Agrupamento de Fornos de Redução de Alumínio Utilizando Self Organizing Map

Alan Marcel Fernandes de Souza[1], Flávia A. N. de Lima[1], Fábio M. Soares[1], Roberto C. L. de Oliveira[2]

[1] Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal do Pará, Brasil.
[2] Faculdade de Engenharia Elétrica e Computação, Universidade Federal do Pará, Brasil.
alanmarcel2@gmail.com

Abstract. This paper is about clustering aluminum reduction potlines using real data from a production plant of this metal. To achieve this, the algorithm Self Organizing Map was used successfully. This clustering helps to discover hide rulers inside data. Besides, this knowledge that has been acquired may be used to create virtual simulators so that there is no need to run tests on the real plant, avoiding expenses and accidents.

Keywords: clustering, real data, Self Organizing Map.

1 Introdução

A produção de alumínio em larga escala acontece através do processo mundialmente conhecido como Hall-Héroult [1]. Nas indústrias, há centenas de fornos que, em suma, recebem a matéria-prima alumina (Al2O3) e altas cargas de corrente elétrica para quebrar a molécula desta matéria-prima são aplicadas nos fornos, produzindo o alumínio e gás carbônico [2]. Este processo é ininterrupto, ou seja, funciona 24 horas por dia, sete dias na semana, 365 dias ao ano.

A Fig. 1 mostra uma a disposição das salas que contém os fornos de redução de alumínio da fábrica que este trabalho foi baseado. Verifica-se que existem oito salas com 120 fornos cada uma, totalizando 960 fornos.

Fig. 1. Layout completo da fábrica de alumínio.

O objetivo deste trabalho é realizar o agrupamento (*clustering*) dos fornos de redução de alumínio - levando em consideração dados reais disponibilizados por uma fábrica produtora deste metal - através da técnica de inteligência computacional conhecida como SOM (*Self Organizing Map* - Mapa Auto-Organizável). Após o agrupamento, será possível descobrir regras associativas, escondidas nos dados, que ajudam a explicar o processo de produção de alumínio realizado pela fábrica.

Este trabalho está dividido em quatro seções. A primeira introduz a ideia de mineração de dados e o seu uso na indústria de alumínio primário. A segunda apresenta como foram feitas a seleção e o pré-processamento dos dados. A terceira detalha como o agrupamento foi realizado, exibindo os resultados alcançados. Finalmente, a quarta seção apresenta a conclusão.

2 Seleção e pré-processamento dos dados

A produção de alumínio é um processo complexo e requer monitoramento contínuo. Neste sentido, os fornos que constituem a fábrica são equipados com sensores que capturam dados e os armazenam em banco de dados. Mais de 130 variáveis são monitoradas.

É impraticável considerar todas as variáveis do processo para o agrupamento. Sendo assim, as variáveis consideradas foram as mesmas utilizadas por [3], que realizou o estudo para construir um estimados de temperatura para o forno. São elas:

a) Temperatura (TMP);
b) Fluoreto de alumínio (% de ALF no Banho);
c) Quantidade de fluoreto adicionado no banho (ALF3A);
d) Quantidade de alumina alimentada (QALr);
e) Incremento de resistência por temperatura (IncTM);
f) Percentual de tempo em alimentação under-feeding (%TUN);
g) Percentual de tempo em alimentação over-feeding (%TOV).

A partir da grande quantidade de amostras mantidas pela fábrica, dois conjuntos de dados diferentes foram criados:

- Com Filtro;
- Sem Filtro.

O conjunto de dados “sem filtro” leva em consideração somente as amostras que possuem valor diferente de zero e que não são nulos (*null*). A Tabela I exibe a quantidade de dados por ano. Nota-se que este conjunto de dados é composto por mais de dois milhões de amostras no total.

Tabela 1. Quantidades de registros por ano (sem filtro)

Variável	Quantidade
2006	348269
2007	347951
2008	348009
2009	346375
2010	346827
2011	348614
2012	216092

Outro conjunto de dados utilizado foi o "com filtro". Os dados contidos nele são aqueles que então dentro das faixas de valores, para cada uma das sete variáveis, mostrados na Tabela 2. Os padrões que não estão dentro do intervalo especificado foram descartados.

Tabela 2. Faixa de valores "com filtro" por variável

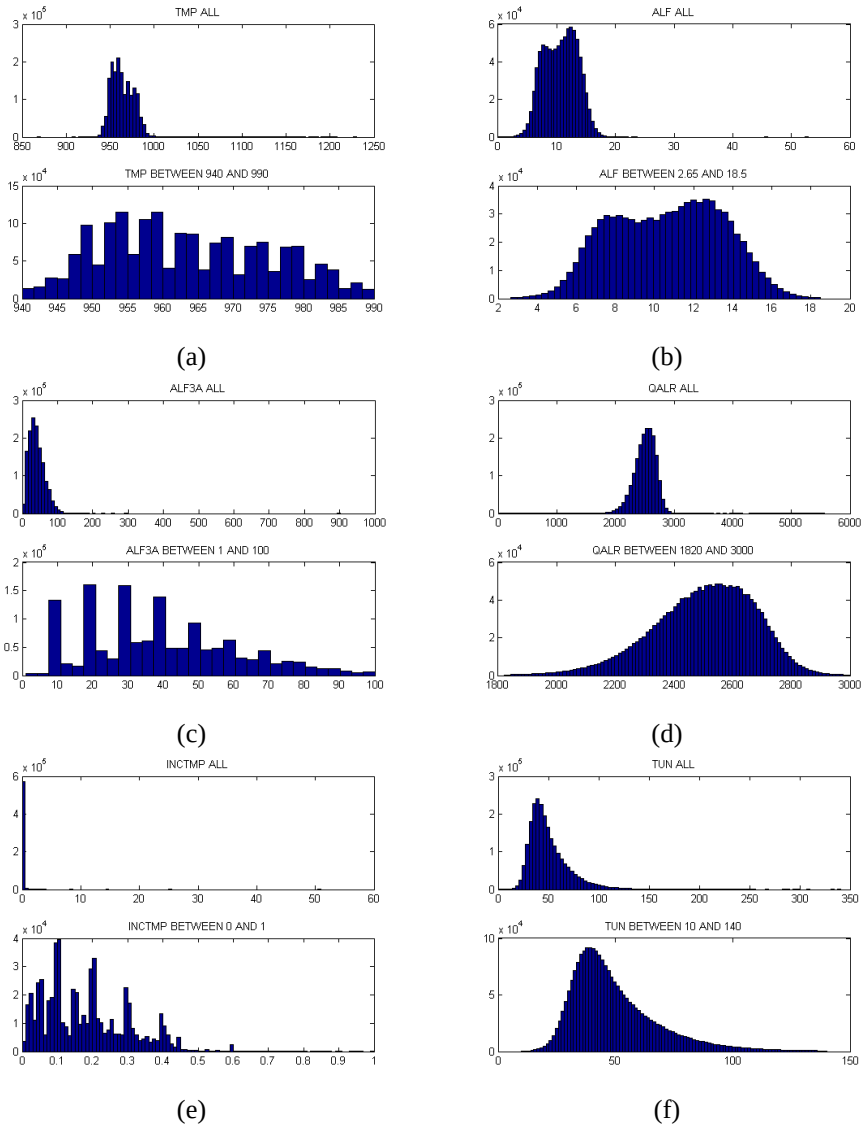
Variável	Faixa de valor
TMP	Entre 940 e 990 ºC
% de ALF no banho	Entre 2,65 e 18,5
ALF3A	Entre 1 e 100
QALr	Entre 1820 e 3000
IncTM	Entre -1 e 1
%TUN	Entre 10 e 140
%TOV	Entre 20 e 80

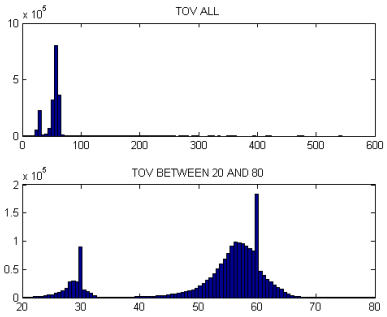
A Tabela 3 mostra a quantidade de dados por ano, levando em consideração a filtragem. Observa-se que o conjunto de dados reduziu para aproximadamente 340 mil registros no total.

Tabela 3. Quantidades de registros por ano (com filtro)

Variável	Faixa de valor
2006	38096
2007	66711
2008	57603
2009	55642
2010	62995
2011	23539
2012	38096

Histogramas foram gerados para visualizar algumas informações estatísticas dos dados para cada variável sem e com filtro. Ao realizar as comparações entre os dois conjuntos de dados diferentes, verifica-se, através da Fig. 2, que os dados com filtro possuem um intervalo de valores menor que os sem filtro, contribuindo para uma melhor distribuição de valores. Isto tende a facilitar o agrupamento dos dados.





(g)

Fig. 2. Histogramas de cada variável do conjunto de dados sem filtro e com filtro. (a) Variável TMP. (b) Variável ALF. (c) Variável ALF3A. (d) Variável QALr. (e) Variável IncTM. (f) Variável %TUN. (g) Variável %TOV.

Além de estabelecer dois tipos diferentes de conjunto de dados, três cálculos estatísticos foram utilizados para realizar o agrupamento: média, mediana e desvio padrão. Isto resultou em seis combinações diferentes de experimentos elencados na Tabela 4. Sabe-se que ao calcular a média de dados, o resultado final é muito influenciado pelos dados *outliers*. Por outro lado, o resultado proveniente da mediana quase não sofre influência dos *outliers*. A exclusão dos *outliers* tende a facilitar o processo de agrupamento. Todos os experimentos utilizam três *clusters* e medida da distância euclidiana para definir os agrupamentos.

Tabela 4. Experimentos realizados

Experimento	Cálculo Estatístico	Filtro
#1	Média	Com filtro
#2		Sem filtro
#3	Mediana	Com filtro
#4		Sem filtro
#5	Desvio Padrão	Com filtro
#6		Sem filtro

3 Agrupamento através de SOM

A clusterização de dados ou análise de agrupamentos é uma prática de mineração de dados que tem por objetivo encontrar similaridades entre as *n* amostras da base dados, usando algoritmo de aprendizado não-supervisionado. No final do processo, *k* grupos, também conhecidos como *clusters*, são identificados [4].

O método de clusterização é definido por um algoritmo específico que determina como será feita a divisão dos dados nos *clusters* distintos e todos os métodos propostos são fundamentados na ideia de distância ou similaridade entre as observações e

definem a pertinência dos objetos a cada *cluster* segundo aquilo que cada elemento tem de similar em relação a outros pertencentes do grupo.

Os elementos que compõem um mesmo *cluster* devem apresentar alta similaridade (i.e., tenham elementos bem parecidos, seguindo um padrão similar), mas devem ser muito dissimilares de objetos de outros *clusters*. Em outras palavras, o agrupamento é feito com objetivo de maximizar a homogeneidade dentro de cada grupo e maximizar a heterogeneidade entre os grupos [7].

Ao agrupar dados similares, pode-se descrever de forma mais eficiente e eficaz as características peculiares de cada um dos grupos identificados. Isso fornece um maior entendimento do conjunto de dados original, além de possibilitar o desenvolvimento de esquemas de classificação para novos dados e descobrir correlações interessantes entre os atributos dos dados que não seriam facilmente visualizadas sem o emprego de tais técnicas.

Em 1982 Teuvo Kohonen apresentou um modelo de rede denominado SOM, com base em determinadas evidências descobertas a nível cerebral. Este tipo de rede possui um aprendizado não-supervisionado competitivo [6].

Um SOM é um conjunto de vetores de dimensão *n* normalmente distribuídos em uma pequena rede (retícula) bidimensional (mesmo que nada impeça que sejam distribuídos em 3 ou mais dimensões). Para cada vetor (que também é chamado "nó" ou "neurônio") é definido uma vizinhança: cada vetor pode ter oito vizinhos (isto é, uma retícula retangular) ou 6 (retícula hexagonal). Existem razões teóricas para preferir uma ou outra; enquanto que a mais popular é a retangular, a hexagonal tem uma base teórica mais sólida [7].

Para treinar um mapa auto-organizável, primeiro é necessário que se tenha um conjunto de dados, que serão divididos em três conjuntos: treinamento, teste e validação. O treinamento, como o próprio nome infere, fará o treinamento do mapa em análise, o teste é feito para selecionar um mapa entre vários, e posteriormente ocorre a validação, cujo o objetivo é definir o erro final. Para definir o erro total de um mapa com determinados parâmetros, é possível usar outra metodologia diferente chamada deixar–k–fora (*leave–k–out*), a qual consiste em dividir o conjunto de dados em *k* partes, onde se usa k-1 para treinamento e 1 para teste; o procedimento se repete para as *k* partes em que se divide o conjunto de treino.

Neste trabalho, a métrica utilizada para determinar o neurônio vencedor foi "o vencedor leva tudo" (*winner-take-all*). A adaptação do parâmetro de aprendizagem foi controlada pela minimização da distância euclidiana, onde o neurônio vencedor é aquela que apresenta o menor valor de distância euclidiana.

A topologia da rede neural utilizada pode ser vista por intermédio da Fig. 3. Nota-se que há sete neurônios na camada de entrada, onde cada neurônio representa uma variável do processo. Além disso, é possível identificar que há nove neurônios na camada de saída, os quais representam os grupos de fornos a serem encontrados. Outras topologias foram testadas, mas a que a apresentou melhores resultados foi esta.

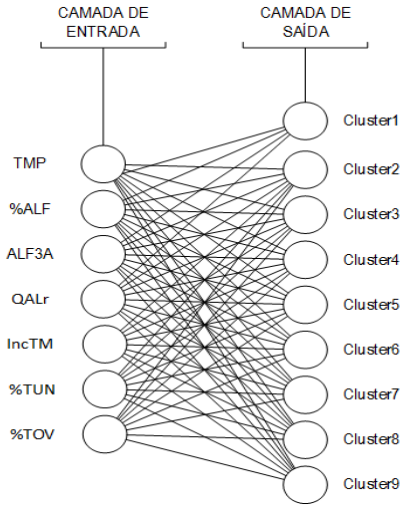


Fig. 3. Topologia da Rede Neural.

3. 1 Resultados do agrupamento

Os agrupamentos foram encontrados por meio de rotinas computacionais programadas no software R Studio (versão 1.0.44), usando o R (versão 3.2-3.4) [8].

O gráfico representado pela Fig. 4 a seguir, mostra a relação entre a distância média para o grupo mais próximo, à medida que as iterações acontecem. Verifica-se que após a iteração 200, as distâncias permanecem no mesmo patamar (abaixo de 0,02) até a última iteração, demonstrando uma distância média pequena.

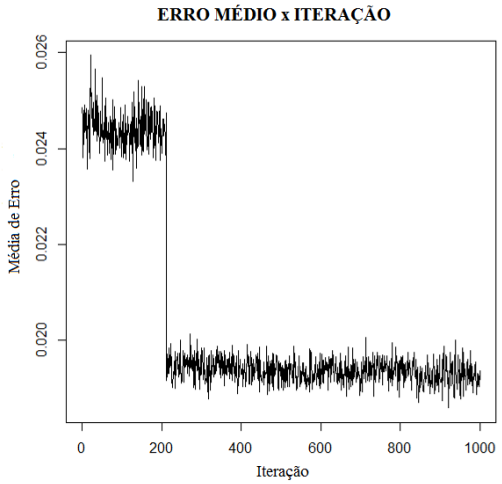


Fig. 4. Evolução do Erro Médio Quadrático por Iteração.

O principal resultado a ser analisado é o gráfico que mostra os grupos encontrados pelo SOM (Fig. 5). Cada círculo representa um grupo de fornos e, dentro deles, há um gráfico de pizza que mostra o nível de determinada variável dentro do cluster. No primeiro grupo, é possível verificar que existe grande influência das variáveis de cor amarela, laranja e verde escuro (ALF3A, QALr, TMP); pequena influência das variáveis representadas pela cor verde claro, laranja claro e rosa (ALF, IncTM, TUN) e ausência da variável de cor cinza (TOV).

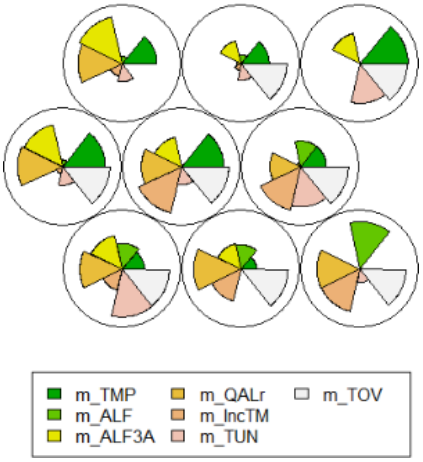


Fig. 5. Grupos de fornos e respectivos níveis de influência.

A Fig. 5 exibe círculos, onde cada um representa um grupo e as fatias representam a intensidade de cada variável existente em cada um dos grupos. A partir da análise feita para o primeiro grupo, foi decidido usar o valor "Alto" quando o tamanho da fatia é grande e acima do valor médio; valor "Baixo" quando o tamanho é pequeno e abaixo do valor médio e valor "Nulo" quando a fatia não aparece no gráfico. Após a análise dos resultados, a Tabela 5 foi gerada, como mostrado abaixo:

Tabela 5. Sumário dos níveis de influência de cada variável nos clusters de fornos.

Grupo	TMP	ALF	ALF3A	QALr	IncTM	TUN	TOV
1	Alto	Baixo	Alto	Alto	Baixo	Baixo	Nulo
2	Alto	Baixo	Baixo	Nulo	Baixo	Baixo	Alto
3	Alto	Nulo	Alto	Nulo	Nulo	Alto	Alto
4	Alto	Baixo	Alto	Alto	Baixo	Baixo	Alto
5	Alto	Nulo	Alto	Alto	Alto	Baixo	Alto
6	Baixo	Baixo	Nulo	Alto	Alto	Alto	Alto
7	Baixo	Baixo	Alto	Alto	Baixo	Alto	Alto
8	Baixo	Baixo	Baixo	Alto	Alto	Nulo	Alto
9	Nulo	Alto	Nulo	Alto	Alto	Baixo	Alto

A quantidade de fornos por grupo e a qualidade do grupo podem ser analisadas através das Fig. 6. Em relação às quantidades, quanto mais escuro é o grupo, menos fornos estão contidos nele. Por outro lado, sobre a qualidade, quanto mais escuro é o grupo, melhor é a qualidade do mesmo, já que a distância média entre seus membros é pequena.

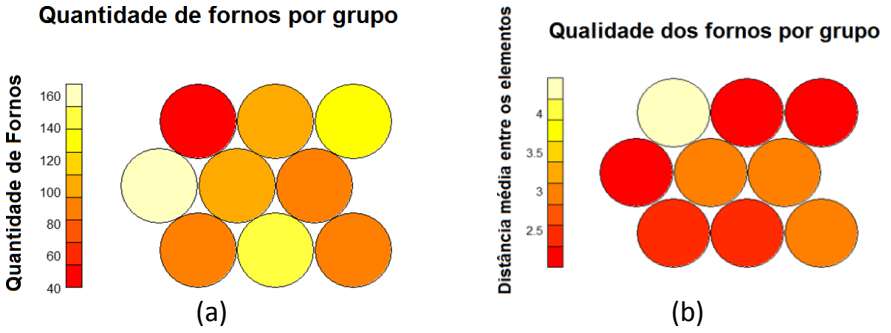


Fig. 6. (a) Quantidade de fornos por cluster. (b) Qualidade de cada cluster.

A Tabela 6 mostra a relação entre as quantidades, as qualidades e as características dos grupos. O grupo vermelho possui de forma estimativa valores entre 40 e 60 fornos, o grupo laranja possui entre 80 e 100 fornos, o grupo amarelo possui uma estimativa de 120 a 140 fornos e o grupo bege possui uma faixa de 160 fornos.

Tabela 6. Relação entre quantidades, qualidades e características dos grupos.

QUANTIDADES							
Grupo	TMP	ALF	ALF3A	QALr	IncTM	TUN	TOV
1	Alto	Baixo	Alto	Alto	Baixo	Baixo	Nulo
2	Alto	Baixo	Baixo	Nulo	Baixo	Baixo	Alto
3	Alto	Nulo	Alto	Nulo	Nulo	Alto	Alto
4	Alto	Baixo	Alto	Alto	Baixo	Baixo	Alto
5	Alto	Nulo	Alto	Alto	Alto	Baixo	Alto
6	Baixo	Baixo	Nulo	Alto	Alto	Alto	Alto
7	Baixo	Baixo	Alto	Alto	Baixo	Alto	Alto
8	Baixo	Baixo	Baixo	Alto	Alto	Nulo	Alto
9	Nulo	Alto	Nulo	Alto	Alto	Baixo	Alto
QUALIDADES							
Grupo	TMP	ALF	ALF3A	QALr	IncTM	TUN	TOV
1	Alto	Baixo	Alto	Alto	Baixo	Baixo	Nulo
2	Alto	Baixo	Baixo	Nulo	Baixo	Baixo	Alto
3	Alto	Nulo	Alto	Nulo	Nulo	Alto	Alto
4	Alto	Baixo	Alto	Alto	Baixo	Baixo	Alto

5	Alto	Nulo	Alto	Alto	Alto	Baixo	Alto
6	Baixo	Baixo	Nulo	Alto	Alto	Alto	Alto
7	Baixo	Baixo	Alto	Alto	Baixo	Alto	Alto
8	Baixo	Baixo	Baixo	Alto	Alto	Nulo	Alto
9	Nulo	Alto	Nulo	Alto	Alto	Baixo	Alto

Ainda de acordo com a Tabela 6, no que diz respeito às quantidades, é possível conferir que o grupo 1 é o que possui menos fornos; o grupo 4 possui o maior número de fornos se comparado a todos os grupos e é exatamente o grupo que não possui nenhuma característica "Nulo"; no entanto, apesar de possuir algumas características "Nulo", os grupos 3 e 8 detêm muitos fornos também. Sobre as qualidades, verifica-se que o grupo 1, além de possuir menos fornos, é o que tem a pior qualidade e que os grupos 2, 3, 4, 7 e 8 são de alta qualidade, sendo que todos os grupos mais numerosos estão entre os melhores, pois possuem menor valor para a métrica distância euclidiana.

Outras interpretações importantes, que constituem as regras associativas descobertas, podem ser encontradas abaixo:

a) TOV aparece alto em todos os grupos, exceto no grupo 1, onde aparece nulo;
b) QALr aparece alto em todos os grupos, exceto nos grupos 2 e 3, nos quais aparece nulo;
c) ALF aparece baixo ou nulo em todos os grupos, exceto no grupo 9;
d) TMP aparece nulo somente no grupo 9;
e) IncTM aparece nulo somente no grupo 3;
f) TUN aparece nulo somente no grupo 8;
g) O grupo com mais variáveis nulos é o 3 (variáveis: ALF, QALr, IncTM);
h) Quando TMP é alto, ALF é baixo ou nulo.
i) Quando TMP é baixo, ALF é baixo;
j) Quando TMP é nulo, ALF é alto;
k) O grupo 1 é o menor (mais escuro) e possui menor qualidade (mais claro);
l) Os maiores grupos (3, 4, 8) são aqueles que tem, também, maior qualidade.

As regras analisadas acima foram adquiridas através dos dados das técnicas de mineração de dados do SOM, e as mesmas podem ser utilizadas para explicar parcialmente o funcionamento do forno.

4 Conclusão

Este trabalho detalha como foi realizado o agrupamento de fornos de redução de alumínio, usando dados reais de uma fábrica. Além do agrupamento, foi possível identificar diversas regras que podem também servir para treinar novos engenheiros, bem como pessoas envolvidas no processo do funcionamento do forno em questão, para que ao invés dos trabalhadores aprenderem a manusear o forno presencialmente, os mesmos podem ter o primeiro contato de forma virtual, através de um programa

computacional, no qual sejam feitas várias simulações e testes, evitando que não haja danos aos fornos e que acidentes não ocorram com os operadores dos mesmos.

Referências

1. Grjotheim, K. e Kvande, H. Introduction to Aluminium Electrolysis Understanding the Hall-Héroult Process, Aluminium-Verlag, 2ª edição, 1997.
2. Shinzato, M. C. Remoção De Metais Pesados Em Solução Por Zeólitas Naturais: Revisão Crítica, Revista do Instituto Geológico, São Paulo, pp. 65-78, 2007.
3. Soares, F. M. Modelagem dinâmica neural e controle fuzzy de banho eletrolítico do forno de redução de alumínio. Qualificação de Doutorado – Universidade Federal do Pará – Belém-Pa, 2015.
4. Castro, L. N. de; Ferrari, D.G. Introdução à Mineração de Dados: Conceitos básicos, algoritmo e aplicações. 1° Edição. São Paulo: Editora Saraiva, 2016.
5. Zaki, M. J., Meira Jr., W. Data Mining and Analysis - Fundamental Concepts and Algorithms. Cambridge University Press, 2014.
6. Kohonen, T. Self-organized formation of topologically correct feature maps. Jornal Biological Cybernetics, Volume 43, Issue 1, pp 59–69, 1982.
7. Merelo, J. J. (2004). Mapa autoorganizativo de Kohonen, Tutorial. Universidad de Granada. Disponível em: <http://geneura.ugr.es/~jmerelo/tutoriales/bioinfo/Kohonen. html>. Útimo acesso:Abril/2017.
8. Matloff, N. The Art of R Programming: A Tour of Statistical Software Design, 1st Edition, No Starch Press, 2011.