Journal of Cybernetics and Informatics

published by

Slovak Society for Cybernetics and Informatics

Volume 14, 2014

http://www.kasr.elf.stuba.sk/sski/casopis/

ISSN: 1336-4774

Journal of Cybernetics and Informatics 14 (2014)

http://www.kasr.elf.stuba.sk/sski/casopis

DATA DIMENSIONALITY REDUCTION BY GENETIC ALGORITHMS

M. Said Abdel Moteleb, Nermin K. Abdel Wahab and Essam W. Helmy

Electronics Research Institute (ERI), Dokki, Cairo, Egypt

Abstract

Pattern recognition generally requires that objects be described in terms of a set of measurable features. Multi-classification process can be significantly enhanced byselecting an optimal set of the features used as input for the training operation. The selectionand quality of the features representing each pattern have aconsiderable bearing on the success of subsequent pattern classification .The selection of such a subset willreduce the dimensionality of the data samples and eliminate the redundancy and ambiguity introduced by some attributes. Here, we present some approaches to both feature selection and feature extraction using some genetic algorithms.

Keywords: Pattern Recognition, Genetic Algorithm, Feature Selection

1 INTRODUCTION

The problem of classification in machine learning consists of using labelled examples to induce a model that classifies objects into a set of known classes. The objects are described by a vector of features, some of which may be irrelevant or redundant and may have a negative effect on the accuracy of the classifier. Potential benefits of reducing the data dimensions include: better modelling (classification/prediction) accuracy, simplification of the developed model, faster learning with fewer parameters, lower measurement costs, and improved reliability of parameter estimation [1]. Here we represent some approaches to feature subset selection: wrapper, filter methods, and by genetic algorithms. The remaining of this paper is organized as follows: first in Section 2 introduction on features selection is presented. In Section 3 an explanation of the filter approach is given. Then in Section 4, Wrapper approach is discussed. A theoretic background about genetic algorithms is in Section 5. The Section 6 summarizes some approaches to use GAs in feature selections. The paper is finally concluded with a summary of the most important points.

2 PROBLEM FORMULATION

A pattern can have a large number of measurable attributes, all of which may not be necessary for uniquely identifying it from other patterns. Thus, the selection of measurable attributes is a crucial step in pattern recognition system design. It has been proved that the reason for feature selection is “to curtail the effect of the ‘curse of dimensionality’ phenomenon on the complexity of the classifier”. Feature selection is the process of reducing input data dimension. By reducing dimensionality, feature selection attempts to solve two important problems: facilitate learning (inducing) accurate classifiers, and discover the most ”interesting” features, which may provide for better understanding of the problem itself [2].

For classical pattern recognition techniques, the patterns are generally represented as a vector of feature values. The selection of features can have a considerable impact on the effectiveness of the resulting classification algorithm [3],[4].

Consider a feature set, F = {f0, f1… fN}. If f0 and f1 are dependent, that is they always move together, then one of these could be discarded and the classifier has no less information to work with. This has the benefit that computational complexity is reduced as there is smaller number of inputs. Often, a secondary benefit found is that the accuracy of the classifier increases. This implies that the removed features were not adding any useful information but they were also actively hindering the recognition process [5]. The problem of feature selection can be seen as a case of feature weighting, where the numerical weights for each of the features have been replaced by binary values. A value of 1 could mean the inclusion of the corresponding feature into the subset, while a value of 0 could mean its absence. In a domain where objects are described by d features, there are 2d possible feature subsets. Obviously, searching exhaustively for the best subset (using any criteria to measure the quality) is futile. For this reason, the genetic algorithms has been identified as the best tools to explore such search space, and produce pseudo-optimal solutions that are sufficient to produce acceptable results[5].

Reducing the dimensionality of the vectors of features that describe each object presents several advantages, irrelevant or redundant features may affect negatively the accuracy of classification algorithms. In addition, reducing the number of features may help decrease the cost of acquiring data and might make the classification models easier to understand. There are numerous techniques for dimensionality reduction. Some common methods seek transformations of the original variables to lower dimensional spaces. For example, principal components analysis reduces the dimensions of the data by finding orthogonal linear combinations with the largest variance. In the mean square error sense, principal components analysis yields the optimal linear reduction of dimensionality. However, it is not necessarily true that the principal components that capture most of the variance are useful to discriminate among objects of different classes. Moreover, the linear combinations of variables make it difficult to interpret the effect of the original variables on class discrimination [6]. For these reasons, in the remainder of this paper we ignore methods that transform the features and we focus on techniques that select subsets of the original variables. There are two basic approaches to feature selection: filter and wrapper methods.

3 FILTER APPROACH

The filter approach [7] to feature selection tries to infer which features will work well for the classification algorithm by drawing conclusions from the observed distributions (histograms) of the individual features. However, the histograms give little insight into the separation between polyps and non-polyps. The correlation structure of the data is responsible for the success of the joint classifier, and a good classification scheme will attempt to utilize this structure. Although filter methods are much faster than wrappers, filters may produce disappointing results, because they completely ignore the induction algorithm [6].

4 WRAPPER METHOD

Wrapper feature selection [8] uses the method of classification itself to measure the importance of a feature or features set. The goal in this approach is maximizing the predicted classification accuracy. This approach, while more computationally expensive, tends to provide better results than the simpler filter methods.

When the major works related to features selection agree that wrapper mode give better results than filter one, this is not always true especially for very large datasets. Training a neural network or an SVM on 100000 samples for each chromosome during each generation of the genetic process became impracticable even on dedicated machines. This approach is useful when the number of training sample is limited according to the features one (the data space dimension) [5]. In both filter and wrapper methods the evaluation function is usually nonlinear and highly multimodal. Furthermore, the search space tends to be astronomically large resulting in a difficult optimization problem [1].

5 GENETIC ALGORITHMS

A. Introduction to GAs

Genetic Algorithms (GAs) are a family of computational models inspired by evolution. Computational studies of Darwinian evolution and natural selection have led to numerous models for computer optimization [9, 10]. GAs comprises a subset of these evolution-based optimization techniques focusing on the application of selection, mutation, and recombination to a population of competing problem solutions. GAs are parallel iterative optimizers, and has been successfully applied to many optimization problems, including pattern recognition and classification tasks. Being a directed search rather than an exhaustive search, population members cluster near good solutions; however, the GA's stochastic component does not rule out wildly different solutions, which may turn out to be better. This has the benefit that, given enough time and a well bounded problem, the algorithm can find a global optimum. This makes them well suited to feature selection problems (they can find near optimum solutions using little or no a priori knowledge) [5].

There are three major design decisions to consider when implementing a GA to solve a particular problem. A representation for candidate solutions must be chosen and encoded on the GA chromosome, an objective (fitness) function must be specified to evaluate the quality of each candidate solution, and finally the GA run parameters must be specified, including which genetic operators to use, such as crossover, mutation, selection, and their possibilities of occurrence[11].

The process of fitness-dependent selection and application of genetic operators to generate successive generations of individuals is repeated many times until a satisfactory solution is found. In practice, the performance of genetic algorithm depends on a number of factors including: the choice of genetic representation and operators, the fitness function, the details of the fitness-dependent selection procedure, and the various user-determined parameters such as population size, probability of application of different genetic operators, etc. [5].

For each GA experiment, the available data were broken into three disjoint sets: training, tuning, and testing. The training and tuning sets were used to train the classifier and provide tuning feedback to the GA. Once the GA run was completed, the

test set was used to perform unbiased testing on the best weight set found by the GA. The holdout testing was done using a variant of the bootstrap test technique [12]–[14].For each weight set w, 100 bootstrap tests were executed. For each bootstrap test, a random bootstrap set was selected from the holdout set using a uniform random distribution of samples with replacement. The weighted classifier was tested on this bootstrap set, and the accuracy for each class, as well as the total accuracy was computed. Finally, after the 100 bootstrap tests for a given weight set were completed, the performance of the weight set was evaluated according to mean bootstrap accuracy and variance of bootstrap accuracy.

B. Feature selection using GAs

Existing work in the field of pattern recognition explores the use of evolutionary algorithms for feature selection [15]–[17], and genetic algorithms are one type of evolutionary algorithms that can be used effectively as engines for solving the feature selection problem. The features selection using genetic algorithms has been studied and proven effective in conjunction with various classifiers, including k-nearest-neighbours, and neural networks [9, 16]. In [11], Yang and Hanovar investigated combinations of genetic algorithm and neural network. Eads et al. [18] and Sepulveda-Sanchez et al. [19] combined genetic algorithm and SVM. Liu and al. in [20] combined the parallel genetic algorithm with classification method proposed by Golub and al. In [21] a combination of SVM and GAs features selection is proposed for gene expression classification. Boudjeloud and Poulet [22] have used the Calinski index value as a fitness measure to evaluate the efficacy of each chromosome representing a dimensions combination. The same binary chromosomes representation is generally used. A binary string represents the set of all existing features, with a value of 1 at the i-th position if the i-th feature is selected, and 0 otherwise. The advantage of this representation is that a standard and well understood GA could be used without any modification. Unfortunately, the model of chromosome is only appropriate for data that have small and medium features. It caused an exponential nature of subsets that exist as the number of features increases. If the number of features is large, it becomes difficult to evaluate all possible combinations of features [5]. Raymer et al. [20] combined the linear transformation with explicit feature selection flags in the chromosomes, and reported an advantage over the pure transformation method. More sophisticated Distribution Estimation Algorithms (DEAs) have also been used to search for optimal feature subsets. DEAs explicitly identify the relationships among the variables of the problem by building a model of selected individuals and use this model to generate new solutions. However, in terms of accuracy, the DEAs do not seem to outperform simple GAs when searching for feature subsets [23,24]. Another idea proposed in [25] is the use of a measure of class separability to select features; it has been used generally in machine learning and computer vision.

The problem of dimensionality reduction is well suited to formulation as an optimization problem. Given a set of d –dimensional input patterns, the task of the GA is to find a transformed set of patterns in an m -dimensional space ($m<d$) that maximizes a set of optimization criteria. Typically, the transformed patterns are evaluated based upon both their dimensionality, and either class separation or the classification accuracy. The following figure shows the structure of a GA-based feature selector using classification accuracy as an evaluation criterion [11].

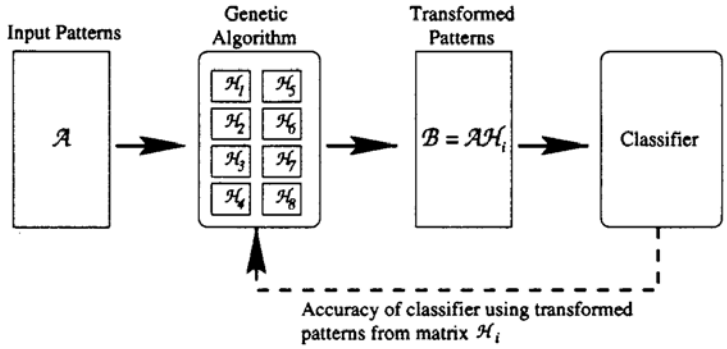

Fig.1. GA-based feature selection using an objective function based onclassification accuracy.

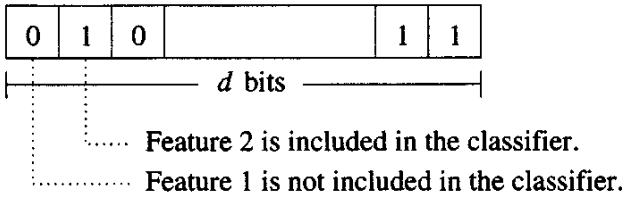

Fig.2. d-dimensional binary vector.

6 UNITS

Feature selection algorithms designed with different evaluation criteria broadly fall into three categories:

Filter model: The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm [32, 33, 30].

Wrapper model: The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than the filter model [35, 36].

Hybrid model: The hybrid model attempts to take advantage ofthe two models by exploiting their different evaluation criteria in different search stages. [37, 38].

7 CASE STUDY (MEMETIC ALGORITHMS FOR FEATURE SELECTION ON MICROARRARY)

In this case study, we present two novel memetic algorithms (MAs) for gene selection. Both are synergies of Genetic Algorithm (wrapper methods) and local search methods (filter methods) under a memetic framework. In particular, the rest MA is a Wrapper-Filter Feature Selection Algorithm (WFFSA) fine-tunes the population of genetic algorithm (GA) solutions by adding or deleting features based on univariate feature filter ranking method. The second MA approach, Markov Blanket-Embedded Genetic Algorithm (MBEGA), fine-tunes the population of solutions by adding relevant features, removing redundant and/or irrelevant features using Markov blanket.

BEGIN
1. **Initialize**: Randomly generate an initial population of feature subsets encodedwith binary string.
2. **While** (*not converged or computational budget is not exhausted*)
3. Evaluate fitness of all feature subsets in the population based on *J*(*Sc*).
4. Select the elite chromosome *cb*to undergo local search.
5. Replace *cb*with improved chromosome *c00b* using Lamarckian learning.
6. Perform evolutionary operators based on restrictive selection, crossover,and mutation.
7. **End While**

END

Fig.4. Memetic algorithm for Gene Selection.

The local search procedure proposed is a recipe of two heuristic operators, namely Add and Del. For a given selected gene subset encoded in chromosome, we define X and Y as the subsets of selected and excluded genes encoded inc, respectively. An Add operator inserts genes from Y into X, while a Del operator removes existing genes from X to Y. The important question is which gene to add or delete from a given chromosome that encodes potential gene subset. Here, we consider two possible scheme for adding or deleting genes in WFFSA and MBEGA.

1. Filter Ranking (WFFSA): All features are ranked using a filter method. In this study the Relief [34] is considered. Add operator selects a feature from Y using the linear ranking selection method described in [31], and moves it to X. Del selects a feature from X also using linear ranking selection and moves it to Y. The outline for Add and Del operators are provided in Figures 5 and 6, 7, respectively.

2. Markov Blanket [29] (MBEGA): Here, both the Add and Del operators select a feature from Y using also the linear ranking selection approach. However, MBEGA divers in the use of the C-correlation measure [30] instead of Relief F in WFFSA for ranking of features (see Figure 2 for the details).

Further for a given Xi, MBEGA proceeds to remove all other features in Xthat have been covered by Xi using the approximate Markov blanket3 [30]. If a feature Xjhas a Markov blanket given by Xi, this suggests that Xj gives no additional useful information beyond Xi on class C. Hence, Xj may be considered as redundant and could be safely removed. If there is no feature in the approximate Markov blanket of Xi, the operator then tries to delete Xi itself. The detailed procedure for Del operator is in Figs. 6, 7.

BEGIN
1. Rank the features in Y in descending order based on Relief in WFFSA while the*C-correlation* measure in MBEGA.
2. Select a feature *Yi* in Y using linear ranking selection [31] such that the higher thequality of a feature in Y, the more likely it will be selected to move to X.
3. Add *Yi* to X.

END

Fig.5. Add operator.

BEGIN
1. Rank the features in X in ascending order using ReliefF.
2. Select a feature *Xi* in X using linear ranking selection [31]such that the lower thequality of a feature in X, the more likely it will be selected to move to Y.

3. Remove *Xi* to Y.
END

Fig.6. Del operator in WFFSA.

BEGIN
1. Rank the features in X in descending order based on *C-correlation* measure.
2. Select a feature *Xi* in X using linear ranking selection [31] such that the higherthe *C-correlation* value of a feature in X, the more likely it will be selected.
3. Eliminate all features in X ¡ *fXig*which are in the approximate Markov blanket
Of*Xi*. If no feature is eliminated, remove *Xi* itself.
END

Fig.7. Del operator in MBEGA.

The results of the feature selection by GA, WFFSA, and MBEGA are tabulated in Table 1 below. Both the WFFSA and MBEGA outperform GA in terms of classification accuracy, showing lower test error rates in Table 1 than the latter. MBEGA obtains the lowest test error rates among all three methods. Both WFFSA and MBEGA also select more compact feature subset than GA.

Algorithm	GA	WFFSA	MBEGA
Test Error	0.0701	0.0250	**0.0202**
#Selected Groups	2.6	**3**	**3**
#Selected Genes	34.1	35.5	**9.7**
#Selected Relevant Genes	3.2	23.2	8.2
#Selected Redundant Genes	0.6	20.2	5.2
#Selected Irrelevant Genes	30.9	12.3	1.5

Table 1. Feature selection by each algorithm [28].

8 CONCLUSION

Feature selection is an important part of pattern recognition. With the help of feature selection process, the computation cost decreases and also the classification performance increases. The wrappers' evaluation of candidate feature subsets can be computationally expensive on large data sets. Filter methods are computationally efficient and offer an alternative to wrappers. Genetic algorithms have been used as filters in regression problems to optimize a cost function derived from the correlation matrix between the features and the target value. GAs has also been used as a filter in classification problems minimizing the inconsistencies present in subsets of the features. An inconsistency between two examples occurs if the examples match with respect to the feature subset considered, but their class labels disagree. Filter method efficiently identifies feature subsets that were at least as predictive as the original set of features (the results were never significantly worse). However, the accuracy on the reduced subset is not much different (better or worse) than with all the features. Many approaches for selecting best features subset using genetic algorithms are presented. The goal is to select the best combination that is sufficient to perform a good classification and obtain acceptable rates. This task cannot be realized with any iterative or exhaustive approach, so we have use an evolutionary genetic algorithm to explore the

huge space of all possible features subsets. In term of feature works, those approaches should be tested with other various datasets with different dimensions. The problem of dataset distribution must also be studied in more depth, as changing the proportionally of each class in the training dataset is shown to change radically the features selection and the classification results.

ACKNOWLEDGMENT

The authors thank the members of the High Institute of Engineering, Culture & Science City, 6th October City, Egypt for their support.

REFERENCES

[1] Alexander Topchy and William Punch, “Dimensionality Reduction via Genetic Value Clustering” E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 1431–1443, 2003.,

[2] Isabelle Guyon and Andr´eElisseeff. , (2003), An introduction to variable and feature selection.Journal of Machine Learning Research, 3:1157–1182.

[3] A. K. Jain and D. Zongker, “Feature selection: Evaluation, application,and small sample performance,” IEEE Trans. Pattern Anal. Machine Intell.,vol. 19, pp. 153–158, Feb. 1997.

[4] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, “Comparative study oftechniques for large-scale feature selection,” in Pattern Recognition inPractice IV, Multiple Paradigms, Comparative Studies and Hybrid Syst.,E. S. Gelsema and L. S. Kanal, Eds. Amsterdam: Elsevier, 1994, pp.403–413.

[5] K.M. Faraoun , A. Rabhi, “Data dimensionality reduction based on genetic selection of feature subsets”,2007.

[6] Erick Cantu-Paz, “Feature Subset Selection, Class Separability, andGenetic Algorithms

[7] G. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problems. In the 11thInternational Conference on Machine Learning,pages 121–129, 1994

[8] R. Kohavi, and G. John. Wrappers for feature subsetselection.Artificial Intelligence journal, volume 97.Special issue on relevance, pp 273-324. Dec. 1997

[9] D. Goldberg. Genetic Algorithms in Search,Optimization, and Machine Learning, AddisonWesley, 1989.

[10] J. Holland. Adaptation in Natural and ArtificialSystems, MIT Press, 1992.

[11] Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., Jain, A.K.: Dimensionality reduction using genetic algorithms. IEEE Transactions on Evolutionary Computation 4 (2000) pp:164-171

[12] A. K. Jain, R. C. Dubes, and C. C. Chen, “Bootstrap techniques for error estimation,” IEEE Trans. Pattern Anal. Machine Intell., vol. 9, pp. 628–633, Sept. 1987

[13] B. Efron, “Bootstrap methods: Another look at the jackknife,” Ann. Statist., vol. 7, pp. 1–26, 1979

[14] “The jackknife, the bootstrap, and other resampling plans,” CBMS–NSF Regional Conf. Ser. Appl. Math., SIAM’91, no. 38, 1982

[15] H. Handels, Th.Rob, J.Kreusch, H.Wolff, and S. Popple. Feature Selection for Optimized Skin Tumor Recognition using Genetic Algorithms. Artificial Intelligence in Medicine, 1999. pp:283-297

[16] F. Brill, D. Brown, and W. Martin, Fast Genetic Selection of Features for Neural Network Classifiers. IEEE Transactions on Neural Networks, 1992. pp:324-328
[17] H. Vafaie, Kenneth D.J., Feature Space Transformation Using Genetic Algorithms. IEEE Transactions on Intelligent Systems, 1998.pp57-65
[18] D. Eads, D. Hill, S. Davis, S. Perkins, J. Ma, R.Porter and J. Theiler, "Genetic algorithms andsupport vector machines for time seriesclassification", 5th Conference on the Applicationand Science of Neural Networks, Fuzzy Systems andEvolutionary Computation, pp. 74-85, 2002.
[19] J. Sepulveda-Sanchis, G. Camps-Valls, E.Soria-Olivas, S. Salcedo-Sanz, C. Bousono-Calzon, G. Sanz-Romero and J. Marrugat, "Support vectormachines and genetic algorithms for detectingunstable angina", Computers in Cardiology, IEEEComputer Society Press, Menphis, USA, 2002
[20] J. Liu, H. Iba and M. Ishizuka, "Selectinginformative genes with parallel genetic algorithms intissue classification", Genome Informatics, vol. 12,pp. 14-23, 2001.
[21] V. D. Nguyen and D. M. Rocke, "Tumorclassification by partial least squares usingmicroarray gene expression data", Bioinformatics,vol. 8, no. 1, pp. 39-50, 2002
[22] L. Boudjeloud and F. Poulet. A geneticapproach for outlier detection in high dimensionaldata sets. In Modelling, Computation andOptimization in Information Systems andManagement Sciences, MCO'04, pages 543–550. LeThi H.A., Pham D.T. Hermes Sciences Publishing,2004
[23] Inza, I., Larranaga, P., Etxeberria, R., Sierra, B.:Feature subset selection by Bayesian networks basedoptimization. Artificial Intelligence 123 (1999)pp:157-184
[24] Cantu-Paz, E.: Feature subset selection byestimation of distribution algorithms. In Langdon,W.B., Cantu-Paz, E., Mathias, K., Roy, R., Davis, D.,Poli, R., Balakrishnan, K., Honavar, V., Rudolph, G.,Wegener, J., Bull, L., Potter, M.A., Schultz, A.C.,Miller, J.F., Burke, E., Jonoska, N., eds.: GECCO2002: Proceedings of the Genetic and EvolutionaryComputation Conference, San Francisco, CA,Morgan Kaufmann Publishers, 2002, pp:303-310.
[25] Guyon, I., Elissee, A.: An introduction to variableand feature selection. Journal of Machine LearningResearch 3 (2003) pp: 1157-1182
[26] Ozdemir, M., Embrechts, M.J., Arciniegas, F., Breneman, C.M., Lockwood, L.,Bennett, K.P.: Feature selection for in-silico drug design using genetic algorithmsand neural networks. In: IEEE Mountain Workshop on Soft Computing in Indus- trial Applications, IEEE Press (2001) 53-57
[27] Lanzi, P.: Fast feature selection with genetic algorithms: a wrapper approach. In: IEEE International Conference on Evolutionary Computation, IEEE Press (1997) 537-540
[28] Memetic Algorithms For Feature Selection OnMicroarrayDataZexuan Zhu1;2 and Yew-Soon Ong1, 1 Division of Information Systems, School of Computer Engineering, NanyangTechnological University, Nanyang Avenue, Singapore 639798, 2 Bioinformatics Research Centre, Nanyang Technological University, Research TechnoPlaza, 50 Nanyang Drive, Singapore 637553
[29] D. Koller and M. Sahami, Toward optimal feature selection, In 13th InternationalConference on Machine Learning, Morgan Kaufmann, Bari, Italy, 1996.
[30] L. Yu and H. Liu, E±cient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research, 5,1205-1224,2004[3
[31] J. E. Baker, Adaptive Selection Methods for Genetic Algorithms, In Proc. Int'lConf. Genetic Algorithm and Their Applications, pp. 101-111, 1985.

[32] M. Dash, K. Choi, P. Scheuermann, and H. Liu.Feature selection for clustering – a filter solution. InProceedings of the Second International Conference onData Mining, pages 115–122, 2002
[33] M.A. Hall.Correlation-based feature selection fordiscrete and numeric class machine learning. InProceedings of the Seventeenth InternationalConference on Machine Learning, pages 359–366,2000
[34] M.Robnic-Sikonja and I. Kononenko, Theoretical and Empirical Analysis of ReliefFandRReliefF, Machine Learning, vol. 53, no. 1-2, pp. 23-69, 2003
[35] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 247–254, 2000
[36] Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 365–369, 2000
[37] S. Das. Filters, wrappers and a boosting-based hybrid
for feature selection. In Proceedings of the Eighteenth