Development of a Model for Estimating and Classifying the Educational Performance of Undergraduate Students Using a Combination of Multilayer Neural Networks
(Case study: Qom University)

Hadiyeh Mahdavi
M.Sc. in Industrial Engineering; Department of Industrial Engineering; Faculty of Engineering; Qom University; Qom, Iran;
Email: h.mahdavi72 @ gmail.com

Jalal Rezaei Noor*
PhD in Industrial Engineering; Associate Professor; Department of Industrial Engineering; Faculty of Engineering; Qom University; Qom, Iran Email: J.rezaee@qom.ac.ir

Mohammad Amini
PhD in Information Technology Engineering; Faculty of Industrial Engineering; University of Science and Technology; Tehran, Iran
Email: moha_amini@ind.iust.ac.ir

Iranian Journal of Information Processing and Management

Received: 18, Mar. 2021 **Accepted: 16, Nov. 2021**

Abstract: Students' growth and development have always been considered important by the education system as they are the future assets of their country. Although many students can flourish their talents and creativity, we are faced with a large number of students each year whose talents are squandered leading them to a position far from success. Today, with the more students entering in different academic levels and the variety of study disciplines, the necessity to properly guide students is felt more than ever. For this purpose, educational data mining has received special attention from the educational system's officials in recent years. So far, various classification methods and techniques in data mining and machine learning have been used to predict student performance. However, these individual classifiers have limitations such as complexity and instability for predicting performance in the education process. To tackle this problem, ensemble classification has been proposed as a new and efficient method. Ensemble classification systems combine the results of several individual classifiers to provide

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 37 | No. 4 | pp. 1319-1350
Summer 2022
https://doi.org/10.35050/JIPM010.2022.012

* Corresponding Author

a model with better performance. In this paper, a new Ensemble classification system is presented using multilayer neural networks and SOM clustering in order to estimate and classify the grade point average of undergraduate students. In addition, we used averaging and majority voting as combination methods for aggregating the results of individual classifiers. Evaluation results on real university data show that our proposed ensemble system provides better accuracy and performance compared to prevalent individual classification methods. Also, the proposed ensemble system obviously outperforms other popular ensemble methods in classifying students' GPA.

Keywords: Clustering, Educational Data Mining, Ensemble Classification, Neural Networks

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

توسعهٔ مدلی برای تخمین و دسته‌بندی عملکرد آموزشی دانشجویان کارشناسی با استفاده از ترکیب شبکه‌های عصبی چندلایه (مطالعهٔ موردی: دانشگاه قم)

هدیه مهدوی
کارشناسی ارشد مهندسی صنایع؛ گروه مهندسی صنایع؛ دانشکدهٔ فنی و مهندسی؛ دانشگاه قم؛ قم، ایران؛
h.mahdavi72 @ gmail.com

جلال رضائی‌نور
دکتری مهندسی صنایع؛ دانشیار؛ گروه مهندسی صنایع؛ دانشکدهٔ فنی و مهندسی؛ دانشگاه قم؛ قم، ایران؛
پدیدآور رابط J.rezaee@qom.ac.ir

محمد امینی
دکتری مهندسی فناوری اطلاعات؛ دانشکدهٔ مهندسی صنایع؛ دانشگاه علم و صنعت؛ تهران، ایران؛
moha_amini@ind.iust.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۹/۱۲/۲۸ **پذیرش:** ۱۴۰۰/۰۸/۲۵ مقاله برای اصلاح به مدت ۲ ماه نزد پدیدآوران بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۷ | شمارهٔ ٤ | ص‌ص ۱۳۱۹-۱۳۵۰
تابستان ۱۴۰۱
https://doi.org/10.35050/JIPM010.2022.012

چکیده: رشد و پیشرفت دانشجویان به‌عنوان سرمایه‌های آیندهٔ کشور همواره مورد توجه و اهمیت نظام آموزش و پرورش بوده است. چه بسیار دانشجویانی که خلاقیت و استعدادشان شکوفا شده و در مقابل، سالیانه با تعداد کثیری از دانشجویان مواجه هستیم که استعدادهایشان تلف شده و از موفقیت دور مانده‌اند. امروزه، با افزایش دانشجویان در مقاطع مختلف دانشگاهی و تنوع رشته‌های تحصیلی، لزوم هدایت صحیح دانشجویان بیشتر از پیش احساس می‌شود. بدین‌منظور، داده‌کاوی آموزشی در سال‌های اخیر مورد توجه ویژهٔ مسئولان نظام آموزش و پرورش قرار گرفته است. تاکنون روش‌ها و تکنیک‌های دسته‌بندی متنوعی در حوزهٔ داده‌کاوی و یادگیری ماشین به‌منظور دسته‌بندی و پیش‌بینی عملکرد دانشجویان به‌کار رفته است. اما این دسته‌بندهای منفرد برای پیش‌بینی عملکرد در فرایند آموزش، دارای محدودیت‌هایی از قبیل پیچیدگی و عدم ثبات هستند. برای مقابله با این مشکل، دسته‌بندهای جمعی به‌عنوان روش‌های نوین و کارآمد مطرح می‌شوند. سیستم‌های دسته‌بندی جمعی نتایج چندین دسته‌بند منفرد را ترکیب می‌کنند و مدلی با عملکرد بهتر

ارائـه می‌دهنـد. در ایـن پژوهـش یـک دسـته‌بند جمعـی جدیـد بـا اسـتفاده از شـبکه‌های عصبـی چندلایـه و خوشـه‌بندی SOM به‌منظـور تخمیـن و دسـته‌بندی معـدل دانشـجویان دورهٔ کارشناسـی ارائـه شـده اسـت. همچنیـن، از روش ترکیبـی میانگین‌گیـری و رأی اکثریـت بـرای ترکیـب نتایـج دسـته‌بندهای منفـرد اسـتفاده شـده اسـت. نتایـج ارزیابـی بـر روی داده‌هـای واقعـی دانشـگاه نشـان می‌دهـد کـه مـدل پیشـنهادی ارائه‌شـده در ایـن پژوهـش دقـت و کارایـی بیشـتری نسـبت بـه روش‌هـای دسـته‌بندی منفـرد مشـهور و پرکاربـرد دارد. همچنیـن، مـدل پیشـنهادی در مقایسـه بـا روش‌هـای جمعـی معـروف، عملکـرد بهتـری در دسـته‌بندی معـدل دانشـجویان داشـته اسـت.

کلیدواژه‌ها: خوشه‌بندی، داده‌کاوی آموزشی، دسته‌بندهای جمعی، شبکه‌های عصبی

۱. مقدمه

در سـال‌های گذشـته روش‌هـای داده‌کاوی در حوزه‌هـای متفاوتـی ماننـد تجـارت الکترونیـک، سـلامت، بهداشـت و درمـان مـورد اسـتفاده قـرار گرفتـه و توجـه پژوهشـگران و نویسـندگان حـوزهٔ آمـوزش را بـه خـود جلـب کـرده اسـت. در حـوزهٔ آمـوزش بـرای بررسـی سـؤالات کلیـدی و یافتـن شـواهد تجربـی بـرای بهبـود و توسـعه از ایـن روش‌هـا اسـتفاده شـده اسـت کـه ایـن امـر، موجـب پدیدآمـدن حـوزه جدیـدی از پژوهـش بـا نـام داده‌کاوی آموزشـی شـده اسـت. امـروزه ارزیابـی و بهبـود عملکـرد پژوهشـی دانشـجویان در دانشـگاه‌ها، بـه منظـور ارتقـا و پیشـرفت سـطح تحقیقاتـی، از موضوعـات بسـیار اساسـی و کلیـدی در حـوزه آمـوزش عالـی می‌باشـد. در واقـع، اطلاعـات کشف‌شـده از داده‌هـای دانشـجویان و بررسـی پارامترهـای مؤثـر بـر موفقیـت دانشـجویان بـه یافتـن الگوهـای خـاص، برنامـه یـا روش موفـق تدریـس و یافتـن نقـاط ضعـف دانشـجو منجـر می‌شـود و می‌توانـد در عرصـهٔ پیشـرفت آمـوزش عالـی و برنامه‌ریزی‌هـای درسـی دانشـجو در جهـت بهبـود نمـرات نقـش مهمـی ایفـا کنـد. یکـی از اقداماتـی کـه دانشـگاه‌ها می‌تواننـد بـرای جلوگیـری از شکسـت دانشـجویان و بلکـه پیشـرفت آن‌هـا انجـام دهنـد، تخمیـن و دسـته‌بندی معـدل اسـت. تخمیـن معـدل، به‌عنـوان یـک روش ارزیابـی مؤثـر می‌توانـد کمـک زیـادی بـه سرنوشـت تحصیلـی دانشـجویان بکنـد. به‌عنـوان مثـال، در مـواردی مشـاهده شـده اسـت کـه دانشـجویان در ترم‌هـای اولِ سـال تحصیلـی روند روبه‌رشـدی داشـته، ولـی در ترم‌هـای بعـدی بـا افـت شـدیدی مواجـه می‌شـوند و همیـن عامـل منجـر بـه تـرک تحصیـل و یـا انـزوای اجتماعـی می‌شـود. علـت ایـن امـر، برنامه‌ریـزی نادرسـت نظـام آموزشـی و نادیـده گرفتـن شـرایط فـردی، تحصیلـی، اجتماعـی و حتـی خانوادگـی دانشـجویان اسـت. تخمیـن معـدل دانشـجویان ایـن امـکان را فراهـم می‌کنـد کـه دانشـگاه‌ها قبـل از شکسـت آن‌هـا در نتایـج

به‌دست‌آمده و بروز مشکلات آتی، به فکر حل مشکلات افتاده و راهی برای موفقیت دانشجو بیابند. می‌توان گفت، شرایط خاص دانشجو می‌تواند بر نتایج وعملکرد وی تأثیر مضاعفی داشته باشد. ویژگی‌های جسمی، روحی، خانوادگی و نمرات درسی دانشجویان می‌تواند منجر به کسب نتایج عالی و یا ضعیف بشوند و همین امر می‌تواند هشداری برای نظام آموزشی در دانشگاه‌ها باشد. به‌عبارت دیگر می‌توان گفت که پیش‌بینی و دسته‌بندی عملکرد دانشجویان یکی از مشهورترین و مفیدترین هدف‌های داده‌کاوی آموزشی است. این امر شامل تخمین زدن مقادیر ناشناختهٔ عملکرد دانشجویان مانند نمره یا رتبهٔ تحصیلی می‌شود (Romero et al. 2013). حل تعداد زیادی از مسائلی که بر عملکرد دانشجویان تأثیر می‌گذارند، مانند مسائل مربوط به وضعیت اقتصادی-اجتماعی، تجارب گذشتهٔ دانشجو، تعامل میان دانشگاه‌ها، ویژگی‌های جغرافیایی، وضعیت روحی و روانی و پیش‌زمینه فرهنگی (Araque, Roldn and Salguero 2009; Richardson, Abraham and Bond 2012) یک مسئلهٔ چالشی و بحث‌برانگیز است. تحقیقات متنوعی با موضوع داده‌کاوی در حیطهٔ آموزش در حوزه‌های کاربردی مختلف انجام شده است. هدف برخی از مطالعات پیش‌بینی این امر است که آیا دانشجو فارغ‌التحصیل می‌شود یا خیر (Aluko et al. 2016). در حالی که هدف مطالعات دیگر پیش‌بینی نمرهٔ نهایی آن‌ها بوده است (Guruler, Istanbullu and Karahasan 2010; Laugerman et al. 2015). مطالعات دیگری نیز برای تعیین عملکرد دانشگاهی در انتهای سال اول تحصیلی انجام شده‌اند (Hoffait and Schyns 2017; Gray, McGuinness and Owende 2014). در حوزهٔ پیش‌بینی ترک تحصیل دانشجویان نیز می‌توان به مطالعات (Bydzovska 2016; Arquez-Vera et al. 2016; Tinto 1982) اشاره کرد.

تاکنون روش‌های داده‌کاوی همچون درخت تصمیم[1]، شبکهٔ عصبی چندلایه[2]، ماشین بردار پشتیبان[3]، نزدیک‌ترین همسایگی[4]، «نایو بیز»[5] و ... در این زمینه به‌کار رفته است. به‌عنوان مثال، Natek and Zwilling (2014) با تمرکز بر داده‌های کوچک دانشجویی به پیش‌بینی نرخ موفقیت دانش‌آموزان آموزش عالی و مدیریت دانش پرداختند. Ramaswami and Bhaskaran (2010) حدود ۱۰۰۰ دادهٔ دانش‌آموزی را از چند مدرسهٔ مختلف در مقطع دوم دبیرستان جمع‌آوری کرده و ویژگی‌هایی مانند جنسیت دانش‌آموز، وضعیت خانوادگی

1. decision tree
2. multilayer perceptron
3. support vector machine
4. k-nearest neighbor
5. naïve base

وی و نوع مدرسه را بررسی نمودند. Huang et al. (2009) ابتدا داده‌های شخصی دانش‌آموزان را جمع‌آوری کردند، و سپس، با روش درخت تصمیم و شبکه‌های عصبی عوامل مؤثر را شناسایی و تعیین نمودند. Lykourentzou et al. (2009) با استفاده از روش‌های شبکهٔ عصبی بر اساس فعالیت‌های درسی از جمله نمرات کوئیز به دسته‌بندی دانش‌آموزان پرداختند. Agarwal, Pandey and Tiwari (2012) در پژوهش خود داده‌های دانشجویان را از دانشگاه جمع‌آوری کرده و روش‌های متنوع داده‌کاوی از قبیل شبکه‌های عصبی چندلایه و ماشین بردار پشتیبان را با نرم‌افزار وکا[1] اجرا و نتایج مقایسه شدند. Şen, Uçar and Delen (2012) با استفاده از داده‌های دانشگاهی و ویژگی‌هایی از قبیل جنسیت، نمرات تحصیلی گذشته، تعداد خواهر و برادر، بورسیه، نوع مدرسه و مذهب دانشجو، به پیش‌بینی و آنالیز حساسیت عملکرد دانش‌آموزان سال دوم دبیرستان پرداختند. Mueen, Zafar and Manzoor (2016) با مقایسهٔ سه روش «نایو بیز»، درخت تصمیم و شبکه‌های عصبی چندلایه و با استفاده از نرم‌افزار «وکا» به پیش‌بینی و ارزیابی عملکرد پرداختند.

روش‌های به‌کاررفته در مطالعات در پیش‌بینی موفقیت‌آمیز هستند، اما هر کدام از آن‌ها به‌تنهایی ضعف‌هایی دارند. به همین علت، به‌تازگی دسته‌بندهای جمعی[2] و هیبریدی توسط محققان برای انواع مسائل دسته‌بندی با هدف بالاتر بردن دقت عملکرد پیشنهاد شده‌اند (Twala 2010; Polat and Güneş 2009; Moschopoulos et al. 2009).

رویکرد ترکیبی چندین روش داده‌کاوی را به‌منظور افزایش عملکرد کلی و نهایی سیستم ترکیب می‌کند. به‌عبارتی، هدف الگوریتم‌های دسته‌بندی جمعی بهبود از طریق ترکیب خروجی‌های چندگانه دسته‌بندهای تکی[3] ضعیف است (Kittler et al. 1998). در واقع، دسته‌بندهای جمعی با ترکیب نتایج دسته‌بندهای منفرد، دقتی بالاتر و دقیق‌تر ارائه می‌دهند.

مطالعات زیادی درگذشته انجام شده است که نشان می‌دهد دسته‌بندهای جمعی، عملکرد دقیق‌تر و باثباتی نسبت به دسته‌بندهای منفرد ارائه داده‌اند (Oza and Tumer 2001; Krogh and Vedelsby 1995; Ho, Hull and Srihari 1994). بعضی دیگر از محققان راه‌های متفاوت ترکیب دسته‌بندها را برای ایجاد سیستم جمعی ارائه داده‌اند (Ji and Ma 1997; Dzeroski and Zenko 2002). یکی از بهترین روش‌هایی که می‌تواند دقت خوب و انعطاف‌پذیری بالا در

1. Weka
2. ensemble classifiers
3. individual classifiers

دسته‌بندی داده‌ها داشته باشد، روش شبکه‌های عصبی مصنوعی است. این‌ها گزینه‌های خوبی به‌عنوان یادگیرنده‌های پایه برای روش‌های جمعی هستند. ترکیب جمعی شبکه‌های عصبی در مطالعات زیادی از جمله Amini, Rezaeenour and Hadavandi (2016); Amini and Rezaeenour (2014); Amini, Rezaeenour and Hadavandi (2014); Amini, Rezaeenour and Hadavandi (2015) استفاده شده است. در همین راستا (Hansen and Salamon 1990) توانایی ترکیب جمعی شبکه‌های عصبی را شرح دادند. بنابراین، مشهور بودن و عملکرد بالای این روش‌ها موجب شده است که محققان در حوزهٔ داده‌کاوی آموزشی از این روش‌ها بهره ببرند.

هدف این پژوهش، ارائهٔ یک دسته‌بند جمعی برای تخمین و دسته‌بندی عملکرد دانشجویان بر اساس ویژگی‌های تحصیلی، فردی، اجتماعی و خانوادگی آن‌هاست که بر اساس زیرمجموعه‌های مختلف از کل شبکهٔ عصبی ترکیبی آموزش می‌بینند.

در واقع، می‌توان گفت که مقالهٔ حاضر، ترکیبی از یک روش حل نوین ارائه‌شده با استفاده از الگوریتم‌های خوشه‌بندی و ماژول‌های ترکیب است که ورودی آن ویژگی‌های جدیدی از دانشجویان است. هدف از این مطالعه پاسخگویی به نیاز مسئولان آموزشی در جهت پیش‌بینی و دسته‌بندی عملکرد تحصیلی دانشجویان است تا با استفاده از آن بتوان به تصمیم‌گیری مدیران در قبال وضعیت دانشجویان کمک نماید. در واقع، مدل دسته‌بندی ارائه‌شده، در چارچوب یک سیستم تصمیم‌یار عمل می‌کند که خروجی آن برآوردی از وضعیت تحصیلی هر دانشجوست.

از آنجا که معدل تحصیلات یک دانشجو می‌تواند برآورد مناسبی برای بررسی وضعیت تحصیلی او باشد، در این مقاله بر روی تخمین معدل تأکید شده است. با توجه به حساسیتی که این نوع تخمین دارد و اطلاعاتی که به تصمیم‌گیر ارائه می‌دهد، لازم است از مدل‌هایی استفاده شود که حداکثر دقت و صحت را داشته باشند. به همین جهت، رویکرد اتخاذشده در این مقاله از روش دسته‌بندی جمعی استفاده می‌کند.

اصولاً در مسائلی با جامعهٔ آماری بالا فرایند دسته‌بندی به‌منظور شناسایی الگوها و تجزیه و تحلیل داده‌های کلان صورت می‌گیرد. با استفاده از عملیات دسته‌بندی یا طبقه‌بندی، داده‌های دانشجویان در دسته‌هایی مجزا قرار می‌گیرند و می‌توانیم روابط بین داده‌ها را کشف و بر اساس آن‌ها الگوهای خروجی را تحلیل کنیم. در واقع، در عملیات طبقه‌بندی، سیستم بر اساس یک مجموعهٔ آموزشی یاد می‌گیرد که داده‌ها را به گروه‌های

درست با کمترین خطا تقسیم و دسته‌بندی کند. مجموعهٔ آموزش حاوی داده‌هایی است که دستهٔ آن‌ها مشخص است، هر الگو یا دسته یک برچسب دارد و داده‌هایی با برچسب هدف یکسان در یک گروه قرار می‌گیرند. با مقایسهٔ برچسب پیش‌بینی با برچسب واقعی، دقت عملکرد مدل تعیین می‌شود. اهمیت دسته‌بندی عملکرد داده‌های دانشجویان زمانی مشخص می‌شود که دانشجویان با ویژگی‌های متنوع و متفاوت درکنار یکدیگر قرار می‌گیرند و با یک جامعهٔ آماری عظیمی روبه‌رو هستیم. با استفاده از تکنیک‌های دسته‌بندی می‌توانیم تخمین و برآوردی نسبت به عملکرد تحصیلی آن‌ها داشته باشیم. دانشگاه‌ها و اساتید با استفاده از روش‌های دسته‌بندی می‌توانند از وضعیت دانشجویان مطلع شده و شیوه‌هایی همچون برگزاری دوره‌های کمک‌آموزشی، استفاده از اساتید برتر و اولویت قرار دادن مشکلات دانشجویان را به‌کار گیرند تا سطح کیفی دانشگاه و عملکرد دانشجویان در سطح مطلوب قرار گیرد.

ویژگی‌هایی که برای دسته‌بندی عملکرد دانشجویان مورد بررسی قرار گرفته‌اند، به‌شرح زیر است:

ویژگی‌های تحصیلی:

- ◇ سال ورود به دانشگاه: ورودی‌های سال‌های ۱۳۸۵ تا ۱۳۹۵؛
- ◇ رشتهٔ تحصیلی: شامل ۶ گروه الهیات و معارف اسلامی، علوم پایه، ادبیات و علوم انسانی، مدیریت، حقوق و فنی و مهندسی که در مجموع، ۲۶ رشتهٔ تحصیلی است؛
- ◇ دورهٔ تحصیلی: شامل دورهٔ الکترونیک، بورسیهٔ غیرایرانی، روزانه، روزانه با هزینه، روزانه با هزینهٔ غیرایرانی، غیرایرانی غیربورسیه و دورهٔ شبانه؛
- ◇ تعداد مشروطی متناوب: دانشجویانی که هیچ مشروطی نداشتند یا از ۱ تا ۵ مشروطی داشتند؛
- ◇ تعداد مشروطی متوالی: دانشجویانی که هیچ مشروطی متوالی نداشته و یا به‌طور متوالی از ۱ تا ۴ مشروطی داشتند؛
- ◇ نوع ورود به دانشگاه: از طریق آزمون ورودی سازمان سنجش، انتقال، انتقال دائم، بورسیه، غیرایرانی غیربورسیه، مجوز سازمان سنجش-انفرادی، معرفی‌نامهٔ وزارت علوم و مهمان؛
- ◇ تعداد واحد مردودی: از ۰ تا ۵۷ در نظر گرفته شده است؛

- ◇ تعداد ترم مرخصی: دانشجویانی که هیچ ترمی مرخصی نگرفته‌اند و یا ۱ تا ۲ بار مرخصی داشتند؛
- ◇ معدل کل: معدل کل دانشجویان به‌عنوان ویژگی مورد تخمین در مقاله در ۴ دسته مورد بررسی قرار گرفته است.

ویژگی‌های فردی:

- ◇ سن یا تاریخ تولد: دانشجویان متولد ۵۶ تا ۷۸؛
- ◇ جنسیت: مرد و زن

ویژگی‌های اجتماعی:

- ◇ بومی بودن: دانشجویانی که بومی شهر مورد مطالعه بوده و یا از شهرهای دیگر در دانشگاه مشغول به تحصیل بوده‌اند؛
- ◇ تحت پوشش کمیتهٔ امداد: دانشجویانی که تحت پوشش کمیتهٔ امداد قرار دارند و دانشجویانی که در این گروه قرار ندارند؛
- ◇ تحت پوشش بهزیستی: دانشجویانی که تحت پوشش بهزیستی قرار دارند و دانشجویانی که در این گروه قرار ندارند.

ویژگی‌های خانوادگی:

- ◇ درآمد خانواده: خانواده‌هایی که ماهانه بیش از ۶ میلیون تومان، بین ۴ تا ۶ میلیون تومان، بین ۲ تا ۴ میلیون تومان و کمتر از ۲ میلیون تومان و یا فاقد درآمد بوده‌اند؛
- ◇ وضعیت تأهل: مجرد یا متاهل بودن دانشجویان.

در این روش پیشنهادی جمعی از رویکرد خوشه‌بندی جهت تقسیم مجموعهٔ داده‌ها به زیرمجموعه‌های کوچک‌تر بهره گرفته شده است؛ به‌گونه‌ای که دسته‌بندهای مبنا بتوانند آسان‌تر و سریع‌تر آموزش ببینند. همچنین، پراکندگی و تنوع خوبی از طریق آموزش دسته‌بندها با زیرمجموعه‌های متفاوت از داده‌ها که اشتراکی با هم ندارند، در بین اعضای سیستم جمعی ایجاد شده است. روش‌های ترکیبی قدرتمندی نیز برای بهبود دقت نهایی دسته‌بندی ارائه شده است.

در حقیقت نوآوری‌های ارائه‌شده در این پژوهش عبارت‌اند از:

۱. استفاده از ویژگی‌های جدید دانشجویان که شامل ویژگی‌های تحصیلی، فردی، اجتماعی و خانوادگی است؛

۲. ارائهٔ یک روش حل نوین به‌منظور دسته‌بندی عملکرد، که با نام دسته‌بندهای جمعی مطرح شده و در ۳ قسمت زیر عنوان می‌شود:

الف) فضای کل داده‌های آموزشی توسط شبکه عصبی خودسازمانده[1] به فضاهای کوچک‌تری افراز می‌شود. این عامل باعث می‌شود که آموزش هر شبکهٔ عصبی با سرعت و دقت بالاتر انجام شود. تاکنون بیشتر مطالعات دسته‌بندهای جمعی در حوزهٔ داده‌کاوی آموزشی بر استفاده از همهٔ داده‌های آموزشی متمرکز بوده‌اند؛

ب) آموزش دسته‌بندهای مبنا از نوع شبکه‌های عصبی چندلایه؛

ج) استفاده از ماژول‌های ترکیب میانگین‌گیری و رأی اکثریت به‌منظور بهبود دقت دسته‌بند جمعی.

نتایج آزمایش‌ها نیز با دسته‌بندهای منفرد معروف مانند درخت تصمیم، نزدیک‌ترین همسایگی، «نایو بیز»، ماشین بردار پشتیبان و همچنین، دو روش مشهور جمعی شامل «بگینگ»[2] و بوستینگ»[3] درخت تصمیم مقایسه شده است.

در ادامه، ساختار پژوهش به این صورت است: در بخش ۲، پیشینهٔ پژوهش و در واقع، مروری بر مطالعات دسته‌بندهای جمعی در حوزهٔ داده‌کاوی آموزشی ارائه شده است. در بخش ۳، روش پژوهش که روش جمعی پیشنهادی است، به همراه آزمایشات تجربی و اجرای مدل شرح داده شده است. در بخش ۴، نتایج ارزیابی شده و آزمایش‌ها در قالب نمودار به تصویر کشیده شده‌اند. بخش ۵، شامل نتیجه‌گیری نهایی از یافته‌های پژوهش است. در بخش ۶، نیز پیشنهادها برای پژوهش‌های آتی ارائه شده‌اند.

۲. پیشینهٔ پژوهش

مطالعات زیادی در حیطهٔ دسته‌بندهای جمعی صورت گرفته است. به‌عنوان مثال، می‌توان به مطالعات زیر اشاره کرد: در مطالعهٔ «الدن» و همکاران روش جمعی «آدابوست»[4]

1. self organized maps (SOM) 2. bagged tree 3. boosted tree 4. Adaboost

بـه همـراه الگوریتـم ژنتیـک ارائـه شـده اسـت. اسـتفاده از تکنیک‌هـای دسـته‌بندی و پیش‌بینـی عملکـرد موجـب شـد تـا ریسـک شکسـت دانشـجویان کاهـش یابـد (ElDen et al. 2013). در مطالعـه‌ای دیگـر «پاریـس، افنـدی و مصطفـی» بـا اسـتفاده از روش رأی اکثریـت نتایـج روش‌هـای منفـرد را بـا یکدیگـر ترکیـب کردنـد. آن‌هـا در نهایـت، نتایـج حاصـل از روش جمعـی را بـا دسـته‌بندهای تکـی بـرای پیش‌بینـی ۲، ۳ و ۴ کلاس بـا یکدیگـر مقایسـه کـرده و نتایـج بهبـود در اسـتفاده از روش جمعـی را نشـان دادنـد (Paris, Affendey and Mustapha 2010). «کورتـز و سـیلوا» آزمایش‌هـای خـود را بـا اسـتفاده از ۴ روش داده‌کاوی ماشـین بـردار پشـتیبان، جنـگل تصادفـی، شـبکه‌های عصبـی، و درخـت تصمیـم و بـر اسـاس ویژگی‌هـای وضعیـت اجتماعـی، نمـرات ریاضـی، جمعیت‌شناسـی، وضعیـت اشـتغال و تحصیـلات والدیـن انجـام دادنـد (Cortez and Silva 2008). «امیـره، هامتینـی و الجـراح» یـک مطالعـهٔ ترکیبـی بـر اسـاس دسـته‌بندهای پایـه و ترکیبـی بـرای پیش‌بینـی کلاس نهایـی و پاییـن آوردن نـرخ شکسـت انجـام دادنـد. دسـته‌بندهای پایـه شـامل «نایـو بیـز»، شـبکه‌های عصبـی و درخـت تصمیـم و دسـته‌بندهای ترکیبـی «بوسـتینگ، بگینـگ» و جنـگل تصادفـی بودنـد (Amrieh, Hamtini and Aljarah 2016). «تریـودی، پـاردوس و هفرنـان» به‌منظـور بهبـود پیش‌بینـی نمـرهٔ دانشـجویان از طریـق روش‌هـای ترکیبـی از خوشـه‌بندی بـرای آمـوزش هـر دسـته‌بند مبنـا اسـتفاده کردنـد. خروجی‌هـا بـا اسـتفاده از روش میانگین‌گیـری بـا یکدیگـر ترکیـب شـدند (Trivedi, Pardos and Heffernan 2011). «هـارا و اوکادا» از طریـق سیسـتم آموزشـی آنلایـن یـک مـدل جمعـی از شـبکه‌های عصبـی را توسـعه دادنـد. آن‌هـا از تعـداد بی‌شـماری شـبکهٔ عصبـی بـا اسـتفاده از روش «بگینـگ» بـرای آمـوزش اسـتفاده کـرده و سـپس، از روش میانگین‌گیـری وزن‌دار بـرای ترکیـب خروجی‌هـا بهـره بردنـد (Hara and Okada 2005). سـاخت سیسـتم جمعـی شـبکه‌های عصبـی بـا اسـتفاده از خوشـه‌بندی در مطالعـهٔ Qiang, Shang-Xu and Sheng-Ying (2005) انجـام شـده اسـت. همچنیـن، اسـتفاده از خوشـه‌بندی ترکیبـی بـرای پیش‌بینـی تـرک تحصیـل دانشـجویان در مطالعـهٔ Iam-On and Boongoen (2017) صـورت گرفتـه اسـت.

«پونلومجیـک» و همـکاران روشـی بـرای بهبـود دسـته‌بندی داده‌هـای نامتـوازن دانشـجویان بـا اسـتفاده از روش‌هـای ترکیبـی رأی اکثریـت، «بگینـگ» و «آدابوسـت» و به‌کارگیـری روش متوازن‌سـازی ارائـه دادنـد. روش جمعـی استفاده‌شـده بـر روی دو دسـته‌بند تکـی درخـت تصمیـم و شـبکه‌های عصبـی اجـرا شـد. روش‌هـای «بگینـگ» و جنـگل تصادفـی بـا دقـت کلـی و صحـت پیش‌بینـی بـالا دارای بهتریـن نتایـج بودنـد (Punlumjeak et al. 2018). رویکـرد پژوهش«چـوا

و فونـگ» اسـتفاده از نمونه‌گیـری و روش‌هـای جنـگل تصادفـی بـرای دسـته‌بندی داده‌هـای نامتـوازن بـود. مطالعـهٔ آن‌هـا بـر اسـاس عملکـرد دانشـگاهی دانش‌آمـوزان انجـام گرفـت و رویکـرد پیشـنهادی در پیش‌بینـی وضعیـت نهایـی دانش‌آمـوزان بسـیار مؤثـر بـوده اسـت (Chau and Phung 2013). در مطالعـهٔ «پانـدی و تارونـا» بـرای پیش‌بینـی عملکـرد دانشـجویان در سـطوح متفـاوت تحصیـلات مهندسـی، سـه دسـته‌بند «نایـو بیـز»، شـبکه‌های عصبـی، و درخـت تصمیـم بـا یکدیگـر ترکیـب شـدند و یـک روش جمعـی جدیـد هیبریـدی بـا اسـتفاده از روش رأی اکثریـت فراهـم شـده اسـت. نتایـج نشـان داد کـه روش جمعـی پیشـنهادی بهتـر و همگون‌تـر از دسـته‌بندهای تکـی عمـل می‌کنـد (Pandey and Taruna 2018). «سـانزانا، گاریـدو و پوبلتـه» تـلاش کردنـد کـه فاکتورهـای عملکـرد و ویژگی‌هـای دانش‌آمـوزان کشـور شـیلی را بـر اسـاس عملکـرد دانشـگاهی آزمایـش و تحلیـل کننـد. جنـگل تصادفـی و درخـت تصمیـم دو روشـی بودنـد کـه بـرای دسـته‌بندی دانش‌آمـوزان بـر اسـاس سـوابق آن‌هـا در درس ریاضـی مـورد اسـتفاده قـرار گرفـت (Sanzana, Garrido and Poblete 2015). «سـوپربی، واندامـه و مسـکنز» مطالعـهٔ خـود را در مـورد دانشـجویان سـال اول دانشـگاه انجـام دادنـد. بیشـترین پارامترهـای مؤثـر در موفقیـت دانشـجویان بـا اسـتفاده از سـه روش جنـگل تصادفـی، درخـت تصمیـم، و شـبکه‌های عصبـی تعییـن و بررسـی شـد. نتایـج حاصـل نشـان داد کـه روش‌هـای جنـگل تصادفـی و شـبکهٔ عصبـی بـه مراتـب روش‌هـای بهتـری هسـتند (Superby, Vandamme and Meskens 2006). در تحقیقـات Kotsiantis, Patriarcheas and Xenos (2010) و Satyanarayana and Nuckowski (2016) نیـز پیش‌بینـی بـا اسـتفاده از روش‌هـای جمعـی صـورت گرفتـه اسـت. «آصـف» و همـکاران در مطالعـهٔ خـود از روش‌هـای داده‌کاوی به‌منظـور مطالعـهٔ عملکـرد دانشـجویان دورهٔ کارشناسـی اسـتفاده کردنـد. روش‌هـای درخـت تصمیـم، شـبکه‌های عصبـی، نزدیک‌تریـن همسـایگی و جنـگل تصادفـی در تحقیـق آن‌هـا بررسـی و آزمایـش شـد (Asif et al. 2017). در همیـن راسـتا Pandey and Taruna (2014) یـک مطالعـهٔ مقایسـه‌ای بیـن روش‌هـای جمعـی «بگینـگ»، «بوسـتینگ» و جنـگل تصادفـی بـرای پیش‌بینـی عملکـرد دانشـجویان انجـام داده‌انـد. (Miguéis et al. 2018) بـا اسـتفاده از اطلاعـات در دسـترس دانشـجویان در انتهـای سـال اول دانشـگاه مدلـی پیشـنهاد دادنـد. جنـگل تصادفـی قوی‌تریـن روش در میـان روش‌هـای ماشـین بـردار پشـتیبان، درخـت تصمیـم، نایـو بیـز، درخـت «بگینـگ» و درخـت «بوسـتینگ» بـود. خلاصـهٔ مطالعـات دسـته‌بندهای جمعـی در حـوزهٔ داده‌کاوی آموزشـی و نـوآوری پژوهـش در جـدول ۱، خلاصـه شـده اسـت.

جدول ۱. خلاصه‌ای از مطالعات دسته‌بندهای جمعی در حوزهٔ داده‌کاوی آموزشی

(نام نویسنده و سال نشر)	دسته‌بندهای پایهٔ سیستم جمعی		روش ایجاد گوناگونی در اعضای سیستم جمعی		روش ترکیب		ویژگی‌های مورد استفاده			
	دسته‌بندهای دیگر	از نوع شبکه عصبی چندلایه	خوشه‌بندی	بگینگ	میانگین‌گیری	رأی اکثریت	تحصیلی	فردی	اجتماعی	خانوادگی
(ElDen et al. 2013)	✓					✓	✓	✓		
(Paris, Affendey and Mustapha 2010)	✓					✓	✓			
(Cortez and Silva 2008)	✓			✓		✓	✓	✓	✓	✓
(Amrieh, Hamtini and Aljarah 2016)	✓	✓		✓		✓	✓	✓	✓	✓
(Trivedi, Pardos and Heffernan 2011)	✓		✓		✓		✓			
(Hara and Okada 2005)		✓		✓	✓		✓			
(Qiang, Shang-Xu and Sheng-Ying 2005)		✓	✓							
(Iam-On and Boongoen 2017)	✓		✓			✓	✓		✓	
(Punlumjeak et al. 2018)	✓	✓		✓		✓	✓			
(Chau and Phung 2013)	✓			✓	✓	✓	✓			
(Pandey and Taruna 2018)	✓	✓				✓	✓			
(Sanzana, Garrido and Poblete 2015)	✓			✓	✓	✓	✓	✓	✓	✓
(Superby, Vandamme and Meskens 2006)	✓			✓		✓	✓			
(Kotsiantis, Patriarcheas and Xenos 2010)	✓					✓	✓	✓		

(نام نویسنده و سال نشر)	دسته‌بندهای پایهٔ سیستم جمعی		روش ایجاد گوناگونی در اعضای سیستم جمعی		روش ترکیب	ویژگی‌های مورد استفاده			
	دسته‌بندهای دیگر	از نوع شبکه عصبی چندلایه	خوشه‌بندی	بگینگ	میانگین گیری	رأی اکثریت	تحصیلی	فردی	اجتماعی	خانوادگی
(Satyanarayana and Nuckowski 2016)	✓			✓		✓	✓	✓	✓	✓
(Asif et al. 2017)	✓	✓		✓		✓	✓		✓	
(Pandey and Taruna 2014)	✓			✓		✓	✓	✓	✓	
(Miguéis et al. 2018)	✓			✓	✓	✓	✓	✓		✓
پژوهش حاضر	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

۳. روش پژوهش

سیستم جمعی شامل سه بخش اصلی است: ۱) خوشه‌بندی و افراز فضای کل داده‌ها ۲) شبکه‌های عصبی منفرد ۳) ماژول‌های ترکیب.

چارچوب سیستم جمعی در دو مرحله، یعنی فاز آموزش[1] و فاز آزمایش[2] ساخته می‌شود. می‌توان گفت، سیستم جمعی پیشنهادی شامل گام‌های زیر است:

گام ۱: از مجموعهٔ داده‌های آموزش D، تعداد K زیرمجموعهٔ مختلف با استفاده از روش خوشه‌بندی ایجاد می‌شود. در این پژوهش از الگوریتم خوشه‌بندی SOM استفاده شده است (Vesanto and Alhoniemi 2000). علت استفاده از این الگوریتم دقت بالا در خوشه‌بندی و بهینه‌بودن خوشه‌هاست؛ به‌نحوی که پس از استفاده از این الگوریتم، داده‌هایی با بیشترین شباهت در یک خوشه قرار می‌گیرند و تجزیه و تحلیل داده‌های کلان بسیار راحت‌تر و دقیق‌تر صورت می‌گیرد. از آنجا که در خوشه‌بندی، به‌دنبال داده‌هایی با مشخصات مشابه و نتایج قوی در تفکیک داده‌ها هستیم، روش خوشه‌بندی SOM می‌تواند داده‌هایی با ویژگی‌های مشابه و یکسان را به‌خوبی در خوشه‌هایی مجزا قرار دهد. گروه‌بندی داده‌ها با استفاده از این الگوریتم قدرتمند، سبب

1. training phase
2. testing phase

می‌شود که به‌جای استفاده از حجم عظیمی از داده‌ها، با گروه‌های مشخصی از داده‌های تمیز و تفکیک‌شده کار کرد. پس از انجام خوشه‌بندی با روش مذکور، داده‌های آموزشی به k جزء به‌صورت $D_1, \dots D_k$ تقسیم می‌شوند و مرحلهٔ بعدی آموزش k شبکهٔ منفرد با استفاده از k خوشه است.

گام ۲: برای هر زیرمجموعهٔ آموزش D_i که $i=1,2,\dots,K$، شبکهٔ عصبی ANN_i[1] با استفاده از این زیرمجموعه آموزش می‌بینند. به این شکل K مدل دسته‌بندی مبنا ایجاد می‌شود.

لازم به ذکر است که هر شبکهٔ عصبی مصنوعی از لایه‌های ورودی، پنهان و خروجی تشکیل شده و هر لایه شامل گروهی از سلول‌های عصبی یا نورون است که با نورون‌های لایه‌های دیگر در ارتباط هستند. یک «پرسپترون»، در واقع، تعدادی از ورودی‌ها را دریافت کرده و پس از تجمیع و اعمال تابع فعال‌سازی، آن‌ها را به لایهٔ خروجی منتقل می‌کند. در این مطالعه، از شبکهٔ عصبی چندلایه استفاده شده است که ورودی‌های شبکه را ویژگی‌های دانشجویان تشکیل داده و خروجی نهایی، کلاس‌ها یا برچسب‌های دسته‌بندی هستند که به‌صورت ۴ گره نمایش داده می‌شوند.

گام ۳: بخشی از داده‌ها برای آموزش کل مجموعه به همهٔ ANN_jها داده می‌شوند و خروجی‌های پیش‌بینی‌شده به‌دست می‌آید. به‌عبارتی، مجموعهٔ داده‌ها به‌عنوان ورودی به آن‌ها اعمال می‌شود و خروجی آن‌ها به شکل برچسب کلاس‌های پیش‌بینی‌شده دریافت می‌گردد و سپس، خروجی‌ها توسط ماژول‌های ترکیب رأی اکثریت (Lam and Suen 1997) و میانگین ساده (Naftaly, Intrator and Horn 1997) با یکدیگر ترکیب می‌شوند. پیش‌بینی برای نمونهٔ i طبق فرمول (۱) است:

$$P^i = [p_1^i, \dots, p_k^i] \qquad i = 1,2, \dots, m \qquad (۱)$$

m تعداد نمونهٔ آموزشی، p_j^i بردار پیش‌بینی برای i امین نمونه توسط j امین MLP[2] است. عملیات به‌ترتیب زیر انجام می‌شوند:

الف) بخشی از داده‌ها به تک‌تک شبکه‌های منفرد داده می‌شود و از هر کدام خروجی گرفته می‌شود.

ب) با تنظیمات بهینه به یک دقت مطلوب در هر شبکه می‌رسیم. هر شبکه در یک

1. artificial neural network (ANN) 2. multi-layer perception

حلقه تعریف می‌شود و آموزش شبکه تا زمانی که به‌دقت مطلوب برسد، ادامه می‌یابد.

ج) در صورت رسیدن به‌دقت مطلوب، آموزش شبکهٔ بعد شروع می‌شود و تا شبکهٔ آخر این روند تکرار می‌شود.

د) بعد از اتمام آموزش تکی هر شبکه توسط ماژول‌های ترکیب دقت آموزش الگوریتم جمعی با استفاده از برچسب پیش‌بینی‌شده الگوریتم ترکیبی و برچسب واقعی داده‌ها به‌دست می‌آید.

سیستم پس از فرایند آموزش الگوریتم ترکیبی به‌منظور پیدا کردن تنظیمات صحیح و کاهش خطا، برای آزمایش نهایی آماده می‌شود.

گام ۴: در مرحلهٔ بعد برای بررسی کارایی و دقت نهایی کل سیستم جمعی این بار، داده‌های تست T به همهٔ دسته‌بندهای پایه داده می‌شوند و خروجی‌هایشان توسط ماژول‌های ترکیب با یکدیگر ترکیب می‌شوند. عملیات به‌ترتیب زیر انجام می‌شوند:

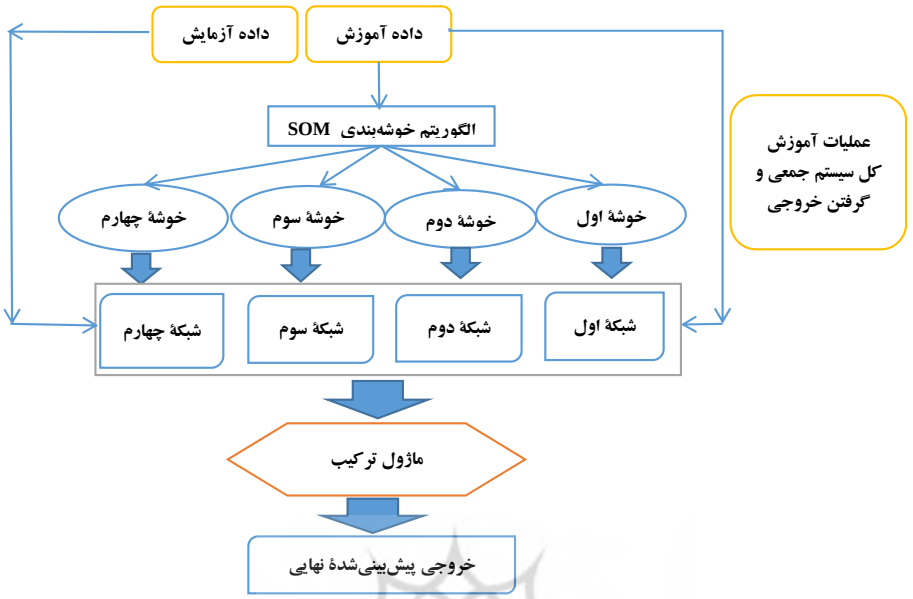
الف) کل داده‌های آزمایش به تک‌تک دسته‌بندهای منفرد داده می‌شود و از هر کدام خروجی گرفته می‌شود.

ب) مطابق حالت قبل هر شبکه در یک حلقه تعریف می‌شود و آزمایش شبکه تا زمانی که به‌دقت مطلوب برسد، ادامه می‌یابد.

ج) در صورت رسیدن به‌دقت مطلوب آزمایش شبکهٔ بعد شروع می‌شود و تا شبکهٔ چهارم این روند تکرار می‌شود.

د) بعد از آزمایش نهایی هر شبکه توسط ماژول‌های ترکیب، دقت نهایی الگوریتم جمعی با استفاده از برچسب پیش‌بینی‌شده الگوریتم ترکیبی و برچسب واقعی داده‌ها به‌دست می‌آید. خروجی ماژول ترکیب پیش‌بینی نهایی الگوریتم جمعی است.

ساختار روش جمعی پیشنهادی در شکل ۱، به تصویر کشیده شده است:



شکل ۱. چارچوب سیستم دسته‌بندی جمعی

۳-۱. آزمایش‌های تجربی

این پژوهش به‌صورت یک مطالعهٔ موردی در «دانشگاه قم» انجام گرفته است. به این صورت که از حدود ۴۸۰۰ دانشجو در مقطع کارشناسی رشته‌های مختلف تحصیلی تحقیق به عمل آمده است. بدین منظور، ابتدا اطلاعات در دسترس مربوط به دانشجویان را از بانک اطلاعاتی دانشگاه به همراه فیلدهای اطلاعاتی جمع‌آوری کردیم. مطالعهٔ حاضر بر روی نرم‌افزار «متلب»[1] ۲۰۱۶ اجرا شده است. همچنین، سیستم مورد استفاده ASUS Core i7 است. همان‌طور که در بخش روش پژوهش ذکر شد، نتایج به‌دست‌آمده پس از تکرار اجراهای متعدد مدل است و به‌عبارتی، دقت نهایی و بهینه حاصل اجرای چندبارهٔ مدل جمعی است.

پس از پیش پردازش داده‌ها، سرانجام، ۴۸۱۶ دانشجو برای انجام داده‌کاوی نهایی تحقیق انتخاب شدند. ویژگی‌های مورد استفاده در این تحقیق سال ورود، سن یا تاریخ تولد، جنسیت، بومی بودن، رشته، دورهٔ تحصیلی، تعداد مشروطی، تعداد مشروطی متوالی، نوع ورود به دانشگاه، درآمد خانواده، وضعیت تأهل، تحت پوشش کمیته و بهزیستی، تعداد واحد مردودی، تعداد ترم مرخصی و معدل کل به‌عنوان ویژگی‌های مورد تخمین

1. MATLAB

است. از آنجا که ویژگی‌ها انواع مختلفی (اسمی، عددی، پیوسته و گسسته) دارند، همهٔ ویژگی‌های اسمی مورد استفاده در تحقیق به مقادیر عددی تبدیل شده و مقدار گرفتند. مقادیر عددی نیز بدون تغییر باقی ماندند. به این صورت که تحقیق بر روی دانشجویان ورودی ۸۵ تا ۹۵ و متولدین ۵۶ تا ۷۸ انجام گرفت و به‌دلیل عددی بودن بدون تغییر و تبدیل باقی ماندند. همچنین، تعداد مشروطی متناوب ۰ تا ۵، تعداد مشروطی متوالی ۰ تا ۴، تعداد واحدهای مردودی ۰ تا ۵۷ و تعداد ترم مرخصی که از ۰ تا ۲ موجود هستند نیز بدون تغییر باقی ماندند. از طرفی، مقادیر کیفی و اسمی به مقادیر عددی تبدیل شدند و از صفر مقدار گرفتند. به این صورت که ویژگی‌های جنسیت، بومی بودن، وضعیت تأهل، تحت پوشش کمیته و بهزیستی ۰ و ۱، دورهٔ تحصیلی از ۰ تا ۶، نوع ورود به دانشگاه از ۰ تا ۷ و رشته‌های تحصیلی از ۱ تا ۲۶ و میزان درآمد خانواده از ۰ تا ۴ مقدار گرفتند. پس از تبدیلات، تمامی ویژگی‌ها نرمال شده و در بازهٔ (۱ و ۰) تعریف شدند. وضعیت دانشجو در هر کلاس معدل در جدول ۲، تعریف شده است:

جدول ۲. وضعیت دانشجو در هر کلاس معدل

دستهٔ معدل	وضعیت دانشجو
<۱۲	مشروط / ضعیف
۱۲-۱۴	متوسط
۱۴-۱۶	خوب
۱۶-۲۰	خیلی خوب

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

۳-۲. اجرای مدل

تعداد خوشه‌ها برای آزمایش، K در نظر گرفته می‌شود که برابر با تعداد شبکه‌های عصبی ANN است و با تعداد ۴ خوشه (K=4) آزمایش‌ها انجام شده است. به‌منظور رسیدن به عدد ۴ در تعداد خوشه و دسته‌بند، آزمایش‌های متعددی صورت گرفت. به‌عبارتی، با استفاده از روش خوشه‌بندی بهینه به عدد ۴ رسیدیم و نتایج نشان داد که این تعداد خوشه، بهترین تمایز را میان داده‌های هر خوشه ایجاد می‌کند. تمایز میان داده‌ها سبب شده است که داده‌های هر خوشه بیشترین شباهت را به یکدیگر داشته باشند و عملیات خوشه‌بندی و آموزش و آزمایش سیستم جمعی راحت‌تر و با نتایج مطمئن‌تری صورت بگیرند.

داده‌های دانش‌آموزان با ۱۵ ویژگی به‌عنوان ورودی به شبکه‌های عصبی داده می‌شوند و شبکه‌ها با تعداد نورون‌های لایهٔ پنهان که به‌وسیلهٔ کاربر تعیین می‌شود، اجرا می‌شوند. پارامترهای مهم برای شبکه‌های عصبی میانگین مربع خطا، گرادیان و تعداد نورون‌های لایهٔ پنهان هستند. لایهٔ ورودی هر شبکه ۱۵ گره و لایهٔ خروجی آن ۴ گره است که خروجی برای هر داده به‌صورت یک بردار ۴تایی تعریف می‌شود که عددی بین صفر و یک است و احتمال تعلق هر داده به هر کلاس را مشخص می‌کند.

لازم به ذکر است که الگوریتم‌های مورد استفاده در مقاله، کدنویسی شده و از تولباکس[1] استفاده شده است.

از مجموعهٔ کل داده‌های ۴۸۱۶تایی، ۳۴۰۰ داده برای خوشه‌بندی و آموزش انتخاب شده‌اند. داده‌ها در ۴ خوشه افراز شدند که هر خوشه برای آموزش هر دسته‌بند منفرد شبکهٔ عصبی استفاده شده است. تعداد ۶۰۰ داده برای آموزش کل سیستم جمعی انتخاب شدند و ۸۱۶ داده برای آزمایش نهایی و کارایی دقت کل سیستم جمعی کنار گذاشته شدند. تعداد کل داده‌ها در هر کلاس پس از پیش پردازش نهایی و افراز داده‌ها در جداول ۳ و ۴، ارائه شده است.

جدول ۳. تعداد کل داده‌ها در هر کلاس

	دستهٔ معدل	کلاس معدل	تعداد داده‌ها
معدل کل	۱۶-۲۰	۴	۱۴۶۸
	۱۴-۱۶	۳	۱۶۸۴
	۱۲-۱۴	۲	۱۲۳۲
	<۱۲	۱	۴۳۲

جدول ٤. افراز داده‌ها با روش SOM

خوشه‌ها	تعداد داده‌ها
C1	۴۱۰
C2	۴۷۸
C3	۱۴۴۸
C4	۱۰۶۴

1. toolbox

٤. تجزیه و تحلیل یافته‌ها

در آزمایش، مجموعهٔ بهینهٔ پارامترها به‌صورت میانگین مربع خطا ۰/۰۰۷، گرادیان ۰/۰۰۱ و بیشترین تعداد نورون‌های لایهٔ پنهان ۲۰ نورون تعیین شدند. توابع انتقال، تعداد دورهای شبکه و گرادیان نیز طی آزمایش‌های بسیار مشخص شدند. به‌دلیل صرفه‌جویی در زمان و نتایج با دقت و کارایی بالاتر، از تابع آموزش Trainscg استفاده شده است. آزمایش‌ها چندین بار تکرار و بهترین نتایج ثبت شده‌اند. بهترین دقت نهایی سیستم جمعی با استفاده از دو روش میانگین‌گیری و رأی اکثریت به‌ترتیب ۶۹ و ۷۰ شدند.

به‌منظور ارزیابی نتایج آزمایش‌ها از سه معیار کارایی متفاوت به نام‌های صحت[1]، یادآوری[2] و معیار توأمان[3] نیز استفاده شد. در ادامه، در مورد این سه معیار توضیحاتی داده شده است.

- **معیار صحت**

درصدی از پیش‌بینی‌های درست را به ما نشان می‌دهد که در دنیای واقعی نیز درست بوده‌اند.

- **معیار یادآوری**

در واقع، درصد پیش‌بینی‌های درست را در کل داده‌ها نشان می‌دهد. به‌عبارتی، از کل پیش‌بینی‌های انجام‌شده، داده‌هایی که درست پیش‌بینی می‌شوند، نمایانگر معیار یادآوری هستند.

- **معیار توأمان**

به‌عنوان میانگینی از دو مقدار صحت و یادآوری، معیاری دقیق‌تر از کل جامعهٔ آماری را نشان می‌دهد و معیاری مناسب برای ارزیابی است.

به‌منظور ارزیابی بهتر عملکرد سیستم جمعی پیشنهادی، نتایج خروجی با دیگر روش‌های معروف دسته‌بندی مقایسه شده‌اند. این روش‌ها عبارت‌اند از: درخت تصمیم، «نایو بیز»، نزدیک‌ترین همسایگی، و ماشین بردار پشتیبان. نتایج را در جداول ۵ تا ۷ مشاهده می‌کنید:

1. precision 2. recall 3. F-value

جدول ۵. مقایسهٔ صحت چهار دسته‌بند مستقل با سیستم جمعی

Classifier / Class	<۱۲	۱۲-۱٤	۱٤-۱٦	۱٦-۲۰
[1]SVM	۲۵/۶	۶۷/۸	۲۹/۸	۹۲/۲
[2]DT	۵۶/۱	۵۹/۵	۴۹/۶	۶۰/۷
[3]KNN	۳۵/۴	۶۰	۵۲/۶	۵۳/۷
[4]NB	۱۰/۲	۵۹/۵	۲۳/۵	۶۸/۹
[5]Ensemble ANN (Voting)	۸۱/۳	۶۹/۴	۶۲/۹	۷۴
Ensemble ANN (Averaging)	۸۳/۳	۶۹/۸	۶۲/۷	۷۱/۸

همان‌طور که در جدول ۵، مشاهده می‌شود صحت مدل جمعی شبکه‌های عصبی با رأی اکثریت و میانگین‌گیری در هر ۴ کلاس به‌غیر از کلاس چهارم دسته‌بندهای ماشین بردار پشتیبان عملکرد بالاتری را نشان می‌دهد. دقت دسته‌بندی بالای مدل‌های جمعی کلاس اول در مقایسه با دسته‌بند «نایو بیز» حاکی از عملکرد فوق‌العادهٔ مدل جمعی پیشنهادی در توانایی دسته‌بندی با وجود نمونه‌های بسیار کم است.

جدول ٦. مقایسه یادآوری چهار دسته‌بند مستقل با سیستم جمعی

Classifier / Class	<۱۲	۱۲-۱٤	۱٤-۱٦	۱٦-۲۰
SVM	۹۵/۵	۵۹/۴	۵۴/۴	۵۷/۷
DT	۷۶/۷	۶۳/۲	۴۴/۹	۵۹/۵
KNN	۵۵/۸	۵۶/۴	۴۶/۶	۵۷/۷
NB	۹۰	۵۳/۳	۳۳/۷	۴۴/۷
Ensemble ANN (Voting)	۶۵	۷۲/۳	۵۹/۱	۸۰/۶
Ensemble ANN (Averaging)	۵۵	۶۳/۸	۶۳/۶	۸۳/۶

با توجه به جدول ۶، در مدل جمعی با رأی اکثریت و میانگین‌گیری معیار یادآوری در کلاس‌های دوم، سوم و چهارم نسبت به دسته‌بندهای منفرد برتری کاملاً محسوسی دارد.

1. Support vector machine
2. Decision tree
3. K-nearest neighbors
4. Naive Bayes
5. Ensemble artificial neural network

همچنین، یادآوری مدل جمعی رأی اکثریت نسبت به کلاس اول دسته‌بند نزدیک‌ترین همسایگی بهتر عمل کرده است.

جدول ۷. مقایسهٔ توأمان چهار دسته‌بند مستقل با سیستم جمعی

Classifier / Class	<۱۲	۱۲-۱٤	۱٤-۱٦	۱٦-۲۰
SVM	۴۰/۳۸	۶۳/۳۲	۳۸/۵۱	۷۰/۹۸
DT	۶۴/۸	۶۱/۲۹	۴۷/۱۳	۶۰/۰۹
KNN	۴۳/۳۲	۵۸/۱۴	۴۹/۴۲	۵۵/۶۳
NB	۱۸/۳۲	۵۶/۲۳	۲۷/۶۹	۵۴/۲۲
Ensemble ANN (Voting)	۷۲/۲۴	۷۰/۸۲	۶۰/۹۴	۷۷/۱۶
Ensemble ANN (Averaging)	۶۶/۲۵	۶۶/۶۶	۶۳/۱۴	۷۷/۲۵

مطابق جدول ۷، با در نظر گرفتن اهمیت «صحت» و «یادآوری» به‌طور توأمان در مدل جمعی رأی اکثریت و میانگین‌گیری این معیار مهم در همهٔ کلاس‌ها نسبت به دسته‌بندهای منفرد بالاترین است. همچنین، معیار «توأمان» دو مدل پیشنهادی نسبت به «نایوبیز» در کلاس اول با وجود نمونه‌های کم، عملکرد بسیار عالی داشته است.

پس از تحلیل و ارزیابی معیارهای ۴ دسته‌بند مستقل، این بار دسته‌بندها با روش رأی اکثریت ترکیب شده و از نظر دقت کلی، با مدل‌های جمعی پیشنهادی مقایسه شده‌اند. معیار دقت[1]، در واقع نشان می‌دهد که مدل تا چه اندازه خروجی را درست پیش‌بینی می‌کند. نتایج ترکیب روش‌ها با استفاده از رأی اکثریت[2] و مقایسه با مدل‌های پیشنهادی در جداول ۸ تا ۱۱ ارائه شده‌اند، همچنین، در جدول ۸، مقایسه به دو صورت دقت آموزش[3] الگوریتم و دقت آزمایش[4] انجام شده است:

جدول ۸. مقایسهٔ دقت چهار روش منفرد با ترکیب روش‌ها

طبقه‌بند	دقت آموزش	دقت آزمایش
SVM	۶۱/۴۴	۵۸/۵
DT	۷۱	۵۶/۲۵

1. accuracy
2. ensemble voting
3. train accuracy
4. test accuracy

طبقه‌بند	دقت آموزش	دقت آزمایش
KNN	۷۰/۵۲	۵۳/۰۶
NB	۵۰	۴۴/۰۶
Ensemble voting	۷۵	۶۰/۵

با توجه به جدول ۸، ترکیب ۴ دسته‌بند منفرد با روش رأی اکثریت دقت آموزش و آزمایش بالاتری را نشان می‌دهد.

جدول ۹. مقایسهٔ صحت چهار روش ترکیبی با سیستم جمعی

Class / Classifier	<۱۲	۱۲-۱٤	۱٤-۱٦	۱٦-۲۰
Ensemble Voting	۳۹	۶۹/۳	۴۷/۴	۶۵/۴
Ensemble ANN (Voting)	۸۱/۳	۶۹/۴	۶۲/۹	۷۴
Ensemble ANN (Averaging)	۸۳/۳	۶۹/۸	۶۲/۷	۷۱/۸

طبق جدول ۹، «صحت» مدل جمعی با رأی اکثریت و میانگین‌گیری در مقایسه با ترکیب ۴ روش منفرد در کلاس اول برتری چشمگیری دارد و با اختلاف، عملکرد بسیار بهتری داشته است. همچنین، این برتری در کلاس‌های دوم، سوم و چهارم نیز حفظ شده است.

جدول ۱۰. مقایسهٔ یادآوری چهار روش ترکیبی با سیستم جمعی

Class / Classifier	<۱۲	۱۲-۱٤	۱٤-۱٦	۱٦-۲۰
Ensemble Voting	۸۶/۵	۶۳	۵۱/۷	۶۰/۶
Ensemble ANN (Voting)	۶۵	۷۲/۳	۵۹/۱	۸۰/۶
Ensemble ANN (Averaging)	۵۵	۶۳/۸	۶۳/۶	۸۳/۶

همان‌طور که در جدول ۱۰، مشاهده می‌شود، معیار «یادآوری» دو مدل پیشنهادی در کلاس‌های دوم، سوم و چهارم نسبت به روش ترکیبی، برتری کاملاً محسوسی دارند.

جدول ۱۱. مقایسهٔ «توأمان» چهار روش ترکیبی با سیستم جمعی

Class / Classifier	<۱۲	۱۲-۱٤	۱٤-۱٦	۱٦-۲۰
Ensemble Voting	۵۳/۷۶	۶۶	۴۹/۴۶	۶۲/۹۱
Ensemble ANN (Voting)	۷۲/۲۴	۷۰/۸۲	۶۰/۹۴	۷۷/۱۶
Ensemble ANN (Averaging)	۶۶/۲۵	۶۶/۶۶	۶۳/۱۴	۷۷/۲۵

طبق جدول ۱۱، با در نظر گرفتن اهمیت «صحت» و «یادآوری» به‌طور توأمان مدل جمعی رأی اکثریت و میانگین‌گیری در همهٔ کلاس‌ها به‌طرز چشمگیری بالاترین مقدار را داشته است.

به‌منظور ارزیابی کارایی استراتژی ساخت سیستم‌های جمعی، از دو روش مشهور «بگینگ» درخت تصمیم یا جنگل تصادفی و درخت «بوستینگ» در ادامه آزمایش‌ها استفاده کرده‌ایم. در روش «بگینگ» دسته‌بند جمعی با استفاده از دسته‌بندهایی که بر روی مجموعه‌های دادهٔ گزینش‌شده به روش «بوت‌استرپ»[1] آموزش می‌بینند، ساخته می‌شود و خروجی نهایی به‌وسیلهٔ رأی اکثریت تعیین می‌گردد. در روش جنگل تصادفی دسته‌بندها درختان تصمیم هستند. در روش «بوستینگ»، دسته‌بند جمعی به‌طور افزایشی ساخته می‌شود و در هر بار یک دسته‌بند مبنا به سیستم جمعی اضافه می‌شود. دسته‌بندی که در هر گام به سیستم اضافه می‌شود با یک مجموعه داده‌های گزینش‌شده از مجموعهٔ داده‌های اصلی آموزش می‌بینند که دارای نمونه‌هایی است که دسته‌بندی آن‌ها دشوارتر از نمونه‌های مجموعهٔ داده‌های گام قبلی است. در این پژوهش از یک الگوریتم معروف در روش «بوستینگ» به نام «آدابوست» استفاده شده است. در روش درخت «بوستینگ» مدل نهایی از مجموعه‌ای از مدل‌ها تشکیل شده است که در آن مدل‌های پایه مبتنی بر درختان تصمیم‌گیری هستند.

نتایج دو روش مشهور در مقایسه با مدل‌های پیشنهادی در جداول ۱۲ تا ۱۴ خلاصه شده‌اند.

1. Bootstrap

جدول ۱۲. مقایسهٔ «صحت» دو روش مشهور جمعی با سیستم جمعی

Classifier \ Class	<۱۲	۱۲-۱٤	۱٤-۱٦	۱٦-۲۰
RF[1]	۵۴/۹	۷۱/۷	۶۱	۷۶/۳
Boosted Tree	۵۱/۲	۷۱/۲	۵۶/۳	۸۳/۳
Ensemble ANN (Voting)	۸۱/۳	۶۹/۴	۶۲/۹	۷۴
Ensemble ANN (Averaging)	۸۳/۳	۶۹/۸	۶۲/۷	۷۱/۸

همان‌طور که در جدول ۱۲، مشاهده می‌شود، «صحت» مدل جمعی شبکه‌های عصبی با رأی اکثریت و میانگین‌گیری در کلاس‌های اول و سوم عملکرد بالاتری را نشان می‌دهند. همچنین، «صحت» دو مدل جمعی در مقایسه با دو مدل مشهور جنگل تصادفی و درخت «بوستینگ» در کلاس اول برتری چشمگیری دارد و با اختلاف، عملکرد بسیار بهتری داشته است. همچنین، این معیار در مقایسه با دو روش مشهور در کلاس دوم اختلاف بسیار ناچیزی با مقادیر بهینه دارد.

جدول ۱۳. مقایسهٔ «یادآوری» دو روش مشهور جمعی با سیستم جمعی

Classifier \ Class	<۱۲	۱۲-۱٤	۱٤-۱٦	۱٦-۲۰
RF	۸۳/۳	۶۶/۲	۶۰/۱	۷۴/۲
Boosted Tree	۸۷/۵	۶۴/۹	۶۲/۴	۷۱/۸
Ensemble ANN (Voting)	۶۵	۷۲/۳	۵۹/۱	۸۰/۶
Ensemble ANN (Averaging)	۵۵	۶۳/۸	۶۳/۶	۸۳/۶

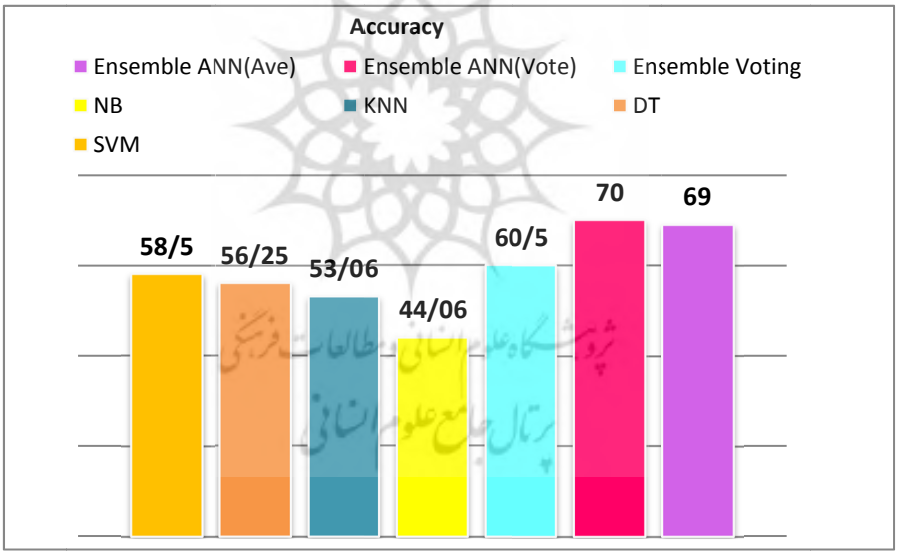
با توجه به جدول ۱۳، معیار «یادآوری» مدل جمعی رأی اکثریت در کلاس‌های دوم و چهارم و مدل میانگین‌گیری در کلاس‌های سوم و چهارم نسبت به دو مدل مشهور برتری کاملاً محسوسی دارند. معیار «یادآوری» (مدل جمعی رأی اکثریت)، اگرچه در کلاس سوم بهترین مقدار نیست، اما با اختلاف بسیار اندک به بهترین مقدار دو مدل نزدیک است.

1. random forest

جدول ١٤. مقایسهٔ «توأمان» دو روش مشهور جمعی با سیستم جمعی

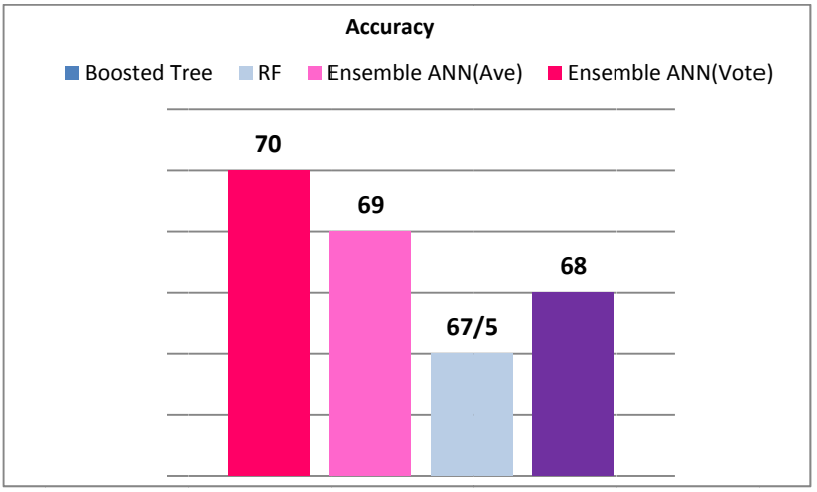
Classifier \ Class	<۱۲	۱۲-۱٤	١٤-١٦	١٦-٢٠
RF	۶۶/۱۸	۶۸/۸۴	۶۰/۵۵	۷۵/۲۴
Boosted Tree	۶۴/۶	۶۷/۹	۵۹/۱۹	۷۷/۱۲
Ensemble ANN (Voting)	۷۲/۲۴	۷۰/۸۲	۶۰/۹۴	۷۷/۱۶
Ensemble ANN (Averaging)	۶۶/۲۵	۶۶/۶۶	۶۳/۱۴	۷۷/۲۵

طبق جدول ۱۴، معیار «توأمان» در مدل جمعی رأی اکثریت در تمام کلاس‌ها بالاترین است. همچنین، این معیار در مدل جمعی میانگین‌گیری نیز در کلاس اول، سوم و چهارم عملکرد بهتری داشته است. اگرچه در کلاس دوم مدل جمعی میانگین‌گیری مقدار «توأمان» بهینه و بالاترین نیست، اما به مقدار بهینه بسیار نزدیک است.



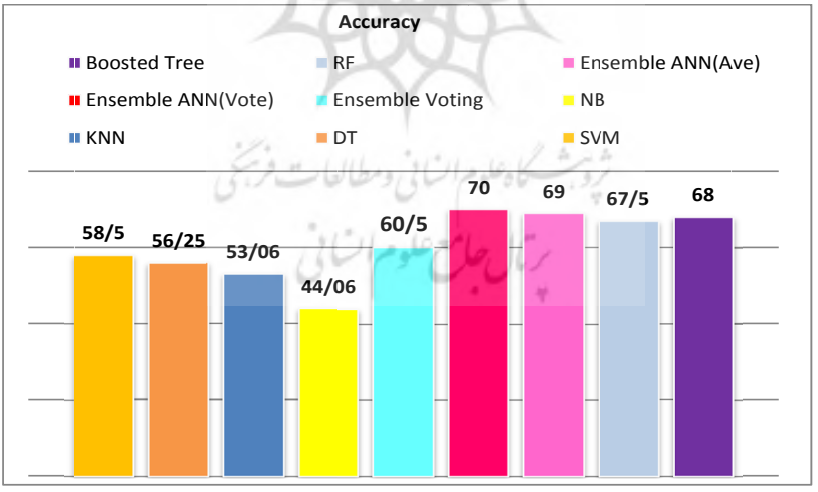
شکل ۲. مقایسهٔ دقت کلی ٤ روش منفرد و ترکیبی با سیستم‌های جمعی

با توجه به شکل ۲، اگر بخواهیم دقت کلی را در هر ۴ کلاس بر روی داده‌های آزمایشی بررسی کنیم، مدل جمعی رأی اکثریت و میانگین‌گیری با اختلاف زیاد عملکرد بهتری را نسبت به ۴ روش منفرد نشان می‌دهند. همچنین، کارایی الگوریتم جمعی نسبت به دقت ترکیب ۴ روش منفرد نیز کاملاً محسوس است.



شکل ۳. مقایسهٔ دقت کلی دو روش مشهور با سیستم‌های جمعی

با مقایسهٔ دقت کلی سیستم‌های جمعی با دو روش «بوستینگ» درخت تصمیم و جنگل تصادفی نیز مطابق شکل ۳، متوجه می‌شویم که در دسته‌بندی معدل، مدل پیشنهادی با وجود دقت بالا و عملکرد خوب این دو روش مشهور بسیار دقیق‌تر عمل کرده است.



شکل ٤. مقایسهٔ دقت کلی همه دسته‌بندها

در یک نتیجه‌گیری نهایی می‌توان گفت که مدل جمعی پیشنهادی دارای عملکرد بهتری نسبت به همهٔ دسته‌بندهای مورد استفاده در آزمایش‌ها دارد. شکل ۴، دقت کل

دسته‌بندها را نشان می‌دهد که نمایش عملکرد آن‌ها می‌تواند به‌عنوان سیستم تخمین معدل برای دانشجویان کارشناسی باشد.

۵. نتیجه‌گیری

امروزه، بررسی وضعیت تحصیلی و عملکرد آموزشی دانشجویان به‌منظور پیشگیری از شکست و ارتقای سطح کیفی عملکرد آن‌ها یک دغدغهٔ مهم در حوزهٔ داده‌کاوی آموزشی است. یادگیری جمعی به‌عنوان یک روش جدید در انواع مسائل دسته‌بندی و حوزه‌های گوناگون، به‌ویژه در داده‌کاوی آموزشی می‌تواند روشی مؤثر باشد.

در این پژوهش یک دسته‌بند جمعی جدید با استفاده از شبکه‌های عصبی چندلایه و بر پایهٔ خوشه‌بندی و ویژگی‌های دانشجویان ارائه شد. مطالعهٔ حاضر به‌عبارتی، یک سیستم جمعی برای دسته‌بندی و تخمین معدل کل دانشجویان مقطع کارشناسی با استفاده از نرم‌افزار «متلب» و الگوریتم‌های خوشه‌بندی و دسته‌بندی ارائه داده است که کارایی بهتر و نتایج دقیق‌تری در مقایسه با روش‌های دسته‌بندی موجود دارد. با توجه به این‌که تعداد نمونه‌ها در دو کلاس اول که معدل‌های زیر ۱۲ هستند و کلاس دوم که بیانگر معدل‌های بین ۱۲ و ۱۴ است، بسیار کمتر از دو کلاس دیگر است، تشخیص و تخمین معدل در این دو کلاس اهمیت بسیار بالایی پیدا می‌کند. همچنین، پیش‌بینی و تخمین معدل‌های بالا که جزء کلاس چهارم قرار می‌گیرند نیز در مواردی دارای اهمیت است. می‌توان گفت که سه معیار ذکرشده در مدل‌های پیشنهادی عملکرد بسیار خوبی در همهٔ کلاس‌ها، به‌خصوص کلاس اول و دوم داشته‌اند که این نشان‌دهندهٔ کارایی مدل جمعی در تخمین معدل دانشجویان کارشناسی است.

پیش‌بینی و تخمین دانشجویانی با معدل‌های زیر ۱۲ که در دستهٔ مشروط قرار می‌گیرند، به دانشگاه‌ها و اساتید و به‌خصوص شخص دانشجو کمک می‌کند که زودتر از این‌که مشروطی اتفاق بیافتد، با اصلاح شیوهٔ درس خواندن، غلبه بر شرایط محیطی و اجتماعی، استفاده از کلاس‌های کمکی و در مورد اساتید، تغییر روش تدریس و مواردی از این قبیل را پیش بگیرند.

از آنجا که مطالعهٔ حاضر در صدد ارائهٔ یک روش دسته‌بندی جمعی برای بررسی و ارزیابی عملکرد تحصیلی دانشجویان مقطع کارشناسی است، می‌توانیم بگوییم که در مقاله، یک نوآوری در این زمینه ارائه شده است.

٦. پیشنهادهایی برای آینده

از جمله زمینه‌های تحقیقاتی آینده که می‌توان به آن اشاره کرد، تغییر روش ترکیب، یعنی استفاده از مکانیزم وزن‌دهی در دو ماژول رأی اکثریت و میانگین‌گیری است. استفاده از دسته‌بندهای پایه متنوع و تغییر روش خوشه‌بندی و همچنین، استفاده از داده‌های مرتبط با اساتید و نمرات درسی دانشجویان نیز می‌تواند مورد پژوهش و تحقیق قرار گیرد. بررسی جامعه‌ای خاص از دانشجویان، به‌عنوان مثال تحقیق و پژوهش بر روی دانشجویان رشته‌های مهندسی و یا دانشجویان مقطع کارشناسی ارشد و دکتری نیز از جمله مواردی است که در مطالعات آتی می‌تواند مطرح شود.

قدردانی

در انتهای مطلب بر خود لازم می‌دانیم از مسئولان حوزهٔ انفورماتیک دانشگاه سراسری قم از بابت در اختیار گذاشتن اطلاعات مرتبط با دانشجویان تشکر نماییم.

References

Agarwal, S., G. N. Pandey, and M. D. Tiwari. 2012. Data mining in education: data classification and decision tree approach. *International Journal of e-Education, e-Business, e-Management and e-Learning* 2 (2): 140.

Aluko, R. O., O. A. Adenuga, P. O. Kukoyi, A. A. Soyingbe, and J. O. Oyedeji. 2016. Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques. *Construction Economics and Building* 16 (4): 86-98.

Amini, M., & J. Rezaeenour. 2014. Early Fraud Detection for Online Auctions Using A Multiple-phased Modeling Method with a Neural Networks Ensemble Classifier. *Applied mathematics in Engineering, Management and Technology*?: 560-567.

Amini, M., J. Rezaeenour, & E. Hadavandi. 2016. A neural network ensemble classifier for effective intrusion detection using fuzzy clustering and radial basis function networks. *International Journal on Artificial Intelligence Tools* 25 (02): 1550033.

_____. 2014. Effective intrusion detection with a neural network ensemble using fuzzy clustering and stacking combination method. *Journal of Computing and Security*, 1 (4): 293-305.

_____. 2015. A cluster-based data balancing ensemble classifier for response modeling in Bank Direct Marketing. *International Journal of Computational Intelligence and Applications* 14 (04): 1550022.

Amrieh, E. A., T. Hamtini, & I. Aljarah. 2016. Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application* 9 (8): 119-136.

Araque, F., C. Roldn, and A. Salguero. 2009. Factors inuencing university dropout rates. *Computers & Education* 53 (3):563-574.

Asif, R., A. Merceron, S. A. Ali, & N. G. Haider. 2017. Analyzing undergraduate students' performance using educational data mining. *Computers & Education* 113: 177-194.

Bydžovská, H. 2016. Course Enrollment Recommender System. In 9th *International Conference on Educational Data Mining (EDM), Raleigh, USA, Jun 29-Jul 2, 2016.,*

Chau, V. T. N., & N. H. Phung. 2013, November. Imbalanced educational data classification: An effective approach with resampling and random forest. *In The 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF)* (pp. 135-140). IEEE. Hanoi, Vietnam

Cortez, P., & A. M. G. Silva. 2008. Using data mining to predict secondary school student performance. In Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008, p. 5-12.

Dzeroski, S., & B. Zenko. 2002. Is combining classifiers better than selecting the best one? *In International Conference on Machine Learning (ICML)* (Vol. 2002, p. 123e30).

ElDen, A. S., M. A. Moustafa, H. M. Harb, & A. H. Emara. 2013. AdaBoost ensemble with simple genetic algorithm for student prediction model. *International Journal of Computer Science & Information Technology* 5 (2): 73.

Gray, G., C. McGuinness, and P. Owende. 2014. An application of classi_cation models to predict learner progression in tertiary education. *In Advance Computing Conference (IACC), 2014 IEEE International*, pages. In *2014 IEEE International Advance Computing Conference (IACC)* urgaon, India, (pp. 549-554).

Guruler, H., A. Istanbullu, and M. Karahasan. 2010. A new student performance analyzing system using knowledge discovery in higher educational databases. *Computers & Education* 55 (1): 247-254.

Hansen, L. K., & P. Salamon. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 10: 993-1001.

Hara, K., & M. Okada. 2005. Ensemble learning of linear perceptrons: on-line learning theory. *Journal of the Physical Society of Japan* 74 (11): 2966-2972.

Ho, T. K., J. J. Hull, & S. N. Srihari. 1994. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1: 66-75.

Hoffait, A.-S., and M. Schyns. 2017. Early detection of university students with potential diculties. *Decision Support Systems 101*: 1-11.

Huang, C.T., W. T. Lin, S. T. Wang, and W. S. Wang. 2009. Planning of educational training courses by data mining: Using China Motor Corporation as an example. *Expert Systems with Applications* 36 (3): 7199-7209.

Iam-On, N., & T. Boongoen. 2017. Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics* 8 (2): 497-510.

Ji, C., & S. Ma. 1997. Combinations of weak classifiers. *In Advances in Neural Information Processing Systems* 9: 494-500.

Kittler, J., M. Hatef, R. P. W. Duin, & J. Matas. 1998. On combining classifiers, *IEEE transactions on pattern analysis and machine intelligence* 20 (3): 226-239.

Kotsiantis, S., K. Patriarcheas, & M.enos. 2010. A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems* 23 (6): 529-535.

Krogh, A., & J. Vedelsby. 1995. Neural network ensembles, cross validation, and active learning. *In Advances in neural information processing systems* 7: 231-238.

Lam, L., & S. Y. Suen. 1997. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 27 (5): 553-568.

Laugerman, M., D. Rover, M. Shelley, and S. Mickelson. 2015. Determining Graduation Rates in Engineering for Community College Transfer Students Using Data Mining. *International Journal of Engineering Education* 31 (6): 1448-1457.

Lykourentzou, I., I. Giannoukos, G. Mpardis, V. Nikolopoulos, & V. Loumos. 2009. Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology* 60 (2): 372–380.

M_arquez-Vera, C., A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura. 2016. Early dropout prediction using data mining: a case study with high school students. *Expert Systems* 33 (1): 107-124.

Miguéis, V. L., A. Freitas, P. J. Garcia, & A. Silva. 2018. Early segmentation of students according to their academic performance: *A predictive modelling approach. Decision Support Systems* 115: 36-51.

Moschopoulos, C. N., P. Tsiatsis, G. N. Beligiannis, D. Fotakis, & S. D. Likothanassis. 2009. Dealing with large datasets using an artificial intelligence clustering tool. *In Tools and Applications with Artificial Intelligence* (pp. 105-120). Berlin, Heidelberg: Springer.

Mueen, A., B. Zafar, and U. Manzoor. 2016. Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *International Journal of Modern Education and Computer Science,* 8 (11): 36-55.

Naftaly, U., N. Intrator, & D. Horn. 1997. Optimal ensemble averaging of neural networks. *Network: Computation in Neural Systems* 8 (3): 283-296.

Natek, S. and M. Zwilling. 2014. Student data mining solution–knowledge management system related to higher education institutions. *Expert systems with applications* 41 (14): 6400-6407.

Oza, N. C., & K. Tumer. 2001. Input decimation ensembles: Decorrelation through dimensionality reduction. In International Workshop on Multiple Classifier Systems (pp. 238-247). Berlin, Heidelberg: Springer.

Pandey, M., & S. Taruna. 2014. A comparative study of ensemble methods for students' performance modeling. *International Journal of Computer Applications* 103 (8): 0975 – 8887.

Pandey, M., & S. Taruna. 2018. An Ensemble-Based Decision Support System for the Students' Academic Performance Prediction. In ICT Based Innovations (pp. 163-169). Singapore: Springer.

Paris, I. H. M., L. S. Affendey, & N. Mustapha. 2010. Improving academic performance prediction using voting technique in data mining. *World Academy of Science, Engineering and Technology* 62: 820-823.

Polat, K., & S. Güneş. 2009. A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications* 36 (2): 1587-1592.

Punlumjeak, W., S. Rugtanom, S. Jantarat, & N. Rachburee. 2018. Improving Classification of Imbalanced Student Dataset Using Ensemble Method of Voting, Bagging, and Adaboost with Under-Sampling Technique. In IT Convergence and Security 2017 (pp. 27-34). Singapore: Springer.

Qiang, F., H. Shang-Xu, & Z. Sheng-Ying. 2005. Clustering-based selective neural network ensemble. *Journal of Zhejiang University-Science A* 6 (5): 387-392.

Ramaswami, M., & R. Bhaskaran. 2010. A CHAID based performance prediction model in educational data mining. arXiv preprint arXiv:1002.1144.

Richardson, M., C. Abraham, and R. Bond. 2012. Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin* 138 (2): 353-387.

Romero, C., M-L. Lpez, J.-M. Luna, and S. Ventura. 2013b. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education* 68: 458-472.

Sanzana, M. B., S. S. Garrido, & C. M. Poblete. 2015. Profiles of Chilean students according to academic performance in mathematics: An exploratory study using classification trees and random forests. *Studies in Educational Evaluation* 44: 50-59.

Satyanarayana, A., & M. Nuckowski. 2016. Data mining using ensemble classifiers for improved prediction of student academic performance. N-913, Dept. of Computer Systems Technology, NewYork City College of Technology (CUNY), 300 Jay St, Brooklyn NY – 11201.

Şen, B., E. Uçar, and D. Delen. 2012. Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications* 39 (10): 9468-9476.

Superby, J. F., J. P. Vandamme, & N. Meskens. 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. In Workshop on educational data mining (Vol. 32, p. 234).

Tinto, V. 1982. Limits of Theory and Practice in Student Attrition. *The Journal of HigherEducation* 53 (6): 687-700.

Trivedi, S., Z. A. Pardos, & N. T. Heffernan. 2011, June. Clustering students to generate an ensemble to improve standard test score predictions. In International Conference on Artificial Intelligence in Education (pp. 377-384). Berlin, Heidelberg: Springer.

Twala, B. 2010. Multiple classifier application to credit risk assessment. *Expert Systems with Applications* 37 (4): 3326-3336.

Vesanto, J., & E. Alhoniemi. 2000. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks* 11 (3): 586-600.

هدیه مهدوی

متولد سال ۱۳۷۲، دارای مدرک تحصیلی کارشناسی ارشد در رشتهٔ مهندسی صنایع از دانشگاه قم است.

داده‌کاوی، فرایندکاوی، یادگیری ماشین، مدیریت و مهندسی دانش از جمله علایق پژوهشی وی است.

جلال رضائی‌نور

متولد ۱۳۵۶، دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی صنایع از دانشگاه علم و صنعت ایران است. ایشان هم‌اکنون دانشیار گروه مهندسی صنایع دانشگاه قم است.

مدیریت و مهندسی دانش، مدیریت فرایندها و داده‌کاوی از جمله علایق پژوهشی وی است.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

محمد امینی

متولد سال ۱۳۶۰، دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی فناوری اطلاعات از دانشگاه علم وصنعت ایران است. ایشان هم‌اکنون پژوهشگر ارشد مرکز مطالعات و تحقیقات شرکت فناپ در پارک فناوری پردیس است.

هوش مصنوعی، یادگیری ماشینی، داده‌کاوی و سیستم‌های تصمیم‌یار از جمله علایق پژوهشی وی است.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی