Review on Wrapper Feature Selection Approaches

Naoual El aboudi
Université Mohammed V
Ecole Mohammadia d'ingénieurs
Rabat, Maroc
Email: nawal.elaboudi@gmail.com

Laila Benhlima
Université Mohammed V
Ecole Mohammadia d'ingénieurs
Rabat, Maroc
Email: benhlima@emi.ac.ma

***Abstract*—The main objective of feature selection process consists of investigating the optimal feature subset leading to better classification quality while spending less computational cost compared to maintaining the whole initial feature set. The problem of feature selection has been extensively researched since the early beginnings of machine learning . Even though several methods were proposed to handle the issue of feature selection using a variety of techniques, it is difficult to identify a specific method as the most fitted one regarding the feature subset selection issue. In this study, we provide an overview of existing wrapper methods pointing out their weaknesses and their strengths.**

I. Introduction

Feature selection constitutes an important preprocessing phase in machine learning. The reason behind performing this task lies in the fact that reducing the dimensionality of the original feature space allows in general gains in both classification accuracy and computational cost. Therefore, such a step is highly recommended in the presence of high dimension data, which seems to be the expected trend in the future. To reach a maximum classification accuracy while spending a minimum computational effort, it is crucial to select the most appropriate feature subset by categorizing features into relevant and irrelevant ones in the initial feature set while being capable of detecting redundant attributes. [1]

In order to label correctly a given instance during classification, it is required to employ relevant features which are highly informative regarding classification process. On the other side, irrelevant features have no impact on classification results, while redundant features are equivalent to some other relevant attributes. The objective of feature selection is to identify relevant features and eliminate irrelevant and redundant attributes seeking the ideal situation where the selected features are the most adequate ones to achieve good results in terms of the classifcation process.

In general, a given FS method belongs to one of two basic families: filters and wrappers [2]. Another family deriving from the two previous ones is hybrid methods. Filter methods do not employ a learning algorithm to evaluate a feature subset since they adopt statistical measures to rank available features, then those achieving scores below a predetermined threshold are automatically rejected. Users often choose filters since they are simple to design and do not require important computational resources, especially when handling large datasets. F-score [3] , mutual information [4] and information gain illustrate examples of popular filters. On the contrary, wrapper methods, involve a particular learning algorithm that will be adopted to evaluate the accuracy performance of a candidate feature subset which lead to better solutions. Nevertheless, they are difficult to deploy in the presence of high dimensional data as heavy computational burden is needed even when using simple learning algorithms. Integrating the complementary aspects of filters and wrappers has led to a third approach that consists of hybrid approach. Hybrid methods attempt to combine the strengths of both filters and wrappers aiming to design more efficient solutions.

In this study, we present a summary of existing feature selection wrapper techniques with a special emphasis on recent evolutions that achieved promising results. Figure 1 illustrate the general principle of a wrapper feature selection model.

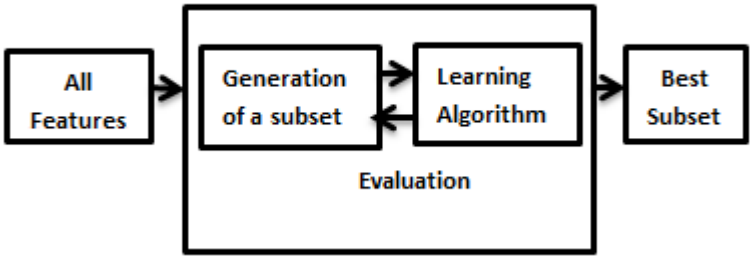

Fig. 1. Wrapper Feature Selection Method

II. Related Works

Since it represents an essential component of data preprocessing, the feature selection problem has been the subject of extensive research, especially in the machine learning framework. In this context, a review on feature selection approaches based on evolutionary techniques was presented by Abd-Alsabour in [5]. Kumar proposed a comprehensive overview of existing feature selection methods in [6] to classify and compare them.

Recently, Xue conducted in [7] an interesting survey on feature selection methods where she focused on evolutionnary techniques highlighting their key aspects which include the model representation, the benefit mechanisms and the fitness

978-1-5090-5579-1/16$31.00

function design. Furthermore, Jovic presented in [8] a review which takes into consideration most of the commonly used FS techniques. In that study, the emphasis is on the application aspects of reviewed solutions. In this paper, we propose a review on feature selection techniques and we focus precisely on wrapper feature selection approaches as they allow to achieve better results.

III. Feature selection based on search strategy

Feature selection, in the case of a search strategy, is performed by selecting an algorithm that looks for the optimal feature subset according to a specific objective function. In other words, the problem of feature selection is translated into an optimization problem where the user attempts in general to maximize the classification accuracy while minimizing the size of the corresponding feature subset. There are many search strategies in the literature, each of them belongs to one of the three following families:

A. Exponential Complexity Methods

As their name suggests, these methods have an exponential complexity, thus the number of evaluated subsets increases dramatically even in the presence of medium size feature space, this phenomenon is known as the curse of dimensionality. For instance, exhaustive search presented in [9] is an intuitive exponential complexity method, Although this solution ensures finding the optimal feature subset, it is impractical to implement due to deterrent computational cost.
It considers all possible subsets which guaranteed to find the optimal solution. However, it is difficult to run, even with moderate size feature sets.

B. Population Based Approaches for Feature Selection

Population based methods for feature selection are very popular, their success may be explained by the good tradeoff they achieve between computational effort and the quality of the provided solution [7]. These methods adopt a particular population based optimization metaheuristic which common principle is to mimic evolution toward better solutions in nature. Each population based algorithm has a set of candidate solutions which are updated iteratively based on a specific mechanism seeking to obtain better solutions according to a given fitness function. After reaching a stop criterion, the algorithm ends running and provides the best solution among its population. In general, every population based metaheuristic has several parameters impacting highly the way the search is performed. In the absence of default values for those parameters, it is up to the user to tune them according to the specific problem been handled. Furtheremore, this kind of algorithms has to be fed with an initial population with respect to a certain initialisation strategy. For instance, a random initialisation strategy may lead to results different from a more biased initialisation policy see Xue. Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) represent population based techniques that are widely adopted in the context of feature selection, thus, those kinds of algorithms will be examined closely in the next sections.

1) Genetic Algorithm for Feature Selection: Genetic Algorithms (GAs) represent a very popular class of optimization metaheuristics that achieved satisfactory results in many applications including feature selection problem since their introduction by Holland in 1970, Figure 2 illustrates the principle of GA in feature selection. For instance, many authors proposed several variants of GAs to handle the feature selection issue in [10], [11], [12], [13], [14].
In genetic algorithm, the population is composed of candidates called chromosomes which are generally coded in the form of a binary sequence where a digit represents a gene. During each iteration, a serie of operations is applied to candidates in order to enhance the quality of individuals forming the next population (Generation). The following operators are applied consecutively during every iteration:
Selection: Based on fitness function scores, this operator will select randomly candidates for the next step organizing them in the form of couples.
Crossover: This operator imitates reproduction, therefore each couple represents parents that will give birth to two childs, each of them shares a part of its sequence with one of its parents. This sequence is obtained thanks to a divide in parents sequences from a given position named crossover point.
Mutation: The newly formed individuals will be subjected to another transformation that consists of modifying their genes according to a predefined probability.
When addressing the problem of feature selection using GAs, the model is straightforward since a binary string is adapted to represent a candidate feature subset. Indeed, each feature subset may be coded by a binary chromosome which length represents the size of the original features set. In this case 0 means that a given feature is ignored, while 1 indicates that the corresponding feature has been selected. As the search proceeds, the quality of individuals improves over generations since the algorithm keeps candidates achieving high scores with respect to the fitness function. The main challenge when adopting GAs is to maintain diversity among population to avoid narrowing the search by a too strong elitist strategy, which will make the search trapped in some local minima. There are few studies dedicated to the topic of population diversity in the context of feature selection based on GAs, for instance, Alsukk proposed in [13] a solution called diverse genetic algorithm that tackles the issue of diversity by modifying the selection and crossover operators. The particular issue of diversity deserves more research effort due to the promising potential of GAs in handling the feature selection problem.Moreover, one of the classical problems when dealing with GAs is their slow convergence. In this context, Jung proposed in [14] a variant of GA for feature selection that minimizes fitness evaluations thanks to a dynamic cost function achieving encouraging outcomes. Another powerful evolutionary algorithm is the particle swarm optimization metaheuristic which will be discussed in the following section.

2) Particle swarm Optimization for feature selection: Particle swarm optimization is a very popular optimization tech-

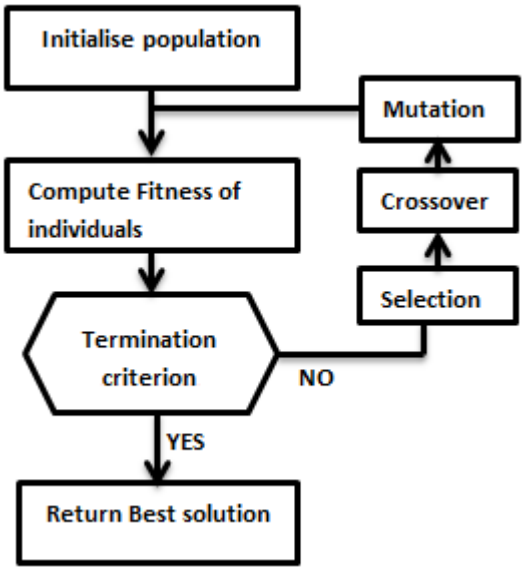

Fig. 2. GA for Feature Selection

nique, its simple benefit mecahnism makes it easy to deploy in several fields. The feature selection problem took advantage from PSO, where it was studied extensively achieving often encouraging results. Depending on the particle representation, PSO has two versions either continuous or binary that is refered as binary PSO (BPSO).

Particles in PSO represent candidate solutions for the problem. Each of them is associated with two main characteristics: velocity and position. During each iteration, the position of a given particle is updated depending on the value assigned to its velocity combined to its best previous own position and the position of the best element among the global population of particles.

The representation of each particle in PSO for feature selection is the same of GA one, it consists of a bit-string which dimension is equal to the cardinality of the feature set. The bit-string is composed either of binary numbers in BPSO variant or real-value numbers in the case of continuous PSO. In the case of BPSO studied in [15], 1 means the corresponding feature is selected and 0 means it is eliminated. When adopting continuous PSO, it is important to define some threshold value in order to decide whether a feature is selected or discarded, such parameter is difficult to set since its value influences to a great degree the performance of the resulting solution. Thanks to its satisfactory results, PSO algorithm has been the subject of steady research which has led to efficient variants more adapted to handle the feature selection problem than the classical version. For instance, Xue proposed in [16] a PSO feature selection model with new benefit mechanisms that take into consideration both the classification accuracy and the number of features. Those proposed benefit mechanisms were used in conjunction with new initialisation strategies inspired from sequential selection algorithms achieving some of the best results in the literature. In a recent contribution, see [17], moradi obtained good results by designing a BPSO based feature selection solution where local search is introduced to guide the search more efficiently.

TABLE I
FEATURE SELECTION METHODS BASED ON SEARCH STRATEGY

Search Strategy	Algorithms
Exponential	Branch and Bound Algorithm [9]
Population Based	Genetic Algorithm [12] Particle Swarm Algorithm [20] Bee Colony Algorithm [21] Advanced Binary Ant Colony Optimisation [22]
Sequential	Sequential Forward Selection [18] Sequential Backward Selection [19]

C. Sequential selection strategies for feature selection

These algorithms belong to the family of greedy algorithms, they were named sequential due to their iterative functionning.

1) The Sequential Feature Selection (SFS) algorithm: This algorithm starts its query by an empty set, then it looks for the feature that allows to reach the best classification accuracy, when identified, this feature is simply added to the empty set which will form the pursued feature subset. This procedure is repeated as necessary until no possible improvement of the classification accuracy is possible by adding any of the remaining features [18]. Although this algorithm returns a solution in a reasonable amount of time, the quality of the provided solution is expected to be poor since the search is very limited as a selected feature can not be removed in further iterations.

2) The Sequential Backward Selection (SBS) algorithm: This algorithm operates similarly to the SFS, exept that the search is performed in the opposite direction. In fact, the algorithm starts with a set containing all the available features, then it removes the feature which elimination increases classification accuracy the most. This task is repeated over again until no improvement of classification performance is possible by removing any of the remaining features [19]. Similiraly to its counterpart represented by SFS, SBS has no guarantee of finding the optimal feature subset, nevertheless its quick convergence toward a solution is assured. The table I recaps the previous section by summarizing some of the well known contributions in the field of feature selection methods based on search strategy.

IV. FEATURE SELECTION APPROACHES FOR HIGH DIMENSIONAL DATA

The amount of available data in almost all applications has become huge over the recent period of time. In fact, data tends to include not only large number of instances but contains also a high dimensional feature space. Such a trend is expected to continue in the years ahead, it is known as the phenomenon of Big Data. When facing data with important volume, conventional approaches fail to deliver acceptable results within a reasonable time period due to poor scalability. In order to overcome such limitations, a bunch of solutions were proposed by researchers. Indeed, parallel computing

solution confirmed itself as a leading option in the domain of Big Data. It consists in splitting the procedure performed by an algorithm into multiple tasks that can be processed simultaneously on different machines. For instance, generating new candidates, evaluating the fitness function are tasks that could be parallelized [23] .

A. Parallel GA approach for feature selection

Due to their inherent properties, GAs are fitted to parallelization. The implementation of parallel GA for feature selection through mapreduce can be realized according to one of the three following approaches:

- Global Parallelisation Model
- Fine-grain Parallelisation Model
- Coarse-grained Parallelisation Model

On the one hand, in the fitness evaluation level (Global Parallelisation Model) only the task of fitness evaluation for each individual is parallelized, other operations remain centralized.
On the other hand, in the individual Level or grid Model (Fine-grain Parallelisation Model), each solution (individual) is given a position on a grid and the GA operations take place in parallel by calculating simultaneously the fitness value while the selection operator is applied only to the small adjacent neighbourhood.
Finally, in the population level or island model (Coarse-grained Parallelisation Model), the population is separated into islands and the GA runs independently on each of them. Seeking a suitable feature subset, Ferruci applies in[24] an island model through the next steps:

- Initialisation phase
- Generation phase

The initialisation phase consists of generating a random population of candidate feature subsets represented in a classical manner by a string of binary variables indicating whether a feature is selected or eliminated, then the fitness function of each candidate solution is computed by determining the classification accuracy during the generation phase, in the following step, GAs operators are applied until a target accuracy is obtained or a maximum number of iterations is reached.
At the end of the algorithm, the final population will be tested with the test dataset.

B. Feature Selection for Streamming Data

In the majority of available work on feature selection, authors assume that the whole feature space is already at the user's disposal, which may contrasts with practical applications where data is made available in a streamming way [25]. In such cases, it is challenging to determine a suitable feature subset as the interaction between current and future features is difficult to assess. The general framework of streaming feature selection method is shown in Figure 3. For instance, grafting method may handle this kind of situations. Indeed, it is a popular feature selection algorithm in the framework of streaming data [26]. It considers the selection of the optimal feature subset as an integral component of designing a prediction model in a regularized learning context. Grafting algorithm runs iteratively in an incremental manner, it builds up a feature set while training a model thanks to a gradient descent. In each iteration, a fast gradient based heuristic is adopted to select the feature that is more likely to enhance the current model, with the model being incrementally optimized using gradient descent.

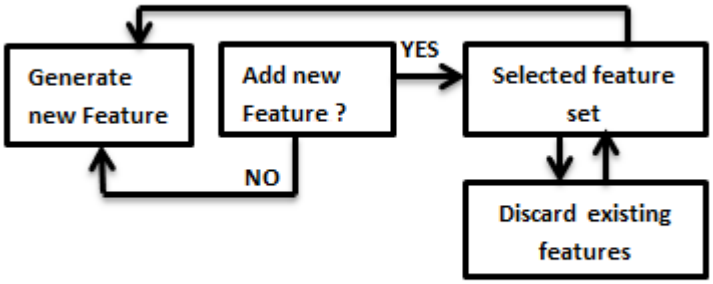

Fig. 3. General Framework of Streaming Feature Selection Method

V. Discussions and Challenges

A. classifier choice

When designing a wrapper feature selection model, the choice of a the classifier influences to a great degree the quality of the obtained feature subset. A recent study realized by Xue in [27] focused on the computational aspects of wrappers using different classifiers by performing extensive experimentation. This investigation has very helpful conclusions, it recommands adopting support vector machine algorithm [28] in the presence of a moderate sized feature space especially when time constraints are in force which seems to be the predominant case in real world applications. On the contrary, when dealing with large features space, KNN classifier proves to perform better.

B. Scalability

In the presence of high dimensional feature space, the existing feature selection methods face difficulties due to a lack of scalability. In fact, most of them obtain modest results regarding feature selection process for large scale classification when the total number of features exceeds thousands or tens of thousands. Therefore, it seems crucial that this area of research deserves more efforts in order to be capable of handling the constantly increasing amount of data especially in the streamming case.

C. Stability

Stability is defined as the sensitivity of a method to variations in the training set. The stability issue is observed in some of the well-known feature selection algorithms. Those methods experience some loss in their efficiency as soon as a small data perturbation is introduced in the training set. Even though feature selection methods have become recently more tailored seeking to adapt more to the problem they adress, those techniques still lack resilience regarding stability. In fact, each feature selection method is based on a particular optimization metaheuristic that has several own parameters needing to be tuned, for instance the population size parameter . A deeper understanding of how those values interact with a particular dataset may improve the ability of the resulting model.

VI. Conclusion and Future Works

Feature selection represents a crucial phase in the field of machine learning since it plays an important role in designing predictors capable of achieving satisfactory classification results. The objective of this overview is not to rank the available wrapper feature selection methods. The aim of this study is rather to put the light on the main issues regarding different aspects of this category of models while highlighting the challenges faced by this technology in the near term which will certainly come under the pressure of Big Data.

References

[1] L. H. Yu Lei, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004.

[2] L. Y. Huan Liu, "Toward integrating feature selection algorithms for classification and clustering," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 4, pp. 491–502, April 2005.

[3] S. Ding, "Feature selection based f-score and aco algorithm in support vector machine," in *Knowledge Acquisition and Modeling, 2009. KAM '09. Second International Symposium on*, vol. 1, Nov 2009, pp. 19–23.

[4] d. B. J. Lee Sungyoung, Park Young-Tack, "A novel feature selection method based on normalized mutual information," *Applied Intelligence*, vol. 37, no. 1, 2012.

[5] A.-A. Nadia, "A review on evolutionary feature selection," in *Proceedings of the 2014 European Modelling Symposium*, ser. EMS '14, 2014, pp. 20–26.

[6] S. M. Vipin Kumar, "Feature selection: A literature review," *Smart CR*, vol. 4, pp. 211–229, 2014.

[7] W. B. X. Y. B. Xue, M. Zhang, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. PP, no. 99, pp. 1–1, 2016.

[8] N. B. A Jovic, K Brkic, "A review of feature selection methods with applications," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*, 2015, pp. 1200–1205.

[9] K. F. P M Narendra, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers*, vol. C-26, no. 9, pp. 917–922, 1977.

[10] P. L. Lanzi, "Fast feature selection with genetic algorithms: a filter approach," in *Evolutionary Computation, 1997., IEEE International Conference on*, Apr 1997, pp. 537–540.

[11] H. S. E. Haupt Randy L, *Practical Genetic Algorithms with CD-ROM*. Wiley-Interscience, 2004.

[12] Y. Z. Jianjiang Lu, Tianzhong Zhao, "Feature selection based-on genetic algorithm for image annotation," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 887 – 891, 2008.

[13] A. A.-A. A AlSukker, R N Khushaba, "Enhancing the diversity of genetic algorithm for improved feature selection," in *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*, Oct 2010, pp. 1325–1331.

[14] V. K. J. D. P. M. A. S. M. J. J. Z. Vassil Alexandrov, Michael Lees, "2013 international conference on computational science a guided hybrid genetic algorithm for feature selection with expensive cost functions," *Procedia Computer Science*, vol. 18, pp. 2337 – 2346, 2013.

[15] T. C.-J. Y. C.-H. Chuang Li-Yeh, Chang Hsueh-Wei, "Improved binary pso for feature selection using gene expression data," *Comput. Biol. Chem.*, vol. 32, no. 1, pp. 29–38, Feb. 2008.

[16] W. N. B. Bing Xue, Mengjie Zhang, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," *Applied Soft Computing*, vol. 18, pp. 261 – 276, 2014.

[17] M. G. Parham Moradi, "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy," *Applied Soft Computing*, vol. 43, pp. 117 – 130, 2016.

[18] A. W. Whitney, "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, vol. C-20, no. 9, pp. 1100–1103, 1971.

[19] D. G. T Marill, "On the effectiveness of receptors in recognition systems," *IEEE Transactions on Information Theory*, vol. 9, no. 1, pp. 11–17, 1963.

[20] Z. M. Tran Binh, Xue Bing, *Overview of Particle Swarm Optimisation for Feature Selection in Classification*. Springer International Publishing, 2014, pp. 605–617.

[21] A. K. Rana Forsati, Alireza Moayedikia, "Article: A novel approach for feature selection based on the bee colony optimization," *International Journal of Computer Applications*, vol. 43, no. 8, pp. 13–16, April 2012.

[22] H. N.-p. Shima Kashef, "A new feature selection algorithm based on binary ant colony optimization," in *Information and Knowledge Technology (IKT), 2013 5th Conference on*, May 2013, pp. 50–54.

[23] F. S. J Silva, A Aguiar, "A parallel computing hybrid approach for feature selection," in *Computational Science and Engineering (CSE), 2015 IEEE 18th International Conference on*, Oct 2015, pp. 97–104.

[24] P. S. F. S. Filomena Ferrucci, M Tahar Kechadi, "A framework for genetic algorithms based on hadoop," *CoRR*, vol. abs/1312.0086, 2013.

[25] H. W. W. D. Xindong Wu, Kui Yu, "Online streaming feature selection," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, 2010, pp. 1159–1166.

[26] J. T. Simon Perkins, "Online feature selection using grafting," in *In International Conference on Machine Learning*, 2003, pp. 592–599.

[27] B. W. N. Xue Bing, Zhang Mengjie, "A comprehensive comparison on evolutionary feature selection approaches to classification," *International Journal of Computational Intelligence and Applications*, vol. 14, no. 02, p. 1550008, 2015.

[28] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.