Rubrique

Distributed Feature Selection modeling:

Impact of Hybridization and Distribution

Esseghir Mohamed Amir

Faculty of Sciences of Tunis
Computer Sciences Department
Tunisia
mohamedemir@gawab.com

RÉSUMÉ. Avec l´ expansion vertigineuse des systèmes dâinformation et des technologies de communication et de stockage associées, les techniques de classification sâavèrent de moins en moins adaptées aux nouvelles tailles et dimensions des données. Les techniques de sélection d'attributs permettent de reduire les la dimensions des données et contribuent à l'amélioration du processus de classification. La nature combinatoire du problème de sélection d'attributs rend les techniques de sélection classiques incapables d´ explorer efficacement les espaces de recherches. Dans ce papier, on propose une nouvelle technique de filtrage distribuées reposant sur le modèle dîlots.

ABSTRACT. The selection of relevant attributes for both classification and forecasting models has became more and more determinant for the majority of learning scheme as the means of data generation, collection, and storage increases in a spectacular manner. Feature Selection (FS) alternatives remain relying on sequential or and centralized approaches while real data set dimensions (attributes and instances) exceed from far the toy benchmark ones. The need to simultaneously tackle overwhelming problem search space and to guide the exploration of relevant subsets, have led us to investigate the distributed perspective for the FS problem. In this paper, we propose an approach which takes advantage of the distribution modeling to tackle FS diversification issues. We study the wrapper-filter hybridization on both centralized and distributed models of as well as the yielding performance improvement.

MOTS-CLÉS : Fouille de données, Optimisation combinatoire, Selection d'attributs

KEYWORDS : Data Mining, Combinatorial Optimization, Feature selection

1. Introduction

Researchers in machine learning, data mining and statistics have developed a number of methods for dimensionality reduction using usefulness/accuracy estimate for individual features or subsets. In fact, feature selection (FS) tries to select the most relevant attributes from row data, and hense guide the construction of the final classification model or decision support systems. From one hand, the majority, of learning scheme, are being relying on feature selection either as independent pre-processing technique or as an embedded stage within the leaning process [3]. On the other hand, both feature section and data mining technique struggle to gain attended reliability, especially when they face high dimensional data [7].

As a result, some trends in feature selection have attempted to tackle this challenge by proposing hybrid approaches or models based on multiple criteria [4, 7].

In this paper, we propose, a new distributed model for feature selection and an extension of it. The main motivations for this proposal are four folds : (i) centralized feature selection resolution has received great attention, whereas few dedicated distributed solution were proposed ; (ii) the combinatorial nature of the problem ramains an exciting challenge to design a distributed alternative models ;(iii) cooperative search space exploration and resolution between the involved strategies in the distributed model ; (iv) ability to tackle problems with high dimensional data.

The main contributions of this paper are the investigation of distributed FS modeling and the study of the impact of hybridization on distribution.

The remainder of this paper is divided into five sections. Section 2 formalizes the feature selection problem and reviews representative approaches. Section 3 describes our proposed distributed model and shows how the diversification will be designed within such a model. Section 4 compares and assesses the model's empirical results. Finally, Section 5 concludes this paper and presents some directions of future research.

2. Feature selection : basics and background

Let D be a data set with N as a set of attributes such that $\parallel N \parallel = n$, and let X $(X \subseteq N)$ be a subset of N. Let $J(X)$ the function that assesses the relevance of the subset X, and involves an evaluation criterion. The problem of feature selection states the selection of a subset Z such that :

$$J(Z) = max_{X \subseteq N} J(X) \quad [1]$$

In other words, the retained feature subset should be compact and representative of the dataset objects or the underlying context. This might be done by both remiving redundant noisy and/or irrelevant attributes by keeping the minimal information loss.

For a given dataset of n features, the exploration would require the examination of 2^n possible subsets. Consequently, the search through the feasible solutions search space is a *np*-hard combinatorial problem [7]. An exhaustive exploration of the feature space seems to be impractical, especially, when n became large.

Numerours reference approaches have been proposed for the identificationof features having the highest predictive power for a given target [6]. The represenative approaches could be divied in two classes : *filters* and *wrappers*.

2.1. Filters as univariate methods

Considered as the earliest approach to feature selection, filter methods discard irrelevant features, without any reference to a data mining technique, by applying independent search which is mainly based on intrinsic attribute properties and their relation with the data set class (*i.e*. Relief, Symetrical uncertainty, *etc*) [3, 4].

The main advantage of the filter methods is their reduced computational complexity which is due to the *simple* independent criterion used for feature evaluation. In most of the cases filetrs rank attributes according to a predefined criterion. Nevertheless, considering one feature at a time cripple the filter to handle with either redundant or interacting features. Such limitations have paved the way to the multivariate approaches (*i.e*. wrappers, embedded alternatives [3], *etc*) taking into consideration a subset of features in both search and evaluation.

2.2. Wrappers : multivariate methods

When feature selection is based on a wrapper, a subset of attributes are simutaniously evaluated using a classification algorithm. The exploration of such feasible solution requires a heuristic search strategy. The wrapper methods often provide better results than filter ones because they consider a classifier within the evaluation process. Kohavi *et al*. [6] were the first to advocate the wrapper as a general framework for feature selection in machine learning. Numerous studies have used the above framework with diffrent combinations for the evaluation (i.e. tree classifier, kernel, neural/bayesien classification scheme) and search components. Featured search technique are ranging from greedy sequential attribute selection methods (i.e. SFS, SBE, Floatting search) to randomized and stochastic methods (*i.e*. GRASP, TABU, BEAM, Genetic, ANT colony) [3, 7].

We should note that feature selection methods based on wrappers are computationally expensive compared to filters, due to the cost of iterative running of the classification algorithm [3]. The folowing section focus on some FS hybridization issues.

2.3. Hybrid approaches : trade-off or problem requirement ?

The motivation to a such orientation is the exhibited multidisciplinary problem property. The simplest form of recombination is to use both filters and wrappers. The common scheme of combination entails a couple of steps. The first one applies filter to reduce the search space, while the second step explores with a wrapper the subsets built from the yielded features returned by the first step[7].

A more sophisticated way of recombination is the use of memetic techniques [6, 7]. In this case, the local search evolves as a component of the whole evolutionary process. These methods which boost the search around solution neigborhood are being shown as promising alternatives in combinatorial optimization [1] and particularly in FS [7].

We should also note that the first attempts in distributed FS modeling was done with parallel scatter search [9] and parallel tabu search[7]. Both approaches, could be considered as distributed wrappers relying on local search.

Theses distribution alternatives motivated us to investigate *diversification*, *distribution* and *hybridization* of the existing approaches.

3. The proposed approach

The global modeling perspective which takes into consideration, not only a unique criterion, but several aspects of the FS problem (attribute interaction, search diversification, nesting effect[3]) might enhance research on feature selection. Recent approaches are confirming such orientation due to the fact that successful ones are based on hybrid approaches [7] In our case, the multi-disciplinary aspect of the FS problem will motivate our modeling consideration. The proposed approach is based on hybrid *evolutionary* wrapper-filter schema and designed as distributed *island model*.

3.1. Model distribution

Since the search effectiveness through the feasible feature subset solutions requires a robust exploratory process, we opt for an optimization scheme : *Genetic algorithms* [3, 7]. Besides, numerous distributed genetic approaches have been proposed for similar problems and combinatorial optimization [1] (*i.e* master-slave, cellular, hierarchical, and island parallel genetic algorithms). Each of which differs for the others by the workload distribution and the underlying collaboration protocol.

Our approach is endowed with distributed genetic algorithm based on island model [1, 10]. The distributed system is made of a set of islands. In such a modeling scheme, a set of genetic instances are distributed over the islands. Each one is evolving in an autonomous way by exploring different regions of the search space. At the same time, they are allowed to communicate by solution exchange.

This choice could be argued with two folds : (i) the multi population scheme could diversify the search, especially when we are handling high dimensional data ; (ii) distributed population could collaborate (solution migration) and exchange valuable informations (*i.e*. escaping local minima).

Indeed, the exploration within the distributed framework will be enhanced by both parallel exploration and collaboration through evolution.

Our island model assigns an instance of genetic algorithm (GA) to each island. The islands are organized in a ring structure. The collaboration between islands (solution exchange) depends on the location of islands. For a given island a migrant solution is sent to the next island and a new one is received from predecessor island. The folowing algorithm details island behaviour with basic evolutionary process and inter-island collaboration stages. The solution are represented by a binary chromosomes in which each bit state corresponds to the presence of the associated attribute in the considered set of features (*i.e* the i^{th} bit is set to 1 implies the selection of the i^{th} attribute in the proposed solution). The iterative genetic process entails evolution of the initial population in such a way that a part of the population will be renewed using a set of evolutionary operators. The regeneration starts with the selection of a sub-population. Here, we opt for a tournament selection [7] and this to keep an equilibrated selection pressure over the whole process. Once the solutions are selected, the reproduction operators will be applied using respectively one point crossover and the classical mutation operators. The resulting new solutions will replace some of the existing ones using a reverse tournament procedure [10] : less fitness solutions will be replaced by new high fittest ones. The evaluation procedure assesses the selected feature according to the classification accuracy on training set (classification error rate).

In order to reduce communication overhead only one solution is allowed to migrate, hence, the communication overhead complexity is in the order of $O(2 * m * n)$ where n is the number of island instances and m is the number of iterations where migration is allowed. Also, we opt for a random migrant elected by tournament

Input:
S : Initial solutions set ; Cla : Classifier ; $Collab_P$: Collaboration protocol parameters ; $Maxgen$: Total number of iterations ; GA_{params} : Initial GA parameter values ; D : Dataset
Output: *S'* : Population of the last generation

```
Begin
    Population P=S, P_tmp=∅ ; i=0
    While i< Maxgen do
        // Evolutionary process
        P_tmp=Select (P, GA_params)
        Crossover(P_tmp, GA_params)
        Mutate(P_tmp, GA_params)
        Evaluate(P_tmp, Cla, D)
        Replace(P_tmp, P, GA_params)
        //Collaboration stage
        M=SelectMigrantSolutions(P, Collab_P)
        Migrate(M, Collab_P)
        F=GetForeign(Collab_P)
        Integrate(F, P, Collab_P)
        i = i + 1
    Return (S'=P)
End
```

Algorithm 1. *Hybrid filter-wrapper instance*

3.2. Hybridization

Since centralized hybrid approaches are showing an acceptable trade off between wrapper and filters [7, 11], we chosen to endow the genetic wrapper with some of the filter capabilities within our distributed and collaborative environment.

In this paper, we opt for a filter that participates to the process generation of the initial population (*i.e*. filter relevance measure). The filter criterion is taken into consideration within the initialization step. As the filter yields an order based on one criterion, a part of the initial set of solutions could be generated with a combination with the relatively relevant features (according to the filter criterion).

4. Empirical study

The aim of this section is to evaluate and compare our distributed model on set of a well known benchmarks [2]. Both fitness assessment procedure and validation of the final solutions are based classification accuracy over separated data (generalization error rate). In the following sections, we will start by shedding some light on technical aspects in relation with model implementation and empirical setting.

4.1. Implementation and empirical settings

In this section, we summarize model implementation and experimental settings. Three data sets provided by the UCI repository [2] were user : Sonar (60 attributes), Audiology (69 attributes) and Arrhythmia (279 attributes). The corresponding search spaces are in-

tractable and prohibitive for exhaustive exploration ($2^{60} - 2^{279}$ solutions). The distributed model was implemented in java using both agent technologies and the classification algorithms provided by the Weka API [8]. JADE [5] was chosen as a Java agent framework for distributed implementation. For experiments enlisting both centralized or distributed GA the same parameter values' were kept for the different experimentation stages (*i.e.* crossover and mutation rate fixed at .6 and .2, selection and replacement technique using tournament, size of the mating pool 25% of population, solution migration and an replacement using the same selction operator. The validation procedure for the experiments are based on generalization error rates for the selected attributes. The search and the validation are done on a separate data. We used the Naive Bayes[7] classifier as a wrapper and for validation of the selected subsets.

4.2. Distribution analysis

	Reference Approaches						**Distributed Model**		
Data Set	**G.A.**	**Relief**	**I.G.**	**S.U.**	**Interact**	**FCBF**	**4 isl.**	**8 isl.**	**16 isl.**
Audiology	13.52	31.85	32.74	31.85	32.89	35.04	**11.62**	–	11.85
Arrhythmia	16.70	28.76	30.97	39.38	28.49	34.51	14.48	**14.33**	–
Sonar	22.33	24.03	33.65	33.65	33.25	28.65	19.84	20.65	**19.71**
Mean	17.51	28.21	32.45	34.96	31.54	32.73	**15.31**	**17.49**	**15.78**

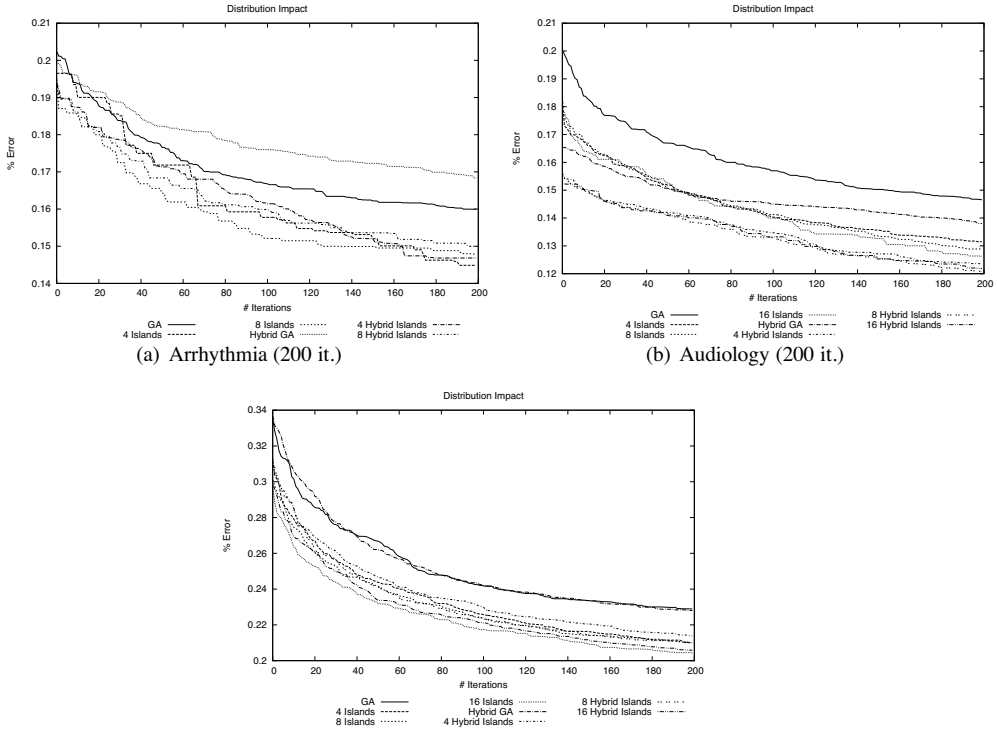
Tableau 1. *Test Error rates* $(\%)$ *for Centralized and Distributed FS approaches.*

For this set of experiments, we empirically assess the distributed model and we compare it to a set of relevant existing approaches as well as centralized genetic algorithm. A set of representative feature selection approaches were compared to a set of distributed instances of our island model. The number of islands is ranging from 4 to 16. Table 4.2 reports the results for six algorithms (i.e. Interact[7], Relief-F[3], Information-Gain(IG)[3], FCBF[11], centralized version of Genetic Algorithm (GA), and Symmetrical Uncertainty filter (SU)[3]) that were compared to our model. The island model instances clearly outperforms both centralized classical genetic wrapper and sequential approaches. The same result is confirmed by average values over the three detests.

4.3. Hybridization analysis

In this section, we detail in depth the assessment of hybridization aspects particularly when it is combined with distribution instances. The current set of experiments will focus on behavioral study of both distributed wrappers and distributed ones boosted with filters.

Figure 1 compares different configurations of the island model by varying the number of islands and the initialization scheme. They show the evolution of error rates (best solution) over the optimisation process. Three sets of instances were compared : genetic algorithm (reference model), distributed wrappers and distributed hybrid wrappers. The first comparison that could be made, is between centralized and distributed models. Indeed, we can clearly note the impact of distribution on performance improvement during the evolution and with final accuracies. Hence for all cases, centralized instances (even hybrid ones) are outperformed by distributed instances. Such results could be explained by both parallel exploration and island collaboration (migation). Furthermore, we can see



(a) Arrhythmia (200 it.) (b) Audiology (200 it.)

(c) Sonar (200 it.)

Figure 1. *Distributed and hybrid genetic instances*

that the hybridization, might also, improve results for both centralized (see Fig.1(b) and Fig.1(c)) and distributed models (see Fig.1(a)and Fig.1(b)).

Table 2 compares distributed hybrid wrappers and illustrates the effect of diffrent filters in a distributed context using four and eight hybrid islands. The generation of the initial population was generated using the following filters ($F1 = \chi_2$, F2=Information Gain and F3=Relief) [3, 7]. Firstly, we can compare results of feature selection approaches and results of classification accuracy with the whole features set. Both models improve results by the selection of the relevant predictors. Also, higher distribution level provide better performance. The best generalization accuracy and subset reduction were provided with 8 islands. The impact of filter initialization on the final results depends on the nature of involved filter. Best results confirm the slight superiority of the hybrid models based on the *Information Gain filter*.

5. Conclusion and perspectives

In this paper, we investigate distributed and hybrid modeling for the feature selection problem. Three features characterize this proposed model : (i) it's distributed ; (ii) it uses many techniques : filters, wrappers and hybrid approaches ; (iii) it takes the advantage of distribution to diversify both search space exploration and selection criterion. Through this distributed model, we studied hybridization impact. Empirical assessment shed lights

		4 Distributed Islands			8 Distributed Islands			Whole set
Data Set		**F1**	**F2**	**F3**	**F1**	**F2**	**F3**	**No FS**
Audiology	$\%\overline{Err}$	11.87	13.78	11.81	11.53	13.14	**11.15**	19.46
	$\#\overline{f}$	27	34	24	26	**23**	25	69
Arrhythmia	$\%\overline{Err}$	14.70	14.81	15.48	15.24	**14.63**	15.32	21.16
	$\#\overline{f}$	126	158	**103**	140	124	141	279
Sonar	$\%\overline{Err}$	21.26	20.71	20.98	21.52	**20.17**	20.89	37.87
	$\#\overline{f}$	29	27	31	**26**	31	28	60

Tableau 2. *Hybrid feature selection effectiveness (F1 = χ_2, F2 = Info. Gain and F3 = Relief)*

on some behavioral aspects of collaborating filters and wrappers. In the future, we will study the scalability of our distributed model. Another issue concerns stability of both search space exploration and solution quality.

6. Bibliographie

[1] E. ALBA AND M. TOMASSINI , « Parallelism and evolutionary algorithms », *IEEE Transactions on, Evolutionary Computation*, vol. 6, p. 443- 462, 2002.

[2] C. BLAKE AND C. MERZ, « UCI repository of machine learning databases », *http :/www.ics.uci.edu/ mlearn/MLRepository.html* .

[3] I. GUYON AND A. ELISSEFF, « An Introduction to Variable and Feature Selection », *Journal of Machine Learning Research* , vol. 3 , p. 1157-1182, 2003 .

[4] I. GUYON AND S. GUNN, M. NIKRAVESH AND L. ZADEH, « Feature Extraction, Foundations and Applications », *Studies in Fuzziness and Soft Computing, Series Springer*, 2006.

[5] JADE, « Java Agent DEvelopment framework. », *http ://jade.cselt.it/* .

[6] R. KOHAVI AND G. H. JOHN, « Wrappers for feature subset selection », *Artificial Intelligence*, vol. 97, p. 273–324, 1997.

[7] H. LIU AND H. MOTODA, « Computational methods of feature selection », *Chapman and Hall/CRC Editions* , 2008.

[8] WEKA MACHINE LEARNING PROJECT, « URL http ://www.cs.waikato.ac.nz/˜ml/weka », *University of Waikato*.

[9] F.G. LOPEZ AND M. G. TORRES, B. M. BATISTA, J. A. M. PEREZ AND J. M. MORENO-VEGA, « Solving feature subset selection problem by a Parallel Scatter Search », European Journal of Operational Research , vol. 2 , p. 477-489, 2006.

[10] D. E. GOLDBERG, « Genetic algorithms in search, optimization and machine learning », *Addison Wesley*, 1989.

[11] L. YU AND H. LIU, « Feature Selection for High-Dimensional Data : A Fast Correlation-Based Filter Solution », *Twentieth International Conference on Machine Leaning (ICML-03)*, p. 856-863, 2003.