Production, xx(x), xx-xx
http://dx.doi.org/10.1590/0103-6513.143613

PRODUCTION
PRODUÇÃO

Seleção de variáveis para clusterização de bateladas produtivas através de ACP e remapeamento *kernel*

Victor Leonardo Cervo[a], Michel José Anzanello[a*]

[a*]Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brasil, michel.anzanello@gmail.com

Resumo

Técnicas de clusterização visam à formação de grupos de observações homogêneas dentro de um mesmo grupo e significativamente distintas das observações inseridas em outros grupos. Em processos industriais cuja produção é apoiada em bateladas, a definição de famílias (grupos) de bateladas com perfis semelhantes auxilia na definição de estratégias de controle e monitoramento desses processos. Este artigo propõe um método para seleção das variáveis de clusterização mais relevantes para formação de famílias de bateladas. Para tanto, integra funções *kernel* a um novo índice de importância de variáveis gerado a partir dos parâmetros oriundos da Análise de Componentes Principais (ACP). A qualidade dos agrupamentos formados é avaliada através do Silhouette Index (SI). Quando aplicada em três processos produtivos, a sistemática proposta reteve em média 5,16% das variáveis iniciais e elevou o SI médio em 235,4% frente à utilização de todas as variáveis. Um estudo de simulação também é realizado para avaliar a robustez do método.

Palavras-chave

Análise de clusterização. Seleção de variáveis. *Kernel*. Processos em batelada.

1. Introdução

Diversas áreas da ciência apoiam-se na coleta e análise de um grande número de variáveis e observações. Tais áreas incluem setores médicos, com o intuito de prever diagnósticos (Wolberg et al., 1994; Detrano et al., 1989), áreas sociais, a fim de estudar aspectos demográficos (Meek et al., 2002), e áreas de gestão de produção, com o propósito de facilitar a programação de manufatura e controle de qualidade. Nessa última área, o agrupamento de bateladas de produção com perfis similares permite que elas sejam encaminhadas a destinos próprios (por exemplo, reprocesso ou clientes específicos), reduzindo esforços e custos de coleta de dados e propiciando maior agilidade nas tomadas de decisão acerca dos grupos (e não das observações individuais), entre outros benefícios. Dessa forma, a formação de famílias de bateladas com características semelhantes entre si permite que conclusões acerca de uma batelada possam ser estendidas às demais bateladas daquele grupo (ou *cluster*).

Bateladas produtivas normalmente são agrupadas com base em variáveis descritivas de processo, as quais podem incluir centenas ou milhares de variáveis. Diversos estudos sugerem que agrupamentos mais consistentes são obtidos quando um conjunto reduzido de variáveis é utilizado, visto que a inserção de variáveis ruidosas tende a degradar significativamente a qualidade do procedimento (Brusco & Cradit, 2001; Milligan, 1980; Li et al., 2008; Maugis et al., 2009). Friedman & Meulman (2004) e Huang et al. (2005) sugerem sistemáticas de seleção baseadas na atribuição de pesos para identificar o subgrupo de variáveis mais relevantes para clusterização, enquanto que Gnanadesikan et al. (1995) e Li et al. (2008) defendem que variáveis irrelevantes devem ser removidas do banco de dados e não ponderadas. Brusco (2004) afirma que a seleção de variáveis elimina por completo o efeito indesejável das variáveis que mascaram a definição das estruturas dos *clusters*. Outras abordagens representativas para seleção de variáveis de clusterização são apresentadas por Raftery & Dean (2006), Bouveyron et al. (2007), Steinley & Brusco (2008b), Dean & Raftery (2010) e Bessaoud et al. (2012).

Recebido: Maio 8, 2013; Aceito: Fev. 7, 2014

Este artigo propõe um método para seleção das variáveis de clusterização mais relevantes para formação de famílias de bateladas com características similares. Para tanto, integra funções *kernel* a um novo índice de importância de variáveis gerado a partir dos parâmetros oriundos da Análise de Componentes Principais (ACP). Nas proposições deste artigo, as funções *kernel* remapeiam os dados originais em um novo espaço com vistas à formação de *clusters* mais consistentes. Na sequência, um índice de importância das variáveis de clusterização é gerado com base nos parâmetros oriundos da ACP aplicada sobre os dados remapeados; as variáveis com índices de importância maior são tidas como mais relevantes (Duda et al., 2001). Um procedimento iterativo de clusterização é iniciado valendo-se da variável mais relevante e a qualidade do agrupamento é avaliada através do Silhouette Index (SI). A segunda variável com maior índice de importância é então inserida no subconjunto de variáveis e a clusterização é novamente executada. Esse processo é repetido até que todas as variáveis sejam inseridas no subconjunto de variáveis de clusterização, sendo o SI médio recalculado a cada nova clusterização (gerando um perfil de qualidade de clusterização com a inserção das variáveis no procedimento). A sistemática é então repetida para um diferente número de *clusters*, permitindo identificar o melhor número de agrupamentos a ser formado e as variáveis a serem consideradas.

Ao ser aplicado em três bancos de dados industriais, o método proposto reteve em média 5,16% das variáveis iniciais e elevou o SI médio em 235,4% frente à utilização de todas as variáveis. Dos três bancos testados em quatro cenários (distinto número de *clusters*), a utilização dos *kernels* mostrou-se superior em nove das 12 possibilidades. Um estudo de simulação também é realizado para avaliar a robustez do método.

O presente artigo está organizado como segue: além desta introdução, a seção 2 apresenta os fundamentos acerca de seleção de variáveis para clusterização e funções *kernel*; a seção 3 apresenta o método proposto, enquanto a seção 4 reporta os resultados numéricos. A seção 5 apresenta os resultados de um experimento de simulação e, por fim, as conclusões e os direcionamentos futuros são apresentados na seção 6.

2. Fundamentação teórica

2.1. *Análise de clusterização*

Técnicas de clusterização inserem observações em grupos (*clusters*), de forma que as similaridades sejam grandes entre observações pertencentes a um mesmo *cluster* e diferentes das inseridas em outro (Hair et al., 1995; Agard & Penz, 2009). Existem dois conjuntos de técnicas usualmente utilizados para clusterização de observações: hierárquicos e não hierárquicos (Hair et al., 1995). Dentre os não hierárquicos, foco deste estudo, destaca-se o *k-means*, o qual agrupa as observações em *k clusters*, sendo esse um valor previamente conhecido para o algoritmo (Kaufman & Rousseeuw, 2005). Matematicamente, as observações são alocadas a um determinado *cluster* de forma a minimizar a soma das distâncias euclidianas entre as observações dentro de um *cluster* e o centroide desse *cluster*.

A avaliação do desempenho da clusterização pode ser realizada através do Silhouette Index (SI), o qual avalia o quanto uma observação é semelhante às demais observações inseridas em seu *cluster*, comparado-a com observações inseridas em outros *clusters* (Kaufman & Rousseeuw, 2005). Cada observação apresenta um SI_n no intervalo [–1; 1], onde n representa a observação avaliada, $n = 1, ..., N$. Valores de SI_n próximos a 1 indicam que a distância, ou dissimilaridade, entre a observação e outras observações alocadas em outros *clusters* é pequena; assim, considera-se que a observação foi corretamente alocada ao *cluster* atual. Valores próximos a –1 indicam que a observação foi alocada a um *cluster* inadequado. Valores intermediários (próximos a 0) denotam observações que não pertencem claramente a um *cluster* ou outro. O SI_n é calculado de acordo com a Equação 1 (Rousseeuw, 1987):

$$SI_n = \frac{b(n) - a(n)}{\max\{b(n), a(n)\}} \quad (1)$$

onde $a(n)$ é a média das distâncias da n-ésima observação a todas as outras pertencentes ao mesmo *cluster* e $b(n)$ é a média das distâncias dessa n-ésima observação a todas as outras alocadas no *cluster* mais próximo. Por ser uma medida baseada em distâncias, o SI independe da técnica utilizada na clusterização, podendo ser utilizada para medir a qualidade global do procedimento de clusterização gerado por qualquer técnica (Kaufman & Rousseeuw, 2005).

2.2. *Análise de componentes principais*

Anderson (2003) define componentes principais como combinações lineares das variáveis originais, as quais são convertidas em um novo sistema de coordenadas ortogonais. Complementarmente, Jolliffe (2002) considera a ACP como um método de redução de dimensionalidade de um conjunto de dados, explicando a maior parte da variabilidade do sistema com base em um número reduzido de combinações

lineares. Os fundamentos matemáticos da ACP serão agora apresentados.

Considere x um vetor de *P* variáveis. O primeiro componente principal é definido como $\alpha_1^T x$ tal que os elementos em x apresentem máxima variância, onde $\alpha_1^T = [\alpha_{11}\ \alpha_{12} \ldots \alpha_{1p}]$ são referidos como pesos. O segundo componente é definido como $\alpha_2^T x$, não correlacionado com $\alpha_1^T x$, e com os elementos de x tendo a máxima variância possível. Os vetores α_j são autovetores da matriz Σ, tida como a matriz de variâncias e covariâncias de x. Por fim, impõe-se à formulação de maximização de variância entre os componentes a restrição $\alpha_j^T \alpha_j = 1$, forçando o comprimento unitário nos autovetores. Nessa notação, cada autovetor α_j está relacionado com λ_j, o *j*-ésimo maior autovalor da matriz Σ. O problema resume-se a maximizar a variância de $\alpha_1^T x = \alpha_1^T \Sigma \alpha_1$, sujeito à restrição $\alpha_1^T \alpha_1 = 1$.

A ACP gera dois parâmetros relevantes para as proposições deste artigo: (i) os componentes principais (autovetores de Σ), representados pelos coeficientes (pesos) das variáveis em cada um dos componentes gerados; e (ii) os autovalores de Σ, λ_k, representando a variância explicada por cada componente retido.

2.3. *Funções kernel*

Funções *kernel* constituem-se em uma classe de algoritmos de análise de padrões que se apoiam no mapeamento dos dados em um espaço de características de alta dimensionalidade. Nesse espaço, cada coordenada corresponde a uma característica dos itens analisados, sendo que uma variedade de métodos pode ser usada para encontrar as relações nos dados. O remapeamento dos dados em um novo espaço permite que relações não evidentes nos dados originais sejam salientadas, possibilitando a identificação de padrões implícitos nos dados (Schölkopf & Smola, 2002).

Segundo Huang et al. (2006), a escolha apropriada da função *kernel* constitui-se num ponto importante de análise, sendo o desenvolvimento de novos *kernels* um tópico de pesquisa recente. Abe (2010) e Schölkopf & Smola (2002) destacam as funções polinomiais, Gaussianas e Sigmóides, como os tipos de *kernel* mais usuais.

A fundamentação matemática das funções *kernel* é agora brevemente introduzida; maiores detalhes podem ser obtidos em Schölkopf & Smola (2002). Considere *N* vetores de observações; sejam x_i e x_j dois vetores desse conjunto. Define-se a matriz quadrada $Z_{ij} = z(x_i, x_j)$, de ordem (*N* x *N*), cujas entradas representam produtos internos entre as observações definidos por uma função *kernel*. Essa matriz é denominada Matriz *Kernel*. A função $z(x_i, x_j)$ gera uma matriz positiva definida; dessa forma, se *z* é definido positivo, existe um mapa Φ, onde $z(x_i, x_j) = [\Phi(x_i), \Phi(x_j)]$ (Schölkopf & Smola, 2002). Porém, para calcular o produto interno no espaço característico, pode-se usar uma função *kernel* sem explicitamente utilizar a função de mapeamento Φ(x) (Park et al., 2006). Para tanto, é necessário definir o tipo de função $Z(x_i, x_j)$ que admite a representação do produto interno no espaço característico. Para isso, considere-se o Teorema de Mercer (Girolami, 2002; Filippone et al., 2008): seja Z uma função simétrica, isto é, $\forall x_i, x_j \in X, X \subseteq R$:

$$Z(x_i, x_j) = \Phi(x_i).\Phi(x_j) \quad (2)$$

onde $\Phi(x): X \rightarrow F$ é uma função de mapeamento não linear dentro do espaço característico F. Neste artigo, a função $z(x_i, x_j) = (x_i, x_j)^d$ é utilizada para substituir o produto interno a ser utilizado no algoritmo de clusterização *k-means*.

Diversos estudos têm avaliado a utilização de *kernels* e mapeamento para espaços de atributos. Filippone et al. (2008) investigam a adaptação de métodos de particionamento, como o *k-means*, através da utilização de *kernels*; os resultados obtidos mostram que a aplicabilidade dos métodos ainda é problemática, tendo em vista o alto custo computacional. Domenicone et al. (2011) estudam técnicas de otimização para os parâmetros da função *kernel* em um contexto de clusterização semissupervisionada, guiada através de restrições dos tipos deve-ligar e não-pode-ligar. Por sua vez, Baghshah & Shouraki (2011) propõem um método de aprendizado com métrica não linear que obtém as funções de *kernel* a partir das restrições para clusterização e da topologia dos dados; os resultados sugerem que o método proposto tem potencial de uso, resolvendo o problema de otimização de forma mais eficiente que outros métodos existentes.

3. Método

O método proposto para seleção de variáveis para clusterização é operacionalizado em seis passos: (1) remapear os dados via funções *kernel*; (2) aplicar a ACP aos dados e gerar um índice de importância das variáveis, utilizando as informações fornecidas pela ACP; (3) definir um número limite de *clusters* a serem formados; (4) para cada número de *clusters* em 3, incluir as variáveis apontadas como mais relevantes pelo índice, seguindo uma sistemática do tipo *forward*, realizar a clusterização utilizando as variáveis selecionadas e avaliar seu desempenho através do Silhouette Index (SI); (5) retornar ao passo 4, alterando o número de *clusters*; e (6) identificar o

número de *clusters* e as variáveis que conduzem ao SI máximo. Esses passos são detalhados na sequência.

Passo 1 – Remapeamento dos dados através de funções kernel

Inicialmente, os dados devem ser normalizados para evitar que as distintas magnitudes das variáveis afetem as distâncias euclidianas utilizadas pela ferramenta de clusterização. Na sequência, os dados normalizados são remapeados através de uma função *kernel*, a qual tem por objetivo inserir no contexto de análise relações não lineares (utilizando para isso um *kernel* polinomial). Espera-se que a nova representação dos dados promova a formação de grupos mais precisos quando comparados à clusterização utilizando os dados originais.

Para efeitos de avaliação, este artigo utiliza 2 *kernels*: X^3 e $X^{1/3}$. A transformação com expoente fracionário com denominador ímpar (nesse caso, equivalente à raiz cúbica) é justificada pelo fato de algumas variáveis originais apresentarem sinal negativo (oriundos de eventuais procedimentos de normalização). Caso fosse escolhida uma potência fracionária com denominador par (1/2, por exemplo), os novos dados poderiam se tornar números complexos, descaracterizando o banco original. A opção por *kernels* do tipo polinomial é justificada por sua simplicidade matemática, facilidade para implementação computacional e bons resultados obtidos em aplicações práticas (Filippone et al., 2008).

Passo 2 – Aplicação da ACP nos dados remapeados e geração do índice de importância das variáveis

Na sequência aplica-se a ACP aos dados e calcula-se o Índice de Importância da Variável *p*, IIV_p, para as *P* variáveis. Esse índice leva em consideração o peso α_{jp} da variável em cada um dos *j* componentes principais e a variância explicada por cada um desses *j* componentes (autovalores λ_j), conforme a Equação 3.

$$IIV_p = \Sigma_{j=1}^{J} |\alpha_{jp}| . \lambda_j \quad (3)$$

Quanto maior o valor do índice, mais importante é considerada a variável para explicação da variabilidade nos dados. Segundo Steinley & Brusco (2008a) e Anzanello & Fogliatto (2011), uma maior variância sugere variáveis mais dispersas e, em consequência, com maior capacidade de diferenciarem observações em grupos quando comparadas a variáveis com menores variâncias. O número de componentes a ser retido pode ser definido através do Scree Graph ou de validação cruzada (Duda et al., 2001).

Uma vez calculado o *IIV* para as variáveis, elas são ordenadas de forma decrescente em função do valor do índice.

Passo 3 – Definição do intervalo de variação do número de clusters (k)

A escolha do intervalo de variação de *k*, número de *clusters*, é um passo importante do método. É bastante lógico que devam existir ao menos dois grupos distintos entre as observações, caso contrário considera-se não haver diferenças significativas entre essas observações; assim, o limite inferior do intervalo de *clusters* será dois. O limite superior é definido por especialistas de acordo com o conhecimento do sistema que está sendo agrupado.

Passo 4 – Inclusão das variáveis relevantes, clusterização das observações e avaliação do SI

O procedimento adotado para a clusterização é não hierárquico, do tipo *k-means*. Os dados remapeados são normalizados para minimizar efeitos de escala das variáveis no cálculo das distâncias euclidianas (Milligan & Cooper, 1988; Steinley, 2004; Anzanello & Fogliatto, 2011).

O subconjunto inicial de variáveis a ser testado parte da variável com o maior valor de *IIV*. Concluída a clusterização, avalia-se a qualidade do agrupamento gerado através do SI médio de todas as observações, conforme a Equação 1. Na sequência, a segunda variável com o maior *IIV* é inserida no subconjunto de variáveis, uma nova clusterização é executada e o valor do SI médio das observações é armazenado. Tal procedimento iterativo é repetido até que todas as *P* variáveis sejam inseridas no subconjunto de variáveis utilizadas para a clusterização. Concluída a clusterização, tem-se o valor do SI médio para cada par ordenado (*m*, *k*), no qual *m* representa o número de variáveis utilizadas na clusterização e *k,* o número de grupos formados.

Passo 5 – Definir k = k + 1 e retornar ao passo 4

Para determinar o número de *clusters* mais adequado, define-se $k = k + 1$ e recomeça-se o procedimento de inserção de variáveis, clusterização de observações e avaliação da qualidade dos agrupamentos via SI. Esse procedimento é repetido até que o limite superior em *k* seja atingido. Neste estudo, investigou-se a separação em até 5 *clusters*, tido por especialistas como o limite superior em termos de implicações práticas.

Passo 6 – Identificar o melhor número de clusters e as variáveis para clusterização

O valor máximo do SI médio para o intervalo de número de *clusters* testados é identificado; tal valor indica o melhor número de *clusters*, assim como as variáveis recomendadas pelo método.

Os resultados globais obtidos com a utilização de funções de *kernel* são comparados aos resultados gerados pelos dados em seu formato original, a fim de avaliar ganhos de desempenho devidos ao remapeamento.

4. Exemplos numéricos

A sistemática proposta foi aplicada em três bancos de dados da indústria química, apresentados na Tabela 1. O banco ADPN refere-se à produção de um subcomponente no processo de produção do nylon; o banco LATEX foi coletado em um estágio de polimerização da fabricação de látex; o banco SPIRA foi obtido em um processo da indústria farmacêutica para produção de um antibiótico. Maiores detalhes acerca de tais bancos podem ser obtidos em Gauchi & Chagnon (2001).

A sistemática proposta foi implementada em MATLAB® 7.0. A função criada para realizar as análises apresenta como parâmetros de saída os valores do SI médio para os dados no espaço original e remapeados para cada banco de dados. Os tempos de execução foram: 22 segundos para o banco ADPN, três minutos e 15 segundos para o banco LATEX e 55 segundos para o banco SPIRA. Um procedimento de validação cruzada recomendou a retenção de três componentes principais na ACP.

A Figura 1 apresenta o perfil de SI médio para o banco ADPN à medida que as variáveis de clusterização são inseridas no procedimento para dois *clusters* e *kernels* distintos. Percebe-se que a inserção de variáveis

Tabela 1. Descrição dos bancos de dados utilizados.

Banco de dados	Número de variáveis	Número de observações
ADPN	100	71
LATEX	117	262
SPIRA	96	145

Fonte: elaborada pelos autores.

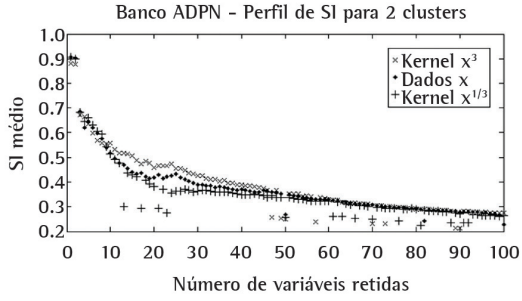

Figura 1. Perfis de SI médio para o banco ADPN na formação de 2 *clusters*.

degrada o valor médio do SI, confirmando que um subconjunto reduzido das variáveis originais conduz a melhores agrupamentos de bateladas produtivas. Percebe-se ainda uma alternância no desempenho de clusterização entre os dois *kernel*s testados. Gráficos semelhantes foram gerados para os demais bancos.

A Tabela 2 apresenta os valores máximos de SI médio da aplicação da sistemática sobre os dados originais (X) e sobre os remapeamentos ($X^{1/3}$ e X^3) para dois, três, quatro e cinco *clusters*. Os valores em negrito apontam o melhor desempenho para cada banco em cada número de *clusters*. Nos casos em que há valores semelhantes na Tabela 2, o desempate foi feito com base nas demais casas decimais, não apresentadas por restrição de espaço na tabela. Observa-se que os dados originais apresentaram melhor qualidade na clusterização em apenas três instâncias, enquanto o remapeamento para $X^{1/3}$ foi melhor em quatro cenários e o remapeamento para X^3 foi melhor em cinco cenários. Melhorias na clusterização quando da utilização do *kernel* $X^{1/3}$ são justificadas pela melhor redistribuição das observações, por conta da compressão dos dados, efeito esse decorrente da raiz cúbica do *kernel*; o oposto é verdadeiro para o *kernel* X^3, no qual a elevação das variáveis ao cubo promove um afastamento das observações no espaço e melhora a formação dos agrupamentos.

Percebe-se a grande recuperação de informações sobre a estrutura dos bancos ADPN e LATEX para dois *clusters*: tanto com a utilização das funções *kernel* quanto dos dados originais a sistemática de seleção de variáveis obteve bons resultados. A utilização dos *kernels* para números de *clusters* maiores do que dois obteve agrupamentos substancialmente melhores do que a utilização dos dados originais para o banco LATEX e SPIRA; para quatro *clusters*, o mapeamento realizado pelo *kernel* $X^{1/3}$ gerou SIs médios maiores do que 0,85, significativamente superiores aos gerados pelos dados originais. Para o banco ADPN, a utilização dos *kernels* obteve um resultado ligeiramente melhor para três *clusters* e resultados similares para quatro e cinco *clusters*.

A Tabela 3 apresenta os percentuais de variáveis retidas pela sistemática, os quais estão atrelados aos valores de SI médio apresentados na Tabela 2.

Observa-se que os percentuais de variáveis retidas ficam abaixo de 5% na maioria dos casos, com uma

Tabela 2. Valores máximos de SI médio para distintos *kernels* e números de *clusters*

Bancos de dados	*k* = 2			*k* = 3			*k* = 4			*k* = 5		
	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3
ADPN	0,90	**0,90**	0,87	0,79	0,80	**0,86**	0,79	**0,80**	0,80	**0,73**	0,73	0,73
LATEX	0,86	**0,94**	0,89	0,87	0,72	**0,89**	**0,92**	0,68	0,81	**0,81**	0,66	0,77
SPIRA	0,59	0,55	**0,85**	0,66	0,55	**0,80**	**0,89**	0,50	0,85	0,77	0,51	**0,80**

Fonte: elaborada pelos autores.

Tabela 3. Percentual de variáveis retidas pela sistemática proposta.

Bancos de dados	$k = 2$			$k = 3$			$k = 4$			$k = 5$		
	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3
ADPN	2,00	2,00	2,00	2,00	2,00	2,00	3,00	3,00	2,00	2,00	3,00	4,00
LATEX	4,27	2,56	3,42	2,56	4,27	3,42	1,71	2,56	3,42	2,56	3,42	5,98
SPIRA	2,08	2,08	93,75	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08
Média	2,78	2,21	33,06	2,21	2,78	2,50	2,26	2,55	2,50	2,21	2,83	4,02

Fonte: elaborada pelos autores.

Tabela 4. Ganho percentual dos SIs médios obtidos pela sistemática de seleção de variáveis.

Bancos de dados	$k = 2$			$k = 3$			$k = 4$			$k = 5$		
	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3	$X^{1/3}$	X	X^3
ADPN	246	309	222	182	175	230	203	166	280	329	265	204
LATEX	230	248	256	262	176	25	283	257	224	285	153	83
SPIRA	195	175	80	371	266	73	709	257	84	600	292	81
Média	223,7	244,0	186,0	271,7	205,7	109,3	398,3	226,7	196,0	404,7	236,7	122,7

Fonte: elaborada pelos autores.

Tabela 5. Fatores e níveis do experimento.

Fatores	Níveis
Correlação das variáveis	$P^{1/3}$; P; P^3
Expoente do *kernel*	1/5; 1/3; 1; 3; 5
Proporção de observações utilizadas	0,5; 1; 5

média geral de 5,16% devida a uma exceção: o mapeamento para X^3 no banco SPIRA, na formação de dois *clusters*.

A Tabela 4 apresenta a elevação de qualidade na clusterização com base no valor de SI médio obtido pela utilização do subconjunto de variáveis de clusterização recomendado pela sistemática; o ganho médio, considerando-se todos os cenários e bancos, é de 235,4% frente à utilização de todas as variáveis. Observa-se que o banco SPIRA é o que apresenta maiores divergências nos ganhos aferidos, sendo responsável pelos maiores e menores níveis percentuais de aprimoramento do desempenho de clusterização. Verifica-se que a função *kernel* $X^{1/3}$ apresentou significativos percentuais de melhora dentro da sistemática de seleção de variáveis; contudo, observando-se em conjunto a Tabela 2, a utilização desse *kernel* levou a melhores resultados do que o *kernel* X^3 em cinco de 12 possibilidades, e em quatro dessas cinco conduziu ao melhor SI médio na análise.

5. Experimentos de simulação

Um estudo de simulação foi realizado para avaliar a robustez do método proposto. Os dados foram gerados tomando-se por base informações reais de um processo de reciclagem de papel, fornecido por Wold et al. (2001). O banco de dados original consiste em 386 observações, descritas por 54 variáveis. Os bancos gerados na simulação seguem distribuições normais multivariadas com média μ, variâncias dadas de acordo com uma matriz Σ de covariâncias e correlações de acordo com uma matriz P. O vetor μ e as matrizes Σ e P são extraídos do banco original.

Os fatores em estudo são (i) correlação entre variáveis, (ii) proporção entre observações e variáveis e (iii) expoente do *kernel* polinomial. O fator correlação é investigado em três níveis: alto, utilizando $P^{1/3}$; nominal, utilizando P; e baixo, utilizando P^3. O fator proporção de observações utilizadas é investigado em três níveis: alto, considerando 400% das observações disponíveis (cinco vezes o número original de observações); nominal, considerando as 386 observações; baixo, considerando 50% do total de observações disponíveis. O fator expoente da transformação *kernel* conta com cinco níveis: $X^{1/5}$, $X^{1/3}$, X, X^3, e X^5. A Tabela 5 apresenta os fatores e os níveis do experimento, enquanto que a Tabela 6 traz os resultados da simulação. Os resultados apoiam-se na formação de dois clusters (k = 2).

Com base nos resultados da Tabela 6 percebe-se uma tendência de elevação no SI médio para transformações do tipo X^5 e X^3 frente aos dados originais (X); tal tendência não é unânime para transformações do tipo $X^{1/3}$ e $X^{1/5}$. Alterações no fator *kernel* não conduziram a tendências claras em termos do percentual de variáveis retidas. Com bases nos dados avaliados, não se pode concluir que diferentes proporções de número de observações e de variáveis impactam sobre o SI médio ou o percentual de variáveis retidas. Por fim, a redução da correlação entre variáveis tende a aumentar o SI médio, sugerindo que menores correlações favorecem a geração dos índices de importância de variáveis e, por consequência, aprimoram o subconjunto de variáveis para clusterização. Não há tendência clara desse fator sobre o percentual de variáveis retidas.

Tabela 6. SI médio e porcentagem de variáveis retidas para os experimentos de simulação.

Proporção observações/ variáveis	Correlação das variáveis	$X^{1/5}$		$X^{1/3}$		X		X^3		X^5	
		SI	% variáveis retidas	SI	% variáveis retidas	SI	% variáveis retidas	SI	% variáveis retidas	SI	% variáveis retidas
0.5	P^3	0,7704	12,96	0,9913	1,85	0,7603	9,26	0,9898	14,81	0,9918	7,41
	P	0,7062	7,41	0,8353	1,85	0,6919	3,70	0,8175	1,85	0,8792	1,85
	$P^{1/3}$	0,6451	1,85	0,6953	3,70	0,6037	1,85	0,6864	3,70	0,7690	9,26
1	P^3	0,7238	9,26	0,9946	3,70	0,7080	5,56	0,9850	37,04	0,9814	40,74
	P	0,7575	5,56	0,7971	3,70	0,7394	1,85	0,7829	3,70	0,9190	3,70
	$P^{1/3}$	0,6639	3,70	0,6340	3,70	0,6378	3,70	0,6190	0,46	0,8038	0,92
5	P^3	0,7687	11,11	0,9640	5,56	0,7198	5,56	0,9633	5,56	0,9990	5,56
	P	0,7534	3,70	0,7976	9,26	0,7141	1,85	0,7777	3,70	0,8849	3,70
	$P^{1/3}$	0,6487	5,56	0,6234	14,81	0,6347	7,41	0,6149	7,41	0,7745	7,41

6. Conclusões

A obtenção de grupos de observações distintos entre si mas com afinidade interna entre seus integrantes é uma tarefa de grande interesse em várias áreas de conhecimento. No contexto da indústria química, mais do que a simples classificação de bateladas de produção, o grande desafio é estabelecer relações de níveis de qualidade entre essas bateladas.

Neste artigo foi proposta uma sistemática para seleção das variáveis de clusterização mais relevantes para a formação de famílias de bateladas com características similares. Sua operacionalização apoia-se na integração de funções *kernel* a um novo índice de importância de variáveis gerado a partir dos parâmetros oriundos da Análise de Componentes Principais. A inserção de relações não lineares teve como objetivo obter melhores agrupamentos para bateladas de produção.

O método proposto foi aplicado a três bancos distintos, apresentando bons resultados. Comparado à utilização da sistemática sem remapeamento, a utilização dos *kernels* elevou, na maior parte dos casos, a qualidade dos agrupamentos, do que conclui-se que a sistemática com remapeamento tem grande potencial de uso. Dos três bancos testados em quatro cenários (distinto número de *clusters*), a utilização dos *kernels* mostrou-se superior em nove das 12 possibilidades. Também realizou-se um experimento de simulação, o qual avaliou variações no SI médio e porcentagem de variáveis retidas frente a diferentes níveis de fatores tidos como relevantes.

Ressalta-se que o objetivo da sistemática proposta é agrupar as observações de forma a maximizar o SI. Estudos futuros podem contemplar o desenvolvimento de uma sistemática que compatibilize elevados valores de SI com reduzido percentual de variáveis retidas. Para tanto, sugere-se a adoção de um critério de distância mínima a um ponto ótimo hipoteticamente arbitrado. Tal proposição priorizaria a seleção de subconjuntos que originam SIs elevados e com poucas variáveis retidas.

Referências

Abe, S. (2010). *Support Vector Machines for Pattern Recognition* (2nd ed.). London: Springer-Verlag. http://dx.doi.org/10.1007/978-1-84996-098-4

Agard, B., & Penz, B. (2009). A simulated annealing method based on a clustering approach to determine bills of materials for a large product family. *International Journal of Production Economics, 117*(2), 389-401. http://dx.doi.org/10.1016/j.ijpe.2008.12.004

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). New Jersey: John Wiley & Sons, Inc.

Anzanello, M. J., & Fogliatto, F. S. (2011). Selecting the best clustering variables for grouping mass-customized products involving workers' learning. *International Journal of Production Economics, 130*(2), 268-276. http://dx.doi.org/10.1016/j.ijpe.2011.01.009

Baghshah, M. S., & Shouraki, S. B. (2011). Learning low-rank *kernel* matrices for constrained clustering. *Neurocomputing, 74*, 2201-2211. http://dx.doi.org/10.1016/j.neucom.2011.02.009

Bessaoud, F., Tretarre, B., Daurès, J. P., & Gerber, M. (2012). Identification of dietary patterns using two statistical approaches and their association with breast cancer risk: a case-control study in southern France. *Annals of Epidemiology, 22*(7), 499-510. PMid:22571994. http://dx.doi.org/10.1016/j.annepidem.2012.04.006

Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis, 52*, 502-519. http://dx.doi.org/10.1016/j.csda.2007.02.009

Brusco, M. J. (2004). Clustering binary data in the presence of masking variables. *Psychological Methods, 9,* 510-523. PMid:15598102. http://dx.doi.org/10.1037/1082-989X.9.4.510

Brusco, M. J., & Cradit, J. D. (2001). A variable-selection heuristic for k-means clustering. *Psychometrika, 66*(2), 249-270. http://dx.doi.org/10.1007/BF02294838

Dean, N., & Raftery, A. E. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics, 62*(1), 11-35. PMid:20827439 PMCid:PMC2934856. http://dx.doi.org/10.1007/s10463-009-0258-9

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology, 64*, 304-310. http://dx.doi.org/10.1016/0002-9149(89)90524-9

Domenicone, C., Peng, J., & Yan, B. (2011). Composite *kernel*s for semi-supervised clustering. *Knowledge and Information Systems, 28*(1), 99-116. http://dx.doi.org/10.1007/s10115-010-0318-8

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). New York: Wiley-Interscience.

Filippone, M., Camastra, F., Masulli, F., & Rovetta, S. (2008). A survey of *kernel* and spectral methods for clustering. *Pattern Recognition, 41*(1), 176-190. http://dx.doi.org/10.1016/j.patcog.2007.05.018

Friedman, J. H., & Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society, Series B, 66*, 815-849. http://dx.doi.org/10.1111/j.1467-9868.2004.02059.x

Gauchi, J. P., & Chagnon, P. (2001). Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics Intelligent Laboratory Systems, 58*, 171-193. http://dx.doi.org/10.1016/S0169-7439(01)00158-7

Girolami, M. (2002). Mercer *Kernel*-Based Clustering in Feature Space. *IEEE Transactions on Neual Networks, 13*(3), 780-784. PMid:18244475. http://dx.doi.org/10.1109/TNN.2002.1000150

Gnanadesikan, R., Kettenring, J., & Tsao, S. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification, 12*(1), 113-136. http://dx.doi.org/10.1007/BF01202271

Hair, J., Anderson, R., Tatham, R. & Black, W. (1995). *Multivariate Data Analysis with Readings* (4th ed.). New Jersey: Prentice-Hall Inc.

Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(5), 657-668. PMid:15875789. http://dx.doi.org/10.1109/TPAMI.2005.95

Huang, T., Kecman, V., & Kopriva, I. (2006). *Kernel based algorithms for mining huge data sets, Supervised, Semi-supervised, and Unsupervised learning.* Berlin: Springer-Verlag.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer-Verlag.

Kaufman, L., & Rousseeuw, P. (2005). *Finding Groups in Data: an Introduction to Cluster Analysis.* New Jersey: Wiley Interscience.

Li, Y., Dong, M., & Hua, J. (2008). Localized feature selection for clustering. *Pattern Recognition Letters, 29*(1), 10-18. http://dx.doi.org/10.1016/j.patrec.2007.08.012

Maugis, C., Celeux, G., & Martin-Magniette, M. (2009). Variable selection for clustering with Gaussian mixture models. *Biometrics, 65*(3), 701-709. PMid:19210744. http://dx.doi.org/10.1111/j.1541-0420.2008.01160.x

Meek, C., Thiesson, B., & Heckerman, D. (2002). The learning-curve sampling method applied to model-based clustering. *Journal of Machine Learning Research, 2*, 397-418.

Milligan, G. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45*, 325-342. http://dx.doi.org/10.1007/BF02293907

Milligan, G., & Cooper, M. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 5*, 181-204. http://dx.doi.org/10.1007/BF01897163

Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association, 101*, 168-178. http://dx.doi.org/10.1198/016214506000000113

Rousseeuw, P. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. http://dx.doi.org/10.1016/0377-0427(87)90125-7

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyound.* Cambridge: The MIT Press.

Steinley, D. (2004). Standardizing variables in *K*-means clustering. In D. Banks, L. House, F. R. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, clustering, and data mining applications* (pp. 53-60). New York: Springer. http://dx.doi.org/10.1007/978-3-642-17103-1_6

Steinley, D., & Brusco, M. J. (2008a). A new variable weighting and selection procedure for K-means cluster analysis. *Multivariate Behavioral Research, 43*(1), 77-108. http://dx.doi.org/10.1080/00273170701836695

Steinley, D., & Brusco, M. J. (2008b). Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika, 73*(1), 125-144. http://dx.doi.org/10.1007/s11336-007-9019-y

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques do diagnose breast cancer from fine-niddle aspirates. Cancer Letters, 77, 163-171. http://dx.doi.org/10.1016/0304-3835(94)90099-X

Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics.*Chemometrics Intelligent Laboratory Systems, 58*(2), 109-130. http://dx.doi.org/10.1016/S0169-7439(01)00155-1

Clustering variable selection for grouping production batches through PCA and *kernel* mapping

Abstract

Clustering techniques are tailored to find internally homogeneous groups of observations. In industrial processes that rely on batches, grouping batches with similar profiles provides valuable information about process control and monitoring. This paper proposes a variable selection approach based on the *kernel* function and Principal Component Analysis (PCA). The clustering quality is assessed through the *Silhouette Index* (SI). When applied to three industrial processes, the proposed approach retained an average of 5.16% of the original variables, yielding on average a 235.4% more precise batch grouping. We also performed a simulation experiment.

Keywords

Clustering analysis. Variable selection. *Kernel*. Batch processes.