین دو خصیصه A_i و A_j را نشان می‌دهد چنان که در رابطه (2) نشان داده شده است.

$$R(A_i : A_j) = \frac{I(A_i : A_j)}{H(A_i : A_j)} \tag{2}$$

که $I(A_i:A_j)$ "اطلاعات متقابل"(Mutual Information) دو خصیصه A_i و A_j را نشان می‌دهد در حالیکه $H(A_i:A_j)$ بیانگر "بی‌نظمی مشاع"(Joint Entropy) آنها است. در نهایت از هر خوشه تعداد معینی از خصیصه‌ها بر اساس (1) انتخاب خواهند شد.

در رویکرد فازی پیشنهاد شده، با الهام از خوشه‌بندی فازی k-means [11]، k-modes را فازی کرده به شکلی که فاصله بین دو خصوصیت ،برابر معکوس همبستگی حاصل از رابطه (2) آنها تعریف می‌شود. با استفاده از عدم قطعیت حاصل از فازی کردن سیستم و حفظ شرایط k-modes، به صورت مرحله‌ای، درجه عضویت هر خصیصه در هر خوشه را محاسبه کرده و در انتهای آن مرحله، مرکز هر خوشه را که دارای بیشترین همبستگی فازی با باقی خصیصه-هاست به عنوان مرکز آن خوشه برمی‌گزینیم. این همبستگی فازی در یک خوشه، که تغییر یافته (1) می‌باشد، طبق رابطه (3) برای خصیصه A_i در خوشه *r*ام محاسبه می‌شود.

$$MR_r(A_i) = \sum_{\substack{j=1, \\ j \neq i}}^{p} u_{rj}^m R(A_i : A_j) \tag{3}$$

که u_{rj} درجه عضویت خصیصه A_j در خوشه *r*ام و *m* یک ثابت می‌باشد که در این مقاله برابر با 1 در نظر گرفته شده است. نهایتاً نیز از بین کل خصیصه‌های موجود، تعداد معینی بنا به بیشترین rank که بر اساس رابطه (4) به دست می-آید، گزینش می‌شود.

$$rank(A_i) = \sum_{r=1}^{k} u_{ri}^m MR_r(A_i) \tag{4}$$

سایر تکنیک‌های نوین پیاده‌سازی

در این مقاله از یک روش گسسته‌سازی نوین استفاده شده است به طریقی که داده‌ها در هر بعد به تعدادی بازه معین تقسیم می‌شود چنان که معیار Fisher[3] در رابطه (5) بیشینه شود.

$$J_F = \frac{\sum_{j=1}^{\|intervals\|} (m_j - m)^2}{\sum_{j=1}^{\|intervals\|} S_j^2} \tag{5}$$

که *m* میانگین تمام داده‌ها در آن بعد، m_j و S_j نیز به ترتیب، میانگین و واریانس داده‌های قرار گرفته در بازه *j*ام می-باشند.

همچنین در این مقاله یک روش مقداردهی اولیه مراکز خوشه‌ها که در [12] ارائه شده است، به کار گرفته شده که فاصله اقلیدسی با همبستگی بین خصیصه‌ها جایگزین شده است.

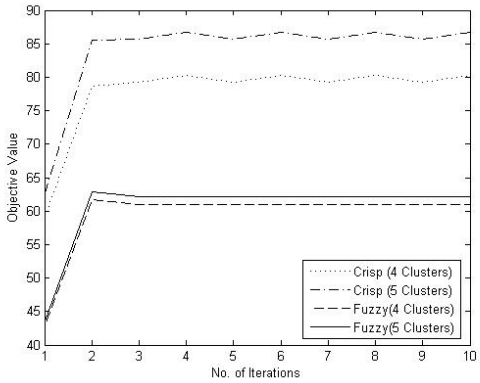
نتایج

در اینجا از درخت تصمیمگیری C4.5 جهت ارزیابی و مقایسه روش انتخاب خصیصه‌ها با رویکرد فازی و غیرفازی استفاده شده است. چنان که در جدول(1) قابل مشاهده است، نتایج رویکرد فازی از بهبود بالایی به نسبت رویکرد غیر فازی برخوردار است. در این جدول که از مجموعه داده‌هایLeukemia استفاده شده است، نتایج کلاسه بندی به ازا 3 و 4 خوشه و 1 تا 4 خصیصه به ازاء هر خوشه بیان شده است اگرچه انتخاب خصیصه‌ها بنا به rank فازی که در (4) بیان شده است، هیچ خصیصه‌ای دقیقاً به یک خوشه تعلق ندارد.

جدول (1): درجه صحت C4.5 بعد از انتخاب خصیصه‌های مناسب

	Crisp		Fuzzy	
FPC/CNo.	3	4	3	4
1	86.3014	80.8219	93.1507	89.0411
2	90.411	84.9315	95.8904	97.2603
3	95.8904	97.2603	98.6301	100
4	100	100	100	100

همچنین نمودار تابع هدف خوشه‌بندی k-modes و نوع فازی آن را برای 4 و 5 خوشه در شکل(1) نمایش داده شده است که نوسان کمتر نشان دهنده پایداری حالت فازی به نسبت حالت غیر فازی است.



شکل (1): نمودار مقدار تابع هدف در مرحله‌های مختلف

مراجع

[1] Isabelle Guyon and Andr´e Elisseeff (2003) An introduction to variable and feature selection. Journal of Machine Learning Research 3: 1157–1182.
[2] I. T. Joliffe (1986) Principal Component Analysis. New York: Springer-Verlag.
[3] K. Fukunaga (,1972) Introduction to Statistical Pattern Recognition. New York: Academic.
[4] F. Z. Bril, D. E. Brown, and N. W. Worthy (1992) Fast genetic selection of features for neural network classifiers. *IEEE Trans. Neural Networks* 3: 324–328.
[5] S.C. Madeira and A.L. Oliveira (2004) Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Trans. Computational Biology and Bioinformatics 1.1:24-45.
[6] M Xiong, W Li, J Zhao, L Jin, and E Boerwinkle (2001) Feature (gene) selection in gene expression-based tumor classification. Mol Genet Metab 73.3:239–47.
[7] Wai-Ho Au, Keith C. C. Chan, Andrew K. C. Wong, Yang Wang (2005) Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2.2:83-101
[8]Quinlan, J. R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.
[9] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. 286.5439:531-7
[10] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. (2004) A local search approximation algorithm for k-means clustering. Comput. Geom. 28:89-112.
[11] Jim C. Bezdek. (1973) Fuzzy Mathematics in Pattern Classification. PhD thesis, Applied Math. Center, Cornell University, Ithaca.
[12]Mic´o, M.L., Oncina, J., Vidal, E. (1973) A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements. Pattern Recognition. 15: 9–17.

Abstract

Classification assigns a discrete value named label to each sample in a dataset regarding its feature values. In this research, we aim to take some datasets into consideration which contains a few samples whereas a huge amount of features are provided for each sample. Most of biological datasets such as micro-arrays has this property. The fundamental contribution of this article is a fuzzy approach of clustering features proposed by Wai-Ho Au et.all which is utilized to select the best features (genes) due to classify such datasets. Our proposed method has two advantages over the crisp method. On the one hand, it leads to more stability and faster convergence; on the other hand, it improves the accuracy resulted by the classifier using the selected features. Moreover, in this paper a novel method has been proposed for the discretization of continuous data using the fisher criterion. The proposed method has reached a considerable improvement in comparison with the crisp version. The leukemia dataset has been used in all experimental results.

Keywords: Bioinformatics, Pattern recognition, Feature Selection, Fuzzy Logic, Clustering

WWW.IBP.IR iranian bioinformatics portal