Published by: Iranian Society for Biomedical Engineering / www.isbme.ir

Iranian Journal of Biomedical Engineering

www.ijbme.org / P-ISSN: 2008-5869 / E-ISSN: 8006-9685

Volume 13, Issue 4, Winter 2020, 337 - 348

Feature Selection based on Information Theory to Select Effective Genes for Diagnosis of Cancer Subtypes using Microarray Data

Tabatabaei, Abolfazl [1] / Derhami, Vali [2*] / Sheikhpour, Razieh [3] / Pajoohan, Mohammad Reza [4]

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

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

[3] - Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Ardakan University, Ardakan, Iran

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

ARTICLE INFO

DOI: 10.22041/IJBME.2019.109466.1490

Received: 14 June 2019 Revised: 11 October 2019 Accepted: 3 November 2019

KEYWORDS

Feature Selection
Effective Genes
Cancer Diagnosis
Microarray Data
Machine Learning
Classification

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.

[*]Corresponding Author

Address	Department of Computer Engineering, Faculty of Engineering, Yazd University, Yazd, Iran		
Postal Code	89195-741	**Tel**	*+98-35-31232365*
E-Mail	*vderhami@yazd.ac.ir*	**Fax**	*+98-35-38200144*

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

ناشر: انجمن مهندسی پزشکی ایران / www.isbme.ir

مجله‌ی مهندسی پزشکی زیستی

شاپای چاپی: ۵۸۶۹-۲۰۰۸ / شاپای الکترونیکی: ۹۶۸۵-۸۰۰۶ / www.ijbme.org

دوره: ۱۳، شماره: ۴، زمستان ۱۳۹۸، ۳۳۷ - ۳۴۸

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

طباطبایی، سیدابوالفضل [۱] / درهمی، ولی [۲*] / شیخ‌پور، راضیه [۳] / پژوهان، محمدرضا [۴]

[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، دانشکده‌ی فنی و مهندسی، دانشگاه اردکان، اردکان، ایران

[۴] – استادیار، گروه مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران

مشخصات مقاله

شناسه‌ی دیجیتال: 10.22041/IJBME.2019.109466.1490

ثبت در سامانه: ۲۴ خرداد ۱۳۹۸ بازنگری: ۱۹ مهر ۱۳۹۸ پذیرش: ۱۲ آبان ۱۳۹۸

واژه‌های کلیدی

انتخاب ویژگی

ژن‌های موثر

تشخیص سرطان

داده‌های ریزآرایه

یادگیری ماشین

دسته‌بندی

چکیده

انتخاب ویژگی یکی از فرایندهای پیش‌پردازش داده‌ها در مباحث مربوط به یادگیری ماشین و داده‌کاوی به شمار می‌رود که در برخی زمینه‌ها مانند کار با داده‌های ریزآرایه در بیوانفورماتیک که با مشکل ابعاد بالای داده‌ها در مقابل تعداد کم نمونه‌ها مواجه است، از اهمیت ویژه‌ای برخوردار می‌باشد. انتخاب ویژگی‌های (ژن‌های) موثر 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 به جای بی‌نظمی از مفهوم پیوستگی خالص برای پیشنهاد یک معیار جدید مرتبط بودن استفاده شده است. در این 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 شده که ویژگی‌های موثر در هر کلاس به تفکیک شناسایی شوند، در حالی که امکان شناسایی ویژگی‌های مشترک نیز وجود دارد. یکی دیگر از مشکل‌های موجود در برخی از روش‌ها، مساله‌ی گسسته‌سازی داده‌ها است. در روش پیشنهادی این مقاله، با استفاده از یک تبدیل مبتنی بر یک‌ریختی، ضمن استفاده از مزایای گسسته‌سازی، از درگیر شدن با پیچیدگی‌های آن نیز اجتناب شده است. برای مقایسه‌ی روش پیشنهادی با تعدادی از روش‌های مرتبط، از هفت مجموعه‌ی داده‌ی ریزآرایه مربوط به انواع سرطان به همراه سه دسته‌بند پرکاربرد بیزین ساده، k-نزدیک‌ترین همسایه و ماشین بردار پشتیبان استفاده شده است. نتایج تجربی نشان دهنده‌ی کارایی روش ارائه شده بر اساس دو پارامتر دقت دسته‌بندی و تعداد ژن‌های انتخابی می‌باشد.

***نویسنده‌ی مسئول**

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

کد پستی ۸۹۱۹۵-۷۴۱ **تلفن** ۰۳۵-۳۱۲۳۲۳۶۵-۹۸+

پست الکترونیک *vderhami@yazd.ac.ir* **دورنگار** ۰۳۵-۳۸۲۰۰۱۴۴-۹۸+

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

۱- مقدمه

پیشرفت فناوری در حوزه‌ی بیوانفورماتیک، به خصوص فناوری ریزآرایه[۱] باعث شده است که داده‌های بیان ژنی[۲] هزاران ژن مربوط به یک نمونه‌ی مبتلا به سرطان به طور هم‌زمان استخراج شوند [۱]. البته این استخراج داده مستلزم هزینه‌های بالا بوده و به همین دلیل تعداد نمونه‌ها در مقایسه با تعداد ویژگی‌های استخراج شده بسیار اندک است. چنین داده‌هایی به اختصار HDLSS[۳] نامیده شده [۲] که در آن معمولا تعداد نمونه‌ها کم‌تر از ۱۰۰ و تعداد ویژگی‌ها بین ۶۰۰۰ تا ۶۰۰۰۰ می‌باشد [۳]. تشخیص ژن‌های موثر در بیماری از داده‌های ریزآرایه، یک موضوع مهم در رابطه با ارتقای دانش در مورد مکانیسم بیماری و بهبود روش‌های مواجهه با بیماری است. مطالعات و روش‌های سنتی مرتبط با موضوع، تعداد زیادی از ژن‌ها را به عنوان ژن‌های موثر در بیماری در نظر می‌گیرند [۴]. به دلیل هزینه‌ی بالای روش‌های تجربی برای تشخیص ژن‌های موثر از میان تعداد زیاد ژن‌های نامزد، به‌کارگیری روش‌های انتخاب ویژگی توصیه می‌شود. علاوه بر این وجود نویز[۴] و تعداد بسیار بالای اطلاعات ژنی، تحلیل داده‌های ریزآرایه را به یک دامنه‌ی مهیج تبدل کرده است [۵]. در مطالعات متعددی نشان داده شده است که تعداد بسیاری از داده‌های بیان ژن ریزآرایه (ویژگی‌ها) از نظر معیارهای دسته‌بندی دارای بار اطلاعاتی نیستند (داده‌ی غیرمرتبط) [۶]. بنابراین انتخاب ویژگی (ژن) و فرایند تشخیص و حذف ویژگی‌های غیرمرتبط، نقشی حیاتی در تحلیل داده‌های ریزآرایه‌ی استخراج شده از DNA[۵] ایفا می‌کنند.

بیش‌برازش[۶] و پدیده‌ی تنگنای ابعاد[۷] از چالش‌های کار کردن با با داده‌های HDLSS است که اولی باعث از دست رفتن قدرت تعمیم مدل [۷] و دومی باعث کاهش چشم‌گیر دقت مدل می‌شود [۸]. یک راه متداول برای غلبه بر این چالش‌ها، کاهش ابعاد داده‌ها[۸] به یکی از دو روش استخراج ویژگی[۹] یا انتخاب ویژگی[۱۰] است. در روش استخراج ویژگی، فضای اولیه با ابعاد بالا به یک فضای جدید با ابعاد پایین تصویر (نگاشت) می‌شود [۹] که PCA[۱۱] و LDA[۱۲] از شاخص‌ترین این روش‌ها به شمار می‌روند [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 می‌شود. با وجود این‌که هر دو روش باعث کاهش فضای مساله می‌شوند اما در کاربردهایی مانند بیوانفورماتیک که تفسیر ویژگی‌های اولیه در مدل اهمیت زیادی دارد، انتخاب ویژگی به استخراج ویژگی ترجیح داده می‌شود. در این مقاله نیز روی روش‌های انتخاب ویژگی تمرکز شده است.

به طور کلی روش‌های انتخاب ویژگی به سه دسته‌ی اصلی فیلتر، دسته‌بند[۱۳] و تعبیه شده[۱۴] (توکار) تقسیم می‌شوند [۱۴، ۱۵]. در روش‌های دسته‌بند از یک دسته‌بند برای یافتن زیرمجموعه‌ای از ویژگی‌ها استفاده شده تا بالاترین دقت دسته‌بندی حاصل شود. برای اجتناب از جست‌وجوی کامل فضای ویژگی‌ها، از روش‌های جست‌وجو مانند GA[۱۵] [۱۶-۱۹] و PSO[۱۶] [۲۰، ۲۱] استفاده می‌شود. در برخی از روش‌های ترکیبی نیز معمولا برای کاهش بیش‌تر فضای جست‌وجو ابتدا از برخی الگوریتم‌های فیلتر استفاده شده و تعداد زیادی از ویژگی‌های نامرتبط حذف می‌شوند [۲۲، ۲۳]. اخیرا بهینه‌سازی چندمنظوره[۱۷] به عنوان ابزاری قدرتمند در انتخاب ویژگی مورد توجه واقع شده است [۲۴]. در کل روش‌های دسته‌بند نسبت به سایر روش‌ها، پیچیدگی زمانی بالاتری دارند. روش‌های تعبیه شده، راه حل متعادلی بین روش‌های فیلتر و دسته‌بند هستند. در این روش‌ها فرایند ارزش‌دهی و انتخاب ویژگی‌ها ضمن بهبود توانایی یادگیری و ساخت مدل اتفاق افتاده و لازم نیست که به طور مکرر مجموعه‌هایی از ویژگی‌ها ارزیابی شوند، بنابراین نسبت به روش‌های دسته‌بند کارایی بیش‌تری دارند [۱۲]. روش‌های انتخاب ویژگی مبتنی بر یادگیری تنک[۱۸] [۲۵، ۲۶] که در سال‌های اخیر به سرعت توسعه پیدا کرده‌اند، یکی از انواع مهم این دسته می‌باشند.

در روش‌های فیلتر برای به دست آوردن ارزش یک ویژگی نسبت به برچسب کلاس، از معیارهای قابل تعریف روی داده‌ها

---

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
۹ Feature Extraction
۱۰ Feature Selection
۱۱ Principle Component Analysis
۱۲ Linear Discriminant Analysis
۱۳ Wrapper
۱۴ Embedded
۱۵ Genetic Algorithms
۱۶ Particle Swarm Optimization
۱۷ Multimodal Optimization (MO) Techniques
۱۸ Sparse Learning

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

استفاده می‌شود. در روش‌های فیلتر به دلیل فقدان یک الگوریتم یادگیری مشخص در طول فرایند ارزش‌گذاری ویژگی‌ها، مجموعه‌ی انتخاب شده لزوما بهترین انتخاب (بهینه‌ی مطلق) نخواهد بود اما این روش‌ها با توجه به قابلیت تعمیم بیش‌تر و سرعت بالاتر، از مقبولیت بیش‌تری نسبت به سایر روش‌ها برخوردار هستند.

الگوریتم‌های مبتنی بر تئوری اطلاعات[۱] که بر اساس مفاهیمی چون مرتبط بودن[۲]، افزونگی[۳] و مکمل بودن[۴] بنا شده‌اند [۲۷]، دسته‌ای بزرگ از الگوریتم‌های انتخاب ویژگی فیلتر را شامل می‌شوند [۱۲]. ایده‌ی اصلی بسیاری از معیارهای خلاقانه‌ی ارائه شده در این الگوریتم‌ها، بیشینه کردن ارتباط ویژگی‌ها با برچسب کلاس و در عین حال کاهش افزونگی بین ویژگی‌ها است [۲۸]. اکثر الگوریتم‌های این دسته از مفهوم بی‌نظمی[۵]، بی‌نظمی مشروط[۶]، بهره‌ی اطلاعاتی[۷] و معیار SU[۸] برای تعریف مفاهیم مرتبط بودن، افزونگی و مکمل بودن استفاده می‌کنند. بی‌نظمی، اندازه‌ی عدم قطعیت[۹] متغیر تصادفی گسسته‌ی X را نشان داده و به صورت رابطه‌ی (۱) تعریف می‌شود [۲۹].

$$H(X) = -\sum_{x_i \in X} p(x_i) \log p(x_i) \qquad (۱)$$

بی‌نظمی مشروط[۱۰]، میزان عدم قطعیت متغیر تصادفی گسسته‌ی X را با شرط حضور متغیر Y نشان داده و به صورت رابطه‌ی (۲) تعریف می‌شود [۲۹].

$$H(X|Y) = -\sum_{y_i \in Y} p(y_i) \sum_{x_i \in X} p(x_i|y_i) \log p(x_i|y_i) \qquad (۲)$$

مفهوم بهره‌ی اطلاعاتی[۱۱] به منظور سنجش وابستگی بین دو متغیر تصادفی X و Y بر اساس مفاهیم بی‌نظمی و بی‌نظمی مشروط، توسط شانون (۲۰۰۱) ارائه شده است [۲۹]. از آن‌جا که بهره‌ی اطلاعاتی نشان دهنده‌ی مقدار اطلاعاتی است که دو متغیر X و Y با هم دارند، به آن اطلاعات متقابل[۱۲] نیز گفته شده و از طریق رابطه‌ی (۳) محاسبه می‌شود.

$$IG(X;Y) = H(X) - H(X|Y) \qquad (۳)$$

معیار SU[۱۳] [۳۰] میزان عدم قطعیت متقارن دو متغیر تصادفی گسسته‌ی X و Y را نشان داده و به صورت زیر تعریف می‌شود.

$$SU(X;Y) = 2 * \left[\frac{IG(X;Y)}{H(X) + H(Y)}\right] \qquad (۴)$$

از مفاهیم فوق در بسیاری از الگوریتم‌های مبتنی بر تئوری اطلاعات استفاده شده است. به عنوان مثال، لوئیس و هم‌کارانش الگوریتم MIM[۱۴] را ارائه کردند که از بهره‌ی اطلاعاتی یک ویژگی به عنوان میزان ارتباط آن با برچسب کلاس استفاده می‌کند (رابطه‌ی ۵) [۳۱]. در این روش، ویژگی‌ها مستقل از هم فرض شده و امتیاز هر ویژگی به طور جداگانه و مستقل از دیگر ویژگی‌ها محاسبه می‌شود. در نظر نگرفتن خاصیت افزونگی، نقطه‌ی ضعف این روش می‌باشد.

$$MIM\_Score(f_i) = IG(f_i; C) \qquad (۵)$$

برای بهبود عمل‌کرد روش MIM، باتیتی با ارائه‌ی روش MIFS[۱۵]، هر دو معیار مرتبط بودن و افزونگی ویژگی‌ها را در فاز انتخاب ویژگی دخالت داد [۳۲]. رابطه‌ی (۶) نحوه‌ی محاسبه‌ی امتیاز هر ویژگی را نشان می‌دهد.

$$MIFS\_Score(f_i) = IG(f_i; C) - \beta . \sum_{x_j \in s} I(x_i, x_j) \qquad (۶)$$

در این رابطه، جمله‌ی اول میزان ارتباط (وابستگی) ویژگی به برچسب کلاس را مشخص کرده، جمله‌ی دوم افزونگی را از طریق کاهش همبستگی بین ویژگی مربوطه با تمام ویژگی‌های تا کنون انتخاب شده در مجموعه‌ی S، کنترل می‌کند و مقدار بتا برابر با یک در نظر گرفته شده است. پنگ و هم‌کارانش الگوریتم MRMR[۱۶] را ارائه کرده و در آن به جای مقدار بتای روش MIFS، از معکوس تعداد اعضای مجموعه‌ی ویژگی‌های تا کنون انتخاب شده استفاده کردند (رابطه‌ی ۷) [۳۳].

$$MRMR\_Score(f_i) = IG(f_i; C) - \frac{1}{|S|} . \sum_{x_j \in s} I(x_i, x_j) \qquad (۷)$$

براون و هم‌کارانش نشان دادند که هر چه مجموعه‌ی ویژگی‌های انتخاب شده بزرگ‌تر باشد، تاثیر جمله‌ی دوم کاهش می‌یابد

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
۱۱ Information Gain, Mutual Information
۱۲ Mutual Information
۱۳ Symmetrical Uncertainty
۱۴ Mutual Information Maximization
۱۵ Mutual Information Feature Selection
۱۶ Minimum Redundancy Maximum Relevance

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

[۳۴]. در الگوریتم‌های CFS[۱] [۳۵] و FCBF[۲] [۳۶] نیز از معیار SU برای ارزیابی میزان ارتباط بین دو ویژگی برای کنترل افزونگی و از ارتباط بین ویژگی و کلاس‌ها برای اندازه‌گیری میزان مرتبط بودن، استفاده شده است. با توجه به مثال‌های ذکر شده، یکی از چالش‌های این الگوریتم‌ها، مواجهه با ویژگی‌های با ارزش بالا اما افزونه است. علاوه بر این بیش‌تر الگوریتم‌های ارائه شده‌ی مبتنی بر تئوری اطلاعات، روی داده‌های گسسته اعمال می‌شوند. در مورد داده‌های پیوسته (عددی) باید از برخی از الگوریتم‌های گسسته‌سازی مانند Chi_merge، MDLP و CAIM استفاده شود [۳۷-۳۹]. در مقاله‌های [۴۰-۴۳] مروری بر الگوریتم‌های مختلف گسسته‌سازی ارائه شده است. گسسته‌سازی یکی از تکنیک‌های مهم در مبحث کاهش ابعاد محسوب شده [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 شده و سپس ویژگی به تعدادی بازه تقسیم شود که هر بازه معرف یک مقدار گسسته است. تعداد و اندازه‌ی بازه‌ها روی کیفیت گسسته‌سازی موثر است [۴۳]. به طور کلی مساله‌ی یافتن بهترین گسسته‌سازی در دسته‌ی الگوریتم‌های با پیچیدگی NP-complete قرار دارد [۴۵].

در این مقاله با استفاده از مفهوم یک‌ریختی[۳] تبدیلی از فضای فعلی به فضای جدید ویژگی ارائه شده که در آن ضمن استفاده از مزایای گسسته‌سازی، از درگیر شدن با پیچیدگی آن اجتناب شده است. سپس در فضای جدید ویژگی‌ها، یک روش انتخاب ویژگی مبتنی بر تئوری اطلاعات پیشنهاد شده که از مفهوم پیوستگی به جای بی‌نظمی برای تعریف یک معیار جدید CRC[۴] استفاده کرده و برای کاهش مشکل افزونگی، ارتباط یک ویژگی با تک‌تک کلاس‌ها به طور جداگانه بررسی شده است. این راه‌کار باعث شده که ژن‌های موثر در هر کلاس به تفکیک شناسایی شده و امکان شناسایی ژن‌های مشترک نیز وجود داشته باشد.

در ادامه‌ی مقاله، در بخش۲ روش پیشنهادی در سه قسمت توضیح داده شده، در بخش ۳ نتایج به دست آمده ارائه شده و در 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. به منظور تبیین بهتر این موضوع، مثالی در جدول (۱) ارائه شده است. 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 و ویژگی‌های مرتبط با هر بیمار (f5-f1) به صورت گسسته در نظر گرفته شده است.

جدول (۱)- مثالی از محاسبه‌ی ارزش ویژگی‌ها بر اساس بی‌نظمی و بهره‌ی اطلاعات (L=low، M=mid و H=high)

	f1	f2	f3	f4	f5	C
۱	L	L	L	H	H	c1
۲	M	L	L	H	H	c1
۳	H	M	L	H	H	c1
۴	L	H	M	L	M	c2
۵	L	H	M	L	M	c2
۶	L	H	M	L	M	c2
۷	M	H	M	L	L	c2
۸	M	H	M	L	L	c2
۹	M	H	M	H	L	c2
۱۰	H	H	M	H	L	c2
۱۱	H	H	M	H	L	c2
۱۲	L	L	H	L	L	c3
۱۳	L	L	H	L	L	c3
۱۴	M	M	H	M	M	c3
۱۵	M	M	H	M	M	c3
۱۶	H	M	H	M	M	c3
	۰/۰۰	۰/۳۱	۰/۴۴	۰/۲۲	۰/۲۲	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 f_1 تا f_5 اختصاص داده شده است. با بررسی اجمالی ویژگی‌ها مشاهده می‌شود ویژگی f_1 که از هیچ نظمی در هیچ‌یک از کلاس‌ها تبعیت نمی‌کند، پایین‌ترین امتیاز را دارد. ویژگی f_3 که یک ویژگی ایده‌آل بوده، به تنهایی هر سه کلاس را توصیف کرده و بالاترین امتیاز را گرفته است. امتیاز بالای بعدی مربوط به ویژگی f_2 بوده که یک توصیف کننده‌ی خوب برای کلاس C_2 است. مقدار ویژگی f_2 در تمام نمونه‌های کلاس C_2 برابر با H است. بنابراین الگوریتم MIM در مورد ویژگی‌های مذکور به خوبی امتیازدهی کرده است اما نحوه‌ی امتیازدهی آن در مورد ویژگی‌های f_4 و f_5 منطقی نیست. در حالی که ویژگی f_5 یک توصیف کننده‌ی خوب برای کلاس C_1 است، امتیازی برابر با

[۱] Correlation Feature Selection
[۲] Fast Correlation-Based Filter
[۳] Homomorphism
[۴] Continuity Based Relevancy Criterion

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

ویژگی f_4 دارد. با وجود این‌که تمام مقادیر ویژگی f_4 در کلاس C_1 برابر با H است اما 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. ریشه‌ی این اشکال به نحوه‌ی گسسته‌سازی ارتباط دارد. اشکال دیگر این است که با مقایسه‌ی امتیازات ویژگی‌های f_2 و f_5 مشخص می‌شود با وجود این‌که f_2 و f_5 هرکدام توصیف کننده‌ی یک کلاس متمایز هستند اما امتیازهای متفاوتی دارند. امتیاز ویژگی توصیف کننده‌ی کلاس بزرگ‌تر، بالاتر است. بنابراین در رتبه‌بندی ویژگی‌ها، ابتدا تمام ویژگی‌های موثر مرتبط با کلاس بزرگ‌تر قرار گرفته و به ترتیب نوبت به ویژگی‌های موثر مرتبط با کلاس‌های کوچک‌تر می‌رسد. این خاصیت باعث تشدید پدیده‌ی افزونگی شده و اگر رویکرد مناسبی برای مقابله با ویژگی‌های افزونه انتخاب نگردد، موجب می‌شود که داده‌های نامتوازن روی کارایی الگوریتم تاثیر مخربی بگذارند. برای کاهش این مشکلات، در این بخش یک روش انتخاب ویژگی مبتنی بر تئوری اطلاعات به نام CRFS[1] برای انتخاب ژن‌های ریزآرایه پیشنهاد شده است که شامل سه قسمت تبدیل فضای ویژگی، امتیازدهی ژن‌ها در هر کلاس و انتخاب ژن‌ها بر اساس امتیاز به دست آمده در هر کلاس می‌باشد.

۲-۱- تبدیل فضای ویژگی‌ها از پیوسته به گسسته

ابتدا متغیرها و نمادهای به کار رفته در مقاله شرح داده می‌شود.

$$\text{ClassLabel} = C_{n*1} = [C_j], |C| = m, f_i \in R^{n*1}, 1 \le i \le d$$

داده‌های ورودی شامل n نمونه و هر نمونه شامل d ویژگی بوده که به یکی از کلاس‌های موجود در ستون کلاس‌ها مربوط شده و تعداد کل کلاس‌ها برابر با m است. منظور از f_i ویژگی i-ام از مجموعه‌ی ویژگی‌ها و منظور از C_j کلاس j-ام است.

برای تبدیل ویژگی پیوسته‌ی f_i به ویژگی گسسته‌ی f'_i از زوج مرتب (f_i,C) استفاده شده است. ابتدا ویژگی f_i به ترتیب صعودی مرتب شده، سپس تمام جابه‌جایی‌های صورت گرفته برای مرتب شدن ویژگی f_i، روی مولفه‌ی دوم (ستون برچسب کلاس) اعمال می‌شود تا f'_i به دست آید. به عبارت دیگر برای به دست آوردن f'_i کافی است تا تمام تغییراتی که برای مرتب‌سازی f_i انجام شده روی ستون برچسب کلاس نیز اعمال شود. به بیان ریاضی، بین زوج مرتب (f_i,C) و $(\text{sorted}(f_i), f'_i)$ رابطه‌ی یک‌ریختی وجود دارد. یک مثال فرضی از داده‌های عددی ویژگی f_i مربوط به ۱۵ نمونه از ۳ کلاس مختلف بیماری (به ترتیب ۷، ۴ و ۴ نمونه در هر کلاس) و ویژگی گسسته‌ی f'_i متناظر با آن در جدول (۲) ارائه شده است.

جدول (۲)- رابطه‌ی یک‌ریختی بین $(f_i,\text{Class Label})$ و $(\text{sorted}(f_i), f'_i)$

f_i	۱/۵	۳/۴	۴/۱	۳/۸	۲/۲	۴/۴	۲/۲	۱/۴	۲/۲	۲/۸	۴/۹	۴/۷	۴/۵	۲/۵	۳/۳
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)	۱/۴	۱/۵	۲/۲	۲/۲	۲/۲	۲/۵	۲/۸	۳/۳	۳/۴	۳/۸	۴/۱	۴/۴	۴/۴	۴/۷	۴/۹
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

۲-۲- امتیازدهی ژن‌ها در هر کلاس

در جدول (۲) چگونگی افراز[2] هر یک از کلاس‌های بیماری توسط f'_i با توجه به رابطه‌ی پیوستگی (هم‌جواری مقداری) قابل مشاهده است. منظور از افراز یک مجموعه، تقسیم آن به تعدادی زیرمجموعه است که اشتراک آن‌ها تهی و اجتماع آن‌ها برابر با مجموعه‌ی اولیه باشد [۴۶]. برای مثال مجموعه‌ی کلاس C_1 به دو زیربخش ۴ و ۳ عضوی افراز شده است. در این مقاله طبق رابطه‌ی (۸) از چگونگی افراز هر یک از کلاس‌ها توسط f'_i به عنوان معیار سنجش امتیاز ژن i-ام استفاده شده است.

$$\text{Score}(f_i|C_j) = \sum_{k=1}^{partition(C_j|f'_i)} P(A_k)\,(\exp(P(A_k)) \qquad (۸)$$
$$P(A_k) = \frac{|A_k|}{|C_j|}$$

در این رابطه، f_i ویژگی مربوط به بیان ژن i-ام، C_j کلاس j-ام سرطان، $\text{Score}(f_i|C_j)$ میزان اهمیت و امتیاز ویژگی i-ام در کلاس j-ام، $\text{partition}(C_j|f'_i)$ تعداد افراز (زیربخش) کلاس C_j توسط ویژگی f'_i، A_k زیربخش k-ام از زیربخش‌های افراز شده‌ی کلاس C_j، $|A_k|$ تعداد نمونه‌های زیربخش k-ام دارای پیوستگی خالص و $|C_j|$ تعداد نمونه‌های کلاس j-ام می‌باشد. در ادامه به تعریف پیوستگی ظاهری و پیوستگی خالص پرداخته شده است.

تعریف ۱: پیوستگی ظاهری

دنباله‌ای از برچسب‌های 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.

[1] Continuity Based Relevancy Criterion Feature Selection

[2] Partition

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

تعریف ۲: پیوستگی خالص

بر اساس هر دو مولفه‌ی زوج مرتب (sorted (f_i), f'_i)، دنباله‌ای از برچسب‌های یک کلاس در f'_i است که مقادیر متناظر آن‌ها در sorted (f_i) در دو طرف، با مقادیر مجاور متفاوت باشد. هر چه 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 رابطه‌ی (۱) امتیاز بیش‌تری به ویژگی f_i در آن کلاس تعلق می‌گیرد.

طبق تعاریف فوق پیوستگی ظاهری می‌تواند خالص نیز باشد اما 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 دنباله‌ی خالص ۱ و ۳ تایی است. اکنون می‌توان با جای‌گذاری مقادیر پیوستگی خالص ویژگی f_i کلاس C_1 در رابطه‌ی (۸)، ارزش ویژگی f_i را در کلاس C_1 به دست آورد.

$$\text{Score}(f_i|C_1) = \sum_{k=1}^{2} P(A_k)\,(\exp(P(A_k))$$
$$= 1/7 * \exp(1/7) + 3/7 * \exp(3/7) = 0.8227$$

به همین صورت امتیاز ویژگی در سایر کلاس‌ها نیز به دست می‌آید. به طور خلاصه مطالب گفته شده را می‌توان در قالب شبه کد ۱ که در جدول (۳) ارائه شده است نشان داد.

جدول (۳)- شبه‌کد ۱: محاسبه‌ی امتیاز ویژگی‌ها در هر کلاس

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

۲-۳- انتخاب ژن‌ها بر اساس 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 تمام امتیازهای به دست آمده در جدول وزنی W قرار داده شود. با توجه به تعداد ویژگی‌ها (d) و تعداد کلاس‌ها (m)، ابعاد جدول برابر با m×d خواهد بود. یک مثال از جدول W با فرض داشتن ۱۰ ویژگی و ۳ کلاس بیماری در جدول (۴) ارائه شده است.

جدول (۴)- مثالی از جدول وزنی ویژگی‌ها

	C_1	C_2	C_3
f_1	۹/۳	۵/۷	۶/۳
f_2	۲/۳	۰/۵	۱/۱
f_3	۱/۵	۵	۸/۹
f_4	۷/۶	۸/۹	۷/۳
f_5	۱/۲	۰/۴	۹/۱
f_6	۶/۲	۷/۴	۹/۵
f_7	۵/۳	۹/۳	۲/۲
f_8	۹/۳	۷/۸	۱/۵
f_9	۱/۹	۹/۴	۵/۱
f_{10}	۹/۳	۰/۴	۰/۲

با در اختیار داشتن جدول وزنی ویژگی‌ها و مشاهده‌ی آن، سوال‌های متنوعی درباره‌ی ارتباط ویژگی‌ها با کلاس‌ها به ذهن خطور می‌کند که پاسخ به بعضی از آن‌ها نیازمند یک پژوهش جدید و مستقل است. به عنوان مثال با مشاهده‌ی جدول (۴) و با فرض این‌که حداکثر امتیاز مرتبط بودن یک ویژگی به یک کلاس برابر با ۱۰ باشد، می‌توان دریافت که ویژگی‌های f_1، f_8 و f_{10} با امتیاز برابر هر کدام یک توصیف کننده‌ی خوب برای کلاس C_1 هستند اما با این تفاوت که f_{10} هیچ ارتباطی با دو کلاس دیگر ندارد در حالی که f_8 می‌تواند به عنوان ویژگی موثر در کلاس C_2 نیز به حساب آمده و f_1 ارتباط نسبی خوبی با دو کلاس دیگر دارد. برای پی بردن به این‌که کدام ویژگی رتبه‌ی بالاتری دارد به تحقیق بیش‌تری نیاز است. هم‌چنین جدول (۴) نشان می‌دهد در حالی که ویژگی f_2 به هیچ یک از کلاس‌ها مرتبط نیست ویژگی f_4 ارتباط خوبی با هر سه کلاس دارد. بنابراین ویژگی‌های مشترک نیز قابل بررسی هستند. می‌توان ضوابط مختلفی برای انتخاب و رتبه‌بندی ویژگی‌ها تعیین کرد. روشی که در این مقاله برای رتبه‌بندی ویژگی‌ها در نظر گرفته شده به این صورت است که ابتدا در هر کلاس ویژگی‌ها بر اساس امتیاز مرتب شده 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 (۴)، ترتیب نهایی ویژگی‌ها به صورت f_1، f_9، f_6، f_8، f_7، f_5، f_{10} f_4، f_3 و f_2 می‌باشد.

۳- نتایج

در این بخش، کارایی روش پیشنهادی با ۵ روش انتخاب ویژگی دیگر مقایسه شده است. از میان این ۵ روش، سه روش انتخاب

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

ویژگی ReliefF [۴۷]، IG [۳۱] و CFS که از روش‌های رایج در ارتباط با داده‌های ریزآرایه هستند، در نرم‌افزار یادگیری ماشین Weka [۴۸] پیاده‌سازی شده است. هم‌چنین دو روش انتخاب ویژگی CAM و DSCFS توسط ونگ و وی (۲۰۱۷) [۴۹] بر اساس اندازه‌گیری توانایی یک ویژگی در دسته‌بندی زیر مسائل دوکلاسه و مفهوم مکمل بودن ویژگی‌ها ارائه شده است. در حالی که تمرکز اصلی در روش انتخاب ویژگی CAM بر حذف ویژگی‌های غیرمرتبط 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[۱]، KNN[۲] و NB[۳] که هر سه در Weka پیاده‌سازی شده‌اند، و هفت مجموعه‌ی داده‌ی ریزآرایه مربوط به انواع سرطان استفاده شده است. هر مجموعه‌ی داده دارای داده‌های آموزش و آزمایش مجزا است که جزئیات مربوط به آن‌ها در جدول (۵) ارائه 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 مجموعه‌ی داده انجام شده است. سپس دقت دسته‌بندی ویژگی‌های انتخاب شده با توجه به داده‌های آزمایش، توسط هر یک از دسته‌بندها به دست آمده است. دقت دسته‌بندی بیان‌گر درصد نمونه‌های درست دسته‌بندی شده در میان تمام نمونه‌ها است. نتایج به دست آمده در جدول‌های (۶)، (۷) و (۸) ارائه شده است. مقادیر پررنگ‌تر در جدول‌ها، معرف بالاترین دقت به دست آمده در هر مجموعه‌ی داده است. به ازای هر مجموعه‌ی داده، دو سطر در جدول‌ها وجود دارد که یک سطر مربوط به دقت و سطر دیگر مربوط به تعداد ویژگی‌هایی از هر الگوریتم است که در دسته‌بندی مشارکت داشته‌اند.

جدول (۵)- مجموعه‌ی داده‌ها

مجموعه‌ی	تعداد	تعداد	تعداد نمونه		مرجع
			آموزش	آزمایش	
Leukemia1	۳	۷۱۲۹	۳۸	۳۴	[۶]
Lung1	۳	۷۱۲۹	۶۴	۳۲	[۵۰]
Leukemia2	۳	۱۲۵۸۲	۵۷	۱۵	[۵۱]
Breast	۵	۹۲۱۶	۵۴	۳۰	[۵۲]
DLBCL	۶	۴۰۲۶	۵۸	۳۰	[۵۳]
Cancers	۱۱	۱۲۵۳۳	۱۰۰	۷۴	[۵۴]
GCM	۱۴	۱۶۰۶۳	۱۴۴	۴۶	[۵۵]

جدول (۶)- مقایسه‌ی دقت دسته‌بندی ۶ الگوریتم با به کارگیری دسته‌بند NB

	OURS	مقاله‌ی ونگ							
مجموعه‌ی داده	CRFS	CAM	DSCFS	CFS	IG	ReliefF	I-CAM	R-CAM	I-CFS	R-CFS
Leukemia1	**۹۷/۰۶**	۸۵/۲۹	**۹۷/۰۶**	**۹۷/۰۶**	۹۱/۱۸	**۹۷/۰۶**	**۹۷/۰۶**	**۹۷/۰۶**	**۹۷/۰۶**	۹۴/۱۲
	۴	۲۰	۲	۷۶	۲	۲	۲۰	۲۰	۷۶	۷۶
Lung1	**۸۷/۵۰**	۸۴/۳۸	۸۱/۲۵	۷۸/۱۳	۷۸/۱۳	۷۱/۸۸	۸۴/۳۸	۸۴/۳۸	۸۱/۲۵	۸۱/۲۵
	۴	۵۹۸	۳	۱۵۱	۳	۳	۵۹۸	۵۹۸	۱۵۱	۱۵۱
Leukemia2	**۱۰۰**	۹۳/۳۳	**۱۰۰**	**۱۰۰**	۷۳/۳۳	**۱۰۰**	**۱۰۰**	**۱۰۰**	**۱۰۰**	**۱۰۰**
	۴	۴۲	۳	۱۷۳	۳	۳	۴۲	۴۲	۱۷۳	۱۷۳
Breast	۸۳/۳۳	۸۰	**۹۰**	۸۳/۳۳	۷۰	۸۳/۳۳	۸۰	۸۳/۳۳	۸۰	۸۰
	۳۰	۱۵۲	۶	۱۸۰	۶	۶	۱۵۲	۱۵۲	۱۸۰	۱۸۰
DLBCL	**۹۶/۶۷**	۹۰	۹۳/۳۳	۹۰	۹۰	۹۳/۳۳	۷۰	۹۳/۳۳	**۹۶/۶۷**	۹۳/۳۳
	۱۸	۲۰۹۲	۱۸۱	۲۵۷	۱۸۱	۱۸۱	۲۰۹۲	۲۰۹۲	۲۵۷	۲۵۷
Cancers	**۹۳/۲۴**	۷۸/۳۸	۷۹/۷۳	۸۵/۱۴	۸۱/۰۸	۸۱/۰۸	۸۶/۴۹	۸۳/۷۸	۸۵/۱۴	۸۶/۴۹
	۵۲	۱۹۲۳	۶۴۸	۳۵۶	۶۴۸	۶۴۸	۱۹۲۳۲	۱۹۲۳	۳۵۶	۳۵۶
GCM	**۶۷/۳۹**	۵۰	۵۶/۵۲	۵۶/۵۲	۵۰	۴۳/۴۸	۵۰	۴۵/۶۵	۳۹/۱۳	۳۹/۱۳
	۳۸	۴۶۷۴	۱۷۳۹	۱۲۸	۱۷۳۹	۱۷۳۹	۴۶۷۴	۴۶۷۴	۱۲۸	۱۲۸

[۱] Support Vectormachine
[۲] K-Nearest Neighbor
[۳] Naive Bayes

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

جدول (۷)- مقایسه‌ی دقت دسته‌بندی ۶ الگوریتم با به کارگیری دسته‌بند SVM

	OURS	مقاله‌ی ونگ								
مجموعه‌ی داده	CRFS	CAM	DSCFS	CFS	IG	ReliefF	I-CAM	R-CAM	I-CFS	R-CFS
Leukemia1	**۹۷/۰۶**	**۹۷/۰۶**	۸۵/۲۹	۸۸/۲۴	۷۹/۴۱	۸۵/۲۹	۹۴/۱۲	۹۴/۱۲	۹۴/۱۲	۹۴/۱۲
	۴	۲۰	۲	۷۶	۲	۲	۲۰	۲۰	۷۶	۷۶
Lung1	۸۱/۲۵	**۸۷/۵**	۷۵	۸۴/۳۸	۷۸/۱۳	۸۱/۲۵	۸۴/۳۸	۸۴/۳۸	۸۴/۳۸	۸۴/۳۸
	۶	۵۹۸	۳	۱۵۱	۳	۳	۵۹۸	۵۹۸	۱۵۱	۱۵۱
Leukemia2	**۱۰۰**	**۱۰۰**	**۱۰۰**	**۱۰۰**	۷۳/۳۳	۹۳/۳۳	**۱۰۰**	**۱۰۰**	**۱۰۰**	**۱۰۰**
	۸	۴۲	۳	۱۷۳	۳	۳	۴۲	۴۲	۱۷۳	۱۷۳
Breast	**۹۶/۶۷**	**۹۶/۶۷**	**۹۶/۶۷**	۸۶/۶۷	۷۶/۶۷	۶۳/۳۳	**۹۶/۶۷**	۹۳/۳۳	**۹۶/۶۷**	۹۳/۳۳
	۳۰	۱۵۲	۶	۱۸۰	۶	۶	۱۵۲	۱۵۲	۱۸۰	۱۸۰
DLBCL	**۱۰۰**	**۱۰۰**	**۱۰۰**	**۱۰۰**	**۱۰۰**	۹۶/۶۷	۵۳/۳۳	**۱۰۰**	**۱۰۰**	**۱۰۰**
	۲۰	۲۰۹۲	۱۸۱	۲۵۷	۱۸۱	۱۸۱	۲۰۹۲	۲۰۹۲	۲۵۷	۲۵۷
Cancers	۸۲/۴۳	۸۹/۱۹	**۹۵/۹۵**	۹۳/۲۴	**۹۵/۹۵**	۹۳/۲۴	۹۳/۲۴	۹۴/۵۹	۹۳/۲۴	۹۴/۵۹
	۶۶	۱۹۲۳	۶۴۸	۳۵۶	۶۴۸	۶۴۸	۱۹۲۳	۱۹۲۳	۳۵۶	۳۵۶
GCM	۵۲/۱۷	۵۲/۱۷	۴۷/۸۳	**۵۸/۷**	۵۰	۴۱/۳۰	۵۴/۳۵	۴۷/۸۳	۴۱/۳۰	۴۳/۴۸
	۳۸	۴۶۷۴	۱۷۳۹	۱۲۸	۱۷۳۹	۱۷۳۹	۴۶۷۴	۴۶۷۴	۱۲۸	۱۲۸
	برد/باخت	۲/۵	۳/۴	۳/۴	۱/۶	۲/۵	۳/۴	۲/۵	۲/۵	۲/۵

جدول (۸)- مقایسه‌ی دقت دسته‌بندی ۶ الگوریتم با به کارگیری دسته‌بند KNN

	OURS	مقاله‌ی ونگ								
مجموعه‌ی داده	CRFS	CAM	DSCFS	CFS	IG	ReliefF	I-CAM	R-CAM	I-CFS	R-CFS
Leukemia1	**۹۷/۰۶**	۷۶/۴۷	**۹۷/۰۶**	۸۵/۲۹	۸۵/۲۹	**۹۷/۰۶**	۹۴/۱۲	۹۴/۱۲	**۹۷/۰۶**	۹۴/۱۲
	۴	۲۰	۲	۷۶	۲	۲	۲۰	۲۰	۷۶	۷۶
Lung1	**۸۷/۵**	۸۱/۲۵	۷۵	۶۸/۷۵	۸۴/۳۸	۷۱/۸۸	۸۴/۳۸	۸۱/۲۵	۸۱/۲۵	۸۱/۲۵
	۱۱	۵۹۸	۳	۱۵۱	۳	۳	۵۹۸	۵۹۸	۱۵۱	۱۵۱
Leukemia2	**۱۰۰**	۹۳/۳۳	۹۳/۳۳	۹۳/۳۳	۶۶/۶۷	۹۳/۳۳	۹۳/۳۳	۹۳/۳۳	۹۳/۳۳	۹۳/۳۳
	۶	۴۲	۳	۱۷۳	۳	۳	۴۲	۴۲	۱۷۳	۱۷۳
Breast	۹۳/۳۳	۹۰	۹۶/۶۷	۹۰	۶۶/۶۷	۸۳/۳۳	۹۶/۶۷	**۱۰۰**	۹۶/۶۷	**۱۰۰**
	۳۰	۱۵۲	۶	۱۸۰	۶	۶	۱۵۲	۱۵۲	۱۸۰	۱۸۰
DLBCL	**۱۰۰**	۹۶/۶۷	۹۶/۶۷	۹۶/۶۷	۹۶/۶۷	۹۶/۶۷	۷۰	**۱۰۰**	۹۶/۶۷	۹۶/۶۷
	۱۸	۲۰۹۲	۱۸۱	۲۵۷	۱۸۱	۱۸۱	۲۰۹۲	۲۰۹۲	۲۵۷	۲۵۷
Cancers	۸۵/۱۴	۸۷/۸۴	۸۹/۱۹	**۹۰/۵۴**	۸۷/۸۴	۸۷/۸۴	۸۷/۸۴	۸۹/۱۹	**۹۰/۵۴**	۸۹/۱۹
	۴۶	۱۹۲۳	۶۴۸	۳۵۶	۶۴۸	۶۴۸	۱۹۲۳	۱۹۲۳	۳۵۶	۳۵۶
GCM	**۶۳/۰۴**	۵۲/۱۷	۴۷/۸۳	۵۲/۱۷	۴۵/۶۵	۴۳/۴۸	۴۷/۸۳	۴۵/۶۵	۵۲/۱۷	۴۱/۳۰
	۳۸	۴۶۷۴	۱۷۳۹	۱۲۸	۱۷۳۹	۱۷۳۹	۴۶۷۴	۴۶۷۴	۱۲۸	۱۲۸

با توجه به این‌که برای ارزیابی روش‌های ReliefF و IG لازم است تا تعداد ویژگی‌های انتخابی تعیین شود، ونگ در مرجع [۴۹] از تعداد ویژگی‌های هر یک از روش‌های DSCFS، CAM و CFS به طور مشابه برای به دست آوردن دقت روش‌های IG و ReliefF استفاده کرده است. بنابراین دقت روش‌های IG و ReliefF در سه حالت به دست آمده است. تعداد ویژگی‌ها در حالت اول (IG و ReliefF) برابر با تعداد ویژگی‌های روش DSCFS، در حالت دوم (I-cam و R-cam) برابر با تعداد ویژگی‌ها در روش CAM و در حالت سوم (I-CFS و R-CFS) برابر با تعداد ویژگی‌های روش CFS در نظر گرفته شده است. در روش پیشنهادی، برای محدود کردن انتخاب تعداد ویژگی‌ها از حد m×۸ (m تعداد کلاس‌های هر مجموعه‌ی داده) به عنوان

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

حداکثر ژن‌های انتخابی استفاده شده است. به عبارت دیگر حداکثر ژن‌های انتخابی از هر کلاس بیش‌تر از ۸ نخواهد بود. دقت‌های به دست آمده برای الگوریتم پیشنهادی در هر یک از جدول‌ها، حاصل مقایسه در این فضای محدود است. در جدول (۶) که از دسته‌بند NB برای ارزیابی الگوریتم‌ها استفاده شده است، برتری محسوس روش CRFS مشهود است. روش CRFS در اکثر مجموعه‌ی داده‌ها به بالاترین دقت دست یافته است. اگر چه در برخی موارد مانند Leukemia1، Leukemia2 و DLBC به طور مشترک برخی دیگر از الگوریتم‌ها نیز به بالاترین دقت رسیده‌اند، اما CRFS در سه مورد Lung1، Cancers و GCM به تنهایی بالاترین دقت را کسب کرده است. نکته‌ی مهم دیگری که در مورد DLBCL، Cancers و GCM باید به آن اشاره کرد، تعداد بسیار پایین ژن‌های انتخابی CRFS در مقایسه با سایر روش‌ها بوده به طوری که تنها روش CFS تا حدودی با نتایج روش پیشنهادی قابل مقایسه است. به خصوص در مورد مجموعه‌ی داده‌ی GCM که دارای ۱۴ کلاس مختلف است، انتخاب تنها ۳۸ ژن که حتی به میانگین ۳ ژن در هر کلاس نیز نمی‌رسد، نشان دهنده‌ی این است که روش CRFS در تشخیص ژن‌های موثر عمل‌کرد قابل قبولی دارد.

اطلاعات جدول (۷) نشان می‌دهد که کارایی CRFS زمانی که از دسته‌بند SVM استفاده شود به خوبی حالت استفاده از NB نیست. بنابراین در این وضعیت برای مقایسه‌ی بهتر بین روش ارائه شده و سایر روش‌ها از سطر آخر جدول (۷) استفاده شده است. مقایسه بر اساس دو پارامتر دقت دسته‌بندی و تعداد ژن‌های انتخابی انجام شده است به این صورت که پارامتر دقت، اولویت بالاتری داشته و در صورت دقت مساوی، الگوریتم با تعداد ژن‌های کم‌تر، بهتر است. در سطر آخر جدول (۷)، نتیجه‌ی مقایسه بین روش CRFS با سایر روش‌ها روی هفت مجموعه‌ی داده به صورت برد/باخت جمع‌آوری شده است که حاکی از برتری نسبی روش CRFS می‌باشد.

در جدول (۸) از دسته‌بند KNN برای ارزیابی الگوریتم‌ها استفاده شده است. در این جدول نیز مانند جدول (۶) برتری محسوس CRFS مشهود بوده و نیازی به تشکیل جدول برد/باخت وجود ندارد. روش CRFS در پنج مجموعه‌ی داده به بالاترین دقت رسیده که در سه مورد Leukemia2، Lung1 و GCM به تنهایی بالاترین دقت را کسب کرده و در مورد DLBCL نیز با توجه به تعداد ژن انتخابی عمل‌کرد بهتری داشته اما در مورد Leukemia1 با وجود دست‌یابی به بالاترین دقت، عمل‌کرد DSCFS با توجه به تعداد ژن انتخابی بهتر است.

۴- نتیجه‌گیری

در این مقاله یک روش انتخاب ویژگی به نام CRFS برای انتخاب ژن‌های موثر در تشخیص انواع سرطان از داده‌های ریزآرایه پیشنهاد شده که در آن از مفهومی به نام پیوستگی خالص برای تعریف معیار مرتبط بودن استفاده شده است. در روش پیشنهادی برای کاهش افزونگی، امتیاز ژن‌ها در هر کلاس به طور جداگانه محاسبه شده و انتخاب نهایی ژن‌ها بر اساس انتخاب ژن‌های بهتر از هر کلاس، انجام شده است. نتایج به دست آمده بیان‌گر کارایی روش پیشنهادی با در نظر گرفتن هم‌زمان دو معیار دقت دسته‌بندی و تعداد ژن انتخابی می‌باشد. با توجه به این‌که مفاهیم مرتبط بودن، افزونگی و مکمل بودن در روش‌های انتخاب ویژگی مبتنی بر تئوری اطلاعات نقش مهمی دارند، توجه هم‌زمان به هر سه مفهوم می‌تواند به بهبود عمل‌کرد روش ارائه شده به خصوص در مواردی که تعداد کلاس‌ها افزایش می‌یابد، کمک کند. نتایج به دست آمده در این مقاله نشان می‌دهد که با وجود برتری نسبی روش پیشنهادی نسبت به سایر روش‌ها، در مواردی مانند مجموعه‌ی داده‌های Cancers و GCM که تعداد کلاس‌ها افزایش یافته (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 مرتبط بودن، به مفاهیم افزونگی و مکمل بودن 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-

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

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. 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-Maroño, 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. Almugren 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 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.

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

[37] R. Kerber, "Chimerge: Discretization of numeric attributes," Proc. tenth Natl. Conf. Artif. Intell., 1992.
[38] K. Irani and U. Fayyad, "Multi-lnterval 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. Data Eng., 2004.
[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 diagnosis using tumor gene expression signatures," Proc. Natl. Acad. Sci., vol. 98, no. 26, pp. 15149–15154, 2001.

Copyright © 2020 by ISBME, http://www.ijbme.org - All rights reserved
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License