Estabilidade de Classificadores de Decisão em Árvore Binária para Dados Imagem em Alta Dimensão

Denis Altieri de Oliveira Moraes[1]

Victor Haertel[1]

Resumo: Neste trabalho é investigada uma abordagem para classificação de dados imagem em alta dimensão utilizando classificadores de decisão em árvore binária. O objetivo é selecionar uma estrutura binária que forneça a maior estabilidade possível em relação à acurácia média de classificação para dados imagem em alta dimensão. A utilização de um classificador hierárquico em estrutura binária, analisando pares de classes em múltiplas etapas ao invés do conjunto total de classes em uma única etapa, permite extrair variáveis mais adequadas para cada subconjunto particular de classes. Contudo, devido às múltiplas estruturas binárias que podem ser produzidas, a seleção de uma árvore binária que seja ótima no sentido de produzir resultados estáveis é uma tarefa complexa. Considerando somente duas classes em cada etapa do processo, é possível implementar a distância de Bhattacharyya para fins de estimação da separabilidade entre classes e também como critério para redução de dimensões. O critério de decisão utilizado no processo de classificação é o da Máxima Verossimilhança Gaussiana (MVG). Experimentos foram realizados utilizando dados imagem coletados pelo sensor AVIRIS, investigando o comportamento da acurácia da imagem temática produzida em função de diferentes valores para o limiar de classificação e para o número de variáveis utilizadas. A performance da metodologia proposta é avaliada segundo a complexidade da árvore de classificação, o tempo de processamento, a acurácia média final da imagem temática produzida e a estabilidade do classificador.

Abstract: This paper deals with the problem of classifying high-dimensional image data image data using a multiple stage classifier structured as a binary tree. The aim here consists in finding the optimal structure for the binary tree in the sense of achieving a stable accuracy. The advantage presented by a multiple stage classifier lies on the fact that only a sub-set of classes is considered at each stage, allowing a better selection of the features to be used at each node. The binary tree is a particular case of a tree structured classifier, on which only two classes are considered at each node. This peculiarity makes possible the direct use of statistical distances for feature reduction (selection or extraction). In this study the criterion used for feature reduction at each node consists in optimizing the Bhattacharyya distance separating both classes in the node. The optimization of Bhattacharyya distance was based

[1] Centro Estadual de Pesquisas em Sensoriamento Remoto e Metereologia UFRGS - Caixa Postal 15044 - CEP 91501-970 - Porto Alegre – RS, Brasil
{d_altieri@yahoo.com.br}
{victor.haertel@ufrgs.br}

on the covariance matrices. Once the final set of features is obtained at each particular node, the classification is performed using the Gaussian Maximum Likelihood decision rule. Tests were performed using high-dimensional image data collected by the sensor system AVIRIS covering a test area. The criteria to evaluate the performance of the classifiers are: the final accuracy yielded by the classifier, its stability, and the required computer time.

Palavras-chaves: reconhecimento de padrões, imagem hiper-espectral, árvores de decisão, distância de Bhattacharyya.

1 Introdução

Métodos de reconhecimento de padrões para dados imagem obtidos por sensores remotos vem se constituindo em uma área de grande interesse. Neste contexto, dados em alta dimensão como os fornecidos pelos chamados sensores hiperespectrais, que adquirem imagens da mesma cena em um número muito grande de bandas espectrais vem despertando um particular interesse da comunidade internacional. Sabe-se que classes muito semelhantes entre si podem ser separadas com boa acurácia em espaços de alta dimensão, desde que as respectivas matrizes de covariância difiram entre si [12]. Por esta razão, estudos vêm sendo desenvolvidos por vários pesquisadores no sentido de desenvolver métodos de classificação adequados a dados desta natureza.

A metodologia tradicional utilizada para classificação de padrões está ilustrada na Figura 1. Esse classificador é denominado de classificador em estágio único. Nesta abordagem, o processo de classificação é efetuado de uma forma global, isto é, considerando todas as classes em uma única etapa. Esse fato tem uma conseqüência direta no problema de seleção das variáveis a serem empregadas no processo de classificação, isto é, as variáveis selecionadas devem satisfazer a um critério de otimização global na performance do classificador, considerando-se simultaneamente todas as classes envolvidas.

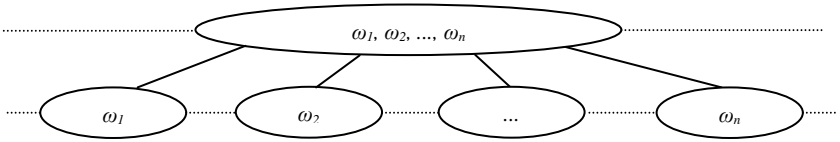

Figura 1. Estrutura geral de um Classificador em Estágio Único

Classificadores em estágio único, como o classificador de Bayes, vêm sendo tradicionalmente utilizados para este propósito. Entretanto, quando aplicados a dados em alta dimensão ocorrem problemas quanto à estimação dos parâmetros das nas funções de decisão que implementam funções densidade de probabilidades condicionadas. Na medida em que a dimensão dos dados cresce, aumenta também o número de parâmetros a serem estimados, especialmente nas matrizes de covariâncias das classes. Por outro lado, em situações reais, o

número de amostras de treinamento disponíveis é limitado, resultando na estimação de parâmetros estatisticamente menos significantes, o que resulta em uma correspondente degradação na performance do classificador. Observa-se que, inicialmente, a acurácia tende a crescer na medida em que a dimensão dos dados também aumenta, ilustrando a contribuição positiva da inserção de informações adicionais no processo de classificação. Eventualmente, este valor atinge um máximo, passando em seguida a declinar, na medida em que a dimensão dos dados continua a aumentar. Esse efeito é conhecido como Fenômeno de Hughes [2], e as metodologias propostas para solucioná-lo seguem três linhas gerais: métodos de análise discriminante regularizada [3], métodos que fazem uso das chamadas amostras semi-rotuladas [4] [5] e métodos que visam reduzir a dimensão dos dados com uma perda mínima de informação [6].

Os métodos para redução na dimensionalidade dos dados inserem-se em duas categorias: seleção de variáveis e extração de variáveis. Na primeira abordagem busca-se definir um sub-conjunto original de variáveis que concentre um razoável poder discriminante entre classes. No segundo caso, é aplicada uma transformação aos dados originais, mapeando-os em um espaço de dimensão menor e mantendo uma parcela significante do poder discriminante entre as classes.

No processo de decisão sobre quais variáveis possuem um maior poder discriminante, os métodos para redução de dimensões fazem uso de medidas de similaridade, ou dissimilaridade, entre classes. Entre as distâncias estatísticas mais utilizadas para esta finalidade estão: a distância de Bhattacharyya, a distância de Jeffries-Matusita, a Divergência e a Divergência Transformada.

Nesse ponto, ocorre uma dificuldade com relação à utilização de classificadores em estágio único, como o classificador de Bayes, pois as distâncias são definidas apenas para um par de classes de cada vez. Nessas situações, costuma-se adotar como critério a média dos valores da distância estatística estimados para pares individuais de classes, o que é um processo obviamente sub-ótimo. Uma possível solução para este problema consiste na adoção de classificadores em estagio múltiplo. Classificadores em estágio múltiplo, como os classificadores de decisão em árvore (CDA) apresentam a vantagem de tratar apenas um subconjunto de classes a cada etapa, possibilitando assim a adoção de subconjuntos diferenciados de variáveis em cada etapa. Assim, em classificadores do tipo CDA, o problema global é particionado em problemas menores – subconjuntos de classes - ao longo dos ramos e níveis da árvore, permitindo otimizar o processo de redução da dimensão dos dados a cada estágio, via seleção ou extração de variáveis, contribuindo assim para reduzir os efeitos do Fenômeno de Hughes. Esta abordagem permite, em princípio, obter probabilidades de erro menores, ou seja, maior acurácia, do que o caso do classificador em estágio único.

Dentre os vários métodos que têm sido propostos para o delineamento da estrutura de um CDA, destacam-se principalmente três abordagens: *top-down*, *bottom-up,* e a união destes dois, conhecido como método híbrido [5] [1], [8], [9] e [10].

Embora esse estudo seja ilustrado com o método de redução de dimensões por extração de variáveis através da utilização de uma otimização específica da distância de Bhattacharyya, outros métodos clássicos para redução de dimensões também foram analisados para comparação dos resultados [13].

2 Metodologia

2.1 Classificador de decisão em árvore

Define-se a estrutura geral de um CDA [11], conforme o esquema apresentado na Figura 2. Nessa estrutura, o nó raiz encontra-se no nível zero, contendo todos os padrões não discriminados pertencentes às *n* classes. Cada nó *t* é composto por uma terna (*C(t)*, *F(t)*, *D(t)*), onde *D(t)* representa a regra de decisão que utiliza o subconjunto de variáveis *F(t)* para discriminar os padrões contidos no nó *t* entre as *C(t)* classes presentes no nó. Esse processo é repetido ao longo dos ramos da árvore, até que apenas uma classe esteja presente no nó. Neste caso, o nó torna-se um nó terminal e recebe um rótulo correspondendo ao da classe em questão.

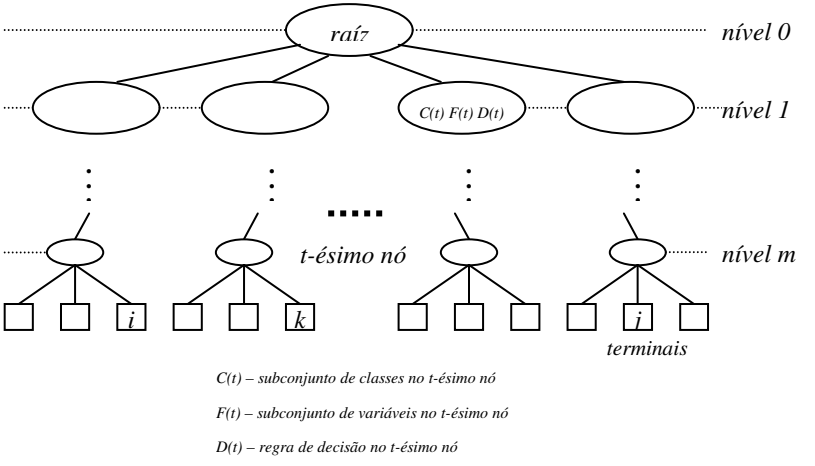

Figura 2. Estrutura geral de um Classificador de Decisão em Árvore (Adaptado de Safavian and Landgrebe, 1991)

Diferentemente do exemplo geral ilustrado na Figura 2, o algoritmo proposto nesse trabalho considera apenas um par de classes a cada nó. Nessa abordagem, dentro do conjunto das classes presentes no nó, são selecionadas as duas que apresentam o maior valor de separação entre si, conforme um determinado critério. Essa estrutura apresenta a notável vantagem de permitir a utilização direta de distâncias estatísticas como, por exemplo, a

distância de Bhattacharyya (1), no processo de seleção/extração de variáveis, o que não seria viável no caso de múltiplas classes. Dessa forma, a etapa de construção da estrutura do CDA binário investigado neste estudo, isto é, a fase de treinamento do classificador utilizando as amostras de treinamento disponíveis, pode ser resumida pelas seguintes etapas:

i) Estimar o grau de separação entre pares de classes, segundo o critério da distância de Bhattacharyya e selecionar o par que apresentar a maior separação. A forma fechada para a expressão da distância de Bhattacharyya para dados com distribuição normal, o que geralmente é satisfeito para classes provenientes de cenas naturais, pode ser lida como a soma de duas parcelas. A primeira, estimando a contribuição dos vetores de médias e a segunda, estimando a contribuição das matrizes de covariância:

$$B = \frac{1}{8}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right)^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) + \frac{1}{2}\left(\frac{\left|\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)/2\right|}{\left|\boldsymbol{\Sigma}_1\right|^{1/2}\left|\boldsymbol{\Sigma}_2\right|^{1/2}}\right) \quad (1)$$

Onde, $\boldsymbol{\mu}_i$ é o vetor de médias e $\boldsymbol{\Sigma}_i$ é a matriz de covariância da classe *i*;

ii) Aplicar um processo de seleção/extração de variáveis com base no par de classes selecionadas no item anterior, para fins de redução da dimensão dos dados dos dados. Neste espaço de dimensão reduzida, estimar os vetores de médias e as matrizes de covariância para as duas classes selecionadas. Esses parâmetros serão utilizados na construção das funções de decisão da Máxima Verossimilhança Gaussiana associadas respectivamente a cada um dos dois nós descendentes;

iii) Classificar as amostras de treinamento das classes restantes em um dos dois nós utilizando a regra de decisão da Máxima Verossimilhança Gaussiana (2), expressa por:

$$G_i\left(\mathbf{X}\right) = -\ln\left|\boldsymbol{\Sigma}_i\right| - \left(\mathbf{X} - \boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{X} - \boldsymbol{\mu}_i\right) \quad (2)$$

Então:

$$\mathbf{X} \in n_L \quad \text{se} \quad G_i\left(\mathbf{X}\right) > G_j\left(\mathbf{X}\right) \;, \forall i \neq j$$

Onde,

n_L e n_R, representam o nó esquerdo e direito;

X é o vetor amostra com *p* variáveis;

$\boldsymbol{\mu}_i$ é o vetor de médias da classe *i*, conforme estimado na etapa (ii);

$\boldsymbol{\Sigma}_i$ é a matriz de covariância da classe *i,* conforme estimado na etapa (ii);

iv) Se desse processo de classificação resultar que o número de amostras pertencentes a uma classe ω_i alocadas a um dos dois nos descendentes superar um *limiar* previamente

estabelecido, a classe ω_i é considerada como pertencente unicamente a este nó. Se esse limiar não for superado, a classe ω_i é atribuída a ambos os nós descendentes. Esse processo continua até que os nós terminais sejam atingidos.

É importante considerar que existem muitas possibilidades de subdivisões em uma árvore binária, tornando-se necessário introduzir a notação ilustrada na Figura 3.

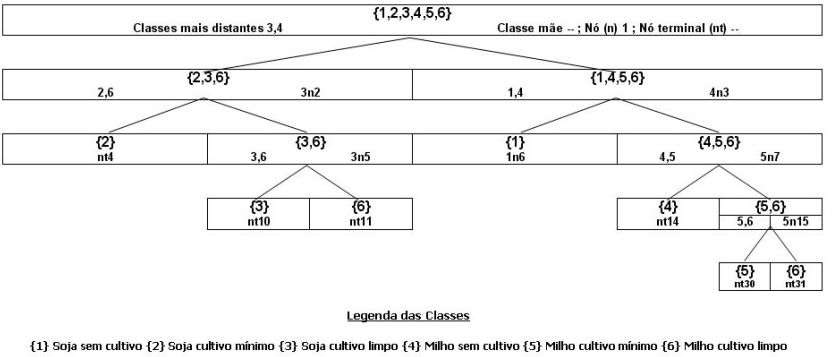

Figura 3. Exemplo de CDA com estrutura binária e cinco classes

Nesta figura, em cada nó do CDA são apresentadas as seguintes informações: os números no primeiro conjunto {1,2,3,4,5,6} se referem à legenda numérica das classes a que pertencem as amostras nesse nível inicial da árvore. O par 3,4 à esquerda, expressa o par de classes que apresenta a maior separação entre si. O número situado à direita do nó refere-se à classe cujos parâmetros (vetor de médias e matriz de covariância) foram utilizados na seleção das amostras que o compõe. O algarismo inicial refere-se à classe cujos parâmetros foram utilizados no processo de seleção de amostras alocadas ao nó; Esta convenção não se aplica ao nó inicial. O número que define o nó encontra-se indicado após o símbolo 'n' e os nós terminais ficam identificados pelo símbolo 'nt'.

2.2 Extração de variáveis

Conforme detalhado na seção 2.1, a distância de Bhattacharyya oferece uma medida conveniente de separação entre duas classes. Um processo de extração de variáveis tendo como critério a otimização dessa distância mostra-se, portanto, promissor. A otimização da distância de Bhattacharyya (1) não se constitui entretanto em uma tarefa trivial, tornando-se necessário adotar soluções sub-ótimas [12]. Neste caso, a otimização pode ser feita levando em consideração (*i*) apenas o primeiro termo em (1), isto é, a contribuição devida à diferença entre os vetores de médias, (*ii*) apenas o segundo termo em (1), isto é, a contribuição devida à diferença entre as matrizes de covariância, (*iii*) otimização de ambos os termos em (1), mas na suposição de dominância do primeiro termo e (*iv*) otimização de ambos os termos em (1), mas na suposição de dominância do segundo termo. Considerando que nesse estudo o objetivo consiste no desenvolvimento de uma metodologia envolvendo a classificação de

padrões pertencentes a classes muito semelhantes entre si, isto é, com vetores de médias muito semelhantes entre si, e ainda que nesta situação o fator preponderante de separação entre as classes está nas matrizes de covariância, adotou-se o critério (*iv*) acima mencionado. O critério (*ii*) não foi utilizado nesse trabalho, pois apesar de ser esperada maior contribuição à separabilidade devido à diferença entre as matrizes de covariância das classes, a contribuição dos vetores de médias, mesmo que pequena, não foi descartada. Um estudo comparativo entre estas quatro possíveis abordagens está apresentado em [14]. No presente estudo o foco maior está na estabilidade da árvore binária.

Na abordagem investigada neste estudo, a otimização da distância de Bhattacharyya (1) é obtida selecionando-se as direções definidas pelos autovetores $\mathbf{\Phi}$, da matriz $\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1$, investigando-se como a separação entre as duas classes devida ao primeiro termo em (1) está distribuida ao longo destes autovetores. Pode-se mostrar [12] que nestas condições a distância de Bhattacharyya pode ser expressa por:

$$B = \sum_{k=1}^{n}\left[\frac{1}{4}\frac{\left\{\varphi_k^T\left(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\right)\right\}^2}{1+\lambda_k}+\frac{1}{4}\left\{\ln\left(\lambda_k+\frac{1}{\lambda_k}+2\right)-\ln 4\right\}\right] \tag{3}$$

Onde φ_k^T, é o k-*ésimo* autovetor transposto, correspondente à λ_k, o k-*ésimo* maior autovalor da matriz $\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1$.

Dessa forma, a redução na dimensão dos dados do valor original *n* para um valor desejado *m* ($n < m$) pode ser obtida selecionando-se os *m* autovetores associados as *m* maiores parcelas em (3).

3 Experimentos

Tabela 1. Total de amostras nos experimentos com seis classes

Código da classe	Nome	Número de amostras
1	Soja sem cultivo	1.006
2	Soja cultivo mínimo	656
3	Soja cultivo limpo	1.821
4	Milho sem cultivo	881
5	Milho cultivo mínimo	1.671
6	Milho cultivo limpo	373

Nos experimentos desenvolvidos neste estudo foram empregados dados coletados pelo sistema sensor AVIRIS, sobre uma área teste denominada de Indian Pines no Estado de Indiana, USA. Esta área é coberta por seis classes de culturas agrícolas espectralmente

muito semelhantes, não separáveis quando se utiliza dados tradicionais com dimensão reduzida, como por exemplo, dados coletados pelo sistema Landsat-TM. Os experimentos foram desenvolvidos utilizando a totalidade dos dados disponíveis, isto é, a cena imageada completa e as 190 bandas espectrais livres de ruído. Inicialmente, foram estudadas seis classes, listadas na Tabela 1.

3.1 Limiar de verossimilhança

Como critério para alocação das classes presentes em um determinado nó aos seus dois nós descendentes, é proposto neste estudo o conceito do *limiar de verossimilhança* (LV). O LV estipula a fração das amostras de treinamento de uma classe que deve ser alocada a um nó descendente para que a classe em questão seja alocada unicamente a este nó. Em outras palavras, se como resultado do processo de classificação das amostras de treinamento de uma dada classe presente em um nó em seus dois nós descendentes, a fração das amostras classificadas em um dos nós descendentes for igual ou superior ao valor estipudado para o LV, então esta classe é alocada unicamente a este nó descendente. Caso contrário, esta classe, com a totalidade de suas amostras de treinamento, é alocada simultaneamente aos dois nós descendentes.

Os experimentos mostraram que o valor adotado para LV apresenta uma influencia grande na estrutura do CDA, maior ainda do que o número de amostras de treinamento utilizadas. Os experimentos mostraram também, que variações no valor adotado para o LV resultam em uma variabilidade significativa na estrutura do CDA resultante, mais especificamente no número de nós terminais, influenciando assim na estabilidade dos resultados obtidos. Os experimentos mostraram também uma influência significativa do LV na acurácia média final produzida pelo classificador.

Uma das estruturas do CDA's produzidas com LV igual a 55% está ilustrada na Figura 4. Para esse mesmo nível de LV, a Figura 5 ilustra o resultado de acurácia individual para cada uma das seis classes e para diferentes valores na dimensão dos dados. A análise destas duas figuras evidencia que com o uso de um LV baixo, igual a 55%, a acurácia individual das classes apresenta grandes variações em função da dimensão dos dados. Para dados com dimensão igual a 25, por exemplo, a acurácia estimada para a classe soja cultivo limpo foi próxima de 25%, enquanto para uma dimensão igual a 30, a acurácia estimada dessa classe foi superior a 85%.

Deve-se notar que a estrutura do CDA ilustrada na Figura 4 é apenas uma das estruturas obtidas com o LV de 55%. A utilização de um valor reduzido para o LV permite que a estrutura da árvore se modifique com o aumento do número de variáveis utilizado em cada nó. Isto acontece porque como em cada nó da árvore são comparados grupos de classes distintos, o incremento no número de variáveis extraídas permite que ora o LV seja ultrapassado, ora não. Desse modo, o número de nós terminais pode ser alterado na medida que o número de variáveis aumenta e o LV permanece constante e baixo (próximo a 50%), classificando um número diferente de amostras de treinamento de uma determinada classe

em apenas um dos nós, superando ou não o LV e produzindo assim um efeito de instabilidade do classificador, conforme pode ser visto na Figura 5.

O comportamento aparentemente periódico na acurácia de classificação individual das classes pode ser devido ao fato de que com o incremento do número de variáveis, a estrutura da árvore binária resulta mais ou menos eficiente para uma classe específica. Dessa forma, torna-se muito difícil prever com exatidão qual é a estrutura específica da árvore, e para que dimensão de variáveis, todas as classes são favorecidas conjuntamente. A estrutura do CDA utilizando um LV igual a 99% é ilustrada na Figura 6. Observa-se que nesse experimento, um valor elevado para o LV resultou em um CDA com a maior estrutura possível, e conseqüentemente com o maior número possível de nós terminais.

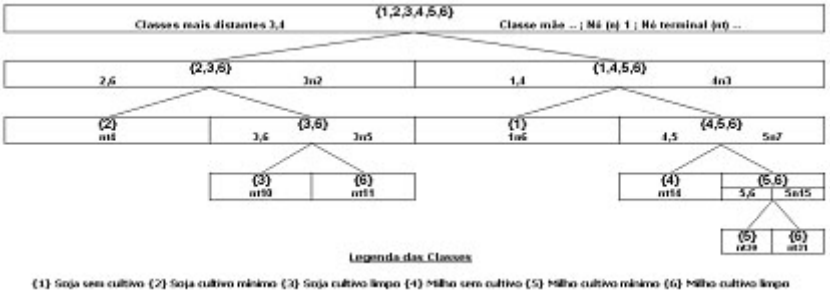

Figura 4. CDA com LV igual a 55%

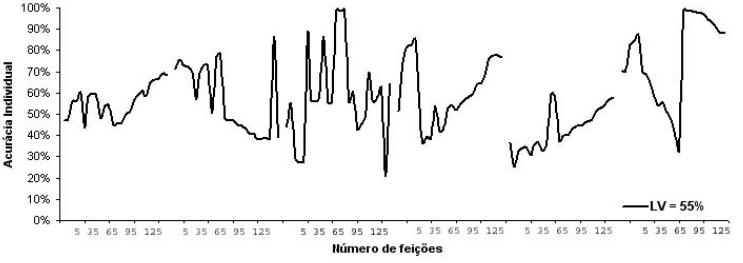

Figura 5. Acurácia do CDA por classe em função do número de variáveis com LV igual a 55%

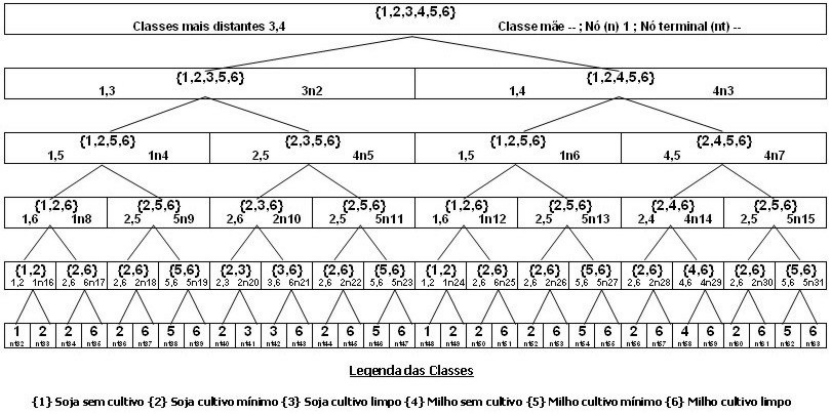

Figura 6. CDA com LV igual a 99%

Examinando a acurácia individual das classes, conforme mostrado na Figura 7, verifica-se que a mesma não é instável como no caso anterior, mas possui um comportamento regular de crescimento na medida em que a dimensão dos dados aumenta, isto é, na medida em que novas variáveis são adicionadas, como seria esperado. Esse resultado se deve ao fato de que, com um valor alto para o LV, um incremento no número de variáveis dificilmente permitirá que o CDA altere sua estrutura, resultando em um comportamento mais estável na acurácia da classificação.

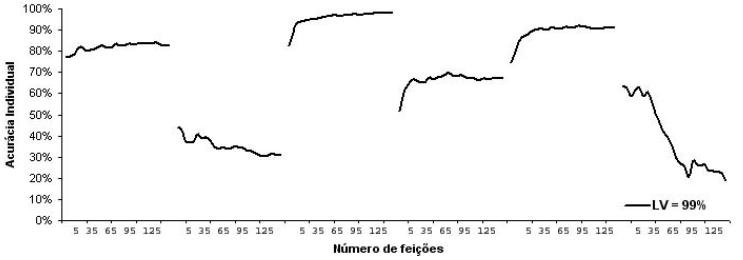

Figura 7. Acurácia do CDA por classe em função do número de variáveis com LV igual a 99%

No caso específico da sexta classe, observa-se nessa figura que há um decréscimo muito acentuado na acurácia de classificação. Esse fato deve-se às características específicas dessa classe, que apresenta um grau de variabilidade muito maior que nas demais classes, além desta também possuir um número de amostras de treinamento bem inferior às demais, conforme pode-se ver na Tabela 1.

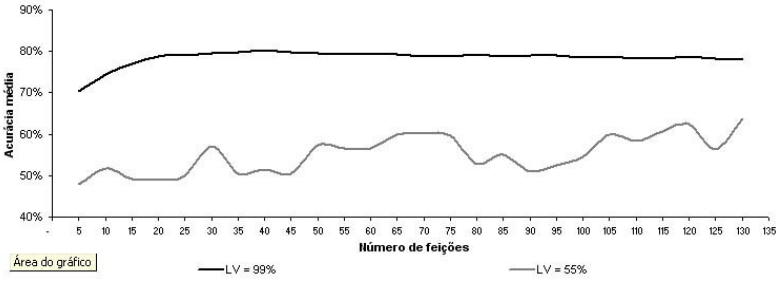

Figura 8. Acurácia média do CDA em função do número de variáveis com LV igual a 55% e 99%

A acurácia média de classificação produzida pelo CDA para as seis classes, como função da dimensão dos dados e dos valores adotados para o limiar LV está ilustrada na Figura 8. Nota-se que com um valor de LV igual a 99%, e a estrutura praticamente invariável que resulta para o CDA, os resultados são superiores ao obtido com um LV igual a 55%, o qual permite maiores variações na estrutura do CDA na medida em que a dimensão dos dados também é alterada.

Observa-se que os valores de acurácia média produzidos pelo CDA com um valor elevado para o LV, apresentam o comportamento esperado, com um crescimento inicial da acurácia na medida em que a dimensão dos dados aumenta, isto é, na medida em que novas variáveis são adicionadas ao CDA, atingindo um máximo e passando a declinar na medida em que o número de variáveis continua crescendo. Esse comportamento caracteriza o fenômeno de Hughes, conformando-se com as predições teóricas.

Os resultados apresentados até aqui sugerem que uma estrutura maior para o CDA e, portanto, um número maior de nós terminais, produz resultados mais acurados e estáveis. Esse fato sugere então uma modificação do algoritmo inicial proposto:

iii) Após a seleção das duas classes que servirão para caracterizar os dois nós descendentes, alocar todas as classes restantes (através de suas amostras de treinamento) em ambos os nós descendentes. Esse processo continua até que reste um subgrupo com apenas duas classes e os nós terminais sejam assim atingidos.

Os experimentos mostraram que esta metodologia modificada produz acurácias mais estáveis e em média mais elevadas no processo de classificação, propiciando ainda um melhor desempenho computacional do algoritmo devido ao fato de que o teste envolvendo o critério de LV torna-se aqui dispensável. Nesta segunda fase dos experimentos decidiu-se eliminar a classe *milho cultivo limpo*, para tornar possível a utilização de um número único de amostras de treinamento e de teste para todas as classes em consideração, conforme descrito na Tabela 2. Da mesma forma que nos experimentos anteriores, são utilizadas aqui todas as 190 feições disponíveis. Para o LV, foram utilizados nestes experimentos três níveis, sendo de 55%, 75% e 100%. Entretanto, este último valor para LV é apenas uma forma de expressar um CDA com a maior estrutura possível, conforme descrito acima.

Tabela 2. Total de amostras nos experimentos com cinco classes

Código da classe	Classe	Amostras de treinamento
1	Soja sem cultivo	500
2	Soja cultivo mínimo	500
3	Soja cultivo limpo	500
4	Milho sem cultivo	500
5	Milho cultivo mínimo	500

Analisando as Figuras 9, 10, observa-se que, na medida em que o valor atribuído ao LV aumenta, diminui a variabilidade no valor estimado para a acurácia de classes individuais, para diferentes valores da dimensão dos dados. Dessa forma, conforme observado também nos experimentos com seis classes, os resultados produzidos com um valor de LV igual a 75% são mais estáveis do que quando utilizando um valor igual a 55%. Pode-se mostrar também que, com o aumento do valor para o LV, são produzidas estruturas mais completas para o CDA, aproximando-se do número máximo de nós terminais.

Figura 9 Acurácia de classificação para o CDA com LV igual a 55%

Figura 10 Acurácia de classificação para o CDA com LV igual a 75%

Figura 11 Acurácia de classificação para o CDA modificado (equivalente a um LV igual a 100%)

Os resultados da Figura 11 ilustram que, como esperado, a metodologia modificada apresenta um padrão de acurácia individual das classes mais estável do que os anteriores. Observa-se na Figura 12, observa-se também que, em média, as acurácias obtidas com um CDA binário com estrutura máxima são superiores aos apresentados por CDA's com estruturas menores.

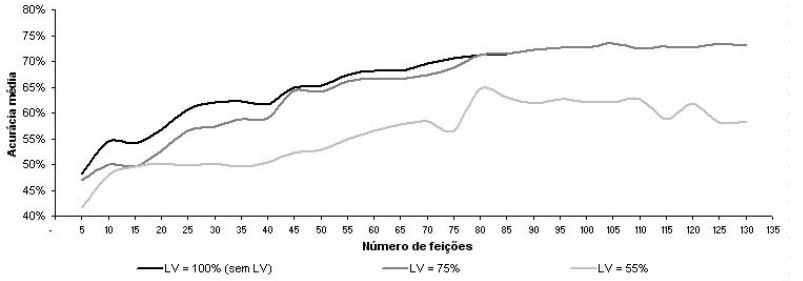

Figura 12 Acurácia média de classificação para o CDA com LV igual a 55%, 75% e 100% (algoritmo modificado)

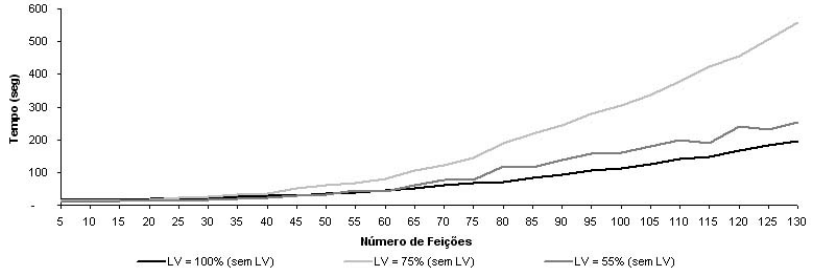

Figura 13 Tempo total de processamento do CDA com LV igual a 55%, 75% e 100% (sem LV)

Na Figura 13 está ilustrado o tempo de processamento do CDA com diferentes estruturas, resultantes de valores distintos para o LV, incluindo aqui o caso equivalente ao de

um valor para o LV igual a 100% conforme o algoritmo modificado descrito acima. Os resultados sugerem a superioridade do algoritmo modificado em termos de desempenho computacional, mostrando um tempo de processamento inferior, até mesmo, no caso de um CDA com as menores estruturas, isto é, com LV igual a 55%.

4 Conclusões

A utilização de classificadores hierárquicos, ao invés da metodologia tradicional de classificação em estágio único tem levantado crescente interesse, na medida em que novos recursos computacionais são disponibilizados. Embora os CDA's sejam teoricamente capazes de produzir resultados de classificação mais acurados, a busca por uma estrutura ideal para o CDA é uma tarefa exaustiva, computacionalmente pesada, mesmo com os recursos computacionais hoje disponíveis. Assim, foi investigado nesse trabalho um CDA estruturado na forma binária que seja capaz de produzir um resultado de classificação estável e ao mesmo tempo com um elevado grau de acurácia.

Os experimentos mostraram que uma metodologia para construção de um CDA com estrutura binária, baseada num valor baixo para o limiar de verossimilhança (LV), pode resultar em diversas estruturas binárias distintas entre si, dependendo do número de variáveis utilizado, caracterizando assim uma instabilidade no classificador. Os resultados inicialmente mostram que um CDA binário com estrutura máxima é capaz não só de produzir um resultado de acurácia mais estável, como também uma acurácia média superior. Os experimentos mostram também que a adoção do método proposto, equivalente a um valor de LV igual a 100%, resulta em desempenho computacional superior ao de CDA's com estruturas menores, na medida em que o tempo de treinamento do classificador é reduzido significativamente.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] You, K. C. and Fu, K. S., **"An approach to the design of a linear binary tree classifier"**, in 3rd. Symp. Machine Processing of Remotely Sensed Data, Purdue Univ., W. Lafayette, IN, 1976.

[2] Hughes, G. F., **"On the mean accuracy of statistical pattern recognizers"**, IEEE Trans. Inform. Theory, vol. IT-14, pp. 55-63, 1968.

[3] Friedman, J. H., **Regularized Discriminant Analysis**, Journal of the American Statistical Association, v. 84, n. 405, p. 165-175, 1989.

[4] Shahshahani, B. M. and Landgrebe D. A., **"The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon"**, IEEE Transactions on Geoscience and Remote Sensing, vol. 32, nº 5, 1087-1095, Sept. 1994.

[5] Jackson, Q., Landgrebe, D. **"An Adaptive Classifier Design for High-Dimensional Data Analysis with a Limited Training data Set"**, IEEE Transactions on Geoscience and Remote Sensing, v. 39, p. 2664-2679, 2001.

[6] Serpico, S. B., D'inca, M., Melgani, F., Moser, G. **"A comparision of feature reduction techniques for classification of hyperespectral remote-sensing data"**, In: Proceedings of SPIE - The International Society for Optical Engineering, 2002. Agia Pelagia, Greece. v 4885, p 347-358.

[7] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., **Classification and Regression Trees (CART)**. Belmont, CA: Wadsworth Int., 1984.

[8] Gelfand, S. B., Ravishankar, C. S., and Delp, E. J., **"An iterative growing and pruning algorithm for classification tree design"**, IEEE Trans. Patt. Anal. Mach. Intell., pp. 163-174, 1991.

[9] Kim, B. and Landgrebe, D. A., **"Hierarquical decision tree classifiers in high-dimensional and large class data**", Ph.D. dissertation and Tech. Rep. TR-EE-90-47, School of Elec. Eng. Purdue Univ., W. Lafayette, IN, 1990.

[10] Kumar, V. and Kanal, L. N., **"A General Branch and Bound Formulation for Understanding and Synthesizing and/or Tree Search Procedures"**, *Artificial Intelligence*, vol. 21, pp. 179-198, 1983.

[11] Safavian, S. R. and Landgrebe, D. A., **"A Survey of Decision Tree Methodology"**, *IEEE Trans. Systems, Man and Cybernetics*, vol. 21, no. 3, May/Jun 1991.

[12] Fukunaga, K., **Introduction to Statistical Pattern Recognition**. 2nd. Ed. Boston: Academic Press, 1990.

[13] Moraes, D. A. O. e Haertel, V., **Extração de Feições em Dados Imagem com Alta Dimensão por Otimização da Distância de Bhattacharyya em um Classificador de Decisão em Árvore.** Dissertação de Mestrado em Sensoriamento Remoto - PPGSR, Universidade Federal do Rio Grande do Sul, CNPq, 99 p., 2005.

[14] Moraes, D. A. O. e Haertel, V., **Extração de Variáveis por Otimização da Distância de Bhattacharyya.** *Artigo submetido à* Revista Brasileira de Geofísica, 2008.