A PCA and SPCA based procedure to variable selection in agriculture

Juscelino Izidoro de Oliveira Jr ¹
Alaine Margarete Guimarães ²
Adriel Ferreira da Fonseca ³
José Carlos Ferreira da Rocha ⁴

Resumo: A mineração de dados agrícolas, frequentemente, envolve o processamento de bases de dados com poucas observações e alta dimensionalidade. Como a complexidade da amostra cresce com a dimensionalidade dos dadoss esses dois fatores podem limitar a confiança nos resultados obtidos ou produzir modelos em que há *overfitting*. Uma forma de reduzir a dimensionalidade dos dados e a complexidade da amostra é selecionar os atributos que são relevantes para a descrição do fenômeno de interesse. Este trabalho apresenta um procedimento que combina métodos de busca e análise de componentes principais supervisionada e não supervisionada para selecionar variáveis. O procedimento remove as variáveis irrelevantes ou com pouca influência sobre a variação dos dados e avalia o impacto da seleção sobre tarefas de regressão e classificação. Sempre que possível, o número de variáveis selecionadas é aquele que atende aos requerimentos da complexidade da amostra. O procedimento foi testado na seleção de variáveis para indução de modelos lineares multivariados e redes neurais artificiais mediante uma base de dados de agricultura de precisão. O procedimento proposto permite uma solução de custo-benefício entre a redução da dimensionalidade e a acurácia do modelo.

Palavras-chave: Análise de componentes principais. Mineração de dados. Seleção de atributos.

Abstract: Agricultural data mining often demands processing data bases with few observations and high dimensionality. As sample complexity grows up with data dimensionality these factors can constraint the confidence on obtained results as also overfitting. An approach for reducing data dimensionality and sample complexity is to select the attributes that are relevant to describe the phenomenon of interest. This work presents a procedure that combines search methods and supervised and unsupervised principal component analysis to select variables. The procedure removes irrelevant variables or variables with small influence on data variation and evaluates the impact of variable selection on tasks of regression and classification. Whenever possible, the number of selected variables attends the sample complexity requirements. The procedure was tested to select variables for the regression of multivariate linear models and artificial neural network classifiers training on a precision agriculture data set. The proposed procedure supports the trades-off between the dimensionality reduction and model accuracy.

Keywords: Attribute selection. Data mining. Principal component analysis.

```
<sup>1</sup>Fundação ABC - PR 151 - Castro (PR) - Brasil. {31011015@uepg.br} - Thanks to CAPES for
```

http://dx.doi.org/10.5335/rbca.2015.3727

^{{31011015@}uepg.br} - Thanks to CAPES for partial financial support to this research.

²UEPG - DEINFO/Infoagro - Campus de Uvaranas - Ponta Grossa (PR) - Brasil.

[{]alainemg@uepg.br}

 $^{^3 \}mbox{UEPG}$ - DESOLO/LNP - $\it Campus$ de Uvaranas - Ponta Grossa (PR) - Brasil.

[{]adriel@uepg.br} - Thanks to CNPq for partial financial support to this research.

⁴UEPG - DEINFO/LCA- Campus de Uvaranas - Ponta Grossa (PR) - Brasil.

[{]jrocha@uepg.br}

1 Introduction

Mathematical modeling is an important tool to capture the interactions among environment, soil and plants and then to get a better understanding about the physiological processes [25]. It usually demands developing multivariate models that relate a set of variables, describing physical and chemical features of plants and soil involved in a study. An important step to proceed the multivariate modeling on agricultural data is selecting the most relevant variables related to the response variable [24, 27]. A suitable variable set allows [11, 14, 12]: (i) to discard redundant attributes or the ones that add noise to the data; (ii) to reduce the risk of overfitting; (iii) to reduce sample complexity; (iv) to make the model simpler; (v) to save time and resources in future data collection.

A point related with the item (iii) is that the confidence on data mining results depends on the size of sample used to learn a pattern/function [7]. From this follows that to know how many observations a procedure needs for inducing a model is a very important practical issue. The computer theorists named such automated learning aspect as sample complexity [15], which establishes the smaller number of cases a learning algorithm requires to learn a concept from data, given a prespecified error limit of its previsions.

Whereas the sample complexity grows with the data dimensionality [8], variable selection procedures could make easier the requirements on the sample size. Basically, if the data analyst removes low relevance variables before learning the model, it would be possible to learn a better model from a smaller sample [18]. Such approach may be useful in domains where is common to have data sets with high dimensionality and few cases. An usual situation in domains as agricultural, medical and biological research [19].

The principal component analysis (PCA) has been a useful tool for variable selection [17] [20] [23] [26]. However, PCA is an unsupervised procedure. That is, variable selection with PCA does not take into account the influence of any input variable on the output one to decide if the former should be discarded or not. To overcome this limitation the supervised principal components analysis (SPCA) proposed by Bair [1] provides an approach to integrate supervised data in principal component analysis. The basic idea in SPCA is applying a filter to preselect attributes that have a stronger association with the response variable and then to use PCA to remove input attributes which are correlated between themselves.

Joliffe [9] presented two PCA based procedures for variable selection. Those methods, called B2 and B4, do not use supervised information to execute that task. Considering it, this work presents a procedure for variable selection that combines B2 and B4 methods with supervised PCA to select attributes for inducing regressors or classifiers. Additionally, the proposed procedure provides information that can be used to trading-off between the sample complexity and model accuracy. Experiments on agricultural data sets allows to evaluate the effectiveness of the approach.

The paper is organized as follows: Section 2 presents the basic theory used in this work; Section 3 describes the materials and methods used to perform the experiment; Section 4 presents the experimental results; Section 5 shows a discussion about the results and Section 6 presents the final remarks.

2 Background review

2.1 PCA-based variable selection

The purpose of the variable selection method is to choose the most relevant variables from a data set to proceed data mining tasks. This approach, also called feature selection, attribute selection and dimensionality

reduction, has been performed with a set of techniques. The PCA procedure provides support information to implement some of them.

Principal component analysis is a procedure for dimensionality reduction of a data set with many interrelated variables [11] [3]. The basic idea is to apply a linear transformation in order to highlight the shape of the data variation. This linear transformation defines a new space where the main axes are called principal components. The components are ordered so that the first components are related with the direction of the largest variation from the data set. More formally, let \mathbf{D} be a data set with multivariate observations and let \mathbf{Y} be a matrix obtained from \mathbf{D} by subtracting the mean of every variable in the original data set. Here, \mathbf{Y} is a matrix of $N \times d$ dimensions so that N is the number of observations and d is the number of variables. Let \mathbf{S} be the sample covariance matrix computed from \mathbf{Y} , $\mathbf{\Lambda}$, the diagonal matrix of the eigenvalues of \mathbf{S} , and, \mathbf{X} , the matrix of the eigenvectors corresponding to the eigenvalues in $\mathbf{\Lambda}$. Furthermore, the matrices are organized so that the first elements in the diagonal of the $\mathbf{\Lambda}$ matrix are the largest ones. From this follows that:

$$\mathbf{S} = \mathbf{X} \mathbf{\Lambda} \mathbf{X}^T \tag{1}$$

The dimensionality reduction is obtained by removing the higher ordered components (less informative). After this data analysis can be proceeded by observing the behavior of the data in the space defined by the first (main) principal components. Another approach is to evaluate the contribution (load) from the original variables on the main components and then to use some criteria to select or remove some of them.

Jolliffe [9, 10] presents two PCA-based variable selection procedures called B2 and B4. B2 and B4 procedures are described in Figures 1 and 2, respectively.

Figure 1: B2 procedure

- 1. execute PCA over the whole data set;
- 2. set an eigenvalue threshold to cut the variables off; this threshold is denoted by λ_0 ;
- 3. sort the principal components in ascending order by eigenvalue; let $u_1, ... u_d$ the ordered sequence of components;
- 4. for i = 1 to d:
 - determine the still non-associated highest loading variable X in u_i ;
 - associate variable X with u_i ;
- 5. remove the variables associated with those principal components which eigenvalues are less than λ_0 .

The main difference between B2 and B4 is that B2 rejects the variables that are strongly related with the components that represent the smallest part of the data variation. B4, otherwise, rejects the variables which are weakly related with the components which represent the most variation of data.

2.2 Supervised principal component analysis

The SPCA intends to concentrate the principal component analysis on input variables which have more influence on a response variable [1]. To that, SPCA pre-selects the input variables which have more influence on the output and then runs PCA on selected ones. The Figure 3 shows the SPCA procedure. L is a threshold on

- 1. execute PCA over the whole data set;
- 2. define an eigenvalue threshold (λ_1) to cut the variables off;
- 3. sort the principal components in descending order by eigenvalue; let $u_1, ... u_d$ the ordered sequence of components;
- 4. for i = 1 to d:
 - determine the still non-associated lowest loading variable X in u_i ;
 - associate variable X with u_i ;
- 5. remove variables associated with those principal components which eigenvalues are greater than λ_1 .

the regression coefficient and s is computed by the expression $s_j = \frac{X_j^T Y}{\sqrt{X_j^T X_j}}$. In this equation, X_j is the column matrix associated with the respective attribute and s_j is the standardized regression coefficient of X_j .

Figure 3: SBCA procedure

- 1. compute the standardized regression coefficient s for each variable;
- 2. create a reduced matrix from \mathbf{D} with only those variables whose absolute value of the standardized regression coefficient is greater than a threshold L (L is estimated by cross validation);
- 3. compute the first (or first few) principal components from the reduced data set;
- 4. use these principal components in a regression model to predict the outcome.

2.3 Regression and classification models

Consider a data set with N observations on a response variable Y and p predictor variables $X_1, X_2, ..., X_d$. The goal of the multiple linear regression (MLR) is to determine a relationship between Y and X in the form [2]:

$$Y = \beta_0 + \sum_{i=1}^{d} \beta_i * X_i + \varepsilon \tag{2}$$

In Equation 2, β_0 is the intercept, β_i , i = 0..d, are the linear coefficients and ε is a random disturbance or error.

Artificial neural network (ANN) is a mathematical model formed by a set of interconnected processing units, called artificial neurons [7]. Each unit u calculates its value using an activation function that integrates messages that the units connected to u send to it. The connections among the units are weighted. ANNs work by getting an input signal and propagating it forward, layer by layer, until it arrives to output layer. When the signal achieves the output layer the activation functions of its units are calculated and the results are related to the response variable and interpreted by the user. A multilayer neural network typically has one layer of input units, one or more intermediate layers, and one layer with output units. The ANNs weights are trained from data sets

using specialized algorithms.

2.4 Sample complexity

In the computer learning theory, the number of examples/cases required by a learning algorithm to learn a concept/pattern is named sample complexity [15]. In this context, the probably approximately correct (PAC) learning framework provides results to calculate bounds on the sample complexity for a given learning task. Summarily, the sample size, m, that allows an algorithm to learn a concept with probability between 0 and δ having error between 0 and ε is:

$$m \ge \frac{1}{\varepsilon} \cdot \left(4 \cdot log\left(\frac{2}{\delta}\right) + 8 \cdot VC(H) \cdot log\left(\frac{13}{\varepsilon}\right) \right)$$
 (3)

In the Inequation 3, VC is the VC dimension of the learning problem [6]. For linear regression the VC dimension is given by VC(H)=d+1, where d is the number of variables (dimensions of the data) [22]. For neural networks, the VC dimension is defined as:

$$VC(H) = 4q (r+1) \log(e q).$$

In this equation, q is the number of nodes in the hidden layers, r is the number of inputs for each hidden node and e is the base of natural logarithm.

3 Materials and methods

Since this work intends to explore procedures for selecting variables on supervised data sets, particularly with respect to the relevance of the selected variables and the reduction on sample complexity demands, experiments were done with standard and supervised PCA. The data set used was obtained from an experiment on precision agriculture carried out in Campos Novos Paulista (Brazil). The data set contains 2416 samples related to the yield of maize (*Zea mays* L.) and the following physico-chemical soil attributes:

- chemical (0-20 cm layer): active acidity (pH) and total (H+Al); total organic matter (OM); available phosphorus (P); exchangeable calcium (Ca), magnesium (Mg) and potassium (K); sum of bases (SB=Ca+Mg+K); cation exchange capacity (CEC) in pH 7.0 (CEC=H+Al+SB); base saturation $(V=\frac{SB}{CEC}\times 100)$; aluminum saturation $(m=\frac{Al}{(Al+SB)})$.
- physical: (i) electrical conductivity (EC) in the 0-20 cm (surface EC) and 20-40 cm (subsurface EC) layers; (ii) cone index (CI) in the 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35 and 35-40 cm layers.

The dataset was generated as described in [16]. Before processing the data outliers were removed if data values exceed the mean plus or minus three the standard deviation [4]. Every record having at least one outlier in a variable was cut off from the dataset. After this preprocessing, the number of cases in the dataset was reduced from 2416 to 2138.

In next, the methods B2 and B4, proposed by [9], were applied to determine the variables that would be used to regress the multivariate linear models (MLR) and to the train neural networks. The aim was to generate models to predict the crop yield from the input variables in the dataset (chemical and physical soil properties). The

determination coefficient (R^2) was used to choose the best linear model generated by each procedure. The best generated ANNs were determined by evaluating the sum of squared error (SSE).

The generated networks were multilayer perceptron [21] with one input layer, one hidden layer and one output layer. The input layer was defined as the variables selected by B2 an B4. The number of nodes in the hidden layers was $\frac{1}{10}$ of the number of neurons in the input layer, with a minimum equal to 1. The output layer was set to have just one neuron: the crop yield. The algorithm used in the learning task was Backpropagation [7] and the number of epochs during training was 100.

In the sequence two methods for variable selection were developed. These methods, called B2+S and B4+S, combine SPCA with B2 and B4 procedures. The Figure 4 presents a pseudocode that describes the procedures B2+S and B4+S⁵.

Figure 4: Method for variable selection through supervised principal component analysis (B2+S, B4+S)

- 1. input a dataset \mathbf{D} with n attributes;
- 2. input a and b so that $a, b \in \mathbb{N}$ and $a \ge 2$; these inputs represent the minimum and the maximum number of variables to be rejected;
- 3. for each z from a to b do:
 - (a) for p going from 1 to z 1 do:
 - i. remove p variables using the supervised criterion;
 - ii. remove q = z p variables with B2 or B4;
 - iii. form a group G_p with the p+q selected (rejected) variables and form a subgroup of remaining variables $R_p = D G_p$;
 - iv. build a model with the remaining varibles ${\cal R}_p$ and execute it to get the performance (${\cal R}^2$ or sum of squared error or relative error)
 - v. save the performance V_p into a vector V, and go to next p;
 - (b) Save the V_p with the best performance value in V into a vector Z_n . Then, go to the next n;
- 4. Return vector Z with the best selections and analyze them to decide what the best subset of variables Z_n has provided the more accurate model.

As previously, the B2+S and B4+S procedures were applied to select variables for the linear regression and neural network training. Once again, the task was to generate models to predict the crop yield from the input variables. The neural network topology was constrained as described in the first test. As before, linear and ANN performances were evaluated by R^2 and SSE results, respectively.

The data set was split, many times, in training data (66% of the cases in the original data set) and validation data (33% of the cases). At all, each test was repeated several times on 35 different randomly generated training and evaluation sets. The overall result of each procedure was computed as the mean and the standard deviation of R^2 and SSE scores. We consider to remove up to 15 variables from the original data. The sample complexity was estimated assuming $\varepsilon=0.1$ and $\delta=0.05$.

⁵Note that these methods select at least two variables.

4 Results

The experiments were carried out aiming to compare the influence of the selection criteria on the predictive performance of the generated classifiers and regressors. The results are presented in the Tables 1 and 2. The first column of the Table 1 indicates the number of selected attributes, columns two to five present the mean and standard deviation of determination coefficient of MLR learned from attributes selected with B2, B4, B2 + S and B4 + S, respectively. The last column shows the sample complexity (m). Table 2 has a similar structure except that columns two to five reports the means and variance of sum of squared error of ANNs learned in each test set.

Table 1: Experimental results for MLR - R^2

#Attr.	B2		B4		B2+S		B4 + S		
Selected	μ	σ	μ	σ	μ	σ	μ	σ	\overline{m}
18	0.7062	0.0218	0.7047	0.0206	0.7097	0.0172	0.7059	0.0236	3277
17	0.7105	0.0184	0.7129	0.0198	0.7131	0.0178	0.7128	0.0191	3108
16	0.7095	0.0222	0.7048	0.0169	0.7209	0.0193	0.7194	0.0135	2939
15	0.7109	0.0205	0.7042	0.0215	0.7277	0.0145	0.7202	0.0125	2770
14	0.6964	0.0228	0.7031	0.0204	0.7278	0.0157	0.7268	0.0151	2601
13	0.7018	0.0218	0.6956	0.0197	0.7214	0.0119	0.7220	0.0130	2432
12	0.7009	0.0203	0.7032	0.0206	0.7235	0.0111	0.7234	0.0105	2263
11	0.6740	0.0248	0.6712	0.0159	0.7237	0.0115	0.7128	0.0125	2093
10	0.6553	0.0229	0.6743	0.0258	0.7126	0.0123	0.7048	0.0120	1924
9	0.6538	0.0229	0.6576	0.0232	0.6936	0.0110	0.6910	0.0156	1755
8	0.6488	0.0236	0.6513	0.0197	0.6880	0.0120	0.6867	0.0141	1586
7	0.6511	0.0252	0.6623	0.0230	0.6837	0.0112	0.6816	0.0099	1417
6	0.6443	0.0268	0.6397	0.0178	0.6846	0.0129	0.6762	0.0086	1248
5	0.6291	0.0290	0.5826	0.0278	0.6713	0.0137	0.6728	0.0164	1079

Table 2: Experimental results for ANN Classifier - SSE

#Attr.	B2		B4		B2 + S		B4 + S		
Sel.	μ	σ	μ	σ	μ	σ	μ	σ	\overline{m}
18	132.4737	22.0656	135.8639	22.6306	130.7489	23.3395	128.0359	23.4387	17652
17	134.1064	28.3094	144.1744	35.8705	115.2205	18.1805	119.9280	15.7663	16637
16	127.3671	23.6653	123.4054	22.9159	112.9119	18.0349	117.1445	14.0980	15792
15	130.6060	29.9840	126.7183	27.3079	108.7788	15.2239	109.3643	13.2144	14777
14	132.1723	35.5059	144.5409	38.7460	108.3647	12.4868	102.8849	13.7427	13932
13	141.8383	22.8159	119.0600	20.2282	100.0727	10.4200	108.2721	13.2380	6998
12	134.5904	20.0038	129.3881	18.5590	98.1271	12.2075	106.5018	8.1089	6490
11	117.4640	12.9698	122.1069	10.3275	111.0897	10.0081	115.0733	9.2617	5983
10	117.1820	9.1407	107.7079	10.7350	111.2271	7.9370	104.4065	9.2708	5476
9	113.0290	9.2985	105.1911	8.6089	110.0339	10.1347	110.3219	7.9812	4968
8	114.5403	11.8241	106.6504	8.0707	109.0685	6.7065	115.1356	9.1634	4461
7	108.6058	10.9559	107.9545	12.3594	110.3832	7.2711	121.2551	7.6308	4123
6	116.5220	12.1823	110.1580	10.2396	109.7526	6.4882	123.0404	8.0108	1079
5	157.6165	16.2754	168.4776	15.6080	127.5329	7.3648	117.7466	8.4202	910

Variable selection using B2+S and B4+S improved the performance of MLR learning. This conclusion is supported by observing that, in Table 1, the mean of the determination coefficient of regressors generated by B2+S and B4+S selection was higher than those generated with B2 and B4 (except in one case). Probably these better results were achieved because supervised procedure discards attributes which are irrelevant for prediction before applying PCA to select the variables that have more influence on uncorrelated principal components. That is, even a variable contributing to data variation can be discarded if it has a low correlation with the output variable.

Furthermore, the results indicate that variable selection can provide a criteria to deal with tradeoffs decisi-

ons between sample complexity requirements and maximization of accuracy score (or reversely, minimization of error based scores) when learning from small datasets. Table 1 shows that the sample complexity exigencies (m) decrease as less variables as selected. Particularly, the sample size demand is reached by the MLR models with five and six variables. The better mean for R^2 was obtained by models with six variables selected by B2+S. In this learning configuration the PAC learning theory specifies that the sample data must have 1248 records. Since the training database has 2416 cases, it satisfies the PAC learning set ($\varepsilon = 0.1$ and $\delta = 0.05$). Additionally, it must be observed that the supervised selection procedure allowed to induce a model that attends the sample complexity constraints with a not so large decline in the determination score ($R^2 \approx 0.684$).

The results of tests made with neural networks indicated that the supervised selection procedure contributed to generate better classifiers in contrast to those obtained only by applying B2 and B4. In particular, the results also showed that B2 + S variable selector also allowed to build ANNs with five or six input variables. These networks showed a small decline of performance and a model structure requiring a sample size of 1079 cases.

5 Discussion

B2+S and B4+S procedures reduces data dimensionality by applying a sequential search to remove irrelevant variables. The search procedure explores supervised information (standardized correlation coefficient) and principal component analysis combined with classifier/regressor performance to decide about attribute removal. The main advantage of this approach is to make possible to evaluate how much the use of a smaller set of attributes downsizes the requirements on sample size and fullfils the machine learning expectations.

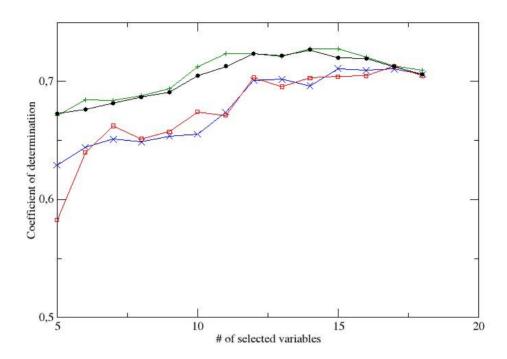
The attributes selected by B2+S and B4+S procedures preserved the most of the variation from data and influence on crop yield. It positively has impacted on reducing the computational effort when proceeding the regression or classification. The heuristic of B2+S and B4+S procedures favors the removal of bad regressors and correlated attributes providing a strategy o trade off between dimensionality reduction and model precision.

When intending select relevant variables for linear regression and to attend PAC learning requirements, the results showed that B2+S and B4+S performed better than B2 and B4. It can be more easily viewed in the figures 5 and 6 - for models with five or six variables. In those figures it is possible to observe that the supervised PCA scores are superior to PCA scores. The difference between the results comes from the fact that SPCA selects the variables which retains the most of the variance of the original data and have higher influence on the output variable. The performance of models derived from supervised procedures was significantly better than those generated with unsupervised ones - 5% of significance for paired T test. For neural networks the results are mixed. However, supervised PCA criteria obtained the best result for models with less attributes.

Krzanowski [13], Jollife [9] and King *et al* [12] also used PCA and sequential search to execute variable selection. In those works the efficacy of the selection process was measured by proceeding a Procrustes analysis to determine how much of the variation in original data was preserved by selected attributes. King *et al* [12] also used a similarity score that summarizes the correlation matrix between principal components computed from the original and reduced data. Those works do not consider the effect of variable selection in tasks as classification and regression. In contrast, the procedures presented here use the performance of machine learning process as the main score to quantify the effectiveness of the supervised/unsupervised strategies applied in variable selection.

The sequential search strategy employed by the proposed procedure can demand a considerable computational [5]. Such framework could be unsuitable for variable selection in large highly dimensional data sets. However,

Figure 5: Results of experiments with B2 (blue), B4 (red), B2+S (green) and B4+S (black) to variable selection for regression



the proposed procedure could be sped up by using a local search strategy as hill-climbing with multiple re-starts [21]. In such implementation it is necessary to consider the trade off between learning performance and speed.

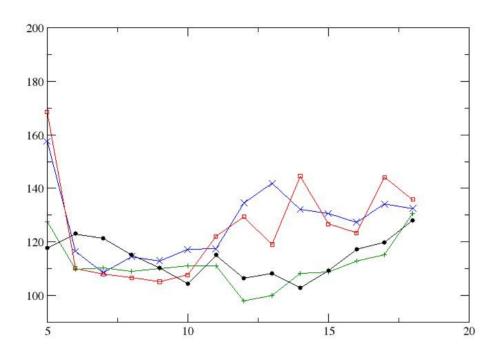
6 Conclusions

This work presented a procedure that combines supervised PCA and Joliffe's B2 and B4 procedures to select variables. The proposed procedure is compared with the original B2 and B4 methods considering two factors: the model quality and the sample complexity demands. The experiments show that, in general, the performance of B2+S and B4+S procedures was higher than B2 and B4 performance for selecting variables to proceed linear regression and neural network classification on a agricultural dataset. The experiments also allow to argue that, usually, the use of PCA based selection methods to reduct data dimension and sample complexity exigencies produces better regressors and classifiers when they are combined with supervised information.

When considering the application domain, that reducing the sample complexity requirements is an important issue when applying multivariate methods to analyze agricultura data sets because in many situations they have not many cases. The proposed procedure can contribute to extend the range of multivariate data analysis in agricultural research by applying dimensionality reduction.

In future, we intend to extend this work in two ways. Firstly, by developing procedures that allow to

Figure 6: Results of experiments with B2 (blue), B4 (red), B2+S (green) and B4+S (black) to variable selection for regression



incorporate prior knowledge. Secondly, by evaluating the use of heuristic search methods instead of a sequential one.

7 References

- [1] E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, 101:119–137, 2006.
- [2] S. Chatterjee and A. S. Hadi. *Regression analysis by example*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley-Interscience, 2006.
- [3] D.F. Ferreira. Multivariate Statistic. Ed. UFLA, Lavras, MG, 1 edition, 2008.
- [4] D. Gonçalves. *Physics*. Ao Livro Técnico, Rio de Janeiro, 2a edition, 1965.
- [5] Q. Guo, W. Wu, D.L. Massart, C. Boucon, and S. de Jong. Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(1?2):123 132, 2002.
- [6] D. Haussler. Probably approximately correct learning. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 1101–1108. AAAI Press, 1990.

- [7] S.S. Haykin. Neural Networks Principles and Practice. Bookman, 2001.
- [8] T. Hemant and V. Cevher. Active learning of multi-index function models. In P. Bartlett, F.c.n. Pereira, C.j.c. Burges, L. Bottou, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems* 25, pages 1475–1483. 2012.
- [9] I.T. Jolliffe. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21(2):160 173, 1972.
- [10] I.T. Jolliffe. Discarding variables in a principal component analysis. ii: Real data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(1):21 31, 1973.
- [11] I.T. Jolliffe. Principal Components Analysis. Springer Verlag, New York, 2 edition, 2002.
- [12] J.R. King and D.A. Jackson. Variable selection in large environmental data sets using Principal Components Analysis. *Environmetrics*, 10(1):67 77, 1999.
- [13] W.J. Krzanowski. Selection of variables to preserve multivariate data structure, using principal components. *Applied Statistics*, 36:22 33, 1987.
- [14] V. Kumar, M. Steinbach, and P. N. Tan. *Introduction to data mining*. Pearson Addison Wesley, 2006.
- [15] T. Mitchell. Machine Learning. McGraw-Hill, New York, 1997.
- [16] J.P. Molin, H.T.Z. Couto, L.M. Gimenez, V. Pauletti, R. Molin, and S.R. Vieira. Regression and correlation analysis of grid soil data versus cell spatial data. In *Third European Conference on Precision Agriculture:**Agro Montpellier*, pages 449 453, 2001.
- [17] Y. Mori, M. Iizuka, T. Tarumi, and Y. Tanaka. Variable selection in principal component analysis. In *Statistical Methods for Biostatistics and Related Fields*, pages 265–283. Springer Berlin Heidelberg, 2007.
- [18] A. Nouri and M.L. Littman. Dimension reduction and its application to model-based exploration in continuous spaces. *Mach. Learn.*, 81(1):85–98, October 2010.
- [19] A. Onisko, M.J. Druzdzel, and H. Wasyluk. Learning bayesian networks from small data sets: application of noise-or gates. *International Journal of Approximate Reasoning*, 27(2):165 182, 2001.
- [20] J. Pacheco, S. Casado, and S. Porras. Exact methods for variable selection in principal component analysis: Guide functions and pre-selection. *Computational Statistics and Data Analysis*, 57(1):95 111, 2013.
- [21] S. Russell and P. Norvig. *Artificial Intelligence: A modern approach*. Prentice Hall, Upper Saddle River, 1995.
- [22] X. Shao, V. Cherkassky, and W. Li. Measuring the vc-dimension using optimized experimental design. *Neural Computation*, 12:2000, 1969.
- [23] Siwadi, A. Muslin, and T. Bakhtiar. Discarding variables in a principal component analysis. ii: Real data. *Journal of of Asian Scientific Research*, 2(12):856 865, 2012.
- [24] S.Roberts and M.A. Martin. Using Supervised Principal Components Analysis to assess multiple pollutant effects. *Environmental Health Perspectives*, 114(12):1877 1882, 08 2006.

- [25] J.H.M. Thornley and I.R. Johnson. *Plant and crop modelling: a mathematical approach to plant and crop physiology*. Oxford science publications. Clarendon Press, 1990.
- [26] F. Westad, M. Hersletha, P. Lea, and H. Martens. Variable selection in pca in sensory descriptive and consumer data. *Food Quality and Preference*, 14(5?6):463 472, 2003.
- [27] S. Yu, K. Yu, V. Tresp, H-P. Kriegel, and M. Wu. Supervised Probabilistic Principal Component Analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 464–473, New York, NY, USA, 2006. ACM Press.