Utilizando simulação para auxiliar a classificação em métodos de seleção de variáveis

Alessandro Kahmann (PPGEP/UFRGS)
Rodolfo Reinaldo Hermes Petter (PPGEP/UFRGS)
João Francisco da Fontoura Vieira (PPGEP/UFRGS)
Liane Werner (PPGEP/UFRGS)

Resumo
A melhoria da qualidade e precisão no monitoramento e controle dos processos industriais faz-se cada vez mais exigente, frente a constante evolução do dinamismo e exigências do mercado consumidor, demandando assim um grande esforço da indústria no desenvolvimento de seus métodos de garantia da qualidade. Frente a isso, a presente pesquisa utilizou simulações para gerar mais dados para análise do comportamento de processos industriais, os quais possuem dados insuficientes para uma análise com acurácia satisfatória. Assim, os dados originais, bem como os gerados via simulação, foram classificados e analisados com auxílio das ferramentas KNN (K-Nearest Neighbor) e IVV Índices de Importância de Variáveis. De forma geral, conseguiu-se por meio desta pesquisa, verificar uma aplicabilidade eficiente no teste ora proposto, entretanto a pesquisa limita-se somente para bancos de dados contendo poucas observações.
Palavras Chave: Simulação, seleção de variáveis, qualidade.

1 Introdução

Com o acelerado avanço de tecnologias de monitoramento de processos, emerge a possibilidade de geração de bancos de dados com o objetivo da construção e estabelecimentos de padrões que auxiliem na previsão e controle de processos nos mais diversos segmentos industriais.

Entretanto, tal avanço acaba gerando bancos de dados excessivamente grandes, dos quais interferem negativamente na qualidade da análise e monitoramento dos processos, fazendo com que ocorra uma redução na acurácia deste monitoramento (Kettaneh et al., 2005, Liu e Yu 2005).

A situação supracitada se agrava em processos que produzem por bateladas pelo fato de se observarem menos amostras para a análise em relação a outros processos, resultando desta forma na geração de bancos de dados com maior quantidade de variáveis do que de amostras.

Ressalta-se, porém que, em alguns processos industriais não há a necessidade de se analisar todas as variáveis que o constituem, pois como se tem variáveis em excesso, gera-se o chamado ruído na análise dos dados, do qual caracteriza a diminuição de acurácia na análise, fazendo-se necessária a escolha das variáveis a serem analisadas.

Tal escolha é realizada via técnicas para seleção de variáveis, das quais possuem por função básica a diminuição do ruído do banco de dados em análise e remoção de dados correlatos visando uma melhora substancial da acurácia e qualidade da análise e previsão do comportamento de processo e, ainda, tratam-se de técnicas estruturadas sobre modelos estatístico/matemáticos menos complexos, fazendo com que o tempo de análise seja diminuído consideravelmente (Gauchi e Chagnon, 2001, Anzanello et al., 2011, Westad et al., 2003, Urtubia et al., 2007).

São diversas as técnicas existentes e utilizadas para seleção de variáveis, no entanto, de acordo com Anzanello et al. (2009) a denominada KNN (k-Nearest Neighbor) destacou-se no quesito acurácia de análise, apresentando melhores resultados quando comparada a outras técnicas. Mas esta acurácia, independente da técnica selecionadora, pode ser afetada pela existência de dados correlatos, como também por dados ruidosos.

Frente a estas desvantagens, Anzanello et al. (2011) utiliza o Índice de Importância de Variáveis (IIV), do qual possui por objetivo central ordenar a remoção de variáveis para que, ao longo das exclusões, a acurácia do método classificador melhore como consequência da eliminação dos dados ruidosos e correlatos.

Com base nestes pressupostos, a presente pesquisa teve por objetivo gerar dados por meio da técnica de simulação de dados aleatórios em dois bancos de dados apresentados por Gauchi e Chagnon (2001), provindos

de um processo industrial que produz em bateladas. A geração de dados via simulação, dá-se em função dos bancos de dados utilizados possuírem reduzida quantidade de amostras, comprometendo assim a acurácia do método KNN.

Por conseguinte, objetivou-se utilizar da simulação de dados para testar a possibilidade de melhoria dos bancos de dados, frente à técnica classificatória, com o intuito de melhorar a qualidade da análise.

2 Referencial Teórico

2.1 K-Nearest Neighbor (KNN)

O KNN é uma técnica simples de data mining para classificação de dados, da qual se baseia na distância do ponto dos dados em análise em relação aos pontos gerados pelos dados de determinado banco de treino (DUDA et al., 2001).

A técnica considera observações *n*-dimensionais de um banco de dados, correspondendo a *n* atributos, e duas classes de produtos, neste caso, bom e ruim. O objetivo é classificar as novas observações em um destes dois grupos.

Inicialmente, é definido o tamanho de *k*, que é a quantidade de amostras que serão retiradas do banco de dados de teste para a classificação. Por conveniência, atribui-se a *k* um valor impar. Para a classificação de uma amostra são tomados então os *k* pontos do banco de treino mais próximos a esta amostra, de acordo com a distância euclidiana entre estes pontos. Destes *k* pontos são verificadas a qual grupo pertencem, e o grupo que possuir a maior quantidade de pontos, destes mais próximos, será o classificador da amostra analisada.

A vantagem de utilizar esta técnica é sua intuitiva aplicação e simplicidade teórica. Outro quesito que sobrepõe esta técnica em relação a outras é sua presença em diversos softwares estatísticos (Anzanello et al., 2009, Duda et al., 2001).

2.2 Índices de Importância de Variáveis (IVV)

Westad et al. (2003) cita que o principal objetivo de ranquear as variáveis é prevenir correlações espúrias que podem ser tratadas como informações relevantes. Para isso, o IIV ordena as variáveis para futura remoção, com o intuito de retirar informações não importantes ou que prejudiquem a classificação. O IIV utilizado neste artigo será o encontrado em Anzanello et al. (2011). A geração do índice é composta por 2 passos:

1 – Análise de componentes principais (ACP) nos dados já existentes (banco de treino).

2 – Soma dos pesos w da ACP para cada variável n

O índice baseia-se nos pesos *w* das variáveis nos componentes principais somados de forma absoluta. Os números devem ser analisados na forma absoluta uma vez que pesos negativos anulam pesos positivos ou reciprocamente, o que não é o objetivo deste índice, pois este analisa apenas o quanto a variável influencia no componente incondicionada a sua positividade. Ou seja, o IIV associado a cada variável n é dado pela equação (1):

$$IIV_n = \sum_{m=1}^{p} |w|_n \qquad (1)$$

Desta forma, as variáveis são ordenadas em forma decrescente de acordo com o IIV e, nesta ordem, as últimas variáveis são uma a uma excluídas. Ou seja, uma variável com menor IIV tem uma tendência a explicar menos da variabilidade do processo, ou ser redundante a outra variável, que uma variável com maior IIV, o que pode resultar em uma análise menos precisa, e por esta razão é retirada antes.

2.3 Geração de números aleatórios

A geração de números aleatórios permite simular dados de processos em que o andamento dependa de fatores aleatórios. A partir de amostras do processo serão extraídas as médias e desvios padrão de cada variável que serão utilizadas para gerar novos dados a partir da distribuição dos dados recolhidos.

Neste contexto Garcia et al. (2010) argumenta que a simulação torna-se um fator colaborativo na tomada de decisões, uma vez que permite a proposição de inferências, através dos experimentos gerados pela simulação, no que tange às variáveis envoltas à tomada de decisão que está sendo analisada, bem como apoiada pela simulação.

Adicionalmente, Moore & Weatherford (2005) argumentam que os dados gerados pela simulação permitem avaliar diversas possibilidades de novos e possíveis cenários dos quais poderão ocorrer sobre a realidade que está sendo testada, podendo ainda ser apontadas as melhores soluções com base no mínimo de desvantagens e no máximo de vantagens presentes nas hipóteses geradas via simulação.

Especificamente a simulação de dados, segundo Gentle (2003) atenta-se na geração de números aleatórios tendo por objetivo de adição de valores nas variáveis do cenário e/ou sistema que está sendo analisado. Os valores gerados pelo método são obtidos com o auxílio de tabelas eletrônicas e/ou softwares estatísticos especializados.

A cada ciclo simulado, os valores obtidos são armazenados e ao final da simulação os dados são transformados, da qual permite inferir nestes por meio de estatística descritiva, o que permite a análise e construção das novas hipóteses para o cenário em análise (Gujarati, 2002).

3 O Banco de Dados

Os bancos de dados utilizados nesta pesquisa foram primeiramente utilizados por Gauchi e Chagnon (2001), com o intuito de predição de comportamento do resultado do processo. Já Anzanello et al. (2012) utilizou-os num intuito diferente, tendo por base um valor definido para classificação, onde os valores finais do processo são ignorados, considerando então apenas sua classificação, podendo ser a interpretação resumida no caso de uma batelada, se ela se apresenta boa ou ruim.

Originalmente o banco de dados contém dados de 5 processos, ADPN, LATEX, SPIRA, OXY e GRANU. Em Anzanello et al. (2012), há uma simulação de dados, consequência da pequena quantidade de dados originais nos bancos OXY e GRANU. Por possuirem poucos dados, neste trabalho o método proposto foi realizado apenas nestes dois bancos de dados.

Os dados do banco de dados OXY são derivados do processo industrial de óxido de titânio, já os dados do banco de dados GRANU foram retirados do processo de fabricação de emulsões antiespumantes, utilizados nas indústrias de papéis. Mais informações sobre os bancos de dados podem ser encontradas em Gauchi e Chagnon (2001).

Tabela 1 – Bancos de dados utilizados no estudo

Banco de dados	**Número de variáveis no processo**	**Número de observações**	
		Treino	**Teste**
OXY	95	20	5
GRANU	78	23	6

Fonte: Gauchi e Chagnon (2001).

4 O Método

Tendo por objetivo a correta classificação dos dados através dos métodos IIV e KNN, é necessário o estabelecimento correto do ponto de corte dos dados, neste caso utilizou-se o mesmo encontrado em Anzanello et al. (2012). Assim, torna-se possível separar os conjuntos de dados em grupo de amostras "bons" e "ruins".

O método de seleção de variáveis ocorre através da seguinte sequência de ações:

1ª – Separar o banco de dados em grupos de treino e teste: separação realizada com base no tamanho de amostras pré-definido, identificando o banco de treino que tem por objetivo definir as propriedades numéricas do método e o banco de teste é utilizado para verificação da acurácia do método.

A partir deste momento considera-se que os dados existentes são os que compõem o banco de treino, enquanto os pertencentes ao banco de teste serão as "novas amostras".

2ª – Gerar o IIV: Com o banco de treino definido, são utilizadas as amostras contidas neste grupo para executar o

IIV. Neste estudo foi utilizado o IIV proposto em Anzanello et al (2011), que baseia-se nos pesos resultantes da Análise de Componentes Principais (ACP) dos dados do banco de treino.

Cada um dos componentes resultantes da análise possui um percentual da variância total. Para não ser afetada por componentes que pouco explicam a variabilidade do processo, foram utilizados apenas os dois primeiros componentes da análise.

3ª – Gerar novos dados, baseados no banco de treino: para isto, obtiveram-se a média e desvio-padrão do conjunto de dados e, na sequência, executou-se a simulação com auxílio de uma planilha eletrônica no software Microsoft Excel®, operacionalizado pela função expressa em (2):

= INV.NORM(ALEATÓRIO(); média; desvio padrão) (2)

Desta forma, geraram-se novas amostras para cada uma das variáveis do banco de treino, sendo simuladas 280 novas amostras para o banco de dados OXY e 277 para o banco de dados GRANU, totalizando 300 observações para ambos. Porém estas novas amostras, que derivam da simulação, carecem de ser classificadas em um dos dois grupos, bom ou ruim. Para isso é utilizada a KNN, neste caso é utilizado k=3.

Este passo acaba criando uma "nuvem gráfica" de amostras que define melhor cada um dos grupos. Este é o objetivo principal deste método, criar um banco de treino melhor definido por possuir maior quantidade de amostras.

4ª – Classificar o banco de testes utilizando a KNN: uma vez o banco de treino definido classificam-se individualmente as amostras do banco de testes. Neste passo foram atribuídos para k valores ímpares entre 3 e 11, sem que houvesse diferença. Valores maiores que estes não foram testados, pois apesar de haverem 300 observações no banco de treino, foi observado que estes bancos de dados possuem um desequilíbrio de amostras entre os dois grupos, onde o grupo das amostras classificadas como boas representam mais de 75% do total. Como a técnica KNN é sensível a tal desiquilíbrio, utilizar valores de k muito altos tende a resultar em classificações exclusivas ao grupo com mais amostras.

5ª – Verificar a acurácia: após todo o banco de teste classificado, as classificações resultantes são comparadas com as classificações reais. A acurácia do método em cada banco de dados é verificada através da equação (3).

$$ACC = \frac{\text{Classificações corretas}}{\text{Total de amostras}} \quad (3)$$

6ª – Remover uma variável de acordo com o IIV e repetir o método: das variáveis analisadas retira-se a com menor IIV e repete-se o método até que reste apenas uma variável. Desta forma o método será repetido 95 vezes para o banco de dados OXY e 78 vezes para o banco de dados GRANU.

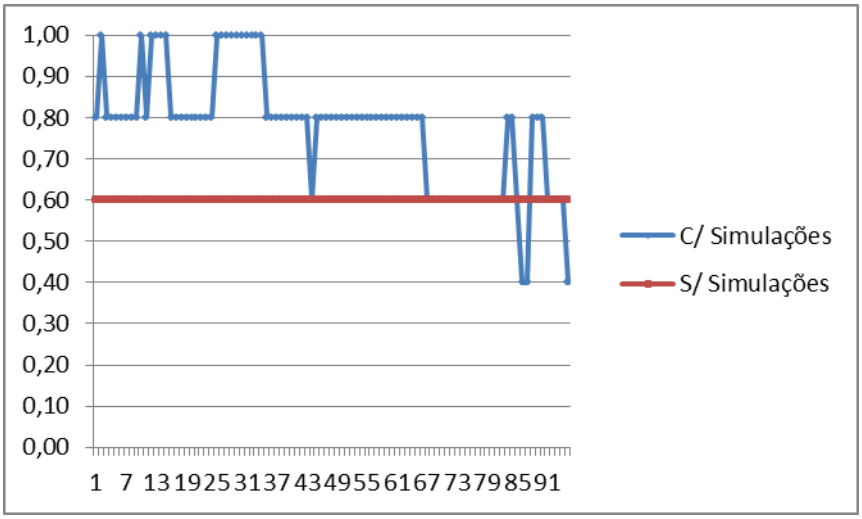
A cada variável removida a tendência é de que a acurácia alcance um ponto ótimo onde a acurácia melhore ou, pelo menos, se mantenha. Porém este comportamento se mantém até determinado ponto, pois raramente o processo pode ser analisado através de apenas uma variável. A partir deste ponto a acurácia tende a cair.

Os resultados do método serão comparados aos obtidos através da classificação dos dados utilizando o mesmo IIV e a mesma ferramenta classificatória, porém sem o acréscimo dos dados simulados ao banco de treino. Logo serão comparados os picos de acurácia entre a classificação com as amostras simuladas e sem as amostras simuladas, para assim determinar a validade do método.

5 Resultados

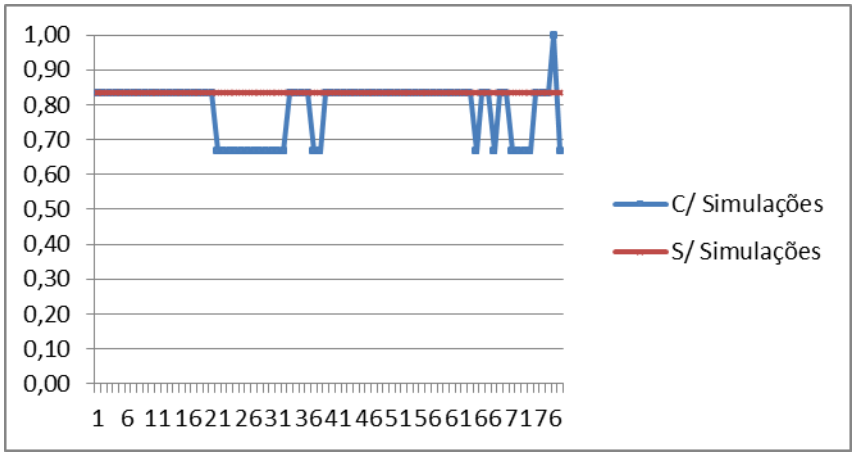
Como citado anteriormente, para a verificação dos resultados serão comparadas as acurácias das classificações utilizando o IIV de Anzanello et al (2011) para ordenação de remoção das variáveis e a classificação através da técnica KNN, com e sem as observações simuladas. Na figura 1, é observado que o acréscimo dos dados simulados acarretou em uma melhor classificação, onde todos as amostras do banco de teste foram classificadas corretamente, resultando em uma acurácia igual a 1.

Figura 1. Comparação de acurácias com o banco de dados OXY



Na figura 2 o ganho na acurácia aparece apenas no ponto ótimo. Porém, apesar de ao longo da remoção as variáveis não haver ganho de acurácia na classificação, o ponto ótimo estar com uma acurácia melhor, e o fato de este valor ser 100%, já é suficiente para justificar a utilização do método proposto neste artigo.

Figura 2. Comparação de acurácias com o banco de dados GRANU



Das figuras 1 e 2 observa-se que em ambos os bancos de dados, a acurácia máxima foi maior em ambos. A acurácia foi constante para os dois sem as observações simuladas, porém com o acréscimo das simulações houve alguma combinação de variáveis retidas que proporcionou 100% de acurácia nos resultados dos dois bancos de dados.

No gráfico da figura 1, onde se tem as acurácias dos dados OXY, é constante o fato de as simulações não causarem perda de acurácia. Sempre houve mais acertos, ou no mínimo igual quantidade de acerto em relação à execução do método sem o auxílio das simulações.

Já no gráfico da figura 2, dos dados GRANU, notou-se uma vantagem em apenas um momento. Tal diferença pode ser explicada pelo fato de o banco de dados GRANU ter amostras com poucos valores diferentes dentro da mesma variável, mas mesmo assim bastante dispersos entre si.

Portanto, na geração de dados simulados, os valores do banco GRANU não foram tão semelhantes à realidade. Já o banco de dados OXY, possui os valores de suas variáveis diferentes em todas as amostras, resultando em um melhor resultado.

6 Considerações Finais

Neste trabalho foi proposto o acréscimo de amostras, por meio de simulação, no momento de realizar a

classificação através da ferramenta KNN. Em processos no formato de bateladas a obtenção de observações é baixa. Considerando o valor de k maior ou igual a 3 a quantidade de pontos classificadores selecionados é proporcionalmente grande.

As novas amostras são incluídas no intuito de diminuir tal proporção evitando que dados atípicos tenham influencia significativa na classificação de novas amostras. Como o banco de dados utilizado continha muitas variáveis foi adicionalmente utilizado o IIV visto em Anzanello et al (2011), tanto na classificação das amostras simuladas como nas amostras separadas no banco de teste.

A proposta de realizar o método de seleção de variáveis e classificar as amostras com a ferramenta KNN com a adição de amostras simuladas se mostrou de aplicabilidade eficiente nos casos onde existem poucas observações. Em comparação, a adição de amostras simuladas resultou em classificações melhores contra a operação do método apenas com as observações reais.

Apesar de mostrar um resultado positivo esta proposta ainda carece de aperfeiçoamentos. A seleção de variáveis é aplicada também na classificação das amostras simuladas. Portanto para encontrar o resultado ótimo foi rodado o teste com as reclassificações a cada variável removida durante a classificação das amostras. Para tornar este método mais eficiente se torna necessário uma pesquisa para uma classificação das amostras também mais eficiente.

Outra possibilidade de pesquisas futuras relacionadas a este trabalho é a utilização de outras técnicas de classificação baseadas em medidas de distância como a distância de Mahalanobis.

Referências Bibliográficas

ANZANELLO, M.J., ALBIN, S.L., CHAOVALITWONGSE, W.A. 2012. Multicriteria variable selection for classification of production batches. European Journal of Operational Research, v.218, n.1, p.97-105.

ANZANELLO, M.J., ALBIN, S.L., CHAOVALITWONGSE, W.A. 2009. Selecting the best variables for classifying production batches into two quality levels. Chemometrics and Intelligent Laboratory Systems, v.97, n.2, p.111-117.

ANZANELLO, M.J., FOGLIATTO, F.S., ROSSINI, K. 2011. Data mining-based method for identifying discriminant attributes in sensory profiling. Food Quality and Preferences, v.22, n.1, p.139-148.

DUDA, R., HART, P., STORK, D. 2001. Pattern Classification. Second ed. New York, Wiley-Interscience.

GARCIA, S., LUSTOSA, P.R.B., BARROS, N.R. 2010. Aplicabilidade do Método de Simulação de Monte Carlo na Previsão dos Custos de Produção de Companhias Industriais: O Caso da Companhia Vale do Rio Doce. Revista de Contabilidade e Organizações, FEA-RP/USP, São Paulo, v.4, n.10, p.152-173.

GAUCHI, J., CHANGNON, P. 2001. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chemometrics and Intelligent Laboratory Systems, v.58, n.2, p.171-193.

GENTLE, J.E. 2003. Random Number Generation and Monte Carlo Methods. New York, Springer.

GUJARATI, D.N. 2002. Econometria básica. 3 ed. São Paulo, Makron Books.

KETTANEH, N., BERGLUND, A., WOLD, S. 2005. PCA and PLS in very large datasets. Computational Statistics & Data Analysis, v.48, n.1, p.69-85.

LIU, H., YU, L. 2005. Toward integrating feature selection algorithms for classification and clustering. Transactions on Knowledge and Data Engineering, v.17, n.4, p.491–502.

MOORE, J., WEATHERFORD, L.R. 2006. Tomada de decisão em administração com planilhas eletrônicas. 6 ed. Porto Alegre: Bookman Companhia Editora.

WESTAD, F., HERSLETH, M., MARTENS, H. 2003. Variable selection in PCA in sensory descriptive and consumer data. Food Quality and Preference, v.14, n.5-6, p.463–472.