UNIVERSIDADE ESTADUAL DE PONTA GROSSA
MESTRADO EM COMPUTAÇÃO APLICADA

SANDRO TEIXEIRA

DETERMINAÇÃO DE MODELO DE ESTIMATIVA DE TEORES DE CARBONO EM SOLOS UTILIZANDO MÁQUINA DE VETOR DE SUPORTE E REFLECTÂNCIA ESPECTRAL

PONTA GROSSA
2014

SANDRO TEIXEIRA

DETERMINAÇÃO DE MODELO DE ESTIMATIVA DE TEORES DE CARBONO EM SOLOS UTILIZANDO MÁQUINA DE VETOR DE SUPORTE E REFLECTÂNCIA ESPECTRAL

Dissertação apresentada no Programa de Pós-Graduação em Computação Aplicada, curso de Mestrado em Computação Aplicada da Universidade Estadual de Ponta Grossa, como requisito para obtenção do título de Mestre.

Orientação: Drª Alaine Margarete Guimarães

PONTA GROSSA
2014

Ficha Catalográfica
Elaborada pelo Setor de Tratamento da Informação BICEN/UEPG

Teixeira, Sandro
T266 Determinação de modelo de estimativa de teores de carbono em solos utilizando máquina de vetor de suporte e reflectância espectral/ Sandro Teixeira. Ponta Grossa, 2014.
62f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.
Orientadora: Profª Drª Alaine Margarete Guimarães.

1.Carbono. 2.Agricultura de precisão. 3.Espectroscopia. 4.Aprendizado de máquina. 5.SVM. I.Guimarães, Alaine Margarete. II. Universidade Estadual de Ponta Grossa. Mestrado em Computação Aplicada. III. T.

CDD: 631

TERMO DE APROVAÇÃO

Sandro Teixeira

"DETERMINAÇÃO DE MODELO DE ESTIMATIVA DE TEORES DE CARBONO EM SOLOS UTILIZANDO MÁQUINA DE VETOR DE SUPORTE E REFLECTÂNCIA ESPECTRAL."

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Orientadora: *Dra. Alaine Margarete Guimarães*
UEPG

Dr. Vanderlei Porfírio da Silva
Embrapa

Dr. Ivo Mario Mathias
UEPG

AGRADECIMENTOS

A Deus por ter me dado saúde durante esses meses para desenvolver este trabalho.

A minha orientadora Dra. Alaine Margarete Guimarães pelo companheirismo e por ter me passado sua experiência e conhecimento, sempre incentivando a desenvolver o trabalho da melhor forma possível.

A Fundação ABC na pessoa do Carlos Alberto Proença, por todo o apoio, disponibilizando o Laboratório de Solos para análises das amostras.

A Universidade Estadual de Ponta Grossa por ter cedido o Laboratório Multiusuário também disponibilizando todos os recursos necessários para esta pesquisa.

Ao meu amigo Juscelino, que por meio da sua experiência me ajudou em algumas etapas.

Obrigado!

RESUMO

Considerado um indicador de qualidade, o carbono constitui-se em um importante atributo na capacidade produtiva do solo. Porém, as tradicionais metodologias empregadas para sua determinação geram problemas ambientais devido ao uso de reagentes químicos. Diante disso, a substituição desse procedimento por outros que gerem menor ou nenhuma quantidade de resíduos tóxicos tem sido considerada relevante. A espectroscopia é uma das técnicas promissora na Agricultura de Precisão para análises de solos e que pode trazer uma solução viável para análise de teor de carbono. Dentre suas vantagens, destaca-se a preservação da amostra, o não consumo de reagentes, além de sua eficiência na aquisição de dados provenientes de um grande número de amostras. O objetivo deste trabalho foi contribuir com um modelo de regressão capaz de predizer a quantidade de carbono em amostras de solo utilizando a espectroscopia na região do visível e no infravermelho próximo. Para tanto, foi utilizada a técnica de Aprendizagem de Máquina SVM incorporada ao software WEKA como auxílio na criação do modelo. A SVM tem representado uma alternativa melhor aos já consagrados métodos de regressão multivariada por apresentar capacidade de generalização. Nos experimentos realizados foram utilizados dois conjuntos de amostras de solo coletadas na região dos Campos Gerais. A avaliação dos resultados teve como base os erros de previsão e os coeficientes de correlação entre os valores dos teores de carbono preditos pelo modelo. Foram encontrados coeficientes de correlação que variaram entre 0,84 a 0,90. Concluiu-se que a espectroscopia no vis-NIRS aliada à técnica SVM é recomendada como uma alternativa aos métodos convencionais de análise de carbono em solos.

Palavras-Chave: Carbono, Agricultura de Precisão, espectroscopia, Aprendizado de Máquina, SVM.

ABSTRACT

Considered a quality indicator, carbon constitutes an important attribute in the productive capacity of the soil. However the traditional methodologies used for determining carbon cause environmental problems due to the use of chemical reagents. The replacement of this procedure by others that generate little or no amount of toxic waste has been considered important. Spectroscopy is one of the promising techniques in Precision Agriculture for soil analysis and can be used to estimate carbon content. Among its benefits, highlights the sample preservation, no consumption of reagents, and their efficiency acquiring data from a large number of samples. The aim of this work was to contribute to determine a regression model able to predict the carbon content in soil samples using spectroscopy in the visible and near infrared region. The Machine Learning SVM technique available in the WEKA software was used to create the model. Because of their generalization ability SVM has been considered a better alternative than the other methods of multivariate regression. Two sets of soil samples collected in the Campos Gerais region were used to the experiments. The results evaluation was based on the forecast errors and the correlation coefficients between the values carbon content predicted by the model. Correlation coefficients ranging from 0.84 to 0.90 were found. It was concluded that the NIRS-vis spectroscopy combined with SVM technique can be recommended as an alternative to conventional methods for carbon analysis in the soil.

Keywords: Carbon, Precision Agriculture, spectroscopy, Machine Learning, Support Vector Machine.

LISTA DE SIGLAS

AP	Agricultura de Precisão
AM	Aprendizado de Máquina
MAPA	Ministério da Agricultura e Pecuária e Abastecimento
CTC	Capacidade de Troca Catiônica
GPL	General Public License
GPS	Global Positioning System
IA	Inteligência Artificial
LS-SVM	Least Square – Support Vector Machine
KDD	Knowledge Discovery in Database
MIR	Infravermelho Médio
MO	Matéria Orgânica
NM	Nanômetros
NIRS	Espectroscopia no Infravermelho Próximo (do inglês, Near Infrared Spectroscopy)
R^2	Coeficiente de Determinação
SMO	Sequential Minimal Optimization
SVM	Suport Vector Machine
TAE	Teoria da Aprendizagem Estatística
VIS	Espectroscopia no Infravermelho Visível (do inglês, Visible Infrared Spectroscopy)
VIS-NIRS	Espectroscopia no Infravermelho Visível e Próximo (do inglês, Visible and Near Infrared Spectroscopy)
WEKA	Waikato Environment for Knowledge Analysis

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Buscando crescimento da produtividade aliado a sustentabilidade, o Brasil avança nas pesquisas relacionadas à agricultura e pecuária para tornar-se ainda mais competitivo no mercado internacional, e o uso de novas tecnologias torna-se um processo vital para esse avanço. Nesse contexto a Agricultura de Precisão (AP) tem um papel fundamental. Hoje, especialmente no Brasil, as soluções existentes estão focadas na aplicação de fertilizantes e corretivos em taxa variável, porém não se deve perder de vista que AP é um sistema de gestão que considera a variabilidade espacial das lavouras em todos seus aspectos: produtividade, solo(características físicas, químicas, compactação, etc), infestação de ervas daninhas, doenças e pragas (MAPA, 2013). O aumento na produtividade das culturas é visível quando os atributos físicos, químicos e biológicos do solo estão equilibrados e suficientemente disponíveis. Para que haja este equilíbrio a utilização dos princípios e tecnologias da AP torna-se inerente.

Considerado um indicador de qualidade, o carbono é um importante atributo na capacidade produtiva do solo. Portanto o levantamento da quantidade de carbono é de grande importância na agricultura. Dentre as diversas metodologias empregadas para determinação de carbono destaca-se a desenvolvida por Walkley & Black em 1934, que utiliza o princípio da combustão úmida (SATO, 2013). Entretanto, problemas ambientais devido o uso de cromo (BRUNETTO *et al*., 2006), tem estimulado a substituição desse procedimento por outros que geram menor ou nenhuma quantidade de resíduos potencialmente tóxicos. Segundo Alkimin (2011), a espectroscopia é considerada uma técnica promissora para análises de solos. Dentre suas vantagens, destaca-se a preservação da amostra, o não consumo de reagentes, além de sua eficiência na aquisição de dados provenientes de um grande número de amostras. Adicionalmente, um único espectro pode ser usado para avaliar diferentes atributos do solo. Para Shepherd e Walsh (2007), é uma das técnicas analíticas mais eficientes e disponíveis do século 21. Canasveras *et al.*(2012) avaliaram o uso da espectroscopia na determinação do teor de Argila, Matéria Orgânica (MO), Capacidade de Troca Catiônica (CTC), Ferro dentre outros atributos do solo no Mediterrâneo, concluíram que a espectroscopia pode ser utilizada como uma alternativa eficaz para substituir os métodos convencionais de análise do solo.

A espectroscopia gera um conjunto de dados após a leitura das amostras. De posse destes dados é possível extrair conhecimento aplicando-se técnicas de aprendizado de

máquina supervisionada. Uma técnica que vem sendo utilizada pelos pesquisadores é a *Support Vector Machine* (SVM) que contempla algoritmos de Aprendizado de Máquina para regressão. Uma função de regressão objetiva determinar uma função linear que expresse exata ou aproximadamente o relacionamento entre duas ou mais variáveis. O caso mais simples é o da regressão linear o qual envolve apenas duas variáveis, X e Y. Já no caso da regressão linear múltipla uma variável dependente é predita por mais de uma variável independente. A SVM consiste na busca de uma função ótima na separação de um hiperplano de decisão na tentativa de maximizar a capacidade de generalização. A popularidade da SVM vem aumentando por apresentar características necessárias para geração de modelos que apresentam boa generalização em conjunto de dados de alta dimensionalidade. Áreas diversas como medicina, química, agricultura, imagens têm utilizado a SVM como ferramenta de aprendizado de máquina como nos trabalhos de Orru *et al*. (2012), Keqiang *et al*. (2014), Fonseca *et al*. (2010), Chen *et al*. (2012).

Neste contexto, o objetivo deste trabalho foi contribuir com um modelo de regressão capaz de predizer a quantidade de carbono em amostras de solo utilizando a espectroscopia na região do visível e no infravermelho próximo. Para tanto foi utilizada a técnica de aprendizado de máquina SVM como auxílio na criação do modelo.

2 REVISÃO BIBLIOGRÁFICA

2.1 AGRICULTURA DE PRECISÃO

A Agricultura de Precisão faz uso de máquinas e equipamentos que distribuem os insumos em taxas variadas baseados em análise de solo e clima. Para se evitar o desperdício de fertilizantes, aparelhos de GPS fazem o mapeamento para ajudar na distribuição dos adubos e na utilização de agroquímicos na quantidade certa para cada espaço. Dessa forma contribui para a redução do impacto ambiental e aumento da produtividade no campo.

A prática da AP tem sua origem nas divisões das lavouras em talhões, os quais eram tratados de forma individualizada pelos antigos agricultores, baseados no conhecimento empírico, tratavam essas partes de forma diferenciada. Acredita-se que o conceito de Agricultura de Precisão tenha surgido juntamente com o advento dos experimentos de uniformidade *(uniformity trials*), instalados em Rothamsted, Grã-Bretanha,em 1925, e com os estudos de acidez do solo na Universidade de Illinois, em 1929 (MACHADO *et al*.,2004).

Segundo Oliver (2010) o termo Agricultura de Precisão foi usado pela primeira vez em 1990, como título de um *workshop* realizado em Great Falls Montana, patrocinado pela Universidade Estadual de Montana. Antes disso, os termos "*site-specific crop management*" ou "*site-spcecific agriculture*" foram usados em edições anteriores do evento.

A Agricultura de Precisão é um conjunto de tecnologias destinadas ao manejo de solos, culturas e insumos, que visa um melhor e mais detalhado gerenciamento do sistema de produção agrícola em todas as etapas, desde a semeadura até a colheita. Por isso, a AP tem como foco a gestão de sistema produtivo agrícola considerando a variabilidade espacial e temporal visando minimizar o efeito negativo ao meio ambiente e maximizar o retorno econômico (INAMAZU *et al.*, 2012).

A AP caracteriza-se pela elevada quantidade de informações disponibilizadas, podendo contribuir para o estabelecimento de relações espaciais de atributos de solo com a produtividade das culturas (AMADO e GIOTTO, 2009). Segundo TING (2008), a Agricultura de Precisão é um sistema inteligente e poderoso de produção que requer capacidade de coleta, processamento de informações e de tomada de decisões assim como dispositivos mecatrônicos de controle e acionamento. A utilização da técnica de

sensoriamento próximo tem grande potencial no complemento das informações necessárias para o melhor desenvolvimento agrícola junto à AP. Associada a esta técnica esta a espectroscopia de refletância no infravermelho a qual fornece sinais marcantes, que podem ser usados para predizer muitos atributos físicos, químicos e biológicos do solo (MADARI, 2006). Técnicas vibracionais, que utilizam a radiação infravermelha para excitar as moléculas, são comuns em muitos laboratórios e têm certas vantagens sobre alguns métodos tradicionais de análise, como, por exemplo: a pequena quantidade de amostra requerida, a rapidez, a facilidade de preparo de amostra, o curto tempo de análise bem como sua ação não poluente por não fazer uso de reagentes. O número de amostras analisadas pode ser triplicado, com redução de custo, pois elimina o uso de reagentes dispersantes (FERRARESI, 2012).

2.2 CARBONO

Por meio da fotossíntese a todo instante ocorre a troca de carbono ente a atmosfera e as plantas. Esse processo é responsável por iniciar a chamada teia alimentar, a qual produz matéria orgânica suprindo grande parte da necessidade energética do planeta. O aumento da matéria orgânica no solo vem em decorrência de uma quantidade maior do carbono resultante da síntese de compostos orgânicos por meio da *fotossíntese.* O carbono é o principal constituinte da MO. A matéria orgânica é um componente fundamental no potencial produtivo dos solos por exercer diversas funções importantes, como a geração de cargas elétricas negativas, disponibilização de nutrientes e a agregação do solo. Além disso, a matéria orgânica é considerada como a principal reserva de carbono do solo, tornando-se um compartimento chave do ciclo global deste elemento (SATO, 2013). A MO afeta diretamente a capacidade produtiva do solo, com os conseqüentes reflexos ambientais (PAVINATO, 2009). A MO que foi sintetizada pelas plantas contém energia, por conseguinte irá servir de alimento para manutenção de processos indispensáveis e de crescimento para os animais. Preliminarmente, essa fonte de energia é transferida aos herbívoros, logo após, é repassada via cadeia alimentar a todos os outros organismos superiores. Caso esse processo seja bloqueado em algum passo, a decomposição da MO por ação de bactérias e fungos, faz com que todos os nutrientes voltem ao solo.

Segundo Bayer e Mielniczuk (2008), a quantidade de carbono residente no solo é calculada relacionando as quantidades adicionadas e perdidas no sistema, pois a proporção desse componente no solo é muito variável, em função do clima, manejo e tipo de solo. A redução do revolvimento do solo e o incremento de resíduos orgânicos pela adubação

podem contribuir para aumentar os estoques de carbono no solo, melhorando as propriedades físicas e reduzindo as perdas por erosão hídrica (LEMOS, 2011). As quantidades de carbono incorporadas ao solo dependem dentre outros fatores do sistema de cultivo em uso. Dessa forma é necessário desenvolver sistemas de rotação de culturas, associadas com a utilização de espécies para cobertura do solo para aumentar a adição de carbono e, ao mesmo tempo, utilizar o mínimo revolvimento do solo para reduzir a taxa de mineralização do carbono adicionado ao sistema. Nesse sentido, o plantio direto apresenta-se como uma boa alternativa para o manejo do solo (PAVINATO, 2009), por consistir em um sistema que enriquece o solo por manter matéria orgânica de restos vegetais de outra culturas.

Segundo Reis (2012), os efeitos dos sistemas de manejo nos estoques de carbono geralmente são estudados principalmente até 20 cm de profundidade, com ênfase na camada de 0 a 5 cm. No entanto, considerando-se que as camadas de solos abaixo de 20 cm podem constituir importantes reservatórios de carbono (BODDEY *et al*., 2010), é necessário investigar como se comportam estas camadas.

Para que a fertilidade do solo de um determinado local seja avaliada, sua análise química é necessária, que consiste em submeter a amostra de solo a uma solução extratora. Para a determinação do teor de carbono do solo o procedimento mais utilizado é o método proposto por Walkley-Black (WALINGA, 1992), o qual utiliza o dicromato ($Cr_2O_7^{2-}$). Segundo Segnini *et al.*(2007) o método de Walkley-Black é ainda o mais utilizado em laboratórios devido a simplicidade e baixo custo, porém, apresenta problemas analíticos e ambientais devido ao uso do cromo.

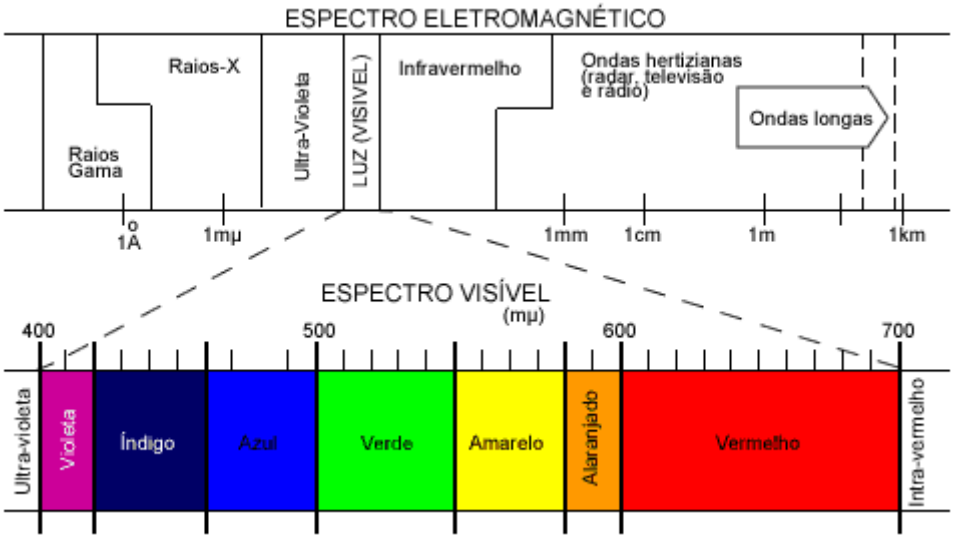
A técnica de espectroscopia de infravermelho próximo (***NIR – Near-infrared***) tem se mostrado uma alternativa mais rápida e limpa para quantificação de carbono demonstrada nos trabalhos de McDowell*et al*.(2012), Fernandes *et al*.(2010) e Filgueiras *et al*.(2012).

2.3 ESPECTROSCOPIA NO INFRAVERMELHO VISÍVEL E PRÓXIMO

A espectroscopia no infravermelho próximo, NIRS (do inglês, Near Infrared Spectroscopy), é uma técnica não-destrutiva utilizada para avaliação rápida de materiais biológicos. Essa técnica, que exige pouca ou nenhuma preparação das amostras, é fundamentada na espectroscopia vibracional, que mede a interação da radiação eletromagnética com a matéria (PASQUINI, 2003).

A radiação eletromagnética pode ser definida como sendo uma propagação de energia, por meio de variação temporal dos campos elétrico e magnético, da onda portadora (NOVO, 1992). O espectro eletromagnético pode ser definido como a distribuição da intensidade da radiação eletromagnética com relação ao seu comprimento de onda ou freqüência. Um espectro eletromagnético pode ser visualizado a partir da Figura 1.

Figura 1 - Faixas do espectro eletromagnético.



Fonte: TENG, 2012

Para a aplicação da espectroscopia são utilizados equipamentos chamados espectrômetros que operam na faixa de 750 a 2.500 nm do espectro eletromagnético. O equipamento NIRS é uma integração da espectroscopia, estatística e computação, tendo o princípio mecânico de iluminar uma amostra com luz de comprimento de onda específico e conhecido da região do espectro eletromagnético. A absorção de luz então é medida por diferenças entre a quantidade de luz emitida pelo NIRS e a quantidade de luz refletida pela amostra, relação pela qual pode-se predizer composição química da mesma, desde que as leituras obtidas possam ser instantâneas, efetivamente comparadas e ajustadas na matriz de um banco de dados armazenado que calibra o software do equipamento (PROENÇA, 2012). Dependendo do modo de funcionamento do equipamento, o espectro pode ser de absorbância, transmitância ou refletância difusa (PASQUINI, 2003).

2.4 ESPECTROSCOPIA E SOLOS

O tradicional levantamento de solos é uma técnica que apresenta custo elevado para aquisição de dados, principalmente, quando existe a demanda por um grande

número de amostras (ODEH *et al.*, 2013). Outra desvantagem é o tempo considerável despendido em laboratórios para alcançar o maior nível de credibilidade dos resultados (FIORO e DEMATTÊ, 2009).

Buscando-se alternativas viáveis para obtenção de resultados mais rápidos e menos poluentes, pesquisadores procuram por meio de análises espectrais uma acurácia mais significativa no levantamento de solos em relação aos métodos tradicionais. No Brasil, o procedimento analítico de determinação de carbono do solo mais tradicional é baseado na oxidação da matéria orgânica a CO2 por íons dicromato, em meio fortemente ácido. É também denominado de determinação por via úmida ou determinação por dicromato (MACHADO *et al.*, 2003). Os avanços no conhecimento da relação entre a refletância espectral e características do solo fornecem uma ferramenta para prever vários atributos físicos e químicos do solo de maneira rápida, segura, não invasiva e quando integrada com o conhecimento agronômico, o valor desta tecnologia pode ser melhor percebido, resultando em um processo contínuo de avaliação, interpretação e operações precisas (KITCHEN, 2008).

Segundo Balena (2011) a demora na obtenção dos resultados das análises e a busca pela maior produtividade geralmente induzem os produtores a aplicar quantidades excessivas de insumos agrícolas. Tal atitude além de aumentar o custo de produção contribui para a contaminação do solo, das águas superficiais e subterrâneas. Por outro lado, analises de solo por técnicas espectroscópicas, em especial, com espectrorradiômetro portátil, são rápidas, não exigem preparo das amostras e podem ser realizadas tanto em laboratório quando no campo. A rapidez na obtenção dos resultados permite detectar a ausência de elementos químicos importantes para as culturas e avaliar as condições físicas do local a tempo de corrigir o solo para o plantio.

Pesquisas para melhor entender as relações existentes entre os diversos componentes do solo e a reflectância são necessárias, mas pode-se afirmar que esta ferramenta tem tido importância crescente na pedologia, tanto que, Demattê *et al.*(2004) sugerem a inclusão de um novo termo, denominado pedologia espectral, fazendo alusão à utilização das propriedades espectrais no auxílio à identificação de classes de solos e seu mapeamento.

Santos *et al.*(2010) não encontrou diferença entre os resultados das análises de MO, argila, teores foliares de silício e nitrogênio realizadas pelo método NIR daqueles realizados pelos métodos convencionais. Em um trabalho desenvolvido por Nanni *et al.*(2013) a espectrorradiometria difusa evidenciou ser capaz de estimar os atributos areia, argila, soma de bases e, principalmente, Fe_2O_3, da área de transição arenito/basalto do noroeste do Estado

do Paraná sendo alcançados melhores resultados para os horizontes superficiais. Souza ***et al.*** (2012) realizou experimento fazendo uma varredura espectral na região do Infravermelho Médio (MIR) em amostras de solo para determinação de MO, com o auxilio da ferramenta LS-SVM (*Least Square - Support Vector Machine*) para calibração multivariada. Obteve erros menores que 1% na classificação dos solos. Concluiu que a associação da espectroscopia MIR com LS-SVM mostra ser uma alternativa limpa, operacional e de baixo custo para predição do teor de MO.

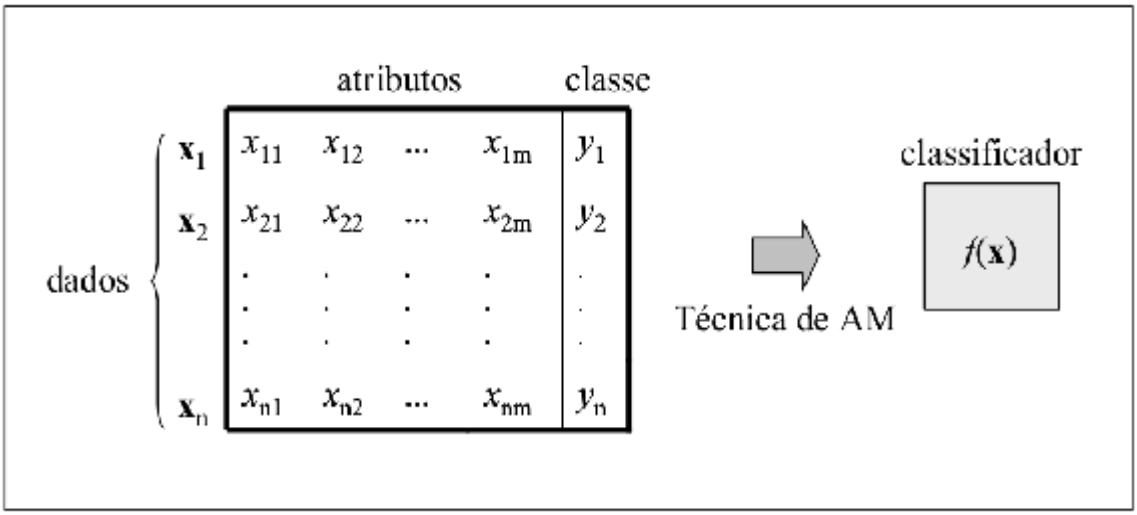
2.5 APRENDIZADO DE MÁQUINA

Aprendizado de Máquina (do inglês, *Machine Learning*) é a área de Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre processo de aprendizado (BISHOP, 2007). Dirigida ao desenvolvimento de programas que permitam ao computador aprender, utilizando o raciocínio indutivo por meio da extração de regras e padrões de um conjunto de dados. É qualquer programa que apresente condição de melhorar sua performance na execução de uma tarefa, a partir da experiência. Em geral, técnicas de Aprendizado de Máquina são divididas em aprendizado supervisionado e não supervisionado. A diferença entre as técnicas reside na existência ou não de conhecimento, ou seja, exemplos anteriormente disponíveis. No caso de aprendizado não supervisionado estes rótulos não existem. Na aprendizagem supervisionada, o objetivo é prever o valor de uma função para qualquer entrada válida utilizando um certo número de exemplos, isto é, dados de treinamento. Tais dados exemplificam as relações entre a entrada e saída. A função de saída pode ser um valor em um espaço discreto, utilizado para classificação ou pode ser um valor em um espaço contínuo, chamado regressão. A análise de regressão estuda a relação que existe entre duas ou mais variáveis. 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 (OLIVEIRA Jr., 2012).

A partir de um conjunto de exemplos rotulados (x_i, y_i), onde x_i representa a entrada e y_i a saída esperada ou também chamada classe meta, será gerado um classificador baseado no conjunto de entrada que deverá estar preparado para predizer precisamente a classe de futuras entradas. Este classificador nada mais é do que uma função. A Figura 2 demonstra esse conceito utilizando o paradigma do aprendizado supervisionado, onde o

chamado conjunto de treinamento é representado pelas instâncias (x_i,... x_{im}), onde x_i possui m atributos e as variáveis y_i representam os atributos meta.

Figura 2 - Indução de classificador em aprendizado supervisionado.



Fonte: Lorena e Carvalho (2007).

O conjunto de dados de treinamento é disponibilizado e analisado, e um modelo de classificação é construído, baseado nesses dados. Então, o modelo é utilizado para classificar outros dados, chamados dados de teste, os quais são contemplados pelo algoritmo durante a fase de treinamento. Cabe ressaltar que o modelo construído a partir dos dados de treinamento só será considerado um bom modelo do ponto de vista de precisão preditiva, se ele classificar corretamente uma significativa porcentagem dos exemplos dos dados de teste. Em outras palavras, o modelo deve representar conhecimento que possa ser generalizado para dados de teste, não utilizados durante o treinamento (CARVALHO, 2005). A otimização de um classificador para maximizar seu desempenho junto a um conjunto de treinamento nem sempre produz um bom resultado para o conjunto de testes.

Um conceito comumente empregado em AM é o de generalização de um classificador, definida como a sua capacidade de prever corretamente a classe de novos dados. No caso em que o modelo se especializa nos dados utilizados em seu treinamento, apresentando uma baixa taxa de acerto quando confrontado com novos dados, tem-se a ocorrência de um superajustamento (*overfitting*). É também possível induzir hipóteses que apresentem uma baixa taxa de acerto mesmo no subconjunto de treinamento, configurando uma condição de subajustamento (*underfitting*). Essa situação pode ocorrer, por exemplo, quando os exemplos de treinamento disponíveis são pouco representativos ou quando o modelo obtido é muito simples (MONARD e BARANAUSKAS, 2003).

Portanto, o objetivo de um algoritmo de aprendizagem é a construção de modelos com considerável capacidade de generalização, ou seja, modelos que permitam prever com precisão os rótulos de classe de registros desconhecidos (TAN *et al*., 2009). Árvores de Decisão, Máquina de Vetores de Suporte, Redes Neurais Artificiais são exemplos de métodos empregados para aplicação da técnica de Aprendizado de Máquina.

2.6 MÁQUINA DE VETORES DE SUPORTE

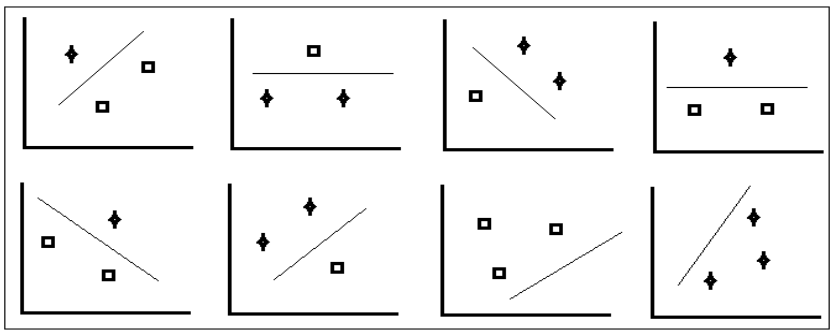
Um classificador baseado em máquina de vetores de suporte consiste em uma técnica fundamentada na teoria de aprendizagem estatística. A Máquina de Vetores de Suporte, do inglês *Support Vector Machine* - SVM foi desenvolvida por Vapnik (1995), com o intuito de resolver problemas de classificação de padrões. Esta técnica originalmente desenvolvida para classificação binária, busca a construção de um hiperplano como superfície de decisão, de tal forma que a separação entre exemplos seja máxima. Isso considerando padrões linearmente separáveis (SOUZA, 2011).

A SVM é também indicada freqüentemente para procedimentos de regressão tradicionais/estatística, pois os processos de derivação de uma função apresenta o menor desvio entre as respostas estimadas e observadas para todos os exemplos de treinamento (BASAK *et al*.,2007).

O termo Máquina de Vetor de Suporte surgiu porque os pontos do conjunto de treinamento que estão mais próximos da superfície de decisão são chamados de vetores de suporte. SVM realiza a maximização da margem de decisão entre as duas classes baseada no princípio de Minimização do Risco Estrutural que é fundamentado no fato de a taxa de erro da máquina de aprendizado no conjunto de teste ser limitada pelo somatório da taxa de erro de treinamento e por um termo que depende da dimensão de Vapnik-Chervonenkis(VC) (ZHUANG, 2006).

Uma função eficaz consegue, potencialmente, combinar todos os arranjos possíveis de classificação num número de 2^m. A dimensão de VC é definida como o maior valor que m pode adquirir, para um conjunto de m combinações nas quais a função classificadora pode se moldar. Se esse limite m não existir, então a dimensão é infinita. O valor de m pode ser traduzido como o valor de capacidade de aprendizagem de um sistema classificador (RODRIGUES, 2007). A Figura 3 mostra um exemplo de dimensão VC onde o valor de m é 3 em R^2.

Figura 3 -Separação em um espaço bidimensional de 3 pontos por meio de retas.



Fonte: Autor

A dimensão VC do conjunto de funções lineares no espaço bidimensional é 3, uma vez que existe (pelo menos) uma configuração de três pontos nesse espaço que pode ser particionada por retas em todas as $2^3 = 8$ combinações binárias de rótulos (LORENA e CARVALHO, 2007).

2.6.1 Máquina de Vetores de Suporte Linearmente Separável com margens rígidas.

Uma Máquina de Vetor de Suporte constrói um classificador binário a partir de um conjunto de padrões, chamados de exemplos de treinamento, em que a classificação é conhecida. Um conjunto de treinamento é dito linearmente separável se for possível separar os padrões de classes diferentes contidos no mesmo por pelo menos um hiperplano (HAYKIN, 2001).

Considere o conjunto de treinamento $\{(\mathbf{x}_i,d_i)\}N_i$=1, onde $\mathbf{x}_i$ é o padrão de entrada para o i-ésimo exemplo e d_i é a resposta desejada d_i= {+1,−1} que representa as classes linearmente separáveis. A equação que separa os padrões através de hiperplanos pode ser definida por:

$$w^t x + b = 0 \quad (1)$$

Onde, $w^t x$ é o produto escalar entre os vetores w e x, em que x é um vetor de entrada que representa os padrões de entrada do conjunto de treinamento, w é o vetor de pesos ajustáveis e b é um limiar também conhecido como bias (TAKAHASHI, 2012).

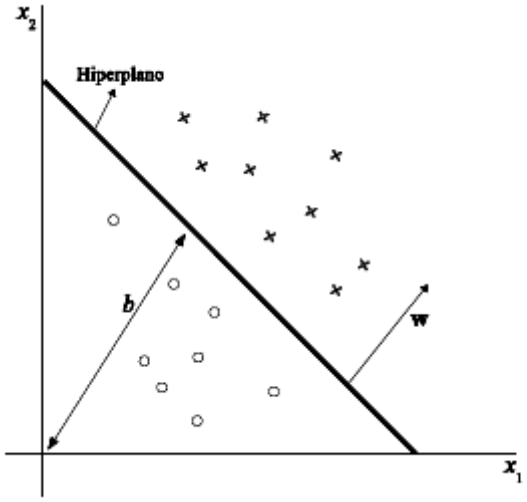
Reescrevendo a equação ela ficará da seguinte forma:

$$w^t.x + b \geq 0 \; para Y = +1 \quad (2)$$

$$w^t.x + b < 0 \; para Y = -1 \quad (3)$$

A Figura 4 mostra o hiperplano de separação (w,b) em um espaço bidimensional para um conjunto de treinamento linearmente separável.

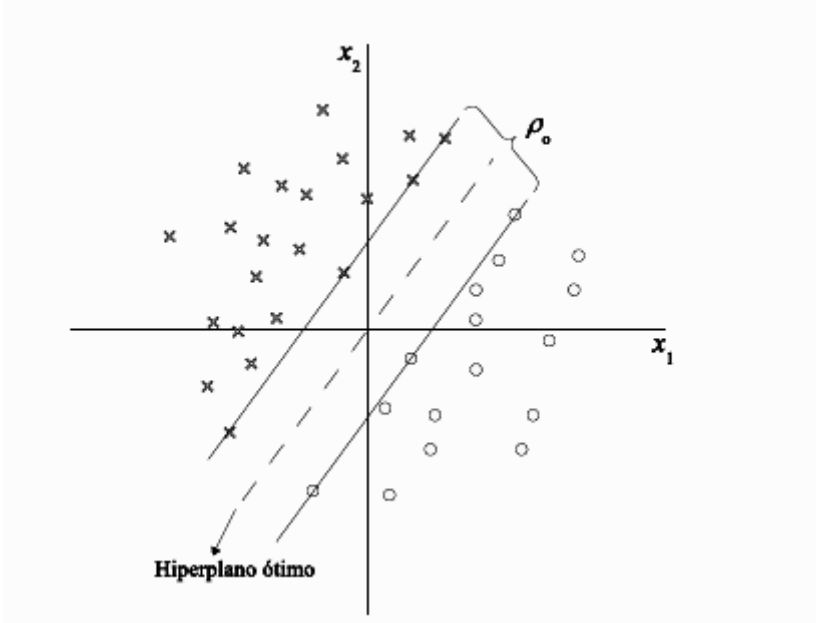
Figura 4 -Hiperplano de separação (w,*b*) para um conjunto de treinamento bidimensional.



Fonte: Takahashi (2012).

Para um dado vetor de peso w e bias *b*, a separação entre o hiperplano definido na Equação1 e o ponto de dado mais próximo é denominada a margem de separação, representada por ρ. O objetivo de uma SVM é encontrar o hiperplano particular para qual a margem de separação é máxima. Sob esta condição, a superfície de decisão é referida como o hiperplano ótimo(MARETTO, 2011).

Figura 5 - Hiperplano ótimo com máxima margem ρ_o de separação dos padrões linearmente separáveis



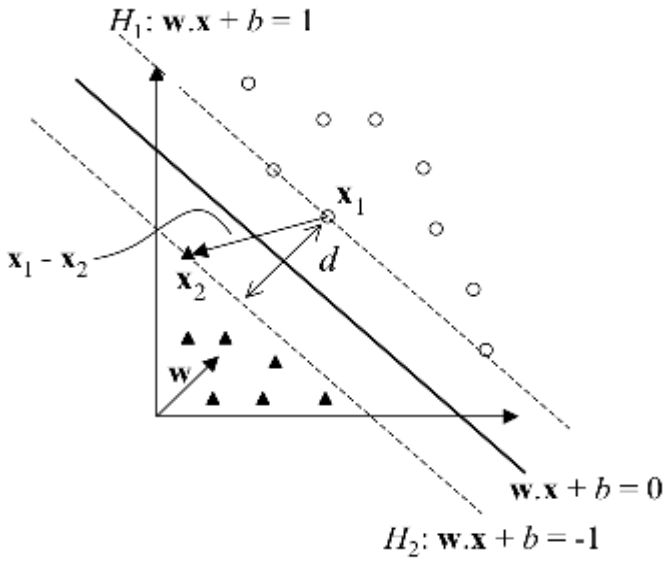
Fonte: Takahashi (2012).

Seja x_1 um ponto no hiperplano H_1:w.x + b = +1 e x_2 um ponto no hiperplanoH_2:w.x + b = -1. Projetando $x_1 – x_2$ na direção de w, perpendicular ao hiperplano

separador w.x + b = 0, é possível obter a distância entre os hiperplanos H_1 e H_2(CAMPBELL, 2000). Essa projeção é apresentada na Equação (4).

$$(x_1 - x_2)\left(\frac{w}{||w||} \cdot \frac{x_1 - x_2}{||x_1 - x_{2||}}\right) \quad (4)$$

Figura 6 - Cálculo da distância *d* entre os hiperplanos H_1 e H_2.



Fonte: Lorena e Carvalho (2007).

A maximização da margem de separação dos dados será obtida com a minimização da chamada função custo (Equação 5).

$$\underset{w,b}{Minimizar}\ \frac{1}{2} \parallel w \parallel^2 \quad (5)$$

Com as restrições: $y_i(w.x_i + b - 1) \geq 0, \ \forall i = 1, \ldots, n$ (6)

O problema acima é chamado de *formulação primal* (HAYKIN, 2001). Devido à natureza das restrições do problema primal, pode-se ter dificuldades na obtenção da solução do problema. Segundo Ales *et al.*,(2009), problemas desse tipo podem ser solucionados com a introdução de uma função Lagrangiana que engloba as restrições a função objetivo associadas a parâmetros denominados Multiplicadores de Lagrange.

$$L(w, b, \alpha) = \frac{1}{2} \parallel w \parallel^2 - \sum_{i=1}^{n} \alpha_i (y_i(w \,.x_1 + b) - 1) \quad (7)$$

Onde α_i (i = 1,...,N) são os multiplicadores de Lagrange. O lagrangeano da equação 7 tem que ser minimizado com respeito a we b e maximizado com respeito a $\alpha_i >= 0$. Tem-se então um ponto de sela[1], no qual:

$$\frac{\partial L}{\partial b} = 0 \;\; e \frac{\partial L}{\partial w} = 0 \tag{8}$$

Resolvendo as equações acima será obtido o seguinte resultado:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{9}$$

$$w = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{10}$$

Substituindo as equações 8 e 9 na equação6, obtém-se o seguinte problema de otimização:

$$\underset{\alpha}{Maximizar} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \,.\, x_j) \tag{11}$$

$$\text{Com as Restrições: } \begin{cases} \alpha_{i>} \geq 0, \quad \forall_i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_I y_i = 0 \end{cases} \tag{12}$$

Esta formulação é chamada de dual enquanto a formulação da equação 5 é chamada de primal. Na solução, os pontos para os quais $\alpha_i > 0$ são chamados vetores de suporte e estão em um dos hiperplanosH_1 ou H_2. Para SVM os vetores de suporte são os elementos críticos do conjunto de treinamento. Eles estão na fronteira, ou seja, mais próximos do hiperplano de decisão. Todos os outros pontos tem α_i i =0. Se todos esses outros pontos forem removidos e o treinamento for repetido, o mesmo hiperplano deve ser encontrado (BURGES, 1998).

Restrições são sempre comuns para busca de soluções. A partir dessa premissa Karush-Kuhn-Tucker (KKT) descreveu a Teoria de Otimização com Restrições (CRISTIANINI e TAYLOR, 2000). As condições KKT dizem que, os multiplicadores de lagrange satisfazem $\alpha^* = 0$.

[1] Um ponto em uma superfície que é um maximum em uma secção transversal planar e um minimum em outra.

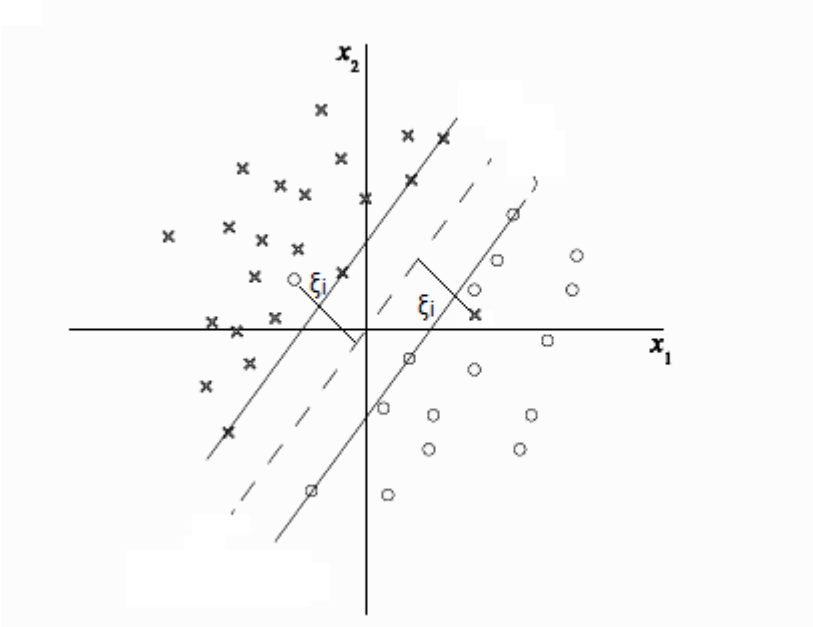
$$g(x) = sgn\big(f(x)\big) = sgn\left(\sum_{x_i \in \mathrm{SV}}^{n} y_i \alpha_i^* \, \mathrm{x}_i \, . \mathrm{x} + \; b^*\right) \qquad (13)$$

Segundo Lorena e Carvalho (2007) esta função linear representa o hiperplano que separa os dados com maior margem, considerado aquele com melhor capacidade de generalização de acordo com a Teoria da Aprendizagem Estatística(TAE). Essa característica difere as SVMs lineares de margens rígidas das Redes Neurais Perceptron, em que o hiperplano obtido na separação dos dados pode não corresponder ao de maior margem de separação.

2.6.2 Máquina de Vetor de Suporte Linearmente Separável com margens suaves.

No caso das chamadas margens rígidas, não se permite erros de classificação. Porém os dados não estão livres de ruídos. Esse caso nomeado SVM linear não separável é definido pela situação em que é tolerável a ocorrência de erros nos dados de treinamento (TAN, 2009). Para que isto seja possível foram introduzidas as chamadas variáveis de folga ou também chamadas variáveis soltas ($\xi_i >= 0$, $i = 1,...,N$) apresentadas na Figura 7. Devido a sua eficiência em trabalhar com dados de alta dimensionalidade, tem sido reportada na literatura como uma técnica altamente robusta, muitas vezes comparada às redes neurais (SUNG e MUKKAMALA, 2003).

Figura 7 – Variáveis Soltas.



Fonte: Adaptado de Takahashi (2012).

Para que alguns erros de classificação possam ser aceitos, ou seja, alguns dados permaneçam entre os hiperplanos H_1 e H_2 é preciso violar algumas restrições da equação 6. Essas variáveis relaxam as restrições impostas ao problema de otimização primal, que se tornam (SMOLA e SCHOLKOPF, 2002)

$$y_i(w \,.\, x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall i = 1, \ldots, n \tag{14}$$

Um erro no conjunto de treinamento é indicado por um valor de ξ_i maior que 1. Logo, a soma dos ξ_i representa um limite no número de erros de treinamento. Para levar em consideração esse termo, minimizando assim o erro sobre os dados de treinamento, a função objetivo da equação 5 é reformulada como(BURGES, 1998):

$$\underset{w,b,\xi}{Minimizar} \frac{1}{2} \| w \|^2 + C\left(\sum_{i=1}^{n} \xi_i\right) \tag{15}$$

O primeiro termo da função tem o objetivo de maximizar a margem, enquanto que o segundo termo C($\sum \xi i$) minimiza o valor das variáveis de folga ξi , reduzindo o número de pontos que ficam do lado incorreto. Isto quer dizer que o parâmetro C destaca maior ou menor importância das variáveis de folga, propiciando que o modelo do SVM não perca a sua capacidade de generalização. A constante C é um termo de regularização que impõe um peso à minimização dos erros no conjunto de treinamento em relação à minimização da complexidade do modelo(PASSERINI, 2004).

Segundo Lorena e Carvalho (2007), a solução da equação envolve passos matemáticos semelhantes aos apresentados anteriormente, com a introdução de uma função Lagrangiana e tornando suas derivadas parciais nulas. Tem-se como resultado o seguinte problema dual (Equações 16 e 17):

$$\underset{\alpha}{Maximizar} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \,.\, x_j) \tag{16}$$

$$\text{Com as Restrições: } \begin{cases} 0 \leq \alpha_i \leq C, \quad \forall_i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_I y_i = 0 \end{cases} \tag{17}$$

Pode-se observar que essa formulação é igual à apresentada para as SVMs de margens rígidas, a não ser pela restrição nos α_i, que agora são limitados pelo valor de C.Seja α* solução do problema dual, enquanto w*, b*e ξ* denotam as soluções da forma

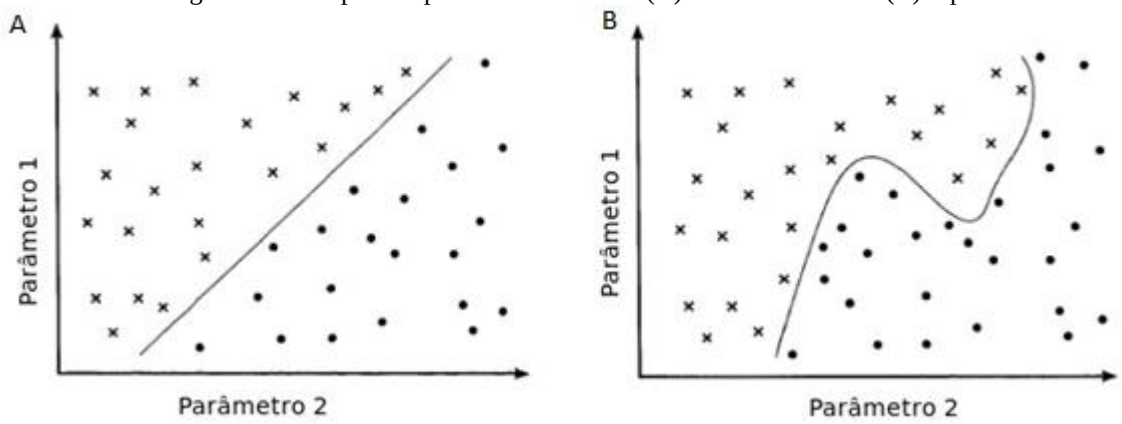
primal. O vetor w^* continua sendo determinado pela equação 10. As variáveis ξ^* podem ser calculadas pela Equação 18(CRISTIANINI e TAYLOR, 2000).

$$\xi_i^* = \max\left\{0{,}1 - y_i \sum_{j=1}^{n} y_j \alpha_j^* x_j . x_i + b^*\right\} \tag{18}$$

2.6.3 Máquina de Vetores de Suporte não lineares

Em problemas reais é difícil encontrar padrões que sejam linearmente separáveis, ou seja, existem situações em que não é possível separar os dados de treinamento por um hiperplano. A Figura 8 apresenta um conjunto de treinamento não separável e outro separável.

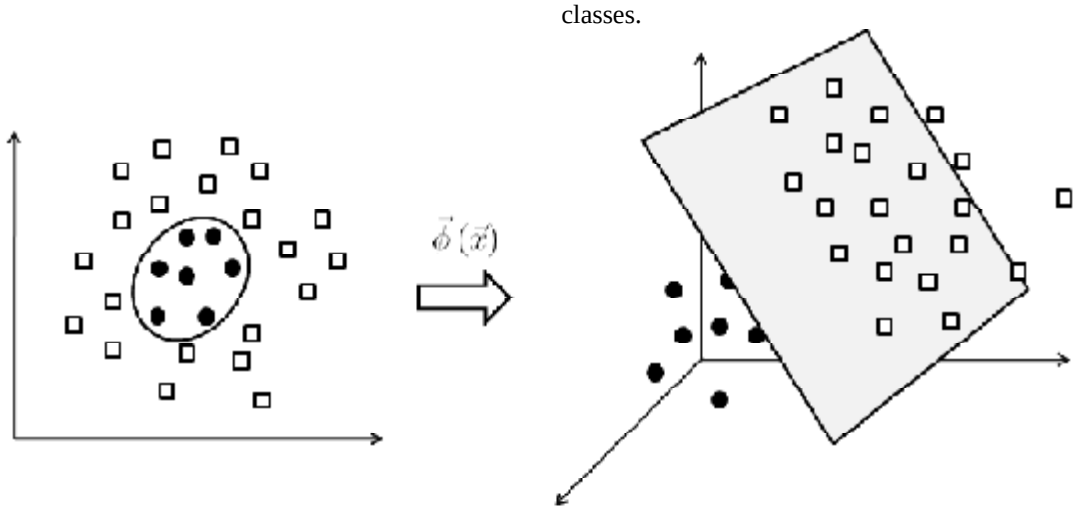
Figura 8 - Exemplos de padrões linearmente (A) e não linearmente (B) separável.



Fonte: Gonçalves (2004).

Para que estes problemas fossem resolvidos criou-se uma extensão da SVM na qual foram criadas funções que mapeiam o conjunto de treinamento em um espaço linearmente separável, o chamado espaço de características. Em muitos problemas, é necessário um mapeamento do espaço de entrada em um outro, geralmente (mas não necessariamente) de dimensão mais elevada, onde pode-se, aí sim, efetuar a separação das classes de maneira linear. Esse novo espaço é denominado espaço de características (SILVA, 2008). Um problema desse tipo é ilustrado na Figura 9.

Figura 9 - Mapeamento do espaço de entrada num espaço de características com separação linear entre as classes.



Fonte: Silva (2008).

A técnica do aumento da dimensionalidade para separação de padrões tem como base o Teorema de Cover, o qual define que um problema complexo de classificação de padrões tem mais probabilidade de ser separável linearmente em um espaço de alta dimensão do que em um espaço de baixa dimensão (HAYKIN, 2001). Considerando-se um espaço de entrada em que os padrões não são linearmente separáveis, o teorema de Cover diz que esse espaço pode ser transformado em um novo espaço de características, onde os padrões têm alta probabilidade de tornarem-se linearmente separáveis, sob duas condições: a transformação deve ser não-linear e a dimensão do espaço de características deve ser muito alta com relação à dimensão do espaço de entrada.

Pode-se então aplicar as SVMs lineares sobre os dados mapeados no espaço de características, determinando um hiperplano de margem maximizada. A SVM não linear incorpora este conceito por meio da utilização de funções *Kernel*, que permitem o acesso a espaços complexos (em alguns casos infinitos) de forma simplificada(LORENA e CARVALHO, 2003).

Um Kernel k é uma função que recebe dois pontos x_i e x_j do espaço de entrada e computa o produto escalar $\phi(x_i).\phi(x_j)$ no espaço de características. O termo $\phi(x_i).\phi(x_j)$representa o produto interno dos vetores x_i e x_j, sendo o kernel representado por:

$$k(x_i x_j) = \Phi(x_i).\Phi(x_j) \quad (19)$$

É comum empregar a função Kernel sem conhecer o mapeamento ϕ, que é gerado implicitamente. A utilidade dos Kernels está, portanto, na simplicidade de seu cálculo e em sua capacidade de representar espaços abstratos (LORENA e CARVALHO, 2007). As funções Kernels empregadas pela SVM são:

- Linear

$$k(x,y) = x.y \quad (20)$$

- Polinomial

$$k(x,y) = ((1+(x_i.y_i))^p \quad (21)$$

- Gaussiano ou RBF (radial basis function)

$$k = exp(-\sigma ||x_i - x_j||^2) \quad (22)$$

- Sigmoidal ou perceptron de duas camadas

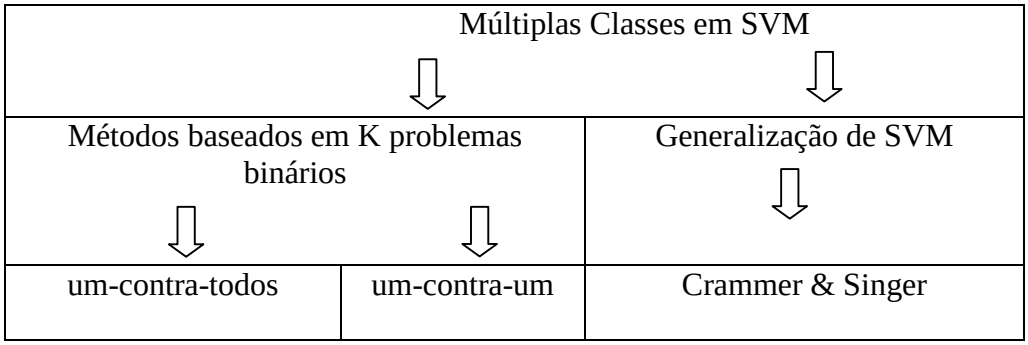
$$k = tanh(\delta(xi.xj) + k) \quad (23)$$

2.6.4 Classificação em múltiplas classes

As SVM foram propostas inicialmente como ferramenta de classificação binária. Porem muitos dos problemas reais possuem características multiclasses. Para que fosse possível a utilização da SVM neste tipo de aplicação, foram propostos alguns procedimentos para estender a SVM binária. Para problemas que possuem mais de duas classes, o conjunto de dados de treinamento deve ser combinado para formar problemas de duas classes (BISOGNIN, 2007).

Para que o problema de classificação em múltiplas classes pudesse ser resolvido criou-se uma extensão da SVM que utiliza duas abordagens (Figura10). Uma é a redução do problema em múltiplas classes a um conjunto de problemas binários, dois métodos utilizam esta técnica de decomposição (um contra um) e (um contra todos). E a outra é uma generalização do algoritmo de classificação binária usando máquinas de vetor suporte para que funcione para mais de duas classes. O método que utiliza essa segunda abordagem é chamado de *Crammer and Singer* (CRAMMER e SINGER, 2000).

Figura 10 - Abordagens para classificação em múltiplas classes.

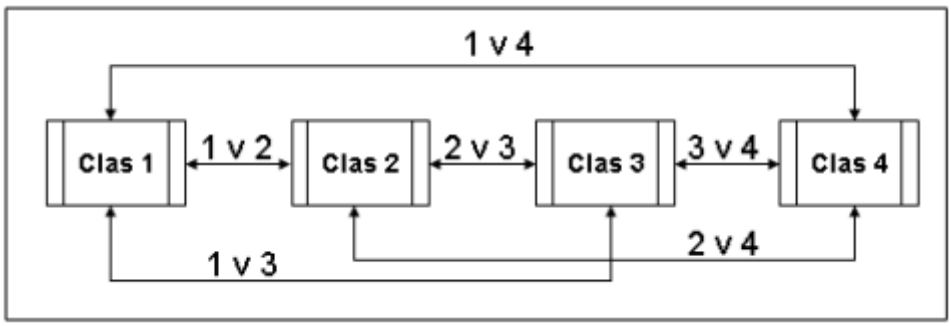


Fonte: Autor

Na primeira abordagem, o método de decomposição por classe um-contra-todos é o mais antigo e o mais empregado. Para uma classificação em k classes, são resolvidos k problemas binários, isto é, é construída uma SVM para cada uma das k classes. A construção da *i*-ésima SVM, $i \in \{1,...,k\}$, é feita usando todos os padrões de treinamento, os exemplos da classe i com saída $y = 1$ e os outros exemplos com saída $y = -1$(CHAVES, 2006).

Já o método um-contra-um constrói k(k-1)/2 classificadores binários, em que cada um é treinado com dados de duas classes. Como para resolução de um problema com k classes são construídas k(k-1)/2 SVMs, esse número, para k > 2, é maior do que o número de SVMs construídas no método de decomposição um por classe. No entanto, os problemas resolvidos são menores. O método de decomposição um por classe usa todos os pontos de treinamento na construção das SVMs. Já o método de separação das classes duas a duas usa, em média, 2N/k pontos de treinamento para cada SVM (CHAVES, 2006). A Figura 11 ilustra o método um-contra-um onde cada ligação entre duas classes representa um classificador binário.

Figura 11- Representação de um método um-contra-um.



Fonte: Bisognin (2007)

2.6.5 Algoritmo de Otimização Sequencial Mínima

O algoritmo *Sequential Minimal Optimization* (SMO) proposto por Platt (1999) busca diminuir a quantidade de operações aritméticas, dessa maneira diminuindo o tempo de processamento das SVM. Além disso, tem capacidade de tratar conjuntos de dados esparsos, que possuem um número substancial de elementos com valor zero, pois empregam maior ênfase na etapa de avaliação da função de decisão. A otimização realizada no algoritmo SMO está na programação quadrática analítica ao invés da abordagem numérica utilizada tradicionalmente nas implementações SVM (SOUZA, 2011). Dos algoritmos que codificam o SVM, o SMO é um dos mais rápidos e um dos que menos consome memória. A complexidade computacional do SMO é menor em relação aos demais algoritmos porque ele não trabalha com inversão de uma matriz de kernel. Esta matriz, que é quadrada, costuma ter um tamanho proporcional ao número de amostras que compõem os chamados vetores de suporte (KINTO, 2011).

O algoritmo SMO escolhe a resolução dos problemas de otimização, optando pelas menores otimizações possíveis em cada passo. Nos problemas de programação quadrática em SVM, a menor otimização possível envolve os multiplicadores de Lagrange, porque eles devem obedecer à restrição de igualdade linear. Em cada passo, o método SMO escolhe a otimização dos dois multiplicadores de Lagrange, buscando valores ótimos para eles e atualizando-os para refletir os novos valores ótimos. A vantagem está em utilizar um otimizador analítico, ao invés de ser chamada toda uma biblioteca de rotinas de programação quadrática. Além disso, não é necessário armazenamento de matrizes extras, o que permite manipular problemas com conjunto de treinamento volumoso (PLATT, 1999).

Inseridos no contexto agronômico, existem trabalhos, como de Martinez *et al.* (2012) o qual utilizou SMO em comparação com demais algoritmos de aprendizagem para estimar o nível de álcool de uma vinha, durante o processo de maturação de uvas. Obtendo resultado satisfatório cujo coeficiente de correlação foi de 0,637. Utilizando dados de sensoriamento remoto, Jachowski*et al.*(2013) avaliaram a biomassa e diversidade de árvores dentro de um ecossistema de mangue na costa da Tailândia. Dentre 18 algoritmos de aprendizado de máquina que foram testados o que apresentou o melhor índice de correlação foi o SMO com 0,81. É importante salientar que nesse trabalho foi extraído um modelo de regressão por meio dos resultados das leituras das bandas espectrais obtidas do satélite. Outro trabalho foi o relatado por Udomsin *et al.* (2014) na previsão do Produto Agrícola Bruto da Tailândia, atingindo um R^2 de 0,968 utilizando SMO como técnica de regressão.

2.7 O SOFTWARE WEKA

O WEKA (*Waikato Environment for Knowledge Analysis*) é uma ferramenta de KDD[2] (*Knowledge Discovery in Database*) que contempla uma série de algoritmos de preparação de dados, de aprendizado de máquina e de validação de resultados (HALL *et al*, 2009). WEKA foi desenvolvido na Universidade de Waikato na Nova Zelândia, sendo escrito em Java e possuindo código aberto com licença de software livre GPL (*General Public License*). Grande parte de seus componentes de software são resultantes de teses e dissertações de grupos de pesquisa daquela universidade. O propósito inicial do desenvolvimento do software visava a investigação de técnicas de aprendizado de máquina, enquanto sua primeira aplicação foi direcionada para a agricultura, uma área chave na economia da Nova Zelândia.

O WEKA suporta várias etapas típicas da mineração de dados, tarefas como:pré-processamento, agrupamento, classificação, regressão e visualização de dados. A ferramenta oferece quatro métodos para avaliar o desempenho de um algoritmo de aprendizagem quanto ao erro de generalização. O erro auxilia o algoritmo de aprendizagem a encontrar um modelo com a complexidade certa que não seja suscetível a *overfitting*. Assim que o modelo tenha sido construído, pode ser aplicado ao conjunto de teste para prever os rótulos de classe de registros não vistos anteriormente (TAN *et al.*, 2009). O WEKA dispõe as seguintes opções para avaliação(PIMENTA e VALENTIM, 2009):

- *Use training set–* Utiliza o próprio conjunto de treinamento como conjunto de teste.
- *Suplied test set*- Essa opção permite ao usuário entrar com um conjunto de teste.
- *Cross Validation*- Esta técnica divide a base de dados em *x*partes iguais. Destas, *x-1* partes são utilizadas para o treinamento e uma parte servirá como base de teste ou validação. O processo é repetido *x* vezes, dessa forma cada parte será usada uma vez como conjunto de validação. Ao término, a correção total é calculada pela média dos resultados obtidos em cada uma das etapas, obtendo-se assim uma estimativa da acurácia do modelo de aprendizagem gerado e permitindo análises estatísticas.

[2]Processo de descoberta de conhecimento em Bases de Dados

- *Percentage Split* - Recebe como entrada, a informação referente à percentagem na qual os dados serão divididos em subconjuntos de treinamento e teste. Cria-se um subconjunto de treinamento com *x*% do tamanho da base de dados fornecida como entrada, sendo *x* a percentagem dada. Com o restante dos dados é criado um subconjunto de teste. O algoritmo de mineração é então aplicado sobre o subconjunto de treinamento. A partir daí, acontece um laço de repetição que itera *j* vezes, sendo *j* o número de instâncias do subconjunto de teste. A cada iteração deste laço, é processada uma predição do classificador sobre a instância de teste corrente.

Uma das medidas de performance disponibilizadas pelo WEKA é o coeficiente de correlação (*correlation coeficient*) cujo valor mede o grau de relacionamentos entre duas variáveis. Seu valor está no intervalo [0,1]. Quanto maior, mais explicativo é o modelo. Também como parâmetro de performance, o WEKA provê coeficientes de medida a partir dos erros de regressão.

- *Mean absolute error*:erro absoluto médio (é a média do erro de predição, é calculado por meio da diferença entre os valores atuais e os preditos);
- *Root mean squared error*:Raiz quadrada do erro médio (é a raiz quadrada do *Mean absolute error*, calculado pela média da raiz quadrada da diferença entre o valor calculado e o valor correto.);
- *Relative absolute error*:erro absoluto relativo (é o erro total absoluto relativo, quanto menor o valor, significa maior precisão do modelo);
- *Raiz quadrada do erro relativo*:raiz quadrada do erro relativo (reduz o quadrado do erro relativo a mesma dimensão da quantidade sendo predita incluindo raiz quadrada;

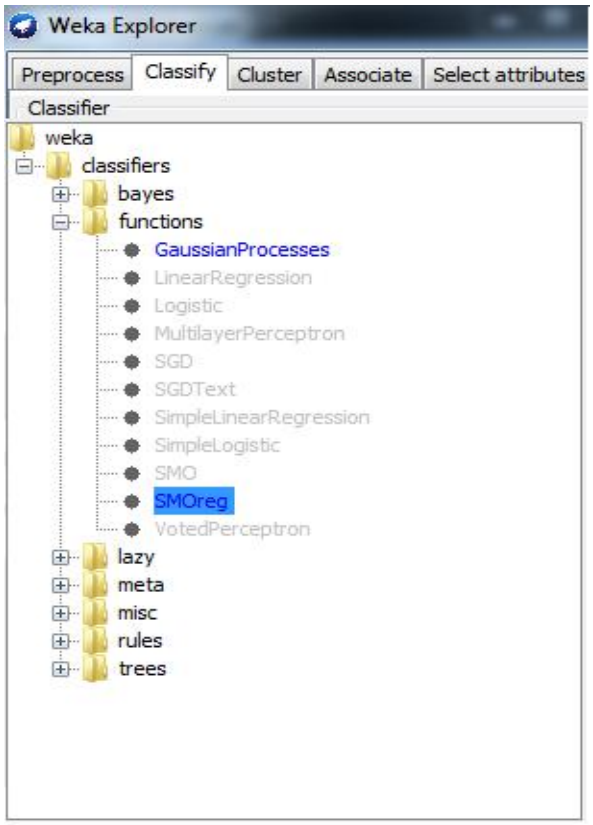
2.7.1 SMOReg no WEKA

De acordo com Cornélio (2012), SMOreg é uma implementação do algoritmo de Otimização Seqüencial Mínima (*Sequential Minimal Optimization* - SMO) para treinamento de um modelo de regressão de vetor de suporte, a idéia é dividir uma programação quadrática complexa em uma série de pequenos programas quadráticos.

O algoritmo SMOreg consiste em uma implementação Java e está disponível no pacote de software WEKA conforme Figura 12. Considerado um método de regressão SMOreg possibilita a opção por alguns parâmetros como por exemplo o parâmetro

de ajuste C. Também é possível indicar ao algoritmo a necessidade ou não da normalização dos dados para o processo de regressão, e ainda optar por funções kernel como Polinomial e RBF.

Figura 12 - Localização do algoritmo SMOReg no WEKA.



3 MATERIAL E MÉTODOS

3.1 CARACTERIZAÇÃO DAS ÁREAS DE COLETA DE AMOSTRAS DE SOLO

As duas áreas pesquisadas situam-se no Estado do Paraná (Figura 13). A primeira área de estudo, denominada Fazenda Estância dos Pinheiros, está localizada no município de Ponta Grossa (25°8’ S, 50°15’ W), e altitude média de 853 m, cujo solo é classificado como Latossolo Vermelho-Escuro distrófico textura média e com alta acidez (JORIS, 2011). Cabe ressaltar que esta área de estudo está há mais de 15 anos sob plantio direto. O clima da região é classificado como Subtropical Úmido Mesotérmico, com temperaturas amenas no verão e ocorrência de geadas severas e freqüentes no inverno, não apresentando estação seca definida. A precipitação pluvial média histórica na região é de 1550 mm anuais, a temperatura média histórica é de 17,8 ^{o}C, com média máxima de 24,1 ^{o}C e média mínima de 13,3 ^{o}C (JORIS, 2011).

A segunda área de estudo foi realizada no município de Piraí do Sul, no Centro-Sul do estado do Paraná, Brasil. Sua posição geográfica tem como coordenadas 24º 22' S,50º04’ W. A gleba, com uma extensão de 110 hectares, é composta predominantemente por Latossolos de textura média a argilosa. A precipitação pluvial média anual varia entre 1400 e 1800 mm, não apresentando estação seca bem definida. A estação mais chuvosa inicia em setembro e se estende até março, mas também ocorrem precipitações pluviais freqüentes durante o inverno. Regionalmente, o mês de janeiro é o mais chuvoso, totalizando médias entre 150 e 210 mm, enquanto que o mês de agosto é o mês mais seco, com precipitação pluvial média entre 50 e 90 mm (PROENÇA, 2012).

Figura 13 - Municípios onde foram retiradas as amostras de solo.

3.2 COLETA DAS AMOSTRAS E ANÁLISE CONVENCIONAL DO SOLO

Na primeira área as amostras de solo foram coletadas em maio de 2004 em 5 diferentes camadas na profundidade de 00–60 cm, o estudo utilizou 38 amostras de 00-05 cm, 38 amostras de 05-10 cm, 38 amostras de 10-20 cm, 59 amostras de 20-40 cm e 58 amostras de 40-60 cm totalizando 231 amostras. Todas as amostras foram secas em estufa com circulação forçada de ar a 40 ^{o}C e passadas em peneira com malha de 2 mm. O carbono orgânico foi determinado de acordo com o método WALKLEY& BLACK descrito em Pavan *et al*.(1992).

A segunda área de estudo compreendeu 111 amostras coletadas à profundidade de 00-20 cm, sendo uma amostra por hectare em um grid georreferenciado. O preparo das amostras consistiu em secagem em estufa com circulação forçada de ar a 40^{0}C durante 12 horas. As amostras secas foram moídas em moinho martelo, homogeneizadas e peneiradas em peneira malha de 2 mm (PROENÇA, 2012). As amostras foram analisadas quanto a quantidade de Matéria Orgânica pelo método colorimétrico, descrito em Raij (2001).

3.3 ANÁLISE DE REFLECTÂNCIA DAS AMOSTRAS

Ambos os conjuntos de amostras foram analisados no espectrômetro marca FOSS, modelo XDS Near-infrared, preparado para leitura de espectros de refletância difusa na região do infravermelho próximo. O software utilizado para aquisição dos epectros foi o ISIscan versão 3.2 (PROENÇA, 2012). A preparação das amostras consistia no acondicionamento da terra em um acessório (Figura 14) do próprio equipamento NIRS. Logo após esse acessório era colocado dentro equipamento para que fosse realizada a leitura. A faixa de comprimento de onda lido foi de 400 a 2500 nanômetros com intervalo de 2 nm a cada varredura.

Figura 14 – Foto de uma amostra de solo pronta para ser analisada pelo equipamento FOSS.

Fonte: Proença (2012).

3.4 CONSTRUÇÃO DOS CONJUNTOS DE DADOS UTILIZADOS.

O primeiro conjunto de dados analisado no espectrômetro continha as leituras de 231 amostras com os resultados das 1051 (um mil e cincoenta e um) faixas de comprimentos de onda. Para cada uma das amostras o correspondente teor de carbono foi adicionado ao conjunto, totalizando dessa forma 1052 atributos.

O segundo conjunto de dados analisado teve origem na junção das 111 amostras da segunda área de estudo com a parte que correspondia a profundidade de 00-20 cm da primeira área analisada. Para que houvesse uma padronização, foi calculada a média

aritmética das amostras com profundidade de 00-05 cm, 05-10 cm e 10-20 cm, bem como o teor de carbono. Como nas 111 amostras o levantamento feito foi da quantidade de Matéria Orgânica, foi necessário um cálculo de conversão transformando a quantidade de MO em teor de carbono. Segundo Primavesi (1979), a MO possui em média 58% de carbono. Dessa forma o segundo conjunto de dados foi formado por 111 registros mais 38 da primeira área analisada. Assim como no primeiro conjunto de dados a faixa de leitura dos espectros comprendeu 1050 comprimentos de onda. A Tabela 1 mostra como ficaram organizados os dois conjuntos de dados.

Tabela 1 – Descrição dos conjuntos de dados.

	Bandas Espectrais (nm)	**Carbono**	**Registros**
Conjunto de dados 1	400 a 2498	1	231
Conjunto de dados 2	400 a 2498	1	149

3.5 GERAÇÃO DE MODELOS DE REGRESSÃO MULTIVARIADA

De posse dos conjuntos de dados formatados de acordo com o padrão utilizado pelo software WEKA versão 3.7.10 foi possível o carregamento e aplicação do algoritmo SMOreg. A validação cruzada em 10 blocos foi utilizada para possibilitar resultados mais confiáveis. A avaliação dos resultados teve como base os erros de previsão e os coeficientes de correlação entre os valores dos teores de carbono preditos pelo modelo utilizando espectros vis/NIR e os valores do atributo de referência das amostras do conjunto de calibração.

Cabe ressaltar que os conjuntos de dados não sofreram nenhum tipo de normalização, ou seja os dados eram originais da leitura dos espectros. Para a avaliação o algoritmo SMOreg utilizou o *kernel* polinomial. Padronizou-se o parâmetro de custo C como 10^6. Valor este que representava os melhores resultados e foi obtido após ter-se aplicado a seguinte variação de C (1, 10, 100, 1000, 100000, 100000, 10^6). Segundo Bonesso (2013) o parâmetro C determina um ponto de equilíbrio entre a maximização da margem e a minimização do erro de classificação.

No primeiro experimento de testes todos os resultados, separadamente da faixa do visível e do infravermelho próximo, foram avaliados para os dois conjuntos de dados. Tan (2009) afirma que a SVM funciona bem com dados de alta dimensionalidade e evita o problema do mal da dimensionalidade, que é ocasionado pela degradação do desempenho de

um algoritmo classificador quando se atinge um numero máximo de características para um dado tamanho de amostras.

No segundo experimento de testes optou-se por reduzir o número de atributos para que um modelo de regressão menor fosse gerado. Procurou-se fazer a redução da dimensionalidade de forma a não perder informação por meio dos atributos mais relevantes. Nos casos de bases de dados de alta dimensionalidade pode ocorrer a existência de atributos redundantes, irrelevantes ou ainda, com pouca influência sobre o atributo de interesse (SILVA, 2013).

4 RESULTADOS E DISCUSSÃO

4.1 PRIMEIRO EXPERIMENTO

Os espectros do vis-NIRS separadamente foram testados utilizando o algoritmo SMOReg disponível no WEKA v. 3.7.10. A opção pela utilização do algoritmo SMOReg é justificada pela sua propriedade em gerar equações de regressão, que caracterizava um dos objetivos do trabalho. O Kernel utilizado foi o polinomial de grau 2 e com o parâmetro de regularização C = 10^6. Os dados não sofreram normalização. A validação cruzada em 10 blocos foi utilizada para avaliar a capacidade de generalização. Erros de previsão e os coeficientes de correlação serviram de parâmetro para avaliar a capacidade de predição dos modelos obtidos. A Tabela 2 apresenta os resultados da predição do Conjunto de Dados 1 com os espectros vis-NIRS.

Tabela 2 - Resultado dapredição do Conjunto de Dados 1 com os espectros vis-NIRS.

Parâmetros	**VIS**	**NIRS**
Nº de Amostras	231	231
Atributos Meta	Carbono	Carbono
Bandas Espectrais (nm)	400 a 700	702 a 2498
Coeficiente de Correlação	0,903	0,896
Erro médio absoluto	1,332	1,429
Erro médio da raiz quadrada	1,686	1,734
Erro relativo absoluto	42,29 %	45,37%
Erro relativo de raiz quadrada	42,96 %	44,18 %

A segunda coluna da Tabela 2 apresentou os resultados da aplicação do algoritmo de regressão nos espectros relativos às leituras da região do visível compreendendo o intervalo de 400 nm a 700 nm e tendo como atributo de referência o teor de carbono determinado em laboratório. A terceira coluna refere-se a leitura feita na região do infravermelho próximo no intervalo de 702 nm a 2498 nm também tendo como atributo de referência o teor de carbono. Os coeficientes de correlação alcançados para ambas as leituras estiveram acima dos 0,89. Os erros médios de raiz quadrada obtiveram melhores resultados dos que os encontrados por Filgueiras *et al.*(2012), o qual tinha o objetivo de avaliar os modelos de calibração multivariada PLS (Partial Least Squares) e SVM na determinação do carbono orgânico tendo obtido para PLS 2,99 e para SVM 1,98.

A Figura 15 demonstra a dispersão dos valores determinados de carbono nas análises tradicionais e os valores preditos pelo modelo vis-NIRS, enquanto que a Tabela 3

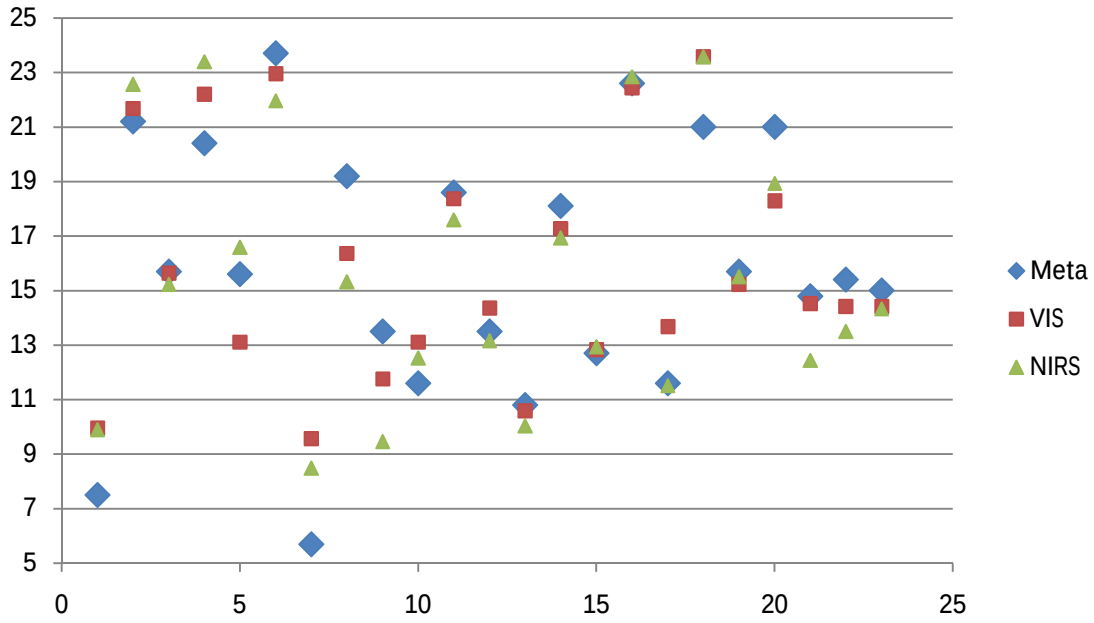
demonstra melhor as estimativas de erros, em aproximadamente 10% das amostras, apontando a diferença do teor de carbono referência e o predito pelo algoritmo.

Tabela 3 - Resultado para o conjunto de dados 1 para 23 amostras de predição e erro dos espectros vis-NIRS.

Meta	Predito VIS	Erro VIS	Predito NIRS	Erro NIRS
7,5	9,96	2,46	9,90	2,40
21,2	21,69	0,49	22,56	1,36
15,7	15,64	-0,06	15,23	-0,46
20,4	22,20	1,80	23.39	2,99
15,6	13,12	-2,47	16,59	0,99
23,7	22,95	-0,74	21,96	-1,73
5,7	9,57	3,87	8,49	2,79
19,2	16,37	-2,82	15,32	-3,87
13,5	11,77	-1,72	9,46	-4,04
11,6	13,12	1,52	12,53	0,93
18,6	18,38	-0,21	17,59	-1,00
13,5	14,36	0,86	13,16	-0,33
10,8	10,59	-0,20	10,04	-0,75
18,1	17,27	-0,82	16,93	-1,16
12,7	12,83	0,13	12,94	0,24
22,6	22,44	-0,15	22,84	0,24
11,6	13,68	2,08	11,51	-0,08
21	23,59	2,59	23,57	2,57
15,7	15,23	-0,46	15,51	-0,18
21	18,29	-2,70	18,93	-2,07
14,8	14,52	-0,28	12,44	-2,35
15,4	14,42	-0,98	13,50	-1,89
15	14,42	-0,57	14,33	-0,67

Por meio da Tabela 3 foi possível observar que amostras com um baixo teor de carbono como 5,7 e 7,5 apresentaram um erro absoluto maior para as leituras da região do VIS e do NIRS, da mesma forma uma concentração mais alta como 19,2 e 21,0 também ocorreram em resultados de erro mais altos. As amostras com menor teor de carbono correspondem às coletas mais profundas, ou seja, de 40 cm a 60 cm. Já amostras com teor mais expressivo de carbono encontram-se mais na superfície com profundidades que variam de 0 a 10 cm. Com base na Figura 15 é possível visualizar a relação existente entre os valores de carbono observados na análises tradicionais versus os valores preditos pelo modelo vis-NIRS para o conjunto de dados 1.

Figura 15 - Gráfico de dispersão com os valores de carbono observados nas análises tradicionais versus os valores preditos pelo modelo vis-NIRS para o Conjunto de Dados 1.



A Tabela 4 refere-se ao Conjunto de Dados 2, o qual apresentou um coeficiente de correlação nos espectros VIS de 0,853 menor que os obtidos em comparação com os espectros VIS do Conjunto de Dados 1. Para as leituras no NIRS o coeficiente de correlação foi de 0,89, valor esse equivalente ao alcançado na faixa do NIRS do Conjunto de Dados 1. Diante desses resultados na faixa do NIRS, pode-se observar que apesar dos conjuntos de dados possuírem uma quantidade de amostras diferente obtiveram coeficientes parecidos, isto em decorrência da técnica SVM possuir a característica de generalizar bem com alta dimensionalidade. Segundo Braun (2012) a SVM sofre menos com o chamado fenômeno de Hughes, que vem a ser uma das principais conseqüências do problema de estimação de parâmetros, frente ao número limitado de amostras e a conseqüente diminuição no valor da acurácia da classificação a partir de uma determinada dimensão dos dados. A Tabela 5 apresenta os erros de regressão para 10% das amostras do Conjunto de Dados 2.

Tabela 4 - Resultado da predição para o conjunto de dados 2 com os espectros vis-NIRS.

Parâmetros	**VIS**	**NIRS**
Nº de Amostras	149	149
Atributos Meta	Carbono	Carbono
Bandas Espectrais (nm)	400 a 700	702 a 2498
Coeficiente de Correlação	0,853	0,897
Erro médio absoluto	1,582	1,303
Erro médio da raiz quadrada	2,072	1,775
Erro relativo absoluto	44,69 %	36,83 %
Erro relativo de raiz quadrada	51,88 %	44,45 %

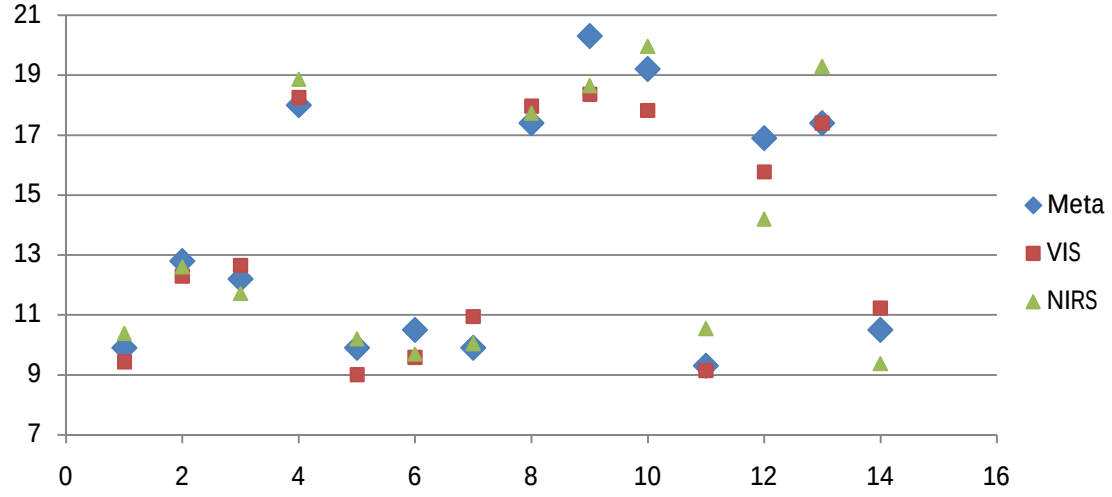
A Figura 16 demonstra a dispersão dos valores determinados de carbono nas análises tradicionais e os valores preditos pelo modelo vis-NIRS, enquanto que a Tabela 5 demonstra melhor as estimativas de erros, em 14 amostras, ou seja, aproximadamente 10%, apontando a diferença do teor de carbono referência e o predito pelo algoritmo de regressão.

Tabela 5 - Resultado para o Conjunto de Dados 2 para 14 amostras de predição e erro dos espectros vis-NIRS.

Meta	**Predito VIS**	**Erro VIS**	**Predito NIRS**	**Erro NIRS**
9,9	9,43	-0,46	10,38	0,48
12,8	12,29	-0,50	12,60	-0,19
12,2	12,66	0,46	11,71	-0,48
18	18,25	0,25	18,86	0,86
9,9	9,0	-0,85	10,20	0,30
10,5	9,587	-0,91	9,69	-0,80
9,9	10,95	1,05	10,04	0,14
17,4	17,96	0,56	17,73	0,33
20,3	18,36	-1,93	18,64	-1,65
19,2	17,83	-1,36	19,96	0,76
9,3	9,14	-0,15	10,54	1,24
16,9	15,78	-1,11	14,19	-2,70
17,4	17,40	0,00	19,29	1,89
10,5	11,22	0,72	9,37	-1,12

Analisando o gráfico da Figura 16 gerado a partir das 14 amostras é possível visualizar que os resultados da predição, tanto para o intervalo VIS quanto o NIRS, estão próximos dos teores de carbono determinados em laboratório.

Figura 16 - Gráfico de dispersão com os valores de carbono observados nas análises tradicionais versus os valores preditos pelo modelo vis-NIRS para o conjunto de dados 2.



4.2 SEGUNDO EXPERIMENTO

No segundo experimento foi aplicada a redução da dimensionalidade para selecionar apenas 10 espectros de cada conjunto de dados. A intenção era gerar uma equação de regressão de tamanho menor, uma vez que, um dos resultados apresentados pelo algoritmo SMOReg é uma equação tendo como parâmetro de entrada as leitura dos espectros e como saída a predição em questão. Para a seleção dos atributos foi utilizado o filtro de seleção baseado em correlação o *CfsSubsetEval (Correlation-based Feature Selection)* com algoritmo *GreedyStpewise* disponível no WEKA. Esse seletor analisa a capacidade de predição de cada atributo no subconjunto juntamente com o grau de redundância entre os atributos. Um subconjunto é considerado bom se os atributos contidos nele são altamente correlacionados com a classe e contém atributos não correlacionados entre si. Para os Conjunto de Dados 1 e 2, com a redução da dimensionalidade foram selecionadas as bandas espectrais mais relevantes para o estudo, conforme apresentado na Tabela 6.

Tabela 6–Espectros selecionados após redução da dimensionalidade

Conjuntos	Bandas Espectrais (nm)
Conjunto de dados 1/vis	438,470,490,502,518,566,596,630,662,694
Conjunto de dados 1/nirs	748,962,1096,1310,1520,1726,1900,2118,2302,2496
Conjunto de dados 2/vis	402,414,450,476,500,512,548,600,630,652
Conjunto de dados 2/nirs	702,900,1086,1304,1514,1720,1932,2108,2318,2498

Da mesma forma que no primeiro experimento os espectros do vis-NIRS foram separadamente testados utilizando o algoritmo SMOReg. O Kernel utilizado foi o polinomial de grau 2 e com o parâmetro de regularização $C = 10^6$. Os dados não sofreram normalização. A validação cruzada em 10 blocos foi utilizada para avaliar a capacidade de generalização. Erros de previsão e os coeficientes de correlação serviram de parâmetro para avaliar a capacidade de predição dos modelos obtidos.

A Tabela 7 demonstra os resultados da aplicação do algoritmo de regressão nas 10 bandas espectrais selecionadas (Tabela 6) nos espectros vis-NIRS, tendo como atributo de referência o teor de carbono.

Tabela 7 - Resultado da predição para o conjunto de dados 1 com os espectros vis-NIRS após a redução de dimensionalidade.

Parâmetros	**VIS**	**NIRS**
Nº de Amostras	231	231
Atributo Meta	Carbono	Carbono
Nº de Bandas Espectrais	10	10
Coeficiente de Correlação	0,891	0,890
Erro médio absoluto	1,394	1,405
Erro médio da raiz quadrada	1,782	1,783
Erro relativo absoluto	44,25%	44,61 %
Erro relativo de raiz quadrada	45,40%	45,42 %

Os resultados obtidos após a redução da dimensionalidade foram expressivos tanto para o intervalo VIS quanto o NIRS alcançando valores de coeficiente de correlação acima de 0,89. Isto atesta que o algoritmo de filtro utilizado demonstrou ser eficiente na seleção dos espectros escolhidos. De posse dos espectros selecionados foi possível gerar as equações 24 e 25 de regressão por meio do algoritmo SMOReg para os espectros vis-NIRS selecionados.

$$\begin{aligned} &\textit{Conjunto de dados 1/vis} = (-314{,}6805 * r438 - 887{,}7766 * r470 - 242{,}2445 * \\ &r490 - 36{,}7462 * r502 + 960{,}5514 * r518 + 391{,}4396 * r566 + 682{,}9357 * \\ &r596 + 925{,}6285 * r630 - 3881{,}1982 * r662 + 2207{,}5724 * r694 + 8{,}7281) \end{aligned} \tag{24}$$

$$\begin{aligned} &\textit{Conjunto de dados 1/nirs} = (-82{,}8045 * r748 + 1028{,}3264 * r962 - 1268{,}0595 * \\ &r1096 + 117{,}4458 * r1310 - 168{,}5211 * r1520 + 1102{,}7026 * r1726 + \\ &162{,}9387 * r1900 - 1177{,}1477 * r2118 + 156{,}2091 * r2302 + 203{,}3498 * \\ &r2496 + 21{,}218) \end{aligned} \tag{25}$$

Para certificar o poder de previsão das equações obtidas, todo o conjunto de dados1foi utilizado como entrada em um teste de validação. Os resultados de 197 amostras mais representativas, ou seja 85% da base, serviu como fonte para que fosse gerado um gráfico de dispersão com o respectivo coeficiente de determinação tanto para a região do Visível (Figura 17), quanto para a região do Infravermelho Próximo (Figura 18).

Figura 17 - Gráfico de dispersão com teor de carbono estimado versus teor de carbono de referência para espectros VIS do conjunto de dados 1.

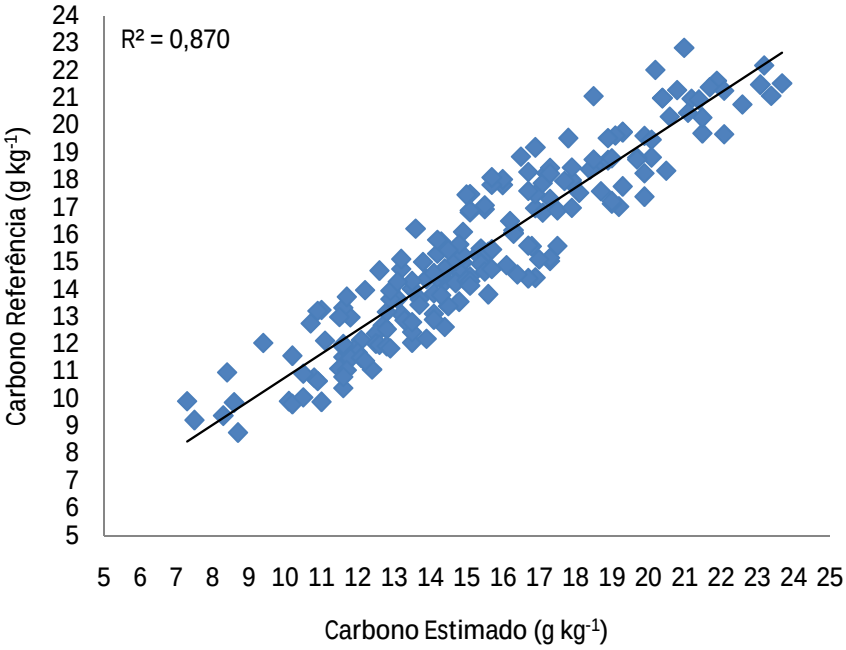
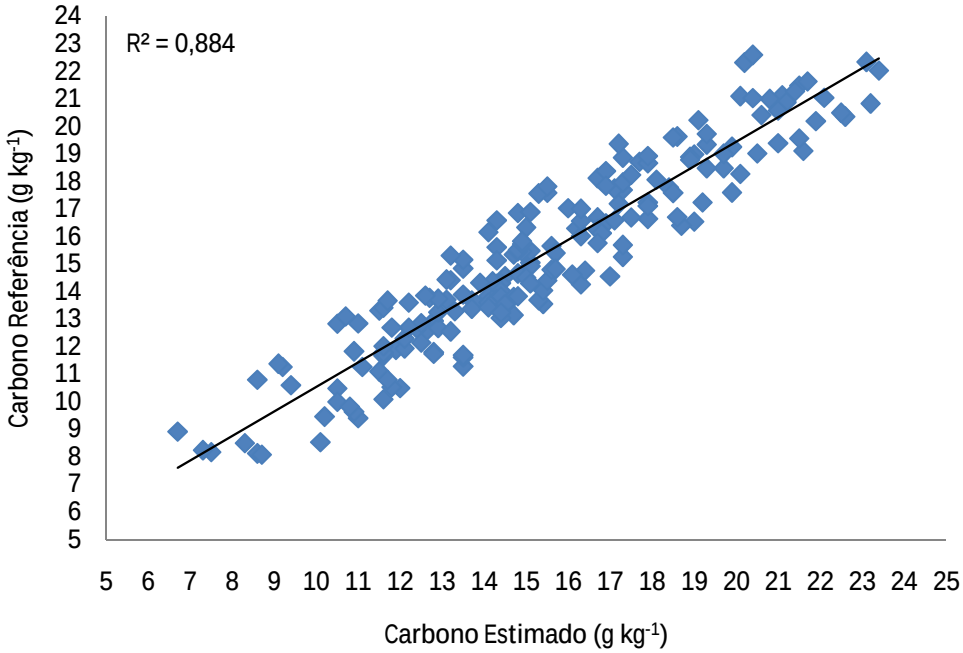


Figura 18 - Gráfico de dispersão com teor de carbono estimado versus teor de carbono de referência para espectros NIRS do conjunto de dados 1.



Equações resultantes da aplicação do algoritmo SMOReg no conjunto de dados 2 dos espectros VIS e NIRS após redução da dimensionalidade.

*Conjunto de dados 2/vis = (47,.3078 * r402 + 105,4959 * r414 + 909,4547 * r450 – 2823,98 * r476 – 424,9005 * r500 + 1104,3096 * r512 + 147,0866 * r548 + 812,4116 * r600 + 1037,9412 * r630 – 1737,3379 * r652 + 29,5467)* (26)

*Conjunto de dados 2/nirs = (276,245 * r702 – 855,6295 * r900 + 592,5297 * r1086 – 203,0655 * r1304 + 830,2072 * r1514 – 538,6452 * r1720 – 151,5565 * r1932 + 133,9365 * r2108 – 259,0109 * r2318 + 165,1943 * r2498 + 12,768)* (27)

Para a obtenção dos resultados (Tabela 8) da predição no conjunto de dados 2após a redução da dimensionalidade, foram utilizados os mesmos parâmetros de kernel e coeficientes de medida empregados no conjunto de dados 1.

Tabela 8 - Resultado da predição para o conjunto de dados 2 com os espectros vis-NIRS após a redução da dimensionalidade.

Parâmetros	**VIS**	**NIRS**
Nº de Amostras	149	149
Atributo Meta	Carbono	Carbono
Nº de Espectros	10	10
Coeficiente de Correlação	0,840	0,875
Erro médio absoluto	1,655	1,425
Erro médio da raiz quadrada	2,165	1,920
Erro relativo absoluto	46,78	40,28 %
Erro relativo de raiz quadrada	54,20	48,06 %

Da mesma maneira foram gerados gráficos de dispersão (Figuras 19 e 20) para o conjunto de dados 2, demonstrando o coeficiente de determinação para 85% das amostras mais representativas.

Figura 19 - Gráfico de dispersão com teor de carbono estimado versus teor de carbono de referência para espectros VIS do conjunto de dados 2.

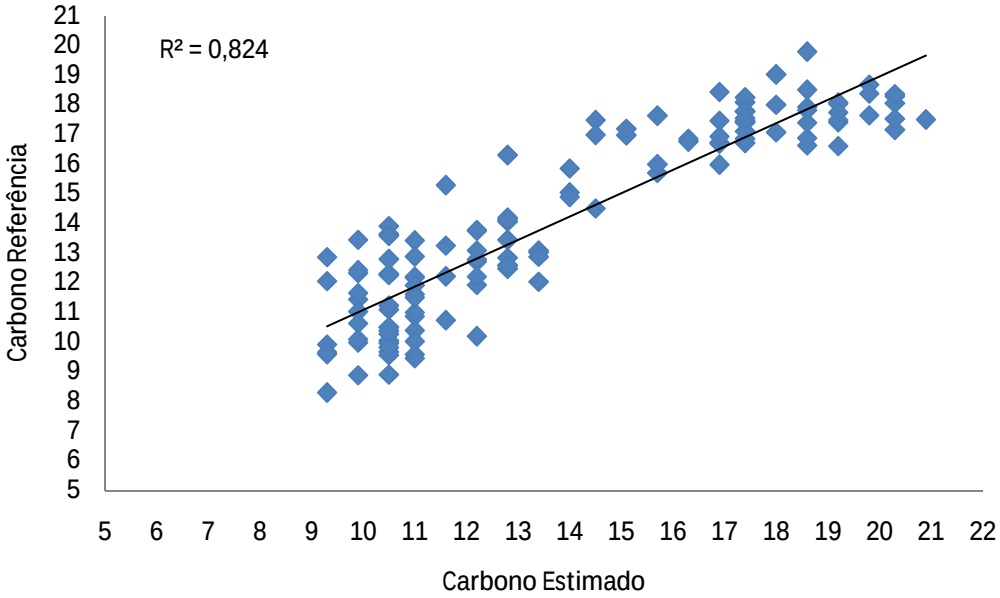
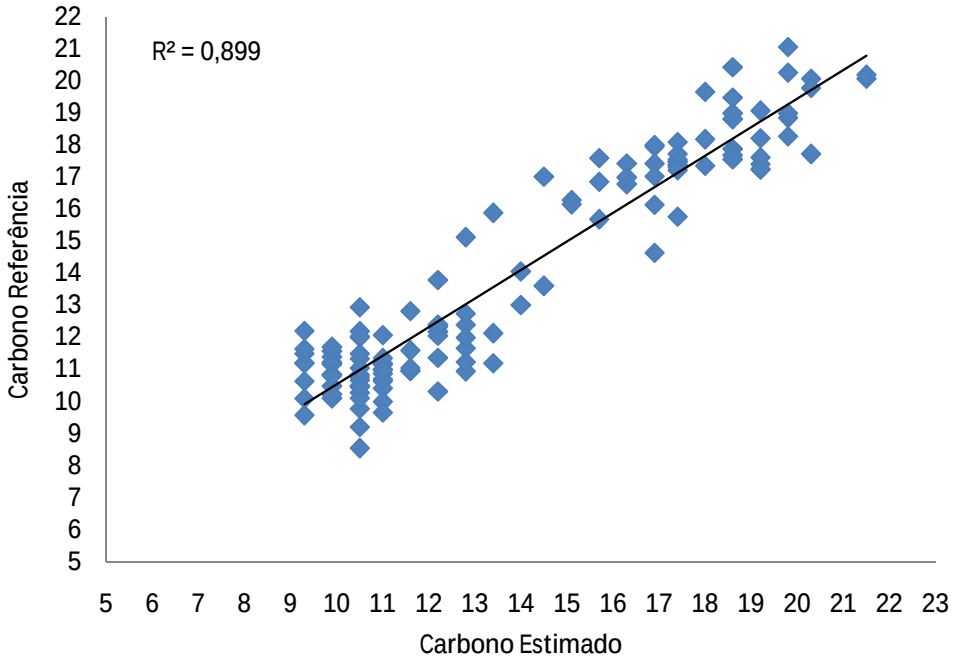


Figura 20 - Gráfico de dispersão com teor de carbono estimado versus teor de carbono de referência para espectros NIRS do conjunto de dados 2.



Ainda que com amostras coletadas de diferentes profundidades o Conjunto de Dados 1 apresentou um coeficiente de determinação melhor na região do visível com R^2 de 0,89 em relação ao conjunto de dados 2 que obteve um R^2 de 0,82. Porém comparando os espectros NIRS dos dois conjuntos, a diferença foi praticamente insignificante.

Os resultados obtidos com a determinação do carbono são semelhantes aos encontrados por Sato (2013) (R^2 0,86) na pesquisa feita em 64 amostras de solo do Cerrado brasileiro utilizando a faixa de 1100 nm a 2500 nm. Já Marchão ***et al.***(2011) obteve resultados mais expressivos, (R^2 0,91) utilizando um conjunto de 390 amostras e os comprimentos de onda compreendidos entre 1100 nm e 2500 nm. Uma regressão pelo método dos mínimos quadrado parciais modificada (mPLS) foi utilizada para desenvolver os modelos de calibração.

Fora do Brasil MCCarty ***et al.***(2010) no oeste da África realizando uma comparação entre espectroscopia na região do infravermelho próximo (NIRS) e médio (MIR) na predição de carbono obteve um coeficiente de determinação também significativo com um R^2 de 0,90.

Utilizando Redes Neurais Artificiais Proença (2012) analisou as mesmas 111 amostras da segunda área de estudo no ano de 2012 e atingiu um R^2 de 0,75 na predição de MO como seu melhor resultado para dados não transformados com redução da dimensionalidade. Com dados transformados, Proença atingiu um R^2 de 0,82 idêntico ao alcançado na predição do carbono guardadas as devidas proporções.

5. CONCLUSÕES

Os resultados alcançados na aplicação da técnica de aprendizado de máquina SVM em conjuntos de dados de espectros na região do visível e do infravermelho próximo obtidos da leitura de amostras de solo para predição de carbono foram satisfatórios, tanto para coletas realizadas em diferentes profundidades quanto para diferentes localizações geográficas, dentro do mesmo Estado da região sul do Brasil e para um mesmo tipo de solo. A principal motivação por utilizar a técnica SVM foi à geração de modelos que apresentam boa generalização em conjunto de dados de alta dimensionalidade, o que foi comprovado nos resultados obtidos no primeiro experimento. Porém, mesmo com a redução da dimensionalidade, o segundo experimento demonstrou a capacidade da técnica em predizer carbono em amostras de solo. Um dos objetivos do trabalho era a geração de equações numéricas de regressão, a qual foi possível com a utilização do algoritmo SMOReg juntamente com o seletor de atributos.

A técnica SVM qualifica a espectroscopia no vis-NIRS para análises de teores de carbono em solos. Desta forma, é possível recomendar a espectroscopia no vis-NIRS aliada a técnica de Máquina de Vetor de Suporte como uma alternativa aos métodos convencionais de análise de carbono em solos, destacando ser mais limpa, ou seja não poluente, e viável também em termos de rapidez no resultado.

6. TRABALHOS FUTUROS

Como trabalhos futuros, algumas sugestões podem ser seguidas para enriquecer a técnica utilizada neste trabalho.

- Utilizar as equações obtidas tanto para VIS quanto para NIRS em amostras de solo de outras regiões com a intenção de uma maior validação.
- Comparar a técnica SVM com outras técnicas de Aprendizado de Máquina na predição de carbono utilizando espectros vis-NIRS.
- Fazer a redução de dimensionalidade do conjunto de espectros utilizando outros algoritmos de filtro.

REFERÊNCIAS

ALES, T. V; GEVERT, G.V; CARNIERI, C; SILVA, L. C. A. Análise de Crédito Bancário Utilizando O Algoritmo Sequential Minimal Optimization. XLI Simpósio Brasileiro de Pesquisa Operacional 2009 – Pesquisa Operacional na Gestão do Conhecimento, p. 2242-2253,2009.

ALKIMIM, A. F., et al. Avaliação da refletância espectral de solos representativos da bacia do rio Benevente com o emprego da análise de componentes principais. Anais XV Simpósio Brasileiro de Sensoriamento Remoto - SBSR, Curitiba, PR, Brasil, 30 de abril a 05 de maio de 2011, INPE página 9064, 2001

AMADO, T.J.C., GIOTTO, E. A sua lavoura na tela. Revista A Granja, Editora Centaurus - São Paulo, SP, edição 732 - p.38-42, 2009.

BALENA, S. P. Correlação de Analises Físico-Quimicas e Espectroscopicas deLaboratório com Dados Obtidos em Campo por Espectrorradiometro. Tese (Doutorado em Química) – UFPR - Curitiba-Pr., 2011.

BASAK, D.; PAL, S.; PATRANABIS, D. C. Support vector regression. Neural Information Processing, v. 11, n. 10, p. 203-224, 2007.

BAYER, C.;MIELNICZUK, J. Dinâmica e funções da matéria orgânica do solo. In: SANTOS, G. de A.; SILVA, L. S.; CANELLAS, L. P.; CAMARGO, F. A. O.(Ed.)Fundamentos da matéria orgânica do solo: Ecossistemas tropicais & subtropicais, 2ª Ed. Porto Alegre, Editora Metrópole, p.7-18, 2008.

BISHOP, C. M. Pattern Recognition and Machine Learning.Springer, 2007, 738 p.

BISOGNIN, G., Utilização de Máquinas de Vetores de Suporte para predição de Estruturas Terciárias de Proteénas. Dissertação (Mestrado em Computação Aplicada) - Universidade do Vale do Rio dos Sinos - São Leopoldo-RS, 2007.

BODDEY, R. M. et al. Carbon accumulation at depth in Ferralsols under zero-till sbtropical agriculture. Global Change Biology, Oxford, v. 16, p. 784-795, 2010.

BONESSO, D., Estimação dos Parâmetros do Kernel em um Classificador SVM na ClassificaÇão de Imagens Hiperespectrais em uma Abordagem Multiclasse. Dissertação (Mestrado em Sensoriamento Remoto) -Universidade Federal do Rio Grande do Sul, 108 páginas, 2013.

BRAUN, A.C.; WEIDNER, U. ; HINZ, S. Classification in High-Dimensional Feature Spaces Assessment Using SVM, IVM and RVM With Focus on Simulated EnMAP. IEEE Journal v.5 Issue 2 Data. Pages. 433-446, 2012

BRUNETTO, G.; MELO, G.W.; KAMINSKI, J.; FURLANETTO, V.; FIALHO, F.B. Avaliação do método de perda de peso por ignição na análise de matéria orgânica em solos da Serra Gaúcha do Rio Grande do Sul. Ciência Rural, 36(6):1936-1939, 2006.

BURGES, C. J. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, p. 243-245, 1998.

CAASVERAS1, J. C.; BARRN, V.; DEL CAMPILLO, M. C. Reflectance spectroscopy: a tool for predicting soil properties related to the incidence of Fe chlorosis. Spanish Journal of Agricultural Research, p. 1133-1142, 2012.

CAMPBELL, C., An introduction to kernel methods. In R. J. Howlett and L. C. Jain, editors, Radial Basis Function Networks: Design and Applications, pages 155-192, Berlin. Springer Verlag, 2000.

CARVALHO, D. R.Árvores de Decisão/Algoritmo Genético para tratar o problema de pequenos disjuntos em classificação de dados. Tese (Doutorado em Computação) - Universidade Federal do Rio de Janeiro, 173 páginas, p.9, 2005.

CHAVES, A. C. F. Extração de Regras Fuzzy para Máquina de Vetor de Suporte (SVM) para Classificação em Múltiplas Classes. Tese (Doutorado em Engenharia Elétrica) -Pontifícia Universidade Católica do Rio de Janeiro,116 páginas, 2006.

CHEN, C.; YANG, E.; JIANG, J.; LIN T.; An imaging system for monitoring the in-and-out activity of honey bees. Computers and Electronics in Agriculture, n. 89, p. 100-109, 2012.

CORNÉLIO, N. N. Estudo sobre os impactos de variáveis do desenvolvimento de software na qualidade do produto. Dissertação (Mestrado em Gestão do conhecimento) - Universidade Católica de Brasília, 94 páginas, 2012.

CRAMMER, K. & SINGER, Y. On the learnability and desing of output codes for multiclass problems. In Computacional Learning Theory, p. 35-46, 2000.

CRISTIANINI, N., SHAWE-TAYLOR, J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000.

DEMATTÊ J.A.M; CAMPOS R.C; ALVES M.C.; FIORIO P.R.; NANNI M.R. Visible-NIR Reflectance: a new approach on soil evaluation. Geoderma, n. 121, p. 95-112, 2004.

FERNANDES, F. A. et al.Uso de espectrometria de refletância no infravermelho próximo (NIRS) na análise de carbono de Neossolos do Pantanal.Comunicado Técnico, Embrapa, Corumbá - MS, 2010.

FERRARESI, Tatiana Maris et al. Espectroscopia de infravermelho na determinação da textura do solo. Rev. Bras. Cinc. Solo, Viçosa , v. 36, n. 6, 2012.

FILGUEIRAS, P. R. et al.Avaliação de modelos de calibração PLS e SVM na determinação do carbono orgânico do solo por espectroscopia NIR. Reunião anual da sociedade Brasileira de Química, 35., Águas de Lindóia, SP, 2012.

FIORO, P. R.; DEMATTE, J. A. M. Orbital and Laboratory Spectral Data to Optimize Soil Analysis. Sci. Agric., v. 66, p.250-257, 2009.

FONSECA, A. de S. et al. Síntese e caracterização estrutural do ligante isatina-3-(N4-benziltiossemicarbazona) e do seu complexo de mercúrio(II). Quím. Nova, São Paulo, v. 33, n. 7,2010.

GONÇALVES, A. R., Máquina de Vetor de Suporte. Disponível em: http://www-users.cs.umn.edu/~andre/arquivos/pdfs/svm.pdf, Acessado em 12 de Novembro de 2013.

HALL, Mark et al. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Witten.Volume 11, Issue 1, 2009.

HAYKIN, S. Redes Neurais – Princípios e práticas. Editora Bookman, Porto Alegre, 2ª edição, 899p, 2001.

INAMAZU, R. Y. et al.Estratégia de implantação, gestão e funcionamento da Rede Agricultura de Precisão. Agricultura de Precisão: um novo olhar, Embrapa, 2012 p 31-40

JACHOWSKI, N. R. A. et al., Mangrove biomass estimation in Southwest Thailand using machine learning. Applied Geography, v.45, p. 311-321, 2013.

JORIS, H. A. W., Calagem Superficial, umidade do solo e comportamento do milho cultivado em sistema plantio direto. Dissertação (Mestrado em Agronomia) – Universidade Estadual de Ponta Grossa, 63 páginas, 2011.

KEQIANG, Y.; YANRU Z.; XIAOLI L.; YONGNI S.; FENGLE Z.; YONG H. Identification of crack features in fresh jujube using Vis/NIR hyperspectral imaging combined with image processing. Computers and Electronics in Agriculture, v. 103, p. 1-10, 2014.
KINTO, E. A. Otimização e análise das máquinas de vetores de suporte aplicadas à classificação de documentos. Tese (Doutorado em Engenharia Elétricao - Escola Politécnica da Universidade de São Paulo, 160 páginas. 2011.

KITCHEN, N. R. Emerging technologies for real-time and integrated agriculture decisions. Computers and Electronics in Agriculture, v. 61, p. 1-3, 2008.

LEMOS, A. M., Matéria Orgânica e perdas de solo, água e nutrientes por erosão em sistemas de preparo e de adubação orgânica e mineral em argissolo vermelho amarelo. Dissertação (Mestrado em Ciência do Solo) - Universidade Federal do Rio Grande do Sul, 104 páginas, 2011.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma introdução às Máquinas de Vetor de Suporte. Relatórios Técnicos do IMC, Universidade de São Paulo Instituto de Ciências Matemáticas e de Computação, 2003.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma introdução às Support Vector Machines. Revista de Informática Teórica e Aplicada, v. 14, p. 43-67, 2007.

MACHADO, P. L. O. A.; BERNARDI, A. C. C.; SILVA, C. A. Agricultura de precisão para o manejo da fertilidade do solo em sistema plantio direto. Rio de Janeiro : Embrapa Solos, p.14, 2004.

MACHADO, P. L. O.; BERNARDI, A. C. C.; SANTOS, F. S. Métodos de Preparo de Amostras e de Determinação de carbono em Solos Tropicais, Circular Técnica 19, Embrapa Solos, p. 9, 2003.
MADARI, B.E.; REEVES, J.B.; MACHADO, P.O.A.; GUIMARÃES, C.M.; TORRES, E. & McCARTY, G.W. Mid and near-infrared spectroscopic assessment of soil compositional parameters and structural indices in two Ferrasols. Geoderma, n.136, p.1-15, 2006.

MAPA. Ministério da Agricultura, Pecuária e Abastecimento. Agricultura de Precisão. Boletim Técnico. Brasília, 36 páginas, 2013. Disponível em:http://www.agricultura.gov.br/arq_editor/Boletim%20T%C3%A9cnico%20AP.pdf. , Acessado em 10 de Dezembro de 2013.

MARCHÃO, R.L.; BECQUER, T.; BRUNET, D. Predição dos teores de carbono e nitrogênio do solo utilizando espectroscopia de infravermelho próximo. Planaltina, DF: Embrapa Cerrados, página 21, 2011.

MARETTO, D. A. Aplicação de máquinas de vetores de suporte para desenvolvimento de modelos de classificação e calibração multivariada em espectroscopia no infravermelho. Tese (Doutorado em Ciências) - Universidade Estadual de Campinas, Instituto de Química, 133 páginas, 2011.

MARTINEZ, R. F. et al. Predictive modelling in grape berry weight during maturation process: comparison of data mining, statistical and artificial intelligence techniques. Spanish Journal of Agricultural Research, n. 9(4), p. 1156-1167, 2011.

MCCARTY, G.W.; REEVES III, J.B.; YOST, R.; DORAISWAMY, P.C.; DOUMBIA, M. Evaluation of methods for measuring soil organic carbon in West African soils. African Journal of Agricultural Research, página 2177, 2010.

McDOWELL, M. L. et al. Effects of Subsetting by Carbon Content, Soil Order, and Spectral Classification on Prediction of Soil Total Carbon with Diffuse Reflectance Spectroscopy. Aollied and Environmental Soil Science. 14 páginas, 2012.

MIELNICZUK, J. Matéria orgânica e a sustentabilidade de sistemas agrícolas. Fundamentos da matéria orgânica do solo: ecossistemas tropicais e subtropicais. Genesis, Porto Alegre, p. 1-8, 1999.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos de aprendizado de máquina. In S. O.Rezende, editor, Sistemas Inteligentes - Fundamentos e Aplicações,Editora Manole, p. 89-114, 2003.

NANNI, M. R.; OLIVEIRA, R. B.; CHICATI, M. L.; CEZAR, E. Utilização de espectrorradiometria na discriminação de solos oriundos da intercessão Basalto/Arenito no Noroeste do Paraná. Simpósio Brasileiro de Sensoriamento Remoto - SBSR, Foz do Iguaçu, PR, Brasil, p. 8988-8995, 2013.

NOVO, E. M. L. M. Sensoriamento Remoto: Princípios e Aplicações. 2. ed. São Paulo: Edgard Blcher Ltda, 1992.

ODEH, I. O. A; TRIANTAFILIS, J.; MCBRATNEY, A. B. Are quantitative methods useful for regional soil inventoriesGeostatitical Aplications for Precision.Springer, United Kingdom, p. 3, 2010.

OLIVEIRA Jr., Seleção de variáveis na mineração de dados agrícolas: Uma abordagem baseada em análise de componentes principais. Dissertação (Mestrado em Computação Aplicada) - Universidade Estadual de Ponta Grossa, Ponta Grossa, 87 páginas, 2012.

OLIVER, M.A., Geostatitical Aplications for Precision, Springer, United Kingdom, p. 3, 2010.

ORRÙ G.; PETTERSSON-YEO W.; MARQUAND A.F.; SARTORI G.; MECHELLI A.Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci Biobehav Rev, n. 36(4), 1140-52, 2012.

PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. Journal of the Brazilian Chemical Society, São Paulo,v. 14, n. 2, p. 198-219, 2003.

PASSERINI, A.; Kernel Methods, multiclass classification and applications to computational molecular biology. PhD thesis, Universit Degli Studi di Firenze, 2004.

PAVAN, M.A.; BLOCH, M.F.; ZEMPULSKI, H.C.; MIYAZAWA, M.; ZOCOLER , D.C. Manual de análise química do solo e controle de qualidade. Londrina: Instituto Agronômico do Paraná(Circular, 76), p. 38, 1992.

PAVINATO, A. carbono e nutrientes no solo e a sustentabilidade do sistema soja-algodão. 118 f. Tese (Doutorado em Ciência do Solo) - Universidade Federal do Rio Grande do Sul. Porto alegre, 2009.

PIMENTA, A.; VALENTIM P. WEKA-G: mineração de dados paralela em grades computacionais. Revista de Sistemas de Informação da FSMA, n.4, p. 2-9, 2009.

PLATT, J. C., Fast training of support vector machines using sequential minimal optimazation, MIT Press Cambridge, MA, USA, p. 185-208, 1999.

PRIMAVESI, A. Agricultura em regiões tropicais. Editora Nobel – São Paulo, 1979.

PROENÇA, C. A. Redes Neurais Artificiais para predição dos teores de matéria orgânica e argila do solo na região dos Campos Gerais utilizando Espectroscopia de Reflectância Difusa. Dissertação (Mestrado em Computação Aplicada) - UEPG, Ponta Grossa – PR., 2012.

RAIJ, B.V.Fertlidade do Solo e Manejo de Nutrientes. Piraciba Internacional Plant Nutrition Institute, 2001.

REIS, C. E. S. Estoque e qualidade da matéria orgânica e retenção de carbono em perfis de dois latossolos subtropicais sob diferentes manejos. Tese (doutorado em Ciência do Solo. Universidade Federal do Rio Grande do Sul. Porto Alegre, 148 páginas, 2012.

RIBEIRO, A. C. C. Diagnóstico de diabetes tipo II por codificação eficiente e máquina de vetor de suporte. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Maranhão, 52 páginas, 2009.

RODRIGUES, J. S. Técnicas de Segmentação e de Classificação em Imagens. Estudo de um Caso na Aplicação. Tese (Doutorado em Informática Industrial) - Universidade do Minho, Portugal, 277 páginas, 2007.

SANTOS, G. A.; PEREIRA A. B.; KORNDÖRFER, G. H. Uso do Sistema de Análises por Infravermelho Próximo (NIR) Para Análises de Matéria Orgânica e Fração Argila em Solos e Teores Foliares de Silício e Nitrogênio em Cana-de-Açucar. Biosci. J.**,** Uberlândia, v. 26, n. 1, p. 100-108, 2010.

SATO, J.H. Métodos para determinação do carbono orgânico em solos do Cerrado. Dissertação (Mestrado em Agronomia) - Faculdade de Agronomia e Medicina Veterinária, Universidade de Brasília, 2013, 79 páginas.

SEGNINI, A. et al. Estudo comparativo de métodos para a determinação da concentração de carbono em solos com altos teores de Fe (Latossolos**).** *Química Nova,* vol.31, n.1, p. 94-97, 2008.

SHEPHERD,K. D.; WALSH, M. G. Infrared sprecroscopy - enabling an evidence based diagnostic survellance approach to agricultural and environmental management in developing countries: Journal of Near Infrared Spectroscopy, Charlton, v.15, p.1-19, 2007.

SILVA, K. S. Uso da mineração de dados para extração de conhecimento agronômico envolvendo o uso de gesso agrícola**.**Dissertação (Mestrado em Computação Aplicada) - UEPG, Ponta Grossa - PR., 2013.

SILVA, M. M. Uma Abordagem Evolucionária Para o Aprendizado Semi-supervisionado em Máquinas de Vetores de Suporte. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal de Minas Gerais, 106 páginas, 2008.

SMOLA, A. J.; SCHOLKOPF, B. Learning with Kernels. The MIT Press, Cambridge, MA, 2002.

SOUZA, D. M.; MADARI, B. E.; GUIMARÃES, F. F. Aplicação de técnicas multivariadas e inteligência artificial na análise de espectros de infravermelho para determinação de matéria orgânica em amostras de solos. Química. Nova, v. 35, n. 9, p. 1738-1745, 2012.

SOUZA, P. B. Uma estratégia baseada em algoritmos de Mineração de dados para validar plano de operação de voo a partir de predições de estados dos Satélites do INPE. Tese (Doutorado em Computação Aplicada), 173 páginas, 2011.

SUNG, A. H.; MUKKAMALA, S. Identifying important features for intrusion detection using support vector machines and neural networks. In:International Symposium on applications and the Internet Technology, p. 209-216,2003.

TAKAHASHI, A. Máquina de Vetor de Suporte Intervalar. Tese(Doutorado em Ciências) - Universidade Federal do Rio Grande do Norte, Rio Grande do Norte, 67 páginas, 2012.

TAN, P-N.; STEINBACH, M.; KUMAR, V. Introdução ao data mining - mineração de dados. 1. ed. São Paulo, Editora Ciência Moderna, 2009.

TING, K.C. Systems Approach to Precision Agriculture - Challenges and Opportunities, 2008, disponível em http://www.docstoc.com/docs/21081830/Systems-Approach-to-Precision-Agriculture---Challenges-and.

UDOMSIN, W. et al. Prediction of agricultural gross domestic product in Thailand using data mining. Biomedical and Applied Tecnology Journal, n. 2,p. 35-50, 2014.

VAPNIK, V. N. The nature of statistical learning theory. Berlin, Springer-Verlag, 1995.

WALINGA, I. et al. Spectrophotometric determination oforganic carbon in soil. Communications in Soil Science an Plant analysis, New York, v.23,p.1935-1944, 1992.

ZHUANG, L. et al. Parameter Optimization of kernel-based One-class Classifier on Imbalance Learning. Journal of Computers, vol. 1 n. 7, outubro, 2006.

APÊNDICE - RESULTADO DAS REFLECTÂNCIAS vis/NIRS DE 20 AMOSTRASPARA O CONJUNTO DE DADOS 1 E CONJUNTO DE DADOS 2 JUNTAMENTE COM O TEOR DE CARBONO REFERÊNCIA.

Tabela com o resultado de 20 amostras aleatórias das leituras realizadas noespectrômetro FOSS do conjunto de dados 1 nos comprimentos de onda NIRS obtidos após a redução da dimensionalidade.

Carbono	r702	r900	r1086	r1304	r1514	r1720	r1932	r2108	r2318	r2498
9,3	0,216433	0,272999	0,306482	0,340219	0,347507	0,368671	0,350873	0,369758	0,319038	0,285794
9,9	0,217231	0,268386	0,298822	0,324853	0,326924	0,34542	0,330282	0,34534	0,291684	0,263791
10,5	0,24435	0,300102	0,330826	0,360534	0,361336	0,380503	0,359448	0,376256	0,318298	0,288633
11	0,237263	0,293899	0,326499	0,356998	0,358917	0,378262	0,355132	0,371835	0,315862	0,281492
11,6	0,220996	0,27016	0,301367	0,329392	0,330162	0,349363	0,329875	0,345895	0,288021	0,258507
12,2	0,226569	0,277358	0,308648	0,332874	0,332946	0,352073	0,333517	0,349067	0,290826	0,261458
12,8	0,207322	0,248743	0,28238	0,306355	0,307309	0,327253	0,312611	0,327123	0,268268	0,242316
13,4	0,211498	0,258096	0,290727	0,313502	0,314878	0,334402	0,321009	0,335666	0,279777	0,255769
14	0,240756	0,29618	0,331409	0,370855	0,371505	0,390816	0,362713	0,37835	0,312829	0,275255
15,1	0,227133	0,288233	0,330684	0,375652	0,379375	0,400822	0,360342	0,381048	0,319524	0,265928
16,3	0,233405	0,29516	0,338589	0,383689	0,386326	0,406681	0,364144	0,384341	0,320534	0,265594
16,9	0,235404	0,298815	0,343057	0,390556	0,392916	0,41414	0,368966	0,390202	0,325015	0,26934
17,4	0,198748	0,246389	0,281772	0,313842	0,318074	0,338064	0,319394	0,332408	0,273132	0,236672
19,2	0,249519	0,31585	0,361421	0,413224	0,414618	0,434571	0,388552	0,406312	0,336679	0,282714
20,3	0,238775	0,301052	0,34471	0,3937	0,396158	0,416775	0,371456	0,391053	0,323534	0,268202
20,9	0,247807	0,314141	0,360054	0,411284	0,411785	0,433288	0,384706	0,40667	0,336463	0,283014
21,5	0,179032	0,234108	0,272566	0,308601	0,315226	0,337021	0,315256	0,329981	0,270718	0,233115
23,3	0,165109	0,216247	0,255346	0,289634	0,297349	0,320487	0,305642	0,320108	0,259823	0,230723
24,4	0,181479	0,241426	0,285244	0,322748	0,33016	0,354926	0,337341	0,352932	0,288094	0,259283
25	0,178934	0,229556	0,268262	0,300115	0,306372	0,328907	0,314442	0,328777	0,269969	0,240223

Tabela com o resultado de 20 amostras aleatórias das leituras realizadas no espectrômetro FOSS do conjunto de dados 1 nos comprimentos de onda vis obtidos após a redução da dimensionalidade.

Carbono	r402	r414	r450	r476	r500	r512	r548	r600	r630	r652
9,3	0,071619	0,069806	0,073008	0,077146	0,083368	0,111773	0,133195	0,151808	0,169057	0,187312
9,9	0,062425	0,060077	0,06248	0,065934	0,071436	0,102651	0,12865	0,149287	0,167554	0,186844
10,5	0,075767	0,073607	0,077415	0,082355	0,089798	0,123313	0,14776	0,169157	0,189339	0,21046
11	0,065254	0,063478	0,066094	0,069558	0,074778	0,096491	0,111517	0,125817	0,140547	0,156082
11,6	0,055806	0,052901	0,054786	0,057827	0,062917	0,096616	0,126381	0,149067	0,168488	0,188659
12,8	0,063022	0,05977	0,061975	0,06554	0,071338	0,10572	0,134982	0,15698	0,17686	0,19737
13,4	0,048452	0,045107	0,045913	0,047795	0,051365	0,08263	0,116744	0,141883	0,161227	0,181143
14	0,050019	0,047298	0,048597	0,050872	0,054841	0,0828	0,109048	0,129209	0,14636	0,164558
14,5	0,045471	0,042616	0,043507	0,045375	0,048749	0,07642	0,103605	0,122776	0,138066	0,153916
15,1	0,061587	0,058406	0,060518	0,063945	0,069509	0,10222	0,12976	0,150467	0,169227	0,189033
15,7	0,062311	0,05908	0,061384	0,065097	0,071167	0,10781	0,139288	0,162291	0,182302	0,20315
16,3	0,062107	0,058974	0,061257	0,064902	0,07085	0,10643	0,136879	0,159442	0,179617	0,200412
17,4	0,064098	0,061063	0,063592	0,067516	0,073872	0,11096	0,14202	0,164912	0,184971	0,206007
18	0,03817	0,034591	0,034621	0,035702	0,037941	0,061009	0,086958	0,105769	0,120625	0,136102
18,6	0,045479	0,042574	0,043218	0,044891	0,048056	0,076415	0,106384	0,126448	0,141318	0,157536
19,2	0,064262	0,06121	0,063876	0,067977	0,074575	0,11126	0,140309	0,162275	0,182617	0,203517
19,8	0,063311	0,060483	0,063192	0,067302	0,0739	0,110462	0,139592	0,161315	0,180894	0,20116
20,3	0,040809	0,037279	0,037524	0,038852	0,041505	0,067535	0,096608	0,117375	0,133148	0,149716
21,5	0,045003	0,042098	0,042893	0,044706	0,048061	0,07532	0,101532	0,120353	0,135979	0,152442
25	0,042176	0,038622	0,03893	0,04032	0,043081	0,069399	0,098029	0,118652	0,134917	0,151893

Tabela com o resultado de 20 amostras aleatórias das leituras realizadas no espectrômetro FOSS do conjunto de dados 2 nos comprimentos de onda vis obtidos após a redução da dimensionalidade.

Carbono	r438	r470	r490	r502	r518	r566	r596	r630	r662	r694
5,7	0,11058	0,068936	0,053844	0,050407	0,054276	0,058087	0,078382	0,148592	0,176807	0,194471
6,7	0,106547	0,066716	0,052628	0,049543	0,053406	0,0571	0,076525	0,14252	0,169027	0,185814
7,3	0,097264	0,059968	0,045289	0,041799	0,044573	0,047581	0,064168	0,123197	0,147771	0,163702
8,6	0,110893	0,069885	0,055462	0,052286	0,056348	0,060228	0,080779	0,151393	0,179174	0,196341
9,6	0,104137	0,066974	0,054284	0,051504	0,055376	0,05898	0,077041	0,130611	0,151958	0,166923
10,1	0,111671	0,07164	0,057265	0,053964	0,05784	0,061588	0,081455	0,149574	0,176498	0,193212
10,2	0,105533	0,065159	0,05062	0,047406	0,051404	0,055282	0,075386	0,13959	0,165864	0,18303
10,9	0,108992	0,069876	0,056311	0,053203	0,057093	0,060788	0,080004	0,14259	0,167365	0,183507
11	0,101047	0,065432	0,053634	0,051046	0,054969	0,058536	0,075976	0,125427	0,145156	0,159242
11,5	0,111709	0,070713	0,057123	0,054106	0,058875	0,063215	0,084732	0,148266	0,173602	0,190545
12	0,099705	0,064072	0,051505	0,04861	0,051944	0,055175	0,071857	0,123363	0,144169	0,1586
13,1	0,10047	0,062372	0,048447	0,045231	0,048918	0,052555	0,071223	0,128078	0,151078	0,166811
14,1	0,11158	0,069662	0,055238	0,052001	0,056731	0,061121	0,083253	0,150557	0,17733	0,194925
14,9	0,105168	0,070969	0,061137	0,059478	0,064249	0,068182	0,086337	0,132684	0,150281	0,162681
16	0,110538	0,0732	0,062619	0,061015	0,066619	0,07116	0,091653	0,141358	0,160018	0,173484
17,3	0,106717	0,069512	0,057886	0,055319	0,06013	0,064333	0,084032	0,134971	0,155241	0,169844
18,9	0,117032	0,078119	0,067772	0,066678	0,072868	0,077704	0,099207	0,150869	0,169756	0,182989
21	0,110442	0,073787	0,064502	0,063231	0,069425	0,074307	0,095598	0,144157	0,162431	0,175759
23,1	0,126937	0,083121	0,072112	0,071106	0,078775	0,084709	0,110189	0,16676	0,188288	0,203989
24,3	0,113262	0,074746	0,064735	0,063639	0,070095	0,075133	0,097183	0,148231	0,166973	0,180149
25,7	0,118486	0,078111	0,067581	0,066487	0,073289	0,078599	0,101764	0,154892	0,174463	0,188288

Tabela com o resultado de 20 amostras aleatórias das leituras realizadas no espectrômetro FOSS do conjunto de dados 2 nos comprimentos de onda NIRS obtidos após a redução da dimensionalidade.

Carbono	r748	r962	r1096	r1310	r1520	r1726	r1900	r2118	r2302	r2496
5,7	0,269361	0,310765	0,354405	0,394051	0,381775	0,400206	0,341402	0,366535	0,29115	0,23425
7,3	0,232616	0,280926	0,323903	0,361941	0,354223	0,373045	0,320799	0,346105	0,277855	0,224787
8,6	0,26894	0,309968	0,353157	0,391703	0,382052	0,400414	0,349306	0,370549	0,294219	0,238113
10,5	0,258891	0,303526	0,345134	0,383606	0,375218	0,393328	0,343035	0,364936	0,292211	0,237305
11,7	0,227657	0,284762	0,32955	0,376926	0,370327	0,389663	0,333906	0,356455	0,284341	0,229391
12,8	0,228085	0,284651	0,328988	0,373505	0,367591	0,386676	0,335793	0,35631	0,285346	0,230734
13,7	0,256562	0,302535	0,347298	0,3948	0,382138	0,401268	0,340287	0,364847	0,287131	0,233289
14,9	0,219132	0,278778	0,320625	0,365357	0,369461	0,390619	0,352172	0,371737	0,310643	0,258631
16	0,23233	0,294637	0,337259	0,382519	0,386323	0,405547	0,365972	0,382503	0,317702	0,264226
18,1	0,228034	0,289829	0,334756	0,386504	0,386509	0,407555	0,355371	0,379026	0,310989	0,25678
19,7	0,244518	0,310684	0,356447	0,408529	0,41139	0,432182	0,38649	0,40562	0,336675	0,281517
21	0,234297	0,294657	0,337618	0,387404	0,389207	0,409725	0,361926	0,383576	0,318192	0,264771
21,5	0,249447	0,317423	0,362488	0,413755	0,417894	0,438847	0,394085	0,412686	0,347938	0,291721
21,7	0,259222	0,327882	0,372678	0,422434	0,424511	0,443544	0,39982	0,416529	0,35149	0,296998
22,5	0,24658	0,31221	0,355283	0,401925	0,405653	0,425657	0,384724	0,401401	0,339026	0,284319
23,2	0,254431	0,324843	0,371018	0,423771	0,428725	0,450304	0,407237	0,426178	0,360514	0,305141
23,4	0,266241	0,33736	0,38346	0,435192	0,438034	0,457649	0,413993	0,430415	0,363433	0,308314
23,7	0,250619	0,318904	0,363147	0,412164	0,416093	0,43705	0,391284	0,4132	0,351551	0,298327
24	0,232201	0,296742	0,339311	0,387347	0,390926	0,411539	0,367142	0,386614	0,323451	0,269039
25,7	0,252069	0,320815	0,365887	0,416833	0,419688	0,441424	0,396014	0,415969	0,349861	0,296412