Aplicação da Árvore Probabilística de Sufixo na Predição de Resultados do Processo de Extração de Café Solúvel

Everton da Silva[1], Elenir Lila Leobet de Lima[2], Fabrício Martins Lopes[3], André Yoshiaki Kashiwabara[3]

[1]Programa de Pós-Graduação em Informática – PPGI – Universidade Tecnológica Federal do Paraná – Cornélio Procópio – PR – Brasil

[2]Supervisão de Produção – Cia Iguaçu de Café Solúvel – Cornélio Procópio – PR – Brasil

[3]Departamento Acadêmico de Computação – Universidade Tecnológica Federal do Paraná – Cornélio Procópio – PR – Brasil.

{evertonvoid,lilaleobet}@gmail.com,
{fabricio,kashiwabara}@utfpr.edu.br

Abstract. *The extraction of instant coffee is an industrial process that generates in real time a large amount of data, such as yield, pH, temperature, concentration, percentage of soluble solids, among others. However, the data collected is still poorly explored to improve the instant coffee process. This work presents a methodology to summarize the results of the coffee extractor using probabilistic suffix trees, in which the observations from the past are used to estimate the probability of each class given a variable length context. These probabilities can indicate if the extractor is operating properly. Our methodology is under study at Cia Iguaçu de Café Solúvel and it would be extended to other applications in near future.*

Resumo. *A extração de café solúvel é um processo industrial que gera grande quantidade de dados em tempo real, como rendimento, pH, temperaturas, concentração, percentual de sólidos solúveis, dentre outros. No entanto, essa grande quantidade de dados é pouco aproveitada na melhoria do processo. Este trabalho apresenta uma metodologia capaz de sumarizar resultados do extrator de café por meio de árvores probabilísticas de sufixo, nas quais o histórico de observações dos resultados é utilizado na estimação de probabilidades de ocorrência de cada classe, indicando se o extrator está operando adequadamente. A metodologia está em estudo na Cia Iguaçu de Café Solúvel e poderá ser estendida para outras aplicações no futuro.*

1. Introdução

É comum que a melhoria no controle de processos industriais seja perseguida pelas empresas que buscam reduzir custos e, em contrapartida, manter a qualidade dos produtos oferecidos aos seus clientes, muitas vezes desconhecendo a metodologia ideal para esse fim [Santos 2014]. Diante da complexidade dos processos de fabricação e produção na indústria de alimentos e bebidas, que atualmente corresponde a 22% da indústria de transformação [CNI 2017], esse setor convive com o constante desafio na

busca por soluções inovadoras, exigindo mais esforços para análise e interpretação de seus dados. Esse cenário remete ao atual desafio da automação industrial em transformar tal volume de dados em informação [De Souza et. al. 2005]. Desse modo, há crescente motivação em identificar características nos processos de fabricação que levem às condições ideais e melhorem a qualidade sensorial na extração de café solúvel.

Neste trabalho, uma etapa do processo chamada de "Extração de Sólidos Solúveis", foi analisada. Percebeu-se a necessidade do desenvolvimento de metodologia computacional para processar e analisar seus indicadores, a fim de que seja possível a extração de conhecimento para a tomada de decisões.

A metodologia desenvolvida visa o auxílio nas ações de operação dos equipamentos para ajustes nas condições que são críticas para a qualidade dos produtos. Neste sentido, este trabalho apresenta uma metodologia baseada em árvores probabilísticas de sufixo [Leonardi 2006, Largeron 2003] e cadeias de Markov de alcance variável [Rissanen 1983], a qual, a partir de estatísticas geradas a cada novo registro, é capaz de estimar as probabilidades de ocorrência de cada classe de resultado de determinada variável, dada a observação do histórico passado.

Este trabalho está em desenvolvimento na Cia Iguaçu de Café Solúvel, uma das três maiores empresas brasileiras exportadoras de café solúvel [Café Iguaçu 2017].

2. Trabalhos Relacionados e Contextualização

Após revisão da literatura, e do melhor conhecimento dos autores, não foram encontrados artigos ou estudos que abordem a estimação de classes sequenciais voltadas à extração de sólidos solúveis de café. No entanto, existem trabalhos que tratam dos temas separadamente, como o trabalho de Ching, Fung and Ng [2002], que apresenta um modelo de cadeia de Markov multivariada para a modelagem de múltiplas sequências de dados categóricos, abordando processos de Markov de tempo discreto, com estados discretos finitos para a modelagem das sequências de dados. Afirma-se que, se for possível modelar sequências de dados categóricos com precisão, então é possível fazer boas previsões e planejamento ótimo nos processos de decisão.

Kashiwabara et. al. (2013) apresenta o ToPS, um *framework* computacional para análise probabilística de dados sequenciais, o qual implementa oito modelos probabilísticos, dentre eles a cadeia de Markov de alcance variável [Rissanen 1983] e o BIC [Schwarz 1978] como método para auxílio na definição dos melhores parâmetros, assim como na metodologia proposta neste trabalho. Diferenciando-se do ToPS, a metodologia proposta apresenta características próprias com relação ao seu sistema de leitura dos dados, conexão ao banco de dados, medida de similaridade entre os elementos e sistema de pesos de histórico, devido à sua aplicação específica na extração de café solúvel, que motivou o trabalho. No início dos estudos, a metodologia do ToPS chegou a ser aplicada em uma parte dos dados utilizados neste trabalho, a fim de se verificar a viabilidade da predição dos resultados com o uso de cadeias de Markov.

Em se tratando da extração de sólidos solúveis de café, o presente trabalho surgiu da constatação de subutilização do grande volume de dados existentes neste processo. Muitos dados, de diversos tipos de instrumentos de medição e de controle do chão de fábrica, são constantemente coletados, mas pouco é aproveitado na melhoria do

processo ou operação em benefícios ao produto, como por exemplo, rendimento, produtividade e qualidade sensorial. Através da automação industrial, esses dados são registrados por um sistema PIMS (*Process Information Management System*) [Alsmeyer 2006], o qual armazena esses dados em banco de dados por eventos, no caso de processos em bateladas, que são aqueles em que as funções de transferência ou processamento de material são cíclicas com resultados repetíveis [Ribeiro 2001]. Os dados são consultados pelo usuário via relatórios ou historiador gráfico.

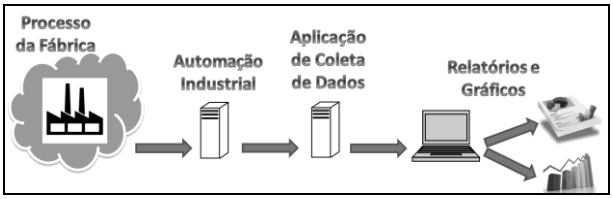

Figura 1. Visão geral do fluxo para coleta dos dados

Dentro do processo de produção do café solúvel, na etapa de extração, objeto de estudo deste trabalho, o café torrado é carregado em colunas extratoras que recebem a circulação de água em altas temperaturas a fim de se obter o extrato de café [Pitchon, Gottesman and Meier 1970]. Tal processo assemelha-se ao que é feito no coador de café doméstico, no qual os grãos torrados e moídos são percolados em água quente.

O processo de extração possui variáveis, ou condições críticas, que afetam as respostas finais do produto, como temperaturas, pressões, pH, concentração e rendimento. Para o setor de Produção, o resultado mais importante no processo de extração é a quantidade de sólidos solúveis produzidos, pois esse resultado determina o rendimento da matéria prima [Zeferino et. al. 2010]. O total de sólidos solúveis produzidos, dado em unidade de peso (kg), em relação à matéria prima consumida pode determinar também o rendimento e condições de operação do equipamento. O peso total de sólidos solúveis (SS) é resultado da multiplicação da concentração do extrato pelo peso total de extrato produzido ao final de cada ciclo.

O produto proveniente do processo de extração contém, em sua maior quantidade, os componentes químicos conhecidos como polissacarídeos, responsáveis pela concentração de sólidos solúveis no extrato de café. Por determinar o rendimento, a concentração de sólidos solúveis no extrato de café é uma das variáveis fortemente controladas e perseguidas no processo de extração de café solúvel [Clark 1985, Clifford 1985], sendo a variável mais indicada para um estudo mais aprofundado, visto que essa é uma variável direta, pura, que não depende de outras variáveis para se obter o seu valor.

3. Materiais e Métodos

3.1. Materiais

Os dados produzidos pelo processo de extração de café solúvel são o alvo deste trabalho. Mais especificamente, os resultados obtidos a cada batelada são adotados para a estimação das probabilidades pela metodologia proposta, a qual pode ser estendida a todos os processos e produtos do equipamento ou de outros equipamentos, nos quais o mesmo conceito se aplica. A obtenção dos dados é feita por conexão direta com o banco de dados do sistema de coleta existente na empresa, via query SQL [Patrick 2009].

Para o desenvolvimento desse trabalho, as informações utilizadas compreenderam o histórico de dados referente a um determinado produto e equipamento, totalizando 29972 registros.

3.2. Método Proposto

Este trabalho apresenta uma metodologia para sumarização de resultados, a qual é baseada no algoritmo Contexto [Leonardi 2006] para estimar uma árvore de sufixo probabilística, utilizada também para prever o resultado do próximo ciclo do processo. Em outras palavras, a metodologia proposta é capaz de realizar tanto a estimação da árvore quanto a predição da classe de resultados ao longo do processo de produção de forma online, isto é, à medida que os resultados do processo são gerados. A Figura 2 apresenta um fluxograma da metodologia proposta, indicando as suas etapas.

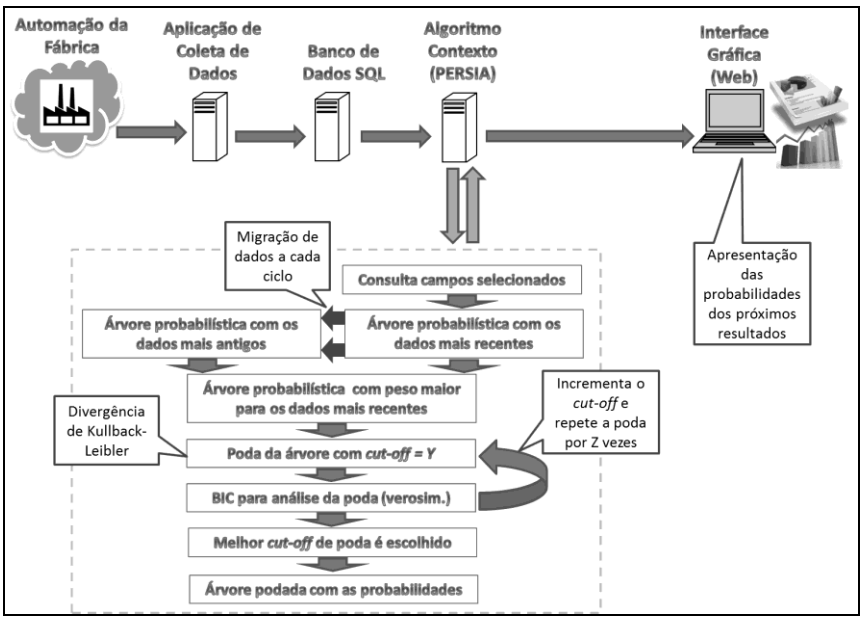

Figura 2. Fluxograma da metodologia proposta

A estrutura da metodologia proposta reconhece cinco rótulos de uma determinada variável, sendo eles: “B”, “b”, “N”, “a” e “A” para “MUITO BAIXO”, “BAIXO”, “NORMAL”, “ALTO” e “MUITO ALTO”, respectivamente. Isso significa que para o algoritmo será usado um alfabeto de cinco caracteres. A faixa considerada “NORMAL” é baseada em informações disponibilizadas pelo setor de Produção, através de técnicas próprias que definem quais são os limites ideais. Foi estipulada, com base na experiência da engenharia de processo da área de Produção, a porcentagem de cinco para acima e abaixo do valor normal para definir os rótulos “ALTO” e “BAIXO”. O que passar de cinco por cento considera-se como “MUITO ALTO” ou “MUITO BAIXO. Os rótulos “ALTO” e “BAIXO” são valores que, apesar de estarem fora dos limites considerados normais, são toleráveis. Acima ou abaixo disso são valores realmente muito discrepantes ou *outliers* [Muñoz-Garcia, Moreno-Rebollo and Pascual-Acosta 1990] que se desviam demais do comportamento esperado.

A Tabela 1 apresenta a definição dos rótulos, onde: “x” é o limite de tolerância máximo, “n1” e “n2” são os limites considerados normais e “y” é o limite de tolerância mínimo.

Tabela 1. Faixas de rótulos

Descrição	Rótulo	Limites de Concentração (%)
MUITO ALTO	A	> x
ALTO	a	> n & <= x (5%)
NORMAL	N	>= n1 & <= n2
BAIXO	b	< n2 (5%) & >= y
MUITO BAIXO	B	< y

A profundidade inicial parametrizada define também o tamanho do alcance do histórico de resultados a ser observado. Por exemplo: em um histórico com a sequência de resultados "BBBNANBNBAN", considerando que a representação do histórico mais recente está à direita, em uma árvore de profundidade quatro, na qual o alcance do histórico também seria 4, a probabilidade de o próximo resultado ser "B" seria estimado observando apenas a sequência "NBAN". Quando um novo resultado chega ao histórico, um "B", por exemplo, a sequência é atualizada, se tornando "BANB".

Também existe a possibilidade de parametrizar o algoritmo para aplicação com apenas três rótulos, sendo eles o "MUITO BAIXO", "NORMAL" e "MUITO ALTO", não havendo, portanto, os rótulos para valores intermediários.

3.2.1. Estrutura da Árvore de Sufixo

A implementação da metodologia proposta foi feita de forma recursiva para construção da estrutura da árvore, onde são criadas todas as combinações possíveis de sequências, sendo que cada elemento possui contadores de ocorrências (*score*) e as probabilidades atuais. Cada elemento armazena a sequência de caracteres a que o elemento pertence.

A Figura 3 apresenta um exemplo simplificado da estrutura da árvore de sufixo, construída em dois níveis (N1 e N2), com uma palavra de três caracteres ("B", "N", "A"). Na raiz (N0) estão os contadores globais de ocorrências de cada caractere, ou seja, toda vez que um resultado novo chega ao histórico o respectivo contador é incrementado, independente do histórico.

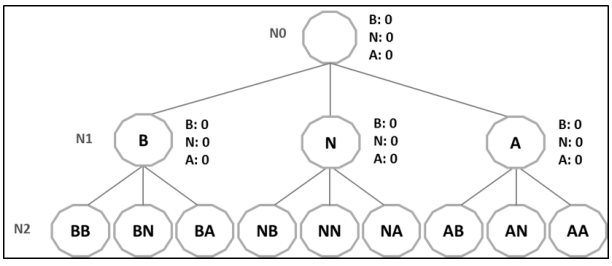

Figura 3. Representação simplificada da criação da árvore de sufixo

Em um exemplo baseado na árvore apresentada da Figura 3, com o histórico começando vazio, a sequência atual seria SEQ= "". Supondo que o primeiro resultado registrado fosse "N", seria incrementado o contador desse caractere na raiz da árvore (N0), já que ainda não existe nenhum histórico a ser observado. "N" é acrescentado à sequência de histórico, ficando agora SEQ="N". Considerando o próximo resultado

como “B”, é incrementado agora o contador de “B” na raiz (N0) e também no nível um (N1) no nó “N”, dado que o histórico anterior foi “N”. Para um próximo resultado “A”, com a sequência atualizada SEQ=”NB”, o caractere “A” é contabilizado então em N0, N1 (nó B) e N2 (nó BN). A Figura 4 exibe como estaria o preenchimento dos contadores da árvore neste ponto.

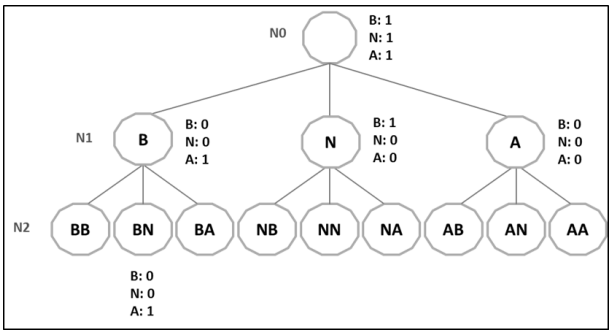

Figura 4. Exemplo de preenchimento da árvore de sufixo baseado no histórico de resultados

Sendo BN um nó folha da árvore, ou seja, um elemento que não possui nós filho, os seus contadores são utilizados para cálculo das probabilidades do próximo resultado. A Figura 5 apresenta um exemplo mais completo deste nó, após diversas iterações do algoritmo, onde 205 ocorrências foram registradas no nó “BN”, distribuídas entre os contadores de cada caractere da palavra do algoritmo. Ao lado direito dos contadores estão as probabilidades para o próximo resultado, dado que o histórico anterior foi “BN”, sendo “B” o resultado com maior probabilidade.

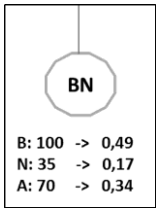

Figura 5. Exemplo de probabilidades calculadas a partir de um nó folha

A Figura 6 apresenta uma amostra de impressão da árvore de sufixo, na qual se observa a representação dos níveis da árvore por meio da identificação no início de cada linha, onde são apresentados os contadores e probabilidades de ocorrência de cada caractere. Os nós folha são indicados pelo “<F>” no final da linha.

```
N0:  -> [ B:1974(0,02) b:3599(0,12) N:12563(0,57) a:8646(0,22) A:3200(0,07) ]
   N1: B -> [ B:1164(0,25) b:485(0,18) N:247(0,29) a:61(0,14) A:25(0,14) ]
      N2: BB -> [ B:812(0,30) b:229(0,20) N:93(0,18) a:23(0,16) A:15(0,16) ]
         N3: BBB -> [ B:595(0,31) b:140(0,19) N:57(0,17) a:15(0,16) A:13(0,16) ]
            N4: BBBB -> [ B:459(0,31) b:90(0,19) N:37(0,17) a:6(0,16) A:11(0,16) ] <F>
            N4: BBBb -> [ B:70(0,28) b:31(0,21) N:12(0,18) a:7(0,17) A:2(0,16) ] <F>
            N4: BBBN -> [ B:38(0,28) b:13(0,20) N:5(0,17) a:6(0,18) A:4(0,17) ] <F>
            N4: BBBa -> [ B:23(0,28) b:9(0,20) N:7(0,19) a:2(0,17) A:2(0,17) ] <F>
            N4: BBBA -> [ B:13(0,27) b:5(0,20) N:4(0,19) a:2(0,17) A:2(0,17) ] <F>
         N3: BBb -> [ B:114(0,28) b:48(0,21) N:19(0,18) a:5(0,16) A:3(0,16) ]
            N4: BBbB -> [ B:43(0,30) b:12(0,20) N:3(0,17) a:3(0,17) A:2(0,16) ] <F>
            N4: BBbb -> [ B:37(0,29) b:11(0,20) N:7(0,18) a:3(0,17) A:2(0,16) ] <F>
            N4: BBbN -> [ B:25(0,25) b:17(0,22) N:9(0,19) a:3(0,17) A:3(0,17) ] <F>
            N4: BBba -> [ B:8(0,23) b:9(0,24) N:5(0,20) a:2(0,17) A:2(0,17) ] <F>
            N4: BBbA -> [ B:9(0,25) b:7(0,23) N:3(0,18) a:2(0,17) A:2(0,17) ] <F>
         N3: BBN -> [ B:58(0,27) b:27(0,21) N:10(0,18) a:7(0,17) A:3(0,16) ]
            N4: BBNB -> [ B:8(0,26) b:3(0,19) N:3(0,19) a:2(0,17) A:3(0,19) ] <F>
```

Figura 6. Amostra da impressão da árvore de sufixo

3.2.2. Processo de Poda da Árvore de Sufixo

A partir da estrutura da árvore de sufixo foi desenvolvido o sistema de poda da árvore, no qual, basicamente, elementos muito semelhantes ao elemento pai devem ser descartados. A medida de similaridade implementada foi a Divergência de Kullback-Leibler [Duda, Hart and Stork 2000a]. A partir das probabilidades do elemento pai e de seus filhos, essa técnica retorna um valor que informa o quão semelhantes ou quão próximos os filhos estão dos pais. A partir de um parâmetro de *cut-off* é feita a decisão pela poda ou não desses elementos filho. Também são removidos nós com baixo número de ocorrências.

Exemplo 1:

PAI		F1			F2			F3			F4			F5		
Q	P(%)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)
20	20,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0
20	20,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0
20	20,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0
20	20,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0
20	20,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0	1	20,0	0,0
			KL:	**0,0**		**KL:**	**0,0**		**KL:**	**0,0**		**KL:**	**0,0**		**KL:**	**0,0**

Exemplo 2:

PAI		F1			F2			F3			F4			F5		
Q	P(%)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)
280	75,1	219	77,1	-2,0	14	56,0	22,0	25	58,1	19,2	17	63,0	13,2	9	64,3	11,6
25	6,7	16	5,6	1,2	3	12,0	-3,9	5	11,6	-3,7	3	11,1	-3,4	2	14,3	-5,1
34	9,1	25	8,8	0,3	4	16,0	-5,1	5	11,6	-2,2	3	11,1	-1,8	1	7,1	2,2
22	5,9	13	4,6	1,5	3	12,0	-4,2	6	14,0	-5,1	3	11,1	-3,7	1	7,1	-1,1
12	3,2	11	3,9	-0,6	1	4,0	-0,7	2	4,7	-1,2	1	3,7	-0,5	1	7,1	-2,6
			KL:	**0,4**		**KL:**	**8,1**		**KL:**	**7,0**		**KL:**	**3,8**		**KL:**	**5,1**

Exemplo 3:

PAI		F1			F2			F3			F4			F5		
Q	P(%)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)	Q	P(%)	ln(x)
21	48,8	10	31,3	21,8	9	69,2	-17,0	1	20,0	43,6	3	42,9	6,4	2	33,3	18,7
7	16,3	7	21,9	-4,8	1	7,7	12,2	1	20,0	-3,4	1	14,3	2,1	1	16,7	-0,4
11	25,6	11	34,4	-7,6	1	7,7	30,7	1	20,0	6,3	1	14,3	14,9	1	16,7	11,0
3	7,0	3	9,4	-2,1	1	7,7	-0,7	1	20,0	-7,3	1	14,3	-5,0	1	16,7	-6,1
1	2,3	1	3,1	-0,7	1	7,7	-2,8	1	20,0	-5,0	1	14,3	-4,2	1	16,7	-4,6
			KL:	**6,7**		**KL:**	**22,4**		**KL:**	**34,2**		**KL:**	**14,2**		**KL:**	**18,6**

Quadro 1. Exemplo de aplicação da Divergência de Kullback-Leibler

O Quadro 1 apresenta uma demonstração da verificação de semelhança entre elementos pai e elementos filho aplicando Divergência de Kullback-Leibler. Foram montados três exemplos através de amostras reais dos dados coletados, onde:

- “Q” indica quantas ocorrências obteve cada caractere do alfabeto, num total de cinco caracteres, seguindo a ordem: “b”, “B”, “N”, “a” e “A”;
- “P(%)” indica a probabilidade de cada caractere, em porcentagem;
- “ln(x)” apresenta o cálculo parcial, dado pela equação:

$$D_{KL(parcial)} = PAI(i) log \frac{FILHO(i)}{PAI(i)}$$

- “KL” é o resultado final da divergência, sendo a somatória dos “ln(x)” calculados para cada caractere. Esse valor indica o quanto esse elemento filho é divergente do pai, sendo que, quanto maior o valor, maior é a divergência.

No Exemplo 1 todos os filhos tem os mesmos valores do pai, por isso os resultados são 0. Nesse caso, o algoritmo poda todos os elementos, pois apenas os valores do pai são suficientes, visto que possuem as mesmas informações. No Exemplo 2 o filho mais divergente é o F2, dado que obteve o maior valor no cálculo de divergência. A diferença pode ser notada através da comparação das suas probabilidades com as do

pai. Já o F1 possui baixa divergência, portanto pode ser podado, dependendo do valor informado no parâmetro de *cut-off*. No Exemplo 3 o elemento mais divergente é o F3. Apesar de possuir valores muito baixos ele é muito divergente do elemento pai, portanto é preservado. Assim, elementos com baixas probabilidades só serão podados quando o elemento pai também possuir essa característica.

O processo de poda só ocorre a partir de elementos pai nos quais todos os filhos são folhas. Ou seja, se um dos filhos não for folha, todos os filhos devem ser preservados. A seguir é apresentado um exemplo do processo de poda, com uma árvore de 3 níveis, onde os elementos são representados por números para facilitar a didática.

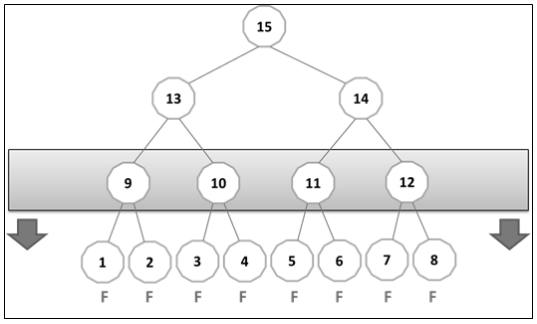

Figura 7. Verificação inicial no exemplo de poda da árvore de sufixo

A Figura 7 apresenta a verificação do nível 2. Caso satisfaçam os critérios de poda, todos os elementos pai podem ter os filhos removidos, pois todos são folhas.

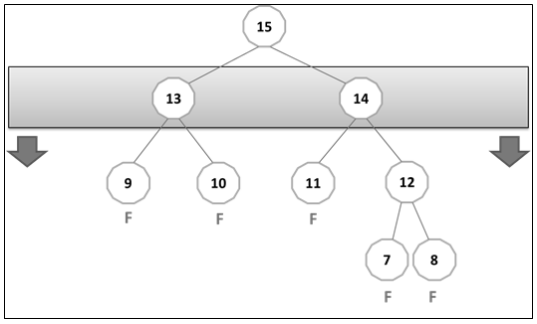

Figura 8. Primeira poda no exemplo de poda da árvore de sufixo

Como exemplo, assumimos que todos os elementos pai, com exceção do elemento 12, tiveram os filhos podados, devido ao critério de divergência. Na verificação do nível 1 (Figura 8), apenas o elemento 13 pode ter os filhos podados, pois todos são folhas. Já o 14 não pode, mesmo que atender aos critérios de poda.

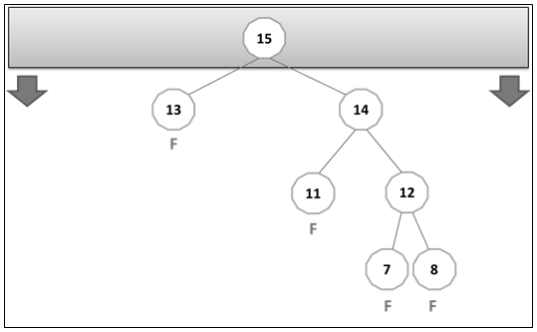

Figura 9. Poda final no exemplo de poda da árvore de sufixo

Assumindo que algum dos critérios de poda foi satisfeito, os filhos do elemento 13 acabam sendo podados (Figura 9). Na verificação do nível 0 nenhuma poda ocorrerá, pois o elemento 14 não é folha.

Na fase de desenvolvimento do sistema, o parâmetro de *cut-off* foi testado inicialmente com valores fixos, sendo necessário informar manualmente diferentes valores para testes. Posteriormente, o algoritmo passou a utilizar o BIC (Bayesian Information Criterion) [Schwarz 1978], técnica que retorna o melhor valor de *cut-off* para poda, independente da técnica utilizada como critério para poda dos elementos, tendendo a deixar os resultados mais precisos.

3.2.3. Pesos do Histórico de Probabilidades

Na indústria, os ajustes de processos para a melhoria contínua são realizados com frequência. Assim, verificou-se a necessidade de que dados antigos, resultantes da coleta realizada no decorrer dos anos, tenham pesos diferentes em relação aos atuais. Dessa maneira, acredita-se que os dados atuais tenham respostas melhores no processo que os dados de tempos atrás. Para resolver essa questão foram implementadas três árvores probabilísticas, onde a primeira contém o histórico mais recente dos registros, com alcance parametrizado. A segunda árvore contém todo o histórico restante, mais antigo. Já a terceira árvore é uma união das duas anteriores, porém com pesos diferentes para as suas probabilidades, como peso 90 para o histórico atual e 10 para o restante, por exemplo. A implementação do sistema de pesos para o histórico de probabilidades se mostra importante para que o algoritmo se adapte a qualquer tipo de variação no processo ao longo do tempo, visto que os dados mais atuais sempre serão reforçados em prioridade, mas ainda assim aproveitando os antigos como referência de aprendizado.

4. Resultados e Discussão

A saída da metodologia proposta grava os resultados das probabilidades em arquivos CSV, permitindo o desenvolvimento de interfaces independentes para a visualização das probabilidades pelo operador do equipamento, atualizadas a cada novo ciclo.

A Figura 10 apresenta um exemplo de exibição das probabilidades em uma interface gráfica, onde “B”, “b”, “N”, “a” e “A” são abreviações para “MUITO BAIXO”, “BAIXO”, “NORMAL”, “ALTO” e “MUITO ALTO”, respectivamente) e suas respectivas probabilidades de que ocorram no próximo ciclo. Nesse exemplo observa-se que há maior probabilidade (0,36) de que o próximo resultado da variável analisada esteja na faixa considerada “NORMAL”, seguido de uma probabilidade de 0,18 de que seja “BAIXO”. Dessa maneira, os resultados poderão ser visualizados pelo operador e transformados em ações para corrigir ou manter o processo sob controle.

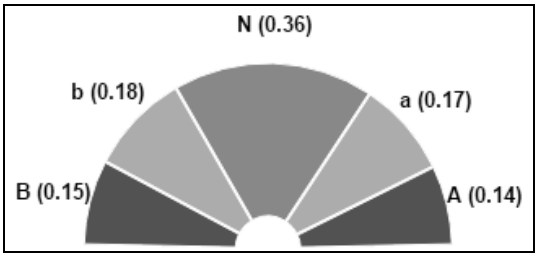

Figura 10. Exemplo de apresentação gráfica das probabilidades do próximo resultado

Dada a dificuldade em se apresentar uma sequência completa de probabilidades em uma árvore de sufixo, devido à grande quantidade de combinações, é apresentada a seguir uma sequência com uma amostra de 520 registros e apenas três rótulos: “B”, “N” e “A”, com uma árvore de sufixo de ordem um. Inicialmente são dadas as probabilidades globais de cada classe e depois as probabilidades de transições entre elas, com todas as combinações possíveis. Cada linha de saída é formada por “ESTADO” | “HISTÓRICO”. Por exemplo, “N | A: 0,34” indica 34% de chances de que de o próximo resultado de concentração seja “NORMAL”, dado que o resultado anterior foi “ALTO”.

Tabela 2. Exemplo simplificado de probabilidades estimadas

Sequências	Probabilid.	Descrição das Probabilidades do Próximo Resultado
A	0,41	Prob. de concentração alta (global)
B	0,15	Prob. de concentração baixa (global)
N	0,44	Prob. de concentração normal (global)
A | A	0,63	Prob. de concentração alta dado que a anterior foi alta
B | A	0,03	Prob. de concentração baixa dado que a anterior foi alta
N | A	0,34	Prob. de concentração normal dado que a anterior foi alta
A | B	0,07	Prob. de concentração alta dado que a anterior foi baixa
B | B	0,55	Prob. de concentração baixa dado que a anterior foi baixa
N | B	0,38	Prob. de concentração normal dado que a anterior foi baixa
A | N	0,33	Prob. de concentração alta dado que a anterior foi normal
B | N	0,13	Prob. de concentração baixa dado que a anterior foi normal
N | N	0,54	Prob. de concentração normal dado que a anterior foi normal

Se, por exemplo, a concentração atual é alta, podemos estimar uma situação em que raramente a concentração será baixa no próximo ciclo, devido à baixa probabilidade estimada (0,03). Em outro exemplo, se a concentração for baixa então há alta probabilidade (0,55) de que a concentração volte a ser baixa no próximo ciclo, o que pode significar, por exemplo, que as colunas estejam trabalhando de forma subótima. Os eventos de baixa probabilidade são eventos raros. Tais eventos são interessantes para a interpretação dos resultados, principalmente no que diz respeito às razões pelas quais ocorreram. Por exemplo, as transições de “ALTO” para “BAIXO” e de “BAIXO” para “ALTO” podem ser erros na medição ou só ocorrem quando há algum problema na máquina ou ainda quando há uma intervenção do operador.

Ao longo dos ciclos, a máquina perde ou ganha gradativamente sua rentabilidade. No funcionamento considerado normal, a mudança de rentabilidade da máquina não pode ser brusca. Se, por exemplo, entrarem grãos de café de baixa qualidade no sistema, esses grãos irão afetar o processo até que eles sejam descartados. Assim, se a quantidade de grãos de baixa qualidade estiver na maioria das colunas é provável que ocorra baixa rentabilidade durante alguns ciclos. Por outro lado, ao entrar grãos de alta qualidade, esses grãos irão afetar positivamente na rentabilidade, até que eles sejam descartados.

5. Conclusões e Trabalhos Futuros

Com o objetivo de extrair conhecimento operacional dos dados existentes do processo de "Extração de Sólidos Solúveis", este trabalho apresentou uma metodologia de sumarização de resultados pela implementação de árvores probabilísticas de sufixo [Leonardi 2006, Largeron 2003], utilizando o histórico de observações dos resultados na estimação de probabilidades de ocorrência de cada classe destes. A previsão dos resultados, tal como, a concentração, faz com que o rendimento e a produtividade da linha sejam resultados estimados indiretamente. Na extração de café solúvel, podemos verificar que a estimação dos resultados trará benefícios para o conhecimento e desenvolvimento de processo.

A implementação da metodologia em um algoritmo demonstrou sua importância para análise de problemas, observação das características do equipamento, melhorias nas ações do operador, entre outras. Trata-se de um recurso adicional na busca da melhoria contínua de processos e produtos, uma ferramenta de apoio para a tomada de ações no processo de produção, obtendo-se mais estabilidade e repetibilidade nas respostas.

Atualmente, todos os parâmetros do algoritmo precisam ser alterados e testados manualmente. Utilizando técnicas de otimização, seria possível estimar melhores valores para os parâmetros [Duda, Hart and Stork 2000b], conseguindo assim mais eficiência nos resultados de estimação e também no desempenho de execução do algoritmo. Dentre as diversas técnicas existentes para otimização ou estimação de parâmetros podemos citar a implementação de algoritmos evolucionários [Linden 2012], como os Algoritmos Genéticos e as Redes Neurais.

Por se tratar de uma metodologia baseada em histórico de resultados, certamente poderá ser aplicada em outras etapas do processo, principalmente aquelas baseadas em bateladas [Ribeiro 2001], como a torrefação, onde hoje já existe uma coleta de dados de saída, como umidade, cor, rendimento e tempo de torra. Mesmo em outras etapas seria possível a aplicação do sistema de predição, bastando identificar as variáveis de saída mais importantes e estabelecer as suas combinações no histórico de resultados.

6. Agradecimentos

Agradecemos à Café Iguaçu por proporcionar as condições para realização deste trabalho com sua moderna infraestrutura.

Referências

Alsmeyer, F. (2006) "Automatic Adjustment of Data Compression in Process Information Management Systems", Computer Aided Chemical Engineering, v.21, p.1533-1538, Aachen, Germany.

Café Iguaçu (2017) "História", http://www.iguacu.com.br/empresa/sobre-nos/, Março.

Ching, W., Fung, E. S. and Ng, M. K. (2002) "A Multivariate Markov Chain Model for Categorical Data Sequences and Its Applications in Demand Predictions", In: IMA Journal of Management Mathematics, p.187-199, Hong Kong.

Clarke, R. J. (1985) "Water and Mineral Contents". In: Clarke, R. J., Macrae, R. "Coffee: Chemistry", Elsevier Applied Science Publishers, v.1, p.42-82, London.

Clifford, M. N. (1985) “Chemical and Physical Aspects of Green Coffee and Coffee Products”. In: Coffee: Botany, Biochemistry and Production of Beans and Beverage, p.305-374, London: M. Chapman and Hall.

CNI, Confederação Nacional da Indústria (2017) “Alimentos e Bebidas”, http:// http://www.portaldaindustria.com.br/agenciacni/noticias/2014/09/alimentos-e-bebidas-1/, Fevereiro.

De Souza, A. J., Bezerra, C. G., De Andrade, W. L. S., Feijo, R. H.; Leitao, G. B. P., Guedes, L. A., Maitelli, A. L., De Medeiros, A. A. D. (2005) “Gerência de Informação da Produção de Petróleo e Gás”, In: 3º Congresso Brasileiro de P&D em Petróleo e Gás. Salvador, Bahia.

Duda, R. O., Hart, P. E. and Stork, D. G. (2000a) “Pattern Classification”, Second Edition, In: Wiley-Interscience, c.3, p.57.

Duda, R. O., Hart, P. E. and Stork, D. G. (2000b) “Pattern Classification”, Second Edition, In: Wiley-Interscience, c.3, p.3.

Kashiwabara, A. Y., Bonadio, Í., Onuchic, V., Amado, F., Mathias, R., Durham, A. M. (2013) “Tops: A Framework to Manipulate Probabilistic Models of Sequence Data”, In: PLOS: Computational Biology.

Largeron, C. (2003) “Prediction Suffix Trees for Supervised Classification of Sequences”, In: Journal Pattern Recognition Letters, v.24, p.3153-3164.

Leonardi, F. G. (2006) “A Generalization of the PST Algorithm: Modeling the Sparse Nature of Protein Sequences”, In: Bioinformatics, v.22, n.11, p.1302-1307.

Linden, R. (2012) “Algoritmos Genéticos”, 3ª Edição, Editora Ciência Moderna, Rio de Janeiro, p.43.

Patrick, J. J. (2009) “SQL Fundamentals”, Third Edition, Pearson Education, USA, p.3.

Pitchon, E., Gottesman, M. and Meier, R. W. (1970) “Process for Manufacture of Coffee Extract”, United States Patent, General Foods Corporation, New York.

Muñoz-Garcia, J., L. Moreno-Rebollo, J., Pascual-Acosta, A. (1990) “Outliers: A Formal Approach”, In: International Statistical Review, v.58, n.3, p.215-226.

Ribeiro, M. A. (2001) “Automação Industrial”, 4 ed, Salvador: Tek Treinamento & Consultoria Ltda.

Rissanen, J. (1983) “A Universal Data Compression System”, In: IEE Transactions on Information Theory, v.29, n.5, p.656-664.

Santos, A. F. S. (2014) “Métodos Facilitadores de Melhoria do Processo e Aumento de Produtividade”, Instituto de Educação Tecnológica - IETEC.

Schwarz, G. (1978) “Estimating the Dimension of a Model”, In: The Annals of Statistics, v.6, n.2, p.461-464.

Zeferino, L. B., Saraiva, S. H., Silva, L. C, Teixeira, L. J. Q., Lucia S. M. D. (2010) “Efeito da Concentração de Sólidos Solúveis do Extrato de Café Conilon no Índice de Refração, na Densidade e na Viscosidade do Extrato”, In: Enciclopédia Biosfera, Centro Científico Conhecer, v.6, n.11, p.1, Goiânia.