JUSCELINO IZIDORO DE OLIVEIRA JR.

SELEÇÃO DE VARIÁVEIS NA MINERAÇÃO DE DADOS AGRÍCOLAS:
Uma abordagem baseada em análise de componentes principais

Ponta Grossa

Julho/2012

JUSCELINO IZIDORO DE OLIVEIRA JR.

SELEÇÃO DE VARIÁVEIS NA MINERAÇÃO DE DADOS AGRÍCOLAS: Uma abordagem baseada em análise de componentes principais

Dissertação submetida ao Programa de Pós-Graduação em Computação Aplicada - Área de concentração Computação para Tecnologias em Agricultura - da Universidade Estadual de Ponta Grossa como requisito para a obtenção do título de mestre.

Orientador:
Profº Dr. José Carlos Ferreira da Rocha

Co-orientador:
Profº Dr. Adriel Ferreira da Fonseca

UNIVERSIDADE ESTADUAL DE PONTA GROSSA

Ponta Grossa

Julho/2012

Ficha Catalográfica Elaborada pelo Setor Tratamento da Informação Belém/UEPG

Oliveira Junior, Juscelino Izidoro de
O48s Seleção de variáveis na mineração de dados agrícolas : uma abordagem baseada em análise de componentes principais / Juscelino Izidoro de Oliveira Junior . Ponta Grossa, 2012.
87 f.

Dissertação (Mestrado em Computação Aplicada – área de concentração Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.
Orientador: Prof. Dr. José Carlos Ferreira da Rocha.
Coorientador: Prof. Dr. Adriel Ferreira da Fonseca.

1. Complexidade da Amostra. 2. Dados rotulados. 3. Modelagem agrícola. 4. Redução de dimensionalidade. I. Rocha, José Carlos Ferreira da. II. Fonseca, Adriel Ferreira da. III. Universidade Estadual de Ponta Grossa. Mestrado em Computação Aplicada. IV. T.

CDD: 006.312

TERMO DE APROVAÇÃO

JUSCELINO IZIDORO DE OLIVEIRA JR.

"SELEÇÃO DE VARIÁVEIS NA MINERAÇÃO DE DADOS AGRÍCOLAS: Uma abordagem baseada em análise de componentes principais"

Dissertação submetida ao Programa de Pós-Graduação em Computação Aplicada - Área de concentração Computação para Tecnologias em Agricultura - da Universidade Estadual de Ponta Grossa como requisito para a obtenção do título de mestre.

Ponta Grossa, 30 de Julho de 2012.

Prof. Dr. José Carlos Ferreira da Rocha - Orientador
Doutor em Engenharia Mecânica
Universidade Estadual de Ponta Grossa

Prof. Dr. Ivo Mario Mathias
Doutor em Agronomia (Energia na Agricultura)
Universidade Estadual de Ponta Grossa

Prof. Dr. Daniel Kikuti
Doutor em Engenharia Mecânica
Universidade de Estadual do Centro-Oeste

AGRADECIMENTOS

Agradeço a DEUS, por estar comigo em todos os momentos, provendo-me saúde e sabedoria para realizar meus estudos.

Agradeço a minha família por me apoiar e incentivar nos estudos e me motivar cada vez mais a aprender.

Agradeço ao meu orientador e grande amigo, Dr. José Carlos Ferreira da Rocha, que me orientou transmitindo o conhecimento necessário para que eu pudesse desenvolver este trabalho da melhor maneira.

Agradeço ao meu co-orientador e amigo, Dr. Adriel Ferreira da Fonseca, pelos ensinamentos transmitidos e disposição em me ajudar a desenvolver meu trabalho de mestrado.

Agradeço ao Dr. José Paulo Molin e ao Dr. Eduardo Fávero Caires, por cederem os conjuntos de dados agrícolas utilizados nos experimentos deste trabalho.

Agradeço a todos os professores e amigos da UEPG, por terem contribuído significativamente com meu progresso em minha formação acadêmica.

Agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de pós-graduação.

RESUMO

A análise multivariada de dados permite verificar a interação de vários atributos que podem influenciar o comportamento de uma variável de resposta. Tal análise utiliza modelos que podem ser induzidos de conjuntos de dados experimentais. Um fator importante na indução de regressores e classificadores multivariados é o tamanho da amostra, pois, esta determina a confiabilidade do modelo quando há a necessidade de se regredir ou classificar a variável de resposta. Este trabalho aborda a questão do tamanho da amostra por meio da Teoria do Aprendizado Provavelmente Aproximadamente Correto, oriundo de problemas sobre o aprendizado de máquina para a indução de modelos. Dada a importância da modelagem agrícola, este trabalho apresenta dois procedimentos para a seleção de variáveis. O procedimento de Seleção de Variáveis por Análise de Componentes Principais, que não é supervisionado e permite ao pesquisador de agricultura selecionar as variáveis mais relevantes de um conjunto de dados agrícolas considerando a variação contida nos dados. O procedimento de Seleção de Variáveis por Análise de Componentes Principais Supervisionado, que é supervisionado e permite realizar o mesmo processo do primeiro procedimento, mas concentrando-se apenas nas variáveis que possuem maior influência no comportamento da variável de resposta. Ambos permitem que informações a respeito da complexidade da amostra sejam exploradas na seleção de variáveis. Os dois procedimentos foram avaliados em cinco experimentos, mostrando que o procedimento supervisionado permitiu, em média, induzir modelos que produziram melhores pontuações do que aqueles modelos gerados sobre as variáveis selecionadas pelo procedimento não supervisionado. Os experimentos também permitiram verificar que as variáveis selecionadas por ambos os procedimentos apresentavam índices reduzidos de multicolinaridade..

Palavras-chave: Complexidade da Amostra. Dados rotulados. Redução de dimensionalidade.

ABSTRACT

Multivariate data analysis allows the researcher to verify the interaction among a lot of attributes that can influence the behavior of a response variable. That analysis uses models that can be induced from experimental data set. An important issue in the induction of multivariate regressors and classifiers is the sample size, because this determines the reliability of the model for tasks of regression or classification of the response variable. This work approachs the sample size issue through the Theory of Probably Approximately Correct Learning, that comes from problems about machine learning for induction of models. Given the importance of agricultural modelling, this work shows two procedures to select variables. Variable Selection by Principal Component Analysis is an unsupervised procedure and allows the researcher to select the most relevant variables from the agricultural data by considering the variation in the data. Variable Selection by Supervised Principal Component Analysis is a supervised procedure and allows the researcher to perform the same process as in the previous procedure, but concentrating the focus of the selection over the variables with more influence in the behavior of the response variable. Both procedures allow the sample complexity informations to be explored in variable selection process. Those procedures were tested in five experiments, showing that the supervised procedure has allowed to induce models that produced better scores, by mean, than that models induced over variables selected by unsupervised procedure. Those experiments also allowed to verify that the variables selected by the unsupervised and supervised procedure showed reduced indices of multicolinearity.

Keywords: Sample Complexity. Labeled Data. Dimensionality Reduction.

LISTA DE SIGLAS

ACP	Análise de Componentes Principais
ACPP	Análise de Componentes Principais Probabilística
ACPS	Análise de Componentes Principais Supervisionada
DCBD	Descoberta de Conhecimento em Banco de Dados
DVS	Decomposição por Valores Singulares
EM	Expectativa e Maximização
FIV	Fator de Inflação da Variância
MD	Mineração de Dados
MRLM	Modelo de Regressão Linear Múltipla
NB	*Naive Bayes*
PAC	Provavelmente Aproximadamente Correto
RNA	Redes Neurais Artificiais
SVACP	Seleção de Variáveis por Análise de Componentes Principais
SVACPS	Seleção de Variáveis por Análise de Componentes Principais Supervisionada
SQE	Soma dos Quadrados dos Erros
VC	Vapnik-Chervonenkis

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A análise multivariada de dados provê métodos que permitem a confecção de modelos que procuram capturar as interações entre elementos envolvidos em processos agrícolas (THORNLEY; JOHNSON, 1990). Um passo precedente à modelagem multivariada, que também deve ser considerada para dados de agricultura, consiste em selecionar as variáveis mais relevantes do conjunto de dados em relação à variável de resposta (ROBERTS; MARTIN, 2006; YU *et al.*, 2006). A seleção de um conjunto adequado de variáveis permite: (i) descartar atributos redundantes ou aqueles que adicionam ruído aos dados; (ii) reduzir o risco de sobre-ajuste[1]; (iii) reduzir a complexidade da amostra; (iv) tornar o modelo simples, reduzindo o número de variáveis; (v) poupar tempo e recursos com coletas de dados futuras (KING; JACKSON, 1999; JOLLIFFE, 2002; WITTEN; FRANK, 2005; KUMAR *et al.*, 2009; HAN; KAMBER; PEI, 2011).

Nesse contexto, técnicas multivariadas que exploram relações, como a Análise de Componentes Principais (ACP), podem prover métodos eficientes para a seleçao de variáveis (JOLLIFFE, 1972). Esse procedimento reduz a dimensionalidade dos dados, removendo as variáveis inter-relacionadas enquanto retém a máxima variação possível presente no conjunto de dados (FERREIRA, 2008). Entretanto, como a ACP é um procedimento não supervisionado, ela usa apenas as variáveis de entrada do sistema durante a etapa de seleção e não explora a informação dada pela variável de resposta durante o processo de seleção. Isto é, a ACP não considera a influência das variáveis de entrada sobre a variável de resposta para decidir qual variável de entrada descartar. Para contornar esta limitação, Bair *et al.* (2006) propuseram a técnica chamada Análise de Componentes Principais Supervisionada (ACPS). Nesta técnica a informação fornecida pela variável de saída é utilizada como um pré-filtro sobre as variáveis de entrada da ACP. O pré-filtro seleciona apenas as que têm associação mais forte com a variável de resposta e a ACP é executada sobre as variáveis pré-selecionadas. Isso permite que a ACPS execute a análise de componentes principais apenas sobre as variáveis de maior influência sobre a variável de resposta.

A indução de modelos de regressão ou classificação sobre os dados agrícolas pode ser abordada com técnicas de Aprendizado de Máquina (GUIMARÃES, 2005; MATHIAS, 2006), que estimam a estrutura e os parâmetros do modelo a partir dos dados coletados. Neste contexto, uma exigência da modelagem multivariada é a complexidade da amostra - o número de amostras (exemplos, instâncias ou registros) necessárias para que um

[1] Do inglês, *overfitting*.

algoritmo de aprendizado aprenda um conceito a partir dos dados, de forma que o erro do modelo seja controlado (HAYKIN, 2001). A complexidade da amostra está relacionada ao tipo de modelo que se pretende induzir ou regredir e, portanto, depende do número de parâmetros - coeficientes ligados às variáveis - que devem ser estimados. Basicamente, quanto mais parâmetros precisam ser estimados maiores são as exigências sobre a complexidade da amostra. Portanto, o número de variáveis que compõem um modelo também tem influência sobre o tamanho da amostra necessária para geração de tal modelo.

Considerando isto, o objetivo deste trabalho é apresentar e avaliar o emprego de dois procedimentos para a seleção de variáveis, para execução de tarefas de mineração de dados agrícolas - em particular para tarefas envolvendo regressão numérica e indução de classificadores. Os procedimentos propostos neste trabalho foram programados em linguagem R e são denominados Seleção de Variáveis por Análise de Componentes Principais (SVACP) e Seleção de Variáveis por Análise de Componentes Principais Supervisionada (SVACPS). O procedimento SVACP implementa um mecanismo de busca que explora duas técnicas de seleção de variáveis por ACP sem supervisão para determinar o conjunto de variáveis que deve compor um regressor ou classificador, tendo em vista o desempenho do modelo gerado e o tamanho da amostra. Os critérios de seleção empregados pelo SVACP são chamados B2 e B4, e foram apresentados por Jolliffe (1972) e Jolliffe (1973). O objetivo destes critérios é selecionar o conjunto de variáveis que está relacionado à variação dos dados observados. O procedimento SVACPS estende o SVACP ao combinar os métodos B2 e B4 com a técnica da ACP Supervisionada proposta por Bair *et al.* (2006). A SVACPS considera o emprego dos critérios B2 e B4 somente sobre as variáveis que têm maior influência em relação à variável de saída. Os procedimentos SVACP e SVACPS permitem que a seleção de variáveis seja realizada sobre bases de dados completas ou incompletas (sendo esta última definida considerando que os valores de alguns atributos não são conhecidos em todos os casos que compõem uma amostra). Além disto, uma vez que muitos experimentos em agricultura procuram avaliar o impacto da manipulação de determinadas variáveis sobre uma variável dependente, os procedimentos SVACP e SVACPS permitem que o processo ocorra de modo que as variáveis de interesse sejam selecionadas obrigatoriamente.

Os procedimentos SVACP e SVACPS foram avaliados em cinco experimentos, em que foram empregados na seleção de variáveis para geração de Modelos de Regressão Linear Múltipla, Classificadores Bayesianos e Redes Neurais Artificiais. O primeiro e o segundo experimentos foram realizados sobre uma base de dados sintética, que permitiu avaliar os resultados obtidos em uma situação em que as características do modelo

alvo eram conhecidas. Nestes experimentos foram consideradas bases de dados completa e incompleta. O terceiro e o quarto experimentos foram realizados sobre um conjunto de dados referente a um experimento agrícola, o que permitiu avaliar os resultados obtidos em uma situação aplicada, em que as características do modelo alvo eram desconhecidas. O quinto experimento foi realizado com um terceiro conjunto de dados, também referente a um experimento agrícola, e permitiu avaliar a situação em que a variável de interesse deve, obrigatoriamente, constar no modelo. Os resultados mostraram que os procedimentos foram capazes de selecionar variáveis que permitiram a geração de modelos, para representar aqueles conjunto de dados, ao mesmo tempo que se reduziam as exigências referentes ao tamanho da amostra.

Este trabalho está organizado da seguinte forma: no Capítulo 2 é feita uma revisão do conteúdo teórico de modo a apresentar o processo de descoberta de conhecimento (Seção 2.1), o processo de redução de dimensionalidade aplicada a trabalhos da agricultura (Seção 2.2), a regressão e a classificação (Seção 2.3) e os modelos utilizados nos experimentos, a Análise de Componentes Principais e suas variações (Seções 2.4, 2.6 e 2.5) e a complexidade da amostra (Seção 2.7). No Capítulo 3 é apresentada a metodologia utilizada, mostrando os métodos para realizar a seleção de variáveis (Seção 3.1) e os conjuntos de dados (Seção 3.2). No Capítulo 4 são apresentados e discutidos os resultados dos experimentos. No Capítulo 5 são apresentadas as conclusões e as considerações finais acerca deste trabalho.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo destaca os principais elementos do problema da redução de dimensionalidade na mineração de dados, em particular na mineração de dados agrícolas. Para tanto, a Seção 2.1 faz uma introdução à descoberta de conhecimento em bancos de dados (DCBD) em que são apresentados os principais conceitos e as etapas de processamento. Em particular, detaca-se a importância da redução de dimensionalidade para a etapa mineração de dados. A Seção 2.2 aborda o uso de métodos de regressão e classificação na pesquisa em Agricultura e destaca a importância da redução de dimensionalidade nestas pesquisas. A Seção 2.3 apresenta três métodos de regressão e classificação frequentemente empregados na mineração de dados agrícolas. As seções 2.4, 2.5 e 2.6 descrevem a Análise de Componentes Principais e suas derivações: a Análise de Componentes Principais Supervisionada e a Análise de Componentes Principais Probabilística. A Seção 2.7 trata da complexidade da amostra em relação à Teoria do Aprendizado Provavelmente Aproximadamente Correto, quando se considera os métodos de regressão e classificação utilizados no desenvolvimento deste trabalho. A Seção 2.8 apresenta as considerações finais do capítulo.

2.1 DESCOBERTA DE CONHECIMENTOS EM BANCOS DE DADOS E REDUÇÃO DE DIMENSIONALIDADE

A descoberta de conhecimento em bancos de dados, segundo Frawley, Piatetsky-Shapiro e Matheus (1992), é a extração não trivial de informações implícitas, desconhecidas e potencialmente relevantes a partir de conjuntos de dados. Seja $\mathbf{F}$ um conjunto de fatos (dados), L uma linguagem e C uma medida de certeza. Um *padrão* é uma afirmação S em L que descreve relações entre os atributos de $\mathbf{F}$ com certeza c. Usualmente, S é mais compacto (representação implícita do conhecimento) do que a enumeração de todos os fatos por ele representados. O *padrão* é chamado de *conhecimento*, pois, concorda com o interesse do usuário a um certo nível de certeza c, de acordo com os critérios estabelecidos pelo próprio usuário. A descoberta de conhecimento pode, então, ser definida como a saída de um programa que verifica o conjunto de fatos em um conjunto de dados e produz padrões.

Segundo Han, Kamber e Pei (2011), a descoberta de conhecimento é basicamente uma série de processos que são executados iterativamente com o objetivo de se obter representações implícitas nos dados. O processamento pode ser abstraído como:

1. **Limpeza de dados:** nesta etapa são realizados procedimentos para remover da-

dos inconsistentes, que são aqueles que não seguem o padrão dos demais dados do conjunto e que podem atrapalhar na busca por padrões. Por exemplo, os valores discrepantes ou registros incompletos podem ser removidos do conjunto ou podem ser estimados por alguma técnica de regressão.

2. **Integração de dados:** múltiplas fontes de dados podem ser combinadas em uma só. Isso permite centralizar o processamento em apenas um conjunto com todos os dados necessários. Também, ajuda a prover um padrão na representação dos dados. Normalmente, são estabelecidas convenções de nomenclatura dos atributos, padrões para a representação dos dados e ajuste dos valores nas mesmas unidades de medidas. Além disso, centralizar os dados em um único conjunto, evita ter que utilizar vários protocolos para acessar outros meios de armazenamento como: fitas magnéticas, gerenciadores de bancos de dados, arquivos de texto ou planilhas eletrônicas;

3. **Seleção de dados:** nesta parte os dados mais relevantes são selecionados para a análise e são separados para o processo seguinte. Por exemplo, pode-se selecionar as variáveis que contribuem significativamente em um determinado padrão, ou eliminar aquelas que contribuem pouco ou não contribuem. A redução de dimensionalidade dos dados pode ser empregada para este fim;

4. **Transformação de dados:** nesta etapa os dados são transformados ou organizados em um formato apropriado para a execução do processamento seguinte, auxiliando na representação do conhecimento implícito e facilitando o reconhecimento do mesmo por algoritmos de aprendizado de máquina. Por exemplo, alguns atributos podem ser discretizados ou normalizados;

5. **Mineração de dados:** neste passo, técnicas de extração de padrões são aplicadas com o objetivo de obter representações implícitas de conhecimentos contidos nos dados;

6. **Avaliação de padrões:** nesta etapa são realizados testes para identificar a validade dos padrões obtidos de acordo com medidas de interesse estabelecidas pelo usuário;

7. **Apresentação de conhecimento:** nesta etapa, técnicas de visualização e representação de conhecimento são empregadas para mostrar os resultados ao usuário.

Os passos de 1 até 4 são formas de pré-processamento que visam preparar os dados em um formato adequado para a mineração de dados (MD). O passo 5 realiza a extração dos padrões por meio da aplicação de algoritmos especializados. Conforme

Han, Kamber e Pei (2011) comentam, a MD é caracterizada pela utilização de técnicas de Inteligência Artificial, Algoritmos de Aprendizagem de Máquina e Estatística, que são combinadas para explorar um conjunto de dados e evidenciar padrões. Os passos 6 e 7 analisam a probabilidade de que os padrões encontrados sejam verdadeiros e apresentam o conhecimento obtido para o usuário de modo compreensível (tabelas, gráficos e grafos).

Conforme Witten e Frank (2005) explicam, a MD tem o foco específico em encontrar e descrever padrões estruturais (modelos), de modo que eles possam ser usados para explicar[1] os dados e, também, realizar tarefas como regressão e classificação. A regressão consiste em inferir a relação entre duas ou mais variáveis por meio de fórmulas matemáticas para serem empregadas em predição numérica. A classificação é a tarefa que consiste em determinar a relação entre duas ou mais variáveis por meio de representações lógicas para serem empregadas em predição de categorias.

Um passo importante na aquisição de conhecimento é o pré-processamento dos dados. Os dados do mundo real são suscetíveis a ruídos, dados incompletos, valores inconsistentes devido a sua origem, de fontes heterogêneas. Witten e Frank (2005), Kumar *et al.* (2009), Han, Kamber e Pei (2011) apresentam várias técnicas de pré-processamento entre elas a seleção de variáveis (ou seleção de atributos, ou seleção de características). O objetivo da seleção de variáveis é eliminar aquelas que pouco influenciam na explicação dos dados e, consequentemente, reter as que melhor sintetizam o comportamento observado. A ideia básica é selecionar um conjunto de variáveis que contribua para obtenção de resultados na MD. Witten e Frank (2005), Kumar *et al.* (2009), Han, Kamber e Pei (2011) enumeram as seguintes estratégias para a seleção de atributos:

- **Seleção *Stepwise Forward*:** Inicialmente o subconjunto é inicializado vazio. Uma medida de ganho de informação é usada para selecionar um dos atributos originais. O atributo selecionado é adicionado ao subconjunto e, assim, sucessivamente até que o último atributo, que provê maior ganho de informação, seja adicionado ao subconjunto.
- **Eliminação *Stepwise Backward*:** Inicialmente o subconjunto é inicializado com todas as variáveis originais. A cada passo, é removida uma variável que provê menor ganho de informação. O subconjunto final contém as variáveis desejadas.

A seleção de variáveis é uma das técnicas existentes para a redução de dimensionalidade, pois, o objetivo desta é permitir representar a mesma explicação dos dados, ou

[1] Explicar os dados significa buscar argumentos que justifiquem determinado valor.

aproximadamente a mesma, por meio de menos variáveis. Jolliffe (2002) apresenta dois tipos de redução de dimensionalidade: 1) transformar o conjunto de dados com as variáveis originais em um conjunto com variáveis latentes, em que estas são formadas por uma combinação daquelas e o número de dimensões é reduzido; 2) selecionar ou eliminar um subgrupo de variáveis originais com base em um critério específico.

Este trabalho utiliza a técnica de eliminação de variáveis baseada em Análise de Componentes Principais, que consiste em transformar os dados pela ACP de modo a produzir uma matriz de pesos (cargas), que é analisada para determinar quais das variáveis são irrelevantes e devem ser descartadas do conjunto original (JOLLIFFE, 1972, 1973). Para tanto, empregam-se dois critérios que têm sido usualmente utilizados nesta tarefa, são eles o critério B2 e o B4. O B2 é realizado executando-se a ACP sobre o conjunto de dados com todas as p variáveis originais; então, inspecionam-se os k autovalores de modo que não ultrapassem um limiar λ_0, definido pelo usuário. Retêm-se os k autovetores correspondentes aos k autovalores. Primeiramente, analisa-se o autovetor de menor autovalor correspondente. Em seguida, analisa-se o autovetor com o segundo menor autovalor e assim por diante até analisar os k autovetores. Escolhe-se uma variável com o maior coeficiente naquele componente (autovetor) que está em análise, considerando que a variável selecionada ainda não foi associada aos componentes analisados anteriormente. Então, as k variáveis selecionadas são rejeitadas, ou seja, elas são removidas do conjunto de dados. O método B4 funciona de forma semelhante ao B2. A diferença é que o B4 analisa os k componentes do primeiro para o último. Finalmente, os métodos B2 e B4 retêm $p - k$ variáveis originais, eliminando as k variáveis que menos contribuem para explicar a variação dos dados.

O número k de variáveis a serem rejeitadas depende do número λ_0 escolhido pelo usuário. Uma alternativa para a eliminação seria ignorar o λ_0 e fixar um número k de variáveis a serem rejeitadas.

Na Seção 2.2, a seguir, é destacada a importância da redução de dimensionalidade para a pesquisa e desenvolvimento em Agricultura.

2.2 REDUÇÃO DE DIMENSIONALIDADE DE DADOS AGRÍCOLAS

Segundo Carvalho (1946), Gomez e Gomez (1984), Mead, Curnow e Hasted (2003), muitas vezes o pesquisador da área agrícola obtém novos conhecimentos por meio da análise de um conjunto de dados que são provenientes de experimentos. Basicamente, os dados são processados com o objetivo de análisar as variáveis que influenciam um fenô-

meno de interesse. O processamento estatístico provê uma maneira de extrair informações de um conjunto de dados, pois, permite tirar conclusões que condizem com a realidade dos fatos, evitando tirar conclusões de características apenas aparentes que podem conduzir a uma interpretação errada da realidade.

Em se tratanto de processamento estatístico, a regressão é uma das técnicas estatísticas utilizada na agricultura para proceder a análise dos dados. Alguns exemplos desse procedimento podem ser vistos nos trabalhos de: Lobell e Asner (2003) que analisaram por meio de regressão a influência do clima no rendimento de grãos; Karkacier, Goktolga e Cicek (2006) que utilizaram a regressão para analisar a relação entre o consumo de energia e o rendimento de grãos; Shuai e Hong (2011) que empregaram análise de regressão sobre dados de pesticidas, fertilizantes, área cultivada, energia consumida por maquinários, para identificar os fatores que mais influenciaram na emissão de gás de efeito estufa sobre a área de produção, em uma província chinesa.

Outra maneira de analisar dados é por meio de modelos de classificação, que podem ser conseguidos com técnicas de mineração de dados. A MD possui algumas aplicações na agricultura no tocante a tarefas que envolvem classificação de amostras. Alguns exemplos de classificação na agricultura podem ser observados nos trabalhos de: El-Telbany, Warda e El-Borahy (2006) que usaram técnicas de MD para descobrir regras e classificar doenças do arroz egípcio usando o algoritmo de árvore de decisão C4.5; Armstrong, Diepeveen e Maddern (2007) que utilizaram técnicas de MD buscar padrões nos dados e classificar perfis de solos; Martins e Fonseca (2009) que empregaram a MD com o propósito de auxiliar na classificação de regiões por meio de imagens de satélite, sobre uma região de atividade agrícola;

Quando o pesquisador dispõe de muitas variáveis para serem analisadas de maneira conjunta, por exemplo, 10 ou mais variáveis, é útil reduzir o número de variáveis para realizar procedimentos de regressão multivariada ou classificação (JOLLIFFE, 1972). Segundo Jolliffe (1972), a redução no número de variáveis serve ao pesquisador como uma maneira de minimizar o custo computacional, reduzir o tempo de processamento e, em experimentos futuros, economizar recursos por poder coletar menos dados. Além disso, os modelos gerados a partir do conjunto reduzido de variáveis são menores e, portanto, mais fáceis de serem interpretados (KUMAR *et al.*, 2009).

Selecionar um subconjunto de variáveis é como aplicar uma *navalha de Ockham* ao processo de análise. Segundo Russell e Norvig (2004), a expressão *navalha de Ockham* significa preferir a hipótese mais simples e consistente com os dados, e descartar qualquer

premissa que não contribua para explicar o fenômeno em estudo. Segundo Han, Kamber e Pei (2011), a redução de dimensionalidade é uma técnica que permite reduzir o volume de dados, mas produzindo os mesmos (ou quase os mesmos) resultados analíticos.

A ACP é uma das técnicas que tem sido utilizada nas pesquisas em agricultura para a redução de dimensionalidade. Ela tem sido utilizada como um pré-processamento à análise multivariada. Alguns exemplos de utilização da ACP na agricultura podem ser vistos nos trabalhos de: Alvarenga e Davide (1999) que usaram a ACP para relacionar mudanças físicas e químicas de um solo sob cinco agroecossistemas, a fim de definir qual dos cinco tinham características de sustentabilidade; Sena *et al.* (2002) que empregaram a ACP para avaliar os efeitos de diferentes práticas de manejo na qualidade do solo e identificar as práticas de manejo sobre oito diferentes parâmetros, além de identificar quais destes parâmetros foram mais importantes em tal análise; Myers *et al.* (2005) que utilizaram a ACP para examinar a relação entre o solo e a concentração de nutrientes das folhas da soja, e verificar a relação de tais elementos com a presença de afídio; Shtangeeva *et al.* (2009) que aplicaram a ACP para visualizar a bioacumulação de vários elementos em diferentes espécies de plantas e avaliar a contribuição de quais fatores podem afetar a interação entre solo e planta; Santos, Santos e Conti (2010) que usaram a ACP como ferramenta para analisar amostras de solos e avaliar o conteúdo de nutrientes e elementos tóxicos sob diferentes métodos agrícolas no cultivo de café; Kummer *et al.* (2010) que empregaram a ACP para verificar a similaridade de amostras de solo e estabelecer relações com o material de origem, profundidade do solo e interferências antrópicas; Rossel *et al.* (2011) que aplicaram a ACP para reduzir a dimensionalidade de espectros de infravermelho próximo, utilizados para medir a composição do solo como uma alternativa à análise de solo convencional.

A redução de dimensionalidade é empregada com o objetivo de comprimir o volume de dados, criando variáveis latentes para a análise dos mesmos. Isso é vantajoso por permitir visualizar menos variáveis. Porém, a técnica também pode ser empregada com a finalidade de reduzir a complexidade da amostra. Quanto mais complexo é um modelo, ou seja, quanto mais variáveis ele possui, mais amostras são necessárias para dar confiabilidade aos resultados (VANDENBERG, 2009). E estatisticamente, quanto mais amostras o conjunto possuir, maior é a probabilidade de que novas amostras tomadas ao acaso, do mesmo universo amostral, permitam tirar as mesmas conclusões (PILLAR, 2004). A redução de dimensionalidade permite selecionar um subconjunto de variáveis que é menor do que o original, possibilitando que um modelo multivariado tenha o mesmo desempenho (ou aproximadamente o mesmo) com menos variáveis. A simplificação do

modelo permite que o treinamento necessite um número menor de amostras.

Como pode ser observado nos trabalhos citados, a ACP tem sido usada na agricultura para a redução de dimensionalidade dos dados. Contudo, ela tem sido empregada como uma técnica multivariada não supervisionada e seu uso não é voltado para a seleção de variáveis. A supervisão é um recurso que poderia ser mais explorado na agricultura, pois, em vários casos o agrônomo está interessado em estudar o impacto de um conjunto de variáveis independentes sobre uma variável de resposta. Por exemplo, nos casos em que se tenta correlacionar atributos de solo com rendimento de grãos, ou ainda, correlacioná-los com a concentração de nutrientes nas folhas das plantas. Uma das vantagens em realizar a redução de dimensionalidade utilizando a supervisão é que a análise dos fatores fica concentrada apenas naquelas variáveis que apresentam maior influência sobre a variável de resposta. Assim, além de permitir simplificar a ACP, a supervisão auxilia a tornar a análise mais expressiva, representando melhor as relações das demais variáveis com a de resposta. A Seção 2.5 apresenta a análise de componentes principais supervisioada (ACPS) que permite explorar dados rotulados.

2.3 REGRESSÃO E CLASSIFICAÇÃO

Esta seção apresenta os conceitos sobre regressão e classificação na análise de dados, bem como as três técnicas utilizadas neste trabalho que foram: Regressão por Modelo Linear Múltiplo, classificação por Redes Neurais Artificiais e classificação por *Naive Bayes*. Estas técnicas foram empregadas para demonstrar o impacto do resultado de procedimentos de seleção de variáveis, sobre modelos de regressão e classificação.

A análise de regressão, estuda a relação que existe entre duas ou mais variáveis (CARVALHO, 1946; SOUNIS, 1971; AFIFI; AZEN, 1979). Uma delas é a variável dependente (ou variável de resposta) e as demais são as variáveis independentes (ou variáveis de entrada). Essa relação é descrita por meio de uma expressão matemática - ou função - que associa a variável dependente às independentes. Segundo Afifi e Azen (1979), os problemas envolvidos na análise de regressão são: (i) obter pontos e estimar parâmetros de modelos de regressão; (ii) testar hipóteses sobre os parâmetros calculados; (iii) determinar o quão adequado é o modelo; (iv) verificar um conjunto de suposições ditas relevantes;

As tarefas de regressão podem ser agrupadas em regressão linear e não linear. Na regressão linear, assume-se que a relação entre as variáveis pode ser descrita como uma função linear $y = a_0 + a_1X_1 \cdots + a_nX_n$ tal que $a_i \in \Re$ e X_i é uma variável real. Quando é detectado não haver uma relação linear entre duas variáveis X e Y, a regressão pode ser

modelada como uma função não linear. Alguns exemplos de modelos não lineares, são: (i) modelo de regressão exponencial representado da forma $y = ae^{bx} + \psi$; (ii) modelo de regressão logística representado da forma $y = \frac{a}{1+be^{cx}} + \psi$;

Seja C_1, C_2 ... C_t um conjunto de rótulos que descrevem as categorias em que os objetos de um determinado domínio podem pertencer. O problema da classificação, segundo Mitchell (1997), diz respeito a determinação da classe de uma amostra/objeto a partir da sua descrição (atributos observados). Na agricultura, procedimentos de classificação automática tem sido usados para classificar doenças na cultura pelos sintomas (EL-TELBANY; WARDA; EL-BORAHY, 2006), classificação de manejo de solo por imagens (MARTINS; FONSECA, 2009), classificação de tipos de solo (VAMANAN; RAMAR, 2011), dentre outras aplicações possíveis.

Existem diversas técnicas para realizar o processo de classificação. Este trabalho considera duas abordagens: (i) a abordagem bayesiana, que consiste em calcular a probabilidade de cada hipótese dado um conjunto de dados e, então, classificar uma instância na categoria mais provável. (ii) a abordagem por redes neurais artificiais, que consiste em gerar uma estrutura com várias unidades de processamento, criando uma função que associa uma instância a uma categoria.

Para testar a efetividade dos procedimentos de seleção descritos no Capítulo 3, este trabalho utiliza Regressão Linear Múltipla, classificação por Redes Neurais Artificiais do tipo Perceptron de Múltiplas Camadas e classificação por Redes Bayesianas com o classificador *Naive Bayes* para ilustrar o impacto da seleção de variáveis sobre o resultado dos modelos. Assim, esta seção está dividida em três subseções: Seção 2.3.1, que apresenta a Regressão Linear Múltipla, Seção 2.3.2, que apresenta o classificador *Naive Bayes* e Seção 2.3.3, que apresenta a Rede Neural Artificial do tipo Perceptron de Múltiplas Camadas.

2.3.1 Regressão Linear Múltipla

A regressão linear objetiva determinar uma função linear que expresse (exata ou aproximadamente) o relacionamento entre duas [ou mais] variáveis. O caso mais simples da regressão linear é aquele que envolve apenas duas variáveis, X e Y. Ao observar um gráfico de dispersão que apresenta $X \times Y$, com n observações, é possível verificar como Y varia em função dos valores de X. Quando a relação entre as variáveis é representada por uma linha reta, diz-se que a equação que a representa é um modelo de regressão linear. Ainda, se a função descreve Y dependente apenas de X, diz-se que a equação representa um modelo de regressão linear simples (CARVALHO, 1946; AFIFI; AZEN, 1979; NETO, 2002).

Quando se estabelece uma função para representar os pontos amostrais de Y em função de X, é importante definir a capacidade da função em estimar (predizer) os valores corretamente, a fim de fortalecer a certeza da relação linear entre as variáveis. Para tal, pode-se calcular o coeficiente de correlação de Pearson r que pode ser estimado pela Expressão 1. O coeficiente de correlação de Pearson é uma das possíveis maneiras de medir a força e a direção do relacionamento linear entre duas variáveis X e Y. O valor de r varia entre -1 e 1. Quando $r = -1$, ele indica correlação linear negativa perfeita, o que implica que se X aumenta, Y diminui. Quando $r = 1$, ele indica correlação linear positiva perfeita, o que implica que se X aumenta, Y também aumenta. Quando $r \approx 0$, ele indica que não há correlação linear entre as variáveis, o que significa que Y varia aleatoriamente em relação a X.

$$r = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}} \tag{1}$$

Quando se pretende estimar a habilidade de um modelo em predizer valores da variável de resposta, pode-se utilizar o coeficiente de determinação R^2. Esse coeficiente é o quadrado do coeficiente de correlação amostral ($R^2 = r^2$) e apresenta valores entre 0 e 1. O R^2 expressa o quanto da variável dependente é explicada pelas variáveis independentes.

Assumindo que a relação entre Y e X é definida aproximadamente por uma reta, a regressão entre as duas variáveis é realizada pelo modelo:

$$Y = a + bX \tag{2}$$

Em que a e b são coeficientes, X é a variável independente e Y é a variável dependente. Sobre a Expressão 2, é possível utilizar o método dos mínimos quadrados (SOUNIS, 1971; AFIFI; AZEN, 1979; NETO, 2002) para minimizar erro entre a reta estimada e os pontos observados. Então, resolve-se o sistema de duas equações e duas variáveis:

$$\begin{cases} a &= \bar{Y} - b * \bar{X} \\ b &= \frac{\sum(X_i * Y_i) - n * \bar{X} * \bar{Y}}{\sum X_i^2 - n * \bar{X}^2} \end{cases} \tag{3}$$

Este sistema de equações permite calcular os coeficientes a e b, da Equação 2, diretamente dos dados amostrais. Considerando que no Sistema 3, $\bar{X}$ e $\bar{Y}$ são as médias das variáveis X e Y, respectivamente.

Quando a variável dependente é predita por mais de uma variável independente,

tem-se um problema de regressão linear múltipla (MRLM). Wooldridge (2009) apresenta uma maneira de estimar os coeficientes de um MRLM com o método dos mínimos quadrados aplicado sobre uma matriz de dados. Seja $\mathbf{Y}$ uma matriz $n \times 1$ com n observações. Seja $\mathbf{X}$ uma matriz de dados $n \times k$, com n observações e k variáveis. A matriz $\mathbf{X}$ é representada como:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{22} & x_{23} & \cdots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix} \tag{4}$$

Então, o modelo de regressão linear múltipla pode ser representado por:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U} \tag{5}$$

Nesta equação, $\mathbf{B}_{k \times 1}$ representa os parâmetros do modelo com k coeficientes. E $\mathbf{U}_{n \times 1}$ é uma matriz de ruídos presentes nas observações[2]. Seja $\mathbf{X}_i$ a i-ésima observação de $\mathbf{X}$ (ou seja, $\mathbf{X}_i$ é uma linha da matriz, ou um registro do conjunto de dados dados, $\mathbf{X}_{i2}, \mathbf{X}_{i3}, ..., \mathbf{X}_{ik}$) e $\mathbf{Y}_i$ a i-ésima observação de $\mathbf{Y}$. O método dos mínimos quadrados aplicado à Expressão 5 fornece a soma dos resíduos quadrados como:

$$\mathbf{S} = \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i \mathbf{B})^2 \tag{6}$$

Então, deve-se minimizar $\mathbf{S}$ de modo a satisfazer:

$$\frac{\partial \mathbf{S}}{\partial \mathbf{B} \equiv 0} \tag{7}$$

Então, para estimar os parâmetros de um MRLM, deve-se empregar:

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{8}$$

[2]Segundo Libralon (2007), o ruído pode ser definido como uma instância que aparenta ser inconsistente com o padrão de comportamento das demais instâncias.

2.3.1.1 Exemplo de Uso de Modelos de Regressão na Agricultura

Um exemplo de uso do MRLM pode ser observado no trabalho de Molin *et al.* (2001), que usaram modelos de regressão múltipla para determinar possíveis causas da variação no rendimento de grãos em função de parâmetros sobre a fertilidade do solo. No trabalho de Molin *et al.* (2001), coletara-se amostras de solo que foram analisadas em laboratório para determinar o teor de cada elemento.

Foram coletadas amostras sobre os atributos: Fósforo, Matéria Orgânica, pH, Hidrogênio + Alumínio, Potássio, Cálcio, Magnésio, Saturação por Base, Capacidade de Troca de Cátions e Soma de Bases. Em cada ponto, além dos atributos mencionados, foi amostrado o rendimento de grãos de uma cultura de Soja.

Os desempenhos foram expressados pelo coeficiente de determinação (R^2). Molin *et al.* (2001) aplicaram diversos modelos de regressão múltipla, dentre os quais, foi utilizado o MRLM.

Na análise dos resultados, eles verificaram que os modelos pouco correlacionavam com as amostras e concluiram que: o rendimento tem variabilidade local dependente da área amostrada, a variabilidade do solo tem sua variabilidade local relacionada à profundidade do solo e as limitações no rendimento de grãos da cultura podem estar atribuídas à fertilidade do solo por relações mais complexas.

2.3.2 Naive Bayes

O classificador *Naive Bayes* (NB) permite estimar a probabilidade de uma instância pertencer a uma determinada categoria. Esse tipo de classificador assume a independência condicional entre os atributos, ou seja, o valor de um atributo em uma determinada categoria independe do valor dos demais atributos. Esta suposição simplifica os cálculos envolvidos na classificação de uma instância, bem como facilita a estimação dos parâmetros do modelo por métodos de mineração de dados (HAN; KAMBER; PEI, 2011). A seguir, é apresentado o algoritmo do classificador NB.

Seja X um conjunto de amostras, ou em termos probabilísticos, evidências. Seja C um conjunto de categorias $C_1, C_2, ..., C_t$, de tal forma que X pertence a uma categoria C_i. Para problemas de classificação, deseja-se determinar a probabilidade $P(C|X)$ da classe C dado as evidências X. Então, verifica-se a probabilidade de X pertencer à uma das possíveis categorias de C, dado que é conhecida a descrição dos atributos de X. A probabilidade $P(C|X)$ é a probabilidade *a posteriori* de C condicionada a X. $P(C)$ é a

probabilidade *a priori* de C. A probabilidade *a priori* pode ser calculada dos dados, caso o pesquisador não saiba informá-la. A probabilidade *a posteriori* é calculada pelo teorema de Bayes:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{9}$$

O classificador NB é treinado da seguinte forma (HAN; KAMBER; PEI, 2011):

1. Seja D um conjunto de dados p-dimensional, de modo que as amostras são dispostas em registros $X = (x_1, x_2, ..., x_n)$, dos respectivos atributos $A_1, A_2, ..., A_n$;

2. Seja t o número de classes, $C_1, C_2, ..., C_t$. Dado uma amostra X, o classificador prediz se X pertence a uma classe C_i se, e somente se, $P(C_i|X) > P(C_j|X)$, para $1 \leq j \leq t, j \neq i$. Respeitando essa restrição, $P(C_i|X)$ é maximizada. Então, $P(C_i|X)$ é calculada aplicando a Expressão 9;

3. Sendo $P(X)$ constante para todas as classes, apenas $P(X|C_i)P(C_i)$ precisa ser maximizada. Se as probabilidades $P(C_i)$ não são conhecidas, elas podem ser calculadas do conjunto de dado por $P(C_i) = |C_{i,D}|/|D|$ (o número de registros C_i em D dividido pelo número de registros em D), no que resulta $P(C_1) = P(C_2) = ... = P(C_k)$. Então, maxima-se apenas $P(X|C_i)$;

4. Neste passo, assume-se a independência condicional entre as variáveis e, considerando que x_k é um valor da variável A_k efetua-se:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{10}$$

verificando-se:

(i) Se A_k é discreta, $P(x_k|C_i)$ é calculado dividindo o número de registros da classe C_i em D, que possuem o valor x_k em A_k, pelo número de registros C_i em D.

(ii) Se A_k é contínua, assume-se que ela possui distribuição gaussiana com média μ e desvio padrão σ, definida por $g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, de modo que $P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$. Os parâmetros μ_{C_i} e σ_{C_i} são estimados dos próprios dados contidos em A_k pertencentes à C_i.

Figura 1: Algoritmo de treinamento do classificador *Naive Bayes*

Calculadas as tabelas de probabilidades, o classificador NB pode ser utilizado para predizer uma provável categoria para X. Efetuando $P(X|C_i)P(C_i)$ para cada cate-

goria C_i, o classificador prediz que o registro X classifica-se como C_i se, e somente se, $P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$, para $1 \leq j \leq t, j \neq i$. Assim, o registro X pertence à C_i, quando esta é a categoria que apresenta maior probabilidade em relação às demais categorias.

2.3.2.1 Exemplo de aplicação do classificador *Naive Bayes*

Exemplo de aplicação de classificadores NB na Agricultura pode ser observado no trabalho de Vamanan e Ramar (2011), que empregaram classificadores bayesianos com a finalidade de verificar uma possível melhoria nos sistemas de uso e manejo do solo nas áreas de agricultura, horticultura e de questões ambientais. Os autores empregaram os classificadores *Naive Bayes*, *Naive Bayes* Atualizável e Rede Bayesiana para classificar tipos de solos no distrito de Kanchipuram, Índia. Tais classificadores foram comparados a classificadores baseados em árvores de decisão (J48 e *Random Forest*) e a métodos estatísticos usuais para a classificação de solos por engenheiros especializados em geologia.

O conjunto de dados utilizado continha os atributos de diagnóstico dos horizontes do solo superficial e sub superficial, regimes de umidade e temperatura do solo e propriedades físico-químicas do solo. Esses atributos foram utilizados para classificar o solo em oito categorias, de acordo com o manual do governo Indiano.

Os resultados mostraram que os classificadores baseados em *Naive Bayes* foram aqueles que forneceram menores erros e com maiores taxas de acerto. Vamanan e Ramar (2011) concluiram que a aplicação de técnicas de mineração de dados aplicadas a perfis de solo podem melhorar a verificação de perfis válidos de solo, bem como a validação de padrões e classificação dos perfis.

2.3.3 Redes Neurais Artificiais

A unidade de processamento básica de uma Rede Neural Artificial (RNA) é chamada neurônio artificial e sua estrutura é representada na Figura 2, por meio de um grafo dirigido e ponderado. Neste grafo os nós $X_1, ..., X_n$ são ditos nós de entrada e permitem a inserção de valores para processamento no neurônio. O nó U executa o processameno dos dados de entrada que são transmitidos a ele pelas arestas (X_i,U), $i = 1..n$. Cada aresta (X_i,U) armazena um peso numérico $w_{i,k}$, chamado peso sináptico, que pondera a influência do valor de uma entrada para o posterior processamento por U. A aresta de saída informa o resultado do processamento das entradas $X_1, ..., X_n$ pela função de transferência (ou função de ativação) codificada em U.

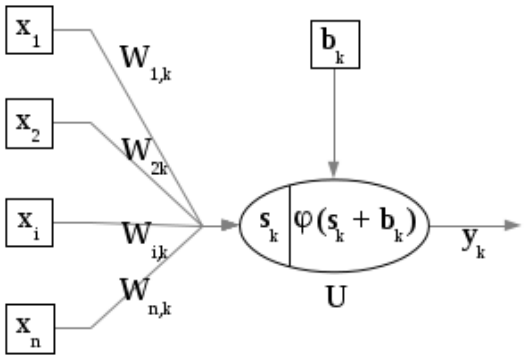

Figura 2: Exemplo de um neurônio artificial

Seja $(x_1..., x_n)$ uma n-upla que descreve um caso (sinal) que deve ser processado pelo neurônio artificial, o processamento se dá da seguinte forma: em primeiro lugar, as variáveis de entrada $X_1..., X_n$ recebem os valores de $(x_1..., x_n)$; em segundo, estes valores são enviados para o nó U por meio das conexões de entrada; em terceiro lugar, o operador de somatório, no interior de U, calcula $s_k = \sum_{i=1}^{n} w_i x_i$; finalmente, o resultado s_k é aplicado sobre a função de ativação φ, o que gera o valor de $y_k = \varphi(s_k)$. Uma vez determinado o valor da saída y_k (sinal de saída) o mesmo pode ser interpretado com a finalidade de regressão ou classificação sobre uma instância do conjunto de dados. Deve ser notado que o neurônio artificial pode conter um viés[3] b_k que ajusta o valor do somatório sobre a função de ativação tal que $\varphi(S_k + b_k)$.

A rede empregada neste trabalho é do tipo acíclica (ou com *encaminhamento para frente*) com múltiplas camadas. Neste tipo de rede, as unidades de processamento são organizadas em camadas - existe uma entrada, uma ou mais camadas ocultas e uma camada de saída. A camada de entrada tem um neurônio de entrada para cada atributo do conjunto de dados a ser processado. O processamento de uma instância ocorre ao longo da rede com a transmissão dos sinais, desde os neurônios da camada de entrada até a camada de saída, que pode funcionar como um regressor ou um classificador. Como regressor, a saída possui um neurônio artificial com uma função de ativação igual a uma função linear. Como um classificador, a saída funciona expressando valores que identificam categorias.

A Figura 3 ilustra uma rede neural acíclica de múltiplas camadas, com uma camada de entrada, uma camada oculta e uma camada de saída. As entradas 1 e 2 representam as variáveis explicativas, cujos neurônios não possuem função de ativação. Os neurônios artificiais 3 e 4, são aqueles que realizam o processamento dos valores de entrada e produzem suas saídas de acordo com uma função de ativação. O neurônio 5, processa os valores dos neurônios artificiais da camada anterior (3 e 4) e produz o valor

[3]Do inglês, *bias*.

de resposta da rede, também, conforme função de ativação.

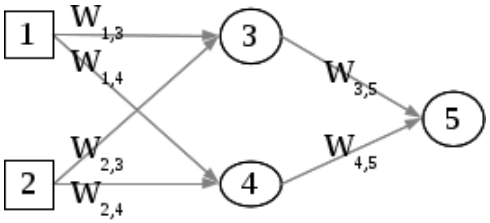

Figura 3: Exemplo de uma rede neural artificial

A função de ativação de um neurônio artificial é representada por uma função que recebe o somatório s_k como parâmetro, mas dependendo da implementação da rede, ela recebe $s_k + b_k$. Segundo Russell e Norvig (2004), a função de ativação é desenvolvida para suportar dois objetivos: 1) ela é projetada para retornar 1 quando a entrada deve ser considerada correta e 0 para quando a entrada deve ser considerada errada; 2) ela precisa ser não linear, senão, a rede neural corresponde a uma função linear comum. Russell e Norvig (2004) comentam a respeito de duas funções de ativação: a função limiar (Figura 4(a)) e a função sigmoide (Figura 4(b)). A função limiar retorna 1 quando o parâmetro de entrada é positivo e retorna 0 em caso contrário. Já a função sigmoide segue a expressão $\varphi(x) = 1/(1 + e^{-x})$. A vantagem desta última função é ela permitir que o neurônio tenha comportamento não linear que, segundo Haykin (2001), Russell e Norvig (2004), permite a rede neural regredir ou representar funções não lineares.

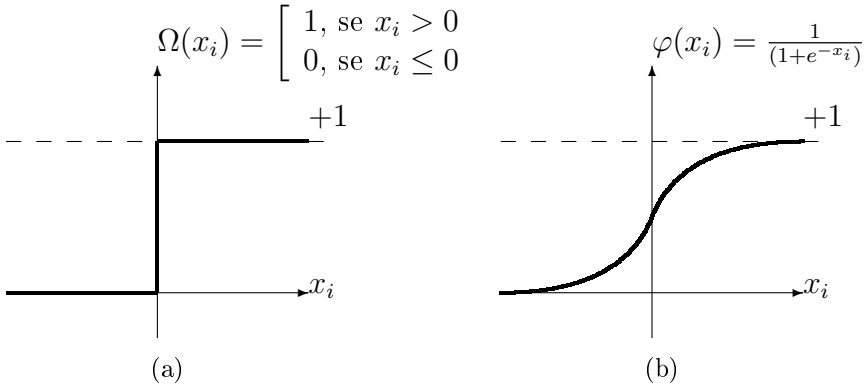

Figura 4: (a) Função limiar; (b) Função sigmoide

Em uma RNA o conhecimento sobre os relacionamentos entre as variáveis são especificados nos pesos das conexões entre as unidades de processamento. Portanto, uma tarefa importante quando da construção deste tipo de modelo é o ajuste dos pesos sinápticos de forma que a rede codifique corretamente as relações do domínio.

Para extrair o conhecimento de um conjunto de dados de modo a armazená-lo na

estrutura da rede, um algoritmo de treinamento deve ser utilizado. Diz-se que um algoritmo de treinamento é quem faz uma rede neural artificial *aprender* os padrões contidos nos dados para que seja possível realizar, posteriormente, tarefas como regressão ou classificação com instâncias inéditas (que ainda não foram apresentadas à rede). Esta etapa de ajuste é chamada de treinamento da rede e ocorre após a definição da topologia do grafo e à escolha das funções de ativação. Quando um pesquisador modela uma RNA, ele considera as variáveis explicativas como neurônios de entrada e a variável de resposta como neurônios de saída. Não há uma regra definida para estabelecer uma arquitetura consistente, ou seja, para calcular quantas camadas ocultas devem ser utilizadas e quantos neurônios devem ser empregados em cada camada oculta. Assim, o pesquisador deve testar várias arquiteturas de acordo com uma metodologia que permita descobrir uma rede neural que apresente resultados consistentes.

Durante o treinamento de uma rede neural é usual que os pesos sinápticos sejam inicializados com valores aleatórios e então um algoritmo de aprendizagem executa o treinamento diversas vezes sobre o mesmo conjunto de dados (MITCHELL, 1997; HAYKIN, 2001; RUSSELL; NORVIG, 2004). A execução do treinamento repetidas vezes objetiva fazer com que os pesos sinápticos venham a convergir para valores que permitam a rede classificar corretamente as intâncias de treinamento. O número de vezes que a rede deve ser treinada é conhecido como número de épocas e pode ser estabelecido pelo pesquisador ou por critérios estatísticos.

2.3.3.1 Modelo Perceptron de Múltiplas Camadas e o Algoritmo de Retropropagação de Erro

Haykin (2001) explica que um modelo de perceptron multicamadas é uma rede com encaminhamento para frente, que consiste de camada de entrada, uma ou mais camadas ocultas e uma camada de saída. A função de ativação utilizadas nos neurônios artificiais é a função sigmoide. As redes perceptrons multicamadas são treinadas por meio do algoritmo de retropropagação do erro. Tal algoritmo é baseado em regras de aprendizado por correção de erros, que é como uma generalização do método dos mínimos quadrados para estimar os coeficientes de uma rede neural. O treinamento é realizado de maneira supervisionada, o que significa que a saída desejada é apresentada à rede juntamente com os dados de entrada, assim, cada resposta estimada pela rede pode ser comparada com a resposta informada e o erro pode ser calculado para ajustar os pesos sinápticos.

O algoritmo de retropropagação de erros realiza o ajuste dos pesos de uma RNA

com base no gradiente de erro. Seja $j = 1, 2, ..., J$ as camadas ocultas da RNA e que $j = 0$ é a camada de entrada (que não possui função de ativação), $k = 1, 2, ..., m_j$ os neurônios da camada j, sendo $m_1, m_2, ..., m_J$ o número de neurônios em cada camada j. A saída $o_k(n)$ é a resposta da função de ativação do neurônio k, especificamente, na camada de saída J na época n. A saída $y_k^{(j)}(n)$ é a saída da função de ativação do neurônio k na camada oculta j em uma época de treinamento n. O valor $d_k(n)$ é o valor observado que se deseja obter na época n.

O algoritmo de aprendizado da retropropagação, segundo Haykin (2001), tem seus passos descritos na Figura 5.

2.3.3.2 Redes Neurais Artificiais Aplicadas à Agricultura - um estudo de caso

Mathias (2006) avaliou o desempenho de redes neurais no reconhecimento de padrões de comportamento de variáveis meteorológicas, relacionando-as com molhamento foliar por orvalho. Naquele trabalho, testou-se o uso de redes neurais do tipo perceptron de múltiplas camadas treinadas por dois diferentes algoritmos de aprendizagem: o algoritmo de retropagação de erro (5) e a retropropagação resiliente (RIEDMILLER; BRAUN, 1993). As redes foram testadas em três conjuntos de dados de modo a possuir entre 2 e 13 neurônios de entrada, sendo as variáveis empregadas:

Tabela 1: Variáveis do conjunto de dados do trabalho de Mathias (2006)

Nome da variável	Unidade de medida
Umidade relativa mínima do ar	%
Umidade relativa média do ar	%
Umidade relativa máxima do ar	%
Temperatura mínima do ar	°C
Temperatura média do ar	°C
Temperatura máxima do ar	°C
Temperatura de bulbo seco	°C
Temperatura de bulbo úmido	°C
Velocidade mínima do vento	m/s
Velocidade média do vento	m/s
Velocidade máxima do vento	m/s
Pressão atmosférica	kPa
Precipitação pluviométrica	mm

Foi utilizada apenas uma camada oculta e o número de neurônios foi calculado por $3e + s$, em que e representa o número de neurônios na camada de entrada e s, o número de neurônios na camada de saída. Para o estudo de caso 1, a camada de saída foi projetada para 10 neurônios, um para cada categoria da variável de resposta. Para

1. **Inicialização dos pesos:** inicializar os pesos com valores aleatórios;

2. **Apresentação as amostras de treinamento:** Apresentar à RNA o conjunto de dados para treinamento. Para cada exemplo do conjunto de dados, realizar a sequência dos Passos 3 e 4;

3. **Propagação para frente:** considere uma amostra de treinamento como sendo o par (x, d), com o vetor de entrada x aplicado na camada de entrada e o vetor de saída d aplicado na camada de saída. Compute os sinais pelas funções de cada neurônio da rede procedendo o passo de propagação para frente, camada por camada. Então, a saída de um neurônio k de uma camada oculta j é dada por

$$s_k^{(j)}(n) = \sum_{i=0}^{m_j} w_{k,i}^{(j)}(n) y_i^{(j-1)}(n)$$

em que $y_i^{(j-1)}(n)$ é a saída do neurônio i na camada $(j-1)$ e $w_{k,i}^{(j)}(n)$ é o peso sináptico do neurônio k na camada j provido pelo neurônio i da camada $(j-1)$. Quando $i = 0$, $y_0^{(j-1)}(n) = 1$ e $w_{k,0}^{(j)}(n) = b_k^{(j)}(n)$, que é o viés aplicado ao neurônio k da camada j na época n. Assumindo que os neurônios utilizam uma função sigmoide φ, a saída de um neurônio k em uma camada oculta j é $y_k^j(n) = \varphi_k(s_k(n))$. Se o neurônio k está na primeira camada oculta $(j = 0)$, então, $y_k^0(n) = x_k(n)$. Se o neurônio k está na camada de saída, então, $y_k^J(n) = o_k(n)$. Então, calcular o erro efetuando $\varepsilon_k(n) = d_k(n) - o_k(n)$.

4. **Propagação para trás:** computar o gradiente local δ definido por

$$\delta_k^{(j)} = \begin{bmatrix} e_k^{(J)} \varphi_k(s_k^{(j)}) & \text{para um neurônio da camada de saída } J \\ \varphi_k(s_k^{(j)}) \sum_k \delta_k^{j+1} w_{i,k}^{(j+1)} & \text{para um neurônio da camada oculta } j \end{bmatrix}$$

Então, atualizar os pesos sinápticos por meio da regra delta generalizada:

$$w_{k,i}^{(j)}(n+1) = w_{k,i}^{(j)}(n) + \alpha[w_{k,i}^{(j)}] + \eta \delta_k^{(j)}(n) y_k^{(j-1)}(n)$$

em que η é a taxa de aprendizado e α é a contante de momento.

5. **Iteração:** iterar os Passos 3 e 4 em novas épocas de treinamento até atingir um critério de parada, que pode ser: atingir a um número máximo de épocas de treinamento ou atingir um erro menor do que um erro máximo informado pelo usuário.

Figura 5: Algoritmo da retropropagação de erro (*Backpropagation*)

os estudos de casos 2 e 3, o número de neurônios na camada de saída foram 2: um para seco e outro para molhado. Após testar diversas redes, foi observado que a menor taxa de acerto obtida nos estudos de caso foi 77%, sendo as melhores redes aquelas com 96%, 80,86% e 96,4% para os estudos 1, 2 e 3 respectivamente.

Foi concluído que as RNAs reconhecem eficientemente os padrões em conjuntos de dados sobre molhamento foliar. Então, elas podem ser utilizadas como uma ferramenta para definir as relações entre variáveis meteorológicas e molhamento foliar por orvalho.

2.4 ANÁLISE DE COMPONENTES PRINCIPAIS

Esta seção apresenta a análise de componentes principais, uma técnica de estatística multivariada cujo objetivo é reduzir a dimensionalidade dos dados (JOLLIFFE, 2002). Na análise de componentes principais, um conjunto de dados é submetido a uma transformação linear, que rotaciona os eixos coordenados do conjunto de dados sobre novos eixos que são combinações lineares das variáveis originais (FERREIRA, 2008). Essas combinações lineares são os componentes principais e são calculados de modo a reter o máximo possível da variação presente nos dados. Adicionalmente, os componentes principais são descorrelacionados entre si e são organizados de maneira que os primeiros poucos componentes retêm a maior parte da variação do conjunto.

Assim, seja $\mathbb{X}$ um vetor coluna com p variáveis. Para analisar a máxima variância entre os p elementos de $\mathbb{X}$ a ACP gera uma função linear $\alpha_1'\mathbb{X}$, que é o componente principal, em que α_1 é um vetor coluna de p coeficientes, $\alpha_{11}, \alpha_{12}, ..., \alpha_{1p}$. Assim:

$$\alpha_1'\mathbb{X} = \alpha_{11}\mathbb{X}_1 + \alpha_{12}\mathbb{X}_2 + \cdots + \alpha_{1p}\mathbb{X}_p = \sum_{i=1}^{p} \alpha_{1i}\mathbb{X}_i$$

O k-ésimo componente $\mathbb{Z}_k = \alpha_k'\mathbb{X}$, α_k é um autovetor calculado a partir da matriz de covariância populacional $\mathbf{\Sigma}$ e está associado k-ésimo autovalor λ_k (FERREIRA, 2008). Contudo, em casos práticos $\mathbf{\Sigma}$ pode ser substituída pela matriz de covariância amostral $\mathbf{S}$. Este trabalho considera que $\mathbb{Z}_k$ é obtido a partir $\mathbf{S}$. A variância de um componente k é dada por $var(\mathbb{Z}_k) = \lambda_k$, que pode ser expressa da forma $var(\alpha_1'\mathbb{X}) = \alpha_1'\mathbf{S}\alpha_1$. A maximização desta última forma faz com que a variância dos dados transformados seja maximizada para aquele componente $\mathbb{Z}_k$ e está sujeita à restrição $\alpha_1'\alpha_1 = 1$. Utilizando a técnica dos multiplicadores de Lagrange e diferenciando em relação à α_1, tem-se que:

$$(\mathbf{S} - \lambda_1 \mathbf{I}_p)\alpha_1 = 0$$

Nesta expressão, $\mathbf{I}_p$ é uma matriz identidade $p \times p$, λ_1 é um autovalor de $\mathbf{S}$ e α_1 é o autovetor correspondente. Então, maximiza-se $\alpha_1' \mathbf{S} \alpha_1 = \lambda_1$ de modo que λ_1 seja o maior possível.

Para calcular um segundo componente principal, deve-se considerar que α_2 não é correlacionado com α_1, então, qualquer uma das restrições $\alpha_1' \mathbf{S} \alpha_2 = 0$, $\alpha_2' \mathbf{S} \alpha_1 = 0$, $\alpha_1' \alpha_2 = 0$ ou $\alpha_2' \alpha_1 = 0$, podem ser utilizadas para a expressão da não correlação de $\alpha_1' X$ com $\alpha_2' X$. Repetindo todo o processo anterior, chega-se a $(\mathbf{S} - \lambda_2 \mathbf{I}_p)\alpha_2 = 0$, de modo que deve-se maximizar $\alpha_2' \mathbf{S} \alpha_2 = \lambda_2$, para que λ_2 seja o maior possível. Então, para p componentes, é possível calcular os autovetores $\alpha_3, \alpha_4, ..., \alpha_p$ de $\mathbf{S}$, obtendo seus respectivos autovalores $\lambda_3, \lambda_4, ..., \lambda_p$.

A porcentagem p_k de variação dos dados explicada por cada componente k é calculada por $p_k = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} \times 100$. A porcentagem acumulada de explicação pelos componentes de 1 até k é dada por $P_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100$ (DUNTEMAN, 1989; FERREIRA, 2008). Assim, é possível usar esta expressão para estipular um limiar L para determinar k tal que a explicação de $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k$ seja a explicação acumulada $P_k \leq L$. Isto permite a escolha de componentes que expliquem a variação dos dados observada em uma amostra.

2.4.1 Decomposição por valores singulares

Os autovetores e autovalores da ACP também podem ser computados por meio da técnica da decomposição por valores singulares (DVS) (JOLLIFFE, 2002; FERREIRA, 2008; KUMAR *et al.*, 2009). A DVS resolve a expressão:

$$\mathbf{X} = \mathbf{ULA}' \tag{11}$$

Nesta expressão, $\mathbf{X}$ é uma matriz de dados dimensões $n \times p$ (n amostras por p variáveis), $\mathbf{U}$ e $\mathbf{A}$ são as matrizes de dimensões $(n \times p)$ e $(p \times p)$, respectivamente, que possuem colunas ortonormais de modo que $\mathbf{U}'\mathbf{U} = \mathbf{I}_p$ e $\mathbf{A}'\mathbf{A} = \mathbf{I}_p$. $\mathbf{L}$ é uma matriz diagonal $(p \times p)$. Definindo $\mathbf{A}$ com colunas a_k, $\mathbf{U}$ com colunas $u_k = l_k^{-1/2} X a_k$, e $\mathbf{L}$ com elementos na diagonal $l_k^{1/2}$, com $k = 1, 2, ..., r$. A matriz $\mathbf{X}$ como pode ser decomposta como:

$$\mathbf{X} = \mathbf{ULA}' = \mathbf{U} \begin{bmatrix} l_1^{1/2} a_1' \\ l_2^{1/2} a_2' \\ \vdots \\ l_p^{1/2} a_p' \end{bmatrix} \tag{12}$$

Substituindo $\mathbf{U}$, obtém-se:

$$\mathbf{ULA}' = \sum_{k=1}^{p} \mathbf{X} a_k a_k' \tag{13}$$

Para interpretar a técnica DVS como análise de componentes principais por autovetores e autovalores, tem-se que o produto matricial $\mathbf{UL}$ representa a matriz de componentes principais, também chamada de projeções (ou dados rotacionados), em que cada coluna identifica um componente (equivalente a $\mathbb{Z}_k$ da Seção 2.4) e cada linha identifica um registro de dados. A matriz $\mathbf{A}$ representa a matriz de autovetores, também chamada de matriz de cargas, ou matriz de rotação, em que cada coluna representa um vetor coluna de coeficientes (equivalente a α_k da Seção 2.4). A matriz $\mathbf{L}$ é uma matriz diagonal constituída da raiz quadrada dos autovalores (equivalente a Λ da Seção 2.4).

2.5 ANÁLISE DE COMPONENTES PRINCIPAIS SUPERVISIONADA

Esta seção apresenta uma abordagem para empregar a supervisão na Análise de Componentes Principais. Um conjunto com dados rotulados é aquele que contém os valores observados das variáveis explicativas associados a valores da variável de resposta (HAYKIN, 2001). A principal motivação para se explorar dados rotulados na ACP tem sido concentrar a análise nas variáveis que têm maior influência sobre a variável de resposta.

Bair *et al.* (2006) apresentaram uma abordagem para utilizar dados supervisionados na ACP. A proposta é executar uma etapa de pré-processamento cujo objetivo é selecionar as variáveis que seriam usadas para construir um modelo de regressão. Assim, os autores sugerem que ao invés de executar o procedimento da ACP em todas as variáveis da base de dados, a mesma seja empregada sobre um subconjunto das mesmas, mais especificamente, sobre aquelas que possuem maior correlação com a variável de saída.

Seja $\mathbf{X}$ uma matriz de dados em que $\mathbf{X}_i$ representa uma coluna (variável) e cada linha representa uma observação. Seja $\mathbb{Y}$ a variável de resposta, representada por um vetor em que cada linha de $\mathbb{Y}$ corresponde à uma linha de $\mathbf{X}$. O procedimento supervisionado

descrito por Bair *et al.* (2006) calcula, no passo 1, o coeficiente $\mathbb{Q}_i$ de regressão padronizados da variável explicativa $\mathbf{X}_i$ em relação à variável de resposta $\mathbb{Y}$, conforme Expressão 14. No passo 2, é determinado um limiar L em valor absoluto sobre os dados, em relação a $\mathbb{Q}$, por meio do processo de validação-cruzada. Então, comparam-se os valores do vetor $\mathbb{Q}$ com o limiar L de modo que aquela variável cujo valor $\mathbb{Q}_i \geq L$ é retida e aquela que possuir $\mathbb{Q}_i < L$ é eliminada do conjunto de dados. Assim, as variáveis retidas do conjunto $\mathbf{X}$ formam o conjunto $\mathbb{R}$. No passo 3, calcula-se a ACP sobre o subconjunto $\mathbb{R}$. Finalmente, no passo 4, os p componentes principais suficientes para a explicação dos dados são utilizados para compor o modelo de regressão.

$$\mathbb{Q}_i = \frac{\mathbf{X}_i' \mathbb{Y}}{\sqrt{\mathbf{X}_i' \mathbf{X}_i}} \tag{14}$$

O procedimento supervisionado pode ser descrito como segue (BAIR *et al.*, 2006):

1. Calcule um coeficiente $\mathbb{Q}_i$ para cada variável (maneira univariada);
2. Forme uma matriz reduzida consistindo apenas nas variáveis cujos seus coeficientes $\mathbb{Q}_i$ univariados excedem um limiar L em valor absoluto.
3. Calcule o primeiro (ou os poucos primeiros) componentes principais da matriz de dados reduzidos;
4. Use os componentes principais obtidos no modelo de regressão para predizer a saída.

A ACP supervisionada pode ser vista como uma maneira de identificar os agrupamentos com preditores relevantes, por seleção baseada em escores, para remover as origens de variações irrelevantes. A habilidade da ACP supervisionada de construir um modelo baseado em um conjunto de dados limitado, ou seja, com poucas observações, é importante para aplicações práticas (BAIR *et al.*, 2006). Além disso, isolar um subconjunto menor de variáveis para a ACP propõe um tipo de Navalha de Ockham, conforme Russell e Norvig (2004), provendo a elaboração de modelos menos complexos por meio da eliminação de variáveis que não influenciam na variável de resposta.

Pelos testes realizados no trabalho de Bair *et al.* (2006), é possível verificar que os resultados obtidos pela ACP supervisionada, demonstraram ter mais correlação com a variável de saída do que os resultados obtidos pela ACP não supervisionada. Os testes estatísticos mostraram que seu método supervisionado para a ACP consegue melhores preditores que a ACP convencional. Resultado este que atesta a efetividade de seu método

quando há a intenção em empregar a ACP para selecionar um conjunto de variáveis preditoras.

2.6 ANÁLISE DE COMPONENTES PRINCIPAIS PROBABILÍSTICA

Esta seção apresenta os conceitos da Análise de Componentes Principais Probabilística (ACPP) (BISHOP, 1999; YU *et al.*, 2006; ZHANG, 2009). Nesta abordagem os componentes principais são determinados com o algoritmo *Expectativa e Maximização (EM)* para indução de modelos. O seu uso permite que a Análise de Componentes Principais seja realizada mesmo quando o conjunto de observações não é completo. Mais detalhadamente, a ACPP permite que a matriz de carga $\mathbf{A}$ seja estimada mesmo quando existem dados incompletos[4].

A Análise de Componentes Principais Probabilística relaciona uma matriz de variáveis $\mathbf{X}$ com um conjunto de variáveis latentes $\mathbf{Z}$ pela expressão:

$$\mathbf{X} = \mathbf{AZ} + \mu + \epsilon \tag{15}$$

Na Expressão 15, $\mathbf{X}$ é a matriz de variáveis observadas de $d \times N$ dimensões em que N é o número de observações e d o número de atributos medidos; $\mathbf{Z}$ é a matriz dos valores dos componentes principais, também considerada matriz de variáveis latentes, de $d \times N$ dimensões; $\mathbf{A}$ é a matriz ortogonal de cargas e direções das projeções de $d \times d$ dimensões; μ é o vetor de médias[5] das variáveis. Deve ser observado que para Bishop (1999), as linhas representam as variáveis e as colunas representam os valores observados. O símbolo ϵ representa o erro.

A abordagem probabilística para ACP assume que o ruído ϵ tem uma distribuição gaussiana isotrópica da forma $N(0, \sigma^2 I)$ e que a priori das variáveis latentes $\mathbf{Z}$ também assume uma gaussiana representada por $p(\mathbf{Z}) = (2\pi)^{-q/2} e \left\{-\frac{1}{2}\mathbf{Z}'\mathbf{Z}\|\right\}$, em que q é o número de variáveis latentes e e é a base do logaritmo natural. Nesta modelagem é considerado, ainda, que o modelo de covariância das variáveis que definem $\mathbf{X}$ é dado por $\mathbf{C} = \mathbf{AA}' + \sigma^2\mathbf{I}$, em que $\mathbf{C}$ é uma matriz $d \times d$. Sob tais condições, a distribuição de probabilidade dos

[4]Bishop (1999) ainda comenta que outra vantagem é que a dimensão das variáveis latentes pode ser menor que a dimensão das variáveis originais. Isso, também, traz benefícios computacionais, pois o número máximo de componentes a serem calculados será menor do que o número de componentes calculados pela ACP convencional.

[5]A Análise Fatorial considera μ como sendo a média populacional, contudo este valor pode ser estimado pela média amostral. Deve-se notar que os pesquisadores, primeiramente, subtraem a média amostral das observações e assumem a média da distribuição igual a zero (ZHANG, 2009).

dados observados pode ser obtida integrando-se as variáveis latentes. O resultado desta marginalização também será uma gaussiana tal que $P(\mathbf{X}) \sim N(\mu, \mathbf{C})$.

As estimativas para os parâmetros $\mathbf{A}$ e σ^2, que estabelecem o relacionamento entre os atributos medidos e as variáveis latentes, podem ser obtidas pela maximização iterativa da Expressão 16 com o algoritmo *EM*. Naquela expressão, $\mathbf{S}$ é a matriz de covariância amostral[6].

$$L = -\frac{N}{2}[d\ ln(2\pi) + ln(|\mathbf{C}|) + tr(\mathbf{C}^{-1}\mathbf{S})] \tag{16}$$

O algoritmo *EM* é realizado em duas etapas (LUNA, 2004; MITCHELL, 1997):

1. **Passo de Estimativa:** calcular os valores esperados das estatísticas para os dados completos, dado os dados incompletos e as estimativas atuais dos parâmetros;

2. **Passo de Maximização:** utilizar estas estatísticas para fazer uma estimativa de máxima verossimilhança;

As duas etapas são repetidas seguidamente até o algoritmo convergir para um ponto estacionário da função de verossimilhança, que é descrita na Expressão 16.

O estimador de máxima verossimilhança 16 é maximizado por:

$$\mathbf{A}_{ML} = \mathbf{U}_q(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R} \tag{17}$$

Nesta expressão, $\mathbf{U}_q$ é uma matrix $d \times q$ com q vetores colunas de autovetores de $\mathbf{S}$, com autovalores correspondentes a $\lambda_1, \lambda_2, ..., \lambda_q$ na matriz λ e $\mathbf{R}$ é uma matriz arbitrária $q \times q$ de rotação ortogonal.

Para $\mathbf{A} = \mathbf{A}_{ML}$, o estimador de máxima verossimilhança para σ^2 é dado por:

$$\sigma^2_{ML} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j \tag{18}$$

Segundo Bishop (1999), na prática, para encontrar o modelo mais provável dado $\mathbf{S}$, primeiramente deve-se estimar σ^2_{ML} a partir da Expressão 18 e, então, $\mathbf{A}_{ML}$ a partir da Expressão 17. Ele comenta que para simplificar a Expressão 17 é possível substituir $\mathbf{R}$ por uma matriz identidade $\mathbf{I}$, já que $\mathbf{R}$ pode ser uma matriz arbitrária.

[6] A matriz de covariância amostral pode ser calculada por $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{X}_n - \mu_n)(\mathbf{X}_n - \mu_n)'$

2.7 COMPLEXIDADE DA AMOSTRA

Esta seção apresenta o conceito de complexidade da amostra aplicada sobre variáveis contínuas. Também, é apresentado o conceito da Teoria do Aprendizado Provavelmente Aproximadamente Correto (PAC). São, ainda, apresentadas as fórmulas necessárias para calcular o tamanho da amostra referente a cada um dos três modelos utilizados neste trabalho. No fim da seção são apresentados um exemplo para cada modelo, para enfatizar a importância do tamanho da amostra no processo de descoberta de conhecimento em bancos de dados.

Em procedimentos estatísticos é importante definir o tamanho do espaço amostral, pois, segundo Pillar (2004) a quantidade de amostras influencia na precisão requerida sobre o procedimento empregado. O tamanho da amostra deve ser suficiente para conferir precisão em uma estimativa. Pillar (2004) também explica que a quantidade de trabalho e materiais usados na coleta dos dados são determinados em função do tamanho da amostra. Assim, se o universo amostral é amplo, a suficiência amostral é uma ferramenta importante para o uso racional dos recursos de um projeto de pesquisa.

Em Aprendizado de Máquina, uma maneira de computar o tamanho da amostra é por meio da complexidade da amostra. Segundo Haykin (2001), a complexidade da amostra permite determinar quantas amostras um algoritmo de aprendizagem requer para aprender um conceito a partir de um espaço de hipóteses. Uma teoria que aborda essa questão é a do Aprendizado Provavelmente Aproximadamente Correto (PAC). Segundo Mitchell (1997), a aprendizagem PAC considera que um algoritmo de aprendizagem L aprende um conceito $c \in C$, por meio de um espaço de hipóteses H com probabilidade mínima de $(1 - \delta)$ e erro ε entre 0 e 1/2, e que L gera uma resposta de hipótese $h \in H$ de modo que o $erro_D(h) \leq \varepsilon$ em tempo polinomial sobre $1/\varepsilon$, $1/\delta$, n e tamanho de c. Segundo essa definição, é possível estabelecer uma estimativa do número de amostras m expresso por:

$$m \geq \frac{1}{\varepsilon}\left(4log\left(\frac{2}{\delta}\right) + 8VC(H)log\left(\frac{13}{\varepsilon}\right)\right) \tag{19}$$

Nesta expressão, ε é o erro com que o algoritmo L responde com uma hipótese h em relação a um conceito c que ele deve aprender dos exemplos de treinamento; δ é a significância do aprendizado, ou seja, $(1 - \delta)$ é a probabilidade com que modelo fornece uma resposta (hipótese h), sem conhecer a distribuição D das variáveis, cujo erro é menor do que ε; VC refere-se à dimensão Vapnik-Chervonenkis (VC). Segundo Mitchell (1997), a

dimensão VC mede a complexidade do espaço de hipóteses H, ou seja, ela mede o número instâncias distintas no conjunto de dados que podem ser completamente discriminadas usando H.

O Aprendizado PAC define uma maneira de calcular a complexidade da amostra em função do espaço de hipóteses de um dado modelo. O problema em calcular a complexidade da amostra recai sobre encontrar a dimensão VC do algoritmo L que se pretende utilizar. Como neste trabalho, são utilizados Modelo de Regressão Linear Múltipla, Redes Neurais Artificiais e classificador *Naive Bayes*, a seguir são apresentadas suas respectivas dimensões VC. De acordo com Shao, Cherkassky e Li (1969), a dimensão VC para um Modelo de Regressão Linear Múltipla pode ser expressa por:

$$VC(C) = d + 1 \tag{20}$$

Em que d é o número de variáveis (dimensões dos dados). A Expressão 20 indica que a dimensão VC para um modelo linear qualquer é expressa diretamente como o número de parâmetros livres. De acordo com Slattery e Craven (1998) a dimensão VC do classificador *Naive Bayes* é calculada da mesma maneira como é feito para um Modelo de Regressão Linear Múltipla. A dimensão VC para RNAs é expressa por Mitchell (1997):

$$VC(C) = 2\ s\ (r+1)\ log(e\ s) \tag{21}$$

Em que s é o número de nós na(s) camada(s) oculta(s), r é o número de entradas para cada nó da camada oculta e a constante e é a base do logarítmo natural. A seguir são apresentados um exemplo de uso da complexidade da amostra para cada um dos modelos utilizados neste trabalho.

Exemplo de complexidade da amostra para Regressão Linear Múltipla

Seja X um conjunto de dados com as variáveis explicativas X_1, X_2, X_3 e a variável de resposta Y. Seja o modelo expresso por $Y = C_0 + C_1X_1 + C_2X_2 + C_3X_3$. Neste caso, o número de amostras necessárias para garantir que o modelo seja estimado, atendendo as exigências do Aprendizado PAC com $\varepsilon = 0,1$ e $\delta = 0,05$, é 741 amostras.

Exemplo de complexidade da amostra para um classificador *Naive Bayes*

Seja X um conjunto de dados com as variáveis explicativas X_1 e X_2 e a variável de resposta Y. Seja o modelo expresso por $P(Y|X) = P(X|Y)P(Y)/P(X)$. Neste caso, são

necessárias 571 amostras para que o classificador *Naive Bayes* seja treinado, respeitando as exigências do Aprendizado PAC com $\varepsilon = 0,1$ e $\delta = 0,05$.

Exemplo de complexidade da amostra para RNA

Seja X um conjunto de dados com as variáveis explicativas X_1, X_2 e X_3 e a variável de resposta Y. Seja o modelo expresso por três variáveis na camada de entrada, dois neurônios em uma única camada oculta e um neurônio na camada de saída. Neste caso, são necessárias 2.093 amostras para treinar o classificador, segundo o Aprendizado PAC com $\varepsilon = 0,1$ e $\delta = 0,05$.

2.8 CONSIDERAÇÕES FINAIS

Como pôde ser observado, a análise multivariada tem sido empregada na agricultura. Contudo, seu emprego demanda que seja observado o tamanho da amostra. Quando este não é atingido, uma das possibilidades para contornar o problema é selecionar as variáveis mais relevantes, que são aquelas que mantêm a variação dos dados, para que o conjunto de dados represente aproximadamente a mesma distribuição, mas com menos variáveis. Uma estratégia é utilizar métodos de seleção de variáveis baseados em ACP, pois, eles podem auxiliar na seleção de variáveis, cuja finalidade é reduzir o conjunto de variáveis, mas mantendo a variação dos dados. Em relação ao tamanho da amostra, pode-se utilizar o Aprendizado PAC, que pode prover a complexidade da amostra necessária para inferir modelos multivariados sobre os dados.

Neste trabalho, o procedimento proposto por Bair *et al.* (2006) foi adaptado para que pudesse ser utilizado na seleção de variáveis. O coeficiente de regressão padronizado $\mathbb{Q}$ foi utilizado de modo que as variáveis com os menores coeficientes (aquelas que menos influenciam na resposta) foram removidas do conjunto. Então, a ACP foi realizada sobre o conjunto restante de variáveis $\mathbb{R}$, chamado de conjunto das variáveis selecionadas (retidas). A essência do procedimento é melhor explicada na Seção 3.1.

Considerando o exposto nesta revisão, o próximo capítulo apresenta dois procedimentos que permitem ao pesquisador de agricultura satisfazer as exigências da complexidade da amostra. Estes procedimentos fornecem uma maneira de selecionar as variáveis mais relevantes e que, neste contexto, são aquelas que mantêm elevada a porcentagem de explicação dos dados.

3 METODOLOGIA

Este capítulo descreve os procedimentos SVACP e SVACPS que exploram a análise de componentes principais e a análise de componentes principais supervisionada, respectivamente, no pré-processamento de bases de dados agrícolas com o objetivo de selecionar variáveis que serão usadas em tarefas de regressão e classificação. Ambos os procedimentos consideram que os dados pré-processados serão empregados nas tarefas de regressão de Modelos Lineares Múltiplos, treinamento de Redes Neurais Artificiais com topologia do tipo encaminhamento para frente e classifcadores to tipo *Naive Bayes*. Além disto, os procedimentos SVACP e SVACPS permitem a qualidade dos modelos gerados e as referentes à complexidade da amostra, sejam consideradas na tomada de decisão.

Também são descritos os experimentos que foram executados, com o objetivo de avaliar o desempenho dos procedimentos propostos para a seleção de variáveis. Os experimentos foram realizados sobre bases de dados sintéticas e sobre bases de dados agrícolas referentes à agricultura de precisão e à análise de fertilidade do solo. Os experimentos também visaram avaliar o desempenho dos procedimentos na seleção de variáveis em bases com dados incompletos.

Este capítulo está organizado da seguinte forma: a Seção 3.1 apresenta os procedimentos SVACP e SVACPS, a Seção 3.2 descreve os conjuntos de dados utilizados nos experimentos e a última seção (Seção 3.3) descreve os experimentos realizados.

3.1 PROCEDIMENTOS DE SELEÇÃO DE VARIÁVEIS: SVACP E SVACPS

3.1.1 Seleção de Variáveis por ACP

O procedimento de Seleção de Variáveis por ACP (SVACP) recebe um conjunto de dados e informações referentes a complexidade da amostra e, então, emprega os métodos B2 ou B4 para selecionar subconjuntos de variáveis para regressão ou classificação. Esta seleção é efetuada de modo a atender as exigências referentes à suficiência amostral e otimizar o desempenho do preditor ou classificador. O procedimento SVACP é apresentado na Figura 6 e como pode ser visto, ele é um procedimento iterativo que em cada iteração elimina um determinado número de variáveis do conjunto de dados e, então, testa o desempenho deste conjunto na regressão ou classificação de uma variável de resposta.

Mais detalhadamente, este procedimento tem como entrada a lista de parâmetros $(\mathbf{X}, B, \delta, \varepsilon, \mathcal{F}, \mathcal{T})$. O parâmetro $\mathbf{X}$ é o conjunto de dados no formato de uma matriz, cujas

- **Entrada:** Um conjunto de dados $\mathbf{X}$; número máximo B de variáveis a serem rejeitadas; probabilidade de Aprendizado PAC δ e o erro ε; fórmula $\mathcal{F}$ da dimensão VC para o modelo pretendido; função $\mathcal{T}$ que permite gerar os modelos de teste;
- **Declaração:** Número mínimo de variáveis A a serem rejeitadas; número de registros L do conjunto $\mathbf{X}$; número de variáveis D do conjunto $\mathbf{X}$; lista de complexidade da amostra $\mathbb{M}$, com D elementos; lista $\mathbb{G}$ que armazena subconjuntos de variáveis a serem rejeitadas; lista $\mathbb{R}$ que armazena subconjuntos de variáveis a serem retidas; lista $\mathbb{Z}$ de desempenhos; lista $\mathbb{U}$ de nomes das variáveis do conjunto $\mathbf{X}$;

1. Guardar o número de registros de $\mathbf{X}$ em L; guardar o número de variáveis de $\mathbf{X}$ em D; guardar o nome das variáveis de $\mathbf{X}$ em $\mathbb{U}$; inicializar $\mathbb{G}$, $\mathbb{R}$ e $\mathbb{Z}$ com conjuntos vazios;
2. Computar a complexidade da amostra por meio da função $\mathcal{F}$, supondo funçoes cujo número de variáveis de entrada seja $D, D-1, D-2, ..., D-B$, e inserir os valores em $\mathbb{M}$;
3. calcular o valor de A; para tanto, para $i = B$ até 1 passo -1 faça
   (a) se $\mathbb{M}_i > L$ então pare
   (b) $A = i$
4. Para n indo de A até B, faça:
   (a) Remover n variáveis utilizando um método de eliminação por ACP;
   (b) Inserir em $\mathbb{G}_n$ o grupo das n variáveis rejeitadas e inserir em $\mathbb{R}_n$ o grupo das $D - n$ variáveis retidas, considerando $\mathbb{R}_n = \mathbb{U} - \mathbb{G}_n$;
   (c) Construir um modelo com as $\mathbb{R}_n$ variáveis retidas e executá-lo para obter a medida de desempenho K;
   (d) Inserir K em $\mathbb{Z}$;
5. Retornar $\mathbb{Z}$, $\mathbb{R}$ e $\mathbb{M}$;

Figura 6: Método para a eliminação de variáveis por meio da ACPS

colunas representam as variáveis e as linhas representam os registros. O parâmetro B é o número máximo de variáveis a serem removidas do conjunto de dados. Os parâmetros δ e ε são a probabilidade δ de se gerar um modelo cujo erro seja menor ou igual ε, de acordo com o Aprendizado PAC. O parâmetro $\mathcal{F}$ define a função usada para determinar a dimensão VC do modelo utilizado e daí calcular a complexidade da amostra, segundo os conceitos do Aprendizado PAC. O último parâmetro $\mathcal{T}$ estabele o tipo de regressor ou classificador que será usado sobre os dados pré-processados.

O procedimento também declara algumas estruturas de dados utilizadas durante o processamento. Em particular, a declaração define um conjunto de listas que são utilizadas para armazenar o resultado do processamento. A lista $\mathbb{G}$ armazena conjuntos (listas) de variáveis rejeitadas, a lista $\mathbb{R}$ armazena conjuntos (listas) de variáveis selecionadas e a lista $\mathbb{Z}$ armazena os desempenhos dos modelos testados.

O passo 1 inicializa as demais variáveis do procedimento. Assim, as listas $\mathbb{G}$ (variáveis removidas), $\mathbb{R}$ (variáveis selecionadas) e $\mathbb{Z}$ (que armazena os valores obtidos pelos classificadores/regressores quando da remoção de determinado número de variáveis) são inicializadas vazias.

O passo 2 cria uma lista $\mathbb{M}$ que contém no máximo $B+1$ registros. Cada registro armazena um par $< b, \mathcal{F}(D-b) >$, onde $b = 0..B$. Isto é, a lista $\mathbb{M}$ armazena as exigências referentes a complexidade da amostra considerando que o número de variáveis selecionadas é $D-b$. Desta forma, esta lista mantém informações referentes a complexidade da amostra para todas as seleções de variáveis cuja cardinalidade estiver entre $D-B$ e D. Em seguida, o passo 3 determina o valor de A, o qual indica o número mínimo de variáveis que devem ser removidas para que sejam atendidas as restrições de complexidade da amostra. Para calcular A, determina-se o maior valor de índice i da lista de complexidades $\mathbb{M}$, tal que $\mathbb{M}_i \leq L$, e faz-se $A = i$.

O passo 4 realiza uma busca sequencial cujo espaço de busca é o número de variáveis a serem removidas. O objetivo desta busca é determinar as variáveis que devem ser retidas para que o desempenho do modelo gerado seja ótimo. Assim, a função que é otimizada neste passo é o escore do procedimento de treinamento $\mathcal{T}$. Para tanto, o passo 4a executa um método de eliminação por ACP sobre os dados em $\mathbf{X}$ e como resultado determina um vetor $\mathbb{G}_n$ com as variáveis removidas nesta iteração. Em seguida, no passo 4b o procedimento insere as variáveis rejeitadas no elemento $\mathbb{G}_n$ da lista $\mathbb{G}$. No mesmo passo o procedimento insere as variáveis que foram selecionadas nesta iteração no elemento $\mathbb{R}_n$ da lista $\mathbb{R}$. Estes elementos são as variáveis contidas no conjunto $\{\mathbb{U}-\mathbb{G}_n\}$. Na sequência o passo 4d utiliza $\mathbb{R}_n$ para gerar um regressor ou classificador usando $\mathcal{T}$. Este passo também calcula o desempenho K do conjunto selecionado, sobre o modelo gerado por $\mathcal{T}$, cujo valor é inserido em $\mathbb{Z}$ (passo 5).

O último passo 5 retorna as listas $\mathbb{Z}$, $\mathbb{R}$ e $\mathbb{M}$ que apresentam os resultados obtidos expressando, respectivamente: os desempenhos, grupos de variáveis selecionadas e complexidades da amostra.

Exemplo 1

Sejam z_1, z_2, z_3 e z_4, quatro variáveis aleatórias com distribuição normal padronizada. Sejam também as variáveis compostas $x_1, ..., x_7$ tal que $x_1 = z_1$, $x_2 = z_2$, $x_3 = z_3$, $x_4 = z_4$, $x_5 = z_1 + z_4$, $x_6 = z_2 + z_3$, $x_7 = z_3 + z_4$. Seja ainda a função $y = 10x_1 - 6x_2$ e a matriz de dados $\mathbf{W_{600\times 8}}$ cujas colunas estão associadas as variáveis definidas anteriormente como segue: $\mathbf{w}^i = x_i$, para $i = 1, ..., 7$, e $\mathbf{w}^8 = y$. Cada linha de $\mathbf{W}$ está associada a um registro conforme a tabela apresentada no Apêndice B. O algoritmo SVACP foi executado para selecionar variáveis, que foram empregadas na regressão de y, a partir de $x_1, ..., x_7$. Neste exemplo, 100% da base de dados foi utilizada para treinamento, onde foi determinado o coeficiente R^2. O resultado da execução do SVACP é descrito na Tabela 2. O critério de seleção utilizado foi o B2. Neste exemplo, foi necessário remover no mínimo 5 variáveis, para atender as restrições da complexidade da amostra. Com isto, foi obtido um regressor cujo coeficiente de determinação foi $0,49$. As variáveis removidas por este método foram x_4, x_2, x_7, x_3 e x_1. As variáveis selecionadas foram x_5 e x_6 definindo um preditor $y' = f(x_5, x_6)$, o que difere da definição de y.

Tabela 2: Resultados do exemplo 1

Nº Vars		**Desempenho**	**Complexidade**
Removidas	Restantes	R^2	da amostra
0	7	1,00	1.417
1	6	1,00	1.248
2	5	1,00	1.079
3	4	1,00	910
4	3	0,87	741
5	2	0,49	571
6	1	0,36	402

3.1.2 Seleção de Variáveis por ACP Supervisionada

O procedimento de Seleção de Variáveis por ACPS (SVACPS) recebe, como entrada, um conjunto de dados e informações necessárias para o cômputo da complexidade da amostra, realiza um pré-filtro para reduzir o conjunto de variáveis àquelas com maior influência sobre a variável de resposta e, então, emprega os métodos B2 ou B4. Assim, a diferença entre a SVACP e a SVACPS é que a SVACPS realiza a seleção considerando a variável de resposta e que os métodos B2 ou B4 são executados sobre as variáveis mais relevantes em relação à resposta. O procedimento SVACPS é apresentado na Figura 7.

Como pode ser visto, este procedimento recebe como entrada a lista de parâmetros $(\mathbf{X}, B, \delta, \varepsilon, \mathcal{F}, \mathcal{T})$ que são definidos como a lista de parâmetros de entrada do procedimento SVACP. Da mesma forma, este procedimento realiza as mesmas declarações de variáveis usadas no procedimento anterior e os passos (1), (2) e (3) executam os mesmos

- **Entrada:** Um conjunto de dados $\mathbf{X}$; número máximo B de variáveis a serem rejeitadas; probabilidade de Aprendizado PAC δ e o erro ε; fórmula $\mathcal{F}$ da dimensão VC para o modelo pretendido; função $\mathbb{T}$ que permite gerar os modelos de teste;
- **Declaração:** Número mínimo de variáveis A a serem rejeitadas; número de registros L do conjunto $\mathbf{X}$; número de variáveis D do conjunto $\mathbf{X}$; lista de complexidade da amostra $\mathbb{M}$, com D elementos; lista $\mathbb{G}$ que armazena subconjuntos de variáveis a serem rejeitadas; lista $\mathbb{R}$ que armazena subconjuntos de variáveis a serem retidas; lista $\mathbb{Z}$ de desempenhos; lista $\mathbb{U}$ de nomes das variáveis do conjunto $\mathbf{X}$;

1. Guardar o número de registros de $\mathbf{X}$ em L; guardar o número de variáveis de $\mathbf{X}$ em D; guardar o nome das variáveis de $\mathbf{X}$ em $\mathbb{U}$; inicializar $\mathbb{G}$, $\mathbb{R}$ e $\mathbb{Z}$ com conjuntos vazios;
2. Computar a complexidade da amostra por meio da função $\mathcal{F}$, supondo funções cujo número de variáveis de entrada seja $D, D-1, D-2, ..., D-B$, e inserir os valores em $\mathbb{M}$;
3. calcular o valor de A; para tanto, para $i = B$ até 1 passo -1 faça
   (a) se $\mathbb{M}_i > L$ então pare
   (b) $A = i$
4. Para n indo de A até B, faça:
   (a) Criar as listas $\mathbb{V}_n$, $\mathbb{G}_p$ e $\mathbb{R}_p$ vazias;
   (b) Para p indo de 1 até $n-1$, faça:
      i. Remover p variáveis de $\mathbf{X}$ utilizando a supervisão;
      ii. Remover $q = n - p$ variáveis de $\mathbf{X}$ utilizando um dos critérios de seleção B2 ou B4;
      iii. Inserir em $\mathbb{G}_p$ o grupo das $p+q$ variáveis rejeitadas e inserir em $\mathbb{R}_p$ o grupo das $D - n$ variáveis retidas, considerando $\mathbb{R}_p = \mathbb{U} - \mathbb{G}_p$;
      iv. Construir um modelo com as $\mathbb{R}_p$ variáveis retidas e executá-lo para obter a medida de desempenho K;
      v. Inserir $< K, p, q, \mathbb{R}_p >$ em $\mathbb{V}_n$;
   (c) Obter $max/min(\mathbb{V}_n)$ e inserir $< K, p, q >$ em $\mathbb{Z}$ e inserir $\mathbb{R}_p$ em R;
5. Retornar $\mathbb{Z}$, $\mathbb{R}$ e $\mathbb{M}$;

Figura 7: Método para a eliminação de variáveis por meio da ACPS

processamentos sobre o conjunto de dados. O passo (1) inicializa as listas $\mathbb{G}$, $\mathbb{R}$ e $\mathbb{Z}$. O passo (2) cria a lista $\mathbb{M}$ que mantém as informações referentes à complexidade da amostra para um determinado número de variáveis selecionadas. O passo (3) calcula o número

mínimo de variáveis que devem ser removidas, para que sejam atendidas as restrições de complexidade da amostra, e armazena este valor na variável A.

O passo (4) realiza uma busca sequencial cujo espaço de busca é o número de variáveis a serem removidas. Contudo, esta busca é realizada sobre os resultados de um procedimento secundário de busca exaustiva, o qual determina quantas variáveis são removidas pela supervisão e quantas variáveis são removidas pelos critérios B2 ou B4. Para tanto, cada iteração do passo (4a) inicia uma lista que $\mathbb{V}_n$ e cujas entradas são da forma $< K, p, q >$, em que K é o escore do modelo gerado por $\mathcal{T}$, p é o número de variáveis removidas por supervisão e q é o número de variáveis removidas pelos critérios B2 ou B4 - deve ser notado que $p+q = n$. Desta forma, $\mathbb{V}_n$ armazena o desempenho do procedimento de seleção considerando todas as possibilidades de se remover n variáveis usado supervisão ou ACP. O passo (4a) também cria as listas $\mathbb{G}_p$ e $\mathbb{R}_p$ que armazenam as variáveis que foram rejeitas ou selecionadas, considerando que p destas variáveis foram escolhidas por supervisão e q por ACP.

Ainda no interior do passo (4), o item 4b realiza a busca secundária para remover p variáveis usando a supervisão (passo 4(b)i) e q variáveis usando os critérios B2 ou B4 (passo 4(b)ii). O passo (4(b)iii) armazena nas listas $\mathbb{G}_p$ e $\mathbb{R}_p$ as variáveis que foram rejeitadas e retidas a cada iteração. O passo (4(b)iv) executa o regressor ou classificador $\mathcal{T}$ sobre as variáveis contidas em $\mathbb{R}_p$ e determina o escore K. O passo (4(b)v) armazena os valores de K, p, q e $\mathbb{R}_p$ em uma das entradas da lista $\mathbb{V}_n$. O passo (4c) armazena em $\mathbb{Z}$ a configuração $< K, p, q >$ que obteve o melhor desempenho e armazena em $\mathbb{R}$ a respetiva lista de variáveis $\mathbb{R}_p$.

O último passo (5) retorna as listas $\mathbb{Z}$, $\mathbb{R}$ e $\mathbb{M}$ que apresentam os resultados obtidos, expressando os maiores desempenhos, grupos de variáveis rejeitadas e complexidades da amostra.

Exemplo 2

Sejam z_1, z_2 z_3 e z_4 quatro variáveis aleatórias com distribuição normal padronizada. Sejam também as variáveis compostas $x_1, ..., x_7$ tal que $x_1 = z_1$, $x_2 = z_2$, $x_3 = z_3$, $x_4 = z_4$, $x_5 = z_1 + z_4$, $x_6 = z_2 + z_3$, $x_7 = z_3 + z_4$. Seja ainda a função $y = 10x_1 - 6x_2$ e a matriz de dados $\mathbf{W_{600\times 8}}$ cujas colunas estão associadas as variáveis definidas anteriormente como segue: $\mathbf{w}^i = x_i$, para $i = 1, ..., 7$, e $\mathbf{w}^8 = y$. Cada linha de $\mathbf{W}$ está associada a um registro conforme a tabela apresentada no Apêndice B. O algoritmo SVACPS foi executado para selecionar variáveis, que foram empregadas na regressão de y, a partir de $x_1, ..., x_7$. Neste exemplo, 100% da base de dados foi utilizada para treinamento. O resul-

tado da execução do SVACPS é descrito na Tabela 3. O critério de seleção utilizado foi o B2. Neste exemplo, foi possível observar que, para atender as restrições da complexidade da amostra, foi necessário remover no mínimo 5 variáveis. Com isto, obtêve-se um regressor cujo coeficiente de correlação foi $1,00$. As variáveis removidas foram x_3, x_4, x_7, x_6 e x_5. As variáveis retidas foram x_1 e x_2, o que permitiu a geração de um preditor definido sobre as mesmas variáveis que eram associadas a y.

Tabela 3: Resultados do exemplo 2

Nº Vars		**Desempenho**	**Complexidade**
Removidas	Restantes	R^2	da amostra
0	7	1,00	1.417
2	5	1,00	1.079
3	4	1,00	910
4	3	1,00	741
5	2	1,00	571
6	1	0,74	402

3.2 CONJUNTOS DE DADOS EXPERIMENTAIS

Esta seção descreve os conjuntos de dados usados para testar o procedimento proposto. Foram realizados experimentos com um conjunto de dados sintéticos e um conjunto de dados agrícolas. O conjunto de dados sintéticos, doravante chamado Conjunto de Dados 1, foi gerado conforme sugerido por Jolliffe (1972). Inicialmente foram definidas variáveis aleatórias z_j $(j = 1, 2, ..., 10)$, com distribuição normal padronizada. Sobre estas variáveis, foram definidas as variáveis compostas $x_1, ..., x_{10}$ conforme a Tabela 4. Nesta tabela também foi definida a variável de resposta y (11ª linha). A partir daí foi criado um conjunto de amostras x_{ik} $(i = 1, 2, ..., 1.500; k = 1, 2, ..., 11)$, segundo a distribuição associada às variáveis z_j (cujos valores foram gerados com o uso da função *rnorm*, do software R).

Tabela 4: Definição das variáveis do Conjunto de Dados 1

Variável	**Modelo**
x_1	z_1
x_2	z_2
x_3	$z_2 + z_3$
x_4	z_4
x_5	$z_4 + 0,75z_5$
x_6	$2z_4 + 0,75z_5 + 1,5z_6$
x_7	z_7
x_8	$z_7 + 0,5z_8$
x_9	$2z_7 + 0,5z_8 + z_9$
x_{10}	$3z_7 + z_8 + z_9 + z_{10}$
y	$z_1 + 0,8z_2 + 0,75z_7$

Conforme Jolliffe (1972) explica, o objetivo é que os grupos de variáveis a serem rejeitadas sejam, idealmente, aqueles cujas variáveis duplicam informações. Por exemplo, a variável x_3 é igual à variável x_2 mais uma variação aleatória. Alternativamente, pode-se considerar o inverso: x_2 igual à variável x_3 mais uma variação aleatória. Então, espera-se que no grupo de variáveis rejeitadas, esteja a variável x_2 ou x_3, independendemente de qual das duas seja a escolhida.

O Conjunto de Dados 2 refere-se a um experimento agronômico realizado em Campos Novos Paulista (SP) - safra de 2003. Naquele experimento agrícola, foram coletados 2.417 registros para os atributos de plantas (rendimento da cultura de milho tipo *Zea mays* L. em Ton/ha) e de solo (descritos na Tabela 5).

Tabela 5: Descrição das variáveis do conjunto de dados 2

Atributo químico	**Descrição**
pH	Acidez ativa
H_Al	Acidez total
MO	Teor de Matéria Orgânica total
P	Teor de Fósforo disponível
Ca	Teor de Cálcio trocável
Mg	Teor de Magnésio trocável
K	Teor de Potássio trocável
SB	Soma de bases
CTC	Capacidade de troca de cátions em pH $7,0$
V	Saturação por base
m	Saturação por alumínio
Atributo físico	**Descrição**
$CE\_Shallow$	Condutividade elétrica superficial (camada de $0-20$ cm)
CE_Deep	Condutividade elétrica subsuperficial (camada de $20-40$ cm)
IC_0_5mp	Índice de cone em 0 a 5 cm de profundidade
IC_5_10mp	Índice de cone da camada de 5 a 10 cm de profundidade
IC_10_15mp	Índice de cone da camada de 10 a 15 cm de profundidade
IC_15_20mp	Índice de cone da camada de 15 a 20 cm de profundidade
IC_20_25mp	Índice de cone da camada de 20 a 25 cm da profundidade
IC_25_30mp	Índice de cone da camada de 25 a 30 cm da profundidade
IC_30_35mp	Índice de cone da camada de 30 a 35 cm da profundidade
IC_35_40mp	Índice de cone da camada de 35 a 40 cm da profundidade

O Conjunto de Dados 3, refere-se a um experimento agrícola realizado em Ponta Grossa (PR) para o estudo da alteração da química do solo pela aplicação de calcário (CAIRES *et al.*, 2003). A avaliação da cultura ocorreu em 2001/2002. O conjunto de dados utilizado neste trabalho possui 96 amostras de solo cujos atributos coletados são apresentados na Tabela 6.

Cada um dos atributos da Tabela 6 possuiu um sufixo A, B, C ou D indicando as camadas de $0-10$, $10-20$, $20-40$ ou $40-60$ cm, respectivamente. Assim, por exemplo, o atributo Ca_A representou o Cálcio coletado na camada A, Ca_B representou o Cálcio coletado na camada B e assim para cada variável, exceto o Gesso, que foi aplicado à

Tabela 6: Conjunto de Dados 3

Variável	Unidade	Descrição
Gesso	$t\ ha^{-1}$	Gesso agrícola
pH	$mol\ L^{-1}$	Acidez ativa
Al	$mmol_c\ L^{-1}$	Teor de Alumínio
Ca	$mmol_c\ L^{-1}$	Teor de Cálcio trocável
Mg	$mmol_c\ L^{-1}$	Teor de Magnésio trocável
K	$mmol_c\ L^{-1}$	Teor de Potássio trocável
m	%	Saturação por Alumínio
Ca_CTCe	$mmol_c L^{-1}$	Saturação por Ca na CTC a pH 7,0

camada superficial. Assim, o Conjunto de Dados 3 possuia 29 variáveis.

Do Conjunto de Dados 2 e 3, foram removidos os valores discrepantes considerando limites de controles iguais à média, mais ou menos três desvios padrões (GONÇALVES, 1965). Todos os registros contendo apenas um valor discrepante, em pelo menos uma das variáveis aleatórias, foram removidos do conjunto de dados. Assim, os totais de amostras utilizadas nos Conjuntos de Dados 2 e 3 foram 2.138 e 79 registros, respectivamente.

3.3 EXPERIMENTOS

Esta seção descreve os experimentos com os procedimentos SVACP e SVACPS, aplicando os critérios de seleção B2 e B4. O objetivo dos experimentos é mostrar a efetividade dos procedimentos na seleção de variáveis. Os experimentos foram realizados em conjuntos de dados sintéticos e agrícolas, como descrito a seguir.

3.3.1 Materiais Empregados nos Experimentos

Para a realização dos experimentos descritos na próxima sub seção, foi utilizada uma implementação em linguagem R (versão 2.13.0) de cada procedimento. Para o cômputo dos componentes principais por meio da técnica DVS, foi utilizada a função *prcomp* do software R. Para o cômputo dos componentes principais por meio da técnica probabilística, foi utilizada a função *ppca* do pacote *pcaMethods*. Os modelos foram gerados utilizando a função *lm* do software R para o Modelo de Regressão Linear Múltiplo, a função *mlp* do pacote *RSNNS* e a função *naiveBayes* do pacote *e1071*. Após a implementação dos dois procedimentos (SVACP e SVACPS), procedeu-se os experimentos descritos a seguir.

3.3.2 Estratégia de Amostragem e Complexidade da Amostra

A estratégia de amostragem foi conduzida de modo que: 66% dos registros foram selecionados aleatoriamente para o treinamento (geração) dos modelos e os 34% remanescentes, foram usados para teste de desempenho dos mesmos. Em cada experimento, cada método foi executado 35 vezes para que fosse possível representar, por média e desvio padrão, a variação dos resultados produzidos. As médias e desvios padrão foram usados para analisar a efetividade dos métodos. A complexidade da amostra foi configurada para que os modelos fossem estimados com erro ε de 10% e probabilidade δ de $0,05$.

3.3.3 Características dos Regressores e Classificadores Usados nos Experimentos

3.3.3.1 Modelo de Regressão Linear Múltipla

Os Modelos de Regressão Linear Múltipla (MRLM) foram construídos sobre as variáveis selecionadas pelos procedimentos SVACP e SVACPS. Isto é, cada variável independente da função a ser regredida representava um dos atributos selecionados. Cada modelo foi induzido utilizando os métodos descritos na Seção 2.3.1. O objetivo de utilizar o MRLM foi verificar se os procedimentos SVACP e SVACPS eram capazes de selecionar subconjuntos de variáveis de forma a reduzir a complexidade da amostra e, ao mesmo tempo, permitir a construção de regressores que apresentassem um desempenho igual ou superior àquele atingido por um regressor definido sobre todos os atributos da base de dados. A medida utilizada para aferir o desempenho do preditor foi o coeficiente de determinação (R^2).

3.3.3.2 *Naive Bayes*

O classificador bayesiano foi construído com base no algoritmo *Naive Bayes*, o qual estimou suas tabelas de probabilidades sobre as variáveis retidas. Cada variável retida representou um parâmetro de entrada e a variável de resposta, a saída do classificador. A variável de resposta foi programada para classificar as instâncias nas categorias abaixo ou acima da média amostral. Estas categorias foram representadas pelos valores 0 e 1, respectivamente.

3.3.3.3 Rede Neural Artificial

As Redes Neurais Artificais foram construídas com base no modelo perceptron de múltiplas camadas, conforme apresentado na Seção 2.3.3. Cada variável selecionada pelo procedimento de seleção, representou um neurônio de entrada. Em cada rede construída, foi utilizada apenas uma única camada oculta com cinco neurônios artificiais. Todos os neurônios da camada oculta foram programados com a função sigmoide, como função de ativação. A camada de saída possuía apenas um neurônio de saída, o qual utilizou a função limiar. Idealmente, a saída esperada deveria ser maior ou igual a um limiar t, sempre que os dados da entrada se referissem a uma resposta maior ou igual à média amostral. A taxa de aprendizado foi fixada em $0,3$ e o treinamento foi realizado em 500 épocas. A medida de desempenho utilizada foi a soma dos quadrados dos erros (SQE).

3.3.4 Experimentos com o Conjunto de Dados 1 (Sintéticos)

O objetivo dos dois experimentos descritos nesta seção foi testar os procedimentos SVACP e SVACPS na seleçao de variáveis sobre o conjunto de dados sintéticos (Conjunto de Dados 1), quando consideradas as operações de regressão e classificação. Como os relacionamentos entre as variáveis são conhecidos, os experimentos sobre este conjunto de dados permite verificar se as variáveis selecionadas pelos métodos SVACP e SVACPS eram as que tinham maior influência na variação dos dados e sobre a variável de resposta.

Os dois experimentos também ilustram duas maneiras de aplicar os procedimentos: quando o conjunto de dados está completo e quando existem dados faltantes. No primeiro caso, emprega-se a ACP implementada por DVS enquanto no segundo caso usa-se a ACP Probabilística.

Para validar os experimentos, um teste T para amostras pareadas foi aplicado para verificar a diferença estatística entre as médias de cada procedimento e calculada a multicolinearidade dos conjuntos de variáveis por meio do fator de inflação da variância (FIV). O teste T foi utilizado para verificar se a SVACPS produziria resultados significativamente diferentes dos resultados da SVACP. O índice FIV foi utilizado para averiguar se os procedimentos rejeitaram grupos de variáveis que não eram relevantes ou que duplicavam dados.

3.3.4.1 Experimento 1

Este experimento consistiu em aplicar os procedimento SVACP e SVACPS no Conjunto de Dados 1. A ACP utilizada nos critérios B2 e B4 foi implementada com a técnica de DVS. O número máximo de variáveis a serem removidas foi limitado a 7. Cada um dos subgrupos de variáveis selecionados foram utilizados para gerar os modelos de Regressão Linear Múltipla, Redes Neurais Artificiais e classificadores bayesianos. O experimento 1 foi elaborado para mostrar a efetividade dos procedimento na seleção de variáveis, quando consideradas as restrições impostas sobre a complexidade da amostra para geração dos respectivos modelos.

3.3.4.2 Experimento 2

Este experimento consistiu em aplicar os procedimentos SVACP e SVACPS sobre um conjunto com dados incompletos. Na primeira etapa deste experimento foram manipulados alguns registros de modo que 15% dos casos tinham um registro com algum atributo não observado. Este conjunto de dados foi gerado a partir do Conjunto de Dados 1. Foram selecionados aleatoriamente 15% dos registros do conjunto original para terem o valor de um dos seus atributos substituído por um rótulo, indicando que o mesmo não foi observado. Para tanto utilizou-se o procedimento apresentado no Apêndice A.

Para possibilitar o emprego dos procedimentos SVACP e SVACPS sobre os conjuntos de dados com dados incompletos, a análise de componentes principais foi executda usando a ACPP. O número máximo de variáveis a serem rejeitadas foi limitado a 7. Os modelos utilizados foram a Regressão Linear Múltipla e o classificador *Naive Bayes*.

3.3.5 Experimentos com o Conjunto de Dados 2 (Agrícolas)

Considerando o interesse pelo uso de métodos multivariados de regressão e classificação na análise de dados agrícolas, foram realizados dois experimentos para avaliar o desempenho dos procedimentos SVACP e SVACPS sobre uma base com dados agrícolas (Conjunto de Dados 2). Mais uma vez, o objetivo foi avaliar o desempenho das variáveis selecionadas por estes procedimentos, quando empregadas na regressão/indução de modelos multivariados, em situações práticas. Os experimentos ilustram duas maneiras de aplicar os procedimentos: quando o conjunto de dados está completo e quando existem dados incompletos (quando faltam alguns valores em alguns campos do conjunto de dados).

Para validar os experimentos, um teste T para amostras pareadas foi aplicado para verificar a diferença estatística entre as médias de cada procedimento e calculada a multicolinearidade dos conjuntos de variáveis por meio do FIV. O teste T foi utilizado para verificar se a SVACPS produziria resultados significativamente diferentes dos resultados da SVACP. O índice FIV foi utilizado para averiguar se os procedimentos rejeitaram grupos de variáveis que não eram relevantes ou que duplicavam dados.

3.3.5.1 Experimento 3

Este experimento consistiu em aplicar os procedimentos SVACP e SVACPS no Conjunto de Dados 2. A ACP empregada neste experimento foi aquela implementada com a técnica de DVS, pois, não existiam dados faltantes no conjunto. Neste experimento, procedeu-se a seleção de variáveis para eliminar no máximo 15 variáveis. Os modelos utilizados neste experimento foram a Regressão Linear Múltipla, a Rede Neural Artificial do tipo Perceptron de Múltiplas Camadas e o classificador *Naive Bayes*. O objetivo deste experimento foi verificar o comportamento dos métodos quando aplicados a um conjunto de dados agrícolas real.

3.3.5.2 Experimento 4

Este experimento consistiu em executar os procedimentos SVACP e SVACPS sobre um conjunto de dados com registros que tinham dados faltantes. Primeiramente, foram manipulados alguns registros de modo que 15% dos casos tinham um atributo não observado ou faltante. Este conjunto de dados foi gerado a partir do Conjunto de Dados 2 pelo mesmo processo utilizado no Experimento 2, apresentado no Apêndice A.

A ACPP foi utilizada para possibilitar o emprego dos procedimentos SVACP e SVACPS. O número máximo de variáveis a serem rejeitadas foi limitado a 15. Os modelos utilizados foram a Regressão Linear Múltipla e o classificador *Naive Bayes*.

3.3.6 Experimentos com o Conjunto de Dados 3 (Agrícolas)

O experimento a seguir foi realizado com o Conjunto de Dados 3, cujo objetivo foi ilustrar uma situação em que o pesquisador necessita que uma variável seja obrigatoriamente selecionada. Tal situação foi observada a partir do trabalho de Caires *et al.* (2003), que analisaram a aplicação de gesso agrícola em relação ao comportamento dos demais atributos do solo. O experimento descrito a seguir considera que, ao gerar um

modelo linear multivariado, é desejável que a variável *Gesso* não seja eliminada do conjunto e que ela apareça no modelo final. Nesta situação, é empregada uma implementação diferenciada dos procedimentos SVACP e SVACPS, a qual tanto a supervisão quanto os critérios B2 e B4 não eliminam a variável em questão.

Para validar o experimento, foi calculada a multicolinearidade dos conjuntos de variáveis por meio do fator de inflação da variância (FIV). O índice FIV foi utilizado para averiguar se os procedimentos rejeitaram grupos de variáveis que não eram relevantes ou que duplicavam dados.

3.3.6.1 Experimento 5

Este experimento consistiu em aplicar os procedimento SVACP e SVACPS no Conjunto de Dados 3, com o objetivo de ilustrar uma situação em que o pesquisador deseja que a variável *Gesso* apareça obrigatoriamente em um modelo de Regressão Linear Múltipla, ou seja, ele deseja que o *Gesso* não seja eliminado do conjunto durante o processo de seleção pelo fato de que tal variável é um fator potencialmente determinante em sua pesquisa.

O objetivo foi selecionar variáveis para gerar um MRLM para predizer a variável Ca_A com base nas demais variáveis. Nesta situação foram descartadas manualmente as variáveis Ca_B, Ca_C e Ca_D, pois tais variáveis são correlacionadas com Ca_A por meio da variável *Gesso*. A aplicação de gesso agrícola influencia diretamente no teor de Cálcio disponível no solo e na planta (CAIRES *et al.*, 2003).

O número máximo de variáveis a serem rejeitadas foi 23, já que o total de variáveis preditoras no conjunto era 25. A variável *Gesso* foi *pré-selecionada*, ou seja, ela não poderia ser eliminada do conjunto. Os procedimentos SVACP e SVACPS foram executados apenas uma vez. Neste experimento não foi executado o teste T, por não terem sido extraídas a média e o desvio padrão dos modelos provenientes dos procedimentos. A multicolinearidade foi utilizada para validar a efetividade dos procedimentos em remover variáveis correlacionadas linearmente.

Especificamente, neste experimento, a estratégia de amostragem foi diferenciada, sendo utilizado 100% do conjunto de dados para estimar o regressor linear múltiplo. Então, calculou-se o coeficiente R^2, do treinamento do regressor gerado sobre os dados, para comparar os diferentes modelos.

4 RESULTADOS

Este capítulo apresenta os resultados dos experimentos realizados com os procedimentos de seleção de variáveis SVACP e SVACPS. Assim, são apresentados os resultados dos experimentos realizados com a base de dados sintética (Conjunto de Dados 1), a base de dados de agricultura de precisão (Conjunto de Dados 2) e com a base de dados sobre análise de solo e nutrição de plantas (Conjunto de Dados 3). Os resultados aqui relatados foram obtidos em bases de dados completas e na presença de dados faltantes.

Os resultados apresentados dizem respeito ao uso dos procedimentos SVACP e SVACPS na seleção de variáveis para a construção de regressores e classificadores, como especificado na Seção 3.3. Os resultados são mostrados separadamente para cada tipo de classificador. Os métodos B2 e B4 foram usados como critérios de seleção não supervisionada. O impacto do emprego da supervisão na seleção de variáveis, em cada experimento, foi avaliada pela comparação do desempenho dos preditores e classificadores. A comparação dos escores foi feita analisando estatisticamente a diferença entre as médias, obtidas pelos modelos oriundos de cada procedimento.

Este capítulo está organizado como segue. Na Seção 4.1 são apresentados os resultados obtidos sobre dados sintéticos. Na Seção 4.2 e 4.3 são apresentados os resultados obtidos sobre conjuntos de dados reais. Na Seção 4.4 são feitas algumas considerações gerais.

4.1 RESULTADOS SOBRE O CONJUNTO DE DADOS SINTÉTICOS

4.1.1 Resultados do Experimento 1 - Dados Completos

Os resultados do Experimento 1 são apresentados nas Tabelas 7, 10 e 12. Em todos os casos, os métodos B2 e B4 usaram a técnica DVS para calcular os componentes principais. Nas referidas tabelas, a primeira coluna indica o número de variáveis selecionadas pelos procedimentos SVACP e SVACPS. As colunas identificadas com μ mostram o valor da média do índice R^2 obtido pelos regressores MRLM calculados sobre as 35 execuções, como descrito na Seção 3.3. As colunas identificadas com σ informam o desvio padrão relacionado à respectiva média μ da coluna anterior. A última coluna, identificada com m informa a complexidade da amostra segundo os critérios estabelecidos pelo aprendizado PAC, conforme descrito na Seção 3.3.2. Deve ser observado que as tabelas de resultados mostram apenas alguns dos casos de seleção, que são aqueles cujo tamanho da amostra disponível atenderam os requisitos da complexidade da amostra m. Como o Conjunto

de Dados 2 possuia 2.138 registros, mas apenas 1.412 foram utilizados para treinamento (66% de 2.138), então, a complexidade da amostra foi atendida quando $m < 1.412$.

4.1.1.1 Resultados sobre MRLM

É possível observar na Tabela 7 que o procedimento SVACPS permitiu a geração de preditores com maiores correlações do que aqueles providos pelos modelos gerados pelo SVACP. Um teste T, para amostras pareadas, foi realizado para verificar se a diferença entre as médias obtidas pelo SVACPS diferem estatísticamente, com nível de significância $\alpha = 0,05$, das médias obtidas pelos modelos gerados pelo SVACP. Este teste indicou que os resultados dos modelos provindos da SVACPS produziram resultados estatísticamente diferentes que os modelos provindos da SVACP. Como os coeficientes de determinação dos modelos da SVACPS foram maiores, então, os resultados mostraram que SVACPS selecionou as variáveis com maior influência sobre a variável de resposta. Estes resultados podem ser observados no gráfico da Figura 8.

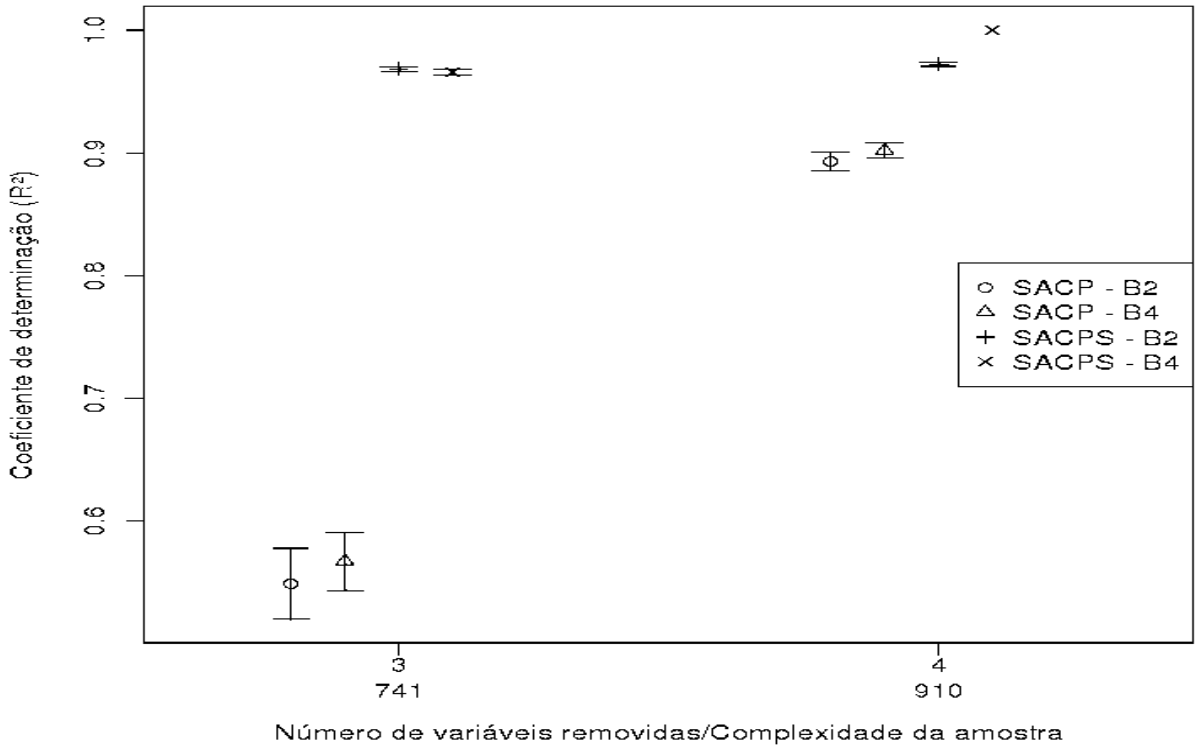

Figura 8: Gráfico dos resultados do Experimento 1 sobre MRLM

As restrições impostas pelo conceito de complexidade da amostra, considerando a abordagem do aprendizado PAC, estipulavam que no máximo 4 variáveis deveriam ser selecionadas para compor um modelo linear múltiplo. Como pode ser observado, o uso dos procedimentos SVACP e SVACPS, permitem a seleção do menor conjunto de variáveis

que maximiza o coeficiente R^2. Nos testes realizados neste experimento, este conjunto tem em média 4 variáveis e foi produzido pelo procedimento SVACPS com o critério B4. O grupo de variáveis mais frequentemente selecionado para tal modelo foi x_1, x_2, x_7 e x_{10}. Ocorreu conforme esperado, que as variáveis x_1, x_2 e x_7 fossem incluídas no grupo de variáveis selecionadas.

Tabela 7: Resultados do Experimento 1 com MRLM

R^2 do MRLM									
	SVACP				SVACPS				
N° Vars.	$B2$		$B4$		$B2$		$B4$		
#Ret.	μ	σ	μ	σ	μ	σ	μ	σ	m
4	0,8930	0,0077	0,9019	0,0063	0,9721	0,0016	1,0000	0,0000	910
3	0,5491	0,0289	0,5674	0,0237	0,9684	0,0018	0,9658	0,0024	741

Foi calculada a multicolinearidade entre todas as variáveis e daquele conjunto que possuiu as variáveis selecionadas, citadas anteriormente. Comparando as multicolinearidades, foi possível observar (Tabelas 8 e 9) que ao reduzir o número de variáveis em um conjunto, a multicolinearidade foi reduzida. Isto indica que as variáveis independentes que foram removidas do conjunto foram aquelas cujo comportamento poderia ser explicado pelas demais.

Tabela 8: Multicolinearidade de todas as variáveis

Variável	x1	x2	x3	x4	x5
FIV	1,0068	2,0512	2,0465	3,5293	3,6289

Variável	x6	x7	x8	x9	x10
FIV	3,1511	6,1880	7,3641	10,6240	11,6830

Tabela 9: Multicolinearidade das variáveis selecionadas

Variável	x1	x2	x7	x10
FIV	1,0022	1,0021	3,8810	3,8747

4.1.1.2 Resultados sobre NB

Dada que a dimensão de Vapnik-Chervonenkis do classificador *Naive Bayes* é a mesma que a da Regressão Linear Múltipla, foram considerados modelos com no máximo 4 variáveis. Foi realizado um teste T, em amostras pareadas, para verificar a diferença entre as médias das taxas de acerto dos classificadores gerados pelos procedimentos SVACP e SVACPS. Este teste permitiu confirmar que os classificadores provindos do SVACPS possuem uma diferença sensível (à significância $\alpha = 0,05$) dos classificadores provenientes do SVACP. Os resultados dos classificadores podem ser observados na Figura 9.

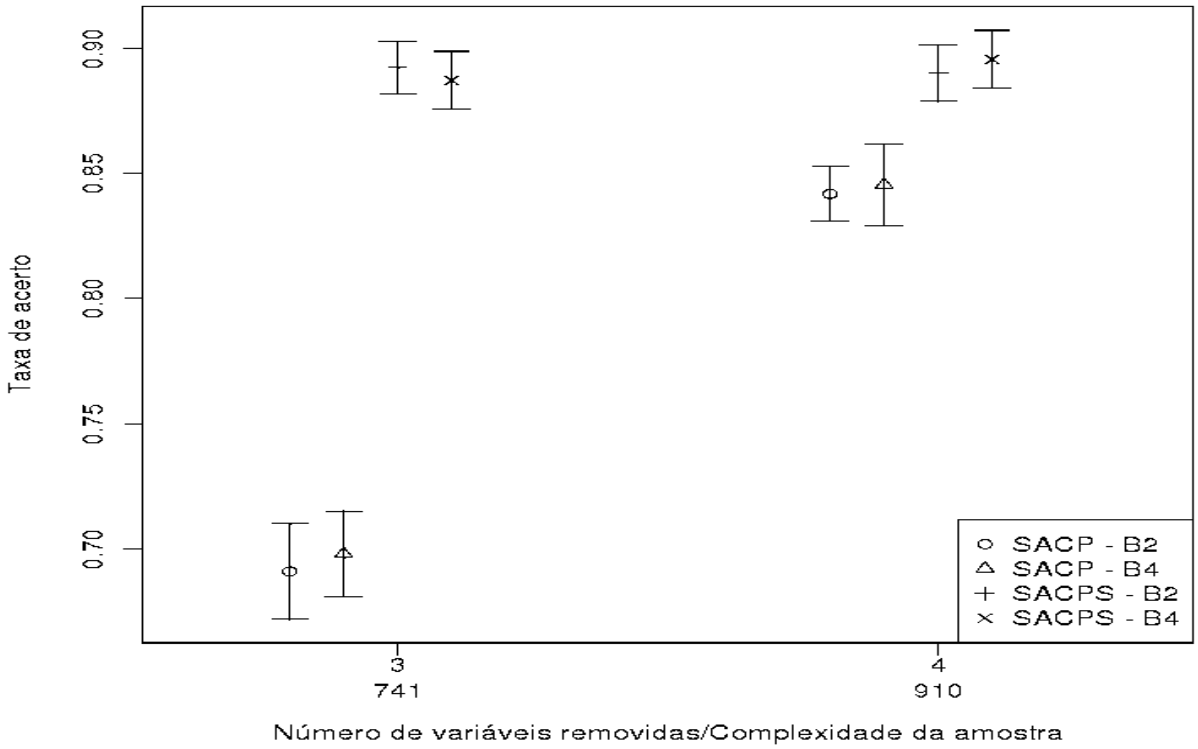

Figura 9: Gráfico dos resultados do Experimento 1 sobre NB

Em média, os melhores resutados, em termos da taxa de acerto do classificador, foram aqueles cujos modelos tinham 4 variáveis, em média, conforme apresentado na Tabela 10. Este resultado foi obtido pelo procedimento SVACPS com critério B4, cuja taxa de acerto média foi $89,56\%$ e complexidade da amostra de 910 registros. As variáveis selecionadas que compuseram o classificador foram x_1,x_2,x_9 e x_{10}. Das três variáveis que deveriam estar no conjunto das selecionadas (x_1, x_2 e x_7), apenas x_1 e x_2 apareceram no grupo de variáveis selecionadas.

Tabela 10: Resultados do Experimento 1 com NB

Taxa de Acerto do classificador NB									
	SVACP				**SVACPS**				
Nº Vars.	***B*2**		***B*4**		***B*2**		***B*4**		
#Ret.	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{m}$
4	0,8418	0,0109	0,8454	0,0164	0,8901	0,0113	0,8956	0,0116	910
3	0,6909	0,0191	0,6979	0,0171	0,8923	0,0104	0,8872	0,0116	741

A multicolinearidade entre as variáveis selecionadas foi calculada (Tabela 11) e comparada com a multicolinearidade extraída do conjunto com todas as variáveis. Foi observado que ao remover algumas variáveis do conjunto, a multicolinearidade das selecionadas foi reduzida, o que permite verificar que aquelas variáveis linearmente correlacionadas às selecionadas foram removidas. Apesar disto, variável x_9 teve seu índice FIV pouco decrementado. O fato de x_9 ser essencialmente composto pelas mesmas variáveis

latentes que x_{10}, pode explicar o porque destas duas variáveis terem o maior FIV do conjunto selecionado.

Tabela 11: Multicolinearidade das variáveis selecionadas

Variável	x1	x2	x9	x10
FIV	1,0016	1,0019	9,2388	9,2374

4.1.1.3 Resultados sobre RNA

É possível observar na Tabela 12, que as RNAs que produziram menores erros foram aquelas geradas pelo procedimento SVACPS. Um teste T foi realizado para avaliar a diferença entre as médias obtidas dos resultados dos modelos produzidos pelos procedimentos SVACP e SVACPS, cujas amostras são pareadas. O nível de significância do teste foi $\alpha = 0,05$. Foi observado que os resultados obtidos pelos modelos gerados pela SVACPS diferem estatisticamente dos resultados produzidos por aqueles modelos gerados pela SVACP. A Figura 10 ilustra os resultados das Redes Neurais Artificiais.

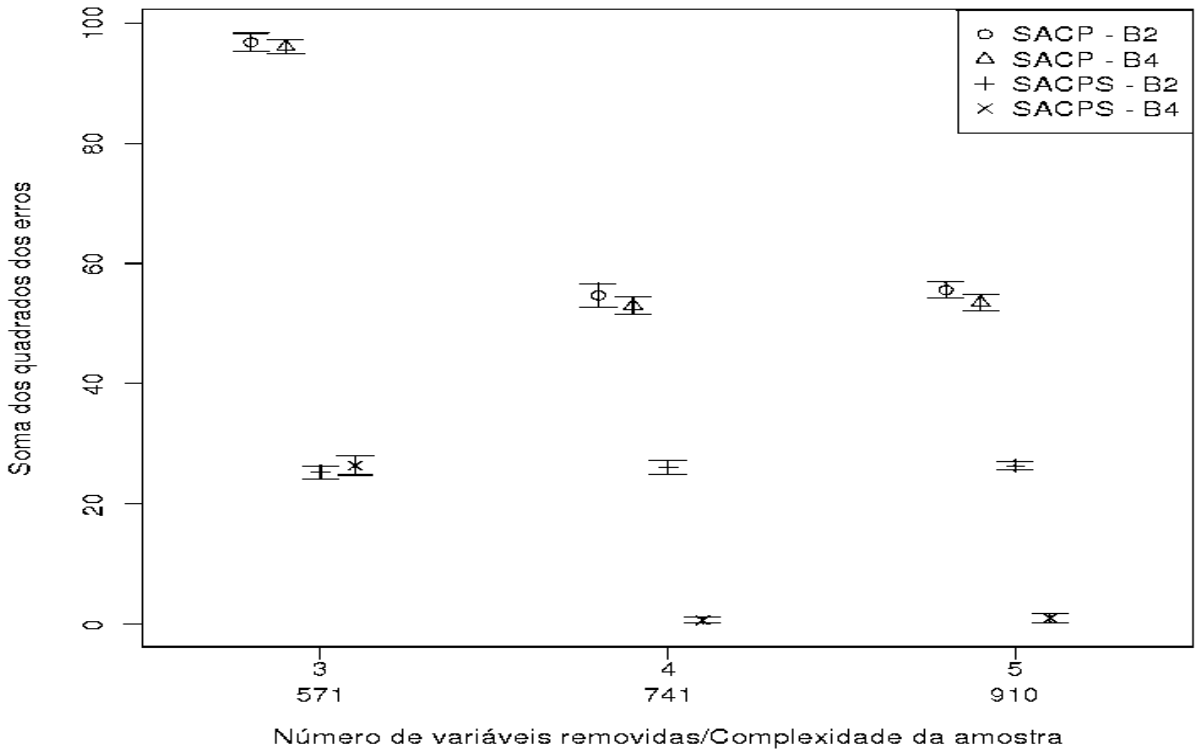

Figura 10: Gráfico dos resultados do Experimento 1 sobre RNA

Para as RNAs atenderem as exigências da complexidade da amostra, pelo aprendizado PAC, foi necessário reter até no máximo 5 variáveis. Assim, os modelos válidos foram aqueles cujas complexidades foram até 910 registros. A RNA que apresentou menor

erro quadrado, em média, foi aquela com 4 variáveis de entrada, selecionada pelo procedimento SVACPS com o critério B4, cuja média da soma dos quadrados dos erros foi $0,6742$ e complexidade da amostra 741 registros. As variáveis selecionadas foram: x_1, x_2, x_7 e x_{10}. Assim, as variáveis que compõem a resposta (x_1, x_2 e x_7) estavam contidas no conjunto selecionado.

Tabela 12: Resultados do Experimento 1 com RNA

Soma dos Quadrados dos Erros do classificador RNA									
	SVACP				**SVACPS**				
Nº Vars.	***B*2**		***B*4**		***B*2**		***B*4**		
#Ret.	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{m}$
5	55,5968	1,3394	53,4930	1,3921	26,3696	0,7448	1,0341	0,7673	910
4	54,7012	1,9771	52,9746	1,4097	26,0996	1,1715	0,6742	0,4908	741
3	96,7718	1,5247	96,0074	1,1253	25,2982	1,0712	26,4232	1,5913	571

Com a redução de dimensionalidade dos dados, a multicolinearidade foi reduzida (comparar Tabelas 8 com 13) entre as variáveis retidas no conjunto de dados. Isso permite verificar que os procedimentos removeram aquelas variáveis mais relacionadas com aquelas que foram retidas. O que mostra que os procedimentos também cumpriram seu propósito quando aplicados à modelagem de RNAs sob a redução de dimensionalidade dos dados e ajuste da complexidade da amostra.

Tabela 13: Multicolinearidade das variáveis selecionadas

Variável	x1	x2	x7	x10
FIV	1,0022	1,0021	3,8810	3,8747

4.1.2 Resultados do Experimento 2 - Dados Incompletos

Os resultados do Experimento 2 são apresentados nas Tabelas 14 e 16. As colunas destas tabelas são descritas como as das tabelas de resultados do Experimento 1. Em todos os casos, os critérios B2 e B4 usaram a técnica da ACPP para calcular os componentes principais. Por meio do algoritmo *EM* a ACPP estimou os valores não observados do conjunto e este foi completado antes da execução do processo de modelagem automática.

Os conjuntos de amostras utilizados neste experimento foram obtidos a partir do Conjunto de Dados 1 com o emprego de um procedimento, descrito no Apêndice A, que atribuiu o rótulo *não observado* a um atributo de 15% dos casos do conjunto original. O objetivo deste experimento foi mostrar a possibilidade de usar os procedimentos SVACP e SVACPS no caso de amostras com dados faltantes em conjunto com dados reais. As tabelas de resultados do Experimento 2 mostram resultados similares aos do Experimento 1, em que o procedimento SVACPS fornece modelos cujo desempenho é maior do que

os desempenhos dos modelos gerados a partir do procedimento SVACP. A seguir são apresentados os resultados para cada tipo de modelo.

4.1.2.1 Resultados sobre MRLM

Como pode ser observado na Tabela 14, o desempenho médio dos modelos gerados a partir das variáveis selecionadas, pelo procedimento SVACPS, foi superior àquela obtido pela SVACP, independentemente do fato do critério B2 ou B4 terem sido utilizados. Também pode ser observado que os resultados do método SVACPS tem uma variânca menor do que aqueles obtidos com a SVACP. O teste T para teste de hipóteses sobre a diferença entre as médias de duas amostras pareadas, foi utilizado verificar se o desempenho dos procedimentos SVACP e SVACPS é estatísticamente significante. O que foi confirmado com um nível de significância de $\alpha = 0.05$. A Figura 11 ilustra os resultados do MRLM.

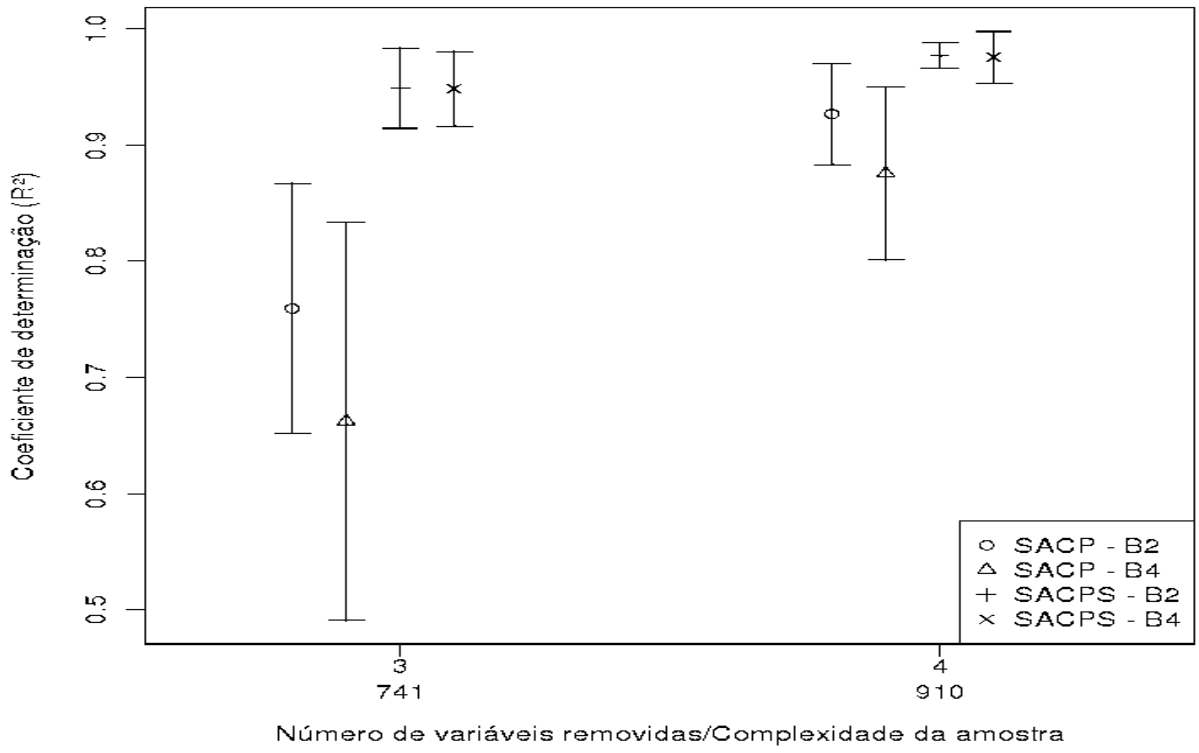

Figura 11: Gráfico dos resultados do Experimento 2 sobre MRLM

É possível observar que o modelo de maior correlação, em média, foi aquele que possuiu 4 variáveis, cujo coeficiente de determinação foi $0,9769$ e a complexidade da amostra foi de 910 registros. O modelo de 4 variáveis que produziu maior correleção foi aquele cujas variáveis selecionadas foram x_1, x_2, x_6 e x_7, por meio do SVACPS utilizando critério B2. Como pode ser observado, as variáveis x_1, x_2 e x_7 estavam presentes dentre as selecionadas, o que foi previsto para um modelo sintético.

Tabela 14: Resultados do Experimento 2 com MRLM

R^2 do MRLM - 15% de dados incompletos									
	SVACP				SVACPS				
N° Vars.	$B2$		$B4$		$B2$		$B4$		
#Ret.	μ	σ	μ	σ	μ	σ	μ	σ	m
4	0,9267	0,0435	0,8755	0,0744	0,9769	0,0110	0,9756	0,0222	910
3	0,7593	0,1078	0,6622	0,1714	0,9490	0,0346	0,9484	0,0321	741

Observando a Tabela 8 e depois a Tabela 15, é possível notar que a multicolinearidade dos dados também pode ser reduzida aplicando a redução de dimensionalidade por meio dos procedimentos SVACP e SVACPS sobre dados incompletos. Isto permite ver que as variáveis removidas foram aquelas que possuiram maior relação com as demais variáveis que foram retidas no conjunto. Ainda, pode-se dizer que não há correlação entre as variáveis selecionadas, o que pode ser confirmado pela Tabela 4, que mostra a composição das variáveis.

Tabela 15: Multicolinearidade das variáveis selecionadas

Variável	x1	x2	x6	x7
FIV	1,0024	1,0018	1,0003	1,0036

4.1.2.2 Resultados sobre NB

Os resultados obtidos com o classificador *Naive Bayes*, Tabela 16, mostra que os modelos gerados a partir do procedimento SVACPS tinham um a taxa de acerto maior do aqueles gerados com SVACP, independentemente do fato do critério B2 ou B4 terem sido utilizados. Novamente, pode ser observado que os resultados do método SVACPS tem uma variânca menor do aqueles obtidos com a SVACP. O teste T para teste de hipóteses, sobre a diferença entre as médias de duas amostras pareadas, foi utilizado verificar se o desempenho dos procedimentos SVACP e SVACPS é estatísticamente significante. O que foi confirmado com um nível de significância de $\alpha = 0.05$. A Figura 12 mostra os resultados do classificador NB.

O classificador que produziu maior taxa de acerto média foi aquele que possuiu 4 variáveis. Sua taxa de acerto média foi $89,19\%$ e sua complexidade da amostra foi de 910 registros. As variáveis selecionadas foram x_1, x_2, x_6 e x_7, as quais foram retidas pelo procedimento SVACPS com o uso do critério B2. O conjunto de variáveis selecionadas possuiu as variáveis previstas em um modelo sintético, que foram x_1, x_2 e x_7.

Como o grupo de variáveis selecionadas foram as mesmas do teste com MRLM, no Experimento 2, pode-se referenciar à Tabela 15, pois a multicolinearidade é a mesma.

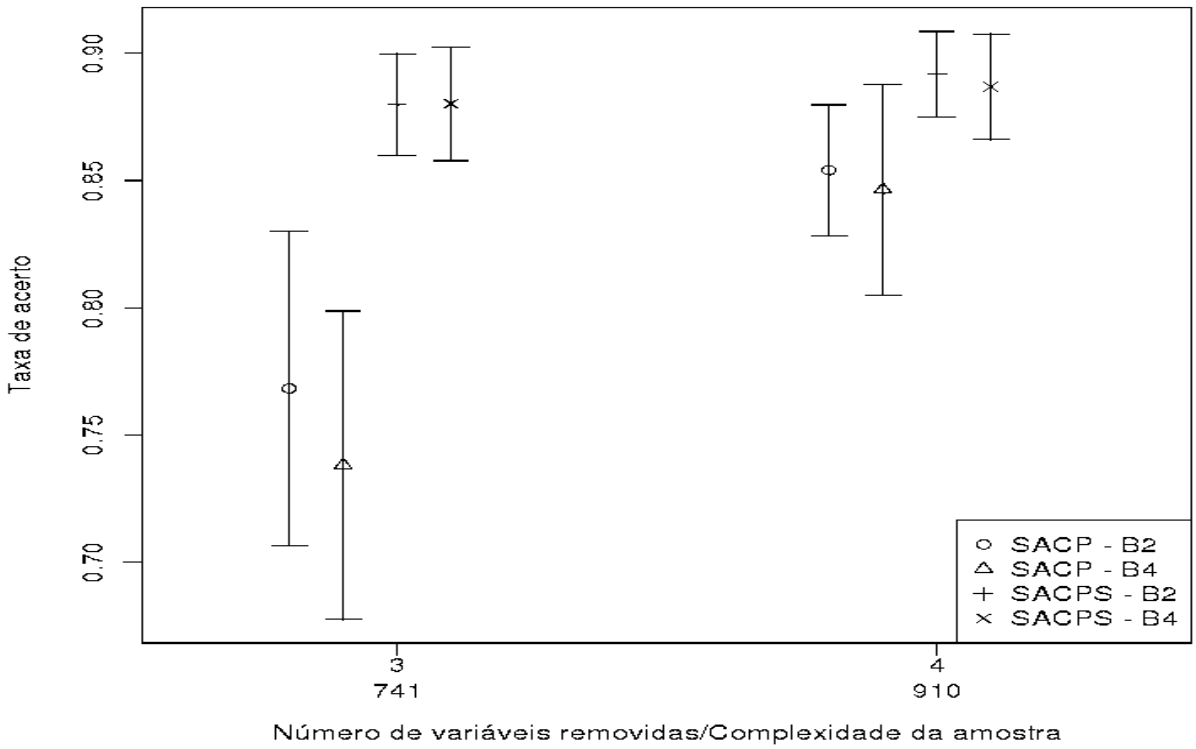

Figura 12: Gráfico dos resultados do Experimento 2 sobre NB

Tabela 16: Resultados do Experimento 2 com NB

Taxa de Acerto do classificador NB - 15% de dados incompletos									
	SVACP				**SVACPS**				
N° Vars.	***B2***		***B4***		***B2***		***B4***		
#Ret.	μ	σ	μ	σ	μ	σ	μ	σ	m
4	0,8541	0,0257	0,8464	0,0414	0,8919	0,0168	0,8869	0,0208	910
3	0,7682	0,0620	0,7380	0,0607	0,8799	0,0200	0,8802	0,0224	741

Isso mostra, que mesmo para dados incompletos, os procedimentos conseguem separar as variáveis com menor relevância e que possuem maior relação com as demais, para serem retiradas do conjunto.

4.2 RESULTADOS SOBRE O CONJUNTO DE DADOS 2

4.2.1 Resultados do Experimento 3 - Dados Completos

Os resultados obtidos do Experimento 3 são apresentados nas Tabelas 17, 20 e 22. Os resultados são apresentados aqui em tabelas cujas colunas são descritas da mesma maneira como no Experimento 1. Os critérios B2 e B4 calcularam os componentes principais por meio da técnica DVS.

Em relação aos modelos MRLM e NB, as maiores correlações e taxas de acerto foram obtidos pelos grupos de variáveis selecionadas pelo procedimento SVACPS. Porém,

em relação à RNA, nem todos os menores erros foram proporcionados pelo procedimentos SVACPS. Os resultados que respeitam o aprendizado PAC, obtidos sobre o Conjunto de Dados 2, diferem estatisticamente à significância de $\alpha = 0,05$.

4.2.1.1 Resultados sobre MRLM

Os resultados do Experimento 3 relacionados à MRLM são apresentados na Tabela 17. Os coeficientes de correlação dos modelos lineares produzidos pelo SVACPS, são maiores do que os produzidos pelos modelos provenientes do procedimento SVACP. Um teste T, para amostras pareadas, foi realizado sobre as médias obtidas para verificar se as diferenças nos resultados entre os dois procedimentos foram significativas. O que foi confirmado a um nível de significância $\alpha = 0,05$. A Figura 13 ilustra os resultados procedentes do MRLM.

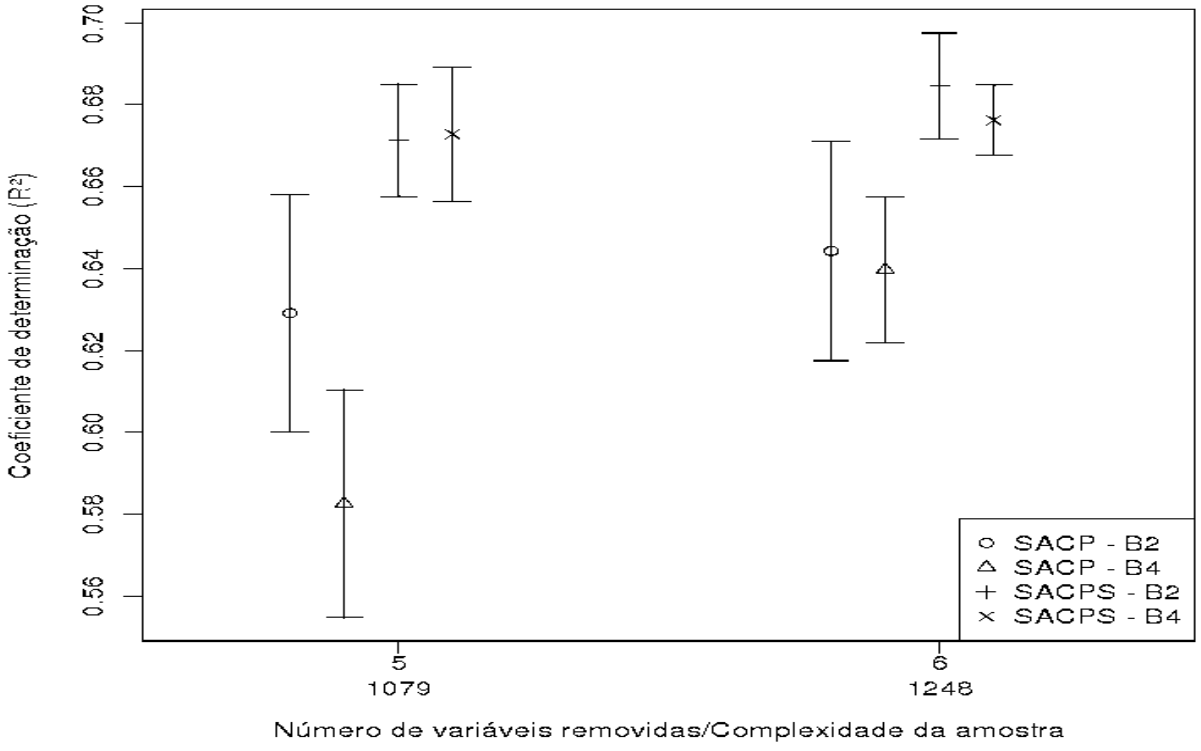

Figura 13: Gráfico dos resultados do Experimento 3 sobre MRLM

O modelo que produziu maior coeficiente de determinação, em média, foi aquele com 6 variáveis. O coeficiente de determinação deste modelo foi, em média, $0,6846$ e sua complexidade da amostra foi de 1.248 registros. Suas variáveis foram selecionadas pelo procedimento SVACPS utilizando o critério B2, as quais foram P, K, MO, Ph, Sb e $CE\_Shallow$.

Foram calculadas as multicolinearidades sobre todas as variáveis do conjunto e

Tabela 17: Resultados do Experimento 3 com MRLM

R^2 do MRLM									
	SVACP				SVACPS				
N° Vars.	$B2$		$B4$		$B2$		$B4$		
#Ret.	μ	σ	μ	σ	μ	σ	μ	σ	m
6	0,6443	0,0268	0,6397	0,0178	0,6846	0,0129	0,6762	0,0086	1.248
5	0,6291	0,0290	0,5826	0,0278	0,6713	0,0137	0,6728	0,0164	1.079

sobre o grupo de variáveis selecionadas, citadas anteriormente. Pelas Tabelas 18 e 19 é possível perceber que a multicolinearidade reduziu quando o conjunto de dados teve sua dimensionalidade reduzida. Isso significa que as variáveis removidas do conjunto de dados foram aquelas que possuiam maior correlação com as que ficaram retidas.

Tabela 18: Multicolinearidade de todas as variáveis

Variável	P	K	Mg	Ca	MO
FIV	1,7074	66,2920	1.023,4000	3.713,9000	6,5914

Variável	H_Al	pH	SB	CTC	V
FIV	6.311,8000	4,8839	10.582,0000	18.506,0000	52,8990

Variável	IC_0_5	IC_5_10	IC_10_15	IC_15_20	IC_20_25
FIV	3,2219	8,2913	7,1792	9,6675	12,5560

Variável	IC_25_30	IC_30_35	IC_35_40	CE_Deep	CE_Shallow
FIV	9,9812	12,0200	6,6914	3,6272	2,9853

Tabela 19: Multicolinearidade das variáveis selecionadas

Variável	P	K	MO	pH	Sb	CE_Shallow
FIV	1,2390	1,2625	3,5855	2,8421	2,1689	1,1276

4.2.1.2 Resultados sobre NB

Os resultados do Experimento 3 originários de classificadores *Naive Bayes* são apresentados na Tabela 20. É possível observar nesta tabela que, em média, as taxas de acerto mais altas foram aquelas obtidas dos modelos gerados pelo procedimento SVACPS. Para averiguar se aquelas médias diferiam entre os procedimento SVACP e SVACPS, foi realizado um teste T, para amostras pareadas, a nível de significância $\alpha = 0,05$. O teste permitiu verificar que a diferença entre as médias foi significativa. A Figura 14 mostra os resultados oriundos dos classificadores NB gerados sobre dados reais.

O modelo bayesiano que forneceu maior taxa de acerto, em média, foi aquele que possuiu 6 variáveis, as quais compuseram o conjunto P, MO, Ph, Sb, V e $CE\_Shallow$. Estas variáveis foram selecionados pelo SVACPS utilizando o critério B2. A referida taxa

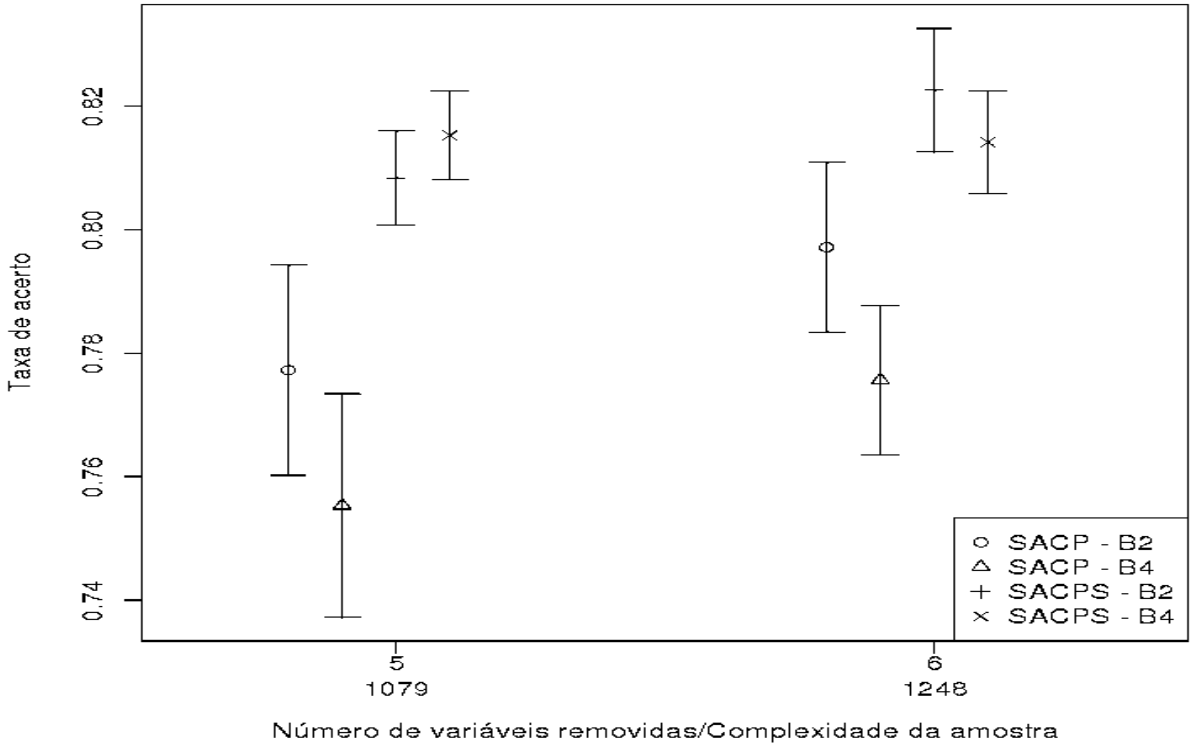

Figura 14: Gráfico dos resultados do Experimento 3 sobre NB

de acerto foi de $82,25\%$. A complexidade da amostra para aquele modelo foi de 1.248 registros.

Tabela 20: Resultados do Experimento 3 com NB

Taxa de Acerto do classificador NB									
	SVACP				**SVACPS**				
Nº Vars.	***B*2**		***B*4**		***B*2**		***B*4**		
#Ret.	μ	σ	μ	σ	μ	σ	μ	σ	m
6	0,7971	0,0137	0,7756	0,0121	0,8225	0,0100	0,8141	0,0083	1.248
5	0,7772	0,0170	0,7553	0,0181	0,8083	0,0076	0,8152	0,0072	1.079

Comparando a multicolinearidade do conjunto de variáveis selecionado, com a multicolinearidade do conjunto com todas as variáveis, é possível observar que a relação das variáveis foi amenizada, o que indica que as variáveis rejeitadas possuiam forte correlação com as variáveis que foram retidas.

Tabela 21: Multicolinearidade das variáveis selecionadas

Variável	P	MO	Ph	Sb	V	CE_Shallow
FIV	1,1306	4,9817	4,0778	4,3749	6,8202	1,1039

4.2.1.3 Resultados sobre RNA

A Tabela 22 apresenta os resultados do Experimento 3 referentes aos testes feitos com RNAs. As menores médias da soma dos erros quadrados são pertinentes aos modelos oriundos do procedimento SVACPS, exceto o modelo de 6 variáveis gerado pelo SVACPS utilizando o critério B4. Foi realizado um teste T, para amostras pareadas, para avaliar se a diferença entre as médias obtidas pelos modelos procedentes do SVACPS contra as dos modelos do SVACP, foram significativas. O teste evidenciou que as diferenças foram significativas, à $\alpha = 0,05$. Os resultados são ilustrados na Figura 15.

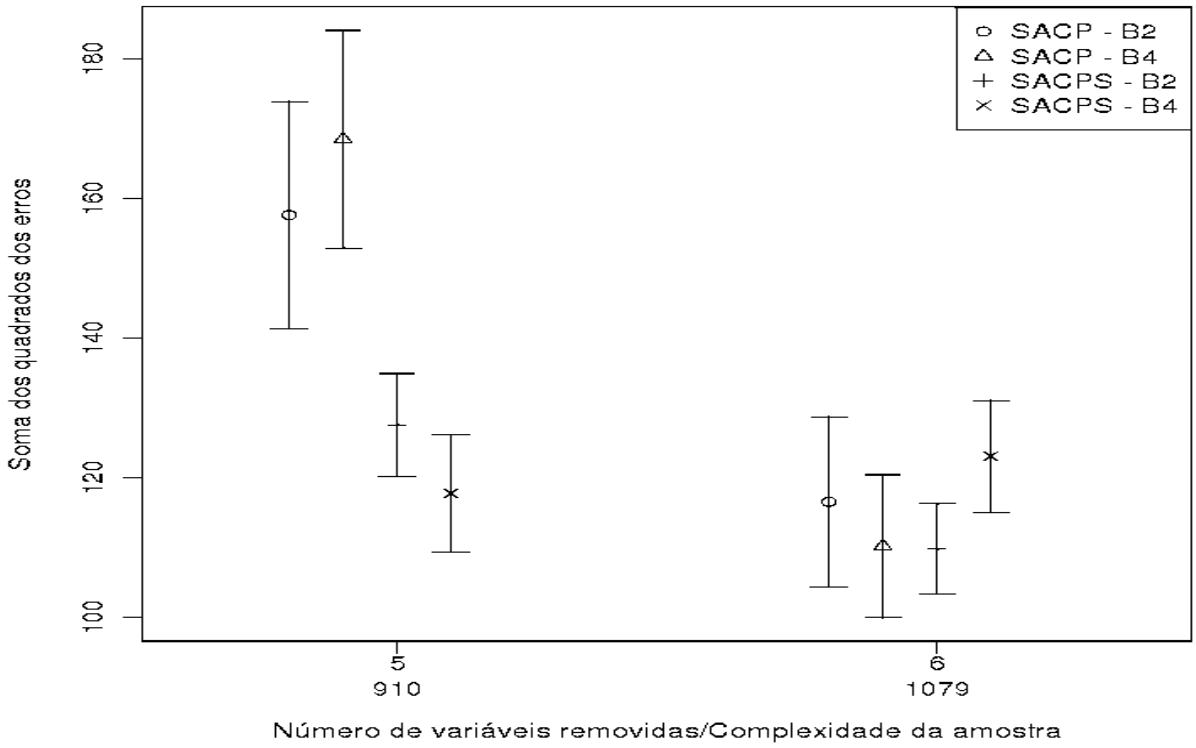

Figura 15: Gráfico dos resultados do Experimento 3 sobre RNA

Para que as RNAs respeitassem as restrições do aprendizado PAC, foi necessário selecionar subconjuntos de no máximo 6 variáveis. É importante observar que nem todos os menores erros foram obtidos pelas redes construídas com o procedimento SVACPS. Porém, o menor erro obtido foi conseguido com uma RNA construída pelo procedimento SVACPS com critério B2, cujo SQE médio foi $109,7526$. Este erro foi produzido por uma RNA que teve 6 variáveis de entrada, as quais formaram o conjunto P, Ca, MO, Ph, V e CE_Deep.

Pelas Tabelas 18 e 23 é possível notar que a multicolinearidade diminuiu quando a dimensionalidade do conjunto foi reduzida. Isto indica que o procedimento de seleção conseguiu rejeitar aquelas variáveis que possuiam maior correlação com as variáveis que

Tabela 22: Resultados do Experimento 3 com RNA

Soma dos Quadrados dos Erros do classificador RNA									
	SVACP				**SVACPS**				
Nº Vars.	***B*2**		***B*4**		***B*2**		***B*4**		
#Ret.	μ	σ	μ	σ	μ	σ	μ	σ	m
6	116,5220	12,1823	110,1580	10,2396	109,7526	6,4882	123,0404	8,0108	1.079
5	157,6165	16,2754	168,4776	15,6080	127,5329	7,3648	117,7466	8,4202	910

foram retidas.

Tabela 23: Multicolinearidade das variáveis selecionadas

Variável	P	Ca	MO	Ph	V	CE_Deep
FIV	1,0825	4,4076	5,1920	4,1110	6,9665	1,3220

4.2.2 Resultados do Experimento 4 - Dados Incompletos

Os resultados do Experimento 4 são apresentados nas Tabelas 24 e 26. As colunas destas tabelas são descritas como as das tabelas de resultados do Experimento 1. Em todos os casos, os critérios B2 e B4 usaram a técnica da ACPP para calcular os componentes principais, que utiliza o algoritmo *EM* para estimar os valores não observados do conjunto. Este foi completado antes da execução do processo de modelagem automática.

Os conjuntos de amostras utilizados neste experimento foram obtidos a partir do Conjunto de Dados 2, empregando o procedimento descrito no Apêndice A, que atribuiu o rótulo *não observado* a um atributo em 15% do total de casos do Conjunto de Dados 2. O objetivo deste experimento foi mostrar a possibilidade de usar os procedimentos SVACP e SVACPS em conjunto de dados reais que possui dados não observados (dados faltantes). As tabelas de resultados do Experimento 4 mostram resultados similares aos do Experimento 1, em que o procedimento SVACPS produz modelos cujo desempenho é maior do que dos modelos gerados pelo SVACP. A seguir são apresentados os resultados para cada tipo de modelo utilizado.

4.2.2.1 Resultados sobre MRLM

Os modelos cujos coeficientes de determinação foram maiores, de acordo com a Tabela 24, foram aqueles gerados pela SVACPS. Um teste T, para amostras pareadas, foi realizado para verificar se a diferenças entre as médias obtidas no experimento, entre os procedimentos SVACPS e SVACP, foram estatisticamente significativas. O teste T permitiu verificar que a diferença foi significativa. Estes resultados podem ser vistos na Figura 16.

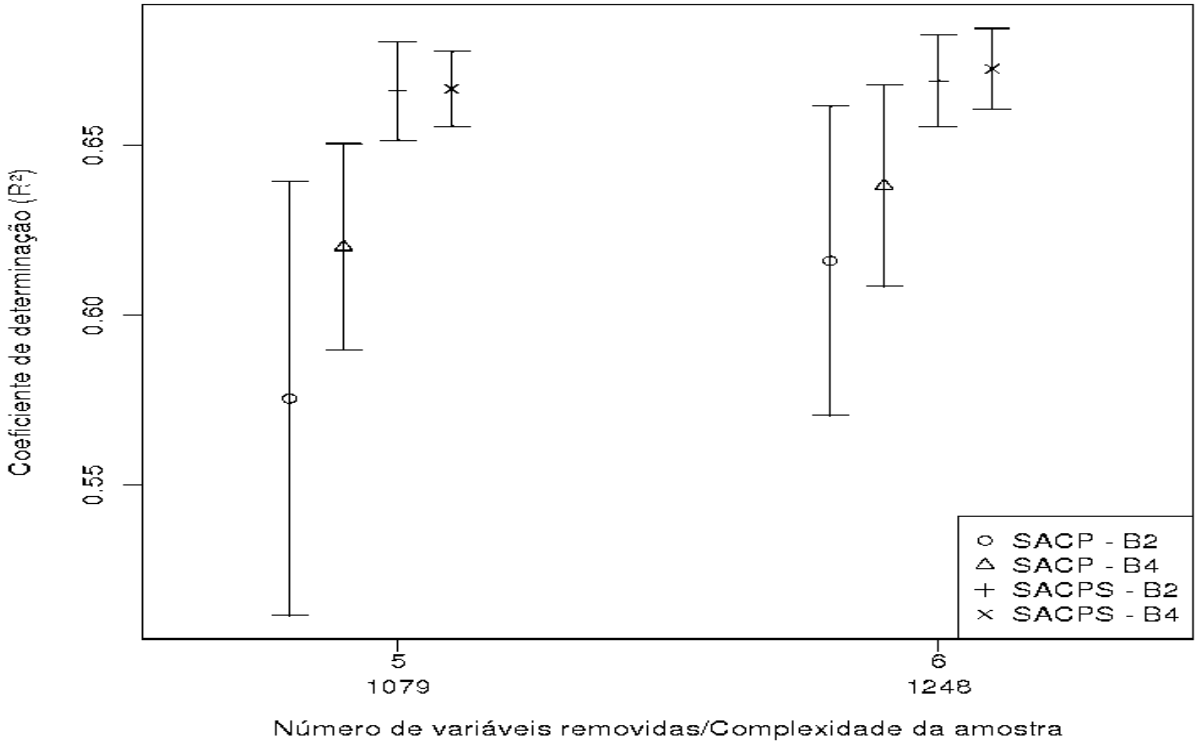

Figura 16: Gráfico dos resultados do Experimento 4 sobre MRLM

A maior correlação média obtida sobre o MRLM foi $0,6724$, produzida pelo modelo que possuiu 6 variáveis e complexidade da amostra de 1.248 registros. Este modelo foi conseguido pelo procedimento SVACPS executando o critério B4. As variáveis selecionadas por este procedimento foram P, MO, CTC, V, IC_35_40 e CE_Deep.

Tabela 24: Resultados do Experimento 4 com MRLM

R^2 do MRLM - 15% de dados incompletos									
	SVACP				**SVACPS**				
N° Vars.	***B2***		***B4***		***B2***		***B4***		
#Ret.	μ	σ	μ	σ	μ	σ	μ	σ	m
6	0,6159	0,0455	0,6380	0,0296	0,6688	0,0135	0,6724	0,0119	1.248
5	0,5754	0,0638	0,6200	0,0304	0,6659	0,0145	0,6665	0,0110	1.079

É possível observar, pelas Tabelas 18 e 23, que a multicolinearidade entre as variáveis selecionadas foi reduzida, em comparação com o índice FIV do conjunto contendo todas as variáveis. Isto mostra que as variáveis rejeitadas pelo procedimento SVACPS, foram aquelas que possuiam maior correlação com as variáveis que foram selecionadas.

Tabela 25: Multicolinearidade das variáveis selecionadas

Variável	P	MO	CTC	V	IC_35_40	CE_Deep
FIV	1,1292	5,5776	4,5751	1,8585	1,3252	1,3193

4.2.2.2 Resultados sobre NB

Na Tabela 26, pode ser observado que as maiores taxas de acerto, em média,foram obtidas por modelos gerados pela SVACPS. Um teste T, para amostras pareadas, foi realizado para verificar se a diferenças entre as médias obtidas no experimento, entre os procedimentos SVACPS e SVACP, foram estatisticamente significativas. O teste T permitiu verificar que a diferença foi significativa. Estes resultados podem ser vistos na Figura 17.

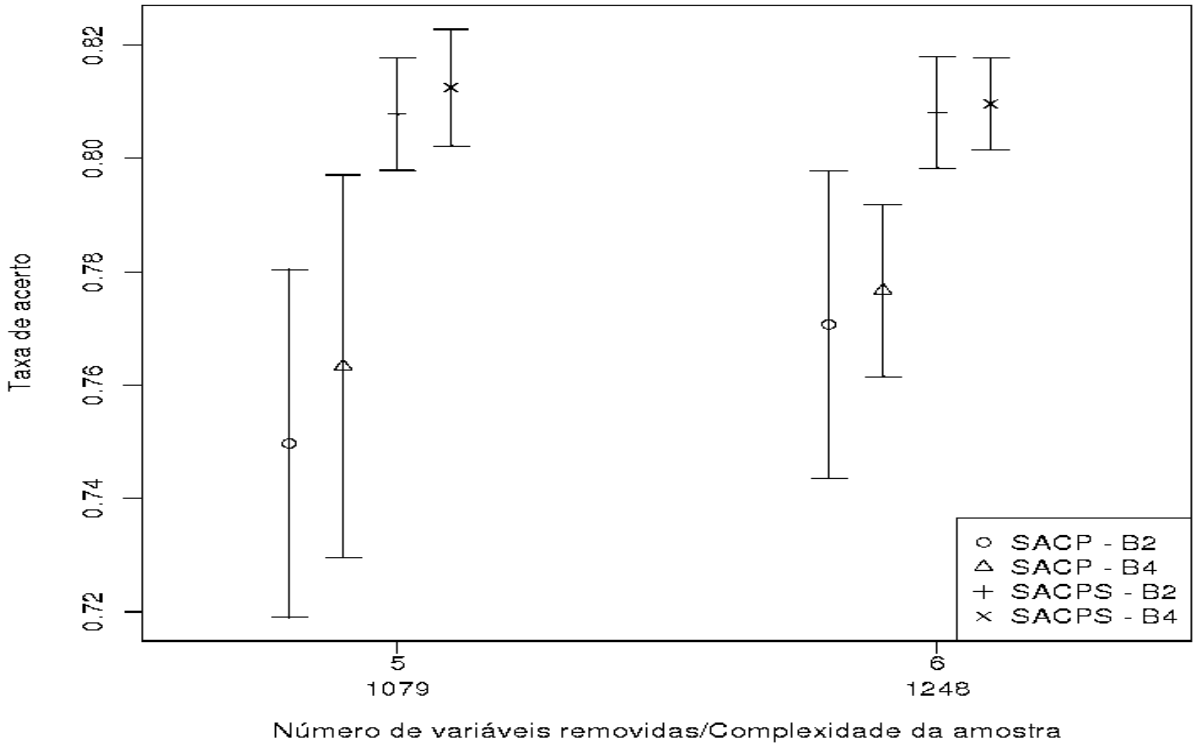

Figura 17: Gráfico dos resultados do Experimento 4 sobre NB

A maior taxa de acerto, em média, foi 81, 25%, produzida pelo modelo que possuiu 5 variáveis, cuja complexidade da amostra foi de 1.079 registros. Este modelo foi gerado pelo procedimento SVACPS utilizando o critério B4, o qual selecionou as variáveis P, K, MO, V e $CE\_Shallow$, para compor o classificador bayesiano.

Tabela 26: Resultados do Experimento 4 com NB

Taxa de Acerto do classificado NB - 15% de dados incompletos									
	SVACP				**SVACPS**				
N° Vars.	***B2***		***B4***		***B2***		***B4***		
#Ret.	μ	σ	μ	σ	μ	σ	μ	σ	m
6	0,7707	0,0271	0,7767	0,0152	0,8081	0,0098	0,8096	0,0081	1.248
5	0,7497	0,0307	0,7633	0,0338	0,8078	0,0099	0,8125	0,0103	1.079

Confrontando a Tabela 27 com a Tabela 18 é possível notar que ao reduzir a

dimensionalidade do conjunto de dados, a multicolinearidade também foi reduzida. O que confirma, mais uma vez, que o procedimento de seleção conseguiu rejeitar as variáveis preditoras que possuiam maior correlação com as que foram retidas.

Tabela 27: Multicolinearidade das variáveis selecionadas

Variável	P	K	MO	V	CE_Shallow
FIV	1,2216	1,2001	1,4373	1,3453	1,1232

4.3 RESULTADOS SOBRE O CONJUNTO DE DADOS 3

4.3.1 Resultados do Experimento 5

Os resultados do Experimento 5 são apresentados nas Tabelas 28 e 32. Nestas tabelas a primeira coluna indica o número de variáveis selecionadas pelos procedimentos SVACP e SVACPS. As colunas identificadas com B2 e B4 apresentam os coeficicientes de determinação dos modelos lineares, cujas variáveis preditoras foram selecionadas ou pelo critério B2 ou pelo critério B4 nos procedimentos SVACP ou SVACPS. A última coluna, identificada com m informa a complexidade da amostra segundo os critérios estabelecidos pelo aprendizado PAC, conforme descrito na Seção 3.3.2. Deve ser notado que neste experimento, a complexidade da amostra, embora seja mostrada, ela não é levada em consideração, pois o número de amostras no Conjunto de Dados 3 é menor do que o mínimo exigido pelo aprendizado PAC. Em todos os casos os métodos B2 e B4 usaram a técnica DVS para calcular os componentes principais.

4.3.1.1 Resultados sobre MRLM - Sem a Seleção Obrigatória da Variável Gesso

A Tabela 28 mostra o resultados quando da execução dos procedimentos SVACP e SVACPS sobre o Conjunto de Dados 3, sem restringir suas execuções de que a variável gesso não poderia ser eliminada. Pode ser observado naquela tabela, que quase todos os resultados produzidos pelos modelos provenientes do SVACPS, foram maiores do que aqueles produzidos pelos modelos provindos do SVACP. Apenas o modelos com 23 variáveis preditoras feitos pelo SVACP forneceram coeficientes de determinação superiores aos modelos de 23 variáveis feitos pelo SVACPS, tanto utilizando o critério B2 quanto com o B4.

Contudo, o modelo que produziu maior coeficiente foi aquele com possuiu 17 variáveis preditoras, oriundo do procedimento SVACPS utilizando critério B4. Naquele

modelo as variáveis selecionadas como preditoras foram *Gesso*, pH_B, Al_C, Al_D, Mg_A, Mg_B, Mg_C, Mg_D, K_A, K_B, K_D, m_C, m_D, Ca_CTCe_A, Ca_CTCe_B, Ca_CTCe_C e Ca_CTCe_D.

Tabela 28: Resultados do Experimento 5 com MRLM - sem a seleção obrigatória da variável Gesso

R^2 do MRLM					
Nº Vars.	**SVACP**		**SVACPS**		
#Ret.	***B*2**	***B*4**	***B*2**	***B*4**	**m**
23	0,9521	0,9521	0,8977	0,9360	4.123
22	0,9516	0,9516	0,9570	0,9565	3.954
21	0,9509	0,9509	0,9517	0,9514	3.785
20	0,9504	0,9504	0,9589	0,9551	3.616
19	0,9499	0,9489	0,9605	0,9681	3.446
18	0,9476	0,9466	0,9574	0,9473	3.277
17	0,9474	0,9464	0,9568	0,9739	3.108
16	0,9391	0,9464	0,9588	0,9671	2.939
15	0,9368	0,9445	0,9624	0,9731	2.770
14	0,9340	0,9443	0,9691	0,9584	2.601
13	0,9315	0,9169	0,9601	0,9634	2.432
12	0,9163	0,8917	0,9604	0,9675	2.263
11	0,8808	0,8914	0,9720	0,9500	2.093
10	0,8808	0,8909	0,9560	0,9541	1.924
9	0,8132	0,8440	0,9610	0,9516	1.755
8	0,8117	0,8372	0,9489	0,9260	1.586
7	0,6075	0,7243	0,9508	0,9370	1.417
6	0,6009	0,6921	0,9358	0,9253	1.248
5	0,5024	0,6784	0,9149	0,9313	1.079
4	0,4972	0,4848	0,8755	0,9012	910
3	0,4650	0,2128	0,8605	0,8735	741
2	0,2156	0,2100	0,8794	0,8725	571

Pode ser observado nas Tabelas 29 e 30 que a multicolinearidade diminuiu entre as variáveis que foram selecionadas em relação ao índice FIV medido sobre o conjunto com todas as variáveis. Isso mostra que as variáveis que foram removidas do conjunto foram aquelas que possuiram correlações com as variáveis que foram retidas.

4.3.1.2 Resultados sobre MRLM - Com a Seleção Obrigatória da Variável Gesso

É possível verificar que quando a variável gesso foi selecionada manualmente, os resultados foram diferentes. Pode ser observado pela Tabela 32 que na maioria dos casos o procedimento SAPCS forneceu modelos que produziram resultados cujos coeficientes de determinação foram maiores do que os dos resultados produzidos pelos modelos provindos do SVACP. Exceto o modelo que possuiu 23 variáveis construído pelo SVACPS com critério B4, no qual o coeficiente de determinação foi menor do que o produzido pelo SVACP para o mesmo critério de seleção.

Tabela 29: Multicolinearidade de todas as variáveis

Variável	gesso	pH_A	pH_B	pH_C	pH_D
FIV	3,0712	3,2186	7,4626	3,2842	6,2075

Variável	Al_A	Al_B	Al_C	Al_D	Mg_A
FIV	54,4585	42,5444	50,7953	24,8376	17,4143

Variável	Mg_B	Mg_C	Mg_D	K_A	K_B
FIV	30,8799	40,7761	22,9870	5,4106	7,1355

Variável	K_C	K_D	m_A	m_B	m_C
FIV	7,1098	4,2312	82,8495	85,3534	108,6054

Variável	m_D	Ca_CTCe_A	Ca_CTCe_B	Ca_CTCe_C	Ca_CTCe_D
FIV	46,1873	27,6278	31,7044	22,8382	16,6321

Tabela 30: Multicolinearidade das variáveis selecionadas

Variável	gesso	pH_B	Al_C	Al_D	Mg_A	Mg_B
FIV	2,5536	1,8924	33,9502	18,9852	7,1314	5,3565

Variável	Mg_C	Mg_D	K_A	K_B	K_D	m_C
FIV	28,5751	20,6570	3,8944	4,4287	1,6728	65,2110

Variável	m_D	Ca_CTCe_A	Ca_CTCe_B	Ca_CTCe_C	Ca_CTCe_D	
FIV	35,3734	10,4700	7,7760	15,4616	15,6130	

O modelo que produziu o maior coeficiente dentre os testados, foi aquele cujo número de variáveis selecionadas foi 14. As variáveis selecionadas para este modelo foram $Gesso, pH\_B, Al\_C, Mg\_A, Mg\_B, Mg\_C, Mg\_D, K\_A, m\_C, m\_D, Ca\_CTCe\_A$, Ca_CTCe_B, Ca_CTCe_C e Ca_CTCe_D.

A Tabela 31, mostra que os índices FIV das variáveis que foram selecionadas, foram reduzidos em relação ao FIV calculado sobre todas as variáveis do Conjunto de Dados 3. Esse fato permite observar que os procedimentos conseguiram rejeitar as variáveis mais correlacionadas linearmente com aquelas que foram retidas.

Tabela 31: Multicolinearidade das variáveis selecionadas

Variável	gesso	pH_B	Al_C	Mg_A	Mg_B
FIV	2,3742	1,6961	28,0958	6,1713	4,7720

Variável	Mg_C	Mg_D	K_A	m_C	m_D
FIV	24,9917	5,0563	1,4140	54,3671	4,3842

Variável	Ca_CTCe_A	Ca_CTCe_B	Ca_CTCe_C	Ca_CTCe_D	
FIV	9,4068	6,4284	13,0722	8,0617	

Tabela 32: Resultados do Experimento 5 - selecionando obrigatóriamente a variável Gesso

R^2 do MRLM					
Nº Vars.	**SVACP**		**SVACPS**		
#Ret.	***B2***	***B4***	***B2***	***B4***	**m**
23	0,9521	0,9521	0,9610	0,9438	4.123
22	0,9516	0,9516	0,9464	0,9388	3.954
21	0,9509	0,9509	0,9622	0,9694	3.785
20	0,9504	0,9504	0,9552	0,9801	3.616
19	0,9499	0,9489	0,9580	0,9533	3.446
18	0,9476	0,9466	0,9698	0,9680	3.277
17	0,9474	0,9464	0,9552	0,9440	3.108
16	0,9391	0,9464	0,9599	0,9608	2.939
15	0,9368	0,9445	0,9684	0,9502	2.770
14	0,9340	0,9443	0,9643	0,9744	2.601
13	0,9315	0,9169	0,9549	0,9632	2.432
12	0,9304	0,9139	0,9726	0,9543	2.263
11	0,9068	0,9139	0,9498	0,9631	2.093
10	0,9068	0,9139	0,9721	0,9718	1.924
9	0,8383	0,8728	0,9484	0,9609	1.755
8	0,8347	0,8594	0,9419	0,9699	1.586
7	0,7037	0,8090	0,9162	0,9446	1.417
6	0,7014	0,7878	0,9330	0,9221	1.248
5	0,5880	0,7875	0,9307	0,9208	1.079
4	0,5850	0,5871	0,9010	0,9427	910
3	0,3529	0,3342	0,8543	0,8747	741
2	0,3155	0,3155	0,8683	0,7797	571

4.4 CONSIDERAÇÕES FINAIS

Este capítulo apresentou os resultados do uso dos procedimentos SVACP e SVACPS para a seleção de variáveis sobre conjuntos de dados sintéticos e agrícolas completos. Os resultados mostraram que o procedimento SVACPS forneceu, em média, resultados melhores do que o procedimento SVACP, tanto utilizando o critério B2 quanto o B4. Exceto quando foi aplicado RNA sobre dados agrícolas, em que foram gerados quatro resultados. Destes, o procedimento SVACPS produziu três resultados melhores do que o SVACP. Também foi possível observar que informações referentes à complexidade da amostra podem ser úteis, para analisar os resultados dos procedimentos propostos.

Os procedimentos também foram empregados para redução de dimensionalidade em bases de dados incompletas. Neste caso, a ACPP foi usada para calcular os componentes principais. Os resultados dos experimentos tanto sobre dados sintéticos quanto sobre dados agrícolas mostraram que a SVACPS produziu melhores resultados que a SVACP, tanto com o critério B2 quanto com o B4.

A multicolinearidade permitiu verificar que quanto menos variáveis forem selecionadas pelos procedimentos, menores são as chances de ocorrer multicolinearidade entre

elas. Embora em alguns experimentos tenham sido usados classificadores, a multicolinearidade é uma propriedade intrínseca dos dados e não depende do tipo de modelo utilizado. Portanto, esta informação pode ser útil mesmo se tratando da indução de classificadores.

Neste sentido, Lam (2010) comenta que quando o índice FIV tem valor menor ou igual a 10, a multicolinearidade pode ser aceitável. Embora o desejável é que tal índice tenha valor de no máximo 4. Pode ser observado que em relação ao Conjunto de Dados 1, os procedimentos SVACP e SVACPS selecionaram grupos de variáveis para os quais o índice FIV ficou abaixo do valor 4, ou seja, na faixa de índice *aceitável*. Exceto para as variáveis x_9 e x_{10}, selecionadas no Experimento 1 para compor o classificador NB, que forneceu a maior taxa de acerto. Em relação ao Conjunto de Dados 2, não foram todos os grupos de variáveis para os quais o índice FIV ficou abaixo do valor 10. Porém, a redução da multicolinearidade foi considerada satisfatória, já que algumas variáveis tiveram redução no índice FIV de cerca de 1.000 vezes, em relação ao FIV calculado sobre todas as variáveis do conjunto.

Os resultados do último experimento mostraram que os procedimentos SVACP e SVACPS podem ser usados em uma situação particular: quando uma ou mais variáveis devem ser selecionadas, obrigatoriamente. A vantagem de tal funcionalidade é permitir que o pesquisador não elimine as variáveis que fazem parte do protocolo experimental, do qual os dados foram obtidos. Se observados os resultados dos dois testes do experimento 5 (sem e com seleção obrigatória), é possível verificar que os resultados dos modelos foram diferentes, mesmo com o *Gesso* não sendo eliminado do conjunto no primeiro teste. Isto mostra que a pré-seleção é uma funcionalidade importante, pois, causa uma alteração no espaço de busca, permitindo obter regressores e classificadores diferentes daqueles obtidos sem o recurso de pré-seleção.

5 CONCLUSÃO

Este trabalho apresentou dois procedimentos para a seleção de variáveis: o SVACP e o SVACPS. Eles podem ser usados na etapa de pré-processamento e em tarefas que envolvam a descoberta de conhecimento em banco de dados. A principal motivação para o desenvolvimento destes procedimentos foi prover estratégias para seleção de variáveis relevantes, para tarefas de regressão e indução de classificadores. Para cada conjunto de variáveis selecionadas, os procedimentos SVACP e SVACPS informam as exigências referentes a complexidade da amostra. Este tipo de informação é relevante, pois, a confiabilidade de modelos gerados por indução de classificadores ou regressão numérica, a partir de dados coletados em pesquisa experimental, depende do tamanho e das características da amostra. Este tipo de tratamento de dados é usualmente empregado na pesquisa experimental na agricultura e, portanto, os procedimentos SVACP e SVACPS podem auxiliar na seleção de variáveis quando do emprego de métodos multivariados em agricultura.

O procedimento SVACP não explora dados rotulados e seleciona as variáveis mais relevantes, considerando apenas a variação contida nos dados, por meio da análise de componentes principais. O procedimento também permite que os requisitos relativos à complexidade da amostra, segundo a teoria do aprendizado PAC, também sejam considerados na seleção de variáveis. O procedimento SVACPS é capaz de usar a supervisão para selecionar as variáveis em duas etapas: primeiramente, ele seleciona as variáveis que possuem maior influência no comportamento da variável de resposta e, então, forma um conjunto de dados reduzido; em seguida, sobre este conjunto reduzido, ele seleciona as variáveis que possuem maior influência sobre a variação observada nos dados. Ambos os procedimentos permitem a utilização de dois critérios de seleção não supervisionados e baseados em ACP, chamados de B2 e B4. Eles também permitem a utilização de outras técnicas de computação dos componentes principais além da técnica DVS, como a ACPP para a estimação de dados incompletos.

Os resultados mostraram que nos conjuntos de dados utilizados nos experimentos, a SVACPS foi em média mais efetiva na seleção de variáveis, quando o objetivo foi realizar regressão multivariada tanto em dados agrícolas quanto em dados sintéticos. O mesmo comportamento foi observado para Redes Neurais Artificiais e classificadores do tipo *Naive Bayes*. Sobre o conjunto de dados sintéticos (Conjunto de Dados 1), a SVACPS selecionou conjuntos de variáveis que geraram os melhores regressores e classificadores em todos os testes. Sobre o primeiro conjunto de dados agrícolas (Conjunto de Dados 2), a SVACPS selecionou grupos de variáveis que geraram os melhores resultados na maioria dos casos.

Exceto nos testes com RNAs, no Experimento 3, em que três dos quatros melhores resultados obtidos foram produzidos pela SVACPS e um pela SVACP. Sobre o segundo conjunto de dados agrícolas (Conjunto de Dados 3), o SVACPS também selecionou variáveis que produziram a maioria dos resultados mais efetivos.

Os resultados também mostraram que tanto a SVACP quanto a SVACPS determinaram subconjuntos de variáveis, os quais tinham menor grau de multicolinearidade, quando comparada com o conjunto original. Isto mostra que eles conseguiram eliminar as variáveis que eram correlacionadas àquelas que foram selecionadas ou, ainda, eliminar aquelas variáveis que não contribuíam com a variação dos dados usados nos experimentos. Para modelos de regressão e classificadores, variáveis correlacionadas duplicam informações e, portanto, não agregam distinção nos resultados do modelo. Também, variáveis com pouca contribuição sobre a variação dos dados causam pouca influência significativa no comportamento de um modelo, não melhorando significativamente a resposta produzida.

Trabalhos Futuros

Como trabalhos futuros, podem ser propostos:

- a realização de mais testes em outros conjuntos de dados agrícolas, considerando testes baseados na técnica de validação cruzada, outros modelos de regressão e outros tipos de classificadores. Tais testes ajudarão a verificar a efetividade dos procedimentos quando submetidos a uma técnica de validação mais restritiva;

- a incorporação de testes estatísticos durante a seleção de variáveis para a validação dos grupos selecionados em tempo de execução. O objetivo disso é validar o grupo de variáveis selecionado quando o pesquisador conhece a distribuição estatística dos dados;

- a pesquisa de novos coeficientes que poderão ser utilizados como critério de supervisão, definindo a influência das variáveis preditoras sobre a variável de resposta. Isso é importante, pois, informaria a correlação adequada que deve ser considerada entre as variáveis preditores e a de resposta.

- a incorporação de heurísticas na seleção de variáveis para agilizar o processo de seleção, pois, os procedimentos atualmente executam busca exaustiva e isso torna o processo de seleção mais lento quando há a necessidade de processar conjuntos com um grande número de variáveis (por exemplo, 1.000 variáveis);

REFEÊNCIAS

AFIFI, A. A.; AZEN, S. P. *Statistical analysis: a computer oriented approach.* [S.l.]: Academic Press, 1979.

ALVARENGA, M. I. N.; DAVIDE, A. C. Características físicas e químicas de um Latossolo Vermelho-Escuro e a sustentabilidade de agroecossistemas. *Revista Brasileira de Ciência do Solo*, v. 23, n. 4, p. 933 – 942, 1999.

ARMSTRONG, L. J.; DIEPEVEEN, D.; MADDERN, R. The application of Data Mining techniques to characterize agricultural soil profiles. In: AUSTRALASIAN CONFERENCE ON DATA MINING AND ANALYTICS, 6., Gold Coast, Australia. *Anais...* Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2007. (AusDM '07, v. 70), p. 85–100.

BAIR, E. *et al.* Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, v. 101, p. 119–137, 2006.

BISHOP, C. M. Latent variable models. In: _____. *Learning in graphical models.* Cambridge, MA, USA: MIT Press, 1999. p. 371–403. ISBN 0-262-60032-3.

CAIRES, E. F. *et al.* Alterações químicas do solo e resposta da soja ao calcário e gesso aplicados na implantação do sistema plantio direto. *Revista Brasileira de Ciência do Solo*, scielo, v. 27, p. 275 – 286, 04 2003.

CARVALHO, M. J. R. de. *A estatística na experimentação agrícola.* [S.l.]: Livraria Sá da Costa, 1946. (Terra e o homen, coleção de livros agrícolas: Os fundamentos das ciências agrárias).

DUNTEMAN, G. H. *Principal Components Analysis.* [S.l.]: Sage Publications, 1989. (Quantitative Applications in the Social Sciences).

EL-TELBANY, M. E.; WARDA, M.; EL-BORAHY, M. Mining the classification rules for egyptian rice diseases. *Int. Arab J. Inf. Technol.*, p. 303–307, 2006.

FERREIRA, D. F. *Estatística Multivariada.* 1. ed. Lavras, MG: Ed. UFLA, 2008. 662 p. ISBN 978-85-87692-52-8.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: An overview. *AI Magazine*, Association for the Advancement of Artificial Intelligence, v. 13, n. 3, 1992.

GOMEZ, K. A.; GOMEZ, A. A. *Statistical Procedures for Agricultural Research.* [S.l.]: Wiley, 1984. (An International Rice Research Institute book).

GONÇALVES, D. *Física.* 2ª. ed. Rio de Janeiro: Ao Livro Técnico, 1965. 250p.

GUIMARÃES, A. M. *Aplicação de Computação Evolucionária na mineração de dados físico-químicos da água e do solo*. Tese (Doutorado) — Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, 12 2005. Tese de Doutorado.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: concepts and techniques*. [S.l.]: Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems).

HAYKIN, S. S. *Redes Neurais - Principios e Prática*. [S.l.]: Bookman Companhia, 2001. ISBN 9788573077186.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Blackwell Publishing for the Royal Statistical Society, v. 21, n. 2, p. 160 – 173, 1972. ISSN 00359254.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. ii: Real data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Blackwell Publishing for the Royal Statistical Society, v. 22, n. 1, p. 21 – 31, 1973. ISSN 00359254.

JOLLIFFE, I. T. *Principal Components Analysis*. 2. ed. New York: Springer Verlag, 2002. 487 p. ISBN 0-387-95442-2.

KARKACIER, O.; GOKTOLGA, Z. G.; CICEK, A. A regression analysis of the effect of energy use in agriculture. *Energy Policy*, v. 34, n. 18, p. 3796 – 3800, 2006.

KING, J. R.; JACKSON, D. A. Variable selection in large environmental data sets using Principal Components Analysis. *Environmetrics*, Department of Fisheries and Oceans, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, British Columbia, Canada V9R 5K6; Department of Zoology, University of Toronto, 25 Harbord Street, Toronto, Ontario, Canada M5S 1A1, v. 10, n. 1, p. 67 – 77, 1999.

KUMAR, V. *et al. Introdução ao Data Mining - Mineração de Dados*. Rio de Janeiro: Ciência Moderna, 2009.

KUMMER, L. *et al.* Uso da análise de componentes principais para agrupamento de amostras de solos com base na granulometria e em características químicas e mineralógicas. *Scientia Agraria*, v. 11, n. 6, 2010.

LAM, L. *An introduction to R*. [S.l.], 2010. 212 p. Acesso em: 30/07/2012.

LIBRALON, G. L. *Investigação de combinações de técnicas de detecção de ruído para dados de expressão gênica*. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2007. Dissertação de Mestrado.

LOBELL, D. B.; ASNER, G. P. Climate and management contributions to recent trends in U.S. agricultural yields. *Science*, v. 299, n. 5609, p. 1032, 2003. ISSN 00368075.

LUNA, J. E. O. *Algoritmos EM para aprendizagem de Redes Bayesianas a partir de dados incompletos*. Dissertação (Mestrado) — Universidade Federal de Mato Grosso do Sul, 06 2004. Dissertação de Mestrado.

MARTINS, V. A.; FONSECA, L. M. G. Classificação de uso de solo baseada na análise orientada a objeto e mineração de dados utilizando imagens SPOT/HRG-5. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO (SBSR), 14., 2006, Natal. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2009. p. 7847–7844.

MATHIAS, I. M. *Aplicação de redes neurais artificiais na análise de dados de molhamento foliar por orvalho*. Tese (Doutorado) — Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu - SP, 12 2006. Tese de Doutorado.

MEAD, R.; CURNOW, R. N.; HASTED, A. M. *Statistical Methods in Agriculture and Experimental Biology*. [S.l.]: Chapman & Hall/CRC, 2003. (Texts in Statistical Science).

MITCHELL, T. *Machine Learning*. New York: McGraw-Hill, 1997.

MOLIN, J. P. *et al.* Regression and correlation analysis of grid soil data versus cell spatial data. In: EUROPEAN CONFERENCE ON PRECISION AGRICULTURE: AGRO MONTPELLIER, 3., 2001. *Anais...* [S.l.], 2001. p. 449 – 453.

MYERS, S. W. *et al.* Effect of soil potassium availability on soybean aphid (hemiptera: Aphididae) population dynamics and soybean yield. *Journal of Economic Entomology*, Entomological Society of America, v. 98, n. 1, p. 113–120, 2005.

NETO, P. L. O. C. *Estatística*. [S.l.]: Editora Edgard Blücher, 2002.

PILLAR, V. D. P. Suficiência amostral. In: _____. *Amostragem em Limnologia*. São Carlos, RS: Editora Rima, 2004. p. 25 – 403.

RIEDMILLER, M.; BRAUN, H. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In: *IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS*. [S.l.: s.n.], 1993. p. 586 – 591.

ROBERTS, S.; MARTIN, M. A. Using Supervised Principal Components Analysis to assess multiple pollutant effects. *Environmental Health Perspectives*, National Institute of Environmental Health Sciences, v. 114, n. 12, p. 1877 – 1882, 08 2006.

ROSSEL, R. A. V. *et al.* On the soil information content of visible-near infrared reflectance spectra. *European Journal of Soil Science*, Blackwell Publishing Ltd, v. 62, n. 3, p. 442–453, 2011.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. [S.l.]: Prentice Hall, 2004. (Prentice Hall series in artificial intelligence). ISBN 9780131038059.

SANTOS, J. S. dos; SANTOS, M. L. P. dos; CONTI, M. M. Comparative study of metal contents in Brazilian coffees cultivated by conventional and organic agriculture applying Principal Component Analysis. *Journal of the Brazilian Chemical Society*, scielo, v. 21, p. 1468 – 1476, 00 2010. ISSN 0103-5053.

SENA, M. *et al.* Discrimination of management effects on soil parameters by using principal component analysis: a multivariate analysis case study. *Soil and Tillage Research*, v. 67, n. 2, p. 171 – 181, 2002.

SHAO, X.; CHERKASSKY, V.; LI, W. Measuring the vc-dimension using optimized experimental design. *Neural Computation*, v. 12, p. 2000, 1969.

SHTANGEEVA, I. *et al.* Multivariate statistical analysis of nutrients and trace elements in plants and soil from northwestern russia. *Plant and Soil*, Springer Netherlands, v. 322, n. 1, p. 219–228, 2009.

SHUAI, S.; HONG, C. Influencing factors regression analysis of low-carbon agriculture in Heilongjiang Province. In: INTERNATIONAL CONFERENCE ON MANAGEMENT SCIENCE AND ENGINEERING, 2011. *Anais...* [S.l.], 2011. (ICMSE), p. 709 –714.

SLATTERY, S.; CRAVEN, M. Combining statistical and relational methods for learning in hypertext domains. In: *In Proceedings of the 8th international Conference on Inductive Logic Programming*. [S.l.]: Springer Verlag, 1998. p. 38–52.

SOUNIS, E. L. de M. *Bioestatística: princípios fundamentais, metodologia estatística: Aplicação às ciências biológicas*. Curitiba: UFPR, 1971.

THORNLEY, J.; JOHNSON, I. *Plant and crop modelling: a mathematical approach to plant and crop physiology*. [S.l.]: Clarendon Press, 1990. (Oxford science publications). ISBN 9780198541608.

VAMANAN, R.; RAMAR, K. Classification of agricultural land soils a data mining approach. *Agricultural Journal*, v. 6, n. 3, p. 82 – 86, 2011.

VANDENBERG, C. E. L. R. J. *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences*. [S.l.]: Routledge, 2009.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques*. [S.l.]: Morgan Kaufman, 2005. (Morgan Kaufmann Series in Data Management Systems).

WOOLDRIDGE, J. M. *Introductory econometrics: a modern approach*. [S.l.]: South Western, Cengage Learning, 2009.

YU, S. *et al.* Supervised Probabilistic Principal Component Analysis. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 12., 2006, Philadelphia, PA, USA. *Anais...* New York, NY, USA: ACM Press, 2006. (KDD '06), p. 464–473.

ZHANG, Y. Smart pca. In: *IJCAI*. [S.l.: s.n.], 2009. p. 1351 – 1356.

APÊNDICE A – GERAÇÃO DE CONJUNTO DE DADOS INCOMPLETO

O procedimento a seguir foi escrito na linguage R e permite manipular um conjunto de dados já existente para criar um novo conjunto de dados simulando dados incompletos. A simulação de dados incompletos ocorre apenas nas variáveis preditoras. A variável de resposta é considerada completa para os experimentos deste trabalho.

O procedimento recebe como parâmetros um conjunto de dados X, na forma de matriz de dados, e uma porcentagem P do total de registros, que indica quantos dos registros devem receber o rótulo de *não observado* (*NO*) em alguma variável. A variável R é inicializado com o número de registros de X. A variável C é inicializada com o número de colunas de X, mas descontando a variável de resposta. A variável RP armazena quantos registros devem ser marcados como *não observados*. A variável I guarda uma lista de índices dos registros que devem receber o rótulo *NO*. A variável J guarda uma lista de colunas, que representam as variável do conjunto, que devem receber o referido rótulo.

O laço faz com que RP registros recebam o rótulo *NO* exatamente na linha I_Z da coluna J_Z da matriz de dados. Deve ser observado que o rótulo *NO* é representado na linguagem R pela palavra reservada *NA* (*Not Available*). Por último, a matriz X é retornada. A seguir é aprensetado o código-fonte:

```
Insere_NO = function ( X , P ) {
  R = nrow ( X );
  C = ncol ( X ) - 1;
  RP = R * P;
  I = sample ( R , RP);
  J = sample ( C , RP , replace=T );

  for (Z in 1:length( I )){ X[ I[Z] , J[Z] ] = NA; }

  return ( X );
}
```

Figura 18: Código-fonte em R para rotular dados como incompletos

APÊNDICE B – CONJUNTO DE DADOS UTILIZADO NOS EXEMPLOS DA METODOLOGIA

O conjunto de dados utilizados nos dois exemplos, apresentados na metodologia, foi gerado utilizando o seguinte código-fonte em R:

```
z_1 = rnorm ( 600 );
z_2 = rnorm ( 600 );
z_3 = rnorm ( 600 );
z_4 = rnorm ( 600 );

x_1 = z_1;
x_2 = z_2;
x_3 = z_3;
x_4 = z_4;
x_5 = z_1 - z_4;
x_6 = z_2 - z_3;
x_7 = z_3 + z_4;

y = (10 * x_1) - (6 * x_2);

W = as.data.frame (
            cbind ( x_1=x_1 , x_2=x_2 , x_3=x_3 , x_4=x_4 ,
                    x_5=x_5 , x_6=x_6 , x_7=x_7 , y=y )
                    );
```

Com este código, foram geradas quatro variáveis latentes (z_1, ..., z_4), as quais compuseram as variáveis de x_1 até x_7. As variáveis x_1, x_2, x_3 e x_4 foram modeladas para serem independentes entre si. As variáveis x_5, x_6 e x_7 foram modeladas para serem dependentes das variáveis x_1, x_2, x_3 e x_4. Embora o conjunto possuisse 7 variáveis independentes, a variável de resposta y foi modelada apenas com x_1 e x_2. Assim, ao realizar qualquer procedimento de seleção de variáveis, esperou-se que no conjunto das variáveis selecionadas estivessem presentes x_1 e x_2. A seguir, é apresentado um trecho do conjunto de dados utilizado nos exemplos:

Tabela 33: Conjunto de dados utilizados nos exemplos de seleção de variáveis

x_1	x_2	x_3	x_4	x_5	x_6	x_7	y	
1	-0.72	-0.08	0.99	0.33	-1.05	-1.08	1.32	-6.72
2	1.15	0.43	-1.34	-0.89	2.04	1.77	-2.23	8.98
3	1.63	0.77	0.71	0.03	1.61	0.07	0.73	11.67
4	0.22	-0.89	-0.12	-0.25	0.47	-0.77	-0.37	7.56
5	-0.50	-0.79	-1.33	1.28	-1.79	0.54	-0.05	-0.28
				⋮				
556	-1.06	0.01	1.92	-0.34	-0.71	-1.91	1.58	-10.65
557	0.49	0.67	0.33	-0.97	1.47	0.34	-0.65	0.91
558	0.85	-0.79	-0.04	1.38	-0.53	-0.75	1.34	13.21
599	0.66	-0.63	-0.47	-1.23	1.89	-0.16	-1.71	10.35
600	-0.50	1.45	-1.91	-0.78	0.27	3.35	-2.68	-13.72