© 2018 IJSRST | Volume 4 | Issue 5 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

A Survey on Feature Selection : in the Perspective of Evolutionary Approaches

Saravanan R[1], Subhasri A[2], Krithiga S[2]
[1]Associate professor, Department of Information Technology, SMVEC,Puducherry, Tamil Nadu, India
[2]UG student, Department of Information Technology, SMVEC, Puducherry, Tamil Nadu, India

ABSTRACT

Feature selection is an important task in data mining and machine learning to reduce the dimensionality of the data and increase the performance of the classification algorithm. However, feature selection is a challenging task to many of the problems mainly to the large search space. There are various methods to solve feature selection problems, where evolutionary computation (EC) techniques have recently added much attention and gave some success. However, the alternative approaches do not have complete guidelines on its strengths and weaknesses which lead to a disjointed and fragmented field with ultimately lost opportunities for improving performance and successful applications. This paper presents a broad survey of the state-of-the-art work on EC for feature selection, which identifies the contributions of these different algorithms. In addition, current issues and challenges are also discussed to identify promising areas for future research.

Keywords: Classification, Data Mining, Evolutionary Computation, Feature Selection, Machine Learning.

I. INTRODUCTION

In data mining and machine learning, real-world problems often involve a large number of features. However, not all features are essential since many of them are redundant or even irrelevant, which may reduce the performance of an algorithm, e.g., a classification algorithm. Feature selection aims to solve this problem by selecting only a small subset of relevant features from the original large set of features. By removing irrelevant and redundant features, feature selection can reduce the dimensionality of the data, speed up the learning process, simplify the learned model, and/or increase the performance [1], [2].

Feature selection is a difficult task due mainly to a large search space, where the total number of possible solutions is 2n for a dataset with n features [1], [2]. The task is becoming more challenging as n is increasing in many areas with the advances in the data collection techniques and the increased complexity of those problems. An exhaustive search for the best feature subset of a given dataset is practically impossible in most situations. However, most existing feature selection methods still suffer from stagnation in local optima and/or high computational cost [3], [4]. Therefore, an efficient global search technique is needed to better solve feature selection problems. Evolutionary computation (EC) techniques have recently received much attention from the feature selection community as they are well-known for their global search ability/potential. However, there are no comprehensive guidelines on the strengths and weaknesses of alternative approaches along with their most suitable application areas. This paper presents a comprehensive survey of the literature on EC for feature selection with the goal of providing interested researchers with the state-of-the-art research.

Feature selection has been used to improve the quality of the feature set in many machine learning tasks, such as classification, clustering, regression, and time series prediction [1]. This paper focuses mainly on feature selection for classification since there is much more work on feature selection for classification than for other tasks [1]. Recent reviews on feature selection can be seen in [5], [6], [7], and [8], which focus mainly on non-EC-based methods.

II. EXISTING WORK ON FEATURE SELECTION

This section briefly summarizes EC techniques from three aspects, which are the search techniques, the evaluation criteria, and the number of objectives.

1) Search Techniques: There are very few feature selection methods that use an exhaustive search [1], [5], [6]. This is because even when the number of features is relatively small (e.g., 50), in many situations, such methods are computationally too expensive to perform. Therefore, different heuristic search techniques have been applied to feature selection, such as greedy search algorithms, where typical examples are sequential forward selection (SFS) [9], sequential backward selection (SBS) [10]. However, both methods suffer from the so called "nesting effect" because a feature that is selected or removed cannot be removed or selected in later stages. "plus-l-take-away-r" [11] compromises these two approaches by applying SFS l times and then SBS r times. This strategy can avoid the nesting effect in principle, but it is hard to determine appropriate values for l and r in practice. To avoid this problem, two methods called sequential backward floating selection (SBFS) and sequential forward floating selection (SFFS) were proposed in [12]. Both floating search methods are claimed to be better than the static sequential methods.

Recently, Mao and Tsang [13] proposed a two-layer cutting plane algorithm to search for the optimal feature subsets. The results show that heuristic search techniques achieved similar performance to the backtracking algorithm but used a much shorter time. In recent years, EC techniques as effective methods have been applied to solve feature selection problems. Such methods include GAs, GP, PSO. Feature selection problems have a large search space, which is often very complex due to feature interaction. Feature interaction leads to individually relevant features becoming redundant or individually weakly relevant features becoming highly relevant when combined with other features. Compared with traditional search methods, EC techniques do not need domain knowledge and do not make any assumption about the search space, such as whether it is linearly or nonlinearly separable, and differentiable.

2) Evaluation Criteria: For wrapper feature selection approaches, the classification performance of the selected features is used as the evaluation criterion. Most of the popular classification algorithms, such as decision tree (DT), support vector machines (SVMs), Naïve Bayes (NB), K-nearest neighbor (KNN), artificial neural networks (ANNs), and linear discriminant analysis (LDA), have been applied to wrappers for feature selection [5], [6]. For filter approaches, measures from different disciplines have been applied, including information theory-based measures, correlation measures, distance measures, and consistency measures [1].Single feature ranking based on a certain criterion is a simple filter approach, where feature selection is achieved by choosing only the top-ranked features [15]. Single feature ranking methods are computationally cheap but do not consider feature interactions, which often leads to redundant feature subsets (or local optima) when applied to complex problems, e.g., microarray gene data, where genes possess intrinsic linkages [1], [2]. To overcome such issues, filter measures that can evaluate the feature subset as a whole have become popular.

Recently, Peng et al. [14] proposed the minimum redundancy maximum relevance method based on mutual information, where the proposed measures have been introduced to EC for feature selection due to their powerful search abilities. Mao and Tsang [8] proposed a novel feature selection approach by optimizing multivariate performance measures (which can also be viewed as an embedded method since the proposed feature selection framework was to optimize the general loss function and was achieved based on SVMs). However, the proposed method resulted in a huge search space for high-dimensional data, which required a powerful heuristic search method to find the optimal solutions. Statistical approaches, such as T-test, logistic regression, hierarchical clustering, and cart classification and regression tree (CART), are relatively simple and can achieve good performance [16]. Sparse approaches have recently become popular, such as sparse logistic regression for feature selection, which has been used for feature selection tasks with millions of features. However, many studies show that filter methods do not scale well to problems with more than tens of thousands of features [8].

1. **Number of Objectives**: Most of the existing feature selection methods aim to maximize the classification performance only during the search process or aggregate the classification performance and the number of features into a single objective function. To the best of our knowledge, all the multi-objective feature selection algorithms to date are based on EC techniques since their population-based mechanism producing multiple solutions in a single run is particularly suitable for multi-objective optimization.

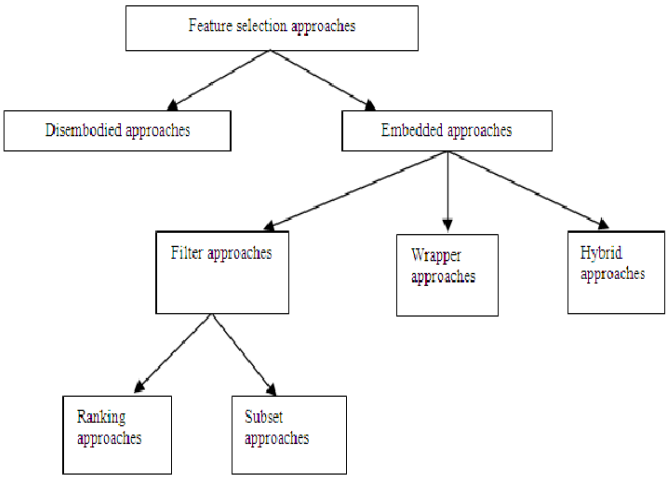

Figure 1

III. EC FOR FEATURE SELECTION

A. GAs for Feature Selection

GAs are most likely the first EC technique widely applied to feature selection problems. One of the earliest works was published in 1989. GAs have a natural representation of a binary string, where 1 shows the corresponding feature is selected and 0 means not selected. Table I shows the typical works on GAs for feature selection. It can be seen that there are more works on wrappers than filters, and more on single objective than multi-objective approaches. For filter approaches, different measures have been applied to GAs for feature selection, e.g., information theory [20], [18], [21], consistency measures [17], [18], rough set theory [19], and fuzzy set theory [99]. Many different new enhancements to GAs have been proposed to improve the performance, which focus mainly on the search mechanisms, the representation, and the fitness function.

Some early works [22], [23] introduced GAs to feature selection by investigating the influence of the population size, mutation, crossover, and reproduction operators, but with limited experiments. Recently, Derrac et al. [24] proposed a cooperative co-evolutionary algorithm for feature selection based on a GA with three populations, where the first focused on feature selection, the second focused on instance

selection, and the third focused on both feature selection and instance selection. The proposed algorithm addressed feature selection and instance selection in a single process, which reduced the computational time. Such approaches should be further investigated in the future given that large datasets (i.e., with thousands or tens of thousands of features) may include not only irrelevant features but also noisy instances.

Such approaches struggle to solve "big data" tasks, whereby both the number of features and the number of instances are huge. This is not only an issue for GAs, but also for other EC techniques for feature selection. To use GAs to address such tasks, a novel representation that can reduce the dimensionality of the search space will be needed. The design of genetic operators, e.g., crossover and mutation, provides opportunities to identify good building blocks (i.e., feature groups) and combine or adjust complementary features to find optimal feature subsets, but this is a challenging task. Furthermore, when and how to apply these operators and the parameter settings in GAs are also key factors that influence their performance on feature selection.

B. GP for Feature Selection

GP is used more often in feature construction than feature selection because of its flexible representation. In feature selection, most GP works use a tree-based representation, where the features used as the leaf nodes of a tree are the selected features. GP can be used as a search algorithm and also as a classification algorithm. In filter approaches, GP is mainly used as the search algorithm. In most wrapper (or embedded) approaches, GP is used as both the search method and the classification algorithm. In a very few cases, GP was used as a classification algorithm only in a feature selection approach [25]. One of the early works on GP for feature selection was published in 1996 [26], where a generalized linear machine was used as the classifier to evaluate the fitness of the selected features. Later, Neshatian and Zhang [27] proposed a wrapper feature selection approach based on GP, where a variation of NB algorithm was used for classification. A bitmask encoding was used to represent feature subsets. Set operators were used as primitive functions. GP was used to combine feature subsets and set operators together to find an optimal subset of features. However, it may suffer from the problem of high computational cost. In most works, GP was used to search for the optimal feature subset and simultaneously trained as a classifier. GP can handle tasks with a very small number of instances [28], which provides an opportunity to better solve feature selection tasks with a small number of instances. When and how to apply genetic operators is also important in GP, but the design and the use of the genetic operators in GP is more difficult than in GAs due to the flexible representation and the different return types of the functions. The parameter settings in GP are also very important. Because of the large population size, GP may suffer from the issue of being computationally expensive.

C. PSO for Feature Selection

Both continuous PSO and binary PSO have been used for both filter and wrapper, single objective and multi-objective feature selection. The representation of each particle in PSO for feature selection is typically a bit-string, whereby the dimensionality is equal to the total number of features in the dataset. The bit-string can be binary numbers in binary PSO or real-value number0s in continuous PSO. When using binary representation, 1 means the corresponding feature is selected and 0 means it is not selected. When using the continuous representation, a threshold θ is usually used to determine the selection of a particular feature, i.e., if the value is larger than θ, the corresponding feature is selected. Otherwise, it is not selected.

The proposed algorithm was shown to be able to significantly reduce the number of features. Lane et al.

[30] further improved the algorithm by allowing the selection of multiple features from the same cluster to further improve the classification performance. Later, Nguyen et al. [29] proposed a new representation, where the dimensionality of each particle was determined by the maximum number of desired features. The dimensionality of the new representation is much smaller than the typical representation; however, it is not easy to determine the desired number of features. Learning from neighbors' experience, i.e., social interaction through gbest, and learning from each individual's own experience through pbest, are the key ideas in PSO. Chuang et al. [31] developed a best resetting mechanism by including zero features in order to guide the swarm to search for small feature subsets.

Xue et al. [32] considered the number of features when updating pbest and gbest during the search process of PSO, which could further reduce the number of features over the traditional updating pbest and gbest mechanism without deteriorating the classification performance. Thus, GAs are likely to be suited to domains in which there are groups of interacting features, potentially with multiple good subsets, to consider. PSO has a more structured neighborhood guiding its recombination method than GAs, as well as a velocity term that enables fast convergence to a solution. PSO should suit domains in which there is a structure in how features interact, i.e., low sensitivity to the inclusion of each feature in a solution, and where fast convergence does not lead to local optima. PSO has an advantage over GAs and GP of being easy to implement. Developing novel PSO algorithms, particularly novel search mechanisms, parameter control strategies and representation for large-scale feature selection, is still an open issue.

IV. ISSUES AND CHALLENGES

Despiting the suitability, success, and promise of EC for feature selection, there are still significant issues and challenges, which will be discussed here.

A. Scalability

Computational intelligence-based techniques have been introduced to feature selection tasks in the ranges of millions [8]. Most of the existing EC-based large-scale feature selection approaches employ a two-stage approach, where in the first stage, a measure is used to evaluate the relevance of individual features, then ranks them according to the relevance value. Only the top-ranked (better) features are used as inputs to the second stage to further select features from them. However, the first stage removes lowly-ranked features without considering their interaction with other features. To solve large-scale feature selection problems, new approaches are needed, including new search algorithms and new evaluation measures.

B. Computational Cost

Most feature selection methods suffer from the problem of being computationally expensive, which is a particularly serious issue in EC for feature selection since they often involve a large number of evaluations. Therefore, it is still a challenge to propose efficient and effective approaches to feature selection problems. To reduce the computational cost, two main factors, an efficient search technique and a fast evaluation measure, need to be considered [1]. A fast evaluation criterion may produce a greater influence than the search technique, since in current approaches the evaluation procedure takes the majority of the computational cost. It is noted that the parallelizable nature of EC is suited as grid computing, graphics processing unit, and cloud computing that can be used to speed up the process.

C. Search Mechanisms

Feature selection is an NP-hard problem and has a large complex solution space [33]. This requires a powerful global search technique and current EC algorithms still have great potential to be improved. The new search mechanisms should have the ability to explore the whole search space and also be able to exploit the local regions when needed. EC algorithms are stochastic approaches, which may produce different solutions when using different starting points. Even when the fitness values of the solutions are the same, they may select different individual features. Therefore, the stability of the algorithms not only involves the difference of the fitness values, but also involves the consistency of the selected features. Therefore, to propose new search algorithms with high stability is also an important task.

D. Multi-Objective Feature Selection

Most of the existing EMO algorithms are designed for continuous problems [34], but feature selection is a discrete problem. This requires the development of novel EMO algorithms. Furthermore, the two main objectives are not always conflicting with each other. This makes it tricky to design an appropriate EMO algorithm. Furthermore, developing new evaluation metrics and further selection methods to choose a single solution from a set of trade-off solutions is also a challenging topic. Finally, besides the two main objectives, other objectives, such as the complexity, the computational time, and the solution size could also be considered in multi-objective feature selection.

E. Feature Construction

Feature selection does not create new features, as it only selects original features. However, if the original features are not informative enough to achieve promising performance, feature selection may not work well, yet feature construction may work well [35]. One of the challenges for feature construction is to decide when feature construction is needed. A measure to estimate the properties of the data might be needed to make such a decision. Meanwhile, feature selection and feature construction can be used together to improve the classification performance and reduce the dimensionality. This can be achieved in three different ways: 1) performing feature selection before feature construction; 2) performing feature construction before feature selection; and 3) simultaneously performing both feature selection and construction [35].

V. CONCLUSION

This paper provided a comprehensive survey of EC techniques in solving feature selection problems, which covered all the commonly used EC algorithms and focused on the key factors, such as representation, search mechanisms, and the performance measures as well as the applications. Important issues and challenges were also discussed.

This survey shows that a variety of EC algorithms have recently attracted much attention to address feature selection tasks. A popular approach in GAs, GP, and PSO is to improve the representation to simultaneously select features and optimize the classifiers, e.g., SVMs. Different algorithms have their own characteristics, such as GAs are able to preserve a small set of features during the evolutionary process because of the nature of genetic operators, PSO is relatively computationally cheap because of its simple updating mechanisms.

The proposal of novel approaches may involve methods or measures from different areas, which encourages research across multiple disciplines. A comprehensive comparison between EC and non EC approaches on a large number of benchmark datasets/problems to test their advantages and disadvantages can help develop novel effective approaches to different kinds of problems. In addition, combining feature selection with feature construction can potentially improve the classification performance,

whereas combining feature selection with instance selection can potentially improve the efficiency.

VI. REFERENCES

[1]. M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., vol. 1, nos. 1-4, pp. 131-156, 1997.

[2]. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157-1182, Mar. 2003.

[3]. A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," Eur. J. Oper. Res., vol. 206, no. 3, pp. 528-539, 2010.

[4]. Y. Liu et al., "An improved particle swarm optimization for feature selection," J. Bionic Eng., vol. 8, no. 2, pp. 191-200, 2011.

[5]. H. Liu and Z. Zhao, "Manipulating data and dimension reduction meth-ods: Feature selection," in Encyclopedia of Complexity and Systems Science. Berlin, Germany: Springer, 2009, pp. 5348-5359.

[6]. H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," in Proc. JMLR Feature Sel. Data Min., vol. 10. Hyderabad, India, 2010, pp. 4-13.

[7]. J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," Neural Comput. Appl., vol. 24, no. 1, pp. 175-186, 2014.

[8]. Y. Zhai, Y.-S. Ong, and I. W. Tsang, "The emerging 'big dimen-sionality,"' IEEE Comput. Intell. Mag., vol. 9, no. 3, pp. 14-26, Aug. 2014.

[9]. A. W. Whitney, "A direct method of nonparametric measurement selection," IEEE Trans. Comput., vol. C-20, no. 9, pp. 1100-1103, Sep. 1971.

[10]. T. Marill and D. M. Green, "On the effectiveness of receptors in recog-nition systems," IEEE Trans. Inf. Theory, vol. 9, no. 1, pp. 11-17, Jan. 1963.

[11]. S. D. Stearns, "On selecting features for pattern classifier," in Proc. 3rd Int. Conf. Pattern Recognit., Coronado, CA, USA, pp. 71-75, 1976.

[12]. P. Pudil, J. Novovicová,ˇ and J. V. Kittler, "Floating search meth-ods in feature selection," Pattern Recognit. Lett., vol. 15, no. 11.

[13]. Q. Mao and I. W.-H. Tsang, "A feature selection method for multivari-ate performance measures," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 9, pp. 2051-2063, Sep. 2013.

[14]. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8.

[15]. W. A. Albukhanajer, J. A. Briffa, and Y. Jin, "Evolutionary multiob-jective image feature extraction in the presence of noise," IEEE Trans. Cybern., vol. 45, no. 9, pp. 1757-1768, Sep. 2015.

[16]. N. C. Tan, W. G. Fisher, K. P. Rosenblatt, and H. R. Garner, "Application of multiple statistical tests to enhance mass spectrometry-based biomarker discovery," BMC Bioinformat., vol. 10, p. 144, May 2009.

[17]. P. L. Lanzi, "Fast feature selection with genetic algorithms: A filter approach," in Proc. IEEE Int. Conf. Evol. Comput., Indianapolis, IN, USA, 1997, pp. 537-540.

[18]. N. Spolaôr, A. C. Lorena, and H. D. Lee, "Multi-objective genetic algo-rithm evaluation in feature selection," in Evolutionary Multi-Criterion Optimization (LNCS 6576). Heidelberg, Germany: Springer, 2011, a.462-476.

[19]. M. Banerjee, S. Mitra, and H. Banka, "Evolutionary rough feature selection in gene expression data," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 37, no. 4, pp. 622-632, Jul. 2007.

[20]. B. Xue, L. Cervante, L. Shang, W. N. Browne, and M. Zhang, "Multi-objective evolutionary algorithms for filter based feature selection in

classification," Int. J. Artif. Intell. Tools, vol. 22, no. 4, 2013, Art. ID 1350024.

[21]. H. Xia, J. Zhuang, and D. Yu, "Multi-objective unsupervised fea-ture selection algorithm utilizing redundancy measure and negative epsilon-dominance for fault diagnosis," Neurocomputing, vol. 146.

[22]. R. Leardi, R. Boggia, and M. Terrile, "Genetic algorithms as a strategy for feature selection," J. Chemometr., vol. 6, no. 5, pp. 267-281, 1992.

[23]. M. Demirekler and A. Haydar, "Feature selection using genetics-based algorithm and its application to speaker identification," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., Phoenix, AZ, USA, 1999, vol. 1, pp. 329-332.

[24]. J. Derrac, S. Garcia, and F. Herrera, "A first study on the use of coevolutionary algorithms for instance and feature selection," in Hybrid Artificial Intelligence Systems (LNCS 5572). Berlin, Germany: Springer, 2009, pp. 557-564.

[25]. S. M. Winkler, M. Affenzeller, W. Jacak, and H. Stekel, "Identification of cancer diagnosis estimation models using evolutionary algorithms: A case study for breast cancer, melanoma, and cancer in the res-piratory system," in Proc. 13th Annu. Conf. Compan. Genet. Evol. Comput. (GECCO), Dublin, Ireland, 2011, pp. 503-510.

[26]. J. Sherrah, R. E. Bogner, and A. Bouzerdoum, "Automatic selection of features for classification using genetic programming," in Proc. Aust. New Zealand Conf. Intell. Inf. Syst., Adelaide, SA, Australia, 1996,284-287.

[27]. K. Neshatian and M. Zhang, "Dimensionality reduction in face detec-tion: A genetic programming approach," in Proc. 24th Int. Conf. Image Vis. Comput. New Zealand (IVCNZ), Wellington, New Zealand, 2009, 391-396.

[28]. H. Al-Sahaf, M. Zhang, and M. Johnston, "Genetic programming for multiclass texture classification using a small number of instances," in Simulated Evolution and Learning (LNCS 8886). Cham, Switzerland: Springer, 2014, pp. 335-346.

[29]. H. B. Nguyen, B. Xue, I. Liu, and M. Zhang, "PSO and statistical clustering for feature selection: A new representation," in Simulated Evolution and Learning (LNCS 8886). Cham, Switzerland: Springer, 2014, pp. 569-581.

[30]. M. C. Lane, B. Xue, I. Liu, and M. Zhang, "Gaussian based par-ticle swarm optimisation and statistical clustering for feature selec-tion," in Evolutionary Computation in Combinatorial Optimisation (LNCS 8600). Berlin, Germany: Springer, 2014, pp. 133-144.

[31]. L.-Y. Chuang, H.-W. Chang, C.-J. Tu, and C.-H. Yang, "Improved binary PSO for feature selection using gene expression data,"Comput. Biol. Chem., vol. 32, no. 1, pp. 29-38, 2008.

[32]. B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," Appl. Soft Comput., vol. 18, pp. 261-276, May 2014.

[33]. E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," Theor. Comput. Sci., vol. 209, nos. 1-2, pp. 237-260, 1998.

[34]. C. A. C. Coello, "Evolutionary multi-objective optimization: A his-torical view of the field," IEEE Comput. Intell. Mag., vol. 1, no. 1,28-36, Feb. 2006.

[35]. A. S. U. Kamath, K. De Jong, and A. Shehu, "Effective auto-mated feature construction and selection for classification of biological sequences," PLoS One, vol. 9, no. 7, 2014, Art. ID e99982.