

Feature selection based on information theory to select effective genes for diagnosis of cancer subtypes using microarray data

A. Tabatabaei¹, V. Derhami^{2*}, R. Sheikhpour³, M.R. Pajooan⁴

¹Ph.D Student, Department of Computer Engineering, Faculty of Engineering, Yazd University, Yazd, Iran

²Associate Professor, Department of Computer Engineering, Faculty of Engineering, Yazd University, Yazd, Iran

³Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran

⁴Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Yazd University, Yazd, Iran

Receipt in the Online Submission System: 14/6/2019, Received in Revised Form: 11/10/2019, Accepted: 3/11/2019

Abstract

Feature selection is a well-known preprocessing technique in machine learning, data mining, and especially bioinformatics microarray analysis with a high-dimension, low-sample-size (HDLSS) data. The diagnosis of genes responsible for disease using microarray data is an important issue to promoting knowledge about the mechanism of disease and improves the way of dealing with the disease. In feature selection methods based on information theory, which cover a wide range of feature selection methods, the concept of entropy is used to define criteria for relevance, redundancy, and complementarity. In this paper, we propose a new relevancy criterion based on the concept of pure continuity rather than the concept of entropy. In the proposed method, to control and reduce redundancy, the relevancy between a feature and each class is separately examined, while in most of the filter methods the value of a feature is measured based on its relation to the entire class. This solution allows us to identify the most efficient features (genes) of each class separately, while identifying common features (genes) is also possible. Discretization is another challenge in some available techniques. Using a homomorphism transformation in proposed method avoids engaging with discretization complexities, while taking advantages of it. Seven types of cancer microarrays with three types of classification models (e.g. NB, KNN, and SVM) are used to establish a comparison between the proposed method and other relevant methods. The results confirm the efficiency of the proposed method in the term of accuracy and number of selected genes as two parameters of classification.

Key words: *Feature selection, Effective genes, Cancer diagnosis, Microarray data, Machine learning, Classification.*

*Corresponding author

Address: Department of Computer Engineering, Faculty of Engineering, Yazd University, Yazd, Iran

Tel: 035-31232365

Fax: 035-38200144

E-mail: vderhami@yazd.ac.ir



انتخاب ویژگی مبتنی بر تئوری اطلاعات برای انتخاب ژن‌های مؤثر در تشخیص نوع سرطان با استفاده از داده‌های ریزآرایه

سیدابوالفضل طباطبایی^۱، ولی درهمی^{۲*}، راضیه شیخ‌پور^۳، محمدرضا پژوهان

^۱ دانشجوی دکتری مهندسی کامپیوتر، گروه مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران

^۲ دانشیار، گروه مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران

^۳ استادیار، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اردکان، اردکان، ایران

^۴ استادیار، گروه مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران

تاریخ ثبت در سامانه: ۱۳۹۸/۳/۲۴، بازنگری: ۱۳۹۸/۷/۱۹، پذیرش قطعی: ۱۳۹۸/۸/۱۲

چکیده

انتخاب ویژگی یکی از فرایندهای پیش پردازش داده‌ها در مباحث مربوط به یادگیری ماشین و داده‌کاوی محسوب می‌شود که در برخی زمینه‌ها نظیر کار با داده‌های ریزآرایه در بیوانفورماتیک که با مشکل ابعاد بالای داده‌ها در مقابل تعداد کم نمونه‌ها مواجه است، از اهمیت ویژه‌ای برخوردار است. انتخاب ویژگی‌های (ژن‌های) مؤثر در تشخیص بیماری از داده‌های ریزآرایه نقش مهمی در تشخیص زودهنگام بیماری و راه‌های مواجهه با آن ایفا می‌کند. در روش‌های انتخاب ویژگی مبتنی بر تئوری اطلاعات که طیف گسترده‌ای از روش‌های انتخاب ویژگی را شامل می‌شوند، از مفهوم بی‌نظمی برای تعریف معیارهای مرتبط بودن، افزونگی و مکمل بودن ویژگی‌ها، استفاده می‌شود. در این مقاله از مفهوم پیوستگی خالص به جای بی‌نظمی برای پیشنهاد یک معیار جدید مرتبط بودن استفاده شده است. در معیار پیشنهادی، برای کنترل و کاهش افزونگی، ارتباط یک ویژگی با تک‌تک کلاس‌ها به طور جداگانه بررسی شده است در حالی که در اکثر روش‌های فیلتر، ارزش یک ویژگی بر اساس ارتباط آن با کل کلاس‌ها سنجیده می‌شود. این راهکار باعث می‌شود که ویژگی‌های (ژن‌های) مؤثر در هر کلاس به تفکیک شناسایی شوند، در حالی که امکان شناسایی ویژگی‌های (ژن‌های) مشترک نیز فراهم است. مشکل دیگری که در برخی روش‌ها وجود دارد، مسئله گسسته‌سازی داده‌ها است. در روش ارائه شده، با استفاده از یک تبدیل مبتنی بر یکرختی ضمن استفاده از مزایای گسسته‌سازی از درگیر شدن با پیچیدگی‌های آن اجتناب شده است. برای مقایسه روش ارائه شده با تعدادی از روش‌های مرتبط، از هفت مجموعه داده ریزآرایه مربوط به انواع سرطان به همراه سه دسته‌بند پرکاربرد بیزین ساده، - کزدیک‌ترین همسایه و ماشین بردار پشتیبان استفاده شده است. نتایج تجربی، کارایی روش ارائه شده را بر اساس دو پارامتر دقت دسته‌بندی و تعداد ژن‌های انتخابی نشان می‌دهد.

کلیدواژه‌ها: انتخاب ویژگی، ژن‌های مؤثر، تشخیص سرطان، داده‌های ریزآرایه، یادگیری ماشین، دسته‌بندی

*نویسنده مسئول

نشانی: گروه مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران

تلفن: ۰۳۵-۳۱۲۳-۲۳۶۵

دورنگار: ۰۳۵-۳۸۲۰-۰۱۴۴

پست الکترونیکی: vderhami@yazd.ac.ir

۱- مقدمه

پیشرفت فناوری در حوزه بیوانفورماتیک، به خصوص فناوری ریزآرایه باعث شده است که داده‌های بیان ژنی^۲ هزاران ژن مربوط به یک نمونه مبتلا به سرطان به طور همزمان استخراج شوند [۱]. البته این استخراج داده مستلزم هزینه‌های بالا است، به همین دلیل تعداد نمونه‌ها در مقایسه با تعداد ویژگی‌های استخراج شده بسیار کم است. چنین داده‌هایی به اختصار HDLSS^۳ نامیده می‌شوند [۲] که در آن معمولاً تعداد نمونه‌ها کمتر از ۱۰۰ و تعداد ویژگی‌ها بین ۶۰۰۰ تا ۶۰۰۰۰ است [۳]. تشخیص ژن‌های مؤثر در بیماری از داده‌های ریزآرایه، یک موضوع مهم در رابطه با ارتقاء دانش در مورد مکانیزم بیماری و بهبود روش‌های مواجهه با بیماری است. مطالعات و روش‌های سنتی مرتبط با موضوع، تعداد زیادی از ژن‌ها را به عنوان ژن‌های مؤثر در بیماری در نظر می‌گیرند [۴]. به دلیل این که روش‌های تجربی برای تشخیص ژن‌های مؤثر از میان تعداد زیاد ژن‌های نامزد بسیار هزینه‌بر است، به کارگیری روش‌های انتخاب ویژگی (ژن) توصیه می‌شود. به علاوه وجود نویز^۴ و تعداد بسیار بالای اطلاعات ژنی، تحلیل داده‌های ریزآرایه را به یک دامنه مهیج تبدیل کرده است [۵]. مطالعات متعددی نشان داده‌اند که تعداد بسیاری از داده‌های بیان ژن ریزآرایه (ویژگی‌ها) از نظر معیارهای دسته‌بندی دارای بار اطلاعاتی نیستند (داده‌ی غیر مرتبط) [۶]. بنابراین انتخاب ویژگی (ژن) و فرآیند تشخیص و حذف ویژگی‌های غیرمرتبط، نقشی حیاتی در تحلیل داده‌های ریزآرایه استخراج شده از DNA^۵ ایفا می‌کنند.

بیش برآزش^۶ و پدیده تنگنای ابعاد^۷ از چالش‌های کار کردن با داده‌های HDLSS است که اولی باعث می‌شود قدرت تعمیم مدل از دست برود [۷] و دومی باعث می‌شود مدل با کاهش چشمگیر دقت مواجه شود [۸]. یک راه معمول جهت غلبه بر چالش‌های اشاره شده، کاهش ابعاد داده‌ها^۸ به یکی از دو روش استخراج ویژگی^۹ و یا از طریق انتخاب ویژگی^{۱۰} است. در روش استخراج ویژگی فضای اولیه با ابعاد بالا به یک فضای جدید با ابعاد پایین تصویر (نگاشت) می‌شود [۹]. PCA^{۱۱} و LDA^{۱۲} از روش‌های شاخص استخراج ویژگی هستند [۱۰]. در روش‌های

استخراج ویژگی، پس از تغییر فضای اولیه، امکان تفسیر و نقش ویژگی‌های اولیه از دست می‌رود. در روش‌های انتخاب ویژگی [۱۱]-[۱۳]، زیرمجموعه‌ای از ویژگی‌های موجود بر اساس معیاری مشخص، برای مدل‌سازی انتخاب می‌شوند. با وجود اینکه هر دو روش باعث کاهش فضای مسئله می‌شود ولی در کاربردهایی مانند بیوانفورماتیک که تفسیر ویژگی‌های اولیه در مدل اهمیت دارد، انتخاب ویژگی بر استخراج ویژگی ارجح است. تمرکز این مقاله هم بر روی روش‌های انتخاب ویژگی است.

به طور کلی روش‌های انتخاب ویژگی به سه دسته اصلی Filter، Wrapper و Embedded تقسیم می‌شوند [۱۴]، [۱۵]. در روش‌های Wrapper از یک دسته‌بند برای یافتن زیرمجموعه‌ای از ویژگی‌ها استفاده می‌شود تا بالاترین دقت دسته‌بندی حاصل شود. برای اجتناب از جستجوی کامل فضای ویژگی‌ها از روش‌های جستجو مثل GA^{۱۳}-[۱۶] و PSO^{۱۴}-[۲۰]، [۲۱] استفاده می‌شود. در برخی از روش‌های ترکیبی نیز معمولاً برای کاهش بیشتر فضای جستجو ابتدا از برخی الگوریتم‌های فیلتر استفاده می‌شود و تعداد زیادی از ویژگی‌های نامرتبط حذف می‌شوند [۲۲]، [۲۳]. اخیراً بهینه‌سازی چند منظوره^{۱۵} به عنوان ابزاری قدرتمند در انتخاب ویژگی مورد توجه واقع شده است [۲۴]. در کل روش‌های Wrapper نسبت به سایر روش‌ها، پیچیدگی زمانی بالاتری دارند.

روش‌های Embedded راه‌حلی متعادل بین روش‌های فیلتر و Wrapper هستند. در این روش‌ها فرایند ارزش‌دهی و انتخاب ویژگی‌ها، ضمن بهبود توانایی یادگیری و ساخت مدل اتفاق می‌افتد و لازم نیست که به طور مکرر مجموعه‌هایی از ویژگی‌ها را ارزیابی کنند. بنابراین نسبت به روش‌های Wrapper کارا تر هستند [۱۲]. روش‌های انتخاب ویژگی مبتنی بر یادگیری تنک^{۱۶} [۲۵]، [۲۶] که در سال‌های اخیر به سرعت توسعه پیدا کرده است، جزء مهمی از این دسته محسوب می‌شوند.

در روش‌های فیلتر برای به دست آوردن ارزش یک ویژگی نسبت به برچسب کلاس از معیارهایی که روی داده‌ها قابل تعریف هستند، استفاده می‌شود. به دلیل فقدان یک الگوریتم یادگیری

^۹ feature extraction^{۱۰} feature selection^{۱۱} Principle Component Analysis^{۱۲} Linear Discriminant Analysis^{۱۳} Genetic algorithms^{۱۴} Particle swarm optimization^{۱۵} Multimodal Optimization (MO) techniques^{۱۶} Sparse learning^۱ Micro Array Data set^۲ Gene expression data^۳ High Dimension Low Sample Size^۴ Noise^۵ Deoxyribonucleic Acid^۶ overfitting^۷ Curse of dimensionality^۸ dimensionality reduction

مفاهیم فوق در بسیاری از الگوریتم‌های مبتنی بر تئوری اطلاعات استفاده شده است. به عنوان مثال، لوئیس^{۱۴} و همکاران، الگوریتمی به نام MIM^{۱۵} ارائه دادند که از بهره اطلاعاتی یک ویژگی به عنوان میزان ارتباط آن با برجسب کلاس مطابق رابطه (۵) استفاده می‌کند [۳۱]. در این روش، ویژگی‌ها مستقل از هم فرض می‌شوند و امتیاز هر ویژگی به طور جداگانه و مستقل از دیگر ویژگی‌ها محاسبه می‌شود. در نظر نگرفتن خاصیت افزونگی نقطه ضعف روش فوق است.

$$\text{MIM_Score}(f_i) = \text{IG}(f_i; C) \quad (5)$$

برای بهبود عملکرد روش MIM، باتیتی^{۱۶} در روشی به نام MIFS^{۱۷}، هر دو معیار مرتبط بودن ویژگی‌ها و افزونگی آن‌ها را در فاز انتخاب ویژگی دخالت داده است [۳۲]. رابطه (۶) نحوه محاسبه امتیاز هر ویژگی را نشان می‌دهد.

$$\text{MIFS_Score}(f_i) = \text{IG}(f_i; C) - \beta \cdot \sum_{x_j \in S} I(x_i, x_j) \quad (6)$$

جمله اول رابطه (۶) میزان ارتباط (وابستگی) ویژگی به برجسب کلاس را مشخص می‌کند و جمله دوم برای کنترل افزونگی از طریق کاهش همبستگی بین ویژگی‌ها با کلیه ویژگی‌های انتخاب شده تاکنون که در مجموعه S قرار دارند، استفاده می‌شود. مقدار بتا در مقاله اصلی ارائه دهنده الگوریتم برابر با یک فرض شده است.

پنگ^{۱۸} و همکاران، الگوریتم MRMR^{۱۹} را طبق رابطه (۷) ارائه دادند که به جای مقدار بتای استفاده شده در روش MIFS، از معکوس تعداد اعضای مجموعه ویژگی‌های انتخاب شده تاکنون استفاده شده است [۳۳].

$$\text{MRMR_Score}(f_i) = \text{IG}(f_i; C) - \frac{1}{|S|} \cdot \sum_{x_j \in S} I(x_i, x_j) \quad (7)$$

براون^{۲۰} و همکاران نشان دادند که هر چه مجموعه ویژگی‌های انتخاب شده بزرگتر باشد، تاثیر جمله دوم کاهش می‌یابد [۳۴].

مشخص در طول فرایند ارزش‌گذاری ویژگی‌ها، مجموعه انتخاب شده لزوماً بهترین انتخاب (بهینه مطلق) نخواهد بود ولی به دلیل قابلیت تعمیم بیشتر و سرعت بالاتر دارای مقبولیت بیشتری نسبت به سایر روش‌ها هستند.

الگوریتم‌های مبتنی بر تئوری اطلاعات^۱ که بر اساس مفاهیمی چون مرتبط بودن^۲، افزونگی^۳ و مکمل بودن^۴ بنا می‌شوند [۲۷]، یک دسته بزرگ از الگوریتم‌های انتخاب ویژگی فیلتر را شامل می‌شوند [۱۲]. ایده اصلی بسیاری از معیارهای خلاقانه ارائه شده در این الگوریتم‌ها، بیشینه کردن ارتباط ویژگی‌ها با برجسب کلاس و در عین حال کاهش افزونگی بین ویژگی‌ها است [۲۸]. بسیاری از الگوریتم‌های این دسته از مفهوم بی‌نظمی^۵، بی‌نظمی مشروط^۶، بهره اطلاعاتی^۷ و معیار SU^۸ برای تعریف مفاهیم مرتبط بودن، افزونگی و مکمل بودن استفاده می‌کنند. بی‌نظمی، اندازه عدم قطعیت^۹ متغیر تصادفی گسسته X را نشان می‌دهد که به صورت رابطه زیر تعریف می‌شود [۲۹]:

$$H(X) = - \sum_{x_i \in X} p(x_i) \log p(x_i) \quad (1)$$

بی‌نظمی مشروط^{۱۰}، میزان عدم قطعیت متغیر تصادفی گسسته X را با شرط حضور متغیر Y نشان می‌دهد که به صورت رابطه زیر تعریف می‌شود [۲۹]:

$$H(X|Y) = - \sum_{y_i \in Y} p(y_i) \sum_{x_i \in X} p(x_i|y_i) \log p(x_i|y_i) \quad (2)$$

مفهوم بهره اطلاعاتی^{۱۱} بین دو متغیر تصادفی X, Y که توسط شانون در سال ۲۰۰۱ ارائه شد [۲۹]، برای سنجش وابستگی بین دو متغیر بر اساس مفهوم بی‌نظمی و بی‌نظمی مشروط است. به دلیل این که بهره اطلاعاتی، مقدار اطلاعاتی که دو متغیر X, Y باهم دارند را نشان می‌دهد به آن اطلاعات متقابل^{۱۲} نیز گفته می‌شود که از طریق رابطه زیر محاسبه می‌شود:

$$\text{IG}(X; Y) = H(X) - H(X|Y) \quad (3)$$

معیار SU^{۱۳} [۳۰] میزان عدم قطعیت متقارن دو متغیر تصادفی گسسته X و Y را نشان می‌دهد که به صورت رابطه ۴ تعریف می‌شود:

$$\text{SU}(X; Y) = 2 * \left[\frac{\text{IG}(X; Y)}{H(X) + H(Y)} \right] \quad (4)$$

^{۱۱} Information gain, mutual information

^{۱۲} Mutual information

^{۱۳} symmetrical uncertainty

^{۱۴} Lewis

^{۱۵} Mutual Information Maximization

^{۱۶} Battiti

^{۱۷} Mutual Information Feature Selection

^{۱۸} Peng

^{۱۹} Minimum Redundancy Maximum Relevance

^{۲۰} Brown

^۱ Information theory

^۲ Relevancy

^۳ Redundancy

^۴ Complementarity

^۵ Entropy

^۶ Conditional entropy

^۷ Information gain, mutual information

^۸ symmetrical uncertainty

^۹ Uncertainty

^{۱۰} Conditional entropy

نتایج بدست آمده ارائه می‌شود و بخش ۴ به جمع‌بندی و نتیجه‌گیری پرداخته شده است.

۲- روش پیشنهادی

در بخش قبل به دو چالش افزونگی و گسسته‌سازی داده‌های پیوسته در الگوریتم‌های مبتنی بر تئوری اطلاعات اشاره شد. مثال ارائه شده در جدول ۱ به تبیین بهتر موضوع، کمک می‌کند. فرض کنید اطلاعات مربوط به ۱۶ بیمار (نمونه) مبتلا به سه کلاس مختلف از یک سرطان (کلاس‌های C1 و C2 و C3) در این جدول گردآوری شده است. در این جدول ویژگی‌های مرتبط با هر بیمار (f1 تا f5)، گسسته فرض شده است.

جدول ۱: مثالی از محاسبه ارزش ویژگی‌ها بر اساس بی‌نظمی و بهره اطلاعات (L=low, M=mid, H=high)

Num	f1	f2	f3	f4	f5	C
1	L	L	L	H	H	c1
2	M	L	L	H	H	c1
3	H	M	L	H	H	c1
4	L	H	M	L	M	c2
5	L	H	M	L	M	c2
6	L	H	M	L	M	c2
7	M	H	M	L	L	c2
8	M	H	M	L	L	c2
9	M	H	M	H	L	c2
10	H	H	M	H	L	c2
11	H	H	M	H	L	c2
12	L	L	H	L	L	c3
13	L	L	H	L	L	c3
14	M	M	H	M	M	c3
15	M	M	H	M	M	c3
16	H	M	H	M	M	c3
	0.00	0.31	0.44	0.22	0.22	IG(f _i ,C)

جدول (۱) اطلاعات ارزشمندی از چگونگی رفتار الگوریتم MIM را نشان می‌دهد. مقدار $IG(f_i, C)$ در ردیف آخر جدول برابر با امتیازهایی است که طبق الگوریتم MIM به ویژگی‌های f1 تا f5 اختصاص داده می‌شود. با بررسی اجمالی ویژگی‌ها متوجه می‌شویم ویژگی f1 که از هیچ نظمی در هیچ یک از کلاس‌ها تبعیت نمی‌کند، پایین‌ترین امتیاز را دارد. ویژگی f3 که یک ویژگی ایده‌آل است و به تنهایی توصیف‌کننده هر سه کلاس است، بالاترین امتیاز را گرفته است. امتیاز بالای بعدی مربوط به ویژگی f2 است که یک توصیف‌کننده خوب برای کلاس C2 است. مقدار ویژگی f2 در کلیه نمونه‌های کلاس C2 برابر با H است. بنابراین در مورد ویژگی‌های مذکور الگوریتم MIM به خوبی امتیازدهی کرده است. اما نحوه امتیازدهی الگوریتم MIM در مورد ویژگی‌های f4, f5 منطقی نیست. در حالی که ویژگی f5

در الگوریتم‌های CFS^۱ و FCBF^۲ [۳۶] نیز از معیار SU برای ارزیابی میزان ارتباط بین دو ویژگی برای کنترل افزونگی و همچنین ارتباط بین ویژگی و کلاس‌ها برای اندازه‌گیری میزان مرتبط بودن، استفاده شده است. یکی از چالش‌های این الگوریتم‌ها با توجه به مثال‌های ذکر شده، مواجهه با ویژگی‌های با ارزش بالا ولی افزونه است. علاوه بر این بیشتر الگوریتم‌های ارائه شده مبتنی بر تئوری اطلاعات، روی داده‌های گسسته اعمال می‌شود. در مورد داده‌های پیوسته (عددی) باید از برخی از الگوریتم‌های گسسته‌سازی مثل Chi_merge، MDLP و CAIM استفاده کرد [۳۷]–[۳۹]. در [۴۰]–[۴۳] مروری بر الگوریتم‌های مختلف گسسته‌سازی ارائه شده است. گسسته‌سازی یکی از تکنیک‌های مهم در مبحث کاهش ابعاد محسوب می‌شود [۴۴] و علاوه بر کاهش ابعاد باعث افزایش دقت و کاهش نویز نیز می‌شود [۴۲]. برای گسسته‌سازی ابتدا باید مقادیر عددی هر ویژگی به ترتیب صعودی مرتب شود و سپس ویژگی به تعدادی بازه تقسیم شود. هر بازه معرف یک مقدار گسسته است. تعداد و اندازه بازه‌ها روی کیفیت گسسته‌سازی موثر است [۴۳] به طور کلی مسئله یافتن بهترین گسسته‌سازی در دسته الگوریتم‌های با پیچیدگی NP-complete قرار دارد [۴۵].

در این مقاله، با استفاده از مفهوم یکرختی^۳ تبدیلی از فضای فعلی به فضای جدید ویژگی، ارائه خواهد شد که ضمن استفاده از مزایای گسسته‌سازی از درگیر شدن با پیچیدگی آن اجتناب می‌شود. سپس در فضای جدید ویژگی‌ها، یک روش انتخاب ویژگی مبتنی بر تئوری اطلاعات پیشنهاد می‌شود که از مفهوم پیوستگی به جای بی‌نظمی برای تعریف یک معیار جدید CRC^۴ استفاده می‌کند و برای کاهش مشکل افزونگی، ارتباط یک ویژگی با تک‌تک کلاس‌ها به طور جداگانه بررسی می‌شود. این راهکار باعث می‌شود که ژن‌های موثر در هر کلاس به تفکیک شناسایی شوند در حالی که امکان شناسایی ژن‌های مشترک نیز فراهم است.

ساختار مقاله به شرح زیر است. در بخش ۲ روش پیشنهادی در سه قسمت ارائه می‌شود. در بخش ۲-۱- چگونگی تبدیل فضای ویژگی‌ها از پیوسته به گسسته شرح داده شده است. در بخش ۲-۲- به چگونگی امتیازدهی ژن‌ها در هر کلاس بر اساس معیار جدید می‌پردازیم و در بخش ۲-۳- انتخاب ژن‌ها بر اساس امتیاز به‌دست آمده در هر کلاس توضیح داده می‌شود. در بخش ۳

^۱ Homomorphism

^۲ Continuity based Relevancy Criterion

^۳ Correlation Feature Selection

^۴ Fast Correlation-Based Filter

برای تبدیل ویژگی عددی (پیوسته) f_i به ویژگی گسسته f'_i از زوج مرتب (f_i, C) استفاده شده است. ابتدا ویژگی f_i به ترتیب صعودی مرتب می‌شود، سپس کلیه جایجایی‌های صورت گرفته برای مرتب شدن ویژگی f_i ، روی مولفه دوم یعنی ستون برچسب کلاس اعمال می‌شود تا f'_i بدست آید. به عبارت دیگر برای بدست آوردن f'_i کافی است تمام تغییراتی که برای مرتب سازی f_i انجام می‌شود روی ستون برچسب کلاس نیز اعمال شود. به بیان ریاضی بین زوج مرتب (f_i, C) و $(sorted(f_i), f'_i)$ رابطه یکرختی وجود دارد.

جدول (۲) یک مثال فرضی از داده‌های عددی ویژگی f_i مربوط به ۱۵ نمونه از ۳ کلاس مختلف بیماری (به ترتیب ۴، ۴ و ۷ نمونه در هر کلاس) و ویژگی گسسته f'_i متناظر با آن را نشان می‌دهد.

۲-۲- امتیازدهی ژن‌ها در هر کلاس

با بررسی جدول (۲) مشاهده می‌شود که متغیر f'_i چگونه هر یک از کلاس‌های بیماری را با توجه به رابطه پیوستگی (همجواری مقداری) افزایش کرده است. منظور از افزایش مجموعه، تقسیم آن به تعدادی زیرمجموعه است که اشتراک آن‌ها تهی و اجتماع آن‌ها برابر مجموعه اولیه باشد [۴۶]. به عنوان مثال مجموعه کلاس C_1 به دو زیربخش ۴ عضوی و ۳ عضوی افزایش شده است. در این مقاله از چگونگی افزایش هر یک از کلاس‌ها توسط f'_i به عنوان معیاری برای سنجش امتیاز ژن نام، طبق رابطه (۸) استفاده شده است.

$$Score(f_i|C_j) = \sum_{k=1}^{partition(C_j|f'_i)} P(A_k) (\exp(P(A_k))), \quad (8)$$

$$P(A_k) = \frac{|A_k|}{|C_j|}$$

که f_i ویژگی مربوط به بیان ژن نام است. C_j ، کلاس نام سرطان را نشان می‌دهد. $Score(f_i|C_j)$ میزان اهمیت و امتیاز ویژگی (ژن) نام در کلاس نام را نشان می‌دهد.

یک توصیف کننده خوب برای کلاس C_1 است، امتیازی برابر با ویژگی f_4 دارد. با وجود این که کلیه مقادیر ویژگی f_4 در کلاس C_1 برابر مقدار H است ولی نمی‌توان صرف H بودن مقدار ویژگی f_4 ، در مورد کلاس بیماری اظهار نظر کرد زیرا چند نمونه از بیماران کلاس C_2 نیز در ویژگی f_4 دارای مقدار H هستند. ریشه این اشکال به نحوه گسسته‌سازی ارتباط دارد. اشکال دیگر با مقایسه امتیازات ویژگی f_2 و ویژگی f_5 مشخص می‌شود: با وجود اینکه f_2 و f_5 هرکدام توصیف کننده یک کلاس متمایز هستند ولی امتیازهای متفاوت دارند. امتیاز ویژگی توصیف کننده کلاس بزرگتر بالاتر است. بنابراین در رتبه‌بندی ویژگی‌ها، ابتدا کلیه ویژگی‌های مؤثر مرتبط با کلاس بزرگتر قرار می‌گیرند و به ترتیب نوبت به ویژگی‌های مؤثر مرتبط با کلاس‌های کوچکتر می‌رسد. این خاصیت باعث تشدید پدیده افزونگی می‌شود و اگر رویکرد مناسبی برای مقابله با ویژگی‌های افزونه انتخاب نگردد، باعث می‌شود که داده‌های نامتوازن روی کارایی الگوریتم تاثیر مخربی داشته باشند. برای کاهش این مشکلات در این بخش، یک روش انتخاب ویژگی مبتنی بر تئوری اطلاعات به نام CRFS^۱ برای انتخاب ژن‌های ریزآرایه پیشنهاد می‌شود که شامل سه قسمت تبدیل فضای ویژگی و امتیازدهی ژن‌ها در هر کلاس و انتخاب ژن‌ها براساس امتیاز به‌دست آمده در هر کلاس است.

۲-۱- تبدیل فضای ویژگی‌ها از پیوسته به گسسته

ابتدا، متغیرها و نمادهای به کار رفته در مقاله شرح داده می‌شوند:

$ClassLabel = C_{n+1} = [C_j], |C| = m, f_i \in R^{n+1}, 1 \leq i \leq d$
داده‌های ورودی شامل n نمونه و هر نمونه شامل d ویژگی است که به یکی از کلاس‌های موجود در ستون کلاس‌ها مربوط می‌شود. تعداد کل کلاس‌ها m است. منظور از f_i ویژگی نام از مجموعه ویژگی‌ها و C_j کلاس نام است.

جدول ۱: رابطه یکرختی بین $(f_i, Class\ label)$ و $(sorted(f_i), f'_i)$

f_i	1.5	3.4	4.1	3.8	2.2	4.4	2.2	1.4	2.2	2.8	4.9	4.7	4.5	2.5	3.3
Class label	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_3	C_3	C_3	C_3
$sorted(f_i)$	1.4	1.5	2.2	2.2	2.2	2.5	2.8	3.3	3.4	3.8	4.1	4.4	4.4	4.7	4.9
f'_i	C_2	C_1	C_1	C_1	C_2	C_3	C_2	C_3	C_1	C_1	C_1	C_1	C_3	C_3	C_2

پیوستگی خالص دارند. $|C_j|$ نیز معرف تعداد نمونه‌های کلاس نام است. برای مشخص شدن مفهوم پیوستگی خالص ابتدا باید پیوستگی ظاهری و سپس پیوستگی خالص را تعریف کرد.

$partition(C_j|f'_i)$ تعداد افزایش (زیربخش) کلاس C_j توسط ویژگی f'_i است. A_k زیر بخش k ام از زیربخش‌های افزایش شده کلاس C_j است. $|A_k|$ ، تعداد نمونه‌های زیر بخش k ام است که

^۱ Partition

^۱ Continuity based Relevancy criterion Feature Selection

تعریف ۱: پیوستگی ظاهری

دنباله‌ای از برجسب‌های یک کلاس در f_i' که از دو طرف به کلاس‌های متفاوت محدود است. البته در ابتدا و انتهای رشته از یک طرف به کلاس متفاوت محدود می‌شود. طبق تعریف ۱ دو پیوستگی ظاهری ۳ و ۴ تایی در مورد کلاس C_1 در جدول ۲ قابل مشاهده است.

تعریف ۲: پیوستگی خالص

پیوستگی خالص بر اساس هر دو مولفه زوج مرتب $(sorted(f_i), f_i')$ تعریف می‌شود: دنباله‌ای از برجسب‌های یک کلاس در f_i' که مقادیر متناظر آن‌ها در $sorted(f_i)$ در دو طرف، با مقادیر مجاور متفاوت باشد. هرچه پیوستگی خالص بزرگتر باشد به معنی وجود نمونه‌های بیشتر از یک کلاس با خصوصیات بیان ژنی نزدیک به هم و متمایز با سایر نمونه‌ها است. بنابراین، طبق رابطه (۱) امتیاز بیشتری به ویژگی f_i در آن کلاس تعلق می‌گیرد.

طبق تعاریف فوق پیوستگی ظاهری می‌تواند خالص نیز باشد ولی در کل پیوستگی خالص همیشه کوچکتر یا مساوی پیوستگی ظاهری است. به عنوان مثال طبق تعریف ۲، پیوستگی ظاهری ۳ تایی کلاس C_1 خالص نیست و پس از حذف دو عضو انتهایی آن که با عضوی از کلاس C_2 دارای مقادیر یکسان هستند، خالص می‌شود. بنابراین در نهایت پس از خالص‌سازی می‌توان گفت که: ویژگی f_i در کلاس C_1 دارای دو دنباله خالص ۳ و ۱ تایی است. حالا می‌توان با جایگذاری مقادیر پیوستگی خالص ویژگی f_i کلاس C_1 در رابطه ۸، ارزش ویژگی f_i را در کلاس C_1 به دست آورد:

$$Score(f_i|C_1) = \sum_{k=1}^2 P(A_k) (\exp(P(A_k))) =$$

$$1/7 * \exp(1/7) + 3/7 * \exp(3/7) = 0.8227$$

به همین صورت امتیاز ویژگی در سایر کلاس‌ها هم به دست می‌آید. به طور خلاصه مطالب گفته شده را می‌توان در قالب شبه کد ۱ نشان داد:

```

Inputs: F={f1,f2,...fd}, fi∈Rn*1
Class_label=Cn*1=[Cj], |C| = m
Output: W={w1,w2,...wd}, wi∈Rm*1
For i=1 to d do
    Create fi' upon homomorphic relation between
    two ordered pairs:
    (fi, Class label) و (sorted(fi), fi')
    For j=1 to m do
        Calculate Score(fi|Cj)
        W(i,j)= Score(fi|Cj)
    End
End

```

شبه کد ۱: محاسبه‌ی امتیاز ویژگی‌ها در هر کلاس

۲-۳- انتخاب ژن‌ها براساس امتیاز به دست آمده در هر کلاس برای رتبه‌دهی نهایی به ویژگی‌ها و مشخص نمودن ترتیبی از ویژگی‌ها به عنوان خروجی الگوریتم، لازم است ابتدا کلیه امتیازهای بدست آمده را در جدول وزنی W قرار دهیم. با توجه به تعداد ویژگی‌ها d و تعداد کلاس‌ها m، ابعاد جدول $m*d$ خواهد بود. جدول ۳ یک مثال از جدول W با فرض داشتن ۱۰ ویژگی و سه کلاس بیماری است.

جدول ۲: مثالی از جدول وزنی ویژگی‌ها

	f ₁	f ₂	f ₃	f ₄	f ₅	f ₆	f ₇	f ₈	f ₉	f ₁₀
C ₁	9.3	2.3	1.5	7.6	1.2	6.2	5.3	9.3	1.9	9.3
C ₂	5.7	0.5	5	8.9	0.4	7.4	9.3	7.8	9.4	0.4
C ₃	6.3	1.1	8.9	7.3	9.1	9.5	2.2	1.5	5.1	0.2

با در اختیار داشتن جدول وزنی ویژگی‌ها و مشاهده آن، سوال‌های متنوعی درباره ارتباط ویژگی‌ها با کلاس‌ها به ذهن خطور می‌کند که پاسخ به بعضی از آن‌ها، خود پژوهشی جدید و مستقل را می‌طلبد. به عنوان مثال با مشاهده جدول ۳ و با فرض این که حداکثر امتیاز مرتبط بودن یک ویژگی به یک کلاس ۱۰ باشد، می‌توان دریافت که ویژگی‌های f₁ و f₈ و f₁₀ با امتیاز برابر هر کدام یک توصیف کننده خوب برای کلاس C₁ است اما با این تفاوت که f₁₀ هیچ ارتباطی با دو کلاس دیگر ندارد. f₈ می‌تواند به عنوان ویژگی موثر در کلاس C₂ نیز به حساب آید و f₁ ارتباط نسبی خوبی با دو کلاس دیگر دارد. اینک بین این سه ویژگی کدامیک رتبه بالاتری داشته باشد نیاز به تحقیق بیشتر دارد. همچنین جدول ۳ نشان می‌دهد در حالی که ویژگی f₂ به هیچ یک از کلاس‌ها مرتبط نیست ویژگی f₄ ارتباط خوبی با هر سه کلاس دارد. بنابراین ویژگی‌های مشترک نیز قابل بررسی هستند. ضوابط مختلفی برای انتخاب و رتبه‌بندی ویژگی‌ها می‌توان تعیین کرد. روشی که در این مقاله برای رتبه بندی ویژگی‌ها در نظر گرفته شده است به این صورت است که ابتدا در هر کلاس ویژگی‌ها بر اساس امتیازشان جداگانه مرتب می‌شوند و سپس ویژگی‌های با امتیاز بالاتر از هر کلاس به ترتیب با حذف تکراری‌ها در لیست خروجی قرار می‌گیرند. به عنوان مثال با توجه به جدول ۳، ترتیب نهایی ویژگی‌ها به صورت f₁, f₉, f₆, f₈, f₇, f₅, f₁₀, f₄, f₃, f₂ خواهد بود.

۳- نتایج

در این بخش، کارایی روش پیشنهادی با ۵ روش انتخاب ویژگی دیگر مقایسه می‌شود. از میان این ۵ روش، سه روش انتخاب ویژگی ReliefF [۴۷]، IG [۳۱] و CFS که از روش‌های رایج در ارتباط با داده‌های ریزآرایه هستند، در نرم‌افزار یادگیری

برابر با تعداد ویژگی‌های روش CFS در نظر گرفته شده است. در مورد روش پیشنهادی برای محدود کردن انتخاب تعداد ویژگی‌ها از حد $8 \times m$ (تعداد کلاس‌های هر مجموعه داده) به عنوان حداکثر ژن‌های انتخابی استفاده شده است. به عبارت دیگر حداکثر ژن‌های انتخابی از هر کلاس بیشتر از ۸ نخواهد بود. دقت‌های بدست آمده الگوریتم پیشنهادی در هر یک از جدول‌ها، حاصل مقایسه در این فضای محدود است. در جدول ۵ که از دسته‌بند NB برای ارزیابی الگوریتم‌ها استفاده شده است، برتری محسوس CRFS مشهود است. CRFS در اکثر مجموعه داده‌ها به بالاترین دقت دست یافته است. اگرچه در برخی موارد مثل DLBC, Leukemia1, Leukemia2 به طور مشترک برخی دیگر از الگوریتم‌ها نیز به بالاترین دقت رسیده‌اند، اما در سه مورد GCM, Cancers, Lung1 به تنهایی بالاترین دقت را کسب کرده است. نکته مهم دیگری که در مورد GCM, Cancers, DLBCL باید به آن اشاره داشت، تعداد بسیار پایین ژن‌های انتخابی CRFS در مقایسه با سایر روش‌ها است و تنها روش CFS تا حدودی با نتایج روش پیشنهادی قابل مقایسه است. به خصوص در مورد مجموعه داده‌ی GCM که دارای ۱۴ کلاس مختلف است، انتخاب تنها ۳۸ ژن که حتی به میانگین ۳ ژن در هر کلاس هم نمی‌رسد، نشان دهنده این است که روش CRFS در تشخیص ژن‌های مؤثر عملکرد قابل قبولی دارد.

ماشین Weka [۴۸] پیاده‌سازی شده است، همچنین دو روش انتخاب ویژگی CAM و DSCFS توسط ونگ^۱ و وی^۲ [۴۹] در سال ۲۰۱۷ بر اساس اندازه‌گیری توانایی یک ویژگی در دسته‌بندی زیر مسائل دو کلاسه و مفهوم مکمل بودن ویژگی‌ها ارائه شده است. در حالی که تمرکز اصلی در روش انتخاب ویژگی CAM بر حذف ویژگی‌های غیر مرتبط است، در روش DSCFS، حذف ویژگی‌های افزونه اهمیت بیشتری دارد. توجه همزمان به هر سه معیار مرتبط بودن، افزونگی و مکمل بودن ویژگی‌ها، دلیل انتخاب این روش‌ها برای مقایسه با روش پیشنهادی است. برای مقایسه از سه دسته‌بند رایج SVM^۳ و KNN^۴ و NB^۵ که هر سه در Weka پیاده‌سازی شده‌اند، و هفت مجموعه داده ریزآرایه مربوط به انواع سرطان، استفاده شده است. هر مجموعه داده، دارای داده‌های آموزش و آزمایش مجزا است که جزئیات مربوط به آن‌ها در جدول ۴ توصیف شده است. روش مقایسه به این صورت است که ابتدا فرایند انتخاب ویژگی توسط هر یک از شش روش انتخاب ویژگی، روی داده‌های آموزش هر یک از هفت مجموعه داده انجام می‌شود. سپس دقت دسته‌بندی ویژگی‌های انتخاب شده با توجه به داده‌های آزمایش، توسط هر یک از دسته‌بندها بدست می‌آید. دقت دسته‌بندی، درصد نمونه‌های دسته‌بندی شده درست را در میان تمام نمونه‌ها نشان می‌دهد. جدول‌های ۵ و ۶ و ۷ نتایج بدست آمده را نشان می‌دهد. مقادیر پررنگ‌تر در جدول‌ها، معرف بالاترین دقت بدست آمده در هر مجموعه داده است. به‌ازای هر مجموعه داده، دو سطر در جدول‌ها وجود دارد. یک سطر مربوط به دقت و سطر دیگر مربوط به تعداد ویژگی‌هایی از هر الگوریتم است که در دسته‌بندی مشارکت داشته‌اند. با توجه به اینکه برای ارزیابی روش‌های Relief و IG لازم است تعداد ویژگی‌های انتخابی تعیین شود، ونگ در [۴۹] از تعداد ویژگی‌های هر یک از روش‌های CAM, CFS, DSCFS به طور مشابه برای بدست آوردن دقت روش‌های Relief, IG استفاده کرد. بنابراین در سه حالت، دقت روش‌های Relief, IG بدست آمده است. در حالت اول که با نام‌های Relief, IG در جدول‌ها مشخص شده است، تعداد ویژگی‌های در نظر گرفته شده برای Relief, IG با تعداد ویژگی‌های روش DSCFS برابر است. در حالت دوم که با نام‌های I-cam, R-cam مشخص شده‌اند تعداد ویژگی‌ها برابر با تعداد ویژگی‌ها در روش CAM در نظر گرفته شده است و در حالت سوم با نام‌های I-CFS, R-CFS تعداد ویژگی‌های Relief, IG

^۴ k-nearest neighbor

^۵ Naive Bayes

^۱ Wang

^۲ Wei

^۳ Support VectorMachine

جدول ۴: مجموعه داده‌ها

Dataset	No. of classes	No. of Features	No. of samples		Reference
			Training	Testing	
Leukemia1	3	7129	38	34	Golub et al.[6]
Lung1	3	7129	64	32	Beer et al.[50]
Leukemia2	3	12582	57	15	Armstrong et al.[51]
Breast	5	9216	54	30	Perou et al.[52]
DLBCL	6	4026	58	30	Alizadeh et al.[53]
Cancers	11	12533	100	74	Su et al.[54]
GCM	14	16063	144	46	Ramaswamy et al. [55]

جدول ۵: مقایسه دقت دسته‌بندی ۶ الگوریتم با بکارگیری دسته‌بند NB

Datasesets	OURS	Wang's paper(report)								
		CRFS	CAM	DSCFS	CFS	IG	ReliefF	I-CAM	R-CAM	I-CFS
Leukemia1	97.06	85.29	97.06	97.06	91.18	97.06	97.06	97.06	97.06	94.12
	4	20	2	76	2	2	20	20	76	76
Lung1	87.50	84.38	81.25	78.13	78.13	71.88	84.38	84.38	81.25	81.25
	4	598	3	151	3	3	598	598	151	151
Leukemia2	100	93.33	100	100	73.33	100	100	100	100	100
	4	42	3	173	3	3	42	42	173	173
Breast	83.33	80	90	83.33	70	83.33	80	83.33	80	80
	30	152	6	180	6	6	152	152	180	180
DLBCL	96.67	90	93.33	90	90	93.33	70	93.33	96.67	93.33
	18	2092	181	257	181	181	2092	2092	257	257
Cancers	93.24	78.38	79.73	85.14	81.08	81.08	86.49	83.78	85.14	86.49
	52	1923	648	356	648	648	1923	1923	356	356
GCM	67.39	50	56.52	56.52	50	43.48	50	45.65	39.13	39.13
	38	4674	1739	128	1739	1739	4674	4674	128	128

جدول ۶: مقایسه دقت دسته‌بندی ۶ الگوریتم با بکارگیری دسته‌بند SVM

Datasesets	OURS	Wang's paper(report)								
		CRFS	CAM	DSCFS	CFS	IG	ReliefF	I-CAM	R-CAM	I-CFS
Leukemia1	97.06	97.06	85.29	88.24	79.41	85.29	94.12	94.12	94.12	94.12
	4	20	2	76	2	2	20	20	76	76
Lung1	81.25	87.5	75	84.38	78.13	81.25	84.38	84.38	84.38	84.38
	6	598	3	151	3	3	598	598	151	151
Leukemia2	100	100	100	100	73.33	93.33	100	100	100	100
	8	42	3	173	3	3	42	42	173	173
Breast	96.67	96.67	96.67	86.67	76.67	63.33	96.67	93.33	96.67	93.33
	30	152	6	180	6	6	152	152	180	180
DLBCL	100	100	100	100	100	96.67	53.33	100	100	100
	20	2092	181	257	181	181	2092	2092	257	257
Cancers	82.43	89.19	95.95	93.24	95.95	93.24	93.24	94.59	93.24	94.59
	66	1923	648	356	648	648	1923	1923	356	356
GCM	52.17	52.17	47.83	58.7	50	41.3	54.35	47.83	41.30	43.48
	38	4674	1739	128	1739	1739	4674	4674	128	128
	Win/Loss	5/2	4/3	4/3	6/1	5/2	4/3	5/2	5/2	5/2

جدول ۷: مقایسه دقت دسته‌بندی ۶ الگوریتم با بکارگیری دسته‌بند KNN

Datasesets	OURS	Wang's paper(report)								
		CRFS	CAM	DSCFS	CFS	IG	ReliefF	I-CAM	R-CAM	I-CFS
Leukemia1	97.06	76.47	97.06	85.29	85.29	97.06	94.12	94.12	97.06	94.12
	4	20	2	76	2	2	20	20	76	76
Lung1	87.5	81.25	75	68.75	84.38	71.88	84.38	81.25	81.25	81.25
	11	598	3	151	3	3	598	598	151	151
Leukemia2	100	93.33	93.33	93.33	66.67	93.33	93.33	93.33	93.33	93.33
	6	42	3	173	3	3	42	42	173	173
Breast	93.33	90	96.67	90	66.67	83.33	96.67	100	96.67	100
	30	152	6	180	6	6	152	152	180	180

DLBCL	100	96.67	96.67	96.67	96.67	96.67	70	100	96.67	96.67
	18	2092	181	257	181	181	2092	2092	257	257
Cancers	85.14	87.84	89.19	90.54	87.84	87.84	87.84	89.19	90.54	89.19
	46	1923	648	356	648	648	1923	1923	356	356
GCM	63.04	52.17	47.83	52.17	45.65	43.48	47.83	45.65	52.17	41.30
	38	4674	1739	128	1739	1739	4674	4674	128	128

هرسه مفهوم می‌تواند به بهبود عملکرد روش ارائه شده به خصوص در مواردی که تعداد کلاس‌ها افزایش می‌یابد، کمک کند. نتایج به‌دست آمده در این مقاله نیز نشان می‌دهد که با وجود برتری نسبی روش ارائه شده نسبت به سایر روش‌ها، در مواردی مثل مجموعه داده‌های Cancers و GCM که تعداد کلاس‌ها افزایش می‌یابد (۱۱ و ۱۴ کلاس)، کارایی روش ارائه شده کاهش یافته است. بنابراین برای افزایش کارایی روش انتخاب ویژگی CRFS پیشنهاد می‌شود علاوه بر تمرکز بر مفهوم Relevancy، به مفاهیم Redundancy و Complementarity نیز در مرحله‌ی انتخاب نهایی ژن‌ها، توجه شود.

۵- مراجع

- [1] G. Piatetsky-Shapiro and P. Tamayo, "Microarray data mining: facing the challenges," *ACM SIGKDD Explor. Newsl.*, vol. 5, no. 2, pp. 1–5, 2003.
- [2] L. Zhang and X. Lin, "Some considerations of classification for high dimension low-sample size data," *Stat. Methods Med. Res.*, vol. 22, no. 5, pp. 537–550, 2011.
- [3] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 1157–1182, 2003.
- [4] A. M. Glazier, "Finding Genes That Underlie Complex Traits," *Science (80-.)*, vol. 298, no. 5602, pp. 2345–2349, 2002.
- [5] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [6] T. R. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science (80-.)*, vol. 286, no. 5439, pp. 531–527, 1999.
- [7] R. Clarke *et al.*, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nat. Rev. Cancer*, vol. 8, no. 1, pp. 37–49, 2008.
- [8] M. Köppen, "The curse of dimensionality," *5th Online World Conf. Soft Comput. Ind. Appl.*, vol. 1, pp. 4–8, 2000.
- [9] L. Huan and H. Motoda, "Feature extraction, construction and selection: A data mining perspective," *Comput. Math. with Appl.*, vol. 38, no. 1, p. 125, 1999.
- [10] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach.*

اطلاعات جدول ۶ نشان می‌دهد که کارایی CRFS موقعی که از دسته‌بند SVM استفاده می‌شود به خوبی حالت قبل که از NB استفاده شد، نیست. بنابراین در این وضعیت برای مقایسه بهتر بین روش ارائه شده و سایر روش‌ها از سطر آخر جدول ۶ کمک می‌گیریم. مقایسه بر اساس دو پارامتر دقت دسته‌بندی و تعداد ژن‌های انتخابی انجام می‌شود به این صورت که پارامتر دقت، اولویت بالاتر دارد و در صورت دقت مساوی، الگوریتم با تعداد ژن‌های کمتر، بهتر است. در سطر آخر جدول ۶، نتیجه مقایسه بین روش CRFS با سایر روش‌ها روی هفت مجموعه داده به صورت برد، باخت جمع‌آوری شده است. نتیجه، برتری نسبی روش CRFS را نشان می‌دهد. در جدول ۷ از دسته‌بند KNN برای ارزیابی الگوریتم‌ها استفاده شده است. در این جدول هم نظیر جدول ۵ برتری محسوس CRFS مشهود است و نیازی به تشکیل جدول برد و باخت نیست. CRFS در پنج مجموعه داده به بالاترین دقت رسیده است که در سه مورد Leukemia2، Lung1، GCM به تنهایی بالاترین دقت را کسب کرده و در مورد DLBCL نیز با توجه به تعداد ژن انتخابی عملکرد بهتری دارد. در مورد Leukemia1 با وجود دستیابی به بالاترین دقت، عملکرد DSCFS با توجه به تعداد ژن انتخابی بهتر است

۴- نتیجه گیری

در این مقاله، یک روش انتخاب ویژگی به نام CRFS برای انتخاب ژن‌های مؤثر در تشخیص انواع سرطان از داده‌های ریزآرایه پیشنهاد شد که در آن از مفهومی به نام پیوستگی خالص برای تعریف معیار مرتبط بودن (Relavancy) استفاده شد. در روش پیشنهادی، برای کاهش افزونگی (Redundancy) امتیاز ژن‌ها در هر کلاس به طور جداگانه محاسبه شد و انتخاب نهایی ژن‌ها براساس انتخاب ژن‌های بهتر از هر کلاس، انجام شد. نتایج به‌دست آمده، کارایی روش ارائه شده را با توجه همزمان به دو معیار دقت دسته‌بندی و تعداد ژن انتخابی نشان می‌دهد.

با توجه به این‌که مفاهیم Relevancy، Redundancy و Complementarity در روش‌های انتخاب ویژگی مبتنی بر تئوری اطلاعات نقش کلیدی ایفا می‌کنند، توجه همزمان به

- filters and GA wrapper approaches for gene selection,” *J. Theor. Appl. Inf. Technol.*, vol. 47, no. 3, pp. 1338–1343, 2013.
- [24] S. Kamyab and M. Eftekhari, “Feature selection using multimodal optimization techniques,” *Neurocomputing*, vol. 171, pp. 586–597, 2016.
- [25] Z. Ma *et al.*, “Discriminating joint feature analysis for multimedia data understanding,” *IEEE Trans. Multimed.*, vol. 14, no. 6, pp. 1662–1672, 2012.
- [26] C. Shi, Q. Ruan, and G. An, “Sparse feature selection based on graph Laplacian for web image annotation,” *Image Vis. Comput.*, vol. 32, no. 3, pp. 189–201, 2014.
- [27] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, 2014.
- [28] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
- [29] C. E. Shannon, “A mathematical theory of communication,” *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
- [30] W. H. Vetterling, William T. and Teukolsky, Saul A. and Press, *Numerical Recipes: Example Book (C)*, 2nd ed. Press Syndicate of the University of Cambridge, 1992.
- [31] D. D. Lewis, “Feature selection and feature extraction for text categorization,” *Proc. Speech Nat. Lang. Work.*, p. 212, 1992.
- [32] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Net Learning,” *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.
- [33] Hanchuan Peng, Fuhui Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [34] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan, “Conditional Likelihood Maximisation: A Unifying Framework for Mutual Information Feature Selection,” *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, 2012.
- [35] M. Hall, “Correlation-based Feature Selection for Machine Learning,” *Methodology*, 1999.
- [36] L. Yu and H. Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution,” *Int. Conf. Mach. Learn.*, 2003.
- [37] R. Kerber, “Chimerge: Discretization of numeric attributes,” *Proc. tenth Natl. Conf. Artif. Intell.*, 1992.
- [38] K. Irani and U. Fayyad, “Multi-Interval Discretization of Continuous-Valued Attributes for Classification learning,” *Proc. Natl. Acad. Sci. U. S. A.*, 1993.
- [39] L. A. Kurgan and K. J. Cios, “CAIM Discretization Algorithm,” *IEEE Trans. Knowl. Intell.*, vol. 23, no. 2, pp. 228–233, 2001.
- [11] VIJAY SUNDER NAGA PAPPU, *Supervised machine learning models for feature selection and classification on high dimensional datasets*. 2013.
- [12] J. Li *et al.*, “Feature selection: A data perspective,” *ACM Comput. Surv.*, vol. 50, no. 6, p. 94, 2017.
- [13] V. Bolón-Canedo, N. Sánchez-Marroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, “A review of microarray datasets and applied feature selection methods,” *Inf. Sci. (Ny)*, 2014.
- [14] B. Venkatesh and J. Anuradha, “A Review of Feature Selection and Its Methods,” *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, 2019.
- [15] N. Almgren and H. Alshamlan, “A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification,” *IEEE Access*, vol. 7, pp. 78533–78548, 2019.
- [16] S. Begum, A. A. Ansari, S. Sultan, and R. Dam, “A Hybrid Model for Optimum Gene Selection of Microarray Datasets,” in *Recent Developments in Machine Learning and Data Analytics*, Springer, 2019, pp. 423–430.
- [17] M. Ghosh, S. Adhikary, K. K. Ghosh, A. Sardar, S. Begum, and R. Sarkar, “Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods,” *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 159–176, 2019.
- [18] A. K. Shukla, P. Singh, and M. Vardhan, “A new hybrid feature subset selection framework based on binary genetic algorithm and information theory,” *Int. J. Comput. Intell. Appl.*, p. 1950020, 2019.
- [19] C. Yan, J. Ma, H. Luo, and A. Patel, “Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets,” *Chemom. Intell. Lab. Syst.*, vol. 184, pp. 102–111, 2019.
- [20] A. Chinnaswamy and R. Srinivasan, “Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data,” in *Innovations in Bio-Inspired Computing and Applications*, Springer, 2016, pp. 229–239.
- [21] P. Ghamisi and J. A. Benediktsson, “Feature selection based on hybridization of genetic algorithm and particle swarm optimization,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 309–313, 2015.
- [22] I. Jain, V. K. Jain, and R. Jain, “Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification,” *Appl. Soft Comput.*, vol. 62, pp. 203–215, 2018.
- [23] S. S. Shreem, S. Abdullah, M. Z. A. Nazri, and M. Alzaqebah, “Hybridizing relief, mRMR

- diagnosis using tumor gene expression signatures,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 26, pp. 15149–15154, 2001.
- [40] J. Dougherty, R. Kohavi, and M. Sahami, “Supervised and Unsupervised Discretization of Continuous Features,” in *Machine Learning Proceedings 1995*, 1995, pp. 194–202.
- [41] S. Kotsiantis and D. Kanellopoulos, “Discretization Techniques: A recent survey,” *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 32, no. 1, pp. 47–58, 2006.
- [42] S. García, J. Luengo, J. A. Sáez, V. López, and F. Herrera, “A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 734–750, 2013.
- [43] S. Sharmin, A. A. Ali, M. A. H. Khan, and M. Shoyaib, “Feature Selection and Discretization based on Mutual Information,” in *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 2017, pp. 1–6.
- [44] H. Liu, F. Hussain, C. L. Tan, and M. Dash, “Discretization: An enabling technique,” *Data Min. Knowl. Discov.*, vol. 6, no. 4, pp. 393–423, 2002.
- [45] B. S. Chlebus and S. H. Nguyen, “On Finding Optimal Discretizations for Two Attributes BT - Rough Sets and Current Trends in Computing,” 1998, pp. 537–544.
- [46] C. C. Pinter, *A Book of SET THEORY*. Dover Publications, 2014.
- [47] I. Kononenko, “Estimating attributes: analysis and extensions of RELIEF,” in *European conference on machine learning*, 1994, pp. 171–182.
- [48] E. Witten, Ian H. and Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.
- [49] S. Wang and J. Wei, “Feature selection based on measurement of ability to classify subproblems,” *Neurocomputing*, vol. 224, pp. 155–165, 2017.
- [50] D. G. Beer *et al.*, “Gene-expression profiles predict survival of patients with lung adenocarcinoma,” *Nat. Med.*, vol. 8, no. 8, p. 816, 2002.
- [51] S. A. Armstrong *et al.*, “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia,” *Nat. Genet.*, vol. 30, no. 1, p. 41, 2002.
- [52] C. M. Perou *et al.*, “Molecular portraits of human breast tumours,” *Nature*, vol. 406, no. 6797, p. 747, 2000.
- [53] A. A. Alizadeh *et al.*, “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,” *Nature*, vol. 403, no. 6769, p. 503, 2000.
- [54] A. I. Su *et al.*, “Molecular classification of human carcinomas by use of gene expression signatures,” *Cancer Res.*, vol. 61, no. 20, pp. 7388–7393, 2001.
- [55] S. Ramaswamy *et al.*, “Multiclass cancer