|  |
| --- |
| **REPUBLIC OF TURKEY**  **ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY**  **FACULTY OF ENGINEERING**  **DEPARTMENT OF COMPUTER ENGINEERING**  **FEATURE SELECTION USING BINARY PSO IN CYBERBULLYING DATASET**  **ZEYNEP ALTIPARMAK**  **FATMA NUR ARSLAN**  **ADANA 2022** |



**REPUBLIC OF TURKEY**

**ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY**

**FACULTY OF ENGINEERING**

**DEPARTMENT OF COMPUTER ENGINEERING**

**FEATURE SELECTION USING BINARY PSO IN CYBERBULLYING DATASET**

**ZEYNEP ALTIPARMAK**

**FATMA NUR ARSLAN**

**SUPERVISOR**

**Asst. Prof. Dr. Esra SARAÇ EŞSİZ**

**ADANA 2022**

**ABSTRACT**

**FEATURE SELECTION USING BINARY PSO IN CYBERBULLYING DATASET**

Zeynep ALTIPARMAK

Fatma Nur ARSLAN

Department of Computer Engineering

Supervisor: Asst. Prof. Dr. Esra SARAÇ EŞSİZ

February 2022

Feature selection is the process by which a subset of relevant features, or variables, are selected from a larger data set for constructing models. In this paper, our aim is to apply a recent optimization technique namely the Binary Particle Swarm Optimization (BPSO) algorithm to select best features for cyberbullying. To investigate the results of our BPSO-based feature selection method, we measured the performance of our method using the kNN classification algorithm. For our dataset 70% accuracy have been achieved.

**Keywords:** feature selection, nature inspired, optimization, machine learning.

**TABLE OF CONTENTS**

[TABLE OF CONTENTS](#_Toc526205815) i

[LIST OF TABLES](#_Toc1748158938) ..ii

[Table 1. Confusion Matrix...................................................................…..](#_Toc263971075) ii

[Table 2. Results...................................................................................…..](#_Toc1817237970) ii

[1. INTRODUCTION](#_Toc934884936) 1

[2. MATERIALS AND METHODS](#_Toc1262443707) 2

[2.1. Feature selection](#_Toc1157666832) 2

[2.2. Performance measures](#_Toc1362838989) 2

[2.3 Dataset](#_Toc59553838) 3

[2.4 PSO(Particle Swarm Optimization) and Binary PSO](#_Toc1107217557) 4

[2.5 K-nearest neighborhood (kNN)](#_Toc1725392877) 6

[3. RESULTS](#_Toc1337427777) 8

[4. CONCLUSIONS](#_Toc1227861038) 9

[REFERENCES 1](#_Toc462081142)0

**LIST OF TABLES**

**Table 1.** Confusion Matrix...................................................................…..

**Table 2.** Results...................................................................................…..

1. **INTRODUCTION**

The rapid development of today’s technologies and the active use of technology by people from all age groups brings with it many problems. The most popular of these is cyberbullying. Cyberbullying is bullying that takes place over digital devices like cell phones, computers and tablets.

Cybebullying can occur through SMS, text, and apps or online in social media, forums or gamin where people can view, participate in or share content. It includes sending, posting or sharing negative, harmful, false or mean content about some one else. It can include sharing personal or private information about someone else causing embarrassment of humiliation. Some cyberbullying crosses the line into unlawful or criminal behavior.

In this study, our aim is to research the success of BPSO based feature selection algorithm on classification algroithms for cyberbullying.

1. **MATERIALS AND METHODS**
   1. **Feature selection**

Feature selection is the process of increasing the performance success of the classification model by selecting the most useful features on the existing data set.

* 1. **Performance measures**

For a classification problem, the success of the model is important. The success of this model is determined by the number of correct and incorrect predictions obtained from all predictions. Confusion matrix is used to measure the success of the model. The confusion matrix has 4 parameters; true positives, true negatives, false positives and false negatives.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ACTUAL** | | |
| **PREDICTED** |  | *POSITIVE* | *NEGATIVE* |
| *POSITIVE* | TP | FP |
| *NEGATIVE* | FN | TN |

**Table 1**. Confusion matrix

Accuracy is used to measure the ratio of accurately estimated samples to the total number of samples. It can be considered that the high accuracy means successful model(1). Error rate is used to measure the ratio of the values of incorrectly estimated samples to the total number of samples(2). Precision is ratio of correctly classiﬁed positive samples to estimated total positive samples(3).

Recall is used to measure the proportion of positive values classiﬁed as true(4). F-score is harmonic mean precision and recall. I takes into account both false positives and false negatives. F-score is more reliable than accuracy(5).

(1)

(2

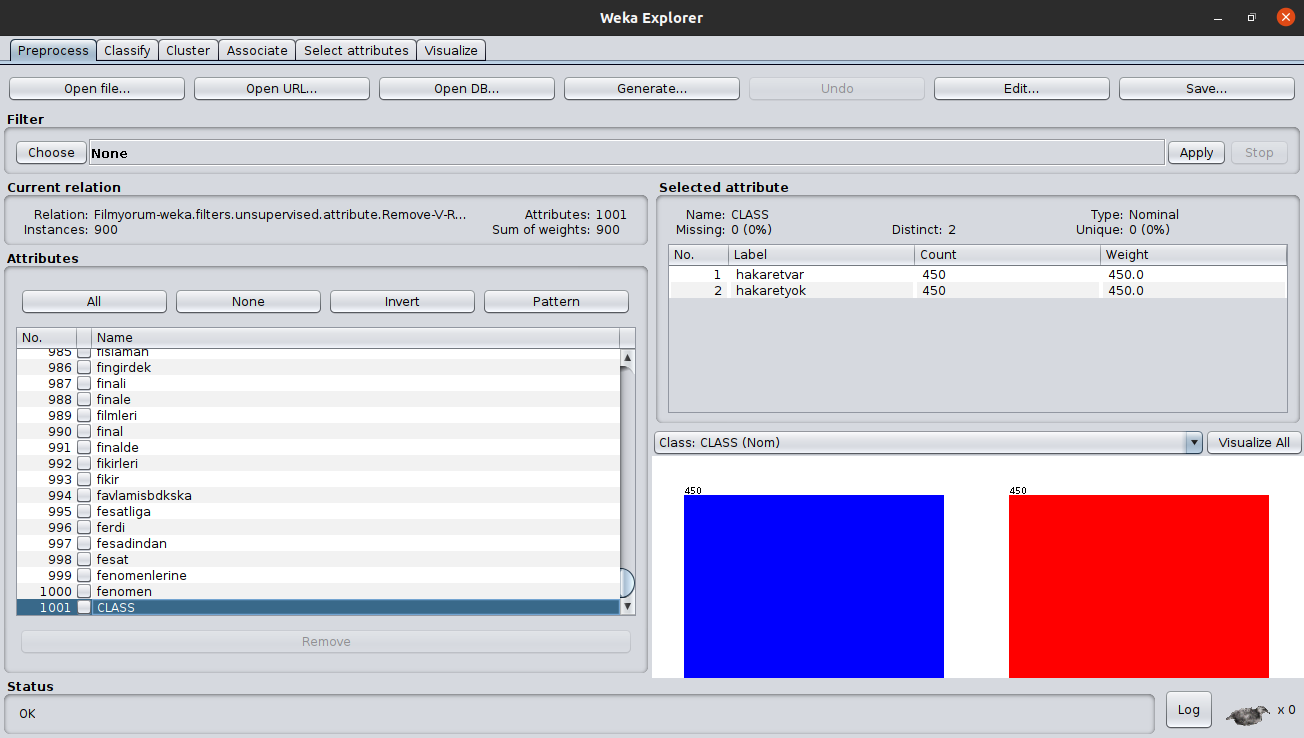
(3)

(4)

(5)

**2.3 Dataset**

We work on the Turkish Twitter dataset without emoji. The dataset contains 900 instances and 1000 features that consists of two different classes (positive and negative).



**2.4 PSO(Particle Swarm Optimization) and Binary PSO**

Particle Swarm Optimization is a population-based metaheuristic algorithm, it was introduced by Kennedy and Eberhart[1] in 1995, that is motivated by the simulation of social behavior such as birds, fishes. PSO is primarily designed for the solution of nonlinear problems and is frequently used to find solutions to multivariate and multiparameter optimization problems.

The PSO was designed by simulating the behavior of flocks of birds while foraging. While flocks of birds do not know exactly where the food is, they try to find out how far away they are from it. For this process, they always follow the bird closest to the food. In PSO, however, each bird is represented as a particle. Particles are constantly in motion and their positions are recalculated after each movement. Each particle must know its position, its velocity, its best position in the solution space. The position and velocities of each particle in the solution are calculated by combining their best coordinates with the best coordinates of other particles.

Binary PSO is a form of PSO applied to binary domains. In the Binary PSO algorithm, each particle represents its position in binary values, 1 and 0. In the binary PSO, the particle’s personal best and global best is updated as in continuous version.

The major difference between binary PSO with continuous version is that velocities of the particles are rather defined in terms of probabilities that a bit will change to one.

**Step 1: Initialization**

• Set *k=0, n=*twice the number of dimensions

• Generate *n* particles randomly as explained before,

• Generate *n* particles randomly as explained before, *{xi⁰ i=1, 2, ….,n}*, where *xi⁰=[ xi1⁰, xi2⁰,…., xid⁰]*.

• Generate the initial velocities of all particles randomly, *{vi⁰ i=1, 2, ….,n}*, where

*vi⁰=[ vi1⁰, vi2⁰,….,vid⁰]*. *vid⁰* is generated randomly with *vid⁰= vimin⁰+ ( vmax⁰- vmin⁰)\*rand()*

• Evaluate each particle in the swarm using the objective function, *f( xi⁰).*

• For each particle *i* in the swarm, set *PBi⁰ = xi⁰*, where *PBi⁰=[ PBi1⁰=xi⁰, PBi2⁰=xi2⁰,…., Pbid⁰=xid⁰]* along with its best fitness value,  *fi =( PBi⁰ ,i=1, 2,….,n)*

pbest

• Set the global best to,

*fi (GB⁰ ) = min{ fi ( PBi⁰ ,i=1, 2,….,n)} with GB⁰ =[gb1, gb2,…., gbd]*

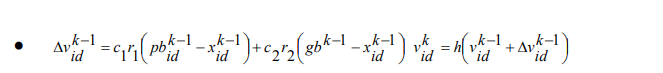
pbest

pbest

**Step 2: Update iteration counter**

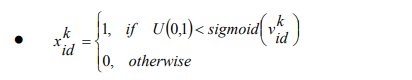
• *k = k +1*

**Step 3: Update velocity by using the piece-wise linear function**



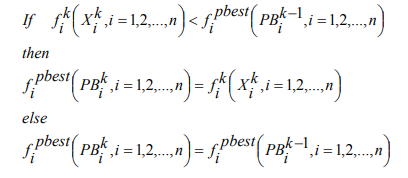
•c1 and c2 are social and cognitive parameters r1 and r2 are uniform random numbers between (0,1)

**Step 4: Update dimension (position) by using the sigmoid function**

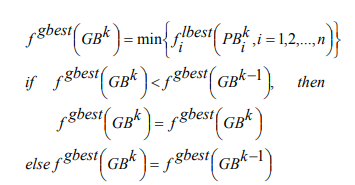


**Step 5: Update particle best**

Each particle is evaluated again with respect to its updated position to see if particle best will change. That is,



**Step 6: Update global best**



**Step 7: Stopping criterion**

If the number of iteration exceeds the maximum number iteration, then stop, otherwise go to step2. [2]

**2.5 K-nearest neighborhood (kNN)**

KNN is a supervised machine learning algorithm used in classification. In the kNN algorithm, similar things are close to each other. The algorithm works with two main parameters:

***Distance:*** *The distance of the point to be estimated from other points is calculated. Usually, the Minkowski distance calculation function is used for this.*

***k(number of neighborhood):*** *The number of neighbors determines how many nearest neighbor points will be used for calculations. Each different k values results in different results. therefore, determining the optimal k value is important to obtain successful results. Generally, the k value is taken as 5 or 10.*

1. **RESULTS**

In the research, using kNN classification model and feature selection were made with different features for data set. The algorithm was run 10 iterations for data set. In each iteration, the accuracy, error rate and f-score values of 10 randomly selected feature swarms were kept. Again in each iteration, these values were averaged and the performance measures were updated, and the result was obtained.

|  |  |
| --- | --- |
| Performance Measures | *Binary PSO* |
| *Accuracy* | 70% |
| *Error rate* | 0.28 |
| *F-score* | 68% |

**Table 2**. Results

1. **CONCLUSIONS**

In this study, we have tried to understand and developed an swarm-based, nature inspired BPSO based feature selection system for the detection of cyberbullying. We run our experiments using the Turkish Twitter dataset without emojies.

The feature selection algorithm used measures the accuracy of the classification model with the obtained F-score values and whether the data distribution is regular or not. When these values are examined, it is seen that the model is successful at an average of 70% and the data distribution in the data set is regular in parallel with this value.

**REFERENCES**

[1] Kennedy, J. and Eberhart, R.C., "A discrete binary version of the particle swarm algorithm", Systems, Man, and Cybernetics, 1997, vol.5, 4104-4108.

[2] M. Fatih Taşgetiren & Yun-Chia Liang “A Binary Particle Swarm Optimization Algorithm for Lot Sizing Problem” , Journal of Economic and Social Research 5 (2), 1-20 , · January 2004

[3] https://www.mathworks.com/matlabcentral/fileexchange/71470-binary-particle-swarm-optimization-for-feature-selection