Effective Streaming Evolutionary Feature Selection Using Dynamic Optimization

Abdennour Boulesnane, Souham Meshoul

HAL Id: hal-01913889
https://inria.hal.science/hal-01913889

Submitted on 6 Nov 2018

Effective Streaming Evolutionary Feature Selection Using Dynamic Optimization

Abdennour Boulesnane[1][2] and Souham Meshoul[1]

[1] Department of Computer Science, College of NTIC, Abdelhamid
Mehri-Constantine 2 University, 25000 Constantine, Algeria
[2] MISC Laboratory, Abdelhamid Mehri-Constantine 2 University, Algeria
`abdennour.boulesnane@univ-constantine2.dz`
`souham.meshoul@univ-constantine2.dz`

Abstract. Feature selection is a key issue in machine learning and data mining. A great deal of effort has been devoted to static feature selection. However, with the assumption that features occur over time, methods developed so far are difficult to use if not applicable. Therefore, there is a need to design new methods to deal with streaming feature selection (SFS). In this paper, we propose the use of dynamic optimization to handle the dynamic nature of SFS with the ultimate goal to improve the quality of the evolving subset of selected features. A hybrid model is developed to fish out relevant features set as unnecessary by an online feature selection process. Experimental results show the effectiveness of the proposed framework compared to some state of the art methods.

Keywords: Dynamic optimization problems, OSFS, Feature selection, Streaming features, Classification

1   Introduction

Feature selection is a crucial task in data mining and machine learning especially when high dimensional datasets need to be processed. The purpose behind feature selection is to select a subset of the most relevant features for building powerful predictive models [1]. All traditional feature selection methods are time consuming and require all input features to be available at the beginning of the learning process. However, with the advent of new information technologies and in the big data era, many real world applications are forced to work with attributes occurring over time or in streaming. Therefore, a new challenge has emerged namely streaming feature selection (SFS) where new features are integrated on their arrival and calculations are carried out at the same time.

In this new challenge, the number of learning samples is fixed whereas the number of features increases with time as new attributes arrive. A critical challenge for SFS is the unavailability of the entire space of features at the beginning of the learning phase. Compared to traditional feature selection problems, there are two properties of SFS [2]. One is that the number of features could grow infinitely over time. Another is that features can be read one by one and each

of them is processed online upon its arrival. However, up today and compared to traditional feature selection methods, few approaches have been proposed in the literature to tackle this new challenge [2] [3] [4] [5]. This can be explained by the fact that the process of SFS requires new, fast and inexpensive methods.

In this context and as a first initiative of this kind, we present in this paper a new approach that deals with the problem of streaming feature selection by introducing dynamic optimization during the selection of the best attributes. Motivated by the fact that the problem of online feature selection is a dynamic problem whose dimension (feature) changes over time, we propose a hybridization between the WD2O dynamic optimization algorithm proposed in [6] and the Online Streaming Feature Selection algorithm (OSFS) [2], whose objective is to find a subset of optimal attributes to ensure better classification of unclassified data. Therefore, the main contribution of this work consists of a new hybrid approach called Dynamic Online Streaming Feature Selection (DOSFS) that exhibits the following features:

- This new feature selection system exhibits two properties: on one hand, the speed of OSFS, helps in providing quality attributes at any time; on the other hand, WD2O's self-adaptivity helps in exploring efficiently the space of redundant features taking into account the importance of the interaction between these features.
- This hybridization helps in fishing out relevant information previously treated as unnecessary data, which in turn helps to improve decision making in the future.
- DOSFS helps in strengthening the exploration capability of the OSFS algorithm and fill its gaps.

The rest of the paper is organized as follows. Section 2 presents some background material and related works. The proposed hybridization DOSFS is described in Section 3. Section 4 describes the conducted experimental study and obtained results. Finally, conclusions and perspectives are given in Section 5.

2   Background and Related Works

2.1   Feature Selection for Classification

Classification is the most common task of data mining and machine learning. It consists in identifying to which of a given set of classes a new incoming instance belongs. The purpose of classification is to obtain a model that can be used to classify unclassified data [7].

Many real world classification problems require supervised learning where the underlying class probabilities and class-conditional probabilities are unknown and each instance is related with a class label, i.e., relevant features are often unknown a priori [8]. Therefore, many candidate features are used in order to improve the representation of the domain, resulting in the existence of irrelevant and redundant attributes for the studied domain. This makes the decision

algorithms complex, inefficient, less generalizable and difficult to interpret. For the majority of classification problems, it is hard to learn good classifiers before deleting these unwanted features because of the huge dimensionality of the data. Decreasing the number of irrelevant / redundant features can dramatically reduce the operating time of the learning algorithms and hence provide a more efficient classifier [7].

Therefore, feature selection techniques is considered one of the best methods to reduce the dimension in the feature space. It consists in finding the best subset among 2^n candidate features according to some evaluation functions. Generally, the feature selection methods used to evaluate a subset of features in the learning algorithms can be classified into three main categories, according on how they introduce the feature selection search with the construction of the classification model: filter methods, wrapper methods and embedded methods [9].

2.2 Streaming Feature Selection

In classical feature selection methods (filter, wrapper and embedded methods), all candidate attributes are assumed to be known a priori. These features are iteratively examined in order to select the best attribute. However, nowadays this does not extend to many real-world applications where one needs to deal with dynamic data streams and feature streams. For example, Twitter generates more than 320 millions of tweets daily and a large amount of words (features) are continually being produced. These new words quickly attract user's attention and become popular in a short period of time [9]. Therefore, because of the ineffectuality of traditional feature selection methods to applications involving streaming features, it should be preferable to use SFS to quickly adapt to the changes. Streaming features involve an attribute vector whose elements flow one by one over time while the number of instances in the training set remains fixed. The particularity of the SFS compared to the traditional feature selection, is as follows [2]:

- **The dynamic and uncertain nature**: Feature space's dimension may grow dynamically over time and may even extend to an infinite size.
- **The streaming nature**: Features flow one by one where each feature must be processed online upon its arrival.

2.2.1 Streaming Feature Selection Approaches For the problem of SFS, the number of instances is considered as constant whereas the features arrive one at a time (in streaming). The task is to select in a timely manner a most relevant subset features from a huge number of available features. Compared to traditional methods and instead of searching in the entire attribute space that is very expensive, the SFS techniques process a new feature on its arrival [10].

In the literature, few approaches have been proposed to tackle the problem of SFS [2] [3] [4] [5]. These approaches have different implementations especially in the way newly arrived features are checked. In the following section, we briefly present one of the most successful techniques widely used to resolve the problem of SFS.

2.2.2 Online Streaming Feature Selection Algorithm (OSFS) Unlike
the existing studies on traditional feature selection, online streaming feature se-
lection aims to deal with feature streams in an online manner. For that purpose,
a new algorithm is proposed in [2] called Online Streaming Feature Selection
(OSFS). The authors in [2] study the SFS problem from an information theo-
retic perspective based on the criterion of Markov blanket. In OSFS, features
are characterized into four types: redundant feature, irrelevant features, weakly
relevant but non redundant features and strongly relevant features. An optimal
feature selection approach should have strongly relevant and non-redundant fea-
tures. OSFS finds an optimal subset of features based on online feature relevance
and feature redundancy analysis [10]. A general framework of OSFS is presented
as follows:

- Initialize the list of Best Candidate Features (BCF) in the model, $BCF = \emptyset$;
- **Step 1**: Generate a new feature x;
- **Step 2**: Online Relevance Analysis;
  - If x is relevant for the class label: $BCF = BCF \cup x$;
  - Otherwise, reject attribute x;
- **Step 3**: Online Redundancy Analysis;
- **Step 4**: Alternate Step 1 to Step 3 until some stop criteria are met.

In step 2 and in the relevance analysis phase, OSFS consists of discovering
strongly and weakly relevant features, in order to add them into the best can-
didate features (BCF). If a new coming feature is relevant to the class label, it
is added to BCF, otherwise it is discarded. In the redundancy analysis (step 3),
OSFS dynamically removes the redundant features in the BCF subset. For each
feature x in BCF, if there exists a subset within BCF making x and the class
label conditionally independent, x is eliminated from BCF [2].

2.3 Optimization in Dynamic Environments

In every day life, each type of optimization problem has specific characteristics
that make it distinct from others. However, almost all of them have a com-
mon feature which is their dynamic nature. Static optimization has known its
limitations in solving such problems and therefore sophisticated methods are
needed. More specifically, the field of research that addresses this kind of prob-
lems is commonly known as: *Optimization in Dynamic Environments* or simply:
Dynamic Optimization [11].

Solving a dynamic optimization problem, requires not only finding the opti-
mal global solution in a specific environment, but also following the trajectory of
the evolution of this optimum in dynamic landscapes. Therefore, the main chal-
lenge for optimization algorithms in dynamic environments is how to increase or
maintain the search diversity in such environments.

2.3.1   Dynamic Optimization Problems We can formally describe a dynamic optimization problem (DOP) as the task that aims to find the sequence $(x_1^*, x_2^*, \ldots, x_n^*)$ that:

$$
\begin{aligned}
&\text{Optimize } f(x,t) \\
&\text{subject to. } h_j(x,t) = 0 \text{ for } j = 1,2,...,u \\
&g_k(x,t) \leq 0 \text{ for } k = 1,2,...,v. \\
&\text{with } x \in \mathbb{R}^n
\end{aligned}
\tag{1}
$$

Where $f(x,t)$ is a time dependent objective function, $(x_1^*, x_2^*, ..., x_n^*)$ is the sequence of n optima found as the fitness landscape changes. In other ways, it depicts optima tracking, $h_j(x,t)$ denotes the jth equality constraint and $g_k(x,t)$ denotes the kth inequality constraint. All these functions may change over time, as indicated by the dependence on the time variable t. A comprehensive review on dynamic optimization can be found in [12].

2.3.2   Wind Driven Dynamic Optimization Algorithm (WD2O) In the literature, several algorithms have been proposed to deal with DOPs [12]. Each one has features that make it appropriate for solving specific problems than others. In other words and according to the No Free Lunch theorem, there is no universal algorithm which solves in the best way all optimization problems. Therefore, each optimization problem requires a thorough study that allows to find the best algorithm.
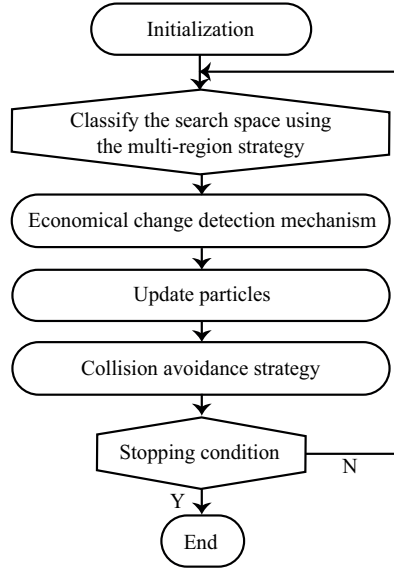

Fig. 1. Flow chart of WD2O algorithm.

Recently, a new dynamic optimization algorithm has been proposed in [6], called "Wind Driven Dynamic Optimization Algorithm (WD2O)". The characteristic property of this metaheuristic is the classification, it suggests to set regions of the search space as promising and non-promising regions with accordance to low and high pressure regions in the natural model. This new framework has been inspired from the meteorology. Compared to other dynamic optimization algorithms, the powerful feature of WD2O essentially resides in its efficiency and scalability against high-dimensional dynamic problems due to the new multi-region classification of search space. It is a multi-population metaheuristic.

Formally and as shown in Figure 1, the WD2O algorithm can be described as follow: in the initialization phase, the particle's positions and velocities are randomly set for each dimension within the corresponding bounds. WD2O proceeds iteratively as described in Figure 1. Firstly, the whole search space is divided into promising and non-promising areas using a multi-region strategy. This classification was beneficial in helping to find and track the global optimum as quickly as possible. Next to the objective function value, the pressure value is calculated for each particle in the population. These values have been exploited to achieve such a classification. A change detection mechanism is a significant step as it allows the algorithm to rapidly react to the possible environmental changes. In the next step, particle's positions are updated. Since this metaheuristic uses multiples populations, collision avoidance strategy is used in order to maintain several sub-populations on several peaks. This process continues in this manner till a stopping criterion is satisfied.

3   Research Problematic and Proposed Approach

3.1   Research Problematic

Generally, the evaluation criterion and the search strategy are two key elements in feature selection. According to the evaluation criterion, most feature selection algorithms (offline) are based either on filter methods or on wrapper methods. Furthermore, since the size of the search space for n features is 2^n, it is impossible to carry out an exhaustive search for the feature selection [13]. Therefore, the search strategy can drastically influence the results of a feature selection algorithm. Evolutionary techniques including particle swarm optimization, evolutionary algorithms, etc. have been widely applied to the traditional feature selection problems [14].

However, in online feature selection, these two problems need to be revisited. For the evaluation criterion, there is no general classification of the related methods, where most of them are based on information theory. On the other hand, since the feature space grows continuously over time, the size of the search space will be unknown or infinite. Therefore, it is impossible to adopt a global search technique because of the unavailability of the entire feature space.

Therefore, the questions that arise are, how to introduce computational intelligence to carry out streaming feature selection? And, to what extent will it

be possible to solve the aforementioned issues? In a first initiative of its kind and motivated by the fact that the problem of the streaming feature selection is a dynamic problem whose dimension (feature) changes over time, we propose to treat this problem by introducing dynamic optimization during the selection of the best attributes.

3.2 Proposed Approach

As explained previously, the OSFS algorithm has been applied successfully to the problem of online feature selection. The idea of this simple technique is mainly focused on the exclusive selection of highly relevant features and weakly relevant non-redundant features. Whereas, irrelevant and redundant features are eliminated during the selection process. Although the results of this approach have been very encouraging in terms of accuracy, the interaction between the eliminated attributes and the future attributes is completely ignored in this model. Therefore, the logical question we have thus to ask is to what extent it would have been interesting to take these interactions into account?

To address the issues raised above, we propose a new model for selecting features. More precisely, we propose a new hybridization between the OSFS algorithm and the dynamic optimization algorithm WD2O.

The proposed approach can be summarized as shown in Figure 2. We adopt the acronym DOSFS (Dynamic Online Streaming Feature Selection) to refer to the proposed approach. In this model, we propose a new system structure of SFS. The gist of this approach can be described as follows. In the first phase, which can be considered as an initialization phase, all parameters and components of either WD2O or OSFS are initialized. It should be mentioned that the selection process is automatically triggered by the arriving of new features, which carries the information relating to each Xi instance in the training data.

As soon as a new feature arrived (feature F44 in Figure 2), the second phase is launched, in which the OSFS algorithm will run. As explained above, this algorithm retains only the relevant features for the classification task in question (class C in Figure 2), which will be added to the set of best candidate feature (BCF) found so far. But before being selected, it must be proved that there is no subset S within BCF for which the features in BCF will be redundant.

Otherwise, irrelevant features are simply eliminated. However, eliminating redundant features is the key task for an optimal feature selection process, because according to the definition in [15], a redundant feature is a weakly relevant feature. Therefore, a redundant feature deleted at time t may become highly relevant in the future, through interaction with other newly arrived attributes.

In order to avoid such a scenario, or at least alleviate the problem, we propose to create a new feature space named Best Redundant Candidate Feature (BRCF). This set includes only the redundant attributes that are dependent on each other. Once the BRCF set is created, the third phase is launched, in which the dynamic optimization algorithm WD2O should find the best sequence of features independently of that found by OSFS.
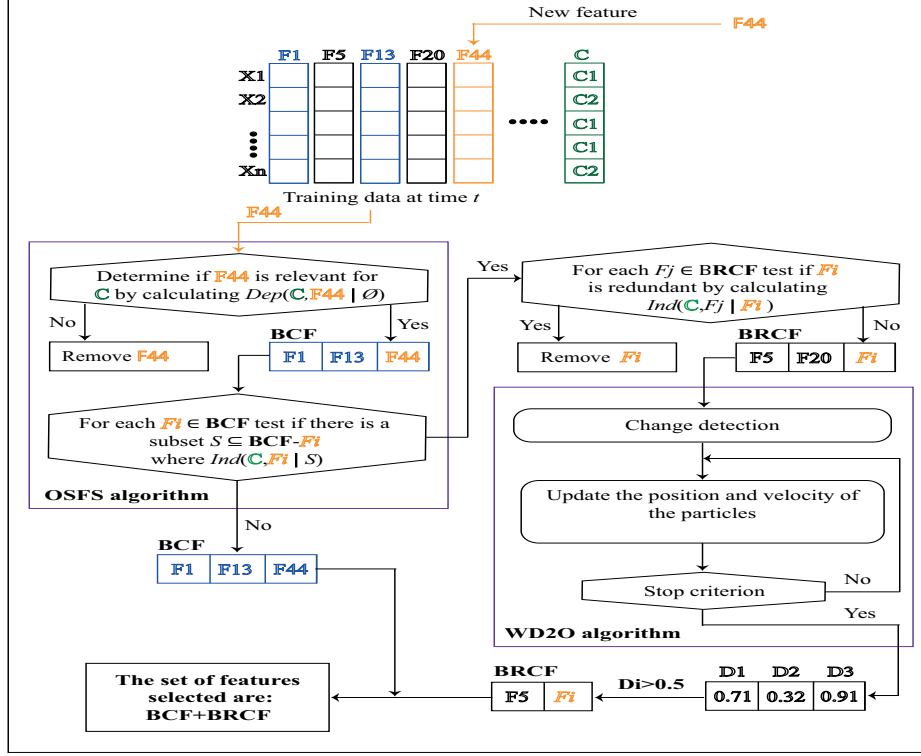
Fig. 2. Functional diagram of the proposed model DOSFS.

3.2.1   Objective Function and Solution Encoding First of all, let us recall that the SFS problem is a dynamic problem whose dimension (D) is the only element that changes over time. In this type of problem, the objective function is always static, it consists, on the one hand, in maximizing the relevance between the features (F1, F2,..., Fn) and the class labels (C1, C2), on the other hand, to minimize the redundancy between the selected features. The objective function used in this problem is inspired by the one proposed in [16] which can be defined as follows:

$$Max\left(Fitness = \sum_{x \in X} Z\left(x, C\right) - \sum_{x_i, x_j \in X} Z\left(x_i, x_j\right)\right) \qquad (2)$$

Where X is the set of selected attributes (BRCF) and C is the class label. In our case, since the data used are continuous, we will use the same objective function defined by equation (2), but with Fisher's Z test instead of mutual information as used in [16].

Otherwise, in WD2O, each particle's position is a vector of D dimension representing the number of features available in the subset BRCF. Each dimension

corresponds to a decision variable, we adopt a real coding (between 0 and 1) because we are in a context of continuous optimization. A value greater than or equal to (respectively less than) 0.5 indicates that the attribute is selected (respectively not selected).

3.2.2   Change Detection In this step, WD2O detects new environmental changes in the problem's dimension (D). Therefore, a simple technique has been adopted in WD2O in order to check the size of the dimension each time a new feature has arrived. If a change is detected, WD2O exploits the best solution found so far (stocked in the memory) (*gbest*) to quickly follow the new optimum. The idea is to change the *gbest*'s dimension to be compatible with the new problem, but without selecting the newly arrived feature (i.e., generating a value between 0 and 0.5 in this dimension). Furthermore, all particles in the population will be reinitialized to increase the diversity level in the search space.

3.2.3   Global Search Once a change is detected (i.e., a new feature has been added to the BRCF set), the WD2O algorithm proceeds iteratively until a stopping criterion is met. At the end of the optimization process, the best features selected that maximize the objective function will be incorporated into the BCF set and considered as important features retained by the dynamic optimization algorithm.

4   Experimental Study

4.1   Experimental Setup

In order to assess the performance of the proposed approach and to verify its usefulness in practice, seven large-scale biological data sets were used (Feature Selection Datasets[3]; Kent ridge biomedical data set repository[4]) as shown in the Table 1. These biological datasets were provided for the purpose of selecting relevant features to a classification problem.

For the datasets: Breast-Cancer, Lung-Cancer, Leukemia-ALLAML, we use the training and validation datasets provided in Kent ridge biomedical data set repository. For the other three datasets, we adopt 2/3 of the instances for the training and 1/3 of the remaining instances for the test. Our comparative study compares the proposed model DOSFS algorithm with α-investing [4] and standard OSFS algorithm [2]. In order to evaluate a selected feature subset in the experiments, the SVM classifier is used. Two performances measures to evaluate our algorithm with the standard OSFS and α-investing are compactness (the number of selected features) and the prediction accuracy (the percentage of correctly classified test instances). All experiments were conducted on a computer with Intel i5-2450, 2.50 GHz CPU and 6 GB memory.
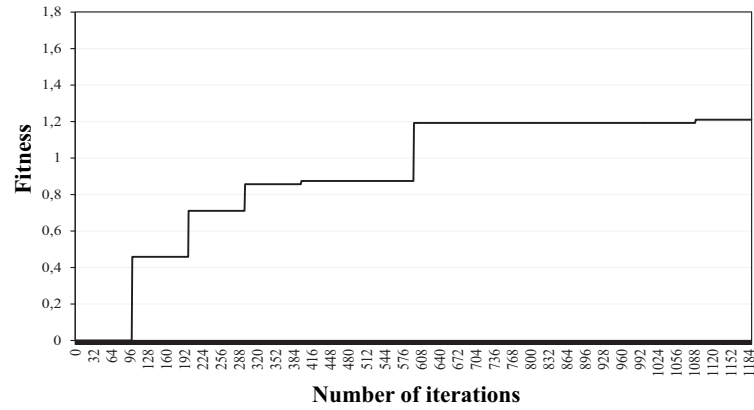
---

[3] http://featureselection.asu.edu/datasets.php
[4] http://datam.i2r.a-star.edu.sg/datasets/krbd/

Table 1. Summary of biological datasets used for evaluation.

Dataset	# Features	# Instances	Training data	Test data
Breast-Cancer	24481	97	78	19
Colon-Tumor	2000	62	41	21
Central Nervous System	7129	60	40	20
Lung-Cancer	12533	181	32	149
Leukemia-ALLAML	7129	72	38	34
Prostate-GE	5966	102	68	34
Arcene	10000	200	134	66

4.2   Experimental Results and Comparisons

4.2.1   Usefulness of WD2O for Streaming Feature Selection To analyze closely how the dynamic optimization algorithm WD2O react to the arrival of a new streaming feature, we recorded the evolution of WD2O over time during the streaming processing of the "Leukemia-ALLAML" dataset, as shown in Figure 3. Similar behaviors have been noticed for the other datasets.

Fig. 3. Evolution of the best fitness value (corresponding to gbest) during a run of the proposed model using the Leukemia-ALLAML dataset.

From Figure 3, one can observe the gradual improvement of the best fitness value corresponding to *gbest*, which implies both the high exploration capacity of WD2O in the search space and its efficiency to handle dynamic optimization problems. Furthermore, the performance stability recorded over certain periods

of time can be explained in two ways: either the new arrival feature is redundant in the set BRCF, which implies the deletion of this attribute, or this new streaming feature is added in BRCF but its interaction with the other features does not significantly improve the classification task.

4.2.2   Comparison and Analysis In this section we compare our proposed hybrid approach with the standard OSFS and α-investing algorithms in terms of prediction accuracy and the number of selected features, on the seven high-dimensional datasets previously presented in Table 1.

Using the SVM classifier, Table 2 presents the prediction accuracy values of the proposed DOSFS vs. OSFS and α-investing algorithms. The highest predictive accuracy values are shown in bold.

Table 2. Prediction accuracy of SVM (Acc) and the number of selected features (#).

Dataset	Algorithm					
	α-investing		OSFS		DOSFS	
	Acc	#	Acc	#	Acc	#
Breast-Cancer	0.63158	1	0.78947	2	**0.84211**	4
Colon-Tumor	0.57143	2	**0.71429**	2	0.57143	5
Central Nervous System	**0.8**	1	0	2	0.05	4
Lung-Cancer	0.92617	2	0.95302	2	**0.96644**	6
Leukemia-ALLAML	0.52941	1	0.91176	2	**0.94118**	6
Prostate-GE	0.88235	2	0.67647	2	**0.91176**	6
Arcene	0.56061	6	0.56061	3	**0.59091**	7

From the obtained results in Table 2, it can be seen that the proposed hybrid approach DOSFS has succeeded in improving the performance of OSFS on several cases, due to the features recovered by the dynamic optimization algorithm as shown in Table 2. According to the SVM classifier, the results indicate that the proposed DOSFS is very competitive and promising approach compared to OSFS and α-investing on most datasets.

5   Conclusion and Future Work

The major contribution of this paper is that we proposed a new model to solve streaming feature selection, the goal of which is to develop an online classifier involving only a small number of features. More precisely, we propose for the first time in the literature, a new hybridization between the streaming feature selection algorithm OSFS and the WD2O dynamic optimization algorithm.

Thanks to the proposed hybrid model, the exploration capability of the OSFS algorithm has been enhanced and many of the relevant features previously treated by the OSFS algorithm as unnecessary data has been fished out,

which in turn helps to improve decision making. To analyze the performance of the proposed approach, high-dimensional biological data were used. The results obtained are encouraging and confirm that this new model is a promising way for a more stable and precise selection of streaming features.

As for future work, we intensify our work on studying the relationship between the dynamic problem of data streaming and dynamic optimization. This new convergence could bring many new lines of research in the field of data mining.

References

1. Wang, J., Zhao, P., Hoi, S., Jin, R.: Online feature selection and its applications. IEEE Transactions on Knowledge and Data Engineering. 26(3), 698–710 (2014)
2. Wu, X., Yu, K., Ding, W., Wang, H., Zhu, X.: Online feature selection with streaming features. IEEE Trans. Pattern Anal. Mach. Intell. 35(5), 1178–1192 (2013)
3. Perkins, S., Theiler, J.: Online feature selection using grafting. In: Proceedings of the 20th International Conference on Machine Learning (ICML), pp. 592-599 (2003)
4. Zhou, J., Foster, D., Stine, R., Ungar, L.: Streamwise feature selection. The Journal of Machine Learning Research 7, 1861-1885 (2006)
5. Yu, K., Wu, X., Ding, W., Pei, J.: Towards scalable and accurate online feature selection for big data. In: 2014 IEEE International Conference on Data Mining, pp. 660-669. IEEE, (2014)
6. Boulesnane, A., Meshoul, S.: WD2O: a novel wind driven dynamic optimization approach with effective change detection. Appl Intell. 47(2), 488-504 (2017)
7. Jiliang, T., Salem, A., Huan, L.: Feature Selection for Classification: A Review. In: Data Classification: Algorithms and Applications, CRC Press, 37-64 (2014)
8. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis. 1(1-4), 131-156 (1997)
9. Li, J., Hu, X., Tang, J., Liu, H.: Unsupervised streaming feature selection in social media. In: Proceedings of the 24th ACM international on conference on information and knowledge management,ACM, pp. 1041-1050 (2015)
10. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R., Tang, J., Liu, H.: Feature selection: a data perspective. arXiv preprint arXiv:1601.07996 (2016)
11. Cruz, C., González, J., Pelta, D.: Optimization in dynamic environments: a survey on problems, methods and measures. Soft Comput. 15(7), 1427-1448 (2011)
12. Yang, S., Yao, X. (eds.): Evolutionary Computation for Dynamic Optimization Problems. SCI, vol. 490. Springer, Heidelberg (2013)
13. Kohavi, R., John, GH.: Wrappers for feature subset selection. Artif Intell. 97(1), 273-324 (1997)
14. Xue, B., Zhang, M., Browne, W., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. 20(4), 606–626 (2016)
15. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res. 5, 1205-1224 (2004)
16. Cervante, L., Xue, B., Zhang, M., Shang, L.: Binary particle swarm optimisation for feature selection: A filter based approach. In: IEEE Congress on Evolutionary Computation (CEC 2012), pp. 881-888 (2012)