ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

A Robust Feature Subsection Selection Algorithm Using GRASP

S.Kayasandigai, S.Sowmyadevi

M.E Student, Department of CSE, Arunai Engineering College, Tiruvannamalai, India

Asst. Professor, Department of CSE, Arunai Engineering College, Tiruvannamalai, India

ABSTRACT: Feature selection in high-dimensional records is one of the active areas of research in model recognition. In view of the substantial number of surviving feature selection algorithms.In this paper, a new method for feature selection algorithm in high-dimensional can control the trade-off between correctness and classification time. This scheme is found on a greedy metaheuristic algorithm called greedy randomized adaptive search procedure (GRASP).It uses an wide-ranging version of a simulated annealing (SA) algorithm for local search. In this version of SA, new parameters are fixed that allow the algorithm to control the trade-off between correctness and classification time. New results show domination of the proposed method over previous versions of GRASP for feature selection. Also, they show how the trade-off between correctness and classification time is appropriate by the bounds introduced in the proposed method.

KEYWORDS: Feature selection, irrelevance, redundancy, high dimensionality ,robust.

I. INTRODUCTION

Data mining is the process of discovering knowledge patterns from large amounts of data. The data sources can include databases, data warehouses, other information repositories, or data that are streamed into the system dynamically. Data mining is also called as knowledge discovery and data mining (KDD).Data mining is extraction of useful patterns from data sources.

Patterns must be valid, novel, potentially useful and understandable. Data Mining functionalities specify the patterns to be found in data mining everyday jobs.Data Mining tasks can be classify into two category descriptive and predictive. Descriptive: describe general properties of data in the database .Predictive: perform deduction on data to make predictions Data can be associated with classes or concepts that can be described and yet precise, terms.This discourse of a concept or class are called class/concept descriptions. These descriptions can be derived via Data Characterization and Data Discrimination.

Frequent patterns are the patterns that occur frequently in the data. Patterns includes itemsets, sequences and subsequences. Frequent itemset refers to a set of items that often appear together in a transactional data set. ex: bread and milk. Classification is the process of finding function that describes data classes or concepts. This model is derived based on the analysis of a set of training data and is used to predict the class label of objects for which the class label is unknown. Clustering analyzes data objects without consulting class labels. Clustering is used to generate class labels for a group of data which did not exist at the beginning. Objects are clustered or grouped on the law of maximizing the intra-class similarity and minimizing the inter-class similarity.

II. RELATED WORK

In addition, in many statistics analysis tasks, multiple interacting features are ignored assuming independency between features or considering only pairwise features interaction address the problem of considering multiple interacting features in high-dimensional statistics sets. Their occupation was inspired by studies on hypergraph clustering to evaluate feature subdivisions.

They proposed a new evaluation measure based on information theory called multidimensional interaction information that determines significance of different conditional feature subdivisions with respect to the decision feature. In the case of statistics sets with a combination of insignificant and numerical features, Michalak et al. prospect a new feature

DOI: 10.15680/IJIRCCE.2016. 0402065

similarity measure based on the probabilistic dependence between features. A considerable number of the reported works focus on efficient search in feature subdivisions space. Each point in feature subdivisions space is a subsection of features. Ahila, Sadasivam, and Manimala ,proposed an evolutionary algorithm based on particle swarm optimization (PSO) to perform simultaneous feature and model selection. They used a probabilistic neural network as the classifier and PSO as the searching algorithm to explore feature subsection space and model parameters. Bermejo, Gámez, and Puerta proposed a new search method that reduces the number of package estimates by iteratively moving between filter and wrapper estimates. Their scheme is based on the Greedy Randomized Adaptive Search Procedure (GRASP) metaheuristic algorithm. GRASP is a two-step iterative algorithm in which in each repitition a elucidation is initially created and is improved later.

In this algorithm, in the building step, evaluation is carried out using a lighter filter measure, while during the improvement step a more costly wrapper measure is utilize. Bermejo, Gámez, and Puerta proposed a new combinatorial method. Their algorithm iteratively alteration between filter and wrapper evaluators. It uses a filter evaluator for constructing a ranked list of the features and considers the first block of the ranked features as selected candidates for inclusion in the wrapper evaluator's selected features. The algorithm stops when there are no new feature candidates for inclusion in the wrapper evaluator's feature selection. Feature selection can be formulated as an optimization problem.

In this case, feature selection is a search procedure for finding a subsection of conditional features with the most relevancies to the object feature. This sector is devoted to reviewing the application of GRASP for feature selection in high-dimensional statistics. GRASP is a two-phase optimization algorithm introduced by Feo and Resende. Extensions of this algorithm and its applications can be found in their successive works (Feo and Resende; Festa and Resende; Resende and Ribeiro. The two phases of the GRASP include

● Construction phase: In this phase, a greedy randomized heuristic algorithm is used to construct a elucidation with which to begin. The progression starts from an empty set and increases the set by selecting items at random selected list of promising candidates.

● Improvement phase: Here, the algorithm improves the output of the construction phase using a local search algorithm. The constructed elucidation is fed to the local search algorithm as an initial point.

GRASP is a iterative algorithm. During each repitition, after the completion of these phases, both the constructed and the improved elucidations are added to nondominated elucidations. The nondominated elucidations set is a set of elucidations in which no item has a complete priority over any other one. This is because in GRASP we are dealing with a multiobjective optimization difficulty.

III. SYSTEM ARCHITECTURE

A. Feature Selection

Given a statistics set with n features, there are $2n-1$ nonempty feature sub divisions, each with the potential to become the optimal subsection for representing that statistics set within the specific problem area. Subsection generator and subsection evaluator are main components of a feature selection method. Hence, feature selection methods can be distinguished from each other according to these components.

B. Construction Phase

In this phase, a greedy randomized heuristic algorithm is used to construct a elucidation with which to begin. The progression starts from an empty set and increases the set by selecting items from a randomly selected list of promising candidates.

C. Improvement phase

Here, the algorithm improves the output of the construction phase using a local search procedure. The constructed elucidation is fed to the local search algorithm as an initial point.

D. *Subset generator*

Subset generator is equivalent to a searching method that determines the sequence of subset evaluations. By definition, the best subset generator reaches to optimal feature subset as fast as possible. There is no unique method with optimal performance for all problem areas.

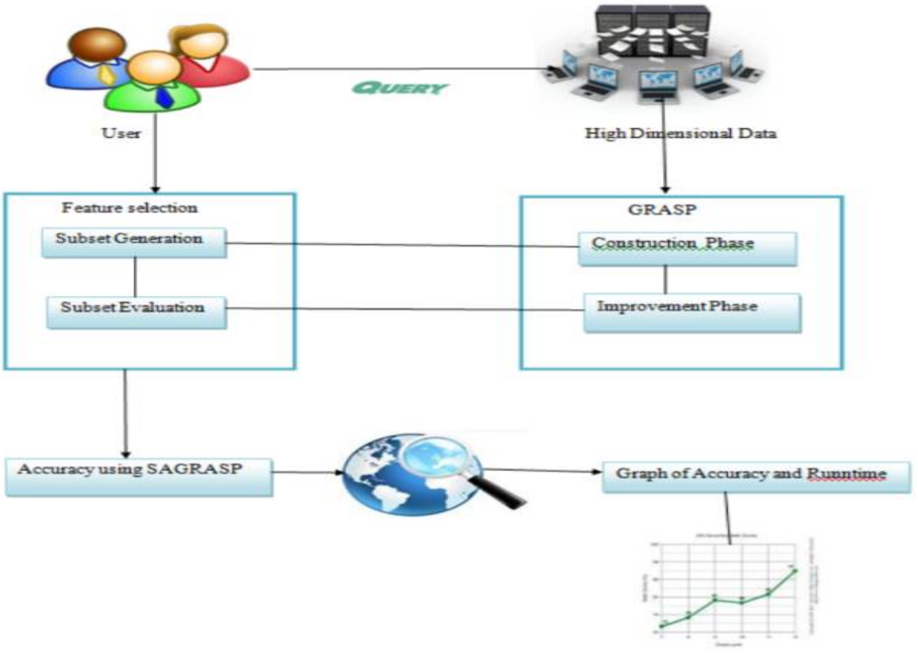


Fig 1. System Architecture

E. Subset evaluator

Subset evaluator is the second component of a feature selection method.It determines the merit of each feature subset. Ideally, it should assign the best merit to the optimal feature subset, which contains the most relevant features to the target feature and excludes irrelevant and redundant features as much as possible.

F. High-dimensional data

Clustering high-dimensional data is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. Such high-dimensional data spaces are often encountered in areas such as medicine, where DNA microarray technology can produce a large number of measurements at once, and the clustering of text documents, where, if a word-frequency vector is used, the number of dimensions equals the size of the vocabulary.

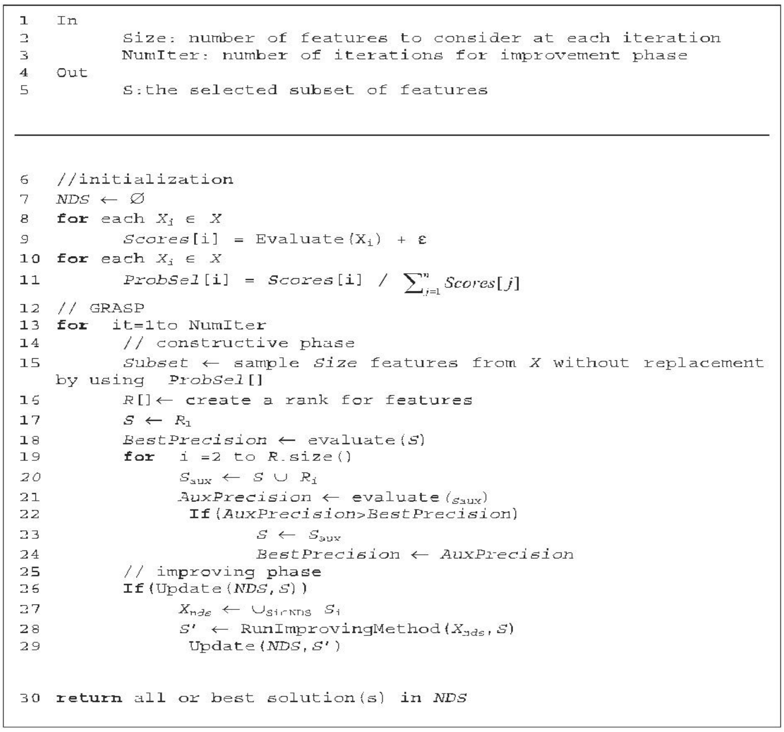
IV. PROPOSED ALGORITHM

GRASP is a iterative algorithm. Each iteration, after the completion of these phases, both the constructed and the improved solutions are added to non dominated solutions. The non subject solutions set is a set of solutions in which no item has a complete precedence over any other one. This is because in GRASP we are dealing with a multiobjective optimization difficulty.

Given a dataset with n features, there are $2n-1$ nonempty feature subsets,each with the potential to become the optimal subset for representing that dataset within the specific problem area. Subset generator and subset evaluator are main components of a feature selection method. Hence, feature selection methods can be distinguished from each other according to these components.

Filter measures are usually used when a rough evaluation is acceptable, and the wrapper evaluator is utilized when more precise evaluation is needed. Hybrid approaches are meant to establish an intermediate solution by combining the advantages of filters and wrappers and managing a trade-off between speed and precision.

DOI: 10.15680/IJIRCCE.2016. 0402065



```
In
        Size: number of features to consider at each iteration
        NumIter: number of iterations for improvement phase
Out
        S:the selected subset of features
-----------------------------------------------------------------
//initialization
$NDS \leftarrow \varnothing$
for each $X_i \in X$
        $Scores$[i] = Evaluate($X_i$) + $\varepsilon$
for each $X_i \in X$
        $ProbSel$[i] = $Scores$[i] / $\sum_{j=1}^{n} Scores[j]$
// GRASP
for  it=1to NumIter
        // constructive phase
        $Subset \leftarrow$ sample $Size$ features from $X$ without replacement
by using  $ProbSel$[]
        $R$[]$\leftarrow$ create a rank for features
        $S \leftarrow R_1$
        $BestPrecision \leftarrow$ evaluate($S$)
        for  i =2 to $R$.size()
              $S_{aux} \leftarrow S \cup R_i$
              $AuxPrecision \leftarrow$ evaluate($_{Saux}$)
               If($AuxPrecision$>$BestPrecision$)
                     $S \leftarrow S_{aux}$
                     $BestPrecision \leftarrow AuxPrecision$
        // improving phase
        If(Update($NDS$,$S$))
              $X_{nds} \leftarrow \cup_{S_i \in NDS} S_i$
              $S' \leftarrow$ RunImprovingMethod($X_{nds}$,$S$)
               Update($NDS$,$S'$)

return all or best solution(s) in $NDS$
```

GRASP ALGORITHM FOR FEATURE SELECTION (FCGRASP)

The output of each wrapper evaluation is the average accuracy obtained by 10 decision trees that trained independently Experiments are performed on 9 high-dimensional datasets selected from different areas such as cancer prediction based on mass spectrometry,and text classification. Also, they are different enough in different aspects such as number of instances, number of features, and number of classes. Their characteristics are listed in Table 1.

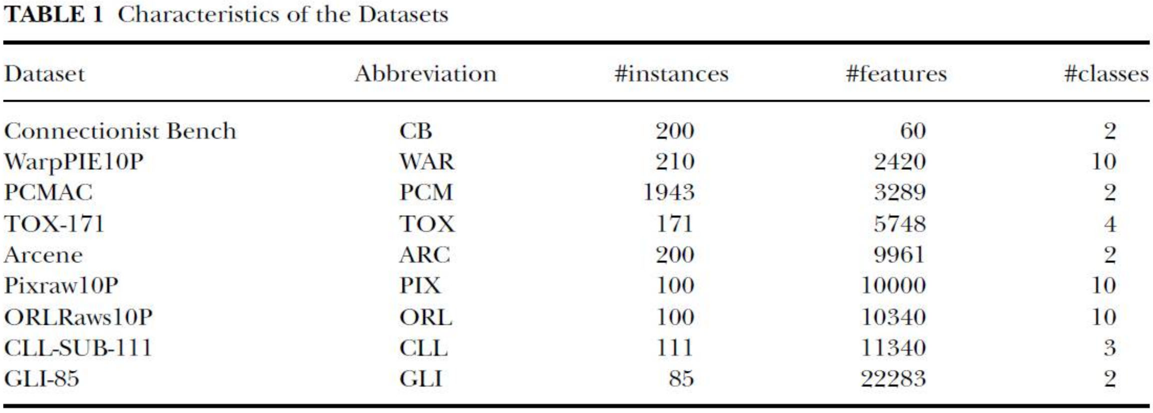


TABLE 1 Characteristics of the Datasets

Dataset	Abbreviation	#instances	#features	#classes
Connectionist Bench	CB	200	60	2
WarpPIE10P	WAR	210	2420	10
PCMAC	PCM	1943	3289	2
TOX-171	TOX	171	5748	4
Arcene	ARC	200	9961	2
Pixraw10P	PIX	100	10000	10
ORLRaws10P	ORL	100	10340	10
CLL-SUB-111	CLL	111	11340	3
GLI-85	GLI	85	22283	2

V. SIMULATION RESULTS

.

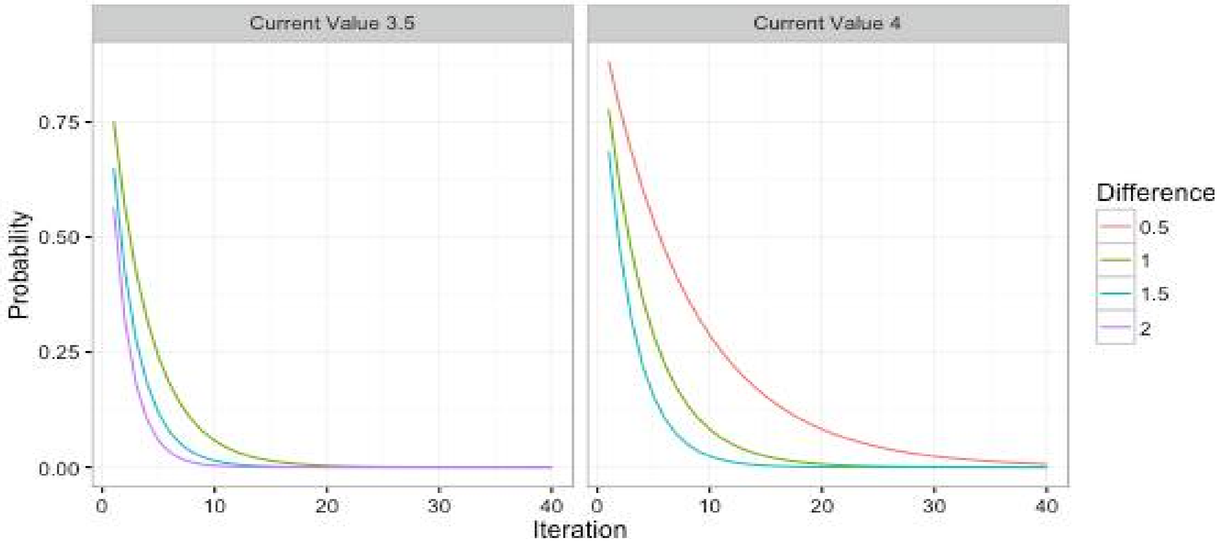
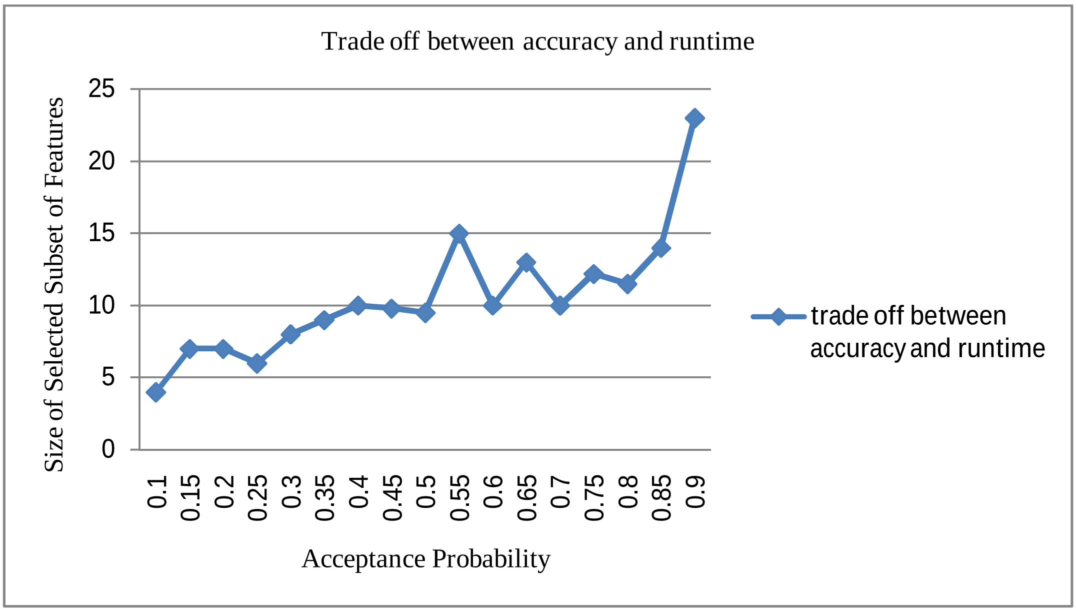


Fig .2 The acceptance probability against Iteration

Simulated annealing (SA) is a global search process that makes small random changes (i.e. perturbations) to an initial candidate solution.When the performance value for the disturbed value is better than the previous solution, the fresh solution is accepted. If not, an acceptance probability is determined based on the difference between the two performance values and the present iteration of the search. From this, a best solution can be accepted on the off-change that it may eventually produce a better solution in subsequent iterations.



.

Fig .3 Effect of acceptance probability on the size of the selected subset of features

VI. CONCLUSION AND FUTURE WORK

In this paper a new feature selection algorithm for high-dimensional datasets, named SAGRASP, was proposed. It uses the GRASP algorithm as its basic frame t with an extended version of SA as a local search algorithm. The proposed algorithm has two advantages over FSGRASP: one is better accuracy and another is the controllability. SAGRASP outperforms FSGRASP significantly over two datasets. It is having *P_add* parameter to control the trade-off between accuracy and classification time. Although the proposed method can implicitly control the trade-off between accuracy and classification time, although its not capable to control explicitly. In other words, it is possible to increase or decrease the number of special features by changing *P_add* parameter, but it is not clear how *P_add* must be set in order to obtain a specific number of features. An unlock line of research is to extend the proposed method in order to perform the control explicitly.

REFERENCES

1. Akand, E., M. Bain, and M. Temple 2010..”Learning with gene ontology annotation using feature selection and construction”. Applied Artificial Intelligence 24(1–2):5–38,.
2. Estévez, P. A., M. Tesmer, C. A. Perez, and J. M. Zurada. 2009. “Normalized mutual information feature selection”. IEEE Transactions on Neural Networks 20(2):189–201.
3. Liu, H., and L. Yu. 2005. “Toward integrating feature selection algorithms for classification and clustering”. IEEE Transactions on Knowledge and Data Engineering 17(4):491–502.
4. Peng, H., F. Long, and C. Ding. 2005. “Feature selection based on mutual information: criteria of maxdependency, max-relevance, and min-redundancy”. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8):1226–1238.
5. Ratanamahatana, C. A., and D. Gunopulos. 2003. “Feature selection for the naive Bayesian classifier using decision trees”. Applied Artificial Intelligence 17(5–6):475–487.
6. Ruiz, R., J. C. Riquelme, J. S. Aguilar-Ruiz, and M. García-Torres. 2012. “Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches”. Expert Systems with Applications 39(12):11094–11102.
7. Waad, B., B. M. Ghazi, and L. Mohamed. 2013. “A three-stage feature selection using quadratic programming for credit scoring”. Applied Artificial Intelligence 27(8):721–742.
8. Zhang, Z., and E. R. Hancock. 2012. “Hypergraph based information-theoretic feature selection”. Pattern Recognition Letters 33(15):1991–1999.