Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

Research Article

A Classifier Based on K-Nearest Neighbors Using Weighted Summation of Reconstruction Errors

Rassoul Hajizadeh [a,*], Mohammad Ali Hosseinzadeh [b]

[a] *Machine Learning and Deep Learning Research Laboratory, Faculty of Engineering Modern Technologies, Amol University of Special Modern Technologies, Amol, Iran*
[b] *Faculty of Engineering Modern Technologies, Amol University of Special Modern Technologies, Amol, Iran*

PAPER INFO

Paper history:

Received: 15 April 2023
Revised: 31 August 2023
Accepted: 04 September 2023

Keywords:

Classifier,
Recognition rate,
K-nearest neighbors,
Linear reconstruction,
Weighted combination.

ABSTRACT

In this paper, a classifier is introduced based on the nearest neighbor classifier and the reconstruction error for data classification. In the proposed method, first, K nearest data points (neighbors) from each category in the training data are calculated for the test data point. Then, the reconstruction of the test data is performed based on different numbers of nearest neighbors (from one to K) in each category, and the reconstruction error is calculated separately for each number of neighbors. In the next step, for each category, the error is calculated as the weighted sum of the errors obtained from all reconstructions. The weight of the reconstruction error is proportional to the number of neighbors involved in it, so the reconstruction error is multiplied by the number of neighbors. Finally, the test data belongs to the category with the lowest overall error. This process allows a combination of K nearest neighbor classifiers to play a role in data classification. In this paper, 10 datasets from the UCR time series database and five datasets from the UCI classification database are used to evaluate the proposed method. The results of these evaluations show that the proposed method significantly improves the performance of the minimum reconstruction error based KNN classifiers, achieving approximately 5% better recognition rate for some K values and an average recognition rate improvement of about 1.6% for all K values (from 2 to 15).

DOI: https://doi.org/10.22075/jme.2023.30380.2437

*** Corresponding author.**
E-mail address: r.hajizadeh@ausmt.ac.ir

How to cite this article:
Hajizadeh, R., & Hosseinzadeh, M. A. (2024). A classifier based on K-nearest neighbors using weighted summation of reconstruction errors. Journal of Modeling in Engineering, 22(76), 55-68. doi: 10.22075/jme.2023.30380.2437

مقاله پژوهشی

طبقه‌بند مبتنی بر K نزدیکترین همسایه‌ها با استفاده از جمع وزن‌دار خطاهای بازسازی

رسول حاجی‌زاده[۱،*]، محمدعلی حسین‌زاده[۲]

اطلاعات مقاله

دریافت مقاله: ۱۴۰۲/۰۱/۲۶
بازنگری مقاله: ۱۴۰۲/۰۶/۰۹
پذیرش مقاله: ۱۴۰۲/۰۶/۱۳

واژگان کلیدی:
طبقه‌بند،
نرخ بازشناسی،
K نزدیکترین همسایه،
بازسازی خطی،
جمع وزن‌دار.

چکیده

در این مقاله، طبقه‌بندی مبتنی بر طبقه‌بند K نزدیکترین همسایه‌ها و خطای بازسازی، جهت دسته‌بندی داده‌ها معرفی شده است. در روش پیشنهادی، ابتدا K نزدیکترین داده (همسایه) به داده‌ی آزمون، از هر دسته موجود در داده‌های آموزش، محاسبه می‌گردد. سپس به بازسازی داده‌ی آزمون، بر حسب تعداد مختلفی از نزدیکترین همسایه‌ها) از یک تا (K ، در هر دسته پرداخته شده و میزان خطای بازسازی به ازای هر تعداد همسایه به طور مجزا محاسبه می‌گردد. در گام بعد، در هر دسته، میزان خطا به صورت جمع وزندار خطای حاصل از تمامی بازسازی‌ها محاسبه می‌گردد. وزن خطای بازسازی، متناسب با تعداد همسایه‌های دخیل در آن در نظر گرفته شده است بدین ترتیب که خطای بازسازی در تعداد همسایه‌های آن ضرب می‌شود. در آخر، داده‌ی آزمون به دسته‌ای تعلق دارد که کمترین میزان خطای کل را دارا است. این عمل موجب می‌گردد تا ترکیبی از طبقه‌بندهای مبتنی بر K نزدیکترین همسایه به صورت هم‌افزایی در طبقه‌بندی داده‌ها نقش ایفا نمایند. در این مقاله از ۱۰ دسته‌مجموعه متعلق به پایگاه داده‌ی سری-زمانی UCRو پنج دسته-مجموعه متعلق به پایگاه داده‌ی دسته‌بندی UCI جهت ارزیابی روش پیشنهادی استفاده شده است. نتایج بدست آمده از این ارزیابی‌ها نشان می‌دهد که روش پیشنهادی، عملکرد طبقه ندهای KNN مبتنی بر کمترین خطای بازسازی را به میزان زیادی بهبود بخشیده و نرخ بازشناسی در برخی K ها را در حدود ۵ درصد بهتر نموده و متوسط نرخ بازشناسی به ازای تمامیK ها (از ۲ الی ۱۵) در حدود ۱.۶ درصد بهبود یافته است.

DOI: https://doi.org/10.22075/jme.2023.30380.2437

۱-مقدمه[۱]

امروزه با پیشرفت روزافزون ابزار جمع‌آوری و اکتساب داده ها، همواره حجم بالایی از داده‌ها در اختیار بوده که نیازمند پردازش و ارزیابی هستند. روش‌های سنتی تحلیل داده‌ها، به ویژه داده‌هایی که از حجم و ابعاد بالایی برخوردارند، زمان‌بر بوده و از کارایی پایینی برخوردار هستند. یکی از روش‌های نوین تحلیل و ارزیابی این دسته از داده‌ها، بهره‌گیری از ماشین و هوش مصنوعی است.

آموزش ماشین و هوش مصنوعی، جزئی جدایی‌ناپذیر در دنیای امروز و آینده است. امروزه، کاربردهای متعددی از حضور ماشین و هوش مصنوعی در حوزه‌های گوناگون از جمله تحلیل تصاویر و داده‌های پزشکی، تحلیل داده‌های بزرگ، طبقه‌بندی داده‌ها و خوشه‌بندی داده‌ها را می‌توان نام برد [۱-۵]. آموزش ماشین و هوش مصنوعی از گام‌های متعددی برخوردارند. پیش‌پردازش داده‌ها و استخراج ویژگی، گام هایی متداول بوده و همواره در تحلیل داده‌ها

* پست الکترونیک نویسنده مسئول: r.hajizadeh@ausmt.ac.ir
۱. استادیار، آزمایشگاه آموزش ماشین و یادگیری عمیق، دانشکده مهندسی فناوری‌های نوین، دانشگاه تخصصی فناوری‌های نوین آمل، آمل، ایران
۲. استادیار، دانشکده مهندسی فناوری‌های نوین، دانشگاه تخصصی فناوری‌های نوین آمل، آمل، ایران

استناد به این مقاله:
حاجی زاده, رسول و حسین زاده, محمد علی. (۱۴۰۳). طبقه بند مبتنی بر K نزدیکترین همسایه ها با استفاده از جمع وزن دار خطاهای بازسازی. *مدل سازی در مهندسی*, ۲۲(۷۶), ۵۵-۶۸، doi: 10.22075/jme.2023.30380.2437 .

نقش ایفا می‌نمایند [۶-۸]. البته، امروزه با ظهور شبکه‌های عمیق، بخش مهندسی ویژگی نیز به ماشین سپرده شده و نتایج بسیار چشم‌گیری فراهم نموده است [۹-۱۱]. یکی از بخش‌های متداول در آموزش ماشین، *طبقه‌بند* است که در بسیاری از کاربردها، حضوری پررنگ دارد. وظیفه طبقه بندها، دسته‌بندی و تعیین دسته مربوط به داده‌ی آزمون با استفاده از داده‌های آموزش است. طبقه‌بندها با استفاده از دانش و ویژگی‌هایی که از داده‌های آموزش و آزمون استخراج شده است به دسته‌بندی اطلاعات و داده‌ها می پردازند [۱۲-۱۴]. طبقه‌بندهای متعددی تاکنون معرفی شده‌اند از جمله، طبقه‌بند مبتنی بر K نزدیکترین همسایه ها (KNN) [۱۵]، طبقه‌بند ماشین بردار پشتیبان (SVM) [۱۶]، طبقه‌بند جنگل تصادفی [۱۷] و طبقه بندهای مبتنی بر شبکه عصبی [۱۸].

طبقه‌بند مبتنی بر نزدیکترین همسایه‌ها، به دلیل نحوه پیاده‌سازی راحت و همچنین کارایی بسیار بالا، از متداول ترین و پرکاربردترین روش‌های طبقه‌بندی داده‌ها است. اولین روش مبتنی بر KNN در سال ۱۹۶۷ توسط کور[۲] و هارت[۳] معرفی شد [۱۵]. طبقه‌بند KNN متداول، با رای‌گیری حداکثری بر روی دسته‌ی همسایه‌های داده آزمون به طبقه‌بندی داده‌ی آزمون می‌پردازد. روش‌های مبتنی بر KNN متعددی در راستای بهبود عملکرد این روش معرفی شده است از جمله: طبقه‌بند KNN مبتنی بر بازنمایی وزن‌دار[۴] (WRKNN) [۱۹]، طبقه‌بند KNN مبتنی بر جمع وزن‌دار میانگین محلی[۵] (WLMRKNN) [۱۹]، طبقه‌بند مبتنی بر شبه نزدیکترین همسایه[۶] (PNN) [۲۰]، طبقه‌بند مبتنی بر میانگین محلی شبه نزدیکترین همسایه[۷] (LMPNN) [۲۱]،طبقه‌بند KNN مبتنی بر میانگین محلی[۸] (LMKNN) [۲۲]، طبقه‌بند نزدیکترین همسایه‌های مبتنی بر میانگین‌های چند-محلی[۹] (MLMNN) [۲۳]، طبقه‌بند نزدیکترین همسایه مبتنی بر بازنمایی مشارکتی[۱۰] (CRNN) [۲۴]، طبقه‌بند مبتنی بر K نزدیکترین همسایه‌ی هارمونیک[۱۱] (KHNN) [۲۵]، طبقه‌بند KNN مبتنی بر فاصله وزن‌دار[۱۲] (WKNN) [۲۶] و طبقه‌بند KNN مبتنی بر همسایگی توام و رای-گیری حداکثری اصلاح شده(MNMKNN)[13] [۲۷]. در همه این طبقه‌بندها، از K نزدیکترین همسایه مربوط به داده آزمون، جهت دسته‌بندی داده آزمون بهره گرفته می‌شود.

طبقه‌بندهای CRNN و WKNN از معیاری مشابه طبقه بند KNN متداول بهره برده و مبتنی بر رای‌گیری حداکثری به طبقه‌بندی داده آزمون می‌پردازند. در طبقه‌بند CRNN، به بازنمایی داده‌ی آزمون بر حسب تمامی داده های آزمون پرداخته و K داده‌ی آموزش متناظر با K بزرگترین ضریب بازنمایی، به عنوان همسایه‌های داده‌ی آزمون در نظر گرفته می‌شود [۲۴]. در WKNN، به همسایه‌ها وزنی متناسب با فاصله آنها از داده آزمون اتلاق می‌گردد تا در طبقه‌بندی به کار گرفته شود [۲۶].

طبقه‌بندهای PNN، LMPNN، KHNN و LMKNN با استفاده از K نزدیکترین همسایه‌ی هر دسته، داده‌ای جدید به عنوان نماینده داده‌های همسایه‌ی هر دسته ایجاد می نمایند. این داده‌ی جدید، ترکیبی وزن‌دار از داده‌های همسایه است که در روش‌های PNN، LMPNN، KHNN و LMKNN متفاوت است [۲۰-۲۲، ۲۵]. در انتها، داده‌ی آزمون در دسته‌ای طبقه‌بندی می‌گردد که نماینده‌ی آن دسته از داده‌ی آزمون کمترین فاصله را دارا است.

از طرف دیگر، در طبقه‌بندهای MLMNN، WRKNN و WLMRKNN معیار تصمیم‌گیری، کمینه خطای بازسازی داده‌ی آزمون بر حسب همسایه‌های آن از هر دسته است. در این طبقه‌بندها نیز، ابتدا K نزدیکترین همسایه به داده آزمون در هر دسته تعیین می‌گردد. سپس به بازسازی خطی داده‌ی آزمون بر حسب داده‌های همسایه پرداخته می‌شود. تابع هزینه خطای بازسازی و در نتیجه نحوه تعیین ضرایب بازسازی در طبقه‌بندهای مذکور متفاوت بوده و منجر به تفاوت در عملکرد آنها می‌شود [۱۹، ۲۳]. در انتها، داده‌ی آزمون در دسته‌ای طبقه‌بندی می‌گردد که از کمترین خطای بازسازی برخوردار است. در [۱۹] نشان داده شده است که روش‌های WRKNN و WLMRKNN از بهترین عملکرد در میان روش‌های مبتنی بر KNN

[2] Cover
[3] Hart
[4] Weighted Representation-based KNN
[5] Weighted Local Mean Representation-based KNN
[6] Pseudo Nearest Neighbor
[7] Local Mean based PNN
[8] Local Mean based KNN
[9] Multi-Local Means based Nearest Neighbors
[10] Collaborative Representation-based Nearest Neighbor
[11] K-Harmonic Nearest Neighbor
[12] Distance-Weighted K-Nearest Neighbor
[13] Mutual Neighborhood and Modified Majority Voting based KNN

برخوردار هستند.

طبقه‌بند MNMKNN نیز روشی است که بر روی نحوه‌ی انتخاب داده‌ها به عنوان همسایه تمرکز داشته و در آن نحوه توزیع داده‌ها در انتخاب همسایه‌ها نقشی مستقیم دارد. در MNMKNN، یک داده (فرض، داده شماره یک) زمانی به عنوان همسایه‌ی داده دیگر (فرض، داده شماره دو) انتخاب می‌گردد که داده شماره دو، خود نیز به عنوان همسایه داده‌ی اول انتخاب گردد. به عبارت دیگر، داده‌هایی که به عنوان همسایه انتخاب می‌گردند باید در همسایگی یکدیگر قرار گیرند. منظور از همسایگی، K نزدیکترین داده‌ها به هر داده است. در [۲۷] نشان داده شده است که عملکرد طبقه‌بندهای مبتنی بر KNN با استفاده از مفهوم همسایگی توام بهبود یافته و موجب افزایش نرخ بازشناسی گردیده است. همچنین در [۲۷] روشی برای بهبود عملکرد طبقه‌بندهای مبتنی بر رای اکثریت نیز معرفی شده که باز هم موجب بهبود عملکرد طبقه‌بندها گردیده است. در [۲۷]، هنگامیکه دو دسته یا بیشتر با رای اکثریت وجود داشته باشند، داده به دسته‌ای از بین آنها متعلق است که کمترین فاصله را از میانگین همسایه‌های آنها برخوردار است.

در تمامی دسته‌بندهای مبتنی بر K نزدیکترین همسایه نام‌برده شده، تصمیم‌گیری‌ها تنها مبتنی بر K همسایه صورت می‌گیرد، در حالیکه با در اختیار داشتن K همسایه، می‌توان به ازای حالتهای با تعداد همسایه‌های کمتر از K همسایه نیز به ارزیابی و استخراج اطلاعات پرداخت. در روش پیشنهادی، به دنبال هم‌افزایی اطلاعات موجود به ازای تعداد همسایه‌های کوچکتر مساوی K است تا بتواند عملکرد طبقه‌بندهای KNN مبتنی بر کمترین خطای بازسازی را بهبود بخشد. در نتیجه می‌توان با هم‌افزایی معیارهای بدست آمده توسط نزدیکترین همسایه الی K نزدیکترین همسایه به طبقه‌بندی داده‌های آزمون پرداخته و دسته مربوطه را تعیین نمود. در این مقاله، مبتنی بر روش پیشنهادی، سه طبقه‌بند جدید که تعمیم طبقه‌های مبتنی بر MLMNN، WRKNN و WLMRKNN است معرفی گردیده و عملکرد آنها بر روی پایگاه داده‌های متداول حوزه دسته‌بندی مورد ارزیابی قرار گرفته است.

در بخش‌های بعدی، ابتدا به بیان طبقه‌بندهای مبتنی بر KNN مرتبط پرداخته خواهد شد. سپس، در بخش سوم، روش پیشنهادی معرفی می‌گردد. در بخش چهارم نتایج شبیه‌سازی و تحلیل آنها بیان شده و در آخر نیز، به جمع‌بندی و نتیجه‌گیری پرداخته شده است.

۲- طبقه‌بندهای مبتنی بر K نزدیکترین همسایه (روش‌های مرتبط)

فرض کنید $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n]\in\Re^{d\times n}$، بیانگر ماتریس داده‌های آموزش است که شامل n داده آموزش d بعدی از m دسته‌ی مختلف ($C=\{c_1,c_2,\dots,c_m\}$) است و $\boldsymbol{X}^j\in\Re^{d\times n_j}$ نشان‌دهنده n_j داده‌ی آموزش از دسته‌ی j-ام است ($\sum_{j=1}^{m}n_j=n$). همچنین $\boldsymbol{y}$ بیانگر داده آزمون است که متعلق به یکی از دسته‌های c_1 تا c_m است. در ادامه طبقه‌بندهای KNN مبتنی بر کمترین خطای بازسازی WRKNN، WLMRKNN و MLMNN معرفی می‌گردند.

۲-۱- طبقه‌بند WRKNN

در این طبقه‌بند، ابتدا به ازای هر دسته، K نزدیکترین داده‌ی آموزش ($\boldsymbol{X}^j_{KNN}(\boldsymbol{y})=[\boldsymbol{x}^j_{1N},\boldsymbol{x}^j_{2N},\dots,\boldsymbol{x}^j_{KN}],\ j=1,2,\dots,m$) به داده‌ی آزمون ($\boldsymbol{y}$)، مبتنی بر فاصله اقلیدسی تعیین و به عنوان K نزدیکترین همسایه‌های داده آزمون در نظر گرفته می‌شود. سپس، داده آزمون با استفاده از رابطه (۱)، به صورت خطی بر حسب همسایه‌های هر دسته بازسازی می‌گردد تا کمترین میزان خطای بازسازی حاصل گردد:

$$\begin{aligned}&min\left\|\boldsymbol{y}-\sum_{i=1}^{K}\eta_1^j\boldsymbol{x}^j_{iN}\right\|_2\\&=min\left\|\boldsymbol{y}-\left(\boldsymbol{X}^j_{KNN}(\boldsymbol{y})\right)\boldsymbol{\eta}^j\right\|_2,\end{aligned}\qquad(۱)$$

که در آن، $\boldsymbol{\eta}^j=\left[\eta_1^j,\eta_2^j,\dots,\eta_K^j\right]^T$ بردار ضرایب بازسازی خطی بر حسب داده‌های همسایه‌ی دسته j-ام است. در روش WRKNN، $\boldsymbol{\eta}^j$ در هردسته با استفاده از رابطه زیر و به ازای قید اعمال شده محاسبه می‌گردد [۱۹]:

$$\begin{aligned}\boldsymbol{\eta}^{j*}=\underset{\boldsymbol{\eta}^j}{arg\,min}\Big\{&\left\|\boldsymbol{y}-\left(\boldsymbol{X}^j_{KNN}(\boldsymbol{y})\right)\boldsymbol{\eta}^j\right\|_2^2\\&+\alpha\left\|\boldsymbol{D}^j\boldsymbol{\eta}^j\right\|_2^2\Big\}\end{aligned}\qquad(۲)$$

که α ضریب تنظیم و $\boldsymbol{D}^j$ یک ماتریس قطری است که عناصر روی قطر اصلی آن بیانگر فاصله اقلیدسی بین داده‌ی آزمون و همسایه‌ها است. ماتریس $\boldsymbol{D}^j$ در رابطه (۳) نشان

داده شده است.

$$\boldsymbol{D}^{j} = \begin{bmatrix} \left\|\boldsymbol{y}-\boldsymbol{x}_{1N}^{j}\right\|_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left\|\boldsymbol{y}-\boldsymbol{x}_{KN}^{j}\right\|_2 \end{bmatrix} \quad (۳)$$

رابطه (۲) نشان می‌دهد که در طبقه‌بند WRKNN، علاوه بر دست‌یابی به حداقل میزان خطای بازسازی، میزان فاصله همسایه‌ها از داده‌ی آزمون نیز مهم بوده و در محاسبه ضرایب بازسازی موثر است. در اینجا، ضریب تنظیم α میزان نقش‌آفرینی خطای بازسازی و فاصله را تعیین می‌نماید. رابطه (۲)، دارای جواب بهینه به صورت زیر است:

$$\boldsymbol{\eta}^{j^*} = \left(\left(\boldsymbol{X}_{KNN}^{j}(\boldsymbol{y})\right)^T \boldsymbol{X}_{KNN}^{j}(\boldsymbol{y}) + \alpha\left(\boldsymbol{D}^{j}\right)^T\boldsymbol{D}^{j}\right)^{-1}\left(\boldsymbol{X}_{KNN}^{j}(\boldsymbol{y})\right)^T\boldsymbol{y}. \quad (۴)$$

که $\boldsymbol{\eta}^{j^*}$ ضرایب بازسازی بهینه به ازای دسته j-ام است. در گام بعدی، میزان خطای بازسازی، با استفاده از رابطه (۵) در هر دسته محاسبه شده و داده‌ی آزمون، متعلق به دسته‌ای است که دارای کمترین میزان خطای بازسازی است [۱۹].

$$r_{WRKNN}^{j}(\boldsymbol{y}) = \left\|\boldsymbol{y} - \left(\boldsymbol{X}_{KNN}^{j}(\boldsymbol{y})\right)\boldsymbol{\eta}^{j^*}\right\|_2^2. \quad (۵)$$

۲-۲- طبقه‌بند WLMRKNN

طبقه‌بند WLMRKNN، مشابه طبقه‌بند WRKNN، مبتنی بر کمترین میزان خطای بازسازی عمل نموده و در بیشتر گام‌ها کاملا مشابه طبقه‌بند WRKNN است. در طبقه‌بند WLMRKNN، پس از تعیین K نزدیکترین همسایه‌های داده‌ی آزمون به ازای هر دسته، با استفاده از رابطه (۶) به محاسبه K همسایه جدید، مبتنی بر میانگین محلی همسایه‌ها می‌پردازد [۱۹].

$$\bar{\boldsymbol{x}}_{iN}^{j} = \frac{1}{i}\sum_{l=1}^{i}\boldsymbol{x}_{lN}^{j}, \qquad i = 1, \dots, K. \quad (۶)$$

در ادامه، داده آزمون بر حسب همسایه‌های جدید ، به صورت خطی بازسازی می‌گردد.

$$\boldsymbol{min}\left\|\boldsymbol{y} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{K}}\boldsymbol{\beta}_{\boldsymbol{1}}^{\boldsymbol{j}}\bar{\boldsymbol{x}}_{\boldsymbol{iN}}^{\boldsymbol{j}}\right\|_{\boldsymbol{2}} = \boldsymbol{min}\left\|\boldsymbol{y} - \left(\bar{\boldsymbol{X}}_{\boldsymbol{KNN}}^{\boldsymbol{j}}(\boldsymbol{y})\right)\boldsymbol{\beta}^{\boldsymbol{j}}\right\|_{\boldsymbol{2}}, \quad (۷)$$

که $\boldsymbol{\beta}^{j} = \left[\beta_1^j, \beta_2^j, \dots, \beta_K^j\right]^T$ بردار ضرایب بازسازی و $\bar{\boldsymbol{X}}_{KNN}^{j} = \left[\bar{\boldsymbol{x}}_{1N}^{j}, \bar{\boldsymbol{x}}_{2N}^{j}, \dots, \bar{\boldsymbol{x}}_{KN}^{j}\right]$ است. گام‌های بعدی کاملا مشابه WRKNN است با این تفاوت که تنها داده‌های جدید $\bar{\boldsymbol{X}}_{KNN}^{j}$ به عنوان همسایه‌های داده‌ی آزمون مورد استفاده قرار می‌گیرند. البته باید توجه داشت که ماتریس قطری فاصله در WLMRKNN بر حسب $\bar{\boldsymbol{x}}_{iN}^{j}$ ، $(i = 1, \dots, K)$، محاسبه گردیده و در روابط به کار گرفته می‌شود. در انتها نیز، داده‌ی آزمون متعلق به دسته‌ای است که از کمترین خطای بازسازی برخوردار است.

۲-۳- طبقه‌بند MLMNN

در این روش نیز، ابتدا K نزدیکترین داده از هر دسته به داده آزمون تعیین می‌گردد. MLMNN مشابه طبقه‌بند WLMRKNN به محاسبه همسایه‌های میانگین محلی با استفاده از رابطه (۶) پرداخته و مشابه رابطه (۷) به بازسازی خطی داده‌ی آزمون بر حسب همسایه‌ها در هر دسته می پردازد. اما در MLMNN، ضرایب بازسازی از رابطه زیر حاصل می‌گردد که در آن با اعمال قیدی بر روی اندازه بردار ضرایب، به کنترل ضرایب بازسازی می‌پردازد:

$$\boldsymbol{\theta}^{j^*} = \underset{\boldsymbol{\theta}^{j}}{arg\,min}\left\{\left\|\boldsymbol{y} - \left(\bar{\boldsymbol{X}}_{KNN}^{j}(\boldsymbol{y})\right)\boldsymbol{\theta}^{j}\right\|_2^2 + \gamma\left\|\boldsymbol{\theta}^{j}\right\|_2^2\right\} \quad (۸)$$

که، γ ضریب تنظیم بوده و $\boldsymbol{\theta}^{j} = \left[\theta_1^j, \theta_2^j, \dots, \theta_K^j\right]^T$ بردار ضرایب بازسازی بر حسب همسایه‌های میانگین محلی دسته j-ام است. مقدار بهینه ضرایب بازسازی به صورت فرم بسته قابل محاسبه است که در رابطه (۹) بیان شده است:

$$\boldsymbol{\theta}^{j^*} = \left(\left(\bar{\boldsymbol{X}}_{KNN}^{j}(\boldsymbol{y})\right)^T\bar{\boldsymbol{X}}_{KNN}^{j}(\boldsymbol{y}) + \gamma\boldsymbol{I}\right)^{-1}\left(\bar{\boldsymbol{X}}_{KNN}^{j}(\boldsymbol{y})\right)^T\boldsymbol{y}. \quad (۹)$$

که در آن، $\boldsymbol{I}$ ماتریس همانی است. پس از محاسبه ضرایب بازسازی به ازای هر دسته، میزان خطای بازسازی در هر دسته با استفاده از رابطه (۱۰) محاسبه می‌گردد.

$$r_{MLMNN}^{j}(\boldsymbol{y}) = \left\|\boldsymbol{y} - \left(\bar{\boldsymbol{X}}_{KNN}^{j}(\boldsymbol{y})\right)\boldsymbol{\theta}^{j^*}\right\|_2^2. \quad (۱۰)$$

در انتها، داده آزمون متعلق به دسته‌ای است که از کمترین میزان خطای بازسازی برخوردار است.

۳- روش پیشنهادی: طبقه‌بند مبتنی بر جمع وزن‌دار خطاهای بازسازی با استفاده از K نزدیکترین همسایه

در این مقاله، روشی مبتنی بر کمینه خطای بازسازی پیشنهاد می‌گردد که به طبقه‌بندی داده‌ها به صورت هم افزایی خطاهای بازسازی می‌پردازد. در روش پیشنهادی، ابتدا K نزدیکترین داده به داده‌ی آزمون، به ازای هر دسته تعیین می‌گردد. سپس با استفاده از همسایه‌های مربوط به هر دسته، خطاهای بازسازی بر حسب نزدیکترین همسایه، دو نزدیکترین همسایه و الی K نزدیکترین همسایه‌ها محاسبه گردیده و در دسته‌بندی داده‌ی آزمون مشارکت می‌نمایند. روش پیشنهادی در طبقه‌بندهای مبتنی بر کمترین بازسازی خطی به کار گرفته شده است.

نوآوری اصلی روش پیشنهادی، در هم‌افزایی و بکارگیری طبقه‌بندهای KNN مبتنی بر کمترین خطای بازسازی در فضاهای بازسازی مختلف است، که به ازای تعداد همسایه های متفاوت حاصل می‌گردد. در طبقه‌بندهای KNN مبتنی بر کمینه خطای بازسازی، داده‌ی آزمون بر حسب داده‌های هر دسته به صورت مجزا بازسازی می‌گردد و داده‌ی آزمون متعلق به دسته‌ای است که از کمترین خطای بازسازی برخوردار است. ایده بکارگرفته شده در این مقاله نیز بدین شرح است که داده‌ی آزمون، عموما، بر حسب داده‌های هم‌دسته و به ازای هر زیرفضایی که با استفاده از داده‌های هم‌دسته ایجاد می‌گردد، بایستی بهتر بازسازی گردد. بنابراین، استفاده از هم‌افزایی مطرح شده در روش پیشنهادی می‌تواند کمک بیشتری به تمیز دادن داده‌ها بپردازد. از دیگر نوآوری‌های مقاله می‌توان به نرمال نمودن خطاهای بازسازی در زیرفضاهای بازسازی متفاوت اشاره داشت که در آن خطای بازسازی بر حسب تعداد داده‌های دخیل در بازسازی، نرمالیزه شده است.

در ادامه به بیان روش پیشنهادی مبتنی بر طبقه‌بند WRKNN پرداخته خواهد شد که نام طبقه‌بند پیشنهادی را WS-WRKNN[14] قرار می‌دهیم. البته باید اشاره داشت که روش پیشنهادی می‌تواند برای بهبود طبقه‌بندهای WLMRKNN و MLMNN نیز به کار گرفته شود که در انتهای این بخش الگوریتم آنها نیز بیان شده است.

در روش پیشنهادی، پس از تعیین همسایه‌های داده‌ی آزمون، رابطه (۱۱) بیانگر بازسازی خطی داده آزمون بر حسب همسایه‌های آن است:

$$\begin{aligned} &min \left\| \boldsymbol{y} - \sum_{i=1}^{z} \eta_1^j \boldsymbol{x}_{iN}^j \right\|_2 \qquad (۱۱) \\ &= min \left\| \boldsymbol{y} - \left(\boldsymbol{X}_{zNN}^j(\boldsymbol{y}) \right) \boldsymbol{\eta}_z^j \right\|_2, \end{aligned}$$

که در آن z تعداد همسایه‌های دخیل در بازسازی خطی داده‌ها و $\boldsymbol{x}_{iN}^j$ نشان‌دهنده‌ی i-امین همسایه (از نظر فاصله اقلیدسی) داده‌ی آزمون، از دسته‌ی j-ام است. همچنین، $\boldsymbol{X}_{zNN}^j(\boldsymbol{y})$ بیان‌گر ماتریس حاصل از z نزدیکترین همسایه بوده و $\boldsymbol{\eta}_z^j$ ضرایب بازسازی متناظر با همسایه‌ها است. ضرایب بازسازی خطی مشابه روابط ۲ الی ۴، به صورت زیر قابل محاسبه است:

$$\begin{aligned} \boldsymbol{\eta}_z^{j^*} = &\left(\left(\boldsymbol{X}_{zNN}^j(\boldsymbol{y}) \right)^T \boldsymbol{X}_{zNN}^j(\boldsymbol{y}) + \right. \qquad (۱۲) \\ &\left. \alpha \left(\boldsymbol{D}_z^j \right)^T \boldsymbol{D}_z^j \right)^{-1} \left(\boldsymbol{X}_{zNN}^j(\boldsymbol{y}) \right)^T \boldsymbol{y}, \end{aligned}$$

که در آن، $\boldsymbol{D}_z^j$ به صورت زیر تعریف شده است:

$$\begin{aligned} &\boldsymbol{D}_z^j \qquad (۱۳) \\ &= \begin{bmatrix} \left\| \boldsymbol{y} - \boldsymbol{x}_{1N}^j \right\|_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left\| \boldsymbol{y} - \boldsymbol{x}_{zN}^j \right\|_2 \end{bmatrix} \end{aligned}$$

پس از تعیین مقادیر بهینه ضرایب بازسازی، خطای بازسازی با استفاده از z همسایه دخیل در آن به صورت زیر قابل بیان است:

$$r_z^j(\boldsymbol{y}) = \left\| \boldsymbol{y} - \left(\boldsymbol{X}_{zNN}^j(\boldsymbol{y}) \right) \boldsymbol{\eta}_z^{j^*} \right\|_2^2. \qquad (۱۴)$$

در روش پیشنهادی میزان خطای بازسازی خطی داده آزمون به ازای $z = 1, \dots, K$ محاسبه گردیده و خطای بازسازی کل در هر دسته، به صورت زیر محاسبه می‌گردد:

$$\begin{aligned} &r_{WS-WRKNN}^j(\boldsymbol{y}) = \qquad (۱۵) \\ &\frac{r_1^j(\boldsymbol{y}) + 2r_2^j(\boldsymbol{y}) + \cdots + Kr_K^j(\boldsymbol{y})}{K} \end{aligned}$$

که در آن، $r_{WS-WRKNN}^j(\boldsymbol{y})$ بیانگر میزان خطای بازسازی کل دسته j-ام است. در رابطه (۱۵) میزان خطا به ازای هر تعداد همسایه دخیل در بازسازی، در تعداد همسایه‌های دخیل در آن ضرب می‌گردد تا خطاها به نوعی نرمالیزه گردیده و سپس با یکدیگر جمع می‌شوند. در انتها، داده آزمون به دسته‌ای $(c_{\boldsymbol{y}})$ تعلق دارد که از کمترین میزان

[14] Weighted Summation WRKNN

خطای بازسازی کل برخوردار است:

$$c_{\boldsymbol{y}} = \underset{j}{arg\,min}\{r^{j}_{WS-WRKNN}(\boldsymbol{y})\}, \quad \forall j = 1, \dots, m. \qquad (۱۶)$$

به همین ترتیب، روش پیشنهادی می‌تواند بر اساس طبقه-بندهای MLMNN و WLMRKNN بیان گردد که به ترتیب روش‌های پیشنهادی WS-MLMNN و WS-WLMRKNN نامگذاری شده‌اند. گام‌های طبقه‌بندهای پیشنهادی WS-WRKNN، WS-WLMRKNN و WS-MLMNN، به ترتیب در الگوریتم‌های ۱ الی ۳ بیان شده است.

الگوریتم ۱: طبقه‌بند WS-WRKNN

ورودی: داده‌های آموزش ($\boldsymbol{X}$)، ضریب تنظیم (α)، تعداد همسایه‌ها (K) و داده آزمون ($\boldsymbol{y}$)

خروجی: تعیین دسته داده آزمون

۱- تعیین K همسایه داده آزمون به ازای داده‌های هر دسته ($\boldsymbol{X}^{j}_{KNN}(\boldsymbol{y})$)،
۲- K مرتبه تکرار مراحل ۳ الی ۵ به ازای تعداد همسایه‌های دخیل در بازسازی از ۱ الی K،
۳- تشکیل ماتریس قطری فاصله همسایه‌ها از داده آزمون به ازای هر دسته ($\boldsymbol{D}^{j}_{z}$)،
۴- محاسبه ضرایب بازسازی خطی ($\boldsymbol{\eta}^{j}_{z}$) به ازای هر دسته با استفاده از رابطه (۱۲)،
۵- محاسبه خطای بازسازی با استفاده از رابطه (۱۴) به ازای هر دسته،
۶- محاسبه خطای بازسازی کل به صورت جمع وزن‌دار خطاهای بازسازی به ازای هر دسته با استفاده از رابطه (۱۵)،
۷- داده آزمون به دسته‌ای تعلق دارد که از کمترین میزان خطای بازسازی کل برخوردار است.

الگوریتم ۲: طبقه‌بند WS-WLMRKNN

ورودی: داده‌های آموزش ($\boldsymbol{X}$)، ضریب تنظیم (α)، تعداد همسایه‌ها (K) و داده آزمون ($\boldsymbol{y}$)

خروجی: تعیین دسته داده آزمون

۱- تعیین K همسایه داده آزمون به ازای داده‌های هر دسته ($\boldsymbol{X}^{j}_{KNN}(\boldsymbol{y})$)،
۲- K مرتبه تکرار مراحل ۳ الی ۶ به ازای تعداد همسایه‌های دخیل در بازسازی از ۱ الی K،
۳- محاسبه همسایه‌های محلی ($\bar{\boldsymbol{X}}^{j}_{zNN}$) با استفاده از رابطه (۶) به ازای هر دسته،
۴- تشکیل ماتریس قطری فاصله همسایه‌های محلی از داده آزمون به ازای هر دسته ($\bar{\boldsymbol{D}}^{j}_{z}$)،
۵- محاسبه ضرایب بازسازی به ازای هر دسته: $\boldsymbol{\beta}^{j^*}_{z} = \left(\left(\bar{\boldsymbol{X}}^{j}_{zNN}(\boldsymbol{y})\right)^{T}\bar{\boldsymbol{X}}^{j}_{zNN}(\boldsymbol{y}) + \alpha\left(\bar{\boldsymbol{D}}^{j}_{z}\right)^{T}\bar{\boldsymbol{D}}^{j}_{z}\right)^{-1}\left(\bar{\boldsymbol{X}}^{j}_{zNN}(\boldsymbol{y})\right)^{T}\boldsymbol{y}$،
۶- محاسبه خطای بازسازی به ازای هر دسته: $r^{j}_{z}(\boldsymbol{y}) = \left\|\boldsymbol{y} - \left(\bar{\boldsymbol{X}}^{j}_{zNN}(\boldsymbol{y})\right)\boldsymbol{\beta}^{j^*}_{z}\right\|^{2}_{2}$،
۷- محاسبه خطای بازسازی کل به صورت جمع وزن‌دار خطاهای بازسازی به ازای هر دسته با استفاده از رابطه (۱۵)،
۸- داده آزمون به دسته‌ای تعلق دارد که از کمترین میزان خطای بازسازی کل برخوردار است.

الگوریتم ۳: طبقه‌بند WS-MLMNN

ورودی: داده‌های آموزش ($\boldsymbol{X}$)، ضریب تنظیم (γ)، تعداد همسایه‌ها (K) و داده آزمون ($\boldsymbol{y}$)

خروجی: تعیین دسته داده آزمون

۱- تعیین K همسایه داده آزمون به ازای داده‌های هر دسته ($\boldsymbol{X}^{j}_{KNN}(\boldsymbol{y})$)،
۲- K مرتبه تکرار مراحل ۳ الی ۵ به ازای تعداد همسایه‌های دخیل در بازسازی از ۱ الی K،
۳- محاسبه همسایه‌های محلی ($\bar{\boldsymbol{X}}^{j}_{zNN}$) با استفاده از رابطه (۶) به ازای هر دسته،
۴- محاسبه ضرایب بازسازی به ازای هر دسته: $\boldsymbol{\beta}^{j^*}_{z} = \left(\left(\bar{\boldsymbol{X}}^{j}_{zNN}(\boldsymbol{y})\right)^{T}\bar{\boldsymbol{X}}^{j}_{zNN}(\boldsymbol{y}) + \gamma\boldsymbol{I}\right)^{-1}\left(\bar{\boldsymbol{X}}^{j}_{zNN}(\boldsymbol{y})\right)^{T}\boldsymbol{y}$،
۵- محاسبه خطای بازسازی به ازای هر دسته: $r^{j}_{z}(\boldsymbol{y}) = \left\|\boldsymbol{y} - \left(\bar{\boldsymbol{X}}^{j}_{zNN}(\boldsymbol{y})\right)\boldsymbol{\beta}^{j^*}_{z}\right\|^{2}_{2}$،
۶- محاسبه خطای بازسازی کل به صورت جمع وزن‌دار خطاهای بازسازی به ازای هر دسته با استفاده از رابطه (۱۵)،
۷- داده آزمون به دسته‌ای تعلق دارد که از کمترین میزان خطای بازسازی کل برخوردار است.

۱-۳- آنالیز فضای برداری

در این بخش، نشان داده شده است که میزان خطای بازسازی با افزایش تعداد همسایه‌ها در بازسازی خطی کوچکتر مساوی میزان خطای بازسازی با تعداد همسایه‌های کمتر است. با فرض اینکه $r^{j}_{z}(.)$ نشان‌دهنده میزان خطای بازسازی خطی هر داده به ازای z داده همسایه آن از دسته

j است، می‌توان به بیان گزاره زیر پرداخت.

گزاره. اگر a و b دو عدد صحیح مثبت و y داده آزمون باشد بطوریکه $a \geq b$، آن گاه داریم: $r_a^j(\boldsymbol{y}) \geq r_b^j(\boldsymbol{y})$

اثبات. فرض کنید $\boldsymbol{\eta}_a^{j^*} = [\eta_1^j, \eta_2^j, \dots, \eta_a^j]$ است. در اینصورت $\left[\eta_1^j, \eta_2^j, \dots, \eta_a^j, \underbrace{0, \dots, 0}_{b-a}\right]$ یکی از بردارهایی است که در پیدا کردن مقدار $\boldsymbol{\eta}_b^{j^*} = arg\ \underset{\boldsymbol{\eta}_b^j}{min}\left\{\left\|\boldsymbol{y} - \left(\boldsymbol{X}_{KNN}^j(\boldsymbol{y})\right)\boldsymbol{\eta}^j\right\|_2^2 + \alpha\left\|\boldsymbol{D}^j\boldsymbol{\eta}^j\right\|_2^2\right\}$، مورد بررسی قرار می گیرد و به ازای آن، مقدار $r_a^j(\boldsymbol{y})$ حاصل می شود. لذا اثبات گزاره کامل می شود. ■

در روش پیشنهادی، انگیزه استفاده از تمامی K خطای $r_1^j(\boldsymbol{y})$، ...، $r_K^j(\boldsymbol{y})$ این است که در روش های پیشین، زمانیکه به محاسبه $r_K^j(\boldsymbol{y})$ می پردازند، تمامی $K\text{-}1$ خطای $r_1^j(\boldsymbol{y})$، ...، $r_{K-1}^j(\boldsymbol{y})$ نیز در دسترس هستند. از طرفی طبق تعریف، $r_i^j(\boldsymbol{y})$ بیان کننده میزان خطا در یک فضای i بعدی (به شرط ناهمبسته بودن داده‌های همسایه) است که طبق گزاره قبل، با افزایش تعداد همسایه‌های دخیل در بازسازی داده آزمون و در نتیجه با افزایش احتمالی بعد فضای برداری تولید شده ، میزان خطای بازسازی به صورت کاهشی خواهد بود. لذا در محاسبه خطای کلی از رابطه زیر، که همان رابطه (۱۵) است، استفاده شده است:

$$r_{WS-WRKNN}^j(\boldsymbol{y}) = \frac{r_1^j(\boldsymbol{y})}{K} + \frac{2r_2^j(\boldsymbol{y})}{K} + \cdots + \frac{Kr_K^j(\boldsymbol{y})}{K}$$

که با قرار دادن ضریب بزرگتر برای میزان خطایی که از تعداد داده های آموزش بیشتری استفاده می کند، تاثیر آن را بر روی میزان خطای بازسازی کل، کمتر می‌نماید.

۴- نتایج شبیه‌سازی

در این بخش، عملکرد روش پیشنهادی و طبقه‌بندهای پیشنهادی مبتنی بر آن، با بهترین طبقه‌بندهای KNN مبتنی بر کمینه بازسازی خطی شامل MLMNN، WRKNN و WKMRKNN مورد مقایسه و ارزیابی قرار خواهد گرفت. در ارزیابی‌های صورت گرفته، از پایگاه داده‌های UCI Machine Learning Repository [۲۸] و UCR Time Series Classification Archives [۲۹] استفاده شده است. در ادامه، به معرفی پایگاه داده‌های بکارگرفته شده، نحوه تشکیل داده‌های آموزش و آزمون، نحوه ارزیابی و بیان نتایج پرداخته شده است.

۴-۱- پایگاه داده UCI

این پایگاه داده شامل ۶۲۲ مجموعه دسته‌داده‌های استاندارد است که در کاربردهای دسته‌بندی، خوشه‌بندی[15] و رگرسیون مورد استفاده قرار می‌گیرد. تعداد نمونه‌ها، ابعاد داده‌ها و تعداد دسته‌ها در مجموعه دسته‌داده‌های UCI از تنوع بالایی برخوردار است [۲۸]. در این مقاله از پنج دسته-داده با کاربرد طبقه‌بندی بهره گرفته شده است که مشخصات آنها در جدول ۱ بیان شده است.

جدول ۱- مجموعه دسته‌داده‌های UCI بکارگرفته شده و مشخصات آنها [۲۸].

عنوان دسته داده	تعداد نمونه‌ها	بُعد داده‌ها	تعداد دسته
Parkinsons	۱۹۵	۲۲	۲
Climate	۳۶۰	۱۸	۲
Vowel	۵۲۸	۱۰	۱۱
Wifi Localization	۲۰۰۰	۷	۴
Planning Relax	۱۸۲	۱۳	۲

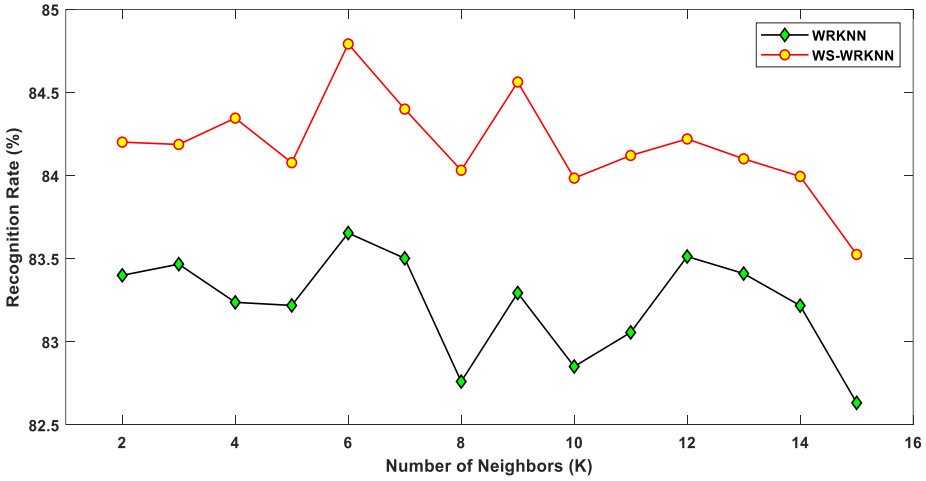
جهت ارزیابی عملکرد طبقه‌بندها، هر دسته‌داده از این پایگاه داده به دو گروه آموزش و آزمون تقسیم گردیده و مقدار دقت طبقه‌بندها (درصد نرخ بازشناسی) محاسبه می‌گردد. تعداد داده‌های آموزش و آزمون به ترتیب ۶۷ و ۳۳ درصد از کل داده‌های هر دسته‌داده را تشکیل داده و به صورت تصادفی انتخاب می‌گردند. همچنین با توجه به تعداد کم داده‌ها در دسته‌داده‌های جدول ۱، نتایج به ازای ۵۰ تقسیم بندی تصادفی متفاوت (آموزش و آزمون) از دسته-داده‌ها تکرار شده و متوسط نرخ بازشناسی به ازای ۵۰ تکرار محاسبه می‌گردد. متوسط مقادیر نرخ بازشناسی بر روی پنج دسته‌داده‌ی پایگاه UCI، در شکل‌های (۱) الی (۳) نشان داده شده است. در شکل (۱)، میزان بهبود نرخ بازشناسی با استفاده از روش پیشنهادی، بر روی طبقه‌بند WRKNN نشان داده شده است. در شکل (۱)، نرخ بازشناسی WRKNN و WS-WRKNN به ازای تعداد همسایه‌های از ۲ الی ۱۵ رسم شده است. نتایج نشان می‌دهد که به ازای تمامی تعداد همسایه‌ها، روش پیشنهادی نرخ بازشناسی را

[15] Clustering

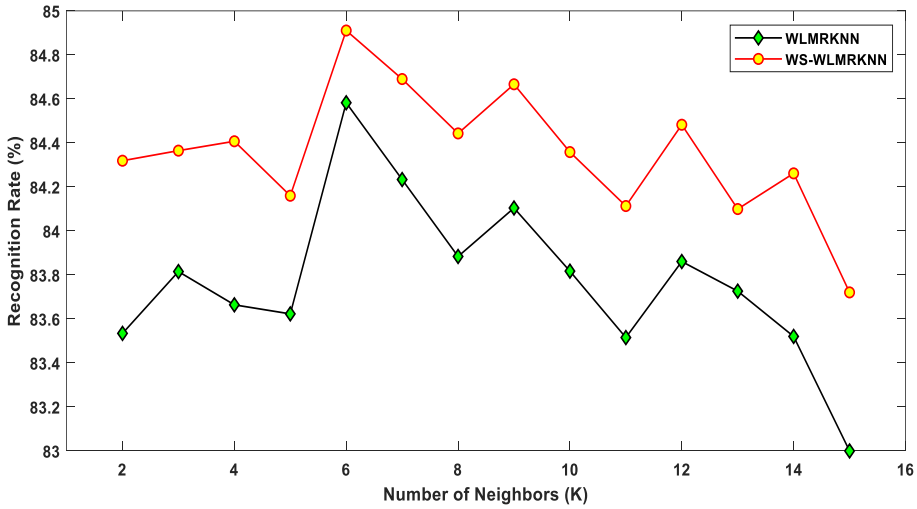
بهبود داده و موجب بهبود عملکرد طبقه‌بند WRKNN شده است.

در شکل (۲)، نتیجه عملکرد روش پیشنهادی بر روی طبقه‌بند WLMRKNN نشان داده شده و مجددا نیز به ازای تمامی تعداد همسایه‌ها موجب بهبود نرخ بازشناسی شده است. همچنین، مقایسه نتایج حاصل از طبقه‌بندهای MLMNN و WS-MLMNN نشان می‌دهد که روش پیشنهادی، نرخ بازشناسی را به ازای تعداد همسایه‌های ۸ الی ۱۵ به میزان ۳ درصد بهبود می‌بخشد.

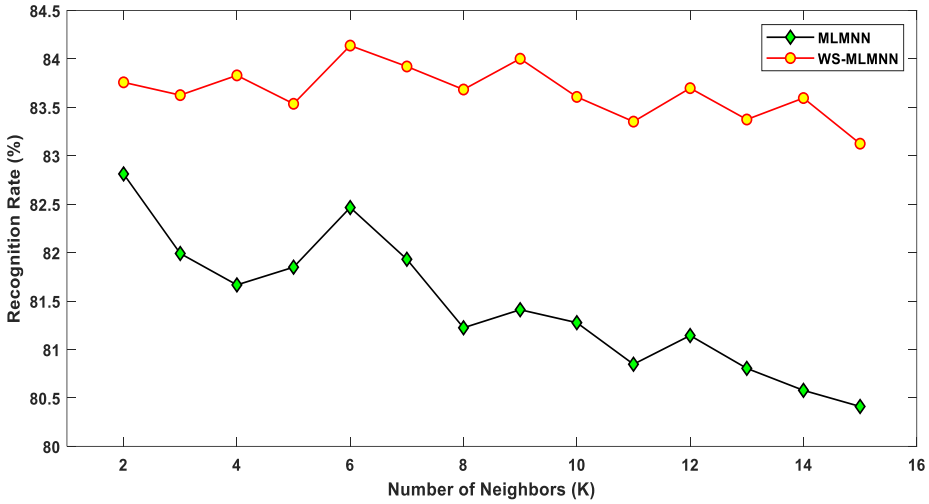
در جدول ۲، مقادیر انحراف معیار مقادیر نرخ بازشناسی حاصل از پنج دسته‌داده UCI تحت بررسی، به ازای تمامی همسایه‌های بیان شده است. مقادیر انحراف معیار نشان می‌دهد که روش پیشنهادی، موجب بهبود عملکرد WRKNN و WLMRKNN در تمامی دسته‌داده‌های تحت بررسی می‌شود و تقریبا به ازای تمامی تعداد همسایه‌های تحت بررسی، نرخ بازشناسی را بهبود می‌بخشد. همچنین، WS-MLMNN به ازای تعداد همسایه‌های کمتر از ۶، به ازای تمامی دسته‌داده‌ها از عملکرد بهتری از MLMNN برخوردار بوده و نرخ بهبود یکسانی را فراهم می‌سازد. اما به ازای تعداد همسایه‌های بیشتر، عملکرد آن کمی وابسته به دسته‌داده بوده و میزان بهبود عملکرد در برخی دسته‌داده‌ها بیشتر از بقیه بوده است.



شکل ۱- متوسط مقادیر نرخ بازشناسی بر روی پنج دسته‌داده‌ی پایگاه UCI با استفاده از طبقه‌بندهای WRKNN و WS-WRKNN به ازای تعداد همسایه‌های ۲ الی ۱۵.



شکل ۲- متوسط مقادیر نرخ بازشناسی بر روی پنج دسته‌داده‌ی پایگاه UCI با استفاده از طبقه‌بندهای WLMRKNN و WS-WLMRKNN به ازای تعداد همسایه‌های ۲ الی ۱۵.



شکل ۳- متوسط مقادیر نرخ بازشناسی بر روی پنج دسته‌داده‌ی پایگاه UCI با استفاده از طبقه‌بندهای MLMNN و WS-MLMNN به ازای تعداد همسایه‌های ۲ الی ۱۵.

جدول ۲- انحراف معیار مقادیر نرخ بازشناسی بدست آمده توسط طبقه‌بندهای MLMNN، WS-MLMNN، WRKNN، WS_WRKNN، WLMRKNN و WS-WLMRKNN به ازای تعداد همسایه‌های ۲ الی ۱۵ بر روی دسته‌داده‌های UCI.

تعداد همسایه (*K*)	انحراف معیار مقادیر نرخ بازشناسی					
	WS-MLMNN	MLMNN	WS-WLMRKNN	WLMRKNN	WS_WRKNN	WRKNN
۲	**۱۵.۸۰**	۱۷.۱۱	**۱۵.۴۸**	۱۶.۷۱	**۱۵.۶۵**	۱۶.۸۵
۳	**۱۶.۴۷**	۱۷.۸۶	**۱۶.۲۵**	۱۶.۹۶	**۱۶.۴۵**	۱۷.۱۲
۴	**۱۶.۵۴**	۱۸.۳۸	**۱۶.۵۶**	۱۷.۹۸	**۱۶.۶۸**	۱۸.۳۲
۵	**۱۷.۲۵**	۱۷.۶۸	**۱۷.۲۴**	۱۸.۰۲	**۱۷.۵۱**	۱۸.۳۹
۶	۱۶.۴۸	**۱۶.۲۴**	**۱۶.۳۹**	۱۶.۹۰	**۱۶.۵۱**	۱۷.۶۶
۷	۱۷.۰۸	**۱۶.۵۳**	**۱۷.۱۱**	۱۷.۵۵	**۱۷.۲۶**	۱۷.۹۳
۸	۱۷.۱۱	**۱۶.۳۹**	**۱۷.۳۱**	۱۸.۰۱	**۱۷.۸۲**	۱۸.۸۸
۹	۱۶.۶۱	**۱۶.۰۸**	**۱۷.۰۴**	۱۷.۶۲	**۱۷.۳۰**	۱۸.۰۸
۱۰	۱۷.۲۵	**۱۶.۴۰**	**۱۷.۷۳**	۱۸.۲۳	**۱۸.۳۳**	۱۸.۷۳
۱۱	۱۷.۷۳	**۱۶.۷۷**	**۱۸.۲۳**	۱۸.۸۲	**۱۸.۰۴**	۱۸.۳۵
۱۲	۱۷.۳۳	**۱۶.۰۴**	**۱۷.۷۴**	۱۷.۹۹	۱۷.۸۱	**۱۷.۶۳**
۱۳	۱۷.۳۴	**۱۵.۸۸**	**۱۸.۰۲**	۱۸.۳۲	۱۷.۹۵	**۱۷.۷۵**
۱۴	۱۷.۵۶	**۱۶.۴۱**	**۱۸.۲۹**	۱۸.۷۰	۱۸.۲۲	**۱۷.۶۰**
۱۵	۱۷.۶۶	**۱۶.۲۰**	**۱۸.۴۹**	۱۸.۷۸	۱۸.۲۷	**۱۷.۸۰**

۲-۴- پایگاه داده UCR

این پایگاه داده شامل دسته‌داده‌های سری زمانی متعددی است که در کاربردهای طبقه‌بندی مورد استفاده قرار می‌گیرد. در این مقاله ۱۰ دسته‌داده به کارگرفته شده، که مشخصات آنها در جدول ۳ بیان شده است. این پایگاه داده شامل داده‌های آموزش و آزمون به صورت مجزا بوده و نرخ بازشناسی بر حسب داده‌های در اختیار تعیین می‌گردد [۲۹]. ابعاد داده‌ها (که همان طول سری زمانی است)، تعداد دسته‌های موجود و تعداد داده‌های آموزش و آزمون در جدول ۳ نشان داده شده است.

نتایج به دست آمده در شکل‌های ۴ الی ۶ نشان داده شده است که شامل متوسط مقادیر نرخ بازشناسی بدست آمده بر روی ده دسته‌داده UCR به ازای تعداد همسایه‌های ۲ الی ۱۵ است. همچنین، در جدول ۴ انحراف معیار مقادیر نرخ بازشناسی بدست آمده توسط طبقه‌بندهای MLMNN، WS-MLMNN، WRKNN، WS-

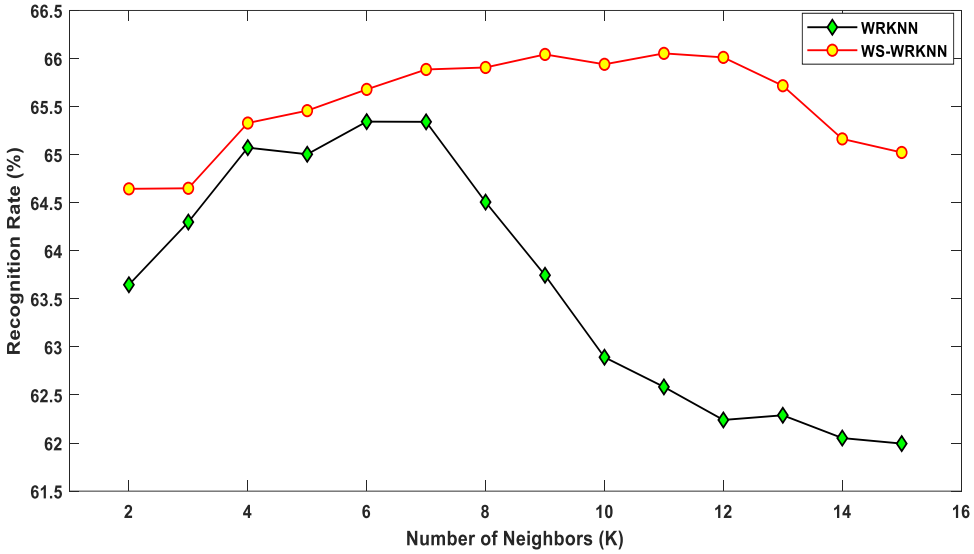
WRKNN، WLMRKNN و WS-WLMRKNN به ازای تعداد همسایه‌های ۲ الی ۱۵ بیان شده است که می تواند در ارزیابی نتایج به کار گرفته شود.

در شکل (۴) دو روش WRKNN و WS-WRKNN مورد ارزیابی قرار گرفته و نتایج نشان از بهبود نرخ بازشناسی مبتنی بر روش پیشنهادی مقاله به ازای تمامی تعداد همسایه‌ها دارد. بهبود حاصله در تعداد همسایه‌های بزرگتر در حدود ۴ درصد است که به میزان قابل توجهی نرخ بازشناسی بهبود می‌یابد. همچنین، انحراف معیار مقادیر نرخ بازشناسی حاصله توسط طبقه‌بند WS-WRKNN (که در جدول ۴ نشان داده شده است)، تقریبا به ازای تمامی همسایه‌ها، کمتر از طبقه‌بند WRKNN است که نتیجه می‌دهد روش WS-WRKNN بر روی تمامی دسته‌داده‌های تحت بررسی از عملکرد بهتری برخوردار بوده است.

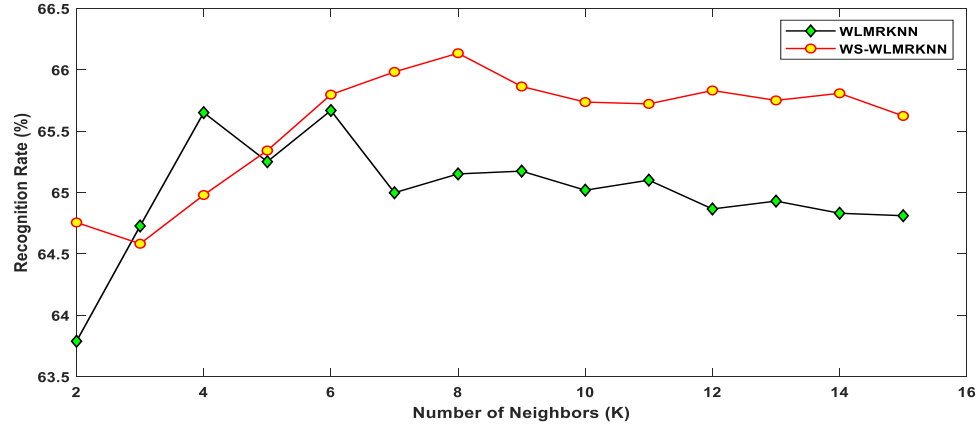
نرخ بازشناسی‌های نشان داده شده در شکل ۵ نیز، بیانگر عملکرد مناسب روش پیشنهادی است و نرخ بازشناسی به جز تعداد همسایه‌های ۳ و ۴ بهبود یافته است. همچنین، مقادیر انحراف معیار جدول ۴ نشان می‌دهد که میزان بهبود روش پیشنهادی بر روی دسته‌داده‌ها مختلف به ازای تعداد همسایه‌های ۳ الی ۱۱ کمی متفاوت بوده است.

جدول ۳- مجموعه دسته‌داده‌های سری زمانی UCR بکارگرفته شده و مشخصات آنها [۲۹].

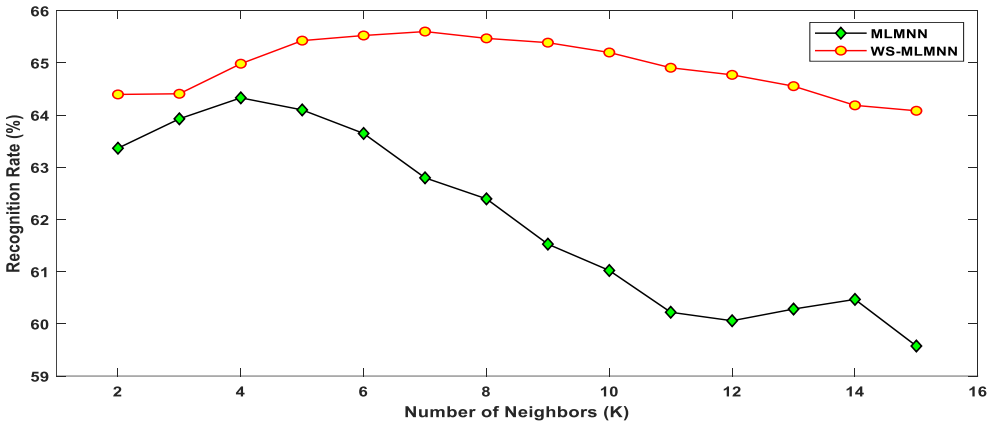
عنوان دسته داده	تعداد نمونه‌های آموزش	تعداد نمونه‌های آزمون	بُعد داده‌ها (طول سری زمانی)	تعداد دسته
ACSF1	۱۰۰	۱۰۰	۱۴۶۰	۱۰
Tow-Patterns	۱۰۰۰	۴۰۰۰	۱۲۸	۴
CricketY	۳۹۰	۳۹۰	۳۰۰	۱۲
CricketZ	۳۹۰	۳۹۰	۳۰۰	۱۲
DistalPhalanxOutlineAgeGroup	۴۰۰	۱۳۹	۸۰	۳
DistalPhalanxOutlineCorrect	۶۰۰	۲۷۶	۸۰	۲
DistalPhalanxTW	۴۰۰	۱۳۹	۸۰	۶
ECG200	۱۰۰	۱۰۰	۹۶	۲
EOGHorizontalSignal	۳۶۲	۳۶۲	۱۲۵۰	۱۲
EOGVerticalSignal	۳۶۲	۳۶۲	۱۲۵۰	۱۲



شکل ۴- متوسط مقادیر نرخ بازشناسی بر روی ده دسته‌داده‌ی پایگاه UCR با استفاده از طبقه‌بندهای WRKNN و WS-WRKNN به ازای تعداد همسایه‌های ۲ الی ۱۵.



شکل ۵- متوسط مقادیر نرخ بازشناسی بر روی ده دسته‌داده‌ی پایگاه UCR با استفاده از طبقه‌بندهای WLMRKNN و WS-WLMRKNN به ازای تعداد همسایه‌های ۲ الی ۱۵.



شکل ۶- متوسط مقادیر نرخ بازشناسی بر روی ده دسته‌داده‌ی پایگاه UCR با استفاده از طبقه‌بندهای MLMNN و WS-MLMNN به ازای تعداد همسایه‌های ۲ الی ۱۵.

جدول ۴- انحراف معیار مقادیر نرخ بازشناسی بدست آمده توسط طبقه‌بندهای MLMNN، WS-MLMNN، WRKNN، WS_WRKNN، WLMRKNN و WS-WLMRKNN به ازای تعداد همسایه‌های ۲ الی ۱۵ بر روی دسته‌داده‌های UCR.

تعداد همسایه (*K*)	انحراف معیار مقادیر نرخ بازشناسی					
	WS-MLMNN	MLMNN	WS-WLMRKNN	WLMRKNN	WS_WRKNN	WRKNN
۲	**۱۷.۳۴**	۱۷.۷۱	**۱۶.۹۱**	۱۷.۱۰	**۱۶.۹۲**	۱۷.۲۷
۳	۱۷.۰۹	**۱۷.۰۵**	۱۷.۰۷	**۱۶.۴۷**	۱۶.۷۱	**۱۶.۳۴**
۴	**۱۶.۷۱**	۱۷.۷۷	۱۶.۵۲	**۱۶.۴۰**	**۱۶.۴۱**	۱۶.۵۱
۵	**۱۶.۵۲**	۱۷.۸۱	۱۶.۴۸	**۱۶.۳۴**	**۱۶.۲۰**	۱۶.۳۵
۶	**۱۶.۵۱**	۱۷.۸۶	۱۶.۳۷	**۱۵.۹۸**	**۱۵.۸۷**	۱۵.۹۲
۷	**۱۶.۴۷**	۱۸.۸۱	۱۶.۲۸	**۱۵.۸۲**	**۱۵.۷۹**	۱۵.۸۷
۸	**۱۶.۵۵**	۱۸.۹۳	۱۶.۲۲	**۱۵.۶۱**	**۱۵.۷۴**	۱۶.۱۷
۹	**۱۶.۴۳**	۱۹.۷۵	۱۶.۱۹	**۱۵.۶۹**	**۱۵.۵۱**	۱۶.۳۳
۱۰	**۱۶.۴۶**	۲۰.۳۴	۱۶.۱۰	**۱۵.۸۲**	**۱۵.۴۵**	۱۶.۸۹
۱۱	**۱۶.۵۵**	۲۰.۹۹	۱۶.۰۱	**۱۵.۷۶**	**۱۵.۵۱**	۱۷.۳۵
۱۲	**۱۶.۸۰**	۲۰.۷۱	**۱۵.۷۵**	۱۵.۷۹	**۱۵.۳۹**	۱۷.۳۷
۱۳	**۱۶.۹۰**	۲۰.۹۰	**۱۵.۷۹**	۱۵.۸۲	**۱۵.۴۹**	۱۷.۲۷
۱۴	**۱۷.۲۲**	۲۰.۶۲	**۱۵.۷۶**	۱۶.۰۱	**۱۵.۷۴**	۱۷.۵۷
۱۵	**۱۷.۲۱**	۲۱.۱۰	**۱۵.۸۵**	۱۶.۱۴	**۱۵.۹۵**	۱۷.۴۵

شکل (۶) و جدول ۴ نیز اثبات می‌نماید که روش پیشنهادی، عملکرد طبقه‌بند MLMNN را بهبود بخشیده و در برخی تعداد همسایه‌ها، میزان بهبود نرخ بازشناسی به ۵ درصد می‌رسد که بسیار ارزشمند است.

۵- نتیجه‌گیری

عمل طبقه‌بندی، بخشی جداناپذیر در بیشتر کاربردهای هوش مصنوعی و آموزش ماشین است. در این مقاله، روشی جهت بهبود عملکرد طبقه‌بندهای KNN مبتنی بر کمترین بازسازی خطی معرفی شده است. در روش پیشنهادی، پس از تعیین همسایه‌ها به ازای هر دسته، مقادیر خطای بازسازی داده آزمون به ازای نزدیکترین همسایه، دو نزدیکترین همسایه، الی K نزدیکترین همسایه محاسبه می‌گردد. سپس مقدار خطای بازسازی کل به صورت جمع وزن‌دار خطاها به ازای هر دسته محاسبه می‌گردد. این عمل موجب هم‌افزایی و بهره‌گیری از اطلاعات قابل استخراج به ازای تمامی تعداد همسایه‌ها می‌گردد. در آخر نیز داده آزمون متعلق به دسته‌ای است که از کمترین خطای بازسازی برخوردار است. سه طبقه‌بند پیشنهادی در این مقاله، WS-WRKNN، WS-WLMRKNN و WS-MLMNN نامگذاری شده‌اند که به ترتیب در الگوریتم‌های ۱ الی ۳ بیان شده‌اند. در این مقاله، از دو پایگاه داده متداول طبقه‌بندی UCI (پنج دسته‌داده) و UCR (ده دسته‌داده) استفاده شده است. نتایج (شکل‌های ۱ الی ۶ و جدول‌های ۲ و ۴) نشان می‌دهد که به ازای تقریبا تمامی ارزیابی‌ها، روش پیشنهادی از عملکرد بهتری برخوردار بوده و در برخی موارد، نرخ بازشناسی در حدود ۵ درصد بهبود یافته است.

به طور کلی، روش پیشنهادی توانسته عملکرد طبقه‌بندهای KNN مبتنی بر کمترین بازسازی خطی را بهبود بخشد. در ادامه، علاقه‌مندان می‌توانند در مورد نحوه انتخاب همسایه‌های تاثیرگذار در خطای بازسازی و همچنین نحوه تعیین وزن‌های ترکیب خطاها به تحقیق و توسعه بپردازند. روش‌های مبتنی بر یادگیری شورایی نیز ایده‌ای است که می‌تواند به کار گرفته شود.

مراجع

[1] Zhang, J. Z., P. R. Srivastava, D. Sharma, and P. Eachempati. "Big data analytics and machine learning: A retrospective overview and bibliometric analysis." *Expert Systems with Applications* 184 (2021): 115561.

[2] Pucchio, A., E. A. Eisenhauer, and F. Y. Moraes. "Medical students need artificial intelligence and machine learning training." *Nature Biotechnology* 39, no. 3 (2021): 388-389.

[3] Hassanat, A. B., H. N. Ali, A. S. Tarawneh, M. Alrashidi, M. Alghamdi, G. A. Altarawneh, and M. A. Abbadi. "Magnetic Force Classifier: A Novel Method for Big Data Classification." *IEEE Access* 10 (2022): 12592-12606.

[4] Nejadshahmohammad. F. "Development of Multi-similarity index clustering algorithm in Mathematical Modelling of Mines". Journal of Modeling in Engineering 17. no. 56 (2019): 267-279. (inPersian)

[5] Tchapga, Tchito C., T. A. Mih, A. Tchagna Kouanou, T. Fozin Fonzin, P. Kuetche Fogang, B. A. Mezatio, and D. Tchiotsop. "Biomedical image classification in a big data architecture using machine learning algorithms." *Journal of Healthcare Engineering* (2021).

[6] Alam, S., and N. Yao. "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis." *Computational and Mathematical Organization Theory* 25, no. 3 (2019): 319-335.

[7] Soto, P. C., N. Ramzy, F. Ocker, and B. Vogel-Heuser. "An ontology-based approach for preprocessing in machine learning." In *2021 IEEE 25th International Conference on Intelligent Engineering Systems (INES)*, 2021, pp. 000133-000138.

[8] Yadav, D. P., A. Sharma, M. Singh, and A. Goyal. "Feature extraction based machine learning for human burn diagnosis from burn images." *IEEE Journal of Translational Engineering in Health and Medicine* 7 (2019): 1-7.

[9] Dong, S., P. Wang, and K. Abbas. "A survey on deep learning and its applications." *Computer Science Review* 40 (2021): 100379.

[10] Janiesch, C., P. Zschech, and K. Heinrich. "Machine learning and deep learning." *Electronic Markets* 31, no. 3 (2021): 685-695.

[11] Alsaqqa, A. H., M. A. Alkahlout, and S. S. Abu-Naser. "Using Deep Learning to Classify Different Types of Vitamin." *International Journal of Academic Engineering Research (IJAER)* 6, no. 1 (2022): 1-6.

[12] Sadeghi. M. H Marvi. A.R Ahmadyfard. "A New and Efficient Feature Extraction Method for Robust Speech Recognition Based on Fractional Fourier Transform and Differential Evolution Optimizer". Journal of Modeling in Engineering 18. no. 61 (2020): 85-96.

[13] Harimi. A. K Yaghmaie. "Improving speech emotion recognition via gender classification". Journal of Modeling in Engineering 15. no. 48 (2017): 183-200.

[14] Javaid, A., M. Sadiq, and F. Akram. "Skin cancer classification using image processing and machine learning." In *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*, 2021, pp. 439-444.

[15] Cover, T., and P. Hart. "Nearest neighbor pattern classification." *IEEE Transaction on Information Theory* 13, no. 1 (1967): 21-27.

[16] You, S., C. Xu, C. Xu, and D. Tao. "Learning with Single-Teacher Multi-Student." In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018, pp. 4390-4397.

[17] Chaudhary, A., S. Kolhe, and R. Kamal. "An improved random forest classifier for multi-class classification." *Information Processing in Agriculture* 3, no. 4 (2016): 215-222.

[18] Uebele, V., S. Abe, and M. S. Lan. "A neural-network-based fuzzy classifier." *IEEE Transactions on Systems, Man, and Cybernetics* 25, no. 2 (1995): 353-361.

[19] Gou, J., W. Qiu, Z. Yi, X. Shen, Y. Zhan, and W. Ou. "Locality constrained representation-based K-nearest neighbor classification." *Knowledge-Based Systems* 167 (2019): 38-52.

[20] Zeng, Y., Y. Yang, and L. Zhao. "Pseudo nearest neighbor rule for pattern classification." *Expert Systems with Applications* 36 (2009): 3587-3595.

[21] Gou, J. P., Y. Z. Zhan, Y. B. Rao, X. J. Shen, X. M. Wang, and W. He. "Improved pseudo nearest neighbor classification." *Knowledge-Based System* 70 (2014): 361-375.

[22] Mitani, Y., and Y. Hamamoto. "A local mean-based nonparametric classifier." *Pattern Recognition Letters* 27, no. 10 (2006): 1151-1159.

[23] Gou, J. P., W. M. Qiu, Q. R. Mao, Y. Z. Zhan, X. Z. Shen, and Y. B. Rao. "A Multi-Local Means Based Nearest Neighbor Classifier." In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2017, pp. 448-452.

[24] Li, W., Q. Du, F. Zhang, and W. Hu. "Collaborative-Representation Based Nearest Classifier for Hyperspectral Imagery." *IEEE Geoscience and Remote Sensing Letters* 12, no. 2 (2015): 389-393.

[25] Pan, Z. P., Y. D. Wang, and W. P. Ku. "A new k-harmonic nearest neighbor classifier based on the multi-local means." *Expert Systems with Applications* 67 (2017): 115-125.

[26] Dudani, S. A. "The distance-weighted k-Nearest Neighbor rule." *IEEE Transaction on Systems, Man and Cybernetics* 6, no. 4 (1976): 325-327.

[27] Hajizadeh, R., A. Aghagolzadeh, and M. Ezoji. "Mutual neighborhood and modified majority voting based KNN classifier for multi-categories classification." *Pattern Analysis and Applications* (2022): 1-21.

[28] Dua, D., and C. Graff. "UCI Machine Learning Repository." Irvine, CA: University of California, School of Information and Computer Science, 2019.

[29] Chen, Y., E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. "The UCR Time Series Classification Archive." 2015.