Feature Weighting Using a Clustering Approach

Mohammad Dousthagh, Mousa Nazari, Amir Mosavi, Shahaboddin Shamshirband, Anthony T. Chronopoulos

***Abstract*— In recent decades, the volume and size of data has significantly increased with the growth of technology. Extracting knowledge and useful patterns in high-dimensional data are challenging. In fact, unrelated features and dimensions reduce the efficiency and increase the complexity of machine learning algorithms. However, the methods used for selecting features and weighting features are a common solution for these problems. In this study, a feature weighting approach is presented based on density-based clustering. This method has been implemented in two steps. In the first step, the features were divided into clusters using density-based clustering. In the second step, the features with a higher degree of importance were selected in accordance to the target class of each cluster. In order to evaluate the efficiency, various standard datasets were classified by the feature selection and their degree of importance. The results indicated that the simplicity and suitability of the method in the high-dimensional dataset are the main advantages of the proposed method.**

***Index Terms*—Feature selection, feature clustering; feature weighting; density-based clustering, machine learning, big data.**

I. Introduction

Machine learning algorithms are required to enable the system to learn new information from the existing data and respond to the new needs. If these algorithms are used in large scales, they will have a higher cost for the system; however, not all features are useful, and some are repetitive or redundant. These repetitive features will lead to the reduction of the accuracy of machine learning algorithms. For this purpose, some features should be selected which have a greater impact on the issue. There are a few algorithms called "feature selection algorithms" which can eliminate the repetitive and redundant features. However, the elimination of these features has a higher cost for the system which cannot be ignored, and weights are assigned values between zero and one. Any features which are closer to the target class have a weight closer to one and any features which are far from the target class have a weight closer to zero and the total of these weights should equal to one. During the last decade, a large number of studies were conducted on the feature selection as follows:

Liu *et al.* (2011) conducted a study on feature selection using a hierarchical clustering of features. The main idea of this method was based on clustering. A new algorithm was provided called FSFC using some criteria such as information and filters, and the advantages of the above-mentioned methods. This algorithm was selected when it had the most connection and the least repetition [1]. In another study, Peng *et al.*(2017) focused on a fast feature weighting algorithm of data gravitation classification. In this study, the features were evaluated by discrimination [Please choose another word for discrimination] and redundancy, and two fuzzy subsets were used. These two sets were solved by Mutual Information (MI) and the Pearson analysis [2]. Eshaghi and Aghagolzadeh (2016) worked on a clustering-based feature selection. In this method, the features were first clustered using the DBSCAN algorithm, and then the representative element from each cluster was selected [3]. Polat (2012) emphasized the classification of Parkinson's disease by using weighting features based on Fuzzy c-means (FCM) clustering and presented a FCM-based method in order to transform the continuous data and discrete data and enhance the efficiency of class differentiation. For this purpose, the center of each cluster was determined by each feature, and then the ratio of these centers was calculated. In this method, the variance in the classes decreased and the difference between classes increased [4].

Modha *et al.* (2003) worked on a weighting feature based on k-means clustering. The study aimed to 1) provide each data a group of multi-feature vectors, 2) assign the measurement of a suitable (and possibly different) complexity to each spatial feature, 3) combine the complexities in a different spatial feature by assigning a weight to each feature, 4) fix the correspondence weight-to-feature of the proposed convex k- means algorithm, and 5) adapt weighing to the optimal features [5]. Sun (2007) investigated the ideal relief for feature weighting. The present study used the mathematical logic of the RELIEF algorithm to present the ideal RELIEF algorithm called I-RELIEF [6]. Dialameh and Jahromi (2017) worked on the proposed general feature weighting function. In this study, a dynamic weighting was presented to be dynamically sensitive to the effect of the features. For this purpose, a dynamic feature weighting function was presented to assign a proper weight to each feature automatically [7]. In another study, Lu et al. (2017) presented a hybrid feature selection algorithm for gene expression data classification. In this study, a hybrid method was introduced for feature selection which combined

Manuscript received October 31, 2018; revised April 12, 2019.

Mohammad Dousthagh, Mousa Nazari are with the Department of Computer Engineering, Faculty Engineering, Rouzbahan Institute of Higher Education, Sari, Iran.

Amir Mosavi is with School of the Built Environment, Oxford Brookes University, Oxford, UK, and Institute of Automation, Kando Kalman Faculty of Electrical Engineering, Obuda University, and Budapest, Hungary.

Shahaboddin Shamshirband is with the Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Viet Nam. And Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Viet Nam, S.S. is the Corresponding author. (e-mail: shahaboddin.shamshirband@tdtu.edu.vn).

Anthony T. Chronopoulos is with the Department of Computer Science, the University of Texas at San Antonio and (Visiting Faculty) Department of Computer Science, University of Patras, Greece.

two algorithms (MIM and AGA) resulting into the proposed MIMAGA-selection algorithm, which significantly reduced the dimensions of the data and eliminated the redundancies [8]. DAS et al. (2017) worked on a hybrid feature selection by using the Feature Association Map for supervised and unsupervised classifications. In this study, a hybrid method was presented for feature selection based on a graph-based approach. The proposed algorithm used the concept of FAM as the basis of the work [9].

As already mentioned, the feature selection has attracted a lot of attention due to its great importance and it aims to select the smallest subset with the least error and cost. In this regard, many algorithms have been provided which have their own advantages and disadvantages. The present study seeks to select the related features associated with the issue which are more influential. In fact, it aims at a feature weighting based on the priority and proximity to the weighted target class. In this regard, the features were first clustered using the clustering algorithm and then a representative was selected from each cluster, and these representatives were weighted as inputs to the weighting function with values from zero to one.

This algorithm was evaluated on a different data set which had a better result, compared to other feature selection algorithms in terms of classification accuracy. For example, the Parkinson dataset could be correctly identified 97% with KNN classification. The result of other data sets is presented in the following sections.

We listed some feature selection algorithm above but each one has problem in selection for example the relief algorithm is not optimum and cannot recognize redundant features. Also, F-DBSCAN determines a single node as a noise node because this algorithm uses simple DBSCAN for clustering and it uses a single node (without any neighbor) as noise. Some of above algorithm use a simple selection method but in our proposed algorithm we finally weighting features.

[Please write 1-2 sentences to say what the new algm does]

The rest of the paper is organized as follows. In Section II, the proposed methods are described. In Section III, evaluation results are presented. Section IV presents conclusions and future work.

II. The Proposed Method

As mentioned before, many studies were conducted on feature selection and each had their own shortcomings such as extensions, ignoring neighboring features and so on. The present study seeks to provide a method having the important features and eliminating the above-mentioned shortcomings.

The proposed WF-DBSCAN algorithm, first, clusters the features through the DBSCAN algorithm. The logic of the DBSCAN algorithm indicates that nodes (features) which are similar, and their numbers equals to the number of m inputs to eps radius are considered as a cluster. Otherwise, it is known as "noise". There are many scales to detect this similarity including the Euclidean distance. However, having no neighboring feature implies that there is no effect on the issue and should be recognized as noise. This is definitely not true. A feature having no neighbor may have a feature having a great impact on the issue. Thus, it is not easy to be judged. Therefore, in the first step, the proposed modify-DBSCAN algorithm is presented to consider non-neighboring features as a separate cluster. The modify-DBSCAN algorithm is presented next.

```
For i=1:size(input)
  If(!visited(i))
    Visited(i)=true;
    Neighbors=find(D(i,:))<=epsilon
    clusterNum++;
    ExpandCluster(I,Neighbors,clusterNum)
end
```

Fig. 1. Modify-dbscan algorithm.

As shown in Fig. 1, modify-DBSCAN algorithm is changed. However, the part of DBSCAN that checks noise, is not needed, as there is no noise and may be an important feature that is ignored.

As already mentioned, the modify-DBSCAN algorithm includes minimum number of points in cluster (minpts) and maximum radius of the neighborhood (eps) parameters. In this study, the values of these two parameters are equal to 2 and 0.5, respectively. In the next step, the modify-DBSCAN algorithm was used to cluster the features. The features were first inserted in the modify-DBSCAN algorithm to compute the Euclidean distance of each feature from the others and then they were clustered. When the clusters are identified, a feature for each cluster should be considered as the representative of that cluster. This candidate feature should have the most dependency on the target class and there should be the least redundancy among the other features. Then, each representative is sent to the weighing function and a calculated weight multiplied by the labeled dataset because weighted matrix influence to the labeled dataset is assigned for and according to the importance of the feature based on the weight and its effect on the data set.

In order to achieve the candidate feature, the relationship of each feature to the class is obtained through the following formula [3]:

$$J(fi)=\frac{SU(fi.c)}{avg\left(SU(fi.F)\right)+std\left(SU(fi.F)\right)} \tag{3-1}$$

where SU (fi.c) represents the uncertainty criterion between the fi feature and the class C, avg(SU(fi.F)) indicates the mean, and std(SU(fi.F)) is the standard deviation. In these two formulas fi means number i feature and F means the set of features in other clusters.

The following formula is used to obtain the uncertainty criterion [3]

$$SU(x.y)=\frac{2I(x.y)}{H(x)+H(y)} \tag{3-2}$$

In the next step, the largest fi feature is regarded as the representative.

$$Fi=\max\left(J(fi)\right)|\ fi\in c \tag{3-3}$$

Therefore, the features of representative have the highest dependency on the target class and the lowest redundancy.

In the final step, a weight is assigned to the features

selected in the previous step:

$$Wi = \frac{Si}{\sum_{i=1}^{len(S)} Si} \quad (3\text{-}4)$$

where si is the ith input parameter.

In the following, we present proposed algorithm in Figure 2 with eps,minpts, features and matrix parameters as algorithm inputs and then we show the proposed method in chart format in Fig. 3.

```
Input eps,minpts,features,D
Start
[labels,cluster] = Modify-DBSCAN(D,eps,minpts);
R(features)=0;
intCount=0;
WeightedR=0;
while intCount<len(cluster)
{
   ClusterR=find(cluster==intCount);
}
R=max(ClusterR);
WeightedR=weight(R);
End
Return WeightedR;
```

Fig, 2. The proposed WF-DBSCAN algorithm.

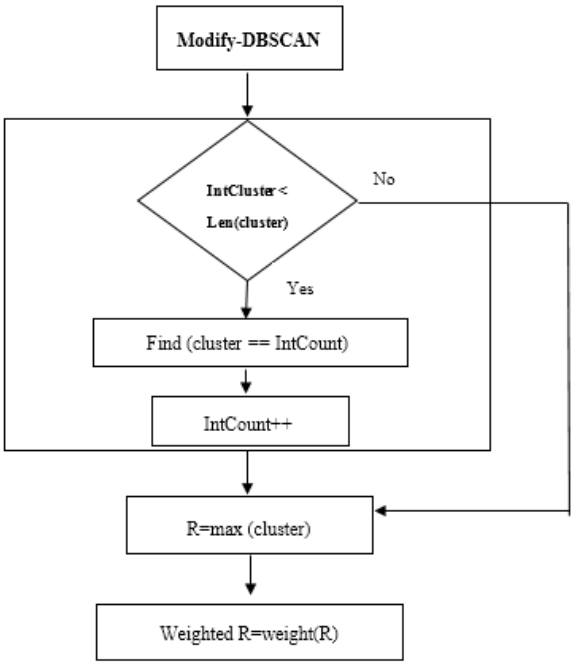

Fig. 3. WF-DBSCAN diagram.

Fig. 3 illustrates the steps of the proposed method. The clustering is performed first, then the relationship of each feature with the target class is obtained and lastly a weight is assigned to each one. For performance analysis of the proposed algorithm, we need compare accuracy of our algorithm to other algorithms. For this, first of all we will consider 70% of the labeled dataset for training and 30% for testing and then split labels and features of each one and create a tree of training dataset. With this action we can check the proposed algorithm, according to training dataset, how accurately it recognizes the test data label.

III. Evaluation of Results

To measure the performance of proposed algorithm, the following critical metrics are used [19].

$$Accuracy = \frac{\sum(TP_i + TN_i)}{\sum(TP_i + TN_i + FP_i + FN_i)}$$

The study was conducted on a Parkinson dataset with 195 samples and 23 features ftp://ftp.ics.uci.edu/pub/machine-learning-databases/parkinsons. The proposed algorithm with KNN and D-Tree classification was examined; the testing phase results of each feature were examined as follows.

TABLE I: The Function of the WF-DBSCAN Algorithm with Different K Kernels of the KNN Classification

KNN classifier	Feature selection method	Result
K= 1	**Relief**	92.30
	F-DBSCAN	87.17
	fsFisher	97.43
	WF-DBSCAN	97.43
K= 3	**Relief**	92.30
	F-DBSCAN	87.17
	fsFisher	97.43
	WF-DBSCAN	97.43
K= 5	**Relief**	89.74
	F-DBSCAN	82.50
	fsFisher	94.87
	WF-DBSCAN	94.87
K= 7	**Relief**	89.17
	F-DBSCAN	87.17
	fsFisher	89.74
	WF-DBSCAN	92.30

As shown in Table I, the proposed algorithm is examined with different k's. The mean of k from 1-7 is 95.30, which is higher than the two algorithms (RELIEF and F-DBSCAN) and in some respects is equal to the Fisher algorithm or better than it.

TABLE II: A Comparison of Different Feature Selection Method with D-Tree Classification

Algorithm	Result
Relief	84.61
F-DBSCAN	87.17
fsFisher	97.43
WF-DBSCAN	97.43

As indicated in Table II, the proposed algorithm on the D-Tree classification classifies a higher percentage of data than the other two algorithms. For more efficiency, the algorithm was tested on three other datasets. The results are presented as follows:

A. Iris Dataset

This dataset is derived from the UCI source including 150 samples and 4 features.

The best dataset found in pattern recognition literature may be Iris data set. This dataset is a collection of plant

information, in which the first feature is related to the length of the stem, the second feature is the width of the stem, the third feature is the length of the petal, and the fourth feature is related to the width of the petal (all measured in cm). Table III presents the results of the three algorithms.

TABLE III: A COMPARISON OF THE KNN CLASSIFICATION ALGORITHM WITH THE IRIS DATA

KNN classifier	Name of algorithm	Result
K= 1	**Relief**	86.66
	fsFisher	90
	F-DBSCAN	90
	WF-DBSCAN	90
K= 3	**Relief**	83.33
	F-DBSCAN	93.33
	fsFisher	93.33
	WF-DBSCAN	93.33
K= 5	**Relief**	86.66
	F-DBSCAN	93.33
	fsFisher	93.33
	WF-DBSCAN	93.33
K= 7	**Relief**	86.66
	F-DBSCAN	93.33
	fsFisher	93.33
	WF-DBSCAN	93.33

B. *Wine Data Set*

Wine is a data set which refers to alcoholic beverages in the same regions of Italy and is derived as the UCI source with 178 samples and 13 features. Table IV present the result of three algorithms and show the proposed method has a better result.

TABLE IV: A COMPARISON OF THE KNN CLASSIFICATION ALGORITHM WITH THE WINE DATA

KNN classifier	Feature selection method	Result
K= 1	**Relief**	91.42
	F-DBSCAN	82.85
	fsFisher	91.42
	WF-DBSCAN	94.28
K= 3	**Relief**	91.42
	F-DBSCAN	85.71
	fsFisher	94.28
	WF-DBSCAN	94.28
K= 5	**Relief**	94.28
	F-DBSCAN	82.85
	fsFisher	94.28
	WF-DBSCAN	97.14
K= 7	**Relief**	91.42
	F-DBSCAN	85.71
	fsFisher	94.28
	WF-DBSCAN	97.14

C. *Isolet Data Set*

isolet is a data set which refers to alphabet expression by different people and, is derived as the UCI source with 1559 samples and 617 features. Table V present the result of the three algorithms and show the proposed method has a better result.

TABLE V: A COMPARISON OF THE KNN CLASSIFICATION ALGORITHM WITH THE ISOLET DATA

KNN classifier	Feature selection method	Result
K= 1	**Relief**	84.88
	F-DBSCAN	7.39
	fsFisher	84.88
	WF-DBSCAN	84.88
K= 3	**Relief**	76.84
	F-DBSCAN	7.07
	fsFisher	76.84
	WF-DBSCAN	76.84
K= 5	**Relief**	79.09
	F-DBSCAN	8.03
	fsFisher	79.09
	WF-DBSCAN	79.09
K= 7	**Relief**	81.67
	F-DBSCAN	7.39
	fsFisher	81.67
	WF-DBSCAN	81.67

As shown in the above table, the relief algorithm and thet WF-DBSCAN algorithm have the same result in this dataset and are better than F-DBSCAN. It is clear that the running time of F-DBSCAN is higher than other existing algorithm. As mentioned earlier in the comparison table, the proposed algorithm has a better classification function than the other two algorithms. The result of the plot and feature classification is as follows.

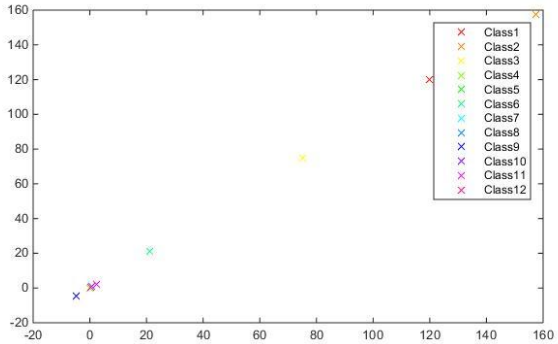

Fig. 4. Classification plot of features of Parkinson's data set.

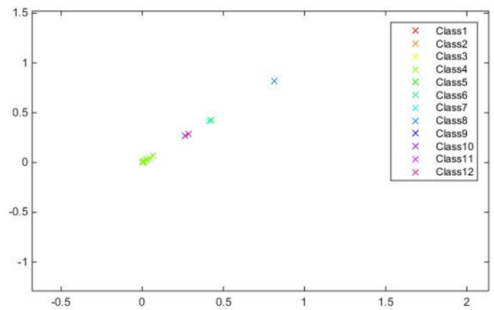

Fig. 5. Classification plot of the features based on Parkinson dataset.

Based on the results, when the WF-DBSCAN algorithm is implemented on a Parkinson's dataset, the proposed DBSCan algorithm includes 12 classes of features, among which some are invisible due to low zoom, while more classes of features are visible (Fig. 4).

As shown in Fig. 5, Class 4 consists of various features because they are neighbors. The modified-DBSCAN algorithm places them in a cluster, and the proposed algorithm is then selected as a representative of this group.

IV. Conclusion and Future Work

Based on the results, redundant and unnecessary features included many disadvantages. Thus, the WF-DBSCAN algorithm was proposed to ignore or diminish the effect of these features so that the features could be first clustered, and a weight could be assigned for each representative. The weakness of this algorithm lies in the essence of the DBSCAN algorithm. This algorithm requires two minpts and eps parameters to determine the minimum points and neighboring radius. These two parameters are adjusted as a trial and error. Further studies can be considered for the automatic adjustment and higher accuracy of these two parameters in this algorithm.

Acknowledgment

This research is funded by the Foundation for Science and Technology Development of Ton Duc Thang University (FOSTECT), website: http://fostect.tdtu.edu.vn, under Grant FOSTECT.2017.BR.19.

References

[1] H. Liu, X. Wu, and S. Zhang, "Feature selection using hierarchical feature clustering," in *Proc. the 20th ACM international conference on Information and knowledge management*, 2011, pp. 979-984.

[2] L. Peng, H. Zhang, H. Zhang, and B. Yang, "A fast feature weighting algorithm of data gravitation classification," *Information Sciences*, vol. 375, pp. 54-78, 2017.

[3] N. Eshaghi and A. Aghagolzadeh, "FFS: A F-Dbscan clustering-based feature selection for classification Data," *Journal of Advances in Computer Research*, 2016.

[4] K. Polat, "Classification of parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering," *International Journal of Systems Science*, vol. 43, pp. 597-609, 2012.

[5] D. S. Modha and W. S. Spangler, "Feature weighting in k-means clustering," *Machine learning*, vol. 52, pp. 217-237, 2003.

[6] Y. Sun, "Iterative RELIEF for feature weighting: algorithms, theories, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, 2007.

[7] M. Dialameh and M. Z. Jahromi, "A general feature-weighting function for classification problems," *Expert Systems with Applications*, vol. 72, pp. 177-188, 2017.

[8] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomputing*, 2017.

[9] A. K. Das, S. Goswami, A. Chakrabarti, and B. Chakraborty, "A new hybrid feature selection approach using feature association map for supervised and unsupervised classification," *Expert Systems with Applications*, vol. 88, pp. 81-94, 2017.

[10] N. Hoque, D. Bhattacharyya, and J. K. Kalita, "MIFS-ND: a mutual information-based feature selection method," *Expert Systems with Applications*, vol. 41, pp. 6371-6385, 2014.

[11] R. Shang, Z. Zhang, L. Jiao, C. Liu, and Y. Li, "Self-representation based dual-graph regularized feature selection clustering," *Neurocomputing*, vol. 171, pp. 1242-1253, 2016.

[12] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, pp. 16-28, 2014.

[13] G. Gan, C. Ma, and J. Wu, Data clustering: theory, algorithms, and applications: SIAM, 2007.

[14] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, pp. 651-666, 2010.

[15] S. Singh and S. Silakari, "An ensemble approach for feature selection of Cyber Attack Dataset," arXiv preprint arXiv:0912.1014, 2009.

[16] X. Zhu and A. B. Goldberg, "Introduction to Semi-Supervised Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning)," *Morgan and Claypool Publishers*, vol. 14, 2009.

[17] X. Zhu, "Semi-supervised learning literature survey," 2005.

[18] H. Liu and H. Motoda, *Computational Methods of Feature Selection*, CRC Press, 2007.

[19] A. Saha and S. Das, "Automated feature weighting in clustering with separable distances and inner product induced norms–A theoretical generalization," *Pattern Recognition Letters*, vol. 63, pp. 50-58, 2015.

Mohammad Dousthagh received the BS and M.S. degrees in computer engineering from Rouzbahan Institute of Higher Education, Sari, Iran, 2015 and 2017, respectively. His research interests include data mining and dimensionality reduction.

Mousa Nazari received the BS degree in computer engineering from Rouzbahan Institute of Higher Education, Sari, Iran and MS degree in Artificial Intelligence from Kharazmi University, Tehran, Iran, in 2012. He is currently a PhD Student in the Faculty of Electrical and Computer engineering, University of Tabriz, Tabriz, Iran. His research interests include data mining, wireless sensor networks, target tracking and fusion.

Amir Mosavi obtained his PhD in data science and currently works as visiting researcher at School of the Built Environment, Oxford Brookes University, Oxford, UK, and Institute of Automation, Kando Kalman Faculty of Electrical Engineering, Obuda University, 1034 Budapest, Hungary, and also at Queensland University of Technology (QUT), Centre for Accident Research Road Safety-Queensland (CARRS-Q), 130 Victoria Park Road, Queensland 4059, Australia.

Shahab Shamshirband obtained a PhD in Computer Science in 2014. He is an associate professor at the Department of Computer Science, Ton Duc Thang University, Vietnam. He has co-authored 100 journals and 30 peer-reviewed conference proceedings publications in the area of Machine Learning and optimization techniques. He is a professional member of IEEE, ACM member.

Anthony T. Chronopoulos obtained a Ph.D. in computer science from the University of Illinois at Urbana-Champaign in 1987. He is a professor at the Department of Computer Science, University of Texas, San Antonio, USA. He has co-authored 80 journals and 71 peer-reviewed conference proceedings publications in the areas of Distributed and Parallel Computing, Grid and Cloud Computing, Scientific Computing. He is a Fellow of the Institution of Engineering and Technology (FIET), ACM Senior member, IEEE Senior member.