Fundamentals of feature selection: an overview and comparison

Amina Benkessirat[1], Nadjia Benblidia[2]
LRDSI Laboratory, University of Blida 1, Algeria
[(1)]aminabks-gsi@outlook.fr,[(2)]benblidia@univ-blida.dz

***Abstract* - Tremendous efforts have been put into the development of Feature Selection (FS) methods by the machine learning community. In this paper, we present basics surrounding this topic, providing its general process, evaluation procedure and metrics. In addition, we thoroughly discuss major application aspects. We also provide a comprehensive overview and comparison of some existing feature selection methods. We conclude this work by highlighting some critics, challenges, and future research directions of feature selection.**

Keywords –feature selection, search strategies, evaluation, relevance, redundancy, real world applications.

I. INTRODUCTION

The challenge of the data mining task is to identify the relation between the features of the data and some endpoint (for instance the target class for a classification task). In most dataset only few features are relevant and contribute to determinate the endpoint. The remaining features contribute to the global dimensionality of the problem space, which involves a large memory required to store all the features, a large processing time to get the wanted result, or biased results because of the noisy feature [1, 2]. Consequently, data pre-processing is of paramount importance. Selecting a sub-set of features is considered an appropriate solution to address the problems above, and has become a primordial component in the machine learning process [3]. Feature selection (FS) is the process that aims to find a "relevant" subset of features from the original set. The relevancy of features depends on the objectives of the system [1][4, 5]. A feature can be classified by their relevancy with three qualifiers: i) an irrelevant feature must be removed from the original set. ii) a feature f is weakly relevant if there exist a sub-ensemble V where the performance of V is better than the performance of $V \cup \{f\}$. iii) a feature f is strongly relevant if its absence from the selected sub-ensemble implies a significant deterioration of the performance of the system. Feature selection approaches evaluate the features individually by assigning a weight to the current feature, or evaluate a subset of features using a search strategy [6]. To select the relevant features, obviously, we have to search in the space of $2^n - 1$ candidates. This problem has been proven to be NP-hard [1][7, 8], and it appears to be difficult to produce a globally optimal solution.

II. FEATURE SELECTION PROCESS

The general procedure for FS has four key steps: subset Generation, subset evaluation, stopping criteria, result validation.

A. SUBSET GENERATION

Before applying a search algorithm, a starting point must be defined. In the present study, the starting point is the set of input features. The starting set could be empty and different features could be added successively or include all the features and some of them are deleted if required. Furthermore, the starting set could be a subset of randomly selected features and one or more features are added or deleted at each iteration [9]. Once the starting set is well defined, a search strategy must be selected. Therefore, the search algorithm is important to select the relevant features. For a space of N features, there are $2^N - 1$ candidate subsets [1] [9]. This search space is massive, and with a moderate N; a thorough search is impossible. Thus, three strategies are proposed in the literature: exponential, sequential and random.

In the sequential procedures, one or more features could be added or deleted [1][9,10]. Several possibilities of this strategy were proposed in the literature: 1) mainly forward variant where the research starts from an empty set and one feature at least is added to each iteration; 2) backward variant starts the research from a set containing all the features and at least one of them is deleted in each iteration; 3) and stepwise variant which mixes the above variants and add or remove at least one feature in each iteration. These algorithms are easy to implement as they are of the order $O(n^2)$ or lower [1][9,10]. 4) Finally, random variant generates a finite number of subsets and the best one is then selected. It seems that this procedure goes through part of the search space. All the algorithms of this category use randomness to escape the local optimum in the search space. It is of the order $O(n^2)$[1].

B. SUBSET EVALUATION

Evaluating the discriminating power of a feature is an important step in the selection process. Several methods are used for the evaluation, based on different criteria. The evaluation criteria are divided into two categories; dependent and independent criteria [1]. The dependent criteria are used by wrapper evaluation procedures, to evaluate the subsets by involving a learning algorithm. Usually it generates better results compared to the independent criteria, but it is computationally [1][9]. The independent criteria are used by the filter evaluation procedure to evaluate the subsets without involving a learning algorithm. In data mining, there are three general approaches for FS: filter, wrapper and embedded.

978-1-7281-5052-9/19/$31.00 ©2019 IEEE

B.1 FILTER ALGORITHMS

They use features independently of the classifier, without involving a learning algorithm [11, 12]. Therefore, this method is considered a preprocessing of data before the learning phase. The filter algorithms are divided into univariate and multivariate categories [2]. The features of the univariate algorithms are evaluated one by one, regardless of the dependence with the other features [2][13, 14]. These methods help to eliminate features that seem useless but are useful if they are combined with others, which negatively affect the system performance. In order to solve this problem, multivariate methods are proposed [2][14]. In the category of filter algorithms, a relevance score is assigned to each feature after the evaluation of its performance. This score measures the power of the feature in discriminating different classes [8]. A threshold is defined to delete the useless features [8]. The remaining set forms the input of the classification algorithm [10]. The performance measures developed for the filtering methods are: information, distance, coherence, similarity and statistical measures [13]. This approach is effective and fast to calculate [1]. Many existing algorithms fall under filter approach such as: B&B [15], BFF [16], MDLM [17], Bobrowski [17], Focus [18], ABB [19], Schlimmer's [20] are based on exponential search strategy.Relief [21], Relief's [22], SFS [23], SBS [24], Koller [25], FG [26], CFS [27], Set cover [28] arebased on sequential search strategy.Liv and QBB [26] are based on random search strategy.

B.2 WRAPPER ALGORITHMS

This category of algorithms depends on the used classifier [2]. The search for a relevant subset is based on a learning algorithm as an evaluation function. They fall into two categories: deterministic or random. It seems that this approach has better performance compared to the filter algorithm [1][13]. However, these methods have two main disadvantages [13]. The selected subset depends strongly on the classifier, and they are both time consuming because of the complexity of the learning algorithms. This complexity makes impossible the use of an exhaustive strategy (NP-complete problem) [1][8]. Many existing algorithms fall under filter approach such as: BS [29], AMB&B [30], FSLC [31], FSBC [31] which are based on exponential search strategy.WSFG and WSBG [32], BDS [29], SS and RACE [33], RC and BSE [34] arebased on sequential search strategy.SA and RGSS [29], LVW [35], GA [36], RVE [37] are based on random search strategy.

B.3 Embedded algorithms

These algorithms interact with the learning algorithms and require shorter computation time, compared to the wrapper algorithm [1]. Many existing algorithms fall under embedded approach such as BBHFS[38], Xing's[39, 40].

C. STOPPING CRITERIA

Selection techniques require a stopping criterion to avoid an exhaustive search of the subsets. Several stopping criteria are proposed in the literature, mainly [1][9]:

- the search completes;
- predefined maximum number of iterations is achieved;
- predefined maximum number of features is achieved;
- adding or removing a feature does not produce a better subset;
- an optimal subset is found.

D. RESULTS VALIDATION

There are two validation alternatives that depend on the nature of the used data [1][9]. If the artificial data are used, the relevant features are known, and the validation is done directly by checking if they belong to the selected subset. If the real data are used, the relevant features are not known; therefore, the accuracy of the selected subset is tested using a classifier.

III. METRIC EVALUATION OF A FEATURE SELECTION ALGORITHM

The evaluation of an algorithm of feature selection is done by comparing the classification results with those of other algorithms. In machine learning, several simple measures are defined to evaluate the performance of a classification model. These measures could be defined according to the provided information by the confusion matrix (Table 1) [41-43].

TABLE I
CONFUSION MATRIX FOR BINARY CLASSIFICATION

	Actual positive class	Actual negative class
Predicted positive class	True positive TP	False negative FN
Predicted negative class	False positive FP	True negative TN

The matrix rows represent the predicted class, while the columns represent the real class. TP and TN are the number of correctly classified positive and negative instances, and FP and FN are the number of positive and negative instances that are misclassified [44]. Several measures can be calculated from the confusion matrix as shown in table 2.

TABLE II
MEASURE FOR BINARY CLASSIFICATION EVALUATION

Name	***Formula***	***Description***
Accuracy (Acc)[45]	$Acc = \frac{TN + TP}{TP + FP + FN + TN}$	The ratio of correct predictions over the total number of instances evaluated
Error rate (Err)	$Err = \frac{FN + FP}{TP + FP + FN + TN}$	The ratio of incorrect predictions over the total number of instances evaluated
Sensitivity (Sn)	$Sn = \frac{TP}{TP + FN}$	The fraction of positive patterns that are correctly classified
Specificity (Sp)	$Sp = \frac{TN}{TN + FP}$	The fraction of negative patterns that are correctly classified
Precision (Pr)[45]	$Pr = \frac{TP}{TP + FP}$	The positive patterns that are correctly

		predicted from the total predicted patterns in a positive class.
Recall (Rec)[45]	$Rec = \frac{TP}{TP + TN}$	The fraction of positive patterns that are correctly classified
F_Measure (FM) [44]	$FM = \frac{2 \times Pr \times Rec}{Pr + Rec}$	The harmonic mean between recall and precision values
Geometric-mean (GM) [44]	$GM = \sqrt{TP \times TN}$	The average type maximize the TPrate and TNrate, and simultaneously keeping both rates relatively balanced

According to [44], the FM and GM measure performs better than the accuracy to optimizing binary classifications. The measures presented in Table 2 are two-class problem measures, however they can be extended to a multi-class problem by considering one of the classes as positive and the others as negative [41]. The average of these measurements for the individual classes becomes the final value of the entire model. Other less used measures were reported in the literature; for instance, the mean square error (MSE), the root mean square error (RMSE), Cohen's KAPPA and area under the ROC curve (AUC) [41, 42]. However, other important evaluation criteria are proposed such as the learning time (T) and the number of selected features (Nb) [46, 47].

IV. Application of Feature Selection in Real World

In any application field, data collection is a key step. However, the huge number of features constituting these data and the irrelevant and redundant feature remain a major problem in data collection. Therefore, several methods for feature selection were developed and successfully applied in different fields, such as image processing [46] [48,49], genomic analysis [2], intrusion detection [50], remote sensing [51] and text mining [52]. Table 3 shows an overview of some of the selection methods proposed in the literature. A discussion is done, based mainly on their scope of application.

In the first work, the authors presented a new nonlinear feature selection method named flexible integration, based on discriminant graphs (FDEFS). The proposed algorithm is for supervised and unsupervised classification. This method incorporates the manifold smoothness, margin discriminant embedding and the sparse regression for feature selection. The test was applied to six bases of public images by comparing the proposed method with 12 other methods. The experiment results showed the superiority of the proposed method in terms of recognition rate. Furthermore, the combination of other techniques mainly sparse regression is an interesting direction. In second work, the authors presented two new methods for feature selection named OSFSMI and OSFSMI-k. The proposed methods aimed to select the features of online streaming. Feature relevance and redundancy are evaluated using mutual information. These methods can be classified as filter methods since no learning model is used in the search process. The test of this selected method was applied on 29 datasets by comparing the methods selected with recent methods reported in the literature. The obtained results showed that the elected methods perform better than the literature methods, in term of running time, number of selected features, accuracy of the classification and stability of the method. In third work, the authors proposed a new selection strategy using genetic algorithm based on random forests GARF. The features are extracted from positron emission tomography (PET) images and clinical data. The test was applied to a dataset collected from 65 patients and results were compared to four competing methods. The experimental results showed a better performance of GARF compared to the other methods, in terms of number of selected features, sensitivity, specificity, and classification error. Since the test was applied to a small number of cases, it would be of great interest to target larger population to assess the robustness of the method. The extension of the system to detect other types of pathologies would be of major interest. In the fourth work, the authors proposed an improvement of the clonal algorithm for the feature selection for intrusion detection. The evaluation of the features performance is based on the combination of the weight measured by deviation minimum redundancy and the biggest correlation analysis between features and label features. The test of the proposed was applied on the basis of KDD CUP99[50], by comparing the improved-clonal method with the traditional clonal method. Experimental results show that the improved algorithm has higher precision compared to the traditional algorithm. However, the improved algorithm increases the execution time. Solving this problem will be the subject of a major future research. In the fifth work, the authors developed a new supervised technique for feature selection, guided by scalable algorithms. The generated subsets are evaluated using a wrapper template in which the fuzzy *knn* classifier is considered. The technique uses the relief algorithm to remove redundant features. The test was applied to three sets of remotely sensed images, comparing the obtained results with those of four other techniques. The results are promising in terms of accuracy of classification and Kappa +coefficient. The improvement of the convergence by varying the population initialization is an interesting track. In the sixth work, the authors proposed a new system of selection and extraction of deep feature D-FES, in order to detect impersonation attacks in Wi-Finetworks. The extraction is conducted by stacking whereas the selection was done by weighting. The test was applied on a Wi-Fi network repository data set, namely the Aegean Wi-Fi intrusion dataset (AWID). The detection accuracy was 99.92%, with false alarm rate of 0.01%. To our knowledge, these results are better than those reported in the literature. This work is limited to the detection of identity theft attacks, the extension of the system to detect other types of attacks would be of major interest. In the seventh, the authors proposed a new method of feature selection, based on the maximum partial correlation information (PMCI). The first step of the method is to extract several orthogonal components from the features space to evaluate them one by one.

TABLE III
SUMMARIZED OF RELEVANT FEATURE SELECTION METHODS IN VARIOUS APPLICATION AREAS

Study	*Area application*	*Subarea*	*Dataset*	*Feature selection method*	*Evaluation metric*
1 [49]	Digital image processing	Classification	six public image datasets including scene, face and object datasets	FDEFS comparing to SDE, SDA, LE, GHFH, RMGT, LLE, MRDL, MFME, JELSR, JELSR-KNN, FSS and FSS-KNN	recognition rate
2 [46]		Online streaming	29 datasets from different domain	OSFSMI and OFSMI-K comparing to mRMR and others	T, Acc, Nb, St
3 [48]		Computer-aided diagnosis	a cohort of 65 patients with a local oesophageal cancer	**GARF** comparing to SFS, HFS, RFE, and Lasso	Err, AUC from ROC, Nb, Se, Sp
4 [50]	Security	intrusion detection	KDD CUP99	Improved-clonal comparing to clonal	Acc, FP, TP, T
5 [51]		Remote sensing	3 hyperspectral remotely sensed images sets	**SADE**	Pr, Kappa
6 [53]		Impersonation Detection	Aegean Wi-Fi Intrusion Dataset (AWID)	**D-FES**	Acc, Err
7 [2]	Genomic analysis	microarray data classification	10 UCI benchmark datasets	**PMCI** comparing to mRMR and others	Acc, kappa, Nb, T
8 [54]	Generic feature selection model	Classification using the available datasets from UCI Machine Learning Repository	6 UCI benchmark datasets	**CSO** comparing to PSO and canonical PSO	Err, Nb, Acc
9 [47]			11 UCI benchmark datasets	ROGA INRSG IBGAFG	Acc, Nb, T
10 [43]			4 UCI benchmark datasets	FS-JMIE comparing to MI, CHI and BPSO-SVM	Acc, Pr, Se, Rec, Nb, F1_score, Q
11 [55]			11 UCI benchmark datasets	mDSM Comparing to CMIM, MIM, MIFS, mRMR, and JMI	Acc, F1_score.

The extraction is based on the correlation between the features and the class. Furthermore, the authors proposed a new generic class-coding scheme for multi-class problems. The test was applied on 10 microarray benchmark datasets by comparing the selected method with six methods. The obtained results show that PMCI is an effective and efficient method and offers a reduced temporal complexity. Applying this method on real data is an interesting field. In the eighth work, the authors used a recent variant of particle swarm optimization (PSO), namely competitive swarm optimizer (CSO). This variant aims to a large-scale optimization. The CSO is originally developed for continuous optimization. In the present study, the authors adapted it to the problem of the selection of features by considering it as a combinatorial optimization problem. The test was applied on six reference datasets. The results show that CSO selected a small number of features compared to PSO while the classification performance was significantly better. In the ninth work, the authors proposed a feature selection model based on granular information. In their research, the authors examined how the level of granularity affects the accuracy of classification and the size of the subset in parallel. The improved binary genetic algorithm with feature granulation (IBGAFG) and the improved neighborhood rough set with sample granulation (INRSG) were used to select the most significant features and the improvement of selected subassembly. In order to determine the optimal granular radius, the authors presented an optimization of granularity based on a genetic algorithm (ROGA). The test was applied to 11 benchmark datasets. Moreover, ROGA algorithm was applied to a financial dataset. The results of the experiments shown that the approaches were effective and could provide higher classification accuracy using granular information. The use of other research strategies and granulation methods may improve the classification accuracy. In the tenth work, the authors proposed a new feature selection method based on the joint maximal information entropy between features and class (FS-JMIE). Firstly, the joint information entropy (JMIE) was defined to generate the subsets. Then, a binary particle swarm optimization (BPSO) algorithm was presented to search the optimal subset. Finally, a classification was carried on four datasets of reference to verify the performance of the proposed method, compared to the traditional method of mutual information (MI), CHI method, as well as the binary version of the function BPSO-SVM. The results shown that FS-JMIE achieves better performance compared to other methods, in terms of classification accuracy and the number of selected features. FS-JMIE is based on a calculation of the MIC between two vectors, which remains a research problem to overcome,

thus, it would be interesting to change this step. In the eleventh work, the authors proposed a new feature selection framework, namely modified discretization and feature selection based on mutual information (mDSM). The aim was to consider corrections when calculating mutual information for the finite samples. The test was applied on 30 reference datasets and compared to five mutual information-based feature selection methods. Experimental results show that in most cases, the proposed methods outperform all the available high-end methods.

In the present paper, we discussed in details the feature selection process. The general procedure for the feature selection involves several key steps mainly, generation and evaluation of subsets, stopping criteria and validation of results. In general, feature selection methods are based on the analysis of the correlation between features and the target class to maximize relevance, and on the correlation between the features to minimize redundancy. Therefore, a good feature selection system must consider the relevance-redundancy trade-off. We found that most tests apply to a small number of datasets on which the test method is running well. Therefore, we suggest expanding the number of datasets, and targeting several domains, especially for generic methods. Furthermore, the calculation time depends strongly on the search strategy and the evaluation procedure. Thus, the era of big data encourages us to develop fast methods that work with millions of features and billions of samples. We think that larger comparative studies, with the same data sets, need to be investigated to obtain more reliable results.

VII. Conclusion and future work

In the present paper, we discussed in details the feature selection process. The general procedure for the feature selection involves several key steps mainly, generation and evaluation of subsets, stopping criteria and validation of results. In general, feature selection methods are based on the analysis of the correlation between features and the target class to maximize relevance, and on the correlation between the features to minimize redundancy. Therefore, a good feature selection system must consider the relevance-redundancy trade-off. We found that most tests apply to a small number of datasets on which the test method is running well. Therefore, we suggest expanding the number of datasets,
and targeting several domains, especially for generic methods. Furthermore, the calculation time depends strongly on the search strategy and the evaluation procedure. Thus, the era of big data encourages us to develop fast methods that work with millions of features and billions of samples. We think that larger comparative studies, with the same data sets, need to be investigated to obtain more reliable results.

Relevance and redundancy are two important feature properties that must be compromised in feature selection problem. The most existing methods eliminate redundancy by measuring the pairwise inter-correlation of features, while the complementarity of features and the interaction between more than two features are not considered. Therefore, it would be interesting to evaluate the features utility considering the complementarity and the interaction between more than two feature; this by introducing a new evaluation criterion.

Finally, deep learning is an important advance in machine learning field, it combines low-level features to form more abstract high-level features. An irrelevant feature might cost a lot of resources during the process of training neural networks. Therefore, we should think about removing the redundant and noisy features before training the depth learning model. Thus, we think that training the depth model with only the relevant features is a challenge to consider by the machine learning community.

References

[1] V. Kumar and S. Minz, “Feature Selection: A literature Review,” Smart CR, vol. 4, no. 3, pp. 211-229, 2014.

[2] M. Yuan, Z. Yang and G. Ji, “Partial maximum correlation information: A new feature selection method for microarray data classification,” Neurocomputing, vol. 323, pp. 231-243, 2019.

[3] A. Kalousis, J. Prados and M. Hilario, “Stability of feature selection algorithms: a study on high-dimensional spaces,” Knowledge and information systems, vol. 12, no. 1, pp. 95-116, 2007.

[4] A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” Artificial intelligence, vol. 97, no. 1-2, pp. 245--271, 1997.

[5] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” Neural computing and applications, vol. 24, no. 1, pp. 175-186, 2014.

[6] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of machine learning research, vol. 3, no. mar, pp. 1157--1182, 2003.

[7] J. a. P. J. Han and M. Kamber, Data mining: concepts and techniques, Elsevier, 2011.

[8] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” Computers & Electrical Engineering, vol. 40, no. 1, pp. 16--28, 2014.

[9] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” IEEE Transactions on Knowledge & Data Engineering, no. 4, pp. 491-502, 2005.

[10] M. Dash and H. Liu, “Feature selection for classification,” Intelligent data analysis, vol. 1, no. 1-4, pp. 131-156, 1997.

[11] G. H. John, R. Kohavi and K. Pfleger, Irrelevant features and the subset selection problem, Elsevier, 1994.

[12] E. Guldogan and M. Gabbouj, “Feature selection for content-based image retrieval,” Signal, Image and Video Processing, vol. 2, no. 3, pp. 241-250, 2008.

[13] A. Joviç, K. Brkiç and N. Bogunoviç, “A review of feature selection methods with applications,” 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1200-1205, 2015.

[14] Z. M. Hira and D. F. Gillies, “A review of feature selection and feature extraction methods applied on microarray data,” Advances in bioinformatics, p. Hindawi, 2015.

[15] P. M. Narendra and K. Fukunaga, “A branch and bound algorithm for feature subset selection,” IEEE Transactions on computers, no. 9, pp. 917-922, 1977.

[16] L. Xu, P. Yan and T. Chang, “Best first strategy for feature selection,” 1988.

[17] J. Sheinvald, B. Dom and W. Niblack, “A modeling approach to feature selection,” Proceedings. 10th International Conference on Pattern Recognition, 1990.

[18] H. Almuallim and T. G. Dietterich, Learning With Many Irrelevant Features, vol. 91, Citeseer, pp. 547-552, 1991.

[19] H. Liul, H. Motoda and M. Dash, A monotonic measure for optimal feature selection, Springer, 1998, pp. 101-106.

[20] Y. Rui, T. S. Huang and S.-F. Chang, “Image retrieval: Current techniques, promising directions and open issues,” Journal of visual communication and image representation, vol. 10, no. 1, pp. 39-62, 1999.

[21] K. Kira and L. A. Rendell, “The feature selection problem: Traditional methods and a new algorithm,” vol. 2, pp. 129-134, 1992.

[22] H. Liu, H. Motoda and L. Yu, “Feature selection with selective sampling,” ICML, pp. 395-402, 2002.
[23] P. Pudil and J. Novoviçovà, “Novel methods for feature subset selection with respect to problem knowledge,” in Feature extraction, construction and selection, Springer, pp. 101-116. 1998.
[24] J. Friedman, T. Hastie and R. Tibshirani, The elements of statistical learning, Springer series in statistics New York, 2001.
[25] D. Koller and M. Sahami, Toward optimal feature selection, Stanford InfoLab, 1996.
[26] Liu, H., Motoda, H. (2012). Feature Selection for Knowledge Discovery and Data Mining. Springer Science & Business Media.
[27] M. A. Hall, Correlation-based feature selection of discrete and numeric class machine learning, University of Waikato, Department of Computer Science, 2000.
[28] M. Dash, “Feature selection via set cover,” in Proceedings 1997 IEEE Knowledge and Data Engineering Exchange Workshop, Chicago, 1997.
[29] J. Doak, “An evaluation of feature selection methods and their application to computer security,” in Techninal Report CSE-92-18, University of California, Department of Computer Science, 1992.
[30] I. Foroutan and J. Sklansky, “Feature selection for automatic classification of non-gaussian data,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 17, no. 2, pp. 187-198, 1987.
[31] M. Ichino and J. Sklansky, “Optimum feature selection by zero-one programming,” IEEE Transactions on Systems, Man, and Cybernetics, no. 5, pp. 737-746, 1984.
[32] P. A. D. a. J. Kittler, Introduction to statistical pattern recognition, London: GB, 1982.
[33] A. W. Moore and M. S. Lee, Efficient algorithms for minimizing cross validation error, Elsevier, 1994.
[34] P. M. Domingos, Why does bagging work? a bayesian account and its implications, Citeseer, 1997.
[35] Liu, H., & Setiono, R.. Feature selection and classification-a probabilistic wrapper approach, 1997.
[36] H. Vafaie and I. F. Imam, Feature selection methods: genetic algorithms vs. greedy-like search, 1994.
[37] D. J. Stracuzzi and P. E. Utgoff, “Randomized variable elimination,” Journal of Machine Learning Research, vol. 5, no. Oct, pp. 1331-1362, 2004.
[38] S. Das, wrappers and a boosting-based hybrid for feature selection, 2001.
[39] S. Loscalzo and L. a. D. C. Yu, Consensus group based stable feature selection, ACM, 2009.
[40] Dash, M., & Liu, H.. Feature selection for clustering. Springer. 2000
[41] M. Hossin and M. Sulaiman, “A review on evaluation metrics for data classification evaluations,” International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, p. 1, 2015.
[42] N. Karayiannis and A. N. Venetsanopoulos, “Artificial neural networks: learning algorithms, performance evaluation, and applications,” no. 209, p. , 2013.
[43] K. Zheng and X. Wang, “Feature selection method with joint maximal information entropy between features and class,” Pattern Recognition, vol. 77, pp. 20--29, 2018.
[44] M. V. Joshi, On evaluating performance of classifiers for rare classes, IEEE, 2002.
[45] M. S. Sorower, “A literature survey on algorithms for multi-label learning,” Oregon State University, Corvallis, pp. 1-25, 2010.
[46] M. Rahmaninia and P. Moradi, “OSFSMI: Online stream feature selection method based on mutual information,” Applied Soft Computing, pp. 733--746, 2018.
[47] H. Dong, T. Li, R. Ding and J. Sun, “A novel hybrid genetic algorithm with granular information for feature selection and optimization,” Applied Soft Computing, vol. 65, pp. 33--46, 2018.
[48] D. Paul, R. Su, M. Romain, V. Sébastien, V. Pierre and G. Isabelle, “Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier,” Computerized Medical Imaging and Graphics, vol. 60, pp. 42-49, 2017.
[49] R. Zhu, F. Dornaika and Y. Ruichek, “Learning a discriminant graph-based embedding with feature selection for image categorization,” Neural Networks, vol. 111, pp. 35--46, 2019.
[50] C. Yin, L. Ma and L. Feng, “A feature selection method for improved clonal algorithm towards intrusion detection,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 30, no. 5, p. 1659013, 2016.
[51] A. Ghosh, A. Datta and S. Ghosh, “Self-adaptive differential evolution for feature selection in hyperspectral image data,” Applied Soft Computing, vol. 13, no. 4, pp. 1969--1977, 2013.
[52] N. D. a. D. S. C. a. F. F. Cilia and A. S. di Freca, “A ranking-based feature selection approach for handwritten character recognition,” Pattern Recognition Letters, 2018.
[53] M. E. Aminanto, R. Choi, H. C. Tanuwidjaja, P. D. Yoo and K. Kim, “Deep abstraction and weighted feature selection for Wi-Fi impersonation detection,” IEEE Transactions on Information Forensics and Security, vol. 13, no. 3, pp. 621-636, 2017.
[54] S. Gu, R. Cheng and Y. Jin, “Feature selection for high-dimensional classification using a competitive swarm optimizer,” Soft Computing, vol. 22, no. 3, pp. 811-822, 2018.
[55] S. a. S. M. Sharmin, A. A. Ali, M. A. H. Khan and O. Chae, “Simultaneous feature selection and discretization based on mutual information,” {Pattern Recognition, vol. 91, pp. 162--174, 2019.