A Novel GA-based Feature Selection Approach for High Dimensional Data

Claudio De Stefano, Francesco Fontanella, Alessandra Scotto di Freca
Universitá di Cassino e del Lazio meridionale
{destefano,fontanella,a.scotto}@unicas.it

ABSTRACT

In this paper we present a novel GA-based approach for feature selection in high dimensional spaces. The proposed system is able to greatly reduce the number of features to be used in the classification phase and can deal with problems involving thousands of features. The system is based on two modules. The first module employs a feature ranking method to reduce the number of features to be taken into account. The second module uses a GA-based search strategy that uses a filter fitness function for finding feature subsets with a high discriminative power.

1. INTRODUCTION

Recent years have seen a strong growth of applications in which a huge number of features is available. In this cases, selecting the features with the most predictive power is a critical task.

Genetic Algorithms (GAs) have been widely used to solve feature selection problems as they have shown to be very effective in solving optimization problems whose search space are discontinuous and very complex[1, 2]. However, when thousands of features are involved, even for a GA it becomes very difficult to find good solutions.

Recently, in order to reduce the search space size, different strategies have been adopted for GA-based algorithms. In most of the cases wrapper fitness functions have been used and problems involving few hundreds of features are taken into account.

In this paper we present a two–module system that combines a feature ranking algorithm with a GA. The first module uses a fast feature ranking algorithm to reduce the number of features to be taken into account by the second module; it provides as output a given number M (a priori fixed) of features that are promising, according to the univariate measure used. The second GA–based module seeks, in the search space provided by the first module, the best feature subset by using a filter fitness function that evaluates feature subsets. The layout of the system is shown in Figure 1.

Because of the reduction performed by the feature ranking, this search space is much smaller than that made of all the possible subsets of the whole feature set, nonetheless this search space contains most of the ”promising areas”, i.e. those containing good and near–optimal solutions (subsets). In practice, the ”filtering” performed by the ranking module allows the second GA–based module to focus its search on these more promising areas. As concerns the univariate measures for the feature ranking, we used the Chi-square measure. As evaluation function for the GA module we adopted the Correlation based Feature Selection function (CFS). This function evaluates the merit of a subset by considering both the correlation between the class labels and the individual features and the inter-correlation among the selected features; The computational complexity of the CFS function is independent of the number of samples making up the training set used for the correlation estimation.

In order to assess the effectiveness of the proposed system, several experiments have been performed and the obtained results have been compared with those achieved by three different feature selection algorithms.

2. SYSTEM ARCHITECTURE

The proposed system is made of two modules. The first module sorts the whole set of features according to a univariate measure. Once the features have been sorted, the first M features are provided as input to the second module. Note that the value of M must be chosen by the user. For our approach, we used the Chi-square univariate measure. This measure estimates feature merit by using a discretization algorithm: if a feature can be discretized to a single value, it has not discriminative power and it can safely be discarded. The discretization algorithm adopts a supervised heuristic method based on the χ^2 statistic.

The second module of the system presented here has been implemented by using a generational GA. As fitness function for the GA we chose a filter one, called CFS (Correlation-

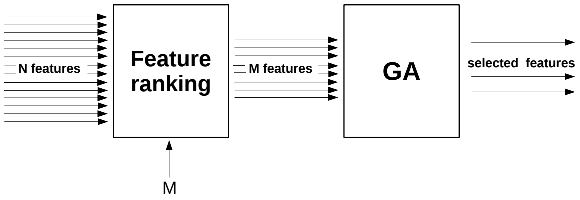

Figure 1: The layout of the proposed system.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

GECCO'16 Companion July 20-24, 2016, Denver, CO, USA

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4323-7/16/07.
DOI: http://dx.doi.org/10.1145/2908961.2909049

Table 1: Arcene and Gisette datasets.

Dataset	M	RNK-GA		GA		RNK	RNK-BF	
		RR	#feat	RR	#feat	RR	RR	#feat
Arcene	100	78.5	8.3	80.1	990	81.4	80.1	8
	200	83	12.5			86.4	82.9	11
	500	85.2	35			87.4*	82.7	18
	1000	88.1	110.7			87.2	83.7	34
	2000	**92.3**	465			79	85.7	53
Gisette	100	92.55	22.6	95.7*	370	94.68	92.8	31
	200	94.55	29.9			95.57	94.7	42
	500	**96.7**	41.9			94.93	95.6*	73
	1000	96.76	103.8			94.52	95.4	74
	2000	95.4	190.8			90.1	95.5	77

based Feature Selection), which uses a correlation based heuristic to evaluate feature subset worth. This function takes into account the usefulness of individual features for predicting class labels along with the level of inter-correlation among them. The idea behind this approach is that good subsets contain features highly correlated with the class and uncorrelated with each other. The CFS function allows the GA to discard irrelevant and redundant features. The former because they are poor in discriminating the different classes at the hand; the latter because they are highly correlated with one or more of the other features. Furthermore, this fitness function is able to automatically find the suitable number of features and does not need the setting of any parameter.

3. EXPERIMENTAL RESULTS

We tested the proposed approach on high dimensional data (ranging from 500 up to 10000 features). For each dataset, a set of values for the parameter M (see Figure 1) has been tested. For each value of M, 30 runs have been performed for the GA module. At the end of every run, the feature subset encoded by the individual with the best fitness, has been used to built a Multilayer Perceptron classifier (MLP in the following), trained by using the back propagation algorithm. The classification performances of the classifiers built have been obtained by using the 10-fold validation approach.

The proposed approach has been tested on the following, publicly available, datasets: *Arcene* (10000 features), *Gisette* (5000), *Madelon* (500) and *Ucihar* (561). In order to test the performances of our system, we compared its results with those obtained by three different feature selection approaches: (i) The feature ranking represented by the first module of our system (Figure 1): given the whole set of N features, it gives as output the best M feature, according to the chi-squared (*RNK* in the following); (ii) the GA used in the second module (Figure 1): given the whole set of N features, it searches for the best solution (subset) by using the GA algorithm; (*GA* in the following); (iii) the third approach instead, is quite similar to our approach but uses the best first technique as search strategy of the second module (*RNK-BF* in the following).

With the purpose of investigating how the value of the parameter M affects the performance of the presented system, we tested several M values. Since the approaches RNK and RNK-BF are deterministic, for each value of M, they generated a single feature subset. However, in order to perform a fair comparison with the proposed approach, for each subset generated, 30 MLP's have been trained with different, randomly generated, initial weights. The trained MLP's have been evaluated by using the 10-fold validation approach. The results reported in the following have been obtained averaging the performance of the 30 MLP's learned.

Comparison results are reported in Tables 1 and 2. In both tables the recognition rate (RR) and the number of selected features (#feat), are reported for each comparing method. In order to statistically validate the comparison results, we performed the non-parametric Wilcoxon rank-sum test ($\alpha = 0.05$) over 30 runs. The values in bold in the recognition rate columns highlight, for each dataset, the results which are significantly better with respect to the second best results (values starred in the table), according to the Wilcoxon test. When the best results do not present statistically significant differences, the best two are starred. Moreover, for each method, in the case that two or even more results do not present statistically significant difference, the result achieved with the minimum number of features has been considered. Finally, note that for our approach, we used the abbreviation *RNK-GA*.

From Table 1 it can be seen that the proposed approach achieves better performance for the datasets Arcene and Gisette. For the Arcene dataset a recognition rate of 92.3% has been obtained by using only 465 out of the 10000 features available. For the Gisette dataset, the proposed system achieved a recognition rate of 96.7%, selecting on average 41.9 features. The results just described seems to confirm that as the search space (exponentially) grows with M, the GA module of the proposed approach is able to locate new areas of the search space containing better solutions, which includes the new features progressively added.

As regards the Madelon and Ucihar datasets, from Table 2 it can be observed that for both datasets the proposed system did not significantly outperform the compared systems.

From the results shown above it can be seen that the univariate measure used in the first module is able to identify most of the relevant features even though the adopted measure evaluates the relevance of each feature, without taking into account any feature interaction.

4. REFERENCES

[1] L. Cordella, C. De Stefano, F. Fontanella, and C. Marrocco. A feature selection algorithm for handwritten character recognition. In *ICPR 2008*, pages 128–131. IEEE Computer Society, 2008.

[2] L. P. Cordella, C. D. Stefano, F. Fontanella, C. Marrocco, and A. S. di Freca. Combining single class features for improving performance of a two stage classifier. In *ICPR 2010*, pages 4352–4355. IEEE Computer Society, 2010.

Table 2: Madelon and Ucihar datasets.

Dataset	M	RNK-GA		GA		RNK	RNK-BF	
		RR	#feat	RR	#feat	RR	RR	#feat
Madelon	20	76.11	9	76.4*	10.7	76.25*	76.14	9
	50	76.17	9			68.82	75.88	9
	100	76.01	8.9			63.97	76.15	9
	200	76.3*	8.7			62.01	76.09	9
	300	76.41	9.1			59.8	76.21*	9
Ucihar	20	81.65	7	95.67*	74.5	86.63	81.83	7
	50	93.85	19.0			94.07	93.8	18
	100	94.21	22.7			93.78	93.7	20
	200	95.51*	27.7			91.9	95.26*	21
	300	94.99	29.6			88.56	93.97	21