نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران
دانشکده مهندسی کامپیوتر، دانشگاه شهید بهشتی، تهران، ۱۵-۱۳ اسفندماه ۱۳۹۲

دسته‌بندی مجموعه داده‌های ریزآرایه براساس تکنیک‌های ترکیبی به‌منظور تشخیص سرطان

محمد مروت پودنک[۱]، علی‌رضا عصاره[۲]، بیتا شادگار[۳]

[۱] دانشجوی کارشناسی ارشد هوش مصنوعی، گروه مهندسی کامپیوتر، دانشگاه شهید چمران اهواز،
Morovvat.mmp@gmail.com

[۲] دانشیار گروه مهندسی کامپیوتر، دانشگاه شهید چمران اهواز،
Alireza.Osareh@scu.ac.ir

[۳] استادیار گروه مهندسی کامپیوتر، دانشگاه شهید چمران اهواز،
Bita.Shadgar@scu.ac.ir

چکیده

در سال‌های اخیر، فناوری ریزآرایه امکان مانیتورینگ بیان هزاران ژن را به‌طور همزمان فراهم آورده است. تحلیل‌هایی که در زمینه داده‌های ریزآرایه صورت گرفته است، بیانگر قدرت این فناوری در زمینه تشخیص بسیاری بیماری‌ها از جمله سرطان است. چالشی که در این زمینه مطرح است، تعداد بالای ویژگی‌ها (ژن‌ها) و از طرفی تعداد پایین نمونه‌ها است. تا به امروز تلاش‌های متعددی در زمینه انتخاب ژن و سپس دسته‌بندی داده‌ها صورت گرفته است که نتایج بدست آمده، بیانگر برتری تکنیک‌های ترکیبی در مقابل تکنیک‌های منفرد هست. لذا در این پژوهش، پس از ارائه روشی کارآمد در زمینه انتخاب ژن، از تکنیک‌های ترکیبی معروف آدابوست، بگینگ و دگینگ جهت کلاس‌بندی داده‌ها کمک گرفته شده است.

به‌علاوه در بخش بعدی این پژوهش، ادغام چندین تکنیک و در نهایت رأی‌گیری اکثریت با هدف بهبود نتایج صورت گرفته است. نتایج بدست آمده، بیانگر کارا بودن روش پیشنهادی در مقایسه با الگوریتم‌های پایه و همچنین هر یک از تکنیک‌های ترکیبی به‌صورت منفرد بوده است.

کلمات کلیدی

فناوری ریزآرایه، کلاس‌بندی داده‌های سرطان، تکنیک‌های ترکیبی، آدابوست، بگینگ، دگینگ.

۱- مقدمه

با کشف فناوری ریزآرایه، به یکباره حجم عظیمی از داده‌های مربوط به ژن‌های انسان در اختیار محققین قرار گرفت. در واقع داده‌های سطوح بیان ژن، اطلاعات ارزشمندی در مورد شبکه بیولوژیک، حالت سلولی و فهم چگونگی ژن‌ها در بردارد که کاربرد عملی آن، تشخیص برخی بیماری‌ها از جمله سرطان هست.

چالشی که در این زمینه موجود است، ابعاد بالای فضای ویژگی و از سوی دیگر تعداد پایین نمونه‌ها است. لذا دو مرحله‌ی انتخاب ژن و سپس دسته‌بندی داده‌ها، دو مرحله‌ی حیاتی به جهت پیش‌بینی احتمال سرطان در یک فرد است. تفسیر داده‌های بیان ژن که پیچیده و نویزی هستند، امری حیاتی است. لذا در سال‌های اخیر، تحقیقات فراوانی در هر دو زمینه صورت گرفته است.

تحقیقات اخیر در زمینه فناوری ریزآرایه نشان‌دهنده این است که دسته‌بند کننده‌های منفرد چندان کارا نیستند و از طرفی سیستم‌هایی با دسته‌بند کننده‌های چندگانه[۱] یا ترکیبی[۲] دارای دقت و پایداری بالاتری نسبت به دسته‌بند کننده‌های منفرد هستند [1].

واقعیت امر این است که علی‌رغم تمامی تلاش‌های صورت گرفته بر روی داده‌های ریزآرایه به جهت تشخیص سرطان، کماکان جای خلاقیت و نوآوری به جهت افزایش دقت نتایج با استفاده از تکنیک‌های رایج یادگیری ماشین است. لذا در این پژوهش، از یک روش دو مرحله‌ای به جهت انتخاب ژن با کمک فیلترهای کارای SU[۳] (عدم قطعیت متقارن) و CFS[۴] کمک گرفته شده است و سپس با