Prediction of anemia with a particle swarm optimization-based approach

Arshed A. Ahmad, Khalid M. Saffer, Murat Sari, Hande Uslu

Department of Computer Science, University of Diyala, Iraq
Department of Mathematical Engineering, Istanbul Technical University, Istanbul, Turkey
Department of Mathematics, Yildiz Technical University, Istanbul, Turkey
arshed980@gmail.com, kha2005ms@yahoo.com, muratsari@itu.edu.tr, usluh@yildiz.edu.tr

ARTICLE INFO

Article History:
Received 23 May 2022
Accepted 8 February 2023
Available 28 July 2023

Keywords:
Anemia
Particle swarm optimization
Support vector machine
Prediction

AMS Classification 2010:
93B52; 93C35; 93C40; 93C42

ABSTRACT

Healthcare enables the maintenance of health through some physical and mental care for the prevention, diagnosis and treatment of disease. Diagnosis of anemia, one of the most common health problems of the age, is also very ambitious. Whereas, pathological individuals could be predicted through various biomedical variables using some appropriate methods. In order to estimate these individuals just by taking into account biological data, particle swarm optimization (PSO) and support vector machine (SVM) clustering techniques have been merged (PSO-SVM). In this respect, the dataset provided has been divided into five clusters based on anemia types consisting of 539 subjects in total, and the anemia values of each subject have been recorded according to corresponding biomedical variables taken as independent parameters. The findings of the PSO-SVM method have been compared to the results of the SVM algorithm. The hybrid PSO-SVM method has proven to be quite effective, particularly in terms of the high predictability of clustered disease types. It is possible to lead the successful creation of appropriate treatment programs for diagnosed patients without overlooking or wasting time during treatment.

1. Introduction

In the past decade computer models have become very popular in the field of biomedicine due to exponentially increasing computer power. So, an efficient healthcare system using computer models can contribute to an important part of a country’s economic development, and industrialization. Healthcare has traditionally been recognized as an important determinant in improving the overall physical and mental health and well-being of people around the world. Anemia is clinically defined as a below hemoglobin value from the appropriate reference range for an individual and it possibly leads to crucial blood diseases. Anemia types are determined depending on symptoms ranged from short episodes to chronic conditions [1–5].

Few have analyzed types of anemia, although it has received much attention recently because of epidemiological studies suggesting that anemia may be associated with worse outcomes in various diseases [5–9].

In the literature, there have been several methods [8–16] to analyze anemia types. Besides to their notable advantages, many common properties, such as being costly, difficulty in usage, time-consuming and having constraints in daily usage, lead to their drawbacks. In this case, optimization modeling should be taken into consideration alongside those methods. It is obviously discovered that by taking into account the mechanics of computer hardware, it can be created novel hybrid approaches that are much more effective for certain calculations as outlined

*Corresponding Author
in the literature [17–24], even though these concepts may appear to be unrelated to physical limitations.

A more efficient algorithm can complete any computation much faster than any inefficient algorithm, even without changing the computer hardware. Therefore, advancements in algorithms can be used to make computers faster, such as pre-computing parts of a problem or other solutions to improve computing efficiency.

Therefore, there are particular techniques for optimization, including particle swarm optimization (PSO) which is becoming increasingly popular as an effective method for many data processing tasks. In this work, the hybrid algorithm of PSO with another important optimization algorithm, support vector machine (SVM), has been created to predict types of anemia. To the best of the authors’ knowledge, this is the first time a combination of PSO and SVM has been used to forecast anemia types using biological data, despite the fact that certain conventional methods are utilized for evaluating anemia types in the literature [7–15]. The PSO-SVM algorithm has been suggested for reliable data treatment and further analysis for interpretation due to its high level of flexibility and lack of requirement for specialized knowledge in statistics [25–27].

The PSO-SVM is a comprehensive optimization clustering technique as a consequence, enabling the estimation of different patterns using the data that is currently available in an area. By forecasting the clusters that are most likely to include individuals who have that form of anemia, it is able to predict anemia types. Investigating pathological individuals from a population was the goal of this study, and using advanced computers to implement the PSO-SVM technique has been attractive [10], [28].

2. Materials and methods

2.1. Biomedical data collection and study design

The dataset has been prepared from observations of anemia disease. The dataset that consists of the anemia measurements contains 5 classes and 539 subjects provided from the literature [9], [11], [12] whose main concerns are creating mathematical models that can predict the type of anemia based on various blood variables by only specific algorithms while in the current study, a hybrid algorithm has been developed using pathological individuals from a population and their corresponding blood variables. The data has been collected from individuals between aged 6 to 56 years. Each sample includes information from blood variables; Red Blood Cells (RBC), a portable protein Hemoglobin (HB) inside the RBC containing iron atoms which carry oxygen from the lungs to the body’s tissues and return carbon dioxide from the tissues back to the lungs, Hematocrit (HCT) which shows the percentage of RBC in the blood, Mean Corpuscular Volume (MCV) which measures average size of the red cells in a sample and [8]. In addition to these, some biophysical variables, sex and age, have also been considered since natural HB in the body varies from male (1) to female (2) and natural HB in the body varies according to age.

The best four variables have been selected from the prediction of anemia through the PSO and machine learning, and are mentioned in Table 1 and Figures [1] and Figures [1]. The relationship between different parameter values of the subjects is illustrated in Figure 2.
The goal of this research is to detect anemia in a population through the utilization of PSO-SVM algorithm and analysis of blood parameters. Since the hybrid algorithm is based on clustering principle, the five clusters have been generated by blood variables as detailed in Table 1 and Figure 3. This study has classified anemia types into five separate clusters based on common characteristics and danger levels. The attributes of anemia dataset can be seen in Table 2. The proposed approach has been implemented on the collected data to identification a healthy or infected person out of involved 539 subjects.

2.2. Particle swarm optimization

Kennedy and Eberhart developed an evolutionary and population-based optimization technique referred to as the particle swarm optimization (PSO) [29] by drawing inspiration from the collective behavior of birds and fish. So, these animals played an important role in the development of an algorithm by escaping dangerous situations and finding food. The PSO is much faster and more effective than other optimization techniques because it requires fewer parameters and is less likely to be stuck in local minimum points as a possible solution.

The PSO seeks to improve a solution to a problem by having each particle trace its coordinates in terms of the best solutions that have already been made. The particle keeps track of its own best, known as pbest, and the best solution in the near surroundings, known as lbest. If the particle considers the entire population for its topological neighbors, then the best value becomes the global best, or gbest. At every step, the particles’ velocities are adjusted according to the pbest and lbest locations.

In the PSO, software agents known as particles move through the related space to find improved results. At each step, the randomly selected particles adjust their velocities using data from their local areas, their own neighborhoods and from randomness in order to search for better spots in the solution space. The position of a particle is an indication of a potential solution to the problem. At the conclusion of each cycle, all particles attempt to find more desirable locations in the search space by altering their speed. For each iteration, the position and velocity vectors have been determined as follows:

\[
V_{p}^{\text{new}} = wV_{p}^{\text{old}} + c_{1} \text{rand}_{1}(P_{\text{best}} - X_{p}^{\text{old}}) + c_{2} \text{rand}_{2}(gbest - X_{p}^{\text{old}}) \tag{1}
\]

\[
X_{p}^{\text{new}} = X_{p}^{\text{old}} + V_{p}^{\text{new}} \tag{2}
\]

where \(w\) represents corresponding weights, \(c_{1}, c_{2}\) are acceleration coefficients (cognitive parameter, social parameter), \(\text{rand}_{1}, \text{rand}_{2}\) are uniformly distributed random numbers between 0 and 1, \(V_{p}^{\text{old}}\) gives velocity of individual \(p\) at the iteration, \(X_{p}^{\text{old}}\) determines position of individual \(p\) at the current iteration, \(P_{\text{best}}\) and \(gbest\) indicate the best local value of each particle and the best value of swarm, respectively [30], [31].
Table 1. Clusters and subject numbers.

<table>
<thead>
<tr>
<th>Cluster no.</th>
<th>MCV</th>
<th>HCT</th>
<th>RBC</th>
<th>HB</th>
<th>Number of Subjects</th>
<th>Anemia Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>≥ 100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>Deficiency B12</td>
</tr>
<tr>
<td>2</td>
<td>&lt; 100</td>
<td>≥ 48</td>
<td>-</td>
<td>-</td>
<td>230</td>
<td>Thalassemia</td>
</tr>
<tr>
<td>3</td>
<td>&lt;100</td>
<td>&lt; 48</td>
<td>≤ 4.5</td>
<td>-</td>
<td>23</td>
<td>Sickle Cell</td>
</tr>
<tr>
<td>4</td>
<td>&lt; 100</td>
<td>&lt; 48</td>
<td>&gt; 4.5</td>
<td>&lt; 11</td>
<td>128</td>
<td>Iron deficiency</td>
</tr>
<tr>
<td>5</td>
<td>&lt;100</td>
<td>&lt; 48</td>
<td>&gt; 4.5</td>
<td>≥ 11</td>
<td>148</td>
<td>Normal</td>
</tr>
</tbody>
</table>

Table 2. Attributes of anemia dataset.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Attribute value</th>
<th>Attribute category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>6-56</td>
<td>0-13 Child, &gt;13 Adult</td>
</tr>
<tr>
<td>Gender</td>
<td>1, 2</td>
<td>1 = Male, 2 = Female</td>
</tr>
<tr>
<td>HB</td>
<td>1.46-18.2</td>
<td>&lt; 11 Severe, 11-15 Moderate</td>
</tr>
<tr>
<td>RBC</td>
<td>0.96-11.9</td>
<td>&lt; 4.5 Low, 4.5-6.5 Normal, &gt; 6.5 High</td>
</tr>
<tr>
<td>MCV</td>
<td>38.6-117</td>
<td>&lt; 80 Microcytic, 80-100 Normocytic, ≥ 100 Macrocytic</td>
</tr>
<tr>
<td>HCT</td>
<td>7.7-51.7</td>
<td>&lt; 35 Low, 35-48 Normal, ≥ 48 High</td>
</tr>
</tbody>
</table>

3. Support vector machine

Vladimir Vapnik began examining Support Vector Machines (SVMs) in the late 1970s, but in the late 1990s, the field began to gain extreme recognition [32]. The SVMs are supervised learning algorithms operated by statistical learning theory to identify patterns and perform suitable regression. Statistical learning theory can accurately identify the components needed for successful learning of certain basic algorithms. However, due to the complexity of more intricate models and algorithms used in real-world applications, it is difficult to analyze them theoretically.

The SVMs can be thought of as a combination of learning theory and practicality, which is simple enough to be understood mathematically. This is due to the fact that an SVM can be regarded as a linear model in a high-dimensional space and the SVMs are able to go beyond the limitations of linear learning machines by incorporating a kernel function, allowing for the discovery of a nonlinear decision function [33].

4. The PSO-SVM algorithm

The process of feature selection can be regarded as a challenge of global combinatorial optimization in machine learning, wherein the amount of features is reduced and also unimportant, noisy and redundant information is eliminated to achieve a satisfactory classification accuracy. The significance of this method is immense in the fields of pattern recognition, medical data analysis, machine learning, and data mining. In order to increase processing rate, accuracy, and reduce incomprehensibility, a reliable feature selection method that takes into account the number of features studied for sample classification is necessary.

In this study, a binary version of the PSO algorithm has been used and the placement of each particle is indicated through a binary string that symbolizes the feature choice situation. The position and velocity of each particle is revised in accordance with the following equations:

\[
V_{new} = wV_{old} + c_1 \text{rand}_1 (P_{best} - X_{old}) + c_2 \text{rand}_2 (g_{best} - X_{old})
\]

\[
S(V_{new}) = \frac{1}{1 + e^{-V_{new}}}
\]

\[
\text{if}(\text{rand} < S(V_{new})) \text{then } X_{new} = 1; \text{ else } X_{new} = 0
\]

where \(V_{new}\) and \(V_{old}\) are the particle velocities, \(X_{old}\) shows the current position, and \(X_{new}\) represents the updated position for the related solution. The values \(P_{best}\) and \(g_{best}\) are defined as local and global best fitness value. The \(\text{rand}_1\) and \(\text{rand}_2\) are randomly generated numbers between 0 to 1, whereas \(c_1\) and \(c_2\) are acceleration factors, usually chosen as \(c_1 = c_2 = 2\). Velocities for each dimension have been tried to reach a maximum velocity \(V_{max}\). If the combined velocities of a given dimension add up to more than the predetermined value of \(V_{max}\), the velocity of that dimension will be restricted to \(V_{max}\), which is a value set by the user.

After renewal, the new feature is calculated as in Eq. 4, where \(V_{new}^d\) represents the velocity. If calculated value \(S(V_{new}^d)\) is greater than a randomly
generated number that is between (0, 1), then its position value $F_n, n = 1, 2, \ldots, m$ is represented as 1 by meaning that this feature is chosen to be a required feature for the upcoming renewal. Otherwise, the value is represented as 0.

A one-layer SVM model has only the capability of distinguishing between two types of anemia, as it is a binary classifier. Because of this limitation, in this study, a four-layer SVM classifier, as illustrated in Figure 4, has been used to predict anemia, which includes four disease states and one normal state.

**Figure 4.** Four-layer SVM classifier.

The anemia dataset has been randomly split into two groups: 80% for training samples and 20% for testing samples. Before the application, the PSO algorithm has been used to determine the best combination of parameters $(c, \delta)$ for each SVM based on the training samples. The testing samples have verified the effectiveness of the multi-layer SVM classifier.

An illustration of the setup of the PSO-SVM model for predicting anemia is presented in Figure 5.

**Figure 5.** Flowchart of the prediction of anemia.

The results of the anemia prediction from the testing samples are presented in Table 3. Comparing the two models, the PSO-SVM exhibits a higher overall accuracy than the SVM.

**5. Results and discussion**

The developed PSO-SVM algorithm has intended to predict anemia outcomes based on the testing samples presented in Table 3 and Figure 6. It has greater overall accuracy than the SVM algorithm, and has been attempted to locate the blood variables of the clustered data for each type of anemia. The algorithm assesses whether the data for each anemia type is categorized precisely.

**Figure 6.** Comparison of the PSO-SVM and SVM algorithms by general accuracy (%).

It can be a wise move to observe the distribution of data both practically and computationally. The comparison between the computed results and the actual values has been displayed in Figures 7–8. It is evident that the calculated results of all clusters are in agreement with the actual results.

As seen in Table 3 and Figure 6, the produced results of the PSO-SVM algorithm are generally good outperform with the results of the SVM algorithm. For example, the accuracy of the PSO-SVM and the SVM algorithms are 80% and 70% and success set 8 and 7 people for cluster 1, respectively. Similarly, in the same order, the algorithm’s accuracy of the cluster 2 are 98.26% and 96.52% and success set 226 and 222 people. The two algorithms generated results of 21 and 20 individuals for Cluster 3, which is 3 lower than the actual number of 23, and so the accuracies are 91.3% and 86.95%, respectively. The success set of the algorithms in the cluster 4 are 123 and
Table 3. Comparison of the PSO-SVM and SVM algorithms by general accuracy (%)  

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Algorithms</th>
<th>Test Set</th>
<th>Success Set</th>
<th>Fail Set</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>SVM</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>PSO-SVM</td>
<td></td>
<td>8</td>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>C2</td>
<td>SVM</td>
<td>230</td>
<td>222</td>
<td>8</td>
<td>96.52</td>
</tr>
<tr>
<td></td>
<td>PSO-SVM</td>
<td></td>
<td>226</td>
<td>4</td>
<td>98.26</td>
</tr>
<tr>
<td>C3</td>
<td>SVM</td>
<td>23</td>
<td>20</td>
<td>3</td>
<td>86.95</td>
</tr>
<tr>
<td></td>
<td>PSO-SVM</td>
<td></td>
<td>21</td>
<td>2</td>
<td>91.3</td>
</tr>
<tr>
<td>C4</td>
<td>SVM</td>
<td>128</td>
<td>120</td>
<td>8</td>
<td>93.75</td>
</tr>
<tr>
<td></td>
<td>PSO-SVM</td>
<td></td>
<td>123</td>
<td>5</td>
<td>96.09</td>
</tr>
<tr>
<td>C5</td>
<td>SVM</td>
<td>148</td>
<td>140</td>
<td>8</td>
<td>94.59</td>
</tr>
<tr>
<td></td>
<td>PSO-SVM</td>
<td></td>
<td>144</td>
<td>4</td>
<td>97.29</td>
</tr>
</tbody>
</table>

120 people and the accuracies 96.09% and 93.75%. Also, for the cluster 5 the success set are 144 and 140 and the accuracies 97.29%, 94.59%, respectively.

The five groups with different parameters without classification can be seen in Figure 7. On the
other hand, the five groups with different parameters after classification can be seen in Figure 8.

The results of Figure 8 demonstrated that the algorithm used in the research was designed to determine the correct number of subjects in the clusters, as well as the anemia types in those clusters. By utilizing the PSO-SVM algorithm, the accuracy of the data clustering and the effects of biomedical information on the anemia types have been examined. In literature, various versions of the PSO-SVM algorithm have been developed for different issues in multiple scientific fields. In comparison to the literature, the PSO and SVM clustering algorithms have been observed to be extremely successful for their own concerns [29–31, 34–43]. This study successfully created a combined version of the algorithm, as well as its computer code.

In this study, the PSO and SVM algorithms and their combination have been used for the first time to detect anemia types. One of the most significant contributions of this study is its application of the algorithm to anemia data for the first time. It has been examined whether it would be successful in this area, like in other areas of science. The combination of the SVM and the PSO resulted in very effective outcomes in the investigation of the anemia types. This application in this field can assist clinicians in predicting the anemia types.

The goal of the clustering was to distinguish the anemia types into normal or pathological classes accurately. The number of clusters for the algorithm was given by the user. Subjects were split into 5 clusters based on their blood variables. The performance of the hybrid algorithm was determined by using the blood variables to predict the anemia types. This demonstrates that the proposed algorithm has been seen to be more effective when there is a well-structured algorithm with a sufficient amount of data.
6. Conclusions

This research has evaluated the feasibility of the PSO-SVM algorithm to predict anemia using biomedical variables. This was the first attempt to use this newly combined approach to predict anemia. The results generated by the PSO-SVM algorithm have been compared to the actual values and have been found to be highly effective in enhancing anemia predictions through biomedical factors. The findings suggest that this method could be clinically valuable for creating suitable treatment programs for patients. More clusters could be used for further research, as the current data structure is limited in terms of medical analysis point of view.

Acknowledgments

The fourth author thanks Dr Aniela Balacescu (Constantin Brancuși University) for her hospitality throughout the conference. The authors also would like to thank the anonymous reviewers for their valuable comments.

References


Arshed A. Ahmad born in Iraq in 1980. He is an Assistant Professor at the University of Diyala,
Iraq. He received a Ph.D. degree from Yildiz Technical University in 2020. His research interests include mathematical modeling, applied mathematics, numerical analysis, and probability and statistics. He is the author of more than 14 research papers since 2008. [https://orcid.org/0000-0003-1393-1253](https://orcid.org/0000-0003-1393-1253)

**Khalid M. Saffer** born in Iraq in 1971. He is a lecturer at the University of Diyala, Iraq. He received a Ph.D. degree from Altinbas University in 2018. His research interests include optimization algorithms, information technology, image processing, information security, and database systems design and implementation. He is the author of more than 12 research papers since 2011. [https://orcid.org/0000-0001-8670-4161](https://orcid.org/0000-0001-8670-4161)

**Murat Sari** is a Professor of the Department of Mathematical Engineering at Istanbul Technical University. He completed his Ph.D. in Mathematics from the University of South Wales, UK, in 2000. His current research interests include artificial intelligence modeling, simulation and computational methods, computational fluid dynamics, modeling of nonlinear behaviors, economical modeling, and biomechanical/biomedical modeling. He has about 100 high-quality scientific papers, over 50 conference proceedings, and has written various chapters in some books. He is a reviewer/editor for some international high-quality journals. [https://orcid.org/0000-0003-0508-2917](https://orcid.org/0000-0003-0508-2917)

**Hande Uslu** is a Research Assistant in the Department of Mathematics at Yildiz Technical University, Istanbul, Turkey since 2017. She received her B.Sc degree in Mathematics from Bogazici University in 2015, and MSc degree in Applied Mathematics from Yildiz Technical University, in 2018. She is currently pursuing her Ph.D. studies in the thesis period in the Department of Mathematics at Yildiz Technical University. Her research interests are mathematical modeling, simulation techniques, stochastic processes, and artificial intelligence. She has six scientific papers and more than fifteen conference proceedings. [https://orcid.org/0000-0002-1642-1120](https://orcid.org/0000-0002-1642-1120)

An International Journal of Optimization and Control: Theories & Applications (http://www.ijocta.org)

This work is licensed under a Creative Commons Attribution 4.0 International License. The authors retain ownership of the copyright for their article, but they allow anyone to download, reuse, reprint, modify, distribute, and/or copy articles in IJOCTA, so long as the original authors and source are credited. To see the complete license contents, please visit http://creativecommons.org/licenses/by/4.0/.