# بهبود الگوریتم ماشین بردار پشتیبان با الگوریتم رقابت استعماری برای دسته‌بندی اسناد متنی

زهرا عاشقی دیزجی*[۱]، سکینه اصغری آقجه‌دیزج[۲] و فرهاد سلیمانیان قره‌چپق[۳]

[۱و۳] گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران

[۲] گروه مهندسی کامپیوتر، واحد بناب، دانشگاه آزاد اسلامی، مراغه، ایران

## چکیده

**با توجه به رشد نمایی متون الکترونیکی، سازماندهی و مدیریت متون، مستلزم ابزاری است که اطلاعات و داده‌های مورد جستجوی کاربران را در کمترین زمان ارائه دهد؛ ازاین‌رو در سال‌های اخیر روش‌های دسته‌بندی اهمیت ویژه‌ای پیدا کرده است. هدف دسته‌بندی متون دست‌یابی به اطلاعات و داده‌ها در کسری از ثانیه است. یکی از مشکلات اصلی در دسته‌بندی متون، ابعاد بالای ویژگی‌هاست. برای کاهش ویژگی‌های متون، انتخاب ویژگی‌ها یکی از مؤثرترین راه‌حل‌هاست. چراکه هزینه محاسباتی که تابعی از طول بردار ویژگی‌هاست، بدون انتخاب ویژگی‌ها افزایش می‌یابد. در این مقاله روشی براساس بهبود الگوریتم ماشین بردار پشتیبان با الگوریتم رقابت استعماری برای دسته‌بندی اسناد متنی ارائه شده است. در روش پیشنهادی، از الگوریتم رقابت استعماری برای انتخاب ویژگی‌های و از الگوریتم ماشین بردار پشتیبان برای دسته‌بندی متون استفاده شده است. آزمایش و ارزیابی روش پیشنهادی بر روی مجموعه داده‌های Reuters21578, WebKB و Cade 12 انجام شده است. نتایج شبیه‌سازی حاکی از آن است که روش پیشنهادی در معیارهای دقت، بازخوانی و F Measure از روش ماشین بردار پشتیبان بدون انتخاب ویژگی عملکرد بهینه‌تری دارد.**

**واژگان کلیدی: انتخاب ویژگی، دسته‌بندی متون، الگوریتم رقابت استعماری، الگوریتم ماشین بردار پشتیبان، بهینه‌سازی**

# An Improvement in Support Vector Machines Algorithm with Imperialism Competitive Algorithm for Text Documents Classification

**Zahra Asheghi Dizaji*[1], Sakineh Asghari Aghjehdizaj [2], Farhad Soleimanian Gharehchopogh[3]**

[1, 3]Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran

[2]Department of Computer Engineering, Bonab Branch, Islamic Azad University, Maragheh, Iran

## Abstract

**Due to the exponential growth of electronic texts, their organization and management requires a tool to provide information and data in search of users in the shortest possible time. Thus, classification methods have become very important in recent years.**

**In natural language processing and especially text processing, one of the most basic tasks is automatic text classification. Moreover, text classification is one of the most important parts in data mining and machine learning. Classification can be considered as the most important supervised technique which classifies the input space to k groups based on similarity and difference such that targets in the same group are similar and targets in different groups are different. Text classification system has been widely used in many fields, like spam filtering, news classification, web page detection, Bioinformatics, machine translation, automatic response systems, and applications regarding of automatic organization of documents.**

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۷/۳/۱۳ • تاریخ پذیرش: ۱۳۹۸/۴/۱۹ • تاریخ انتشار: ۱۳۹۹/۰۴/۰۱ • نوع مطالعه: پژوهشی





[ DOI: 10.29252/jsdp.17.1.117 ]

The important point in obtaining an efficient text classification method is extraction and selection of key features of texts. It is proved that only 33% of words and features of the texts are useful and they can be used to extract information and most words existing in texts are used to represent purpose of a text and they are sometimes repeated. Feature selection is known as a good solution to high dimensionality of the feature space. Excessive number of Features not only increase computation time but also degrade classification accuracy. In general, purpose of extracting and selecting features of texts is to reduce data volume, time required for training, computational time and increase performance speed of the methods proposed for text classification. Feature extraction refers to the process of generating a small set of new features by combining or transforming the original ones, while in feature selection dimension of the space is reduced by selecting the most prominent features.

In this paper, a solution to improve support vector machine algorithm using Imperialism Competitive Algorithm, are provided. In this proposed method, the Imperialism Competitive Algorithm for selecting features and the support vector machine algorithm for Classification of texts are used.

At the stage of extracting the features of the texts, using weighting schemes such as NORMTF, LOGTF, ITF, SPARCK, and TF, each extracted word is allocated a weight in order to determine the role of the words in terms of their effects as the keywords of the texts. The weight of each word indicates the extent of its effect on the main topic of the text compared to other words used in the same text. In the proposed method, the TF weighting scheme is used for attributing weights to the words. In this scheme, the features are a function of the distribution of different features in each of the documents $d_i \in D$.

Moreover, at this stage, using the process of pruning, low-frequency features and words that are used fewer than two times in the text are pruned. Pruning basically filters low-frequency features in a text [18].

In order to reduce the number of dimensions of the features and decrease computational complexity, the imperialist competitive algorithm (ICA) is utilized in the proposed method. The main goal of employing the imperialist competitive algorithm (ICA) in the proposed method is minimizing the loss of data in the texts, while also maximizing the reduction of the dimensions of the features.

In the proposed method, since the imperialist competitive algorithm (ICA) has been used for selecting the features, there must be a mapping created between the parameters of the imperialist competitive algorithm (ICA) and the proposed method. Accordingly, when using the imperialist competitive algorithm (ICA) for selecting the key features, the search space includes the dimensions of the features, and among all the extracted features, $\frac{1}{2}$, $\frac{1}{4}$, or $\frac{1}{5}$ of all the features are attributed to each of the countries. Since the mapping is carried out randomly, there may be repetitive features in any of the countries as well. Next, based on the general trend of the imperialist competitive algorithm (ICA),some countries which are more powerful are considered as imperialists, while the other countries are considered as colonies. Once the countries are identified, the optimization process can begin. Each country is defined in the form of an $1 * N$ array with different values for the variables as in Equations 2 and 3.

$$\text{Country} = |var_1, var_2, \ldots, var_i, var_N| \quad (2)$$

$$\text{Cost} = f(\text{Country}) \quad (3)$$

The variables attributed to each country can be structural features, lexical features, semantic features, or the weight of each word, and so on. Accordingly, the power of each country for identifying the class of each text is increased or decreased based on its variables.

One of the most important phases of the imperialist competitive algorithm (ICA) is the colonial competition phase. In this phase, all the imperialists try to increase the number of colonies they own. Each of the more powerful empires tries to seize the colonies of the weakest empires to increase their own power. In the proposed method, colonies with the highest number of errors in classification and the highest number of features are considered as the weakest empires.

Based on trial and error, and considering the target function in the proposed method, the number of key features relevant to the main topic of the texts is set to $\frac{1}{5}$ of the total extracted features, and only through using $\frac{1}{5}$ of the key features of each text along with a classifier algorithm such as $AdaBoost$, support vector machine (SVM), $K$ nearest neighbors, and so on, the class of that text can be determined in the proposed method.

Since the classification of texts is a nonlinear problem, in order to classify texts, the problem must first be mapped into a linear problem. In this paper, the RBF kernel function along with $\gamma$ is used for mapping the problem.

The hybrid algorithm is implemented on the Reuters21578, WebKB, and Cade 12 data sets to evaluate the accuracy of the proposed method. The simulation results indicate that the proposed hybrid algorithm in precision, recall and F Measure criteria is more efficient than primary support machine carriers.

Keywords: Feature Selection, Text Classification, Imperialism Competitive Algorithm, Support Vector Machines Algorithm, Optimization





[ DOI: 10.29252/jsdp.17.1.117 ]

## ۱- مقدمه

امروزه اطلاعات، بزرگ‌ترین و مهمترین نقش در زندگی انسان‌ها دارند. بیش‌تر سازمان‌ها، دانشگاه‌ها، رسانه‌های اجتماعی و .. بانک‌های اطلاعاتی ارزشمند دارند که از داده‌ها و اطلاعات انبوهی بهره می‌برند. از طرفی با توجه به گسترش فضای مجازی انتشار و تبادل اطلاعات نیز سهل و آسان شده است و روزانه حجم اطلاعات رو به افزایش است. متون خبری، مقالات و کتاب‌های الکترونیکی، نامه‌های الکترونیکی، صفحات وب و... تنها بخشی از این اطلاعات رو به رشد هستند. برای سازمان‌دهی و بازیابی اطلاعات ابزارهایی لازم است که در کمترین زمان اطلاعات و داده‌های مورد جستجوی کاربران را ارائه دهند و از میان حجم بالای داده‌ها دانش نهفته آنها را استخراج کنند. به این منظور روش‌های مختلفی در دسته‌بندی متون توسط پژوهش‌گران ارائه می‌شود؛ چراکه دسته‌بندی اطلاعات به‌عنوان یکی از مهم‌ترین روش‌ها در بازیابی اطلاعات و راه‌حل مؤثری برای سازماندهی پایگاه‌داده‌های متنی است.

دسته‌بندی که اغلب کلاس‌بندی نامیده می‌شود، متون زبان طبیعی را به یک یا بیشتر دسته‌های از قبل معرفی‌شده براساس محتوی نسبت می‌دهد [1,2,3]. هدف دسته‌بندی متون دست‌یابی به اطلاعات و داده‌ها در کسری از ثانیه است. به این صورت که اسناد مرتبط با هم شناسایی می‌شوند و با توجه به ارتباط منطقی هرکدام با یکدیگر در دسته‌های مشابه قرار می‌گیرند. دسته‌بندی یا رده‌بندی یک متن می‌تواند اطلاعات مفیدی برای فرآیندهای همچون تبدیل ترجمه ماشین، فیلترکردن متون، سامانه‌های خودکار پاسخ به سؤالات و کاربردهای مرتبط با سازماندهی خودکار مستندات، افزایش کارایی در موتورهای جستجو و ... به‌دست آورد.

یکی از مشکلات اصلی در دسته‌بندی متون، وجود ابعاد بالا در فضای ویژگی‌های متون است. برای کاهش ویژگی‌های متون، انتخاب ویژگی‌ها و واژگان کلیدی یکی از موثرترین راه‌حل‌هاست. دست‌یابی به روشی دقیق در دسته‌بندی متون رابطهٔ مستقیم به انتخاب درست ویژگی‌های متون و تعداد آنها دارد؛ زیرا عامل اصلی در دسته‌بندی متون انتخاب ویژگی‌های کلیدی متون است [4,5].

ویژگی‌ها و واژگان کلیدی مجموعه‌ای از لغات مهم در یک مستند هستند که توصیفی از محتوای مستند را فراهم می‌آوردند. به‌طورکلی انتخاب ویژگی‌ها، زمان لازم را برای آموزش و هزینه محاسبات که تابعی از طول بردار ویژگی‌هاست، کاهش می‌دهد؛ همچنین سرعت و دقت عملکرد روش‌ها را با کاهش حجم داده‌ها افزایش می‌دهد. ازاین‌رو در این مقاله روشی براساس بهبود الگوریتم ماشین بردار پشتیبان با الگوریتم رقابت استعماری برای دسته‌بندی اسناد متنی ارائه شده که در روش پیشنهادی از الگوریتم رقابت استعماری برای انتخاب ویژگی‌های کلیدی استفاده شده، چون که این الگوریتم به‌عنوان نخستین الگوریتم بهینه‌سازی مبتنی بر یک فرایند اجتماعی- سیاسی است و توانایی بهینه‌سازی هم‌تراز و حتی بالاتر در مقایسه با الگوریتم‌های مختلف بهینه‌سازی، در مواجهه با انواع مسائل بهینه‌سازی دارد و از سرعت مناسب در یافتن جواب بهینه برخودار است. همچنین با توجه به انتخاب ویژگی‌ها از الگوریتم ماشین بردار پشتیبان برای دسته‌بندی متون استفاده شده و در مقالات مختلفی از قبیل [6]، [7]، [8]، [9]، [10]، [11] و [12] از الگوریتم ماشین بردار پشتیبان برای دسته بندی متون استفاده شده است.

ساختار کلی مقاله به‌شرح زیر سامان‌دهی شده است: در بخش ۲ کارهای مرتبط قبلی انجام‌شده در این حوزه بیان شده است. در بخش ۳ روش پیشنهادی به‌صورت کامل ارائه شده است. در بخش ۴ نتایج روش پیشنهادی به‌طور کامل بررسی و ارزیابی می‌شود و درنهایت در بخش ۵ به نتیجه‌گیری و کارهای آینده پرداختیم.

## ۲- کارهای مرتبط

با توجه به اهمیت دسته‌بندی اسناد متنی در گذشته پژوهش‌گران مختلفی با روش‌های مختلفی به حل این امر مهم مبادرت کرده‌اند که در زیر به تعدادی از آنها اشاره شده است.

در [7]، از الگوریتم ماشین بردار پشتیبان با هدف دسته‌بندی متون از روش انتخاب ویژگی چند ردهٔ ماشین بردار پشتیبان برای دسته‌بندی متون استفاده شده است. نتایج حاکی از آن است که روش ارائه‌شده دقت خوبی در دسته‌بندی متون دارد و همچنین در مطالعه [9]، رویکرد خوشه‌بندی بدون نظارت با استفاده از شبکه‌های خود سازمان‌دهی شده و خوشه‌بندی خودکار با استفاده از ضریب همبستگی استفاده شده است. درنهایت، خوشه‌ها به‌عنوان برچسب‌هایی برای آموزش ماشین بردار استفاده می‌شود.




[ DOI: 10.29252/jsdp.17.1.117 ]

در [8]، روشی تحت عنوان TCFP ارائه شده است و از الگوریتم ماشین بردار پشتیبان در ترکیب با الگوریتم‌های نیوی بیز و K نزدیک‌ترین همسایه برای دسته‌بندی استفاده شده است. آزمایش روش پیشنهادی بر روی مجموعه داده‌های Newsgroups, Reuters و WebKB انجام شده است. نتایج معیار F-Measure در مجموعه داده Newsgroups با روش ماشین بردار پشتیبان برابر با ۸۲.۴۹، با روش K نزدیک‌ترین همسایه برابر ۷۹.۹۵، با روش Rocchio برابر با ۸۳ است. در مجموعه‌داده WebKB نتایج همین معیار با روش ماشین بردار پشتیبان برابر با ۷۳.۷۴، با روش K نزدیک‌ترین همسایه برابر ۶۸/۰۴، با روش Rocchio برابر با ۷۵/۲۸ محاسبه شده است؛ همچنین در مجموعه‌داده Reuters نتایج با روش ماشین بردار پشتیبان برابر با ۸۷/۴۱، با در روش K نزدیک‌ترین همسایه برابر ۸۵.۶۵، با روش Rocchio برابر با ۸۶/۲۶ محاسبه شده است. همچنین در [13]، نیز روش‌های الگوریتم ماشین بردار پشتیبان و الگوریتم K نزدیکترین همسایه برای دسته‌بندی ارائه شده است. برای ارزیابی این روش از مجموعه داده 20 News-Groups و برای وزن‌دهی واژگان داده‌های مجموعه‌داده از روش TF-IDF استفاده شده است. در روش ماشین بردار پشتیبان از تابع کرنل برای آموزش داده استفاده شده است. نتایج حاصله بیان‌گر دقت بیشتر روش K نزدیک‌ترین همسایه در مقایسه با روش ماشین بردار پشتیبان بوده است. در [14] نیز، از روش‌های نیوی بیز، ماشین بردار پشتیبان، رگرسیون لجستیک و روش‌های ترکیبی آدابوست و نیوی بیز، ماشین بردار پشتیبان و آدابوست در دسته‌بندی متون استفاده شده است. برای ارزیابی روش‌ها از مجموعه داده ACM استفاده شده است.

ko و همکاران در [15]، روشی براساس دسته‌بندی خودکار متن با استفاده از اهمیت واژگان هر یک از متون ارائه دادند. آزمایش روش پیشنهادی بر روی دو زبان مختلف (انگلیسی و کره‌ای) در مجموعه‌داده‌های خبری انجام شده است. از چهار نوع مختلف روش طبقه‌بندی: بیزین ساده، ماشین بردار پشتیبان، K نزدیک‌ترین همسایه و Rocchio در روش پیشنهادی استفاده شده است. در [16]، یک دسته‌بندی جدید با یک‌پارچه‌سازی الگوریتم K نزدیک‌ترین همسایه با الگوریتم ماشین بردار پشتیبان ارائه شده است. هدف از این روش کاهش تأثیر پارامترها در دقت دسته‌بندی بوده است به این صورت که در مرحله آموزش الگوریتم ماشین بردار پشتیبان برای کاهش نمونه‌های آموزشی در دسته‌بندی‌های موجود در بردار پشتیبان خود استفاده شده است. این بردارهای پشتیبان از دسته‌های مختلف به‌عنوان دادهٔ آموزشی با الگوریتم K نزدیک‌ترین همسایه هستند که در آن معیارهای شباهت یا تابع فاصله برای محاسبه داده‌های آزمون که در کدام دسته قرار می‌گیرد، استفاده می‌شود و همچنین برای تشخیص اینکه کدام یک مصرف زمان را کاهش می‌دهد، کاربرد دارد. Feng و همکاران در [17]، روشی بر مبنای روش نیو بیز ارائه کرده‌اند که از ویژگی‌های انتخاب‌شده برای دسته‌بندی بهینه‌تر استفاده می‌کند. قابلیت‌های منحصربفرد این روش با استفاده از روش شاخص‌گذاری خصایص با روش شاخص‌گذاری انتخاب عمومی است. نتایج این روش حاکی از آن است که با روش‌های دیگر دسته‌بندی شناخته شده‌ای مثل ماشین بردار پشتیبان قابل رقابت است. در [18]، روش انتخاب ویژگی دومرحله‌ای با استفاده از بهرهٔ اطلاعات، آنالیز بخش‌های اصلی و الگوریتم ژنتیک برای دسته‌بندی متون ارائه شده است. برای ارزیابی روش پیشنهادی از مجموعه داده‌های Reuters-21,578 وClassic3 استفاده شده است. به‌طور کلی در این روش برای کاهش ابعاد ویژگی‌های متون از الگوریتم‌های ژنتیک و PCA استفاده شده و بعد از کاهش ابعاد ویژگی‌ها، دسته‌بندهای K نزدیک‌ترین همسایه و درخت تصمیم‌گیر C4.5 برای دسته‌بندی متون به‌کار رفته است.

naH و همکاران در [19]، روشی با استفاده از تعدیل وزن در دسته‌بندی متون از طریق الگوریتم K نزدیک‌ترین همسایه ارائه دادند. در [20]، روشی براساس اسناد نشان‌دار و بدون نشانه با استفاده از EM برای دسته‌بندی متون ارائه شده است. که درستی دسته‌بندی متون می‌تواند با افزایش یا کاهش تعداد اسناد برچسب‌گذاری شده یا بدون برچسب متغیر است و دقت دسته‌بندی با افزایش تعداد متون برچسب‌گذاری افزایش می‌یابد.

در [21]، روش ترکیبی دو الگوریتم K نزدیک‌ترین همسایه و K-Means در دسته‌بندی متون ارائه شده است. که از الگوریتم K-means برای خوشه‌بندی متون و از الگوریتم K نزدیک‌ترین همسایه به‌منظور تشخیص اسناد مشابه استفاده شده است. نتایج حاصل از ارزیابی روش پیشنهادی در مجموعه داده Reuters-21578 بیان‌گر دقت بالای روش ترکیبی در مقایسه با نتایج روش K-means است. در [22]، روشی با عنوان الگوریتم کلونی مورچه





[ DOI: 10.29252/jsdp.17.1.117 ]

پیشرفته برای انتخاب زیرمجموعه ویژگی مطرح شده است. در این روش از مدل گراف استفاده شده، به‌صورتی‌که هر گره دو زیرگره دارد که یکی برای انتخاب و دیگری برای عدم انتخاب ویژگی بوده است. الگوریتم کلونی مورچه برای انتخاب گره‌ها استفاده می‌شود. عملکرد این روش پیشنهادی با الگوریتم ژنتیک دودویی، الگوریتم ازدحام ذرات بهبود یافته، الگوریتم جستجوی دودویی گرانشی بهبود یافته و برخی الگوریتم‌های برجسته بر اساس الگوریتم کلونی مورچه روی انتخاب ویژگی مقایسه شده است. نتایج نشان داده است که الگوریتم پیشنهادی با استفاده از مجموعه ویژگی کوچک‌تر نسبت به روش‌های دیگر دقت بیشتری دارد.

در [23]، موضوع انتخاب ویژگی متن بر اساس الگوریتم ژنتیک بهبودیافته ارائه شده است. در این روش از یک روش جستجوی ترکیبی استفاده می‌شود که مزیت‌های انتخاب ویژگی به‌صورت فیلتر را با الگوریتم ژنتیک افزایش‌یافته ترکیب می‌کند تا به ابعاد بالایی از فضای ویژگی دست یابد و عملکرد را نیز به‌طور هم‌زمان بهبود ببخشد. نتایج در این رویکرد نشان می‌دهد که این روش ترکیبی مؤثرتر از روش فیلتر تکی برای کاهش ابعاد است، زیرا قادر است، بدون اینکه دقت دسته‌بندی را در بیشتر مواقع از دست بدهد، نرخ کاهش بالاتری تولید کند.

Yonge lu و همکاران در [24]، روشی با عنوان الگوریتم بهینه‌سازی ذرات بهبودیافته و کاربرد آن در انتخاب ویژگی متن ارائه کردند. تلاش به‌منظور بهبود اثر انتخاب ویژگی از طریق الگوریتم بهینه‌سازی ذرات است. نتایج نشان می‌دهد که الگوریتم بهینه‌سازی ذرات بهبودیافته ناهمگام یکی از بهترین روش‌ها در دسته‌بندی متن در میان تمام روش‌هاست.

Hong Wang و Ben Niu روشی با عنوان الگوریتم باکتریایی نوظهور با کنترل تصادفی برای انتخاب ویژگی در طبقه‌بندی ارائه دادند [25]. الگوریتم‌های مبتنی بر باکتری مبتنی بر جمعیت هستند که به‌خاطر توانایی جستجوی سراسری‌شان شناخته شده هستند. این مقاله یک الگوریتم باکتریایی جدید را پیشنهاد می‌کند که بر اساس سازوکارهای کنترل و استراتژی‌های، به‌روزرسانی جمعیت برای انتخاب ویژگی در دسته‌بندی است. مطالعات مقایسه‌ای روی پنج الگوریتم باکتریایی نشان می‌دهد که الگوریتم پیشنهادی با دست‌یابی به کارایی بالاتر طبقه‌بندی از الگوریتم های دیگر بهتر است.

رضایی و همکاران در [26]، روشی برای استخراج واژگان کلیدی و وزن‌دهی واژگان برای بهبود طبقه‌بندی متون فارسی ارائه کردند. در این مقاله با استفاده از اصطلاح‌نامه که از نظامی ساختارمند برخودار است، واژگان کلیدی بامعناتری از متون استخراج شده و با آن‌ها طبقه‌بندی متون فارسی صورت گرفته است. در مرحله نخست، واژگان زائد حذف و باقی واژگان ریشه‌یابی شده است؛ سپس به‌کمک اصطلاح‌نامه کلمات هم‌معنی، اعم و اخص‌ها و همچنین وابسته‌ها پیدا و در ادامه برای مشخص‌شدن اهمیت نسبی واژگان یک وزن عددی به هر کلمه منسوب شده است که بیان‌گر میزان تأثیر واژه در ارتباط با موضوع متن و در مقایسه با سایر واژگان به‌کاررفته در متن است؛ درنهایت طبقه‌بندی متون انجام شده است. همچنین در [27]، با استفاده از اطلاعات زبان‌شناختی و اصطلاح‌نامه، واژگان کلیدی بامعناتری استخراج شده است. نتایج آزمایش‌ها روی چندین متن در موضوعات مختلف نشان‌دهندهٔ دقت و توانایی روش در استخراج واژگان کلیدی منطبق با خواست کاربر بوده و درنتیجه خوشه‌بندی متون با دقت بالا صورت گرفته است.

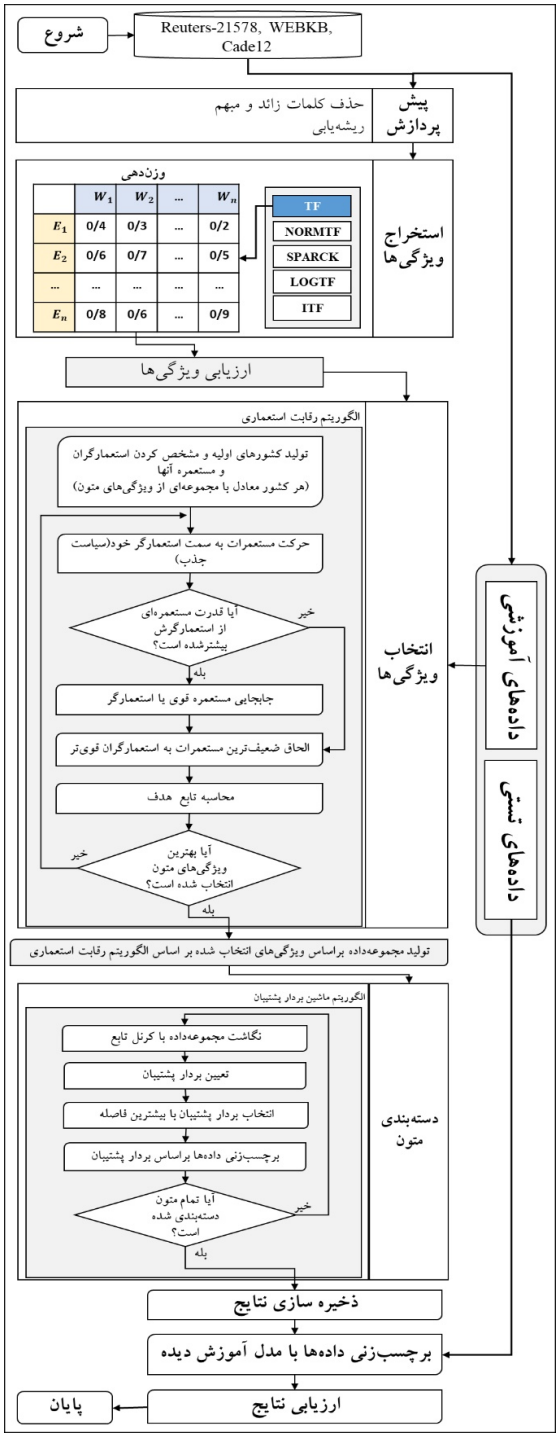
در [28]، برای هماهنگ‌کردن معیار تبدیل ویژگی و نیز معیار دسته‌بندی ماشین بردار پشتیبان روشی برای تخمین تبدیل ویژگی با استفاده از الگوریتم ژنتیک پیشنهاد شده که معیار تبدیل آن کمینه‌کردن خطای دسته‌بندی ماشین‌بردار پشتیبان است. علاوه‌بر آن، روشی برای تخمین تبدیل ویژگی با استفاده از الگوریتم ژنتیک دوهدفه، پیشنهاد شده که معیار این تبدیل بیشینه‌شدن تمایز بین‌دسته‌ای و کمینه‌کردن خطای دسته‌بندی ماشین‌بردار پشتیبان به‌صورت هم‌زمان است.

## ۳- روش پیشنهادی

مسأله دسته‌بندی متون یکی از مسائل اصلی مطرح‌شده در یادگیری ماشین است. روش‌های یادگیری ماشین بر پایه دو ردهٔ آموزش و آزمون هستند. بر این اساس هشتاد درصد داده‌های مجموعه‌داده‌ها برای آموزش و بیست درصد مابقی نیز برای آزمایش و مشاهده نتایج روش پیشنهادی در نظر گرفته شده است. به‌طور کلی روش پیشنهادی شامل چهار مرحلهٔ اصلی پیش‌پردازش، استخراج ویژگی‌ها، انتخاب ویژگی‌ها و دسته‌بندی متون است. در شکل (۱) نحوه عملکرد روش پیشنهادی نشان داده شده است.




[ DOI: 10.29252/jsdp.17.1.117 ]




(شکل-۱): روندنمای روش پیشنهادی
**(Figure-1): Flowchart of Proposed Method**

## ۳-۱- پیش‌پردازش

در نخستین مرحله از روش پیشنهادی، مجموعه‌داده انتخابی کاربر خوانده و پیش‌پردازش می‌شود. در این مرحله واژگانی نظیر حروف ربطی، علایم نشانه‌گذاری، ضمایر، بسیاری از قیدها، صفات و واژگان مبهم که با مفهوم و موضوع اصلی متن مرتبط نیستند، حذف می‌شوند. دراصطلاح به این ویژگی‌های غیرمرتبط با موضوع اصلی متن، فهرست توقف[1] گفته می‌شود. حذف این ویژگی‌های تکراری بسیار مهم است؛ چون وجود این واژگان پرتکرار موجب دسته‌بندی اشتباه متون می‌شود؛ همچنین در این مرحله ریشه‌یابی واژگان نیز انجام می‌شود. در مرحلهٔ ریشه‌یابی، ریشهٔ تشکیل‌دهندهٔ واژگان مشخص می‌شوند؛ بنابراین، واژه‌هایی با ریشهٔ مشابه تعیین می‌شوند، یعنی واژگانی که به‌دلیل داشتن پیشوندها و پسوندها مختلف به نظر می‌رسند نیز در یک رده در نظر گرفته می‌شوند. برای مثال، "put" ، "put up" ، "put on" و "put off" همه ریشهٔ put دارند.

## ۳-۲- استخراج ویژگی‌ها

در مرحلهٔ استخراج ویژگی‌های متون با استفاده از الگوهای وزن‌دهی مانند SPARCK, ITF, LOGTF, NORMTF و TF به هر کلمه استخراج‌شده، وزنی نسبت داده می‌شود و نقش واژگان از نظر میزان تأثیرآنها به‌عنوان واژگان کلیدی متون مشخص می‌شود. وزن هر واژه بیان‌گر میزان تأثیر واژه در موضوع اصلی متن در مقایسه با سایر واژگان به‌کاررفته در متن است. در روش پیشنهادی از الگوی وزن‌دهی TF برای وزن‌دهی به واژگان استفاده شده است. در این روش وزن‌دهی ویژگی‌ها تابعی از توزیع ویژگی‌های مختلف در هریک از مستندات $d_i \in D$ است.

در روش TF در صورت وجود ویژگی $t_k$ در اسناد $d_i$ وزن آن برابر تعداد تکرار آن ویژگی در اسناد مربوطه است. روش TF طبق معادله (۱) تعریف شده که $(t_k, d_i)$ برابر تعداد تکرار هر ویژگی $t_k$ در $d_i$ است.

$$W_{ki} = tf(t_k, d_i) = \begin{cases} (t_k, d_i) & t_k \in \text{vector } d_i \\ 0 & t_k \notin \text{vector } d_i \end{cases} \quad (۱)$$

همچنین در این مرحله از طریق فرآیند هرس‌کردن ویژگی‌های کم‌تکرار و واژگانی که کمتر از دو بار در متن ظاهر شده باشند، هرس شده است. فرآیند هرس‌کردن به‌طوراساسی ویژگی‌های کم‌تکرار در یک متن را فیلتر می‌کند [18].

[1] Stop-List





[ DOI: 10.29252/jsdp.17.1.117 ]

## ۳-۳- انتخاب ویژگی‌ها

با توجه به اینکه تعداد ویژگی‌های متون با فرایند وزن‌دهی کاهش می‌یابد، باز هم مشکل اصلی در دسته‌بندی، فضای پیچیده و ابعاد بالای ویژگی‌هاست. چراکه برای دسته‌بندی سریع نیاز به ویژگی‌های مرتبط با موضوع اصلی هر متن است. ازاین‌رو، برای کاهش ابعاد ویژگی‌ها و کاهش پیچیدگی محاسباتی، الگوریتم رقابت استعماری را در روش پیشنهادی به‌کار برده‌ایم. الگوریتم رقابت استعماری، الگوریتمی نوظهور در حیطه محاسباتی تکاملی و برگرفته از روند تکامل جوامع انسانی است. نخستین مقاله در مورد الگوریتم رقابت استعماری در سال ۲۰۰۷ توسط مبدع روش، ارائه شده است [29]، از این الگوریتم تاکنون در مسائل مختلفی استفاده شده و با گذشت زمان کاربرد آن در مسائل بهینه‌سازی نیز افزایش یافته است [30,31]. این الگوریتم به‌عنوان نخستین الگوریتم بهینه‌سازی مبتنی بر یک فرایند اجتماعی- سیاسی است و توانایی بهینه‌سازی هم‌تراز و حتی بالاتر در مقایسه با الگوریتم‌های مختلف بهینه‌سازی، در مواجهه با انواع مسائل بهینه‌سازی دارد. همچنین از سرعت مناسب در یافتن جواب بهینه برخودار است. با توجه به محاسن این الگوریتم، به‌کارگیری این الگوریتم در روش پیشنهادی به‌عنوان یک راه‌حل قدرتمند برای انتخاب ویژگی‌های کلیدی و کاهش ابعاد ویژگی‌های متون شده است. هدف اصلی از به‌کارگیری الگوریتم رقابت استعماری در روش پیشنهادی این است که از دست رفتن اطلاعات متون به کمینه برسد، در حالی که کاهش ابعاد ویژگی‌ها نیز بیشینه باشد.

در روش پیشنهادی، براساس این‌که از الگوریتم رقابت استعماری در انتخاب ویژگی‌ها استفاده شده است، باید نگاشتی بین پارامترهای الگوریتم رقابت استعماری و روش پیشنهادی ایجاد شود.

براین اساس، هنگام استفاده از الگوریتم رقابت استعماری برای انتخاب ویژگی‌های کلیدی، فضای جستجو ابعاد ویژگی‌هاست و از مجموع کل ویژگی‌های استخراج‌شده، به‌طور تصادفی ۱/۲ یا ۱/۴ یا ۱/۵ از کل ویژگی‌ها به هر یک از کشورها نسبت داده شده، چون ارسال به‌طور تصادفی صورت می‌گیرد ممکن است در یک کشور، ویژگی‌های تکراری هم وجود داشته باشد؛ سپس با توجه به روند کلی الگوریتم رقابت استعماری تعدادی از کشورها که قدرتمندتر هستند به‌عنوان امپریالیست و مابقی به‌عنوان مستعمره در نظر گرفته شده است.

با مشخص‌شدن کشورها می‌توان فرایند بهینه‌سازی را آغاز کرد. هر کشور به‌صورت آرایه **1*N** با مقدار متغیرهای متفاوت به‌صورت معادلات (۲) و (۳) تعریف می‌شوند. متغیرهای نسبت داده‌شده به هر کشور می‌تواند ویژگی‌های ساختاری، لغوی، معنایی، وزن هر کلمه و ... باشد. بدین ترتیب قدرت هر کشور در تشخیص دستهٔ هر متن با توجه به متغیرهایش زیاد یا کم می‌شود.

$$\text{Country} = [var_1, var_2, \ldots, var_i, var_N] \qquad (۲)$$

$$\text{Cost} = f(\text{Country}) \qquad (۳)$$

در معادلات (۲) و (۳) $\boldsymbol{var_i}$: متغیر تصمیم iام، Country: راه‌حل مسأله و Cost: مقدار تابع هدف است.

یکی از مهم‌ترین بخش‌های الگوریتم رقابت استعماری مرحله رقابت استعماری است، که در این مرحله تمام استعمارگران سعی بر افزایش تعداد مستعمرات خود دارند. هر امپراطوری قوی‌تر، سعی دارد مستعمرات ضعیف‌ترین امپراطورها را تحت سلطهٔ خود آورد و بر قدرت خود بیفزایید. در روش پیشنهادی مستعمره‌های که بیشترین تعداد خطا در دسته‌بندی را با بیشترین تعداد ویژگی‌ها داشته‌اند به‌عنوان ضعیف‌ترین امپراطوری در نظر گرفته شده است و به‌صورت معادله (۴) نشان داده می‌شود.

$$T.C_n = Cost\,(imperialist_n) + \xi\, mean\{Cost\,(colonies\ of\ empire_n)\} \qquad (۴)$$

در معادله (۴)، $Cost\,(imperialist_n)$: مقدار تابع هدف امپریالیست nام که برابر با تعداد دسته‌بندی درست، $\xi$: یک عدد مثبت دلخواه کوچک‌تر از یک $T.C_n$: مقدار تابع هدف امپریالست nام که برابر با کمترین تعداد ویژگی‌ها و بالاترین تعداد دسته‌بندی درست در نظر گرفته شده است. بر اساس سعی و خطا و با توجه به تابع هدف در روش پیشنهادی، تعداد ویژگی‌های کلیدی و مرتبط با موضوع اصلی متون برابر با ۱/۵ ویژگی‌ها از کل ویژگی‌های استخراج شده، تعیین شده است و فقط از طریق ۱/۵ ویژگی‌های کلیدی هر متن در کنار یک الگوریتم طبقه‌بند مثل آدابوست، ماشین بردار پشتیبان، K نزدیکترین همسایه و ... می‌توان طبقه‌ی آن متن را در روش پیشنهادی مشخص کرد. که در روش پیشنهادی از الگوریتم ماشین بردار پشتیبان برای دسته‌بندی متون از طریق ویژگی‌های کلیدی استفاده شده است و این الگوریتم فقط براساس ویژگی‌های ارسالی از الگوریتم رقابت استعماری دسته‌بندی متون را انجام می‌دهد.




[ DOI: 10.29252/jsdp.17.1.117 ]

## ۴-۳- دسته‌بندی متون

براساس این‌که دسته‌بندی متون یک مسأله غیر خطی است. برای دسته‌بندی متون ابتدا باید مسأله به‌صورت خطی نگاشت شود. که در این مقاله از تابع کرنل RBF به همراه $\gamma$ که در معادله (۵) نشان داده شده است برای نگاشت مسأله استفاده شده است.

$$K(X_i \times X_j) = \exp\left(-\gamma \|X_i - X_j\|^2\right), g > 0 \qquad (۵)$$

در جدول (۱) پارامترهای روش پیشنهادی نشان داده شده است. مقدار پارامترهای روش پیشنهادی به‌صورت تجربی در پیاده‌سازی تعیین شده‌اند و زیاد یا کم‌بودن مقدار پارامترها در نتایج روش پیشنهادی بسیار تأثیرگذارند.

**(جدول-۱): پارامترهای مورد استفاده در روش پیشنهادی**
**(Table-1): Parameters used in Proposed Model**

| | پارامتر | مقدار |
|---|---|---|
| الگوریتم می گرقابت استعماری | تعداد جمعیت اولیه | ۲۰۰ |
| | تعداد امپریالیست | ۵۰ |
| | تعداد تکرار | ۱۰۰ |
| ماشین بردار پشتیبان | Kernel | Radial basis function(RBF) |
| تابع ارزیابی روش پیشنهادی | | تعداد پیش‌بینی درست و غلط |

## ۴- بررسی و ارزیابی نتایج

ارزیابی مدل پیشنهادی در محیط VS.NET 2015 انجام شده است و به‌منظور مشاهده نتایج روش پیشنهادی از مجموعه‌داده‌های WebKB, Reuters 21578 و Cade12 استفاده شده است. این مجموعه داده‌ها شامل موضوعات و ویژگی‌های مختلف هستند به این‌صورت که مجموعه‌داده WEBKB شامل چهار دسته [13,32] مجموعه‌داده Reuters 2157 شامل ده دسته [33,34] و مجموعه‌داده Cade 12 شامل دوازده دستهٔ مختلف [35] است. در جدوال (۲) تا (۴) انواع دسته‌ها و موضوعات مجموعه داده‌ها در دو مجموعه داده آزمون و آموزش نشان داده شده است.

**(جدول-۲): فهرست دسته‌ها و موضوعات مجموعه‌داده Ruters-21578**
**(Table-2): List of Categories and Topics in Ruters-21578 Data Set**

| Nos. | Class label | Training samples | Testing samples |
|---|---|---|---|
| 1 | Earn | 2877 | 1087 |
| 2 | Acq | 1650 | 719 |
| 3 | Money-fx | 538 | 179 |
| 4 | Grain | 433 | 149 |
| 5 | Crude | 389 | 189 |
| 6 | Trade | 369 | 117 |
| 7 | Interest | 347 | 131 |
| 8 | Ship | 197 | 89 |
| 9 | Wheat | 212 | 71 |
| 10 | Corn | 181 | 56 |

**(جدول-۳): فهرست دسته‌ها و موضوعات مجموعه‌داده WEBKB**
**(Table-3): List of Categories and Topics in WEBKB Data Set**

| Nos. | Class label | Training samples | Testing samples |
|---|---|---|---|
| 1 | Course | 651 | 279 |
| 2 | Faculty | 786 | 338 |
| 3 | Project | 352 | 152 |
| 4 | Student | 1148 | 493 |

**(جدول-۴): فهرست دسته‌ها و موضوعات مجموعه‌داده Cade 12**
**(Table-4): List of Categories and Topics in Cade12 Data Set**

| Nos. | Class label | Training samples | Testing samples |
|---|---|---|---|
| 1 | servicos | 5627 | 2846 |
| 2 | sociedade | 4935 | 2428 |
| 3 | lazer | 3698 | 1892 |
| 4 | informatica | 2983 | 1536 |
| 5 | saude | 2118 | 1053 |
| 6 | educacao | 1912 | 944 |
| 7 | internet | 1585 | 796 |
| 8 | cultura | 1494 | 643 |
| 9 | esportes | 1277 | 630 |
| 10 | noticias | 701 | 381 |
| 11 | ciencias | 569 | 310 |
| 12 | compras-online | 423 | 202 |

همچنین معیارهای ارزیابی دقت، بازخوانی و F Measure را در محاسبه و ارزیابی روش پیشنهادی به‌کار گرفته‌ایم. فاکتورهای دقت و دوباره‌خوانی دقت عملکرد روش پیشنهادی را محاسبه می‌کنند، با استفاده از معادلات (۶) و (۷) مشخص شده‌اند و معیار F Measure میانگین هم‌ساز وزن‌دهی‌شده از معیارهای دقت و بازخوانی است [36]، با استفاده از معادله (۸) ارزیابی می‌شود:

$$\text{Precision} = \frac{true\ positive}{true\ positive + false\ positive} \qquad (۶)$$

$$\text{Recall} = \frac{true\ positive}{true\ positive + false\ negative} \qquad (۷)$$

$$F\ Measure = \frac{2.Recall.Precision}{Recall + Precision} \qquad (۸)$$

برای اطمینان از ضروت به‌کارگیری هم‌زمان دو الگوریتم رقابت استعماری و ماشین بردار پشتیبان برای دسته‌بندی متون، الگوریتم ماشین بردار پشتیبان بدون دخالت الگوریتم رقابت استعماری نیز پیاده‌سازی شده است. همچنین از آنجایی‌که انتخاب ویژگی‌ها اهمیت بالایی در دسته‌بندی اسناد متنی دارد و انتخاب تعداد کمتر یا بیشتر ویژگی‌های متون رابطهٔ مستقیم با دقت روش‌ها دارد، لذا




[ DOI: 10.29252/jsdp.17.1.117 ]

نتایج روش‌ها را با انتخاب ویژگی‌های ۱/۲، ۱/۴ و ۱/۵ از کل ویژگی‌های استخراج‌شده نیز محاسبه و بررسی شده است. در ضمن برای مقایسه بیشتر نتایج سایر روش‌های k نزدیک‌ترین همسایه [18]، درخت تصمیم‌گیری C4.5 [18]، Naïve-Bayes [37], JRip [37]، ماشین بردار پشتیبان [37] که در مقالات دیگر پیاده‌سازی شده نیز در جداول یادشده است. تمام نتایج در جداول (۵) تا (۷) جمع‌آوری و شکل‌های (۲) تا (۵) نشان داده شده است.

**(جدول-۵): نتایج روش پیشنهادی و سایر روش‌ها در مجموعه‌دادهٔ Reuters 21578 با انتخاب مختلف از تعداد ویژگی‌ها**

**(Table-5): The Results of the Proposed Method and Other Methods in the Reuters 21578 Data Set with a Different Choice of Number of Feature**

| روش‌ها | مقدار ویژگی‌های انتخابی از کل ویژگی‌ها | الگوریتم مورد‌نظر برای انتخاب ویژگیهای متون | معیارهای ارزیابی | | |
|---|---|---|---|---|---|
| | | | Precision (%) | Recall (%) | F FMeasu (%) |
| روش پیشنهادی | 1/2 | ICA | 0.9881 | 0.8982 | 0.941 |
| | 1/4 | ICA | 0.9811 | 0.8779 | 0.9015 |
| | 1/5 | ICA | 0.9821 | 0.9077 | 0.91836 |
| ماشین بردار پشتیبان | 1/2 | توابع ریاضی | 0.7751 | 0.7209 | 0.747 |
| | 1/4 | توابع ریاضی | 0.8242 | 0.8579 | 0.8407 |
| | 1/5 | توابع ریاضی | 0.8832 | 0.8684 | 0.8757 |
| نزدیکترین همسایه K [18] | 7542 | توابع ریاضی | 0.7336 | 0.9559 | 0.8302 |
| C4.5 decision tree[18] | 7542 | توابع ریاضی | 0.8464 | 0.8923 | 0.8688 |
| PCA-Based feature reduction [37] | - | Naïve-Bayes | 0.406 | 0.279 | 0.284 |
| | - | JRip | 0.529 | 0.464 | 0.458 |
| | - | J48 | 0.539 | 0.495 | 0.498 |
| | - | ماشین بردار پشتیبان | 0.607 | 0.324 | 0.323 |
| Ontology -based feature reduction [37] | - | Naïve-Bayes | 0.607 | 0.324 | 0.323 |
| | - | JRip | 0.798 | 0.752 | 0.759 |
| | - | J48 | 0.795 | 0.757 | 0.762 |
| | - | ماشین بردار پشتیبان | 0.843 | 0.811 | 0.814 |

**(جدول-۶): نتایج روش پیشنهادی و سایر روش‌ها در مجموعه‌دادهٔ WebKB با انتخاب مختلف از تعداد ویژگی‌ها**

**(Table-6): The Results of the Proposed Method and Other Methods in the Webkb Data Set with a Different Choice of Number of Feature**

| روش‌ها | مقدار ویژگی‌های انتخابی از کل ویژگی‌ها | الگوریتم‌های مورد‌نظر برای انتخاب ویژگی‌های متون | معیارهای ارزیابی | | |
|---|---|---|---|---|---|
| | | | Precision (%) | Recall (%) | F Measure (%) |
| روش پیشنهادی | 1/2 | ICA | 0.9693 | 0.9154 | 0.9343 |
| | 1/4 | ICA | 0.9763 | 0.9067 | 0.9402 |
| | 1/5 | ICA | 0.9724 | 0.9237 | 0.9373 |
| ماشین بردار پشتیبان | 1/2 | توابع ریاضی | 0.9278 | 0.8383 | 0.8808 |
| | 1/4 | توابع ریاضی | 0.9012 | 0.8471 | 0.8733 |
| | 1/5 | توابع ریاضی | 0.9146 | 0.9107 | 0.9127 |

**(جدول-۷): نتایج روش پیشنهادی و سایر روش‌ها در مجموعه‌دادهٔ Cade12 با انتخاب مختلف از تعداد ویژگی‌ها**

**(Table-7): The Results of the Proposed Method and Other Methods in the Cade12 Data Set with a Different Choice of Number of Feature**

| روش‌ها | مقدار ویژگی‌های انتخابی از کل ویژگی‌ها | الگوریتم‌های مورد‌نظر برای انتخاب ویژگیهای متون | معیارهای ارزیابی | | |
|---|---|---|---|---|---|
| | | | Precision (%) | Recall (%) | F Measure (%) |
| روش پیشنهادی | 1/2 | ICA | 0.9733 | 0.915 | 0.941 |
| | 1/4 | ICA | 0.9793 | 0.911 | 0.9379 |
| | 1/5 | ICA | 0.9823 | 0.9569 | 0.9694 |
| ماشین بردار پشتیبان | 1/2 | توابع ریاضی | 0.9048 | 0.7917 | 0.8444 |
| | 1/4 | توابع ریاضی | 0.8471 | 0.7143 | 0.7792 |
| | 1/5 | توابع ریاضی | 0.9444 | 0.7333 | 0.8256 |

با توجه به جدول (۵) تا (۷)، نتایج روش پیشنهادی با ۱/۲ تعداد ویژگی‌های انتخابی از کل ویژگی‌ها، بر روی مجموعه‌داده WEBKB معیار Precision برابر با ۰/۹۶۹۳ معیار Recall برابر با ۰/۹۱۵۴ و معیار F Measure برابر با ۰/۹۳۴۳ می‌باشد. در مجموعه‌داده Reuters 21578 معیار Precision برابر با ۰/۹۸۸۱ معیار Recall برابر با ۰/۸۹۸۲ و معیار F Measure برابر با ۰/۹۴۱ است و همچنین در مجموعه‌داده Cade12 معیار Precision برابر با ۰/۹۷۳۳





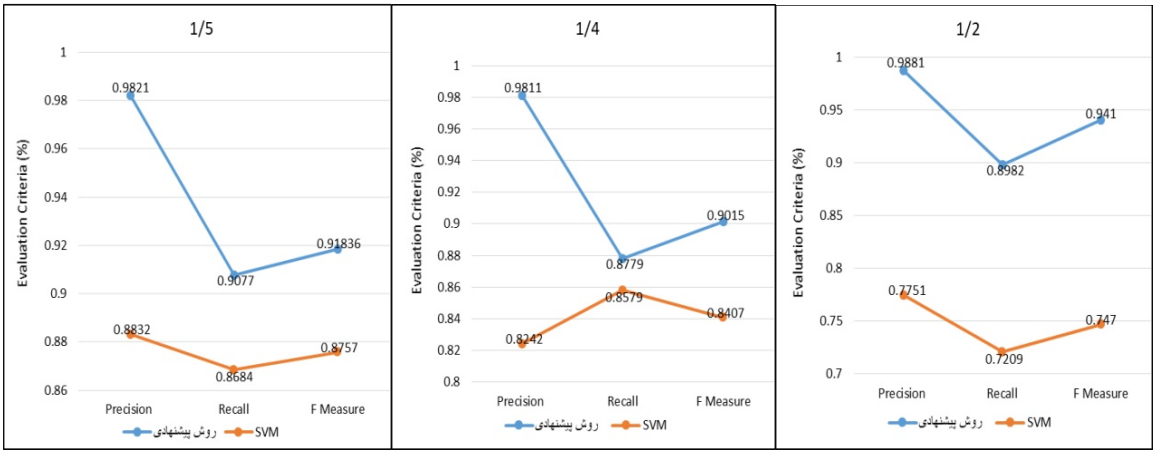
[ DOI: 10.29252/jsdp.17.1.117 ]

معیار Recall برابر با ۰/۹۱۵ و معیار F Measure برابر با ۰/۹۴۱ است.

نتایج روش پیشنهادی با ۱/۴ تعداد ویژگی‌های انتخابی از کل ویژگی‌ها، بر روی مجموعه‌داده WEBKB معیار Precision برابر با ۰/۹۷۶۳ معیار Recall برابر با ۰/۹۰۶۷ و معیار F Measure برابر با ۰/۹۴۰۲ است. در مجموعه‌داده Reuters 21578 معیار Precision برابر با ۰/۹۸۱۱ معیار Recall برابر با ۰/۸۷۷۹ و معیار F Measure برابر با ۰/۹۰۱۵ و همچنین در مجموعه‌داده Cade12 معیار Precision برابر با ۰/۹۷۹۳ معیار Recall برابر با ۰/۹۱۱ و معیار F Measure برابر با ۰/۹۳۷۹ است.

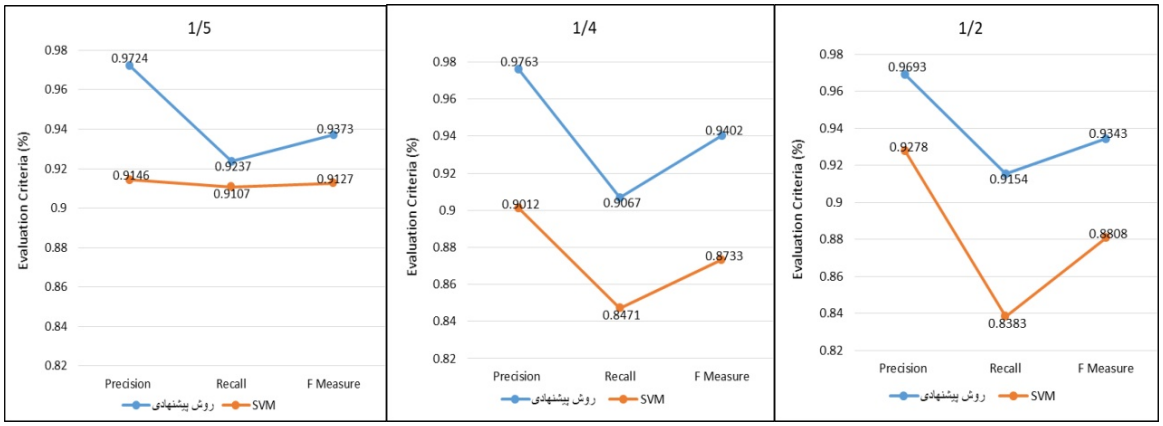
و همچنین نتایج روش پیشنهادی با ۱/۵ تعداد ویژگی‌های انتخابی از کل ویژگی‌ها، بر روی مجموعه‌داده WEBKB معیار Precision برابر با ۰/۹۷۲۴ معیار Recall برابر با ۰/۹۲۳۷ و معیار F Measure برابر با ۰/۹۳۷۳ است. در مجموعه‌داده Reuters 21578 معیار Precision برابر با ۰/۹۸۲۱ معیار Recall برابر با ۰/۹۰۷۷ و معیار F Measure برابر با ۰/۹۱۸۳ و همچنین در مجموعه‌داده Cadc12 معیار Precision برابر با ۰/۹۸۲۳ معیار Recall برابر با ۰/۹۵۶۹ و معیار F Measure برابر با ۰/۹۶۹۴ است.

براساس نتایج حاصله، روش پیشنهادی با ۱/۲، ۱/۴ و ۱/۵ تعداد ویژگی‌های انتخابی از کل ویژگی‌ها در مقایسه با نتایج روش ماشین بردار پشتیبان عملکرد بهینه و دقت بالایی در هرسه مجموعه داده WebKB, Reuters 21578 و cade12 به‌دست آورده است.




**(شکل-۲): ارزیابی روش‌ها بر روی مجموعه داده Reuters 21578 با ۱/۲، ۱/۴ و ۱/۵ ویژگی‌های انتخابی از کل ویژگی‌ها**

**(Figure-2): Evaluation of Methods on the Reuters 21578 dataset with 1.2, 1.4 and 1.5 Features Selected from all Features**



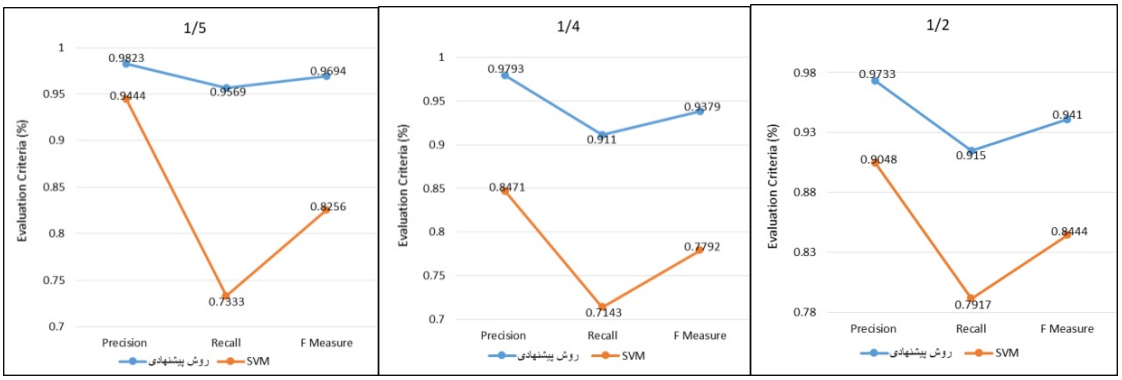

**(شکل-۳): ارزیابی روش‌ها بر روی مجموعه داده WEBKB با ۱/۲، ۱/۴ و ۱/۵ ویژگی‌های انتخابی از کل ویژگی‌ها**

**(Figure-3): Evaluation of Methods on the WEBKB dataset with 1.2, 1.4 and 1.5 Features Selected from all Features**






[ DOI: 10.29252/jsdp.17.1.117 ]




(شکل-۴): ارزیابی روش‌ها بر روی مجموعه داده Cade 12 با۱/۲، ۱/۴ و ۱/۵ ویژگی‌های انتخابی از کل ویژگی‌ها

**(Figure-4): Evaluation of Methods on the Cade12 dataset with 1.2, 1.4 and 1.5 Features Selected from all Features**

## ۵- نتیجه‌گیری

در این مقاله روشی براساس بهبود الگوریتم ماشین بردار پشتیبان با الگوریتم رقابت استعماری برای انتخاب ویژگی‌های کلیدی در دسته‌بندی اسناد متنی با هدف کاهش ابعاد ویژگی‌ها، حذف ویژگی‌های زائد و غیر مرتبط از فضای مسأله که درنتیجه بهبود عملکرد دسته‌بندی متون است، ارائه شده است. در مرحله نخست، هر یک از واژگان در متن متناسب با اهمیت‌شان با استفاده از روش‌های وزن‌دهی SPARCK, ITF, LOGTF, NORMTF و TF رتبه‌بندی شدند؛ سپس براساس الگوریتم رقابت استعماری، ویژگی‌ها و واژگان کلیدی متون انتخاب و کاهش ابعاد ویژگی‌ها انجام شد. بر این اساس، در فرایند دسته‌بندی متون، واژگان و واژگان با اهمیت کمتر، نادیده گرفته می‌شوند؛ درنتیجه انتخاب ویژگی‌های کلیدی متون باعث کاهش هزینه و کاهش زمان محاسباتی، افزایش دقت و بهبود عملکرد روش پیشنهادی در دسته‌بندی متون شد.

با ارزیابی نتایج حاصله می‌توان بیان کرد که روش پیشنهادی عملکرد بهینه‌ای نسبت به سایر روش‌ها در دسته‌بندی متون داشته و عملکرد ماشین بردار پشتیبان در دسته‌بندی متون با استفاده از انتخاب ویژگی‌های کلیدی توسط الگوریتم رقابت استعماری تفاوت مثبتی در نتایج حاصل کرده و باعث افزایش دقت و کارایی روش پیشنهادی در دسته‌بندی متون شده است.

## ۶- مراجع — 6- References


[1] R. Feldman, and J. Sanger, The Text Mining Handbook, "Advanced Approach in Analyzing Unstructured Data", Cambridge University Press, 2007.

[2] D. Chiang, H. Keh, H. Huang, and D. Chyr, "The Chinese text categorization system with association rule and category priority", *Expert System with Applications,* vol. 35, no. 1-2, pp. 102-110, 2008.

[3] L. Khreisat , "A machine learning approach for Arabic text classification using N-gram frequency statistics", *Proceeding of the 3nd Journal of Informetrics*, vol.3(1), pp 72-77, 2009.

[4] A. An, B. Dauletbakov and E. Levner, "Multi-attribute Classification of Text Documents as a Tool for Ranking and Categorization of Educational Innovation Projects", *Lecture Notes in Computer Science*, vol. 8404, pp 404-416, 2014.

[5] A. K. Uysal, "An improved global feature selection scheme for text classification", *Expert systems with Applications*, vol. 43, pp.82-92, 2016.

[6] C. H. Wan, L. H. Lee, R. Rajkumar and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine", *Expert Systems with Applications*, vol. 39(15), pp.11880-11888, 2013.

[7] B. Ramesh and J. G. R. Sathiaseelan, "An advanced Multi Class instance selection based Support Vector Machine for Text Classification", Procedia Computer Science, vol. 57, pp. 1124-1130, 2015.

[8] Y. Ko and J. Seo, "Text classification from unlabeled documents with bootstrapping and feature projection techniques", *Information Processing & Management*, vol. 45(1), pp. 70-83, 2009.

[9] N. Shafiabady, L. H. Lee, R. Rajkumar,, V. P. Kallimani, , N. A. Akram and D. Isa, "Using unsupervised clustering approach to train the Support Vector Machine for text classification", Neurocomputing, vol. 211, pp. 4-10, 2016.




[ DOI: 10.29252/jsdp.17.1.117 ]

[10] L. H. Lee, D. Isa, W. O. Choo and W. Y. Chue, "High Relevance Keyword Extraction facility for Bayesian text classification on different domains of varying characteristic", Expert Systems with Applications, vol. 39(1), pp. 1147-115, 2013.

[11] J. He, A. H. Tan and C. L. Tan, "On Machine Learning methods for Chinese document categorization", Applied Intelligence, vol. 18(3), pp. 311-322, 2003.

[12] D. Isa, L. H. Lee, V. P. Kallimani and R. Rajkumar, "Text document pre-processing with the bayes Formula for classification using the support vector machine", *IEEE Transaction on Knowledge and Data Engineering*, vol. 20(9), pp. 1264-1272, 2008.

[13] D. S. Guru and M. Suhil, "A Novel Term_Class Relevance Measure for Text Categorization", *Procedia Computer Science*, Vol. 45, pp. 13-22, 2015.

[14] A. Onan, S. Korukoğlu and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification", *Expert Systems with Applications*, vol. 57, pp. 232-247, 2016.

[15] Y. Ko, J. Park and J. Seo, "Automatic Text Categorization using the Importance of Sentences", *19th international linguistics-Association for Computational Linguistics*, vol. 1, PP.1-7, 2002.

[16] M. Sivakumar, C. Karthika and P. Renuga, "A Hybrid Text Classification Approach Using KNN And SVM ", *Intenational Journal of Innovative Research In Science Engineering And Technology,Special Issue 3*, vol. 3, pp.1987-1991, 2014.

[17] G. Feng, J. Guo, B. Y. Jing and T. Sun, "Feature Subset Selection Using Naive Bayes for Text Classification", *Pattern Recognition Letters*, vol. 65, pp. 109-115, 2015.

[18] H. Uguz, "A two-stage feature selection method for text categorization by using information gain", *principal component analysis and genetic algorithm. Knowledge-Based Systems*, vol. 24(7), pp. 1024-1032, 2011.

[19] E. H. S. Han, G. Karypis and V. Kumar, "Text Categor ization Using Weight Adjusted k-Nearest Neighbor Classification", *In Pacific-asia conference on knowledge discovery and data mining*, PP: 53-65. Springer Berlin Heidelberg, 2001.

[20] K. Nigam, A. K. McCallum, S. Thrun and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM", Kluwer Academic Publishers, Printed in The Netherlands. Machine Learning, vol.39 (2), pp. 103-134, 2000.

[21] R. Habibpour and K. Khalilpour, "A New Hybrid K-means and K-Nearest-Neighbor Algorithms for Text Document Clustering", *International Journal of Academic Research*, vol.6(3), pp. 7984, 2004.

[22] S. Kashef and H. Nezamabadi-pour, "An advanced ACO algorithm for feature subset selection", Neurocomputing, vol.147, pp. 271-279, 2015.

[23] A. S. Ghareb, A. A. Bakar and A. R. Hamdan, "Hybrid feature selection based on enhanced genetic algorithm for text categorization", *Expert SystemsWith Applications*, vol.49, pp.31-47, 2016.

[24] Y. Lu, M. Liang, Z. Ye and L. Cao, "Improved particle swarm optimization algorithm and its applicationin text feature selection", *Applied Soft Computing*, vol. 35, pp. 629-636, 2015.

[25] H. Wang and B. Niu, "A novel bacterial algorithm with randomness control for feature selection in classification", Neurocomputing, vol. 228, pp. 176-186, 2017.

[۲۶] رضایی، وحیده.، محمدپور، مجید.، پروین، حمید.، نجاتیان، صمد.، ۱۳۹۶. ارائه روشی برای استخراج کلمات کلیدی و وزن‌دهی کلمات برای بهبود طبقه‌بندی متون فارسی. فصل‌نامه‌ی پردازش علائم و داده‌ها، شماره ۴ پیاپی ۳۴.

[26] V. Rezaie, M. Mohammadpour, H. parvin, S. Nejatian, "An Approach for Extraction of Keywords and Weighting Words for Improvement Farsi Documents Classification", *JSDP*, vol. 14 (4), pp.55-78. 2018.

[۲۷] راد، فرهاد.، پروین، حمید.، دهباشی، آتوسا.، مینایی، بهروز.، ۱۳۹۵. ارائه روشی جدید برای شاخص‌گذاری خودکار و استخراج کلمات کلیدی برای بازیابی اطلاعات و خوشه‌بندی متون. فصل‌نامه‌ی پردازش و علائم داده‌ها، شماره ۱ پیاپی ۲۷.

[27] ,F. Rad, H. Parvin, A. Dehbashi, B. Minaee "Improved Clustering Persian Text Based on Keyword Using Linguistic and Thesaurus Knowledge", *JSDP*, vol. 13 (1), pp.87-100, 2016.

[۲۸] حسین‌خانی، فاطمه.، ناصرشریف، بابک.، دو روش تبدیل ویژگی مبتنی بر الگوریتم‌های ژنتیک برای کاهش خطای دسته‌بندی ماشین بردار پشتیبان. فصل‌نامه‌ی پردازش علائم و داده‌ها، شماره ۲ پیاپی ۲۴.

[28] F. Hoseinkhani, B. Nasersharif, "Two Feature Transformation Methods Based on Genetic



[ DOI: 10.29252/jsdp.17.1.117 ]

Algorithm for Reducing Support Vector Machine Classification Error", *JSDP*. Vol. 12 (2), pp. 23-39, 2015

[29] E. Atashpaz-Gargari and C. Lucas, "Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition", *IEEE Congress on Evolutionary Computation*, pp. 4661–4667, 2007.

[30] C. Lucas, Z. Nasiri-Gheidari and F. Tootoonchian, "Application of an imperialist competitive algorithm to the design of a linear induction motor", *Energy Conversion and Management, Elsevier*, vol. 51(7). pp. 1407–1411, 2010.

[31] E. Atashpaz-Gargari, F. Hashemzadeh, R. Rajabioun, and C. Lucas, "Colonial Competitive Algorithm, a novel approach for PID controller design in MIMO distillation column process", *International Journal of Intelligent Computing and Cybernetics*, vol. 1(3). pp. 337–355, 2008.

[32] T.Mitchell ,K.Nigam, D.Freitag, M,Craven , "Learning to extract symbolic knowledge from the world wide web", In: DTIC Document, 1998.

[33] A. Asuncion and D.J. Newmen, UCI Machine Learning Repository, Irvine, CA: Uni-versity of California, Department of information and Computer Science, 2007.

[34]http://archive.ics.uci.edu/ml/datasets/Reuters21778+Text+Categorization+Collection [Last Access: 12-19-2112.

[35] http://ana.cachopo.org/datasets-for-single-label-text-categorization [Lasc Access: 12-19-2112].

[36] J. J. Rocchio, "Document Retrieval Systems - Optimization and Evaluation", PhD thesis, Harvard, 1966.

[37] K. M. Elhadad, Kh. M. Badran, and G. I. Salama, "A Novel Approach for Ontology-based Dimensionality Reduction for Web Text Document Classification", Computer society, pp.373-378, 2017.

**زهرا عاشقی دیزجی** کارشناسی ارشد خود را در رشته مهندسی کامپیوتر از دانشگاه آزاد اسلامی واحد علوم تحقیقات آذربایجان غربی گرفته و تخصص وی پردازش زبان طبیعی، داده‌کاوی و الگوریتم‌های فرابتکاری است.

نشانی رایانامه ایشان عبارت است از:

**zahra_ashegi@yahoo.com**

**سکینه اصغری آقجه‌دیزج** کارشناسی ارشد خود را در رشته مهندسی کامپیوتر از دانشگاه آزاد اسلامی واحد بناب گرفته و تخصص وی پردازش زبان طبیعی، داده‌کاوی و الگوریتم‌های یادگیری ماشین است.

نشانی رایانامه ایشان عبارت است از:

**Sakineh154asghari@gmail.com**

**فرهاد سلیمانیان قره چپق** کارشناسی ارشد و دکترای تخصصی خود را در رشته مهندسی کامپیوتر به ترتیب از دانشگاه چوکوروا و دانشگاه حاجت تپه در کشور ترکیه گرفته و تخصص وی پردازش زبان طبیعی، داده‌کاوی و الگوریتم‌های یادگیری ماشین است. ایشان عضو هیأت علمی دانشکده مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد ارومیه و به تدریس دروس مختلف در حوزه کاری خویش مشغول است. و در حال حاضر سردبیری و مدیرمسئولی نشریه International Journal of Academic Research in Computer Engineering را برعهده دارد.

نشانی رایانامه ایشان عبارت است از:

**farhad@iaurmia.ac.ir**



[ DOI: 10.29252/jsdp.17.1.117 ]