Learning and Nonlinear Models (L&NLM) – Journal of the Brazilian Neural Network Society, Vol. 9, Iss. 1, pp. 66-75, 2011

SELEÇÃO DE CARACTERÍSTICAS APOIADA POR MINERAÇÃO VISUAL DE DADOS

Glenda Michele Botelho, João Batista Neto
Universidade de São Paulo (USP)
Instituto de Ciências Matemáticas e de Computação
{glenda,jbatista}@icmc.usp.br
São Carlos - SP - Brasil

Resumo – A seleção de características é fundamental para minimizar os problemas causados pela alta dimensionalidade. Existem diversos métodos tradicionais de seleção que se baseiam em análises estatísticas dos dados ou redes neurais. Nestes, a qualidade do subconjunto selecionado é dada por meio de alguma função critério. O presente trabalho propõe a inclusão de técnicas de Mineração Visual de Dados, particularmente a projeção de dados multidimensionais, para apoiar o processo de seleção. Os resultados mostram que a técnica é capaz de prover boa redução no espaço de características, ao mesmo tempo que mantém a capacidade de discriminação. A qualidade dos subconjuntos selecionados é comprovada tanto quantitativamente pela medida de silhueta quanto pela qualidade visual das projeções obtidas.

Palavras-chave – Seleção de características, projeção de dados multidimensionais, silhueta, agrupamento.

Abstract – Feature selection (FS) is a fundamental step to minimize problems caused by highly dimensional datasets. Many traditional FS methods are either based on statistical analysis or neural networks. Such methods employ a criterion function to measure the quality of the selected subset. This work proposes visual data mining techniques, multidimensional data projection in particular, to aid the FS process. Results have shown that the proposed technique is capable of providing good reduction of the feature space as well as keeping the discrimination power. The quality of the subset is assessed both quantitatively, by means of the Silhouette measure, and qualitatively, by means of visual inspection of the projections obtained.

Keywords – Feature selection, multidimensional data projection, silhouette, clustering.

1. INTRODUÇÃO

A popularização do uso de tecnologias geradoras de dados digitais, incluindo equipamentos que produzem imagens, quer seja para finalidades médico-terapêuticas, agricultura ou de entretenimento, permitiu o surgimento de grandes banco de imagens digitais, ao mesmo tempo em que impôs grandes desafios aos desenvolvedores de sistemas.

Um destes desafios é a correta recuperação de imagens com base no conteúdo. Via de regra, um sistema de recuperação de imagens por conteúdo tem como entrada não a imagem em si, mas um conjunto de descritores numéricos, organizados no que comumente chamamos de vetor de características. Portanto, cada instância do conjunto de dados (imagens) é representada por um ou mais vetores de características. Características são extraídas de imagens por meio de extratores, que são algoritmos de processamento de imagens que procuram identificar atributos visuais (normalmente cor, forma ou textura) e traduzi-los em descritores numéricos.

Sabe-se ainda que um sistema de recuperação por conteúdo que exiba taxas de acerto razoáveis deve, via de regra, contemplar não apenas um único tipo de atributo visual, mas ao contrário, deve combinar uma série de atributos, gerando vetores de características de alta dimensionalidade, ou seja, com um número muito elevado de características.

No contexto de reconhecimento de padrões, vetores de características de alta dimensionalidade podem suscitar uma série de problemas, a saber: a) ao lançar mão de vários extratores, é possível que muitas das características produzidas sejam correlacionadas e/ou não tenham poder discriminatório esperado; b) a relação desproporcional entre a quantidade de características e o número de instâncias pode levar a um problema conhecido como maldição da dimensionalidade [1].

Em ambas as situações, o resultado é, invariavelmente, a degradação do desempenho das técnicas desenvolvidas para o armazenamento, manipulação e análise dos dados, como também um aumento considerável do custo computacional. Diante disto, a tarefa de redução da dimensionalidade do vetor de características é uma atividade bastante desejável.

A redução da dimensionalidade tem por objetivo eliminar do conjunto de dados características que sejam correlacionadas, irrelevantes ou distorcidas. Essa redução pode ser feita por meio de extração ou seleção. No primeiro, o espaço original de características é transformado ou rotacionado, gerando "novas" características, algumas das quais possuindo maior poder discriminatório que as suas versões originais. Dentre as técnicas de extração destacam-se PCA [2] e LDA [3].

Os métodos de seleção, em contrapartida, procuram encontrar no espaço original o subconjunto de características com maior poder discriminatório, descartando aquelas que exibem pouca ou nenhuma relevância. Na literatura, existem diferentes métodos

de seleção de características [4–6], que seguem o enfoque de reconhecimento de padrões e, portanto, se baseiam em alguma análise estatística dos dados ou, em outros casos, adotam uma abordagem de redes neurais.

Neste trabalho propomos uma abordagem alternativa para a seleção de características que combina técnicas tradicionais de seleção com o auxílio da mineração visual de dados [7]. A mineração visual de dados corresponde à fusão das áreas de mineração de dados (que busca extrair padrões de conjuntos de dados) e visualização de informações (que apóia o processo de mineração por meio de métodos visuais de apresentação e interação com dados abstratos). Dentre as técnicas de mineração visual de dados, e que dão apoio à interpretação visual de informações, estão as projeções de dados multidimensionais [8].

A projeção consiste em mapear dados de um espaço m-dimensional em um espaço p-dimensional, com $p < m$, geralmente igual a 2 ou 3, buscando preservar ao máximo as relações de distância existentes entre os dados. Tradicionalmente, os pontos projetados correspondem às instâncias de dados, por exemplo, imagens. Pontos projetados próximos uns aos outros indicam agrupamentos de instâncias que compartilham as mesmas propriedades. Portanto, agrupamentos visualmente separáveis na projeção, indicam grupos de imagens semelhantes.

Na abordagem proposta, ao invés de projetarmos as imagens como instância de dados, projetam-se as características. Para isso, o conjunto original (composto por n instâncias, com m características cada) deve ser transposto. A projeção resultante consistirá, portanto, de m amostras, cada qual representando uma característica. À exemplo da projeção tradicional, agrupamentos de características poderão ser gerados. Se agrupamentos indicam instâncias com propriedades comuns, então podemos assumir que características pertencentes ao mesmo grupo possuem poder discriminatório semelhante. Ao selecionarmos apenas algumas amostras de cada agrupamento da projeção, teremos um subconjunto de características, configurando um processo de seleção.

A qualidade do subconjunto de características selecionado pelo método proposto será avaliada comparando-se as projeções obtidas para este conjunto e o conjunto original de características. Para tanto será utilizada a medida de silhueta [9], que quantifica a qualidade da projeção, medindo o grau de coesão das amostras pertencentes à mesma classe e o grau de separação entre classes.

Na Seção 2 são apresentados alguns conceitos básicos sobre seleção de características. Na Seção 3 é feita uma revisão sobre projeção de dados multidimensionais, detalhando o cálculo da medida de silhueta usada para avaliar a qualidade das projeções. A Seção 4 apresenta a metodologia proposta. Na Seção 5 são expostos os resultados experimentais, os quais são comparados com o método tradicional de seleção via *k-means*. Por fim, a Seção 6 apresenta as conclusões.

2 SELEÇÃO DE CARACTERÍSTICAS

A seleção de características é um processo importante para minimizar os problemas gerados pela alta dimensionalidade e pela utilização de conjuntos de dados que contenham características correlacionadas ou irrelevantes. Na prática, a seleção de características consiste em selecionar um subconjunto de características mais relevantes, dado o conjunto original de tamanho m. Entretanto, encontrar um subconjunto ótimo só é possível por meio de uma busca exaustiva, o que é inviável computacionalmente, na maioria dos casos. Diante disso, vários métodos utilizam heurísticas para realizar a seleção.

Tradicionalmente, os métodos de seleção de características ou seguem uma perspectiva de reconhecimento de padrões [4–6], onde alguma forma de análise estatística dos dados é conduzida, ou adotam uma abordagem de redes neurais artificiais [10]. Em ambos os casos, a qualidade dos subconjuntos selecionados é medida por meio de uma função critério.

Acredita-se que a combinação de técnicas de mineração visual, por meio de projeção de dados multidimensionais, com o processo de seleção de características seja vantajosa, pois a avaliação da seleção, anteriormente baseada apenas em uma função estatística, conta agora com o ponto de vista de um observador (usuário) que avalia a qualidade do agrupamento fornecido pela projeção para um dado subconjunto de características. Além disso, pode-se explorar a viabilidade de realizar seleção de características de forma automática, combinando projeções com processos de agrupamento.

3 PROJEÇÃO DE DADOS MULTIDIMENSIONAIS

Técnicas de projeção de dados multidimensionais buscam mapear dados m-dimensionais em um espaço p-dimensional, com $p < m$, geralmente igual a 1, 2 ou 3, preservando ao máximo as relações de distância existentes entre os dados, de forma a revelar algum tipo de similaridade ou correlação [8]. Projeções de dados possuem inúmeras vantagens, onde se destaca a possibilidade de identificar grupos de instâncias similares, os relacionamentos entre grupos que compartilham fronteiras e instâncias que pertencem a mais de um grupo. Além disso, possibilitam uma visão geral dos dados e uma visão detalhada das instâncias pertencentes a cada grupo.

Existem diferentes técnicas de projeção na literatura [11–13], as quais podem ser divididas em três grandes grupos [13]: *Multidimensional Scaling* (MDS), *Force-Directed Placement* (FDP) e Técnicas para Redução de Dimensionalidade. As técnicas de MDS realizam um mapeamento injetivo de dados do espaço m-dimensional em pontos do espaço p-dimensional, preservando alguma relação de distância. Como exemplo de técnicas, citam-se *Classical Scaling* e *Sammom's Mapping* [14], as quais foram utilizadas na geração de algumas projeções apresentadas neste artigo, visto que são capazes de preservar bem as distâncias dos dados originais nos dados projetados.

Técnicas de FDP empregam um sistema de molas para calcular a posição dos dados na projeção. Para isso, os dados sofrem ação do sistema de molas até que as forças exercidas sobre eles se anulam, estabilizando o sistema. Como exemplo de técnicas, citam-se Modelo de Molas e *Force-Scheme* [8]. Já as técnicas para Redução de Dimensionalidade buscam transformar o espaço original, de grande dimensão, em um espaço de dimensão reduzida preservando relações de distância. Como exemplo, citam-se as técnicas *Principal Component Analysis* [15] e *Projection Pursuit* [16].

Tem-se também a técnica *Least Square Projection* (LSP) [17], a qual foi utilizada para gerar algumas projeções presentes neste artigo. Esta técnica, desenvolvida mais recentemente e que apresentou bons resultados no mapeamento de documentos [17], não se encaixa perfeitamente em nenhum grupo citado pois, diferentemente das anteriores, esta busca preservar relações de vizinhança (e não de distância) nos dados projetados. Dessa forma, espera-se que dados vizinhos no espaço multidimensional sejam projetados na mesma vizinhança no plano.

Embora cada técnica de projeção adote seu próprio formalismo matemático para mapear um espaço de alta dimensão para o plano ou para o espaço, todas tem como objetivo comum gerar grupos coesos de instâncias semelhantes e que ao mesmo tempo estejam suficientemente separados. A avaliação da qualidade destes agrupamentos é muitas vezes feita de forma qualitativa, por um observador. No entanto, é possível avaliar a qualidade de forma quantitativa, por meio do cálculo da medida de **silhueta** [9]. Seja $S = x_1, x_2, ..., x_n$ um conjunto de n instâncias e, k agrupamentos não sobrepostos representados por $C = C_1, C_2, ..., C_k$. Ao considerar a instância i pertencente ao agrupamento A, a dissimilaridade média de i para todas as outras instâncias de A é dada por $a(i)$.

Considerando todos os agrupamentos $C \neq A$, a dissimilaridade média de i para todas as outras instâncias dos agrupamentos C é dada por $d(i, C)$. Dessa forma, seleciona-se o menor valor $b(i) = \min d(i, C)$, $C \neq A$, o qual representa a dissimilaridade de i para o agrupamento vizinho e calcula-se a silhueta da instância i através da seguinte equação:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

Quanto maior o valor obtido pela medida de silhueta ($-1 \leq s(i) \leq 1$), melhor é a atribuição da instância i a um dado agrupamento. A média de $s(i)$, para $i = 1, 2, ..., n$, pode ser usada como um critério de avaliação da qualidade de uma dada partição. Fazendo isso, o melhor agrupamento é encontrado quando o valor da média é maximizado.

4 METODOLOGIA

A Figura 1 apresenta a metodologia usada na nova abordagem de seleção de características proposta neste artigo.

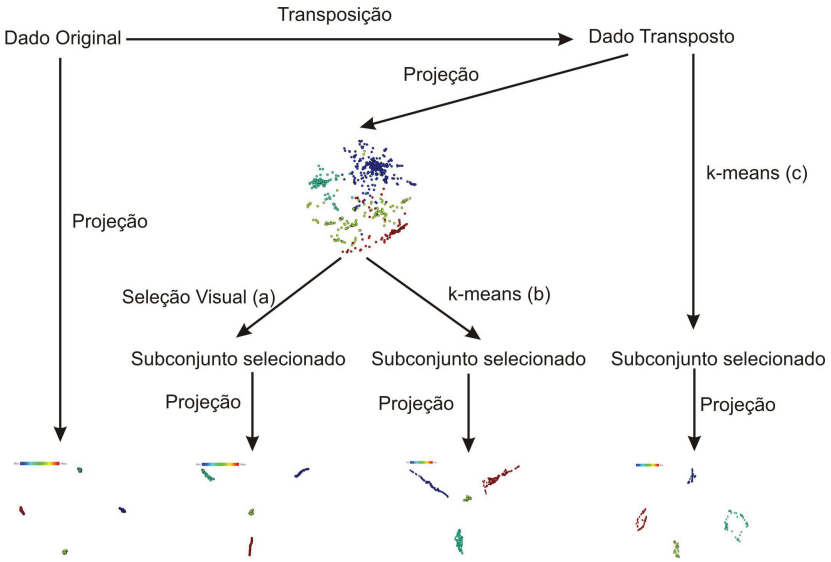


Figura 1: Metodologia. (a) Seleção Visual, com auxílio do usuário. (b) Seleção por meio da aplicação do algoritmo *k-means* sobre os dados projetados. (c) Seleção por meio da aplicação do algoritmo *k-means* diretamente no conjunto transposto.

Inicialmente, o conjunto original de dados é projetado. Este é representado por uma matriz $M_{n \times m}$, composta por n instâncias (imagens) com m características cada. Em seguida, o conjunto de dados original é transposto, obtendo-se agora uma matriz $M_{m \times n}$, ou seja, com m instâncias (características) e n dimensões (imagens). A partir deste conjunto transposto, três processos de seleção são efetuados: a) seleção visual, com auxílio do usuário; b) seleção por meio da aplicação de *k-means* sobre os dados projetados e c) seleção por meio da aplicação de *k-means* diretamente no conjunto transposto.

No primeiro processo (a), o conjunto transposto é projetado, podendo revelar agrupamentos. Cada um dos m pontos, naturalmente, representa uma característica. A partir destes agrupamentos, um subconjunto de características (uma ou mais amostras de cada agrupamento) pode ser selecionado manualmente pelo usuário.

O segundo processo (b) consiste em substituir a seleção manual de características na projeção, pela seleção automática por meio da aplicação do algoritmo *k-means* sobre as características projetadas. Esta solução é normalmente empregada quando a projeção de características não revela agrupamentos bem delineados.

No terceiro processo de seleção (c), o algoritmo de agrupamento não supervisionado *k-means* é aplicado diretamente sobre o conjunto de dados transposto, gerando grupos de características. A partir destes, seleciona-se um subconjunto de características (uma ou mais amostras de cada grupo), geralmente as mais próximas do centróide de cada agrupamento.

Os subconjuntos de características selecionados pelos três processos são, então, utilizados para gerar novas projeções no conjunto original de dados. Finalmente, as projeções obtidas podem ser comparadas tanto subjetivamente, por meio da análise visual de um observador, como por meio da medida de silhueta. Esta comparação também se dará com relação à projeção obtida com o conjunto original de características, em que não ocorreu o processo de seleção.

5 RESULTADOS

O objetivo dos estudos de casos realizados é avaliar a nova abordagem de seleção de características proposta (seleção apoiada por projeção) frente à uma seleção tradicional, realizada por meio do algoritmo de agrupamento *k-means*. Primeiramente, projeta-se o conjunto original de dados, considerando todas as características. Em seguida, realiza-se a seleção visual (apoiada pelo processo de projeção das características) e a seleção por meio do *k-means*. Os subconjuntos de características selecionados são usados para gerar novas projeções. A qualidade das projeções obtidas por meio de todas as características e por meio dos subconjuntos selecionados é avaliada pela medida de silhueta.

Neste artigo são apresentados três estudos de casos. O primeiro corresponde a um experimento "fabricado", que utiliza um mosaico de imagens formado por padrões de textura. O segundo e o terceiro utilizam um conjunto de imagens de cenas naturais. As Subseções 5.1, 5.2 e 5.3 explicam detalhadamente os três estudos de casos.

5.1 Estudo de Caso 1

O primeiro estudo de caso usou a técnica de projeção *Least Square Projection* e amostras (imagens) provenientes de um mosaico composto por padrões de textura em quatro ângulos de rotações diferentes (0°, 30°, 60° e 90°), conforme pode ser observado na Figura 2, do álbum *Rotate* de Brodatz [18]. As características foram extraídas utilizando os Filtros de Gabor [19], considerando 4 escalas, 20 orientações e 3 medidas (média, variância e energia), totalizando 240 características. Dois testes foram conduzidos: o primeiro, com 240 imagens (60 por classe), e o segundo com 784 imagens (196 por classe). O objetivo em se variar a quantidade de amostras é avaliar o método em uma situação típica de ocorrência da maldição da dimensionalidade. Para o conjunto de 784 amostras e 240 características, ao transpor os dados, teremos 240 amostras e um total de 784 características, configurando a situação. As amostras no conjunto de dados original são pré-rotuladas e cada classe é representada na projeção por uma cor distinta.

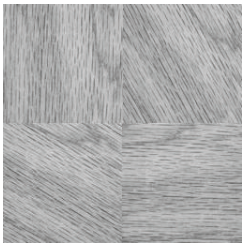
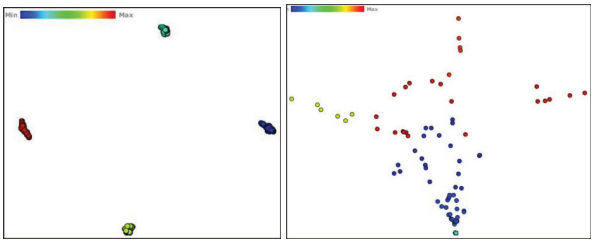


Figura 2: Mosaico de textura de madeira com 4 "classes": 0°(acima esq.); 30°(acima dir.); 60°(abaixo esq.); 90°(abaixo dir.).

A Figura 3(a) ilustra uma projeção do conjunto original de dados com 240 imagens para todas as 240 características. Percebe-se que as 4 classes (0° - azul, 30° - verde, 60° - amarelo e 90° - vermelho) foram perfeitamente separadas na projeção, visto que as características de Gabor conseguiram representar bem as orientações de cada classe. Em seguida, o conjunto original de dados foi transposto e projetado (Figura 3(b)). Entretanto, a projeção não revelou agrupamentos perfeitamente delineados. Ao invés de se aplicar a seleção manual, foi empregado o algoritmo *k-means* (para 4 classes) sobre os dados projetados. Os resultados desta seleção automática podem ser visto na Figura 3(b). Para cada um dos 4 agrupamentos, 3 amostras de cada grupo foram aleatoriamente tomadas, totalizando 12 características. O processo aleatório pode ser facilmente substituído por um esquema automático em que as x amostras mais próximas ao centróide de cada grupo sejam tomadas.

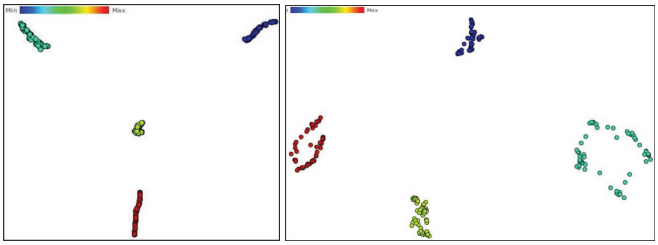


(a) (b)

Figura 3: Projeção das 240 amostras. (a) Projeção do conjunto original de dados. (b) Projeção dos dados transpostos com os 4 agrupamentos computados por *k-means* (cada ponto indica uma característica).

Para avaliar o resultado da seleção por meio de projeções, aplicou-se o algoritmo *k-means* diretamente sobre o conjunto transposto. Para fins de comparação, foram considerados quatro agrupamentos e tomadas 3 amostras de cada agrupamento (as

mais próximas do centróide de cada agrupamento). As Figuras 4(a) e (b) ilustram as projeções obtidas das 240 imagens para *k-means* aplicado sobre a projeção e *k-means* aplicado diretamente sobre o conjunto transposto, respectivamente. As medidas de silhueta obtidas nas projeções são apresentadas na Tabela 1. Os valores entre parênteses na tabela indicam a quantidade de características empregadas em cada projeção.



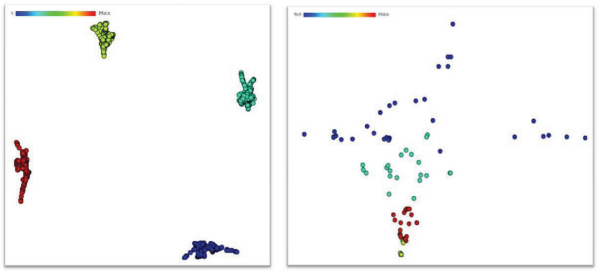
(a) (b)

Figura 4: Projeção das 240 imagens. (a) Subconjunto de características selecionado pela aplicação do *k-means* sobre os dados projetados. (b) Subconjunto de características selecionado pela aplicação do *k-means* diretamente no conjunto transposto.

Tabela 1: Medidas de silhueta obtidas das projeções para o conjunto de 240 imagens do mosaico de textura.

Método de Seleção	Silhueta
Sem seleção (240)	0.9693
Seleção baseada em projeção (12)	0.8952
Seleção por meio do *k-means* (12)	0.8472

O segundo teste empregou as mesmas características e procedimentos do teste anterior. No entanto, a quantidade de imagens foi de 784. A Figura 5(a) ilustra a projeção do conjunto original de dados com 784 imagens para todas as 240 características. A Figura 8(b) ilustra a projeção do conjunto transposto. Entretanto, esta projeção também não revelou agrupamentos perfeitamente delineados, sendo necessária a aplicação do algoritmo de *k-means* (para 4 classes) sobre os dados projetados. O resultado desta seleção automática pode ser visto na Figura 5(b). Para cada um dos 4 agrupamentos, 3 amostras foram tomadas, totalizando 12 características.



(a) (b)

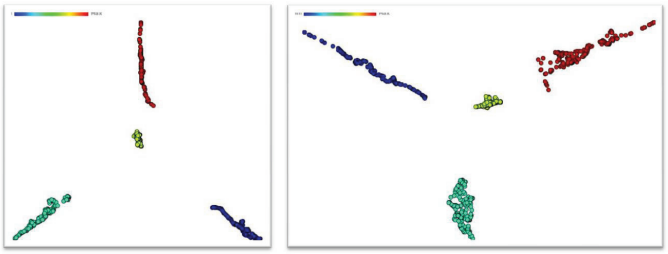
Figura 5: Projeção das 784 amostras. (a) Projeção do conjunto original de dados. (b) Projeção dos dados transpostos com os 4 agrupamentos computados por *k-means* (cada ponto indica uma característica).

Para fins de comparação, aplicou-se o *k-means* diretamente no conjunto transposto, obtendo um subconjunto com 12 características. A Figura 6(a) e (b) ilustram as projeções obtidas das 784 imagens para o método proposto e *k-means* diretamente sobre o conjunto transposto, respectivamente. As medidas de silhueta obtidas nas projeções são apresentadas na Tabela 2.

Tabela 2: Medidas de silhueta obtidas das projeções para o conjunto de 748 imagens do mosaico de textura.

Método de Seleção	Silhueta
Sem seleção (240)	0.9235
Seleção baseada em projeção (12)	0.8312
Seleção por meio do *k-means* (12)	0.7661

Ao observar os resultados dos dois testes realizados percebe-se que as projeções das imagens apresentam as quatro classes distintas (0°, 30°, 60° e 90°) separadamente, ou seja, a qualidade visual das projeções se mantém após a seleção, com consi-



(a) (b)

Figura 6: Projeção das 784 imagens. (a) Subconjunto de características selecionado pela aplicação do *k-means* sobre os dados projetados. (b) Subconjunto de características selecionado pela aplicação do *k-means* diretamente no conjunto transposto.

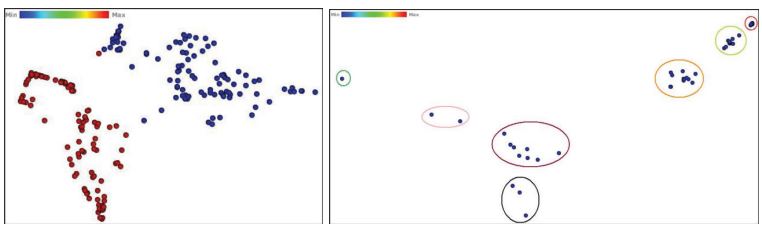
derável redução do número de características (95% das características foram eliminadas). Isto é comprovado ao verificar que a variação do valor de silhueta nas projeções é pequena. Comparando-se os dois testes, observa-se que o valor de silhueta obtido nas projeções do segundo teste diminuiu (em torno de 6 %), demonstrando que a maldição da dimensionalidade exerce alguma influência nos resultados obtidos. No entanto, não chega a causar prejuízos na distinção das classes nas projeções, pois a qualidade visual se mantém.

5.2 Estudo de Caso 2

Um segundo estudo de caso foi realizado usando a técnica de projeção *Least Square Projection* e o banco de imagens de cenas naturais *Scene*, composto por 200 imagens, sendo 100 imagens de construções e 100 de costas oceânicas, conforme ilustra a Figura 7. As características das imagens foram extraídas por meio dos Filtros de Gabor, para 4 escalas, 12 orientações e apenas a função média, totalizando 48 características.

Figura 7: Amostras do banco de imagens de cenas naturais: construções (esq.) e costas oceânicas (dir.).

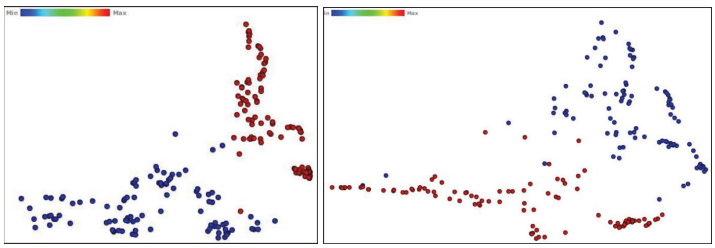
A Figura 8(a) ilustra a projeção do conjunto original de dados com 200 imagens para todas as características, onde pontos azuis representam imagens de construções e pontos vermelhos representam imagens de costas oceânicas. Em seguida, foi realizado o processo de seleção: o conjunto de dados foi transposto e projetado, revelando sete agrupamentos, conforme ilustra a Figura 8(b). O usuário, então, selecionou aleatoriamente uma amostra de cada agrupamento, compondo um subconjunto com 7 características. Para fins de comparação, aplicou-se o algoritmo *k-means* sobre o conjunto transposto não projetado, para 7 agrupamentos, tomando uma única amostra de cada agrupamento (a mais próxima do centróide). A Figura 9 ilustra as projeções das 200 imagens para o subconjunto gerado pela seleção visual e o obtido pelo método de seleção por *k-means*. As medidas de silhueta obtidas nas projeções são apresentadas na Tabela 3.



(a) (b)

Figura 8: Projeção das 200 amostras. (a) Projeção do conjunto original de dados. (b) Projeção dos dados transpostos (cada ponto representa uma característica).

Observando os valores de silhueta obtidos nas projeções percebe-se que ocorre uma pequena variação, ressaltando que houve um aumento no valor da silhueta obtido na projeção que utiliza apenas cerca de 15% das características selecionadas por meio visual. Além disso, a qualidade visual das projeções obtidas se mantém, pois as imagens de classes diferentes permanecem separadas.



(a) (b)

Figura 9: Projeções das 200 imagens. (a) Subconjunto de características selecionado visualmente. (b) Subconjunto de características selecionado pela aplicação do *k-means* diretamente no conjunto transposto.

Tabela 3: Medidas de silhueta obtidas das projeções para o conjunto de 200 imagens de cenas naturais.

Método de Seleção	Silhueta
Sem seleção (48)	0.5353
Seleção baseada em projeção (7)	0.5373
Seleção por meio do *k-means* (7)	0.4031

5.3 Estudo de Caso 3

O terceiro estudo de caso realizado seguiu um procedimento diferente dos anteriores, pois teve como objetivo principal comparar os resultados do método de seleção visual, os quais foram obtidos utilizando a técnica *Least Square Projection* com os resultados obtidos utilizando as técnicas *Classical Scaling* e *Sammon's Mapping*. Estas técnicas foram escolhidas devido à sua capacidade de preservar a distância dos dados originais nos dados projetados.

As amostras utilizadas são provenientes do banco de imagens *Scene*, composto por 200 imagens, sendo 100 imagens de construções e 100 de costas oceânicas. As características foram extraídas por meio dos Filtros de Gabor, considerando 4 escalas, 12 orientações e 3 medidas (média, variância e energia), resultando em 144 características. Além dos Filtros de Gabor, foram extraídas as 6 características (*coarseness*, *contrast*, *directionality*, *linelikeness*, *regularity* e *roughness*) de Tamura [20]. No total, cada imagem foi representada por um vetor de 150 características.

A primeira técnica de projeção utilizada foi a *Least Square Projection*. A Figura 10 ilustra uma projeção das 200 imagens para todas as características. Em seguida, foi realizado o processo de seleção proposto: o conjunto de dados foi transposto e projetado, revelando alguns agrupamentos, conforme ilustra a Figura 11(a). O usuário, então, selecionou aleatoriamente amostras de cada agrupamento, compondo um subconjunto com 30 características. Por fim, a Figura 11(b) ilustra a projeção das 200 imagens para o subconjunto gerado pela seleção visual.

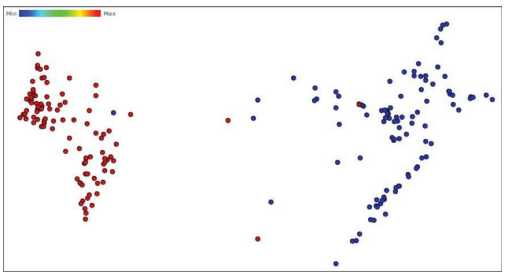
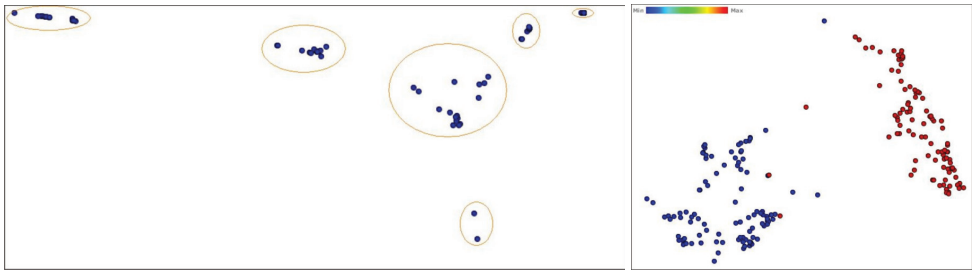


Figura 10: Projeção das 200 imagens para todas as características utilizando a técnica *Least Square Projection*.

Utilizando a técnica *Classical Scaling*, primeiramente projetamos as 200 imagens para o conjunto original de características, conforme apresentado na Figura 12. Em seguida, este conjunto foi transposto e projetado, revelando alguns agrupamentos, conforme pode ser observado na Figura 13(a), permitindo que o usuário selecionasse 30 amostras. A projeção das 200 imagens considerando apenas o subconjunto selecionado é apresentada na Figura 13(b).

Por fim, as projeções foram geradas utilizando a técnica *Sammon's Mapping*. A projeção das 200 imagens considerando todas as características é apresentada na Figura 14. Em seguida, este conjunto foi transposto e projetado, conforme pode ser visto na Figura 15(a). Esta projeção revelou alguns agrupamentos de características, dos quais o usuário selecionou 30 amostras. A Figura 15(b) ilustra a projeção das 200 imagens considerando apenas o subconjunto de características selecionado.

Observando visualmente as projeções é possível verificar variações de acordo com a técnica utilizada. Estas variações são confirmadas pelas medidas de silhueta obtidas, conforme pode ser visto na Tabela 4, onde os valores entre parênteses indicam a quantidade de características empregada em cada projeção. Percebe-se que a técnica *Least Square Projection* projeta as duas



(a) (b)

Figura 11: Técnica de Projeção: *Least Square Projection*. (a) Projeção dos dados transpostos. Cada ponto representa uma característica. (b) Projeção dos dados para o subconjunto de 30 características.

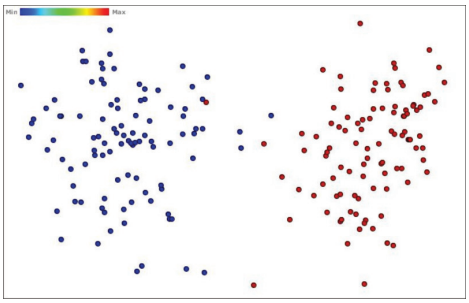
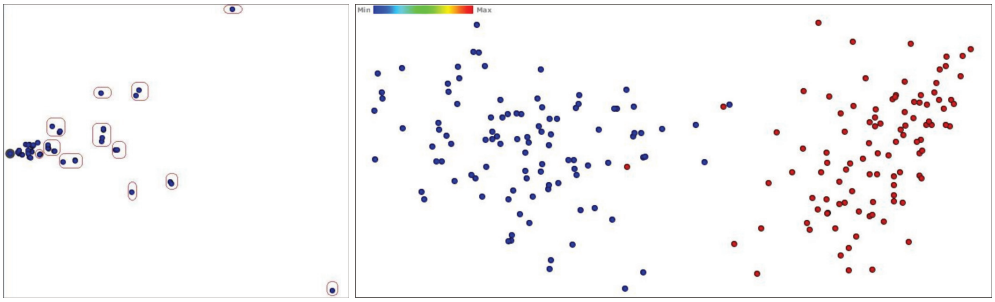


Figura 12: Projeção das 200 imagens para todas as características utilizando a técnica *Classical Scaling*.



(a) (b)

Figura 13: Técnica de Projeção: *Classical Scaling*. (a) Projeção dos dados transpostos. Cada ponto representa uma característica. (b) Projeção dos dados para o subconjunto de 30 características.

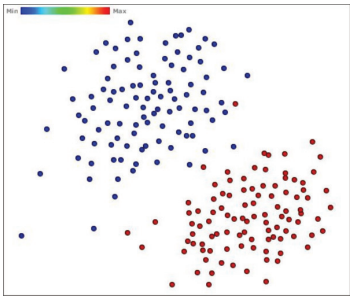
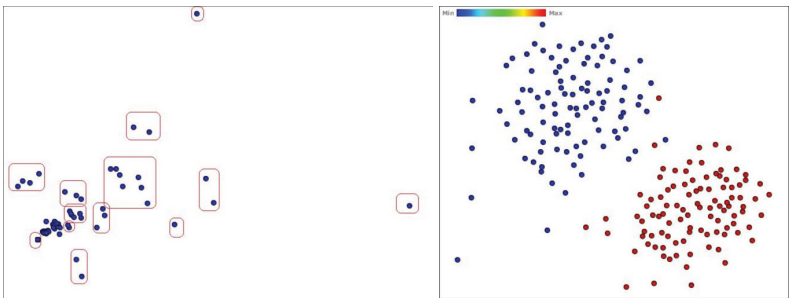


Figura 14: Projeção das 200 imagens para todas as características utilizando a técnica *Sammon's Mapping*.

classes de imagens com maior grau de separação e mantém os elementos de mesma classe mais coesos, o que é comprovado pelo maior valor de silhueta obtido (0.74) em relação às outras técnicas utilizadas. No entanto, quando as técnicas *Classical Scaling* e *Sammon's Mapping* são usadas, o subconjunto de características selecionado, à partir dos agrupamentos obtidos na projeção, consegue gerar uma nova projeção com valor de silhueta maior que àquele obtido quando se utiliza o conjunto completo de características. Além disso, pode-se perceber que o valor de silhueta obtido pelo subconjunto selecionado via *Classical Scaling* é equivalente àquele obtido pelo subconjunto selecionado via *Least Square Projection*. Isto pode ser explicado pelo fato da técnica *Classical Scaling* melhor preservar a distância dos dados originais nos dados projetados. Assim, a projeção resulta em agrupamentos de características com poder discriminatório semelhante, permitindo a seleção de um subconjunto mais



(a) (b)

Figura 15: Técnica de Projeção: *Sammon's Mapping*. (a) Projeção dos dados transpostos. Cada ponto representa uma característica. (b) Projeção dos dados para o subconjunto de 30 características.

representativo.

Tabela 4: Medidas de silhueta obtidas das projeções para o conjunto de 200 imagens de cenas naturais, considerando diferentes técnicas.

Técnica de Projeção	Conjunto Original (150)	Conjunto Selecionado (30)
Least Square Projection	0.74	0.68
Classical Scaling	0.64	0.68
Sammon's Mapping	0.54	0.58

6 CONCLUSÕES

Este trabalho apresenta uma nova abordagem para a seleção de características, baseada no processo de projeção de dados multidimensionais. Esta abordagem foi comparada com um método de seleção tradicional, que usa o algoritmo de agrupamento *k-means* para a obtenção do subconjunto de características. O conjunto total de características e os subconjuntos selecionados são usados no cálculo de projeções de imagens, as quais são avaliadas visualmente e por meio da medida de silhueta.

Os experimentos realizados mostraram resultados interessantes, visto que as projeções obtidas, após o processo de seleção de características, mantiveram qualidade visual semelhante às projeções que usam o conjunto total de características. Isto é importante pois a redução do número de características é sempre bastante significativa. Os resultados visuais são confirmados quantitativamente por meio do cálculo dos valores de silhueta obtidos nas projeções, os quais sofrem pequenas variações e, em alguns casos, aumentam após a seleção.

Ressalta-se que nem sempre é possível visualizar o melhor valor de k (número de grupos de características) na projeção de características, dificultando a seleção de amostras. Diante disso, pode-se aplicar um algoritmo de agrupamento como, por exemplo, o *k-means*, sobre as características projetadas (conforme exemplificado no estudo de caso 1) ou diretamente no conjunto de dados transposto. Ao aplicar o algoritmo de agrupamento sobre o conjunto de características, pode-se considerar diferentes valores de k, o que permite selecionar diferentes subconjuntos de características. Estes subconjuntos são usados no cálculo de novas projeções e seus respectivos valores de silhueta. Assim, o melhor subconjunto será aquele cuja projeção produz o maior valor de silhueta.

Também foi realizada uma comparação dos resultados obtidos considerando três técnicas de projeção diferentes: *Least Square Projection*, *Classical Scaling* e *Sammon's Mapping*. Com isso, pôde-se perceber variações visuais nas projeções, as quais são confirmadas pelos diferentes valores de silhueta obtidos. A técnica *Least Square Projection* é a que melhor separa as diferentes classes na projeção, mantendo os elementos de mesma classe mais coesos. No entanto, as técnicas *Classical Scaling* e *Sammon's Mapping* preservam melhor a distância dos dados originais nos dados projetados, permitindo que os agrupamentos obtidos possuam características com poder discriminatório semelhante. Assim, a seleção de algumas amostras de cada grupo de característica consegue gerar projeções tão boas quanto àquela obtida pela técnica *Least Square Projection*. Além disso, os subconjuntos selecionados geram projeções com valor de silhueta maior que as projeções com o conjunto total de características.

Em trabalhos futuros pretende-se avaliar o método de seleção apoiado por projeção frente à um método de seleção quantitativo baseado em Redes Neurais Artificiais com Saliência [10]. Além disso, os grupos de características obtidos nas projeções serão usados para determinar a quantidade de neurônios da camada oculta da rede neural (quantidade de agrupamentos será igual a quantidade de neurônios) e também influenciarão no cálculo da saliência de cada característica, o qual vai depender apenas das características pertencentes ao mesmo grupo e não mais do conjunto completo de características. Outro tema importante a ser tratado é a obtenção do melhor valor de k automaticamente, o qual corresponderá ao melhor subconjunto de características selecionado.

REFERÊNCIAS

[1] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[2] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[3] R. Fisher. "The statistical utilization of multiple measurements". *In Annals of Egenics*, vol. 8, pp. 376–386, 1938.

[4] D. Jain, A.; Zongker. "Feature Selection: Evaluation, Application and Small Sample Performance". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[5] J. Kudo, M; Sklansky. "Comparison of algorithms that select features for pattern classifiers". *Pattern Recognition*, vol. 33, no. 1, pp. 25–41, 2000.

[6] L. Liu, H; Yu. "Toward integrating feature selection algorithms for classification and clustering". *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[7] P. Wong. "Guest editor's introduction: Visual data mining". *IEEE Compuer Graphics and Applications*, vol. 19, pp. 20–21, 1999.

[8] R. N. L. Tejada, E.; Minghim. "On Improved projection techniques to support visual exploration of multidimensional data sets." *Information Visualization*, vol. 2, no. 4, pp. 218–231, 2003.

[9] P. J. Kaufman, L.; Rousseeuw. *Finding Groups in Data - An Introduction to Cluster Analysis*. Wiley Series in Propability and Mathematical Statistics, 1990.

[10] J. Santos, D.P.; Neto. "Feature Selection with Equalized Salience Measures and its Application to Segmentation". *In SIBGRAPI '07: Proceedings of the Brazilian Symposium on Computer Graphics and Image Processing*, pp. 253–262, 2007.

[11] M. Cox, T.F.; Cox. *Multidimensional Scaling*. Chapman e Hall/CRC, second ed. edition, 2000.

[12] I. Fodor. "A Survey of Dimension Reduction Techniques". *Center for Applied Scientifc Computing, Lawrence Livermore National Laboratory*, 2002. Relatório Técnico.

[13] F. V. Paulovich. "Mapeamento de Dados multi-dimensionais - integrando mineração e visualização". Ph.D. thesis, Universidade de São Paulo, 2008.

[14] J. Sammon. "A nonlinear mapping for data structure analysis". *IEEE Transactions on Computer*, vol. 18, no. 5, pp. 401–409, 1969.

[15] I. Jolliffe. "Principal Component Analysis". *Springer-Verlag*, 1986.

[16] J. W. Friedman, J. H.; Tukey. "A Projection Pursuit Algorithm for Exploratory Data Analysis". *IEEE Transactions on Computers*, vol. C-23, no. 9, pp. 881–890, 1974.

[17] L. M. R. H. Paulovich, F. V.; Nonato. "Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping". *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 3, 2008.

[18] P. Brodatz. *Textures: A Photografic Album for Artists and Designers*. Dover Publications, New York, 1966. http://sipi.usc.edu/database/database.cgi?volume=rotate.

[19] W. Y. Manjunath, B. S.; Ma. "Texture Features for Browsing and Retrieval of Image Data". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[20] S. Y. T. Tamura, H.; Mori. "Textural Features Corresponding to Visual Perception". *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, no. 6, pp. 460–473, 1978.