UNIVERSIDADE PRESBITERIANA MACKENZIE
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA E COMPUTAÇÃO

Thiago Mesquita Rolemberg

Aplicação de conceitos de redes complexas para a descoberta de formação de grupos em mapas auto-organizáveis

São Paulo
2021

UNIVERSIDADE PRESBITERIANA MACKENZIE
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA ELÉTRICA E COMPUTAÇÃO

Thiago Mesquita Rolemberg

Aplicação de conceitos de redes complexas para a descoberta de formação de grupos em mapas auto-organizáveis

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação da Universidade Presbiteriana Mackenzie como parte dos requisitos para a obtenção do título de Mestre.

Orientador: Prof. Dr. Leandro Augusto da Silva

São Paulo
2021

Mesquita Rolemberg

L475a Rolemberg, Thiago Mesquita
Aplicação de conceitos de redes complexas para a descoberta de formação de grupos em mapas auto-organizáveis / Thiago Mesquita Rolemberg.

69 f.: il.; 30 cm
Bibliografia: f. 1-70

Dissertação (Mestrado em Engenharia Elétrica e Computação, São Paulo, 2021.
Orientador: Prof. Dr. Dr. Leandro Augusto da Silva

1. Mapa Auto-organizado 2. Redes Neurais 3. Redes Complexas 4. Grafos 5.medidas de centralidades 6.U-Matrix 7.Dendrogramas, 8.Aprendizado de máquina 9.Agrupamento 10.Aprendizado De Máquina 11. Não Supervisionado. I. Silva, Leandro Augusto da; Orientador. II. Título.

CDD 620.5
Bibliotecária Responsável: Maria Gabriela Brandi Teixeira – CRB 8/6339

$_r ev.pdf$

Thiago Mesquita Rolemberg

Aplicação de conceitos de redes complexas para a descoberta de formação de grupos em mapas auto-organizáveis.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Computação da Universidade Presbiteriana Mackenzie como parte dos requisitos para obtenção do título de mestre.

Aprovada em 22 de Setembro de 2021.

BANCA EXAMINADORA

Prof. Dr. Leandro Augusto da Silva
Universidade Presbiteriana Mackenzie

Prof. Dr. Luiz Henrique Alves Monteiro
Universidade Presbiteriana Mackenzie

Prof. Dr. EDSON CAORU KITANI
FATEC Santo André - CPS

AGRADECIMENTOS

Ao Prof. Dr. Leandro Augusto da Silva, meu orientador, que sempre me incentivou e acreditou, durante todos esses anos do mestrado, me mostrando a beleza de ser um pesquisador.

A minha família, por entender me apoiar e em muitas das vezes compreender minha ausente em cansáveis noites, finais de semanas e até ferias. Acreditaram no meu potencial e de conseguir realizar esse meu grande sonho.

Aos membros da banca Prof. Dr. Luiz Henrique Alves Monteiro e Prof. Dr. Edson Caoru Kitani, que colaboraram durante a fase de qualificação da minha pesquisa com valiosas contribuições e observações.

A Universidade Mackenzie, por sua estrutura acadêmica que me ajudou a chegar a aqui.

E principalmente a Deus por ter me dado forças para não desistir em momentos muito difíceis da minha vida.

RESUMO

Redes Neurais do tipo Mapas Auto-Organizáveis ou SOM (do inglês *Self-Organizing Maps*), em particular, se destaca como um algoritmo que permite analisar as características de agrupamento e a relação topológica dos dados, a partir de um reticulado de neurônios. Contudo, ainda há uma lacuna de pesquisa que consiste em descobrir a relação por de trás dos atributos que levam a formação de grupos. Neste sentido, propõe-se neste trabalho o uso de conceitos de redes complexas no sentido de usar os neurônios do reticulado para a geração de um grafo e complementar a análise no contexto de comunidade, analisando a formação de grupos por medidas de centralidade. Experimentos em três bases de dados demonstram a viabilidade da proposta.

Palavras-chave: mapa auto-organizado, redes neurais, redes complexas, grafos, medidas de centralidades, u-matrix, dendrogramas, aprendizado de máquina, agrupamento, aprendizado de máquina, não supervisionado.

ABSTRACT

Neural Networks of the Self-Organizing Maps type, in particular, stands out as one of the clustering algorithms for allowing the analysis of cluster characteristics and the topological relationship between clusters from a lattice of neurons. However, there is still a research gap, which consists of discovering the relationship behind the attributes that lead to the formation of groups. In this sense, this work proposes the use of complex network concepts in order to use the lattice neurons to generate a graph and complement the analysis in the community context, analyzing the formation of groups by measures of centrality. Experiments using three datasets shown the proposal effectivity.

Keywords: self-organizing map, redes neurais, neural networks, complex networks, graphs, centrality measures, u-matrix, dendrograms, machine learning, grouping, machine learning, unsupervised.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

1 INTRODUÇÃO

Agrupamento de dados consiste em uma tarefa de Mineração de Dados cuja aplicação de um algoritmo resulta em objetos organizados em grupos. Isso se dá a partir de medidas de similaridade ou dissimilaridade entre os objetos que são representados por um conjunto de atributos. A tarefa de agrupamento de dados desperta o interesse de diferentes áreas de conhecimento e da indústria pelo fato de permitir descobrir de maneira não supervisionada algum conhecimento nos dados que são úteis em tomadas de decisões estratégicas (WU, 1993; PEI, 2011; JAIN, 2010). Exemplos típicos de aplicações vão desde segmentação de imagens (WU, 1993) a descoberta de perfil de usuários em conjunto de dados (JAIN, 2010).

Na literatura, uma das taxonomias mais aceitas para algoritmos de agrupamentos é a proposta por (JAIN, 2010) que os separam em particionais e hierárquicos. Os algoritmos particionais separam os objetos a partir de uma parametrização a priori do número de grupos desejado. Por sua vez, os métodos hierárquicos organizam os objetos por similaridade. Ainda existem os métodos bio-inspirados que geralmente tentam descobrir diferentes composições de grupos guiados por alguma função de custo em um processo de otimização (BERKHIN, 2011). Alternativamente aos algoritmos supracitados destaca-se a rede neural de Mapas Auto-Organizáveis SOM (do inglês *Self-Organizing Maps*), como a rede é conhecida, e tem uma caraterística similar ao algoritmo de particionamento, agrupando objetos semelhantes em um reticulado de neurônios, geralmente em 2D, e garantindo a manutenção da topologia dos dados (KOHONEN, 2013).

O aumento do interesse de que os algoritmos de aprendizagem de máquinas sejam cada vez menos "caixa-preta", apesar de sua superioridade em muitas aplicações no muno real, é um obstáculo fundamental para a adoração no meio corporativo. Para dar mais clareza na "caixa-preta", existem alguns métodos ou abordagens conceitualmente chamadas de XAI (do inglês, *eXplainable Artificial Intelligence* (ARRIETA et al., 2020)). É importante pensar em metodologias que permitam ao especialista entender a mecânica do algoritmo em termos de como o resultado final foi gerado. No contexto de agrupamento um interesse que se torna relevante é saber explicar como se deu uma formação de grupos.

Algumas tentativas foram feitas usando conceitos de redes complexas e considerando o reticulado de neurônios como vértices de um grafo. A citar como exemplo a tentativa de geração de um grafo a partir de neurônios com uma abordagem empírica para remover arestas com conexão fraca, gerando particionamentos no reticulado (SILVA; COSTA, 2011). Outras tentativas ainda considerando SOM e redes complexas foram feitas para criar uma reordenação do reticulado permitindo entender sequências de imagens e vídeos a partir da reorientação do reticulado em um grafo orientado (KITANI; DEL-MORAL-HERNANDEZ; SILVA, 2012; KITANI; HERNANDEZ; SILVA, 2013).

1.1 JUSTIFICATIVA

Durante o processo de descoberta do conhecimento no agrupamento, algumas dificuldades podem ser encontradas, a saber: estimar o número de grupos, visualizar os agrupamentos e interpretar as relações dos objetos em cada grupo. Podemos, por exemplo, citar algoritmos particionais que tem a subjetividade na definição do parâmetro k ideal, que define a quantidade de grupos a ser particionado (SILVA; PERES; BOSCARIOLI, 2016).

Diante dessas dificuldades para a descoberta do agrupamento, algumas técnicas podem ser usadas para auxiliar no entendimento dos resultados (JAIN, 2010), (SILVA; PERES; BOSCARIOLI, 2016). As medidas de validação de agrupamento como Silhouette e Davies-Bouldin Index são as mais conhecidas na literatura (PETROVIć, 2006). Essas medidas, de modo geral, avaliam a coesão de cada grupo formado (intra-grupo) e o isolamento entre os grupos (inter-grupo) sem considerar conhecimento prévio sobre o número de grupos do conjunto de dados. Para visualização dos agrupamentos é comum o uso de técnicas como dendrogramas (JR. JOSEPH F.; ANDERSON, 1998) e a *u-matrix* que é uma ferramenta canônica para exibição das estruturas de distâncias dos protótipos em algoritmos *Self-Organizing Map* (VESANTO J.; ALHONIEMI, 2000).

Os índices citados mensuram a compacidade e a separabilidade de agrupamentos, porém não trazem informações do relacionamento em que cada grupo exerce sobre o outro.

Nesse sentido que emerge a motivação em propor uma técnica para a descoberta de formação do agrupamento, em especial com o uso da rede SOM, analisando o reticulado

da camada de saída da rede. Por se tratar de uma saída matricial, a aplicação de técnicas de redes complexas, como grafos e medidas de centralidades, passam a ser uma ferramenta robusta para entender as relações de cada neurônio. Tais análises podem dar ao especialista do negócio um entendimento mais preciso de como o agrupamento foi formado, qual o grupo mais importante e qual o neurônio que tem mais influência sobre os demais.

Embora os trabalhos anteriores tenham caminhado para fornecer informações adicionais sobre os grupos a partir de redes complexas, ainda existe a necessidade de conseguir explicar como se dá uma formação de grupos e como os mesmos se relacionam.

1.2 OBJETIVOS

O objetivo deste trabalho é apresentar uma abordagem híbrida usando SOM e Redes Complexas em que primeiro organiza os dados em um reticulado de neurônios e, depois, reorganiza os neurônios em vértices de um grafo usando como informações de centralidade para que se consiga descobrir como a formação de grupos acontece e explicar como os mesmos se relacionam.

De forma específica, as medidas de centralidades aplicadas nos neurônios do mapa SOM serão usados para exploração na descoberta do agrupamento com duas abordagens:

- A transformação da matriz do espaço de saída do SOM, em grafo não orientado como ferramenta de visualização da relação dos protótipos do SOM.
- Uso de medidas de centralidades para entender a formação do agrupamento.

1.3 HIPÓTESES DE PESQUISA

Diante dos fatos apresentados na introdução e na justificativa, a seguinte hipótese de pesquisa pode ser formulada: O uso de medidas de centralidades pode ser útil para a interpretação dos resultados do algoritmo SOM?

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado de maneira que possa compreender a aplicação do algoritmo SOM para realizar o agrupamento de dados. Posteriormente a aplicação de técnicas de redes complexas, medidas de centralidades para a compreensão do entendimento da formação de grupos no SOM.

No capítulo 2 (Trabalhos Correlatos) são apresentados alguns trabalhos pesquisados que ajudam a entender a inspiração no desenvolvimento desta pesquisa. Principalmente por demonstrar trabalhos que utilizam o algoritmo SOM e, posteriormente, outras técnicas que tendem a trabalhar para o entendimento da formação do agrupamento. O capítulo 3 (Referencial Teórico) apresenta conceitos teóricos para a compreensão do entendimento da aplicação deste trabalho. Desde aspectos sobre agrupamentos de dados, explicação do algoritmo SOM, redes complexas, grafos e medidas de centralidades. No capítulo 4 (Método proposto: Combinação do SOM com Redes Complexas) é apresentado o método usado nesta pesquisa para realizar a combinação do resultado do agrupamento feito pelo SOM com as medidas de centralidades para descoberta da formação do agrupamento. O capítulo 5 (Materiais e Métodos) apresenta os materiais e métodos usados para reproduzir os experimentos desta pesquisa. Com todo o desenvolvimento teórico desta pesquisa no capítulo 6 (Resultados Experimentais) apresenta todos os experimentos e resultados das bases de dados de acordo com o capítulo 5. Por último, no capítulo 7 (Conclusão e Trabalhos Futuros) apresenta a conclusão desta pesquisa e apresenta propostas de trabalhos futuros.

2 TRABALHOS CORRELATOS

Durante o desenvolvimento deste trabalho foram pesquisados trabalhos que usam SOM como um tipo de pré-processamento de dados combinado com algoritmos de agrupamento para a descoberta de grupos.

Um dos primeiros trabalhos desta revisão, que é um dos mais citados na literatura, envolveu a organização dos dados do SOM e, consequentemente, o agrupamento dos vetores de pesos do neurônio usando k-médias (agrupamento particional) e agrupamento aglomerativo (agrupamento hierárquico). No trabalho, os autores justificam a abordagem híbrida como forma de eliminar os *outlier* dos dados e reduzir o número de objetos a partir do SOM (VESANTO; ALHONIEMI, 2000).

Outros trabalhos que se assemelham mais a proposta deste trabalho tiveram como origem o trabalho de (VESANTO; ALHONIEMI, 2000). Contudo, usam conceitos de redes complexas ao invés de algoritmos agrupamento de dados, como os apresentados em (SAXENA et al., 2017).

Como a saída do SOM é um reticulado pode-se representar cada neurônio como vértice e usar alguma medida de similaridade entre os neurônios adjacentes para medir a similaridade. Assim, esta abordagem possibilita transformar o reticulado em um grafo (SILVA; COSTA, 2011). A tentativa dos autores foi de conseguir descobrir o número de grupos no conjunto de dados sem a interferência de parâmetros como proposto por (VESANTO; ALHONIEMI, 2000).

Contudo, ambas as abordagens tinham como proposta descobrir automaticamente o número de grupos e não descobrir como os grupos se formam. O trabalho mais próximo a isto também partiu do reticulado dos neurônios, transformando em um grafo não orientado de forma a criar uma rede complexa. Os autores usaram o algoritmo de Dijkstra de maneira a percorrer pelo grafo a partir de seus vértices (neurônio), com a finalidade de entender como uma sequência de imagens podem ser apresentadas de forma coerente (KITANI; DEL-MORAL-HERNANDEZ; SILVA, 2012; KITANI; HERNANDEZ; SILVA, 2013).

A proposta deste trabalho se assemelha às propostas de (KITANI; DEL-MORAL-HERNANDEZ; SILVA, 2012; KITANI; HERNANDEZ; SILVA, 2013). A diferença é que o objetivo aqui é trabalhar com dados numéricos e com a finalidade de descobrir como os grupos se formam; e para isto se usa medidas de centralidade, como será apresentado no decorrer deste trabalho.

Portanto, as contribuições deste trabalho face aos trabalhos da literatura apresentados podem ser assim definidas:

- Aplicações de métricas de centralidades e redes complexas como uma abordagem para a descoberta de grupos em SOM;
- Uma metodologia que se possa explicar como a formação dos grupos acontecem e os seus relacionamentos para auxiliar a tomada de decisão;
- O uso de medidas de centralidade para definir arestas na transformação do espaço matricial de saída do SOM em um grafo.

3 REFERENCIAL TEÓRICO

Este capítulo ira abordar os tópicos estudados durante o desenvolvimento deste trabalho. Desde tópicos de agrupamento de dados até sistemas complexos.

3.1 AGRUPAMENTO DE DADOS

A execução dos algoritmos que implementam estratégias para identificar agrupamento buscam por similaridades ou dissimilaridades entre os objetos de um conjunto de dados, quantificadas geralmente por medidas de similaridade. Algoritmo de agrupamento de dados procura fazer com que a similaridade intra-grupos seja maximizada e a similaridade inter-grupos minimizada (SILVA; PERES; BOSCARIOLI, 2016).

A descoberta de um agrupamento se constitui, portanto, como um processo de tomada de decisão sobre a associação de um objeto a um grupo. Ao término do processo, se faz necessária uma segunda etapa, que busca explicação que ajude o especialista do domínio de negócio a entender as relações pelas quais os objetos do grupo são similares (SILVA; PERES; BOSCARIOLI, 2016).

Conceitualmente, os métodos de agrupamento se diferenciam pela forma como os grupos são formados e apresentados como resultados (BERKHIN, 2011), (SILVA; PERES; BOSCARIOLI, 2016). Dentre os métodos encontrados na literatura (BERKHIN, 2011), para esta pesquisa será enfatizada os dois principais: hierárquicas (onde os grupos de dados são definidos por hierarquia), particionais (onde por um critério previamente definido o conjunto de dados é dividido por partes) (SILVA; PERES; BOSCARIOLI, 2016), (BERKHIN, 2011), (DUBES, 1988).

3.1.1 AGRUPAMENTO HIERÁRQUICO

O método hierárquico de agrupamento divide os objetos em forma de hierarquia, cuja estrutura resultante pode ser chamada de dendrograma (MURTAGH, 1983). Na Figura 8 está ilustrado um exemplo de como este algoritmo opera sobre um conjunto de dados.

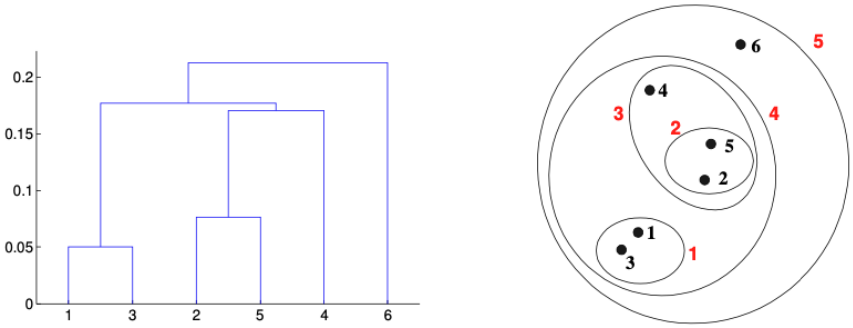


Figura 1 – Exemplo de um dendograma.

O método hierárquico é dividido em duas abordagens:*top-down* e *bottom-up*. Quando é adotada a abordagem por *top-down*, o processo de análise se inicia colocando todos os objetos em um único grupo e, iterativamente, dividindo um grupo em grupos menores, até que cada objeto seja isolado em um único grupo (SILVA; PERES; BOSCARIOLI, 2016). Na abordagem por *bottom-up*, o processo de análise inicia com cada objeto em um grupo separado e, iterativamente, aglomera grupos similares, até que um único grupo com todos os objetos seja formado, como exemplificado na Figura 8 (SILVA; PERES; BOSCARIOLI, 2016).

Independente da abordagem, o resultado será apresentado em forma de hierarquia, como um dendrograma (vide Figura 8), podendo ser uma forma de visualização dos grupos descobertos (MURTAGH, 1983), (SILVA; PERES; BOSCARIOLI, 2016), (DUBES, 1988). Como o algoritmo não particiona os grupos, pode-se aplicar cortes no dendrograma e consequentemente ter os grupos desejados (SILVA; PERES; BOSCARIOLI, 2016).

3.1.2 MÉTODOS PARTICIONAIS DE AGRUPAMENTO DE DADOS

Diferente dos métodos hierárquicos, os particionais usam um parâmetro, a ser definido a priori, para segmentar o conjunto de dados em grupos. Partindo de protótipos inicializados aleatoriamente, os objetos vão sendo segmentados até chegar a partição final. Em princípio, a partição ótima pode não ser descoberta e, portanto, algum critério de parada precisa ser definido (ASSIS, 2018; SILVA; PERES; BOSCARIOLI, 2016). A definição do parâmetro inicial, de definição da quantidade de grupos é considerado um ponto negativo do método, pois esse domínio de conhecimento não é disponível para muitas aplicações (ESTER, 1996).

Descritivamente, o funcionamento de algoritmos particionais é feito da seguinte forma: inicialmente é definido o número de grupos k, sendo este o valor de centro dos agrupamentos. Os objetos então são divididos entre os k grupos de acordo com a medida de similaridade escolhida, em muitos algoritmos essa medida é a distância Euclidiana. Esse processo ocorre de modo que cada objeto fique no grupo que forneça a menor distância entre o objeto e o centro do mesmo. Então, o algoritmo utiliza uma estratégia iterativa de controle para determinar que objetos devem mudar de grupo, de forma que a função objetiva seja otimizada (ESTER, 1996).

Ao terminar a divisão inicial, há duas possibilidades na escolha do objeto que irá representar o centro do grupo, e que será a referência para o cálculo da medida de similaridade. Ou será utilizada a média dos objetos que pertencem ao grupo em questão, também chamado de centro de gravidade do grupo, ou escolhe-se como representante o objeto que se encontra mais próximo ao centro da gravidade do grupo. A primeira escolha é conhecida como k-médias (em inglês k-means) e a segunda escolha como k-medóides (em inglês medoids) (ESTER, 1996).

O algoritmo k-means é bastante utilizado, tanto no mundo acadêmico como no mundo corporativo, principalmente por se tratar de uma abordagem mais simples de particionamento (JAIN, 2010). A função objetiva utilizada para os espaços métricos nos métodos particionais é o erro quadrático.

$$E = \sum_{j=1}^{k} \sum_{x \in C_i} \|p - m_i\|^2, k \in (1, n) \tag{3.1}$$

Na equação 3.1, E e a soma do erro o quadrático para todos os objetos na base de dados, p e o ponto no espaço representando um dado objeto, e m_i e o representante do agrupamento C_i. Tanto p quanto m_i são multidimensionais. Essa função objetivo dividida por n representa a distância media de cada objeto ao seu respectivo representante (ESTER, 1996) e também e conhecida como critério do erro médio quadrático. Os algoritmos terminam quando nao existem atribuições possíveis capazes de melhorar esta função objetivo (ESTER, 1996).

Diferente do que é encontrado em métodos hierárquicos, na qual o agrupamento é produzido por um relacionamento, o método particional produz um agrupamento mais

simples. No artigo, (ANKERST MARKUS M. BREUNIG, 1999), destaca que esses algoritmos são efetivos se o número de grupos k puder ser razoavelmente estimado e, se os grupos são de forma convexa e possuem tamanho e densidade similares.

3.1.3 MAPAS AUTO-ORGANIZÁVEIS

Mapas Auto-Organizáveis ou simplesmente SOM (do inglês, *Self Organizing Maps*) consiste em uma categoria de Redes Neurais Artificiais com arquitetura de duas camadas, sendo a primeira para recebimento dos dados de entrada e a segunda formada por um reticulado com neurônios. Geralmente, esse reticulado é bidimensional (2-D). A estrutura do reticulado 2-D pode ser ainda hexagonal ou regular (KOHONEN, 2013).

Na figura 2 é apresentada uma arquitetura unidimensional do funcionamento do SOM. A camada de entrada é composta por uma série de neurônios sensoriais, sendo responsáveis por estimular a rede neural. O número de neurônios sensoriais é sempre igual ao número de atributos no conjunto de dados analisado.Todos os atributos descritivos de um conjunto de objetos são apresentados a todos os neurônios da saída da arquitetura. Esta apresentação é representada na arquitetura pelas ligações entre os neurônios das duas camadas. Na figura 3 são apresentados três topologias que uma arquitetura SOM pode ser configurada (CASTRO, 2016).

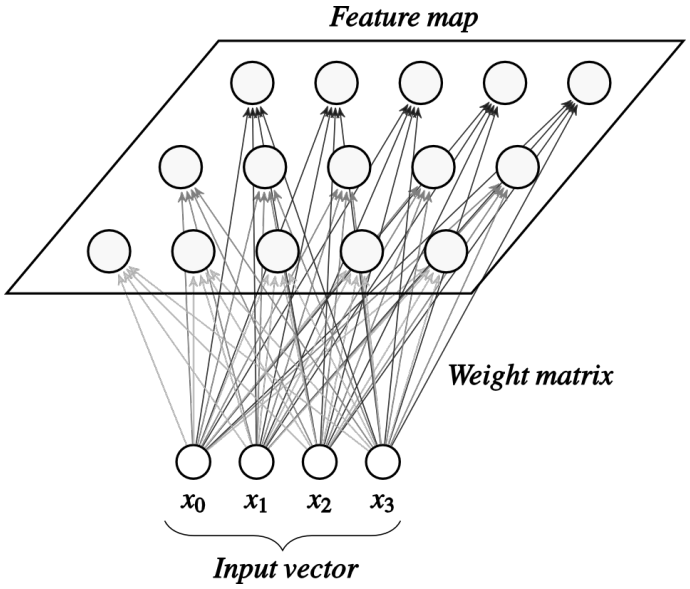


Figura 2 – Camada SOM.

A camada de entrada espera receber um objeto $\mathbf{x}_i$ de um conjunto de dados $\mathbf{X} =$

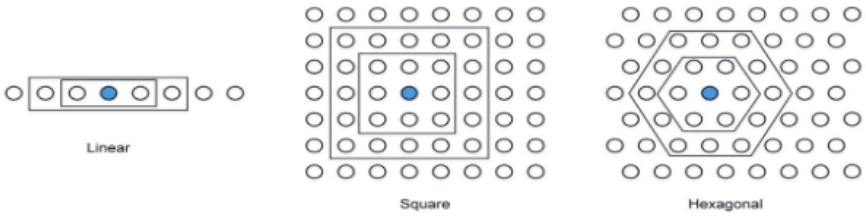


Figura 3 – Tipos de topologias SOM.

$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_N\}$. Cada objeto é definido por um conjunto de M atributos, $\mathbf{x}_i = \{x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{iM}\}$.

Na camada de saída, o reticulado é parametrizado com U unidades, sendo cada neurônio u definido como um vetor de pesos com a mesma dimensão do objeto de entrada, ou seja, $\mathbf{w}_u = \{w_1, w_2, ..., w_M\}$. Estes pesos são inicializados de forma aleatória e ajustados durante o aprendizado.

O treinamento do algoritmo SOM é realizado de forma iterativa. Inicialmente, com $t = 0$, os vetores de peso são inicializados aleatoriamente, de preferência a partir do domínio de vetores de entrada (KOHONEN, 2001). Em cada etapa do treinamento t, um padrão de entrada $\mathbf{x}_i(t)$ é escolhido aleatoriamente do conjunto de treinamento $\mathbf{X}$. Cada objeto é comparado com os vetores de peso por meio de uma medida de distância, geralmente Euclidiana. A distância entre $\mathbf{x}_i(t)$ e todos os vetores de peso $(w)_u$ são calculadas. O neurônio vencedor é o protótipo mais próximo de $\mathbf{x}_i(t)$ sendo este o *Best Match Unit* (BMU). O vetor de peso BMU é atualizado, assim como o vetor de pesos dos neurônios dos seus vizinhos, mas com a intensidade menor. Para detalhe completo do algoritmo recomenda-se a leitura desse artigo (KOHONEN, 2013).

Cabe por fim ressaltar que após o treinamento, os neurônios representam um subgrupo do conjunto de dados com semelhanças em aspectos de atributos e, no que lhe concerne, neurônios vizinhos representam conjunto de dados que mantém algumas características em comum. Esta manutenção topologia dos dados é uma das principais razões para uso do SOM em problemas de agrupamento de dados. Contudo, é importante mencionar que se trata de um reticulado fixo em que situações onde os dados diferem em atributos, ou seja, grupos distintos, ele se mantém próximos no espaço do reticulado. E isso, por sua vez, é uma característica do SOM que dificulta o entendimento da formação de grupos e quais os atributos similares.

3.1.3.1 O processo competitivo

É responsável por definir o neurônio vencedor, neste caso aquele que apresentou o melhor casamento vetor de entrada com os vetores de pesos sinápticos w_j após o cálculo de uma função discriminante. Ela fornece a base para a competição entre os neurônios. O neurônio particular com o maior valor da função discriminante é chamado o neurônio (em inglês *Best Match Unit*, (BMU)), neurônio vencedor para o vetor de entrada (KOHONEN, 2012).

Na figura 4 é possível observar um exemplo de um mapa de dimensão 4 x4, justamente no momento que é informado o BMU.

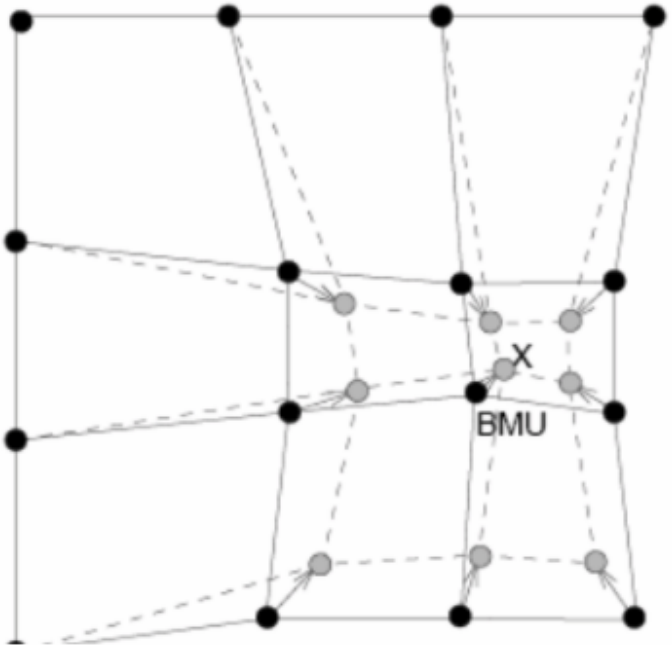


Figura 4 – Grade do mapa auto-organizável com o neurônio vencedor (BMU) para o padrão de entrada x. (VESANTO; ALHONIEMI, 2000)

A distância euclidiana é usualmente usada como função discriminante, e o neurônio vencedor é aquele que apresenta a menor distância entre o vetor de entrada e os pesos w_j (HAYKIN, 2001). Uma ponto importante visto nesse processo é: um espaço contínuo de entrada de padrões de ativação é mapeado para um espaço discreto de saída de neurônios por um processo de competição entre os neurônios da grade (HAYKIN, 2001).

3.1.3.2 O processo cooperativo

O processo cooperativo define quais os neurônios, além do vencedor, terão seus pesos sinápticos alterados. Um grande desafio encontrado nesse processo é como definir uma vizinhança topológica que seja correta do ponto de vista neurobiológico.

Neurobiologicamente, existe uma interação lateral entre os neurônios de forma que o neurônio que dispara tende a excitar mais fortemente os neurônios na sua vizinhança imediata que aqueles distantes dele (HAYKIN, 2001). Uma função de vizinhança topológica, h_{ji}, deve satisfazer duas exigências distintas, considerando d_{ij} a distância lateral entre o neurônio vencedor e o neurônio excitado j (HAYKIN, 2001) :

1. Deve ser simétrica em relação ao ponto máximo definido por $d_{ij} = 0$; em outras palavras, ela alcançar o seu valor máximo no neurônio vencedor i para o qual a distância d_{ij} é zero.

2. Sua amplitude deve decrescer monotonamente com o aumento da distância lateral d_{ij}, decaindo a zero para $d_{ij} \Rightarrow \infty$; esta é uma condição necessária para a convergência.

As funções geralmente usadas são Chapéu-mexicano e a Gaussiana. Na função Chapéu-mexicano o neurônio vencedor estimula lateralmente um apequena vizinhança ao seu redor e a medida que a distancia aumenta a estimulação torna-se inibição. Na função Gaussiana a amplitude da vizinhança tende a zero à medida que a distância lateral aumenta. Na figura 5 é observado essas funções.

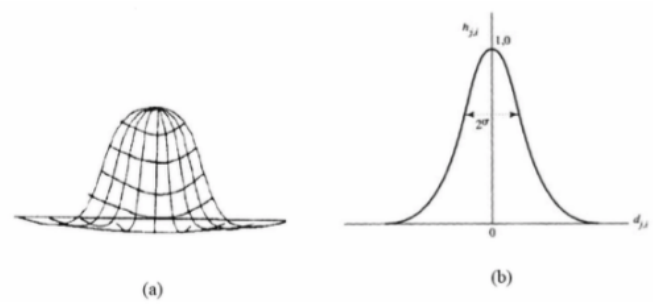


Figura 5 – Função de vizinhança Chapeu Mexicano (a) e Gaussiana(b) (VESANTO; ALHONIEMI, 2000).

A topologia da rede pode assumir diversas formas, na literatura é encontrado com mais frequência experimentos com a topologia quadrada e hexagonal (VESANTO; ALHONIEMI, 2000). A Figura ilustra um exemplo com ambas as topologias 6.

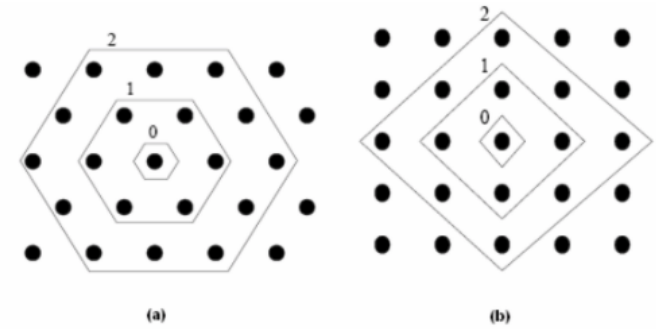


Figura 6 – Topologias: (a) grade com disposição quadrada e (b) grade com disposição hexagonal (VESANTO; ALHONIEMI, 2000).

3.1.3.3 O processo adaptativo

O sináptico é o último passo na formação auto-organizada e um mapa de características para que esse mapa seja auto-organizável é necessário que o vetor de peso sináptico w_{ji} do neurônio j do reticulado se modifique em relação ao vetor de entrada x. Os pesos dos neurônios (vencedor e seus vizinhos) serão atualizados a partir da Equação (KOHONEN, 2012).

$$w_j(n+1) = w_j(n) + \eta(n)h_{ij}(n)(x(n) - w_j(n)) \tag{3.2}$$

Onde $w_j(n)$ é o vetor de pesos no tempo n $w_j(n+1)$ é o vetor de pesos atualizado, x é um padrão de entrada e $\eta(n)$ é a taxa de aprendizado no instante n. A taxa de aprendizado segue as mesmas regras de decaimento do raio da vizinhança, isto é, pode ser calculado por um decaimento exponencial ou ainda variar de acordo com um valor fixo pré-determinado a cada iteração. Logo, o processo adaptativo é dividido, segundo (HAYKIN, 2001), em dois momentos: fase de ordenação e fase de convergência.

A fase de ordenação dos vetores de pesos, iniciados linearmente ou aleatoriamente, são ordenados topologicamente. Este momento exige em torno de 1000 iterações da rede e tem como objetivo organizar os neurônios evidenciando a distribuição dos padrões do espaço de entrada. Deve-se ter cuidado na escolha dos parâmetros da taxa de aprendizagem e

raio de vizinhança. Durante esta fase, a taxa de aprendizagem inicialmente é alta em torno de 1 e reduzida a um valor próximo de 0,1. Quando a vizinhança, deve envolver inicialmente todos ou quase todos os neurônios da rede, sendo reduzida até atingir um raio por de um ou nenhum neurônio (HAYKIN, 2001), (KOHONEN, 2013).

Na convergência é feito um ajuste fino no mapa e tem como objetivo produzir uma quantização estatística precisa do espaço de entrada. Agora é necessário segundo (HAYKIN, 2001), de no mínimo 500 vezes o número de neurônios no reticulado. a taxa de aprendizado, nessa fase, é baixa em torno de 0,01 ou menos. Porém, deve-se evitar que diminua a zero (HAYKIN, 2001), pois caso ocorra, é possível que ao reticulado fique preso em um estado metaestável. Para o raio de vizinhança tem-se apenas um ou nenhum vizinho.

3.1.3.4 Algoritmo SOM

Neste capítulo é apresentado uma explicação da execução do algoritmo SOM, também é apresentado o pseudo-código.

O primeiro passo do algoritmo é a definição dos parâmetros iniciais, e observar que o processo do treinamento pressupõe um laço de controle de sua duração, ou seja, esse laço funciona com a quantidade de épocas que o treinamento deve seguir (SILVA; PERES; BOSCARIOLI, 2016). Abaixo segue a descrição dos parâmetros iniciais.

- X_{tr}: um conjunto de objetos de treinamento não rotulado, ou seja, $X_{tr} = \overrightarrow{x_i}$, $i = 1 \ldots n$
- O: número de neurônios no mapa
- $mapa$: configuração do mapa (número de neurônios por dimensão do mapa, lattice)
- $dist$: uma medida de distância vetorial, aqui, considerada a euclidiana
- $\eta : taxadeaprendizadoinicial$;
- $V_{l,c}$: função de vizinha topológica entre o neurônio l e o neurônio mais próximo ao objeto c
- r: raio de vizinhança

- t_{maximo}: número de iterações ou épocas
- e_{maximo}: valor do erro máximo esperado pela alterações nos pesos
- W: conjunto de pesos sinápticos ajustados

No passo seguinte, enquanto as condições iniciais não atingem o esperado, calcula-se a distância vetorial entre cada objeto do conjunto analisado entre todos os neurônios da camada de saída do SOM. É possível observar que calcular a distância vetorial entre esses dois elementos significa executar o cálculo da distância usando o vetor do objeto x e o vetor de pesos de um neurônio w. Realizado o cálculo de distância do objeto de entrada e todos os neurônios do mapa já calculados, é preciso encontrar qual o neurônio mais próximo do vetor de entrada, ou seja, encontrar qual neurônio minimiza a função de distância $dist(x, w)$. É importante entender que dois vetores que se encontram próximos no espaço dos dados são parecidos entre si. Logo, o neurônio que minimiza a distância é o mais próximo ou similar ao objeto sob análise, e diz que ele é o BMU para o objeto. Essa estratégia implementa um aprendizado do tipo competitivo, ou seja, os neurônios da rede neural estão competindo para representar o objeto (SILVA; PERES; BOSCARIOLI, 2016).

Então, após encontrar o BMU para o conjunto de objetos sob análise, o ajuste dos pesos foi realizado. O ajuste é aplicado ao neurônio vencedor (BMU) e também aos seus neurônios vizinhos no espaço matricial, ou seja, seus vizinhos topológicos. O fato de o ajuste ser feito no BMU e nos seus vizinhos caracteriza esse aprendizado como sendo cooperativo, na estratégia competitiva. O ajuste dos pesos é feito de tal maneira que os neurônios ajustados serão reposicionados no espaço dos dados, de forma a se aproximarem do conjunto, tornando-se mais similar a ele. Uma operação vetorial implementa o reposicionamento do neurônio como vetor $w = w + n * (x, w)$ (SILVA; PERES; BOSCARIOLI, 2016).

A taxa de aprendizado é um termo que regulariza a magnitude da alteração de pesos. Ela pode ser vista como um termo de ponderação para o tamanho do deslocamento que o neurônio sofrerá no ajuste de seus pesos. A taxa de aprendizado deve sempre variar entre [0, 1]. Ela deve assumir valores acentuados, possibilitando que os neurônios assumam uma organização próxima à topologia do conjunto de dados. Na fase de refinamento, a taxa

de aprendizado deve ter o seu valor diminuído, pois assim, contribui para a estabilização dos pesos dos neurônios, permitindo apenas que deslocamentos suaves sejam realizados (SILVA; PERES; BOSCARIOLI, 2016).

O ajuste de pesos deve ser aplicado aos vizinhos do neurônio vencedor. A avaliação para determinar quais são os vizinhos é feita na matriz estrutural matricial, sendo definida logo no início do algoritmo e com a análise tanto da função de vizinhança adotada, quanto do valor do raio de vizinhança (r) vigente na época da execução. A função da vizinhança que realiza o ajuste dos pesos no algoritmo é a função gaussiana (SILVA; PERES; BOSCARIOLI, 2016).

$$v_{lc} = exp(-\frac{da(l,c)^2}{2r^2}) \tag{3.3}$$

Em que $l = 1....0$, c é o BMU e r é o raio de vizinhança (largura da função gaussiana) escolhido no algoritmo. A depender do valor assumido por r, essa função pode permitir que todos os neurônios da camada de saída de um SOM tenham seus pesos ajustados. O decaimento do ajuste sofrido pelo neurônio descreve conforme a expressão $d(l, c)$ (SILVA; PERES; BOSCARIOLI, 2016).

A saída do algoritmo SOM retorna o conjunto de pesos finais do SOM(W), que permite verificar a forma como os neurônios da camada de saída respondem aos estímulos após o processo de aprendizado. Contudo, as configurações básicas do espaço de saída, o mapa, usadas durante o algoritmo devem ser conhecidas, se houver o objetivo de analisar seus resultados em relação a informações de vizinhança topológica, geralmente feito quando se pretende visualizar o mapa gerado (SILVA; PERES; BOSCARIOLI, 2016).

A descrição do pseudo código 1 demonstra o passo a passo do funcionamento de cada

uma das variáveis é descrita para gerar maior entendimento do algoritmo.

Algorithm 1: Pseudo Código SOM

Input: X_{tr}, O, $mapa$, $dist$, n, $V_{l,c}$, r, t_{maximo}, e_{maximo}

defina $t = 0$

while $t < t_{maximo}$ *ou* $e > e_{maximo}$ **do**

Calcule $dist(\overrightarrow{x}, \overrightarrow{w})$ em que $\overrightarrow{w} \in W$ e $l = 1....O$;

Determine $BMU = argmin_l dist(\overrightarrow{x_l}, \overrightarrow{w_l})$

Determine o ajuste de pesos $\overrightarrow{w} = \overrightarrow{w} + v_{lc} * n * (\overrightarrow{x}, \overrightarrow{w})$ para o neurônio BMU e seus vizinhos topológicos.

end

$e = ||W(t+1) - W(t)||$;

$t = t + 1$

Ajuste a taxa de aprendizado (n) e o raio de vizinhança (r), se for o caso

Result: W

3.2 ANÁLISE DE AGRUPAMENTO

Estudos em análise de agrupamento demonstram que existem vários algoritmos que buscam realizar a descoberta do conjunto de objetos (DUBES, 1988), (PEI, 2011). Muitos desses algoritmos realizam o processo de forma simples, com uso de técnicas estatísticas como média, mediana e etc. Essas técnicas são eficazes quando se tem uma quantidade de variáveis pequenas. Porém, quando existe um número de dimensões grande o processo se torna difícil. É comum encontrar em diversos trabalhos, antes do processo de descoberta com um algoritmo tradicional, o uso de técnicas como a de PCA *em inglês Principal Component Analysis*, para realizar o redimensionamento dos dados antes de se aplicar o algoritmo (DUBES, 1988), (HRUSCHKA E. R. EBECKEN, 2001).

A quantidade de dimensões em um conjunto de dados pode ser considerado um problema em análise de agrupamento (BERKHIN, 2011). Além de ser um problema de complexidade computacional também dificulta a análise da visualização (PEI, 2011). Com técnicas tradicionais fica difícil de se realizar a análise, pois se torna complicado traçar um vector ou as relações entre os diferentes vetores existentes. Por essa razão, são necessários outros métodos. A grande desvantagem dos métodos tradicionais é a de não reduzirem a quantidade de dados. Se a quantidade de dados for grande, a visualização dos mesmos

torna-se incompreensível (CASTRO, 2016).

Quando se trata da análise de agrupamento por meio do algoritmo de SOM, esses fatores são minimizados, pelo fato do SOM não ter preocupações com a quantidade de dimensões expostas no conjunto de dados. A camada de entrada acaba sendo criada pela quantidade de dimensões a serem encontradas no conjunto de dados e cria a quantidade de neurônios necessárias para realizar o processo de descoberta de grupos (KOHONEN, 2012).

O ponto comum que podemos encontrar entre os métodos tradicionais de análise de agrupamento e o método de SOM, pode ser visto após a formação dos grupos e como eles devem ser validados. Algumas técnicas, (como já discutido nesta pesquisa) podem ser usadas para realizar essa validação e esse é um ponto que a pesquisa pretende discutir, usar técnicas de redes complexas para visualizar e estudar as relações entre os neurônios de saída.

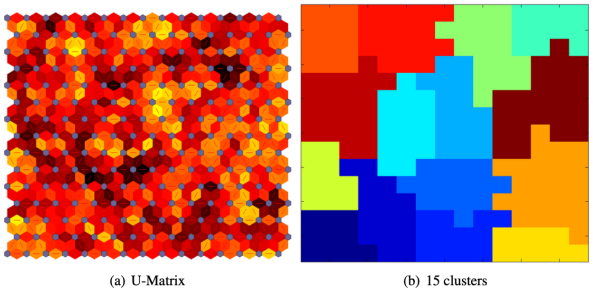
3.2.1 ABORDAGENS USADAS NA VISUALIZAÇÃO DOS RESULTADOS COM SOM

A visualização é definida como a ciência que estuda métodos para facilitar o raciocínio analítico por meio de interfaces visuais (KITANI; DEL-MORAL-HERNANDEZ; SILVA, 2012). A análise visual é considerada uma área multidisciplinar, pois o objetivo é extrair conhecimento a partir de dados brutos e então representar essas informações por meio de gráficos auxiliando assim a tomada de decisão (KEIM D. A., 2006).

Na literatura é possível encontrar diversas técnicas que podem ser usadas para visualizar, analisar e decidir referente ao resultado, nesta pesquisa destaca-se o método de dendrogramas (JR. JOSEPH F.; ANDERSON, 1998) e por Matriz-U (VESANTO J.; ALHONIEMI, 2000).

A abordagem por U-Matrix (matriz de distância unificada) pode ser observada na figura 7, os dados usados gerados foram os mesmo na abordagem por dendograma, onde foi observado 15 grupos.

Para ambas as representações, na maioria das pesquisas, métricas de validação dos agrupamentos é usada, como Davies Bouldin (DAVIES D. L.; BOULDIN, 1979). Outro



(a) U-Matrix (b) 15 clusters

Figura 7 – U-Matrix e segmentação do mapa treinado.

fator a analisar é quando a dimensão dos dados é muito alta (ULTSCH, 2004), nesse caso dendogramas apresentam dificuldades para se visualizar a quantidade de agrupamentos e com isso é usado U-Matrix.

O método proposto por U-matriz ou matriz de distâncias unificadas, foi desenvolvido por Alfred Ulsch com o objetivo de permitir a detecção visual das relações topológicas dos neurônios. usa-se mesma forma de cálculo utilizada durante o treinamento para determinar a distância entre os vetores de peso de neurônios adjacentes. O resultado é uma imagem $f(x, y)$ na qual as coordenadas de cada pixel (x, y) são derivadas das coordenadas dos neurônios da grade de saída do mapa, e a intensidade de cada pixel na imagem $f(x, y)$ corresponde a uma distância calculada. Um mapa bidimensional N x M gera uma imagem $f(x, y)$ de $(2N - 1)$ x $(2M - 1)$ pixel (COSTA, 1999).

Já método por visualização usando o dendrograma, o exemplo pode ser visto na figura 8 que apresenta o resultado da execução do algoritmo SOM. O entendimento da visualização do dendrograma é feito com funções de link, que considera as distâncias entre eles. Para calculara a distância entre os vetores foi escolhida a distância euclidiana e o método do link foi o método de Ward (WARD, 1963). O método foi escolhido, pois, ele minimiza a variância intra-cluster total.

Pode-se criar grupos a partir de dendrogramas traçando linhas horizontais neles, sendo que os grupos ão os ramos da árvore abaixo da linha. Apatir do eixo vertical, pode-se traçar uma linha e com isso obter o número de agrupamentos desejados. Um exemplo usando a figura 8, se traçar uma linha onde o eixo marca 30, será possível obter 5 grupos.

Para ambos os casos, quando se pretende analisar o comportamento dos neurônios não é possível. Por exemplo, pode-se querer entender qual neurônio é responsável por influenciar os demais na formação dos grupos ou até mesmo entender a relação de cada

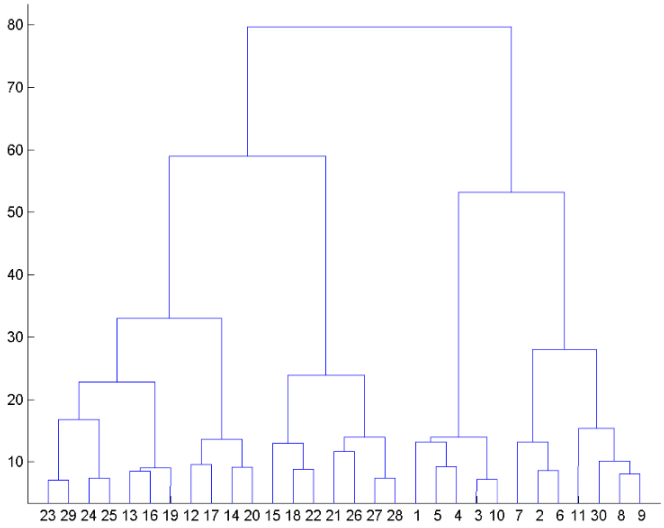


Figura 8 – Topologias: (a) grade com disposição quadrada e (b) grade com disposição hexagonal.

neurônio com seu vizinho. Nesse sentido é que a presente pesquisa através uma nova abordagem para a análise da formação dos grupos, onde se considera a aplicação de redes complexas e suas medidas de centralidade para entender a formação do agrupamento.

3.3 DESAFIOS PARA DESCOBERTA DE GRUPOS

A tarefa de encontrar o número ideal de grupos em um conjunto de dados é considerado um problema difícil. Os autores (HRUSCHKA E. R., 2001), destacam esse problema como um NP-completo e não computacionalmente calculado. Para que o problema não seja classificado como NP-completo o n (número de objetos) e k (número de grupos) devem ser extremamente pequenos, visto que o número de partições distintas em que podemos dividir n objetos em k grupos aumenta aproximadamente como $\frac{k_n}{n!}$ (HRUSCHKA E. R., 2001).

(ANKERST MARKUS M. BREUNIG, 1999), cita diretamente três desafios para a descoberta de grupos.

1. A maioria dos algoritmos de agrupamento exigem parâmetros iniciais de entrada que são difíceis de determinar. Principalmente para conjunto de dados de alta dimensão.

2. Os algoritmos são muito sensíveis aos parâmetros de entrada, com isso acabam produzindo partições muito diferentes do conjunto de dados.

3. E por último, conjunto de dados com alta dimensão geralmente tem uma distribuição com muito distorção que pode ser revelada por um algoritmo de agrupamento.

Um aspecto importante a ser comentado, é que no desafio da descoberta de grupos, está relacionado a como medir a similaridade de cada objeto no conjunto de dados, para isso usam-se medidas de intragrupos e intergrupos para se determinar o agrupamento (SILVA; PERES; BOSCARIOLI, 2016), (HRUSCHKA E. R. EBECKEN, 2001), (ESTER, 1996).

Por mais que algoritmos descrevem que o processo para descoberta de grupos é um processo autônomo, é sempre recomendando a participação do especialista do domínio para validar o resultado, de forma que em alguns casos essa validação possa servir como ajuste dos parâmetros iniciais do algoritmo (SILVA; PERES; BOSCARIOLI, 2016; CASTRO, 2016).

3.4 REDES COMPLEXAS

O estudo de redes complexas se utiliza de conceitos de estatística, sistemas dinâmicos e teoria dos grafos (MONTEIRO, 2014). Muitas áreas têm se interessado nesse estudo como física, matemática, biologia e sociologia, devido a suas aplicações sobre uma grande variedade de problemas, os quais incluem redes sociais, redes biológicas, Internet e World Wide Web e redes de energia elétrica (NEWMAN, 2005; BARABÂSI A.-L.; JEONG, 2002; STROGATZ, 2001). As características topológicas dessas redes não são triviais, nem completamente regulares e nem completamente aleatórias, por isso são denominadas de redes complexas. Medidas são utilizadas para caracterizar a estrutura dessas redes. A utilização das medidas são, frequentemente, usadas para analisar as propriedades estatísticas que descrevem a estrutura e o comportamento de sistemas em rede, enquanto a criação de modelos de rede está normalmente relacionada ao entendimento do significado dessas propriedades.

Com o desenvolvimento das ferramentas computacionais e a visível ruptura dos limites disciplinares, a teoria dos grafos passou a ser cada vez mais utilizada na modelagem de viés dos sistemas complexos. Isto se deu no início de uma época de síntese do conhecimento científico, ressaltando diversas interconexões entre as mais distintas áreas. Uma forma

de analisar as interconexões é utilizando as medidas de centralidades, que avaliam as características de uma estrutura da rede modelada.

A utilização de redes complexas para análise de dados, é geralmente feita por grafos, onde cada vértice representa um nó da rede (pessoa, animal, empresa, objeto, elemento) e cada aresta representa qualquer relação arbitrária entre entidades (GRANDO, 2015).

A definição de um grafo é dada por $G = V, A$, onde V define o conjunto de nós ou vértices e A é o conjunto de arestas, conexões ou ligações (MONTEIRO, 2014; NETTO, 2001).

Num grafo não direcionado ou não dirigido, a aresta que conecta dois vértices quaisquer i e j também conecta j a i; portanto, o par (i, j) é não ordenado. Num grafo direcionado ou dirigido, a presença de aresta que parte de i e chega a j não implica a existência de aresta que parte de j e chega a i; ou seja, o par(i,j) é ordenado.

Como estudado neste trabalho, o espaço de saída de do algoritmo SOM pode ser interpretado como uma matriz adjacente de neurônios, por essa característica entende-se que é possível modelar assim um grafo dando assim a possibilidade de avaliar a conexão de neurônios por medidas de centralidades. Por esta metodologia pretende-se estudar o resultado do algoritmo SOM como uma análise e rede, dando assim a possibilidade de obter informações importantes sobre seus elementos e suas interações.

3.4.1 TEORIA DOS GRAFOS

A Teoria dos Grafos é um ponto importante deste trabalho, por esse ponto se faz importante contextualizar o que é um grafo. Um grafo é um par $G = V, A$ de conjuntos tal que os elementos de V são seus vértices e os elementos de A, suas arestas.

A representação de um grafo visualmente é feita da seguinte forma: cada vértice é indicado por um ponto e cada aresta é indicada por linhas conectando dois pontos. A figura 9 é um exemplo de grafo com 4 nós.

A figura 9 ilustra a modelagem de um grafo. As conexões deste grafo G, pode ser lido da seguinte forma: $G1(a, b)$, $G2(b, c)$, $G3(c, d)$, $G4(d, a)$. O entendimento da formação de um grafo por uma matriz adjacente é importante para este trabalho, justamente porque é a partir de uma matriz adjacente, espaço de saída do SOM, que toda a aplicação das

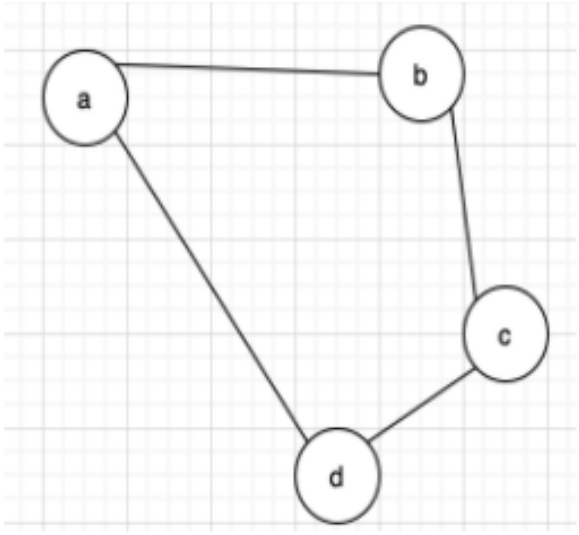


Figura 9 – Exemplo de um grafo.

	a	b	c	d
a	0	1	0	1
b	1	0	1	0
c	0	1	0	1
d	1	0	1	0

Tabela 1 – Matriz de adjacência do exemplo grafo G

medidas de centralidades para o estudo da formação do agrupamento se faz necessário.

A matriz de adjacência guarda informações sobre todas as relações de adjacência de uma rede. Se ela possui n vértices, sua matriz de adjacência M_n,n é construída da seguinte forma:

$M_{i,j} =$ 1, se há uma aresta entre os vértices i e j do grafo. 0, caso contrário.

Com a representação da figura 9, a matriz adjacente pode ser criada como:

Para este trabalho, a matriz de adjacência que formará o grafo G a ser estudado é uma matriz ponderada. A ponderação será os pesos calculados conforme estudado em materiais e métodos. A fim de dar a teoria de um grafo ponderado é quando as arestas possuem um peso. Usando o exemplo da da tabela 1 poderia ficar conforme a tabela 2.

Pode ser estudado aqui duas definições importantes dentro da Teoria dos Grafos estudados neste trabalho. São elas:

- Um grafo é dito como ponderado quando se associa um valor (normalmente, um

	a	b	c	d
a	0	3	0	5
b	2	0	6	0
c	0	10	0	8
d	12	0	12	0

Tabela 2 – Matriz de adjacência do exemplo grafo G ponderado.

número real), conforme tabela 2, a cada uma de suas arestas. Este valor é denominado comumente de peso da aresta.

- Um grafo é dito ser desconexo se há pelo menos um par de vértices para o qual, partindo de um deles e atravessando qualquer sequência finita de arestas, não é possível atingir o outro. Caso esta propriedade não valha para nenhum par de vértices no grafo, ele é conexo.

Por fim, introduziremos o conceito de isomorfismo entre grafos, dois grafos GV, A e $G'V', A'$ são isomorfos se somente for possível obter um a partir do outro apenas via remuneração de seus vértices. Mais formalmente, GV, A e $G'V', A'$ são isomorfos se e somente se existir uma função bijetora (isomorfismo) entre V e V' que preserve as relações de adjacência de G e G'. Uma rede é um grafo utilizado para representação de um sistema complexo.

3.4.2 MEDIDAS CENTRALIDADES DE REDES COMPLEXAS

A análise de uma rede envolve entender aspectos importantes como quais são os vértices mais importantes ou centrais. As medidas de centralidade são uma forma de quantificar essa importância. As medidas de rede são utilizadas para inferir um grafo e destacar a relevância de uma aresta, de um vértice ou definir as características e o tipo de grafo (MONTEIRO, 2014; GRASSI R., 2009).

Uma medida de centralidade de um nó é definida com a função $c_x : V \rightarrow$ que possui certas propriedades. Avaliando as propriedades da função, c_x é um índice estrutural de G, ou seja, G e H são grafos isomorfos por ϕ, então $c_{x(\phi(v))} = c_{x(v)}$, $\vee v$ pertencente a $\vee$. O outro ponto importante nas propriedades das medidas de centralidades é estudar a relação de ordem entre $c_{x(vi)}$ e $c_{x(vj)}$, que deve refletir a percepção de que v_i é mais central

que v_j, em algum sentido. Na maioria das medidas, quanto mais central for o vértice v, maior o valor de $c_{x(v)}$, porém pode valer o oposto.

A noção de centralidade, em várias aplicações, é associada à importância do elemento na estrutura. Esperam-se, por exemplo, altos índices de centralidades de nó para uma pessoa influente num certo círculo social, uma pessoa com cargo de chefia em uma organização, uma página da internet com muitos visitantes ou um roteador que media um grande fluxo de dados. analogamente, para as avenidas principais de uma cidade, reações químicas chave em uma célula ou fibras óticas ligando continentes, são esperadas altas centralidades de arestas.

Um grafo S_n, onde um nó central v_0 é conectado aos n demais nós sem nenhuma conexão entre estes últimos, é um exemplo onde a contratilidade onde a ela é óbvia. O exemplo pode ser visto na figura 10.

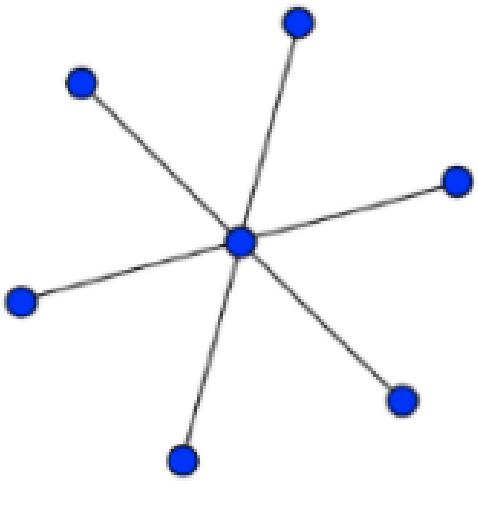


Figura 10 – Exemplo. Grafo estrela S_6

Para redes em geral, o grau nem sempre capta adequadamente a importância de um nó. O grafo na figura 11 é um exemplo. Quando analisada a figura, o grau em destaque, preenchido, apresenta apenas o valor de grau 2 porém ao observar a figura 11 é possível analisar que este no acaba sendo responsável por realizar a ligação entre os demais nós da rede. Em relação aos demais nos, está mais próximo, em média, de um outro no qualquer, ou seja, ele exibe maior proximidade. Além disso, qualquer caminhos entre um dos 4 nós à esquerda e um dos 4 à nós direita (ou vice versa) passa por v, conferindo a v um caráter de intermediação.

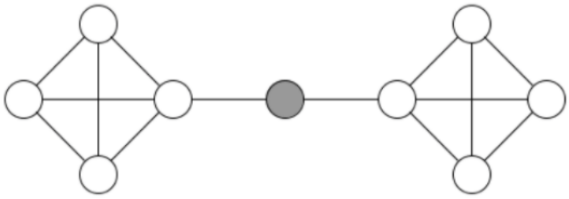


Figura 11 – No (preenchido) em destaque tem papel central, ainda que seu grau seja mínimo

As características estudadas das medidas de centralidades são pontos importantes para serem observados em um agrupamento gerado pelo algoritmo SOM. Para analisar a formação da descoberta do agrupamento, nesta pesquisa foram estudadas as de Grau, Betweenness e Proximação.

3.4.2.1 Centralidade de Grau

A centralidade de grau (em inglês *degree centrality*) é considerada a mais simples de todas as medidas, ela avalia a importância de um nó no grafo, analisando a quantidade de nós a que ele é ligado, ou seja, quanto maior o número de nós ligados a este, maior a importância dele e, portanto, maior o valor atribuído para este (BORGATTI S. P., 1999). A centralidade k do *i*ésimo vértice é calculada como (GRASSI R., 2009):

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{3.4}$$

sendo que i e j são elementos da matriz adjacente A da rede e n o número de vértices na rede.

3.4.2.2 Centralidade de intermediação (*betweenness*)

A medida de intermediação (em inglês *betweenness centrality*) consiste em avaliar a importância de um nó na transmissão de mensagens ou eventos entre os demais, ou de maneira equivalente, como ele se encontra no caminho entre os outros vértices da rede se quiserem trocar informações (NEWMAN, 2005; GRANDO, 2015). A intermediação de

um nó i é dada por:

$$g(v) = \sum_{i \neq v \neq j} \frac{g_{ij}(v)}{g_{ij}} \tag{3.5}$$

onde $g_{ij}(v)$ é o número total de menores caminhos do vértice i para o j e g_{ij} é o número de menores caminhos que passam por v.

3.4.2.3 Centralidade de proximidade (*closeness*)

A centralidade de proximidade (em inglês, *closeness centrality*) proposta por (GRASSI R., 2009) visa avaliar o quanto um determinado vértice esta distante dos demais. Assim, os vértices que possuírem uma menor distância média comparados com os demais, receberão um valor alto para a centralidade (GRASSI R., 2009).

A centralidade de proximidade é calculada como sendo (GRASSI R., 2009):

$$C_i = \frac{1}{l_i} \tag{3.6}$$

$$l_i = \frac{1}{n-1} \sum_{j(\neq i)} d_{ij} \tag{3.7}$$

sendo que n representa o número total de vértices na rede; d_{ij} é o comprimento do menor caminho entre os vértices i e j; l_i representa a média do comprimento das menores distâncias entre i e todos os outros vértices da rede.

4 MÉTODO PROPOSTO: COMBINAÇÃO DO SOM COM REDES COMPLEXAS

O método proposto nesta pesquisa está ilustrado na Figura 12. Na figura mais à esquerda o resultado do SOM está ilustrado com um reticulado em 2D, em que se procurou ilustrar como a manutenção topológica dos dados acontece, permitindo que casos parecidos permaneçam entre neurônios próximos. É possível observar no exemplo que em cada quadrante, as barras de cores representam o agrupamento dos dados após a execução do SOM.

Como se pode notar, o reticulado é uma grade fixa em que para casos em que os neurônios vizinhos representam objetos semelhantes, a interpretação e formação de grupos é natural de fazer. Contudo, para situações em que começam a aparecer dissimilaridade entre os objetos isso pode levar a falsas interpretações pelo fato dos neurônios serem adjacentes no reticulado.

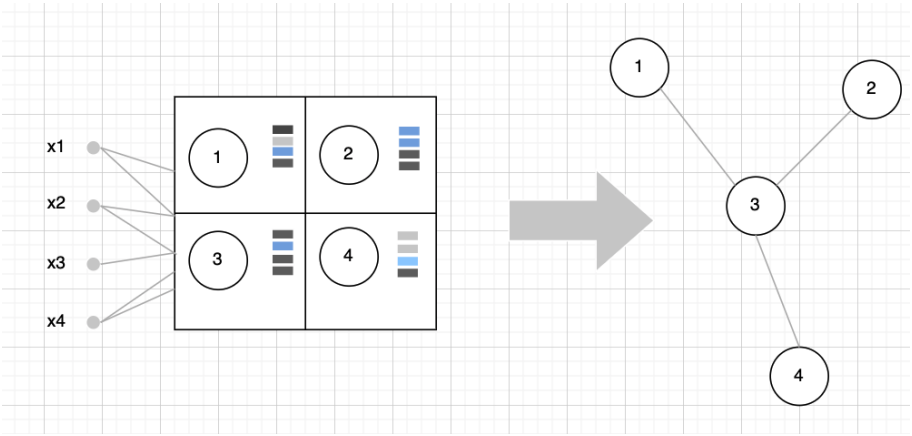


Figura 12 – Representação do processo de transformação do SOM para o grafo.

A proposta deste trabalho é que após treinamento do SOM, o mapa seja transformado em um grafo não direcionado, onde cada neurônio será representado por um vértice e as arestas representada da seguinte maneira:

$$a_{i,j} = \frac{1}{dist(\mathbf{w}_i, \mathbf{w}_j)} \tag{4.1}$$

sendo $a_{i,j}$ a aresta entre os vértices i e j, dois neurônios do mapa, $dist()$ a distância Euclidiana. A proposta de medir pelo inverso da distância é para que menores distâncias sejam transformadas em maiores pesos.

Além da proposta na forma de calcular a aresta, este trabalho ainda contempla a eliminação de arestas com valores pequenos. De forma empírica arestas com quartil menor a 25% serão removidas.

4.0.1 ELIMINAÇÃO DE ARESTAS INCONSISTENTES

Essa é uma parte importante da combinação do algoritmo, pois é nesse momento que ocorre a transformação do espaço de saída do SOM em algum modelo de grafos. Basicamente, o processo de eliminação da arestas busca as arestas que tenham uma distância grande entre os vértices i e j. Isto porque distâncias muito grande entre os vértices para o resultado do SOM representam neurônios que não estão próximos ou seja não seriam vizinhos próximos.

Os passos para a eliminação são descritos abaixo:

1. Dado um mapa treinado, obtenha o inverso das distâncias entre os pesos dos neurônios adjacentes i e j, $d(w_i, w_j)$. O cálculo para o inverso das distâncias pode ser observado na equação 4.1.

2. Neste passo todas as, arestas entre os vértices receberam um peso, conforme calculado no passo 1.

3. A remoção das arestas é feita usando o cálculo de quarties dos pesos das áreas. Para este trabalho foi considerado que o corte deveria ser de 25 porcento, isso faz com que os menores pesos entre as arestas sejam removidos. Um ponto importante a ser destacado neste item é que para trabalhos futuros é importante buscar outras heurísticas para a renovação da arestas ou até mesmo outros valores de quarties.

4. Modelar o grafo final somente com os vértices e suas arestas que não foram removidas.

5 MATERIAIS E MÉTODOS

A Figura 13 apresenta o processo completo de como se dá o experimento usando a proposta apresentada neste trabalho. Todo o processo inicia com a base de dados, depois o mapa SOM é parametrizado com as dimensões do mapa, que neste trabalho será de $3x3$. Uma vez que o mapa é treinado calcula-se as arestas, conforme o capítulo 4, com as arestas calculadas é possível modelar o grafo que será usado para aplicar os estudos de sistemas complexos, centralidades. A interpretação da formação dos grupos é apoiada por uma metodologia que permite visualizar por ferramentas gráficas do SOM, tabelas e grafos.

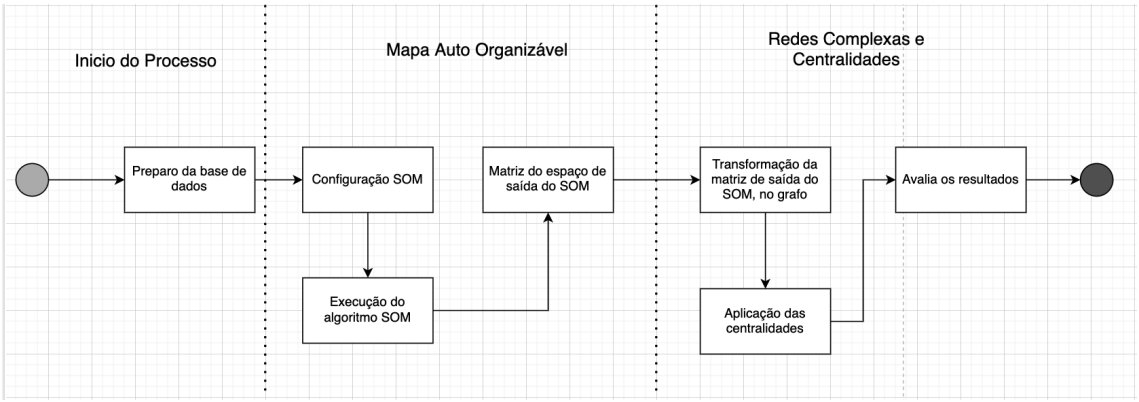


Figura 13 – Representação do processo de transformação do SOM para o grafo.

A construção destes experimentos será codificada com a linguagem de programação R, com as seguintes bibliotecas:

- *SOMbrero*: que é uma biblioteca da linguagem R para treinamento da rede SOM (OLTEANU; VILLA-VIALANEIX, 2015; OLTEANU; VILLA-VIALANEIX; COTTRELL, 2012; MARIETTE et al., 2017)
- *Igraph*: que contém métodos para geração de grafos e execução dos cálculos de centralidades (CSARDI; NEPUSZ, 2006)

Para realizar esta pesquisa foram usados três conjuntos de dados: Animal (DUA; GRAFF, 2017), *Wines*, (FORINA M. ET AL, 2003) e um terceiro conjunto de dados de uma empresa da área de energia.

5.1 CONJUNTO DE DADOS ANIMAL

O conjunto de dados "Animal"tem um total de 101 objetos definidos em 16 atributos: 1 (número, "*número de pernas*"e 15 binários, como "tem pernas", "voa"e etc). A distribuição dos objetos por classes está divida conforme tabela 3. A legenda dessas classes pode ser lida como: 1-Mamíferos, 2-Aves, 3-Répteis, 4-Peixes, 5-Anfíbios, 6-Insetos e 7-Moluscos e Crustáceos.

Classe	1	2	3	4	5	6	7
Distribuição	41	20	13	13	4	8	10

Tabela 3 – Classe dos animais e Distribuição.

O conjunto de dados "Animal"é considerado "bem comportado", com boa separação entre as classes.

5.2 CONJUNTO DE DADOS DE VINHOS (*WINES*)

Os dados do conjunto de dados de Vinhos,(*Wines*), são resultado de uma análise química de vinhos cultivados na mesma região da Itália, e derivados de três cultivos diferentes. Durante a análise foram determinadas as quantidades de grupos nós quais esses cultivos foram separados.

O conjunto de dados de Vinhos (FORINA M. ET AL, 2003), contém 178 linhas, com 13 atributos: Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline. Além desses atributos, o conjunto de dados contém o atributo class responsável por distribuir as linhas em classes de vinhos. A distribuição das linhas nessas classes pode ser observado na tabela 2.

Classe	1	2	3
Distribuição	59	71	48

Tabela 4 – Classes da base de Vinhos e sua Distribuição.

A figura 14 demonstra a distribuição dos dados em relação as três classe de cultivos ao qual o conjunto de dados foi disponibilizado. A analise deste figura é feita por cada

um dos 13 atributos disponibilizados para estudo, nela pode ser observado como os dados estão distribuídos, olhando por uma dimensão de 2D. A dispersão dos dados está bem densa, conforme pode ser observado.

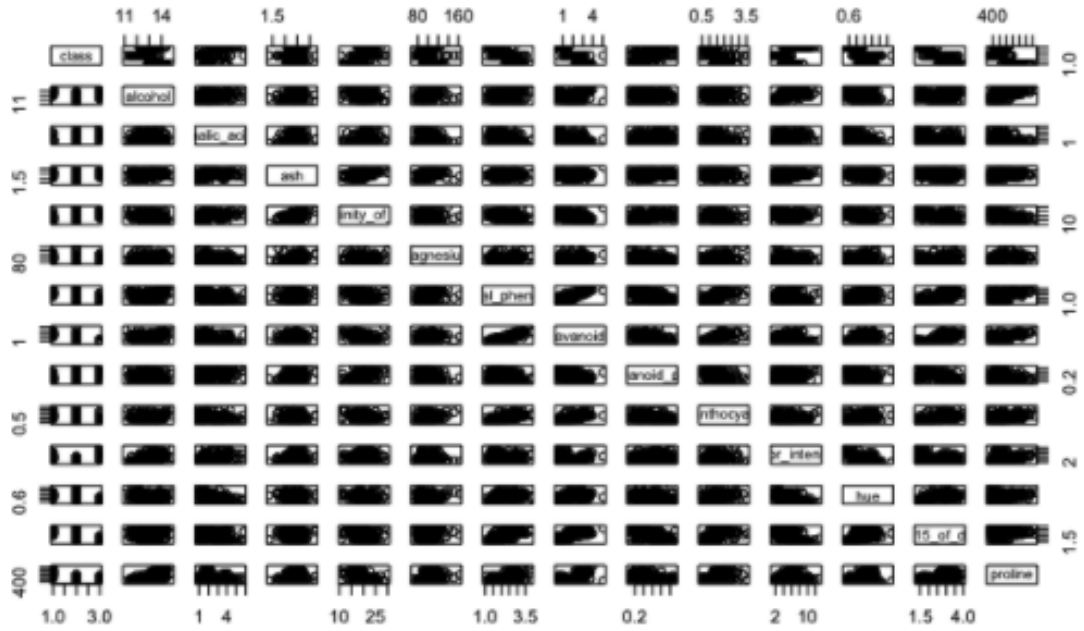


Figura 14 – Dispersão dos dados da tabela **Wines**.

Assim como foi usado nos demais experimentos deste trabalho, para a base de Vinhos também será aplicado a configuração 3 x 3 no SOM. Neste conjunto de dados, a distribuição está em 3 categorias, e com a aplicação do experimento pretende-se estudar a distribuição em 9 grupos, diferente da base original, na sequência com a aplicação de sistemas complexos para avaliar como esses grupos se formaram.

5.3 CONJUNTO DE DADOS CLIENTES

O conjunto de dados de clientes se refere a dados de uma empresa de gás. Para este experimento foi utilizado uma base com 29999 objetos. Esses objetos estão distribuídos pelo atributo classe social, que pode ser observado na tabela 5.

Classe	0	1	2	3	4	5	6	7
Distribuição	352	565	6237	2633	12345	7301	238	328

Tabela 5 – Classe dos clientes e Distribuição.

É importante observar que, diferentemente dos demais experimentos apresentados neste trabalho, este experimento contém atributos que são categóricos. Na tabela 6 é possível observar a distribuição desses dados.

Tipo Contrato	**Distribuição**
1	19298
2	5150
4	738
7	2244
15	344
17	313
18	917
19	995

Tabela 6 – Distribuição dos tipos contratos.

Na figura 15 são apresentados a dispersão de todos os objetos no atributos selecionados para criação do modelo.

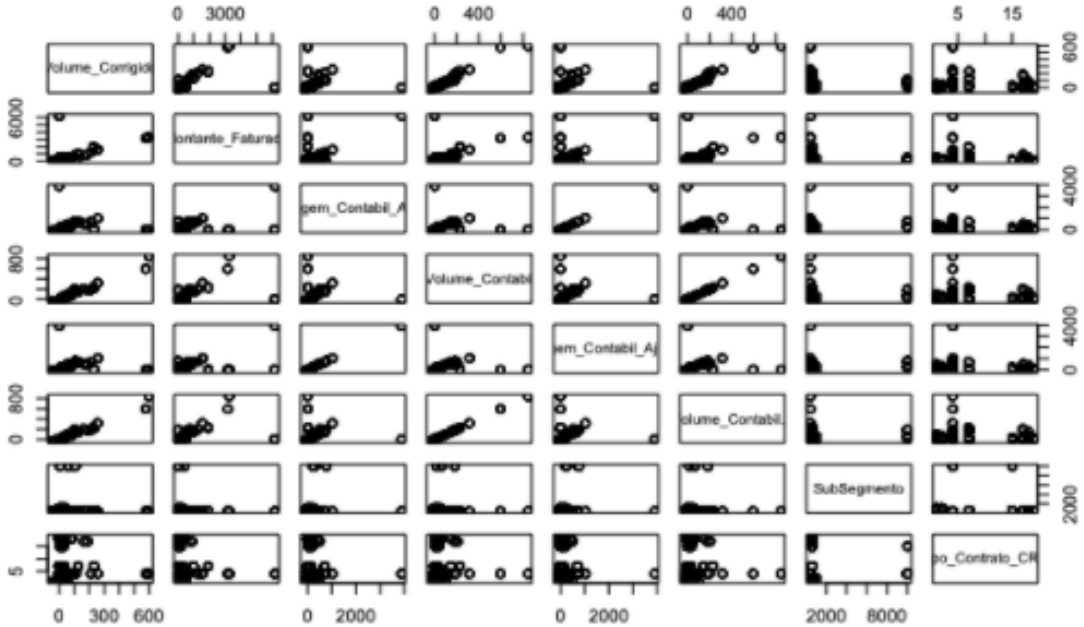


Figura 15 – Dispersão dos dados da tabela **Clientes**.

É importante observar nesta base que a distribuição dos objetos se comportam menos dispersos. Por este motivo, durante o experimento com esta base será observado como o algoritmo SOM distribui esses objetos nos neurônios formando o agrupamento. A configuração do algoritmo SOM para este experimento também será quadrada de 3 x 3.

Entende-se que para trabalhos futuros será importante testar outras configurações. Por fim, sendo uma base com objetos menos dispersos será importante observar a aplicação de sistemas complexos e assim entender como se dá a formação do agrupamento.

6 RESULTADOS EXPERIMENTAIS

Neste capítulo são apresentados os resultados realizados nas bases de dados apresentadas no capítulo de materiais e métodos.

6.0.1 RESULTADOS PARA CONJUNTO DE DADOS ANIMAL

A Figura 16, embora seja para apresentar os resultados com o conjunto de dados animais, sintetiza os principais resultados deste trabalho. As Figuras 16, 17, e 18 traz três informações importantes para o entendimento do treinamento do SOM. Na figura 16 é possível observar a quantidade de neurônio e suas identificações (de 1 a 9) e a relação topologia dos neurônios adjacentes a partir da medida de distância (Euclidiana) de um neurônio e seus adjacentes. Assim, valores menores e em tons azuis-escuros significam neurônios próximos no espaço de vetores de pesos. Cores com intensidades mais claras significam neurônios mais distantes entre si. Assim, nota-se que o neurônio 3 é o mais distante de seus vizinhos 2, 5 e 6. Por sua vez, o neurônio 9 é o que tem o menor valor de distância e, por sua vez, está mais próximo dos seus vizinhos 5, 6 e 8.

Embora os resultados da Figura 16 são valiosos na descoberta de quantidade de grupos, uma vez que neurônios próximos ou distintos possam ser trabalhados no sentido de descobrir automaticamente o número de grupos, ainda não permite analisar a relação entre os grupos e a sua formação.

Por isto, há um mapeamento dos neurônios em um grafo, conforme ilustra na figura 17, em que os neurônios viram vértices e a aresta é definida pelo inverso da distância como indicado na Eq. 5. Ou seja, a Figura 18 ilustra uma matriz simétrica relacionando a aresta de um vértice com outros. A partir do grafo (Figura 17) pode-se fazer algumas associações com a matriz-U como é o caso do neurônio/vértice 3 que é o mais afastado. Isso para o SOM indica a maior distância e para o grafo o menor peso. Contudo, a vantagem que se destaca na neste trabalho é justamente a relação dos vértices com os demais.

Neste sentido, justifica-se o uso das medidas de centralidade apresentadas na Tabela 7. Estas medidas ajudam a explicar o relacionamento da formação dos grupos com será

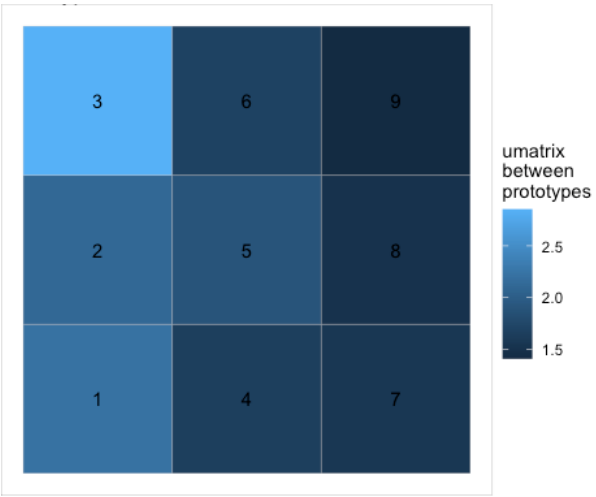


Figura 16 – U-Matrix cuja intensidade de cores indica a distância média entre vetores de pesos do neurônios adjacentes.

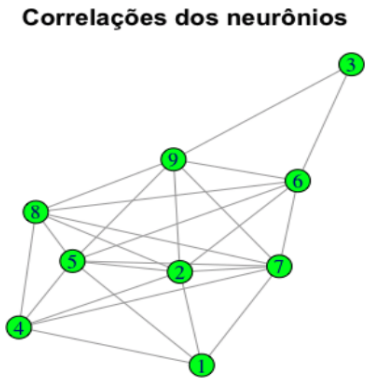


Figura 17 – Grafo tendo com vértices os neurônios do mapa SOM, cuja numeração vem da figura 16

Vértices	1	2	3	4	5	6	7	8	9
1	0	0,409	0	0,733	0,353	0	0,369	0	0
2	0,409	0	0	0,642	0,928	0,464	0,604	0	0,41
3	0	0	0	0	0	0,551	0	0	0,36
4	0,733	0,642	0	0	0,556	0	0,724	0,463	0
5	0,353	0,928	0	0,556	0	0,536	0,596	1,406	0,58
6	0	0,464	0,551	0	0,536	0	0,347	0,636	0,97
7	0,369	0,604	0	0,724	0,596	0,347	0	0,629	0,33
8	0	0	0	0,463	1,406	0,636	0,629	0	0,69
9	0	0,412	0,363	0	0,576	0,968	0,33	0,685	0

Figura 18 – Resultado experimental com a base animais apresentado pelo reticulado da rede SOM junto com as cores da U-Matrix, a geração do grafo e, por fim, o mapa de cores (Heatmap) com as arestas do grafo 17.

Vértice	1	2	3	4	5	6	7	8	9
Grau	7	5	7	7	6	6	6	4	2
Intermediação	2	0.25	2	2	3	0.75	3	0	0
Proximidade	0.111	0.083	0.111	0.111	0.100	0.100	0.100	0.077	0.062

Tabela 7 – Medidas de Centralidades.

neurônio	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
número de objetos	21	0	14	2	7	8	20	0	29
Distribuição dos objetos por classes	4(3), 7(2), 1(13), 3(3)	–	6(8), 7(6)	7(2)	1(7)	5(4), 1(2), 3(2)	2(20)	–	9(29)

Tabela 8 – Distribuição de objetos e classes por neurônios.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
pelo	***0.05***	0	0.29	0	***1***	0.25	0	0	1
penas	0	0	0	0	0	0	1	0	0
ovos	**0.81**	0	0.03	1	0	0.88	**1**	0	0
leite	***0.14***	0	0	1	***1***	0.25	0	0	1
voador	0	0	0.43	0	0.29	0	0.8	0	0
aquatico	**0.86**	0	0.36	0	**0.14**	0.62	**0.3**	0	0.03
predador	**0.81**	0	0.5	0	**0.29**	0.62	**0.45**	0	0.55
dentado	***0.9***	0	0	0	***1***	0.75	0	0	1
respira	**0.24**	0	0.64	1	**1**	1	**1**	0	1
venenoso	0.19	0	0.21	0	0	0.12	0	0	0
barbatana	***0.76***	0	0	0	***0.14***	0	0	0	0
pernas	0	0	2,4	0	2	4	2	0	4
rabo	**0.86**	0	0.07	0	**0.71**	0.5	**1**	0	0.93
domestico	**0.05**	0	0.07	0	**0.14**	0.12	**0.15**	0	0.21
catsize	**0.33**	0	0.07	0	**0.57**	0.25	**0.3**	0	0.83

Tabela 9 – Vetores de pesos após treinamento da rede SOM.

apresentado a seguir. Antes disso, cabe apenas apresentar que cada neurônio e vértice tem uma densidade de objetos agrupados. A distribuição desta densidade (objetos por neurônio/vértice) está apresentada na Tabela 8.

Analisando cada um dos vértices, mais especificamente os objetos que foram mapeados em cada neurônio da rede, conforme tabela (8) observa-se que o vértice 1 é formado por animais de classe peixe (13), moluscos e crustáceos (2), mamíferos (3) e répteis (3). Ana-

lisando as conexões e arestas (Figura 18) é possível observar que as conexões acontecem com os vértices:

- 2: não tem formação de grupos
- 4: grupo formado predominantemente com moluscos e crustáceos;
- 5 grupo formado predominantemente com mamíferos;
- 7 grupo formado predominantemente com anfíbios.

Por sua vez, a Tabela 9 possibilita entender as características de cada neurônio por meio dos valores após treinamento do vetor de pesos. Por exemplo, analisando as características entre neurônios 1 e 5 nota-se que tem em comum pelo, leite, dente, barbatana. Portanto, características de mamíferos, justificando assim a relação com aresta de 0.353. Por sua vez, a relação de 1 e 7, com aresta de 0.369 a característica em comum é o ovo.

Em relação às centralidades estudadas nesta pesquisa, os vértices 1, 3, e 4 apresentaram maiores valores de grau e proximidade, vide Tabela 7. Pegando novamente o exemplo do vértice 1, formado por peixe, moluscos e crustáceos, mamíferos e répteis, pode-se concluir que esses animais são os que mais se destacam como importantes para rede por serem os que mais se aproximam dos demais animais. Já para a centralidade de intermediação os vértices 5 mamíferos e vértice aves, logo pode se concluir que esses animais são considerados responsáveis por estarem mais nos caminhos da formação dos demais grupos do grafo.

Com esse experimento pode ser concluído que o cruzamento entre as técnicas é possível, porém uma observação nesse experimento é que, por exemplo, no neurônio 1 existe características com baixo valor estatístico, isso pode ser observado na figura 9. Neste ponto, entende-se que para trabalhos futuros é importante estudar uma heurística que possa eliminar objetos que estejam fora de contexto no grupo.

6.1 RESULTADOS PARA BASE DE VINHOS

Como apresentado no capítulo de materiais e métodos, o conjunto de dados de Vinhos é referente a valores químicos de vinhos cultivados em uma mesma região na Itália. Cate-

goricamente esse conjunto de dados está dividido em três classes, que pode ser observado na tabela 4.

Assim como nos experimentos anteriores, a figura 19, 20 e 21 traz três informações importantes para o entendimento deste experimento. Assim como no experimento anterior, foi utilizado uma dimensão quadrada 3 X 3, a fim de comparar melhor os resultados. Na figura 19, pode-se analisar os neurônios plotados em forma de uma matriz. A leitura desta matriz pode ser feita como: cores com intensidades mais claras significam neurônios mais distantes entre si. Diante desta explicação entende-se: o neurônio 9 é o com menor valor de distância e, por sua vez, está mais próximo dos seus vizinhos 5, 6 e 8.

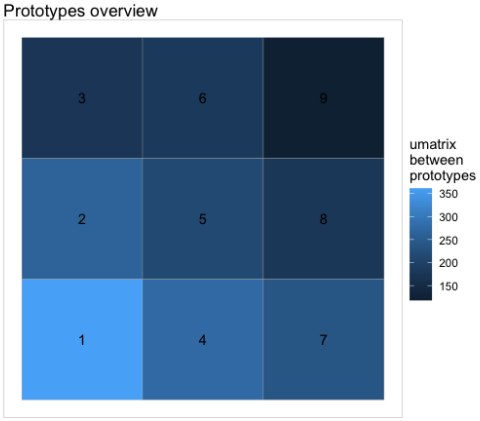


Figura 19 – U-Matrix cuja intensidade de cores indica a distância média entre vetores de pesos do neurônios adjacentes.

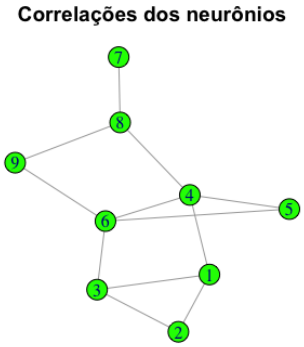


Figura 20 – Grafo tendo com vértices os neurônios do mapa SOM, cuja numeração vem da figura 19

Como apresentado durante o desenvolvimento desta pesquisa, mesmo os resultados apresentados 19, são valiosos na descoberta da quantidade de grupos ideias para este

Vértices	1	2	3	4	5	6	7	8	9
1	0	0.007	0.003	0.002	0	0	0.001	0	0
2	0.007	0	0.006	0	0	0	0	0	0
3	0.003	0.006	0	0	0	0.005	0	0	0
4	0.002	0	0	0	0.069	0.020	0	0.004	0
5	0	0	0	0.069	0	0.029	0	0	0
6	0	0	0.005	0.020	0.029	0	0	0	0.009
7	0.001	0	0	0	0	0	0	0.014	0
8	0	0	0	0.004	0	0	0.014	0	0.010
9	0	0	0	0	0	0.009	0	0.010	0

Figura 21 – Resultado experimental com a base de Vinhos apresentado pelo reticulado da rede SOM junto com as cores da U-Matrix, a geração do grafo e, por fim, o mapa de cores (Heatmap) com as arestas do grafo.

conjunto de dados, uma vez que neurônios próximos ou distintos possam ser trabalhados no sentido de descobrir automaticamente o número de grupos, isso não permite analisar a relação entre os grupos e a sua formação.

A fim de construir o modelo para trabalhar o entendimento da formação dos grupos, o grafo 20 é apresentado. Assim como nos experimentos anteriores, é possível observar uma relação entre a análise do grafo e a matriz-U. Por esta observação, a análise feita na matriz-U para este experimento se confirmar quando usado o grafo mostrando o neurônio/vértice 3 como o mais afastado de todos, essa análise ficará mais clara quando se observado a centralidade de proximidade.

A formação do grafo 20 é interessante para mostrar a relação por arestas entre os neurônios/vértices, e junto deste grafo é possível observar os valores de pesos das arestas que estão disponíveis no mapa de calor, 21. Importante entender que pela biblioteca o R usada, conforme capitulo de materiais e métodos, as arestas graficamente não mostram a distancias dos neurônios/vértices e sim somente a relação deles, conforme relatado. Uma análise que pode ser feita visualmente é observar os neurônios/vértices que se relacionam, como por exemplo o neurônio/vértice 7 que só se relaciona com o neurônio/vértice 8. Esse tipo de análise pode ajudar o especialista de negócio a entender e até mesmo tomar decisões quanto a que tipo de relação os grupos podem exercer uns sobre os outros somente analisando as arestas do grafo. Outro fato muito importante ao estudar a arestas é observar os pesos que cada uma tem, pois assim pode gerar uma maior importância que os grupos têm. Outro exemplo que pode ser analisado é do neurônio/vértice 1 que se relaciona com os neurônios/vértices 2, 3, 4 onde a característica marcante dessa relação é que ambos apresentam um alto teor de álcool, conforme pode ser observado na tabela 11.

A análise das centralidades é apresentada na tabela 10. Antes de analisar as medidas

Vértice	1	2	3	4	5	6	7	8	9
Grau	2	3	4	4	2	3	2	3	1
Intermediação	0	11	11	9	0	9	0	12	0
Proximidade	0.134	0.104	0.155	0.118	0.121	0.143	0.114	0.068	0.043

Tabela 10 – Medidas de Centralidades.

de centralidades, vale entender a densidade de objetos agrupados em cada neurônio/vértice. A distribuição desta densidade (objetos por neurônio/vértice) está presente na Tabela 11.

Pode ser observado neste experimento, que diferente do experimento com os dados da base de Animal, todos os grupos foram preenchidos. O estudo deste experimento demostrou que no futuro pode ser estudado outro fenômeno ao se estudar modelagens por redes complexas, a existência de grupos que se relacionam entre si, gerando pequenos grafos com alta conexão. Na figura 20, isso por ser observado entre os neurônios 3, 9 e 6.

Em relação às centralidades, neste experimento na centralidade de grau os neurônios 3 e 4 apresentaram os maiores valores. Na centralidade de proximidade o neurônio 4 apresentou o maior valor. Na centralidade de intermediação o neurônio 8 apresentou o maior valor. Novamente, analisando as características dos neurônios concluímos que: em relação a centralidade de grau, mostra que vinhos com maior valor de álcool e cinzas são mais relevantes. Nas centralidades de proximidade a característica de valor de álcool chamam mais atenção, mas também a característica de fenóis não flavonóides, que a partir do neurônio 4 em diante, começa a sempre aumentar. Por último, analisando a centralidade de intermediação é possível encontrar vinhos que tem um baixo valor de álcool em relação aos demais grupos. Logo pode se considerar que essas características são consideradas responsáveis por estarem mais nos caminhos da formação dos demais neurônios do grupo.

6.2 RESULTADOS PARA CONJUNTO DE DADOS CLIENTES DE ENERGIA.

O detalhamento da base de clientes foi feito durante o capítulo de materiais e métodos, pode ser observado que o agrupamento natural desta base é baseado no atributo, classe social (CS). Não faz parte do escopo desta pesquisa detalhar a formação deste atributo.

	1	2	3	4	5	6	7	8	9
alcohol	***13.86***	13.76	13.14	13.34	***13.43***	12.67	12.37	12.71	12.76
malic acid	1.79	2.02	2.04	2.8	3.01	2.7	2.2	2.71	2.55
ash	**2.51**	2.35	2.41	2.46	2.41	2.33	**2.29**	2.27	2.4
alcalinity of ash	***17.07***	16.4	19.56	19.54	***18.83***	19.38	20.88	20.26	21.38
magnesium	106	102.95	112.06	111.43	112.67	100.17	91.64	92.78	95.59
total phenols	**2.94**	2.75	2.53	2.24	**2.08**	2	**2.32**	1.74	1.99
flavanoids	**3.11**	2.91	2.29	1.9	**1.4**	1.5	**2.16**	1.35	1.28
nonflavanoid	***0.3***	0.27	0.32	0.36	***0.41***	0.42	0.37	0.4	0.41
proanthocyanins	**1.93**	1.87	1.78	1.66	**1.54**	1.46	**1.55**	1.3	1.38
color intensity	6.26	5.22	5.09	5.67	5.89	5.36	3.43	5.03	5.52
hue	***1.1***	1.05	0.98	0.87	***0.84***	0.94	1.02	0.85	0.83
oD280.OD315	3.04	3.18	2.88	2.79	2.14	2.3	2.74	2.16	2.18
proline	**1338.57**	1072.75	905.5	802.86	**750**	694.28	**395.08**	508.91	609.38

Tabela 11 – Vetores de pesos após treinamento da rede SOM.

Vale ressaltar que essa classificação é conforme regras da empresa, na avaliação dos resultados deste trabalho será analisado de um ponto de vista diferente. Será visto pelos atributos descritos no capítulo de materiais e métodos. A execução do algoritmo SOM criará o agrupamento e com as medidas de centralidades será avaliado a descoberta deste agrupamento e posteriormente a explicação de como esse agrupamento foi formado pelos sistemas complexos.

Importante antes de realizar a análises do resultado do trabalho implementado, se faz necessário entender estatisticamente como ficou a distribuição dos objetos nos neurônios. Na tabela 16 pode ser analisado como ficou a distribuição, a leitura dessa tabela pode ser feita da seguinte forma:

- O neurônio que mais recebeu objetos foi o neurônio 3, com 33,92 porcento dos objetos. Também é o que tem o maior volume corrigido com 10.78 em média. O valor de margem não é o maior com 38.25, porém traz o maior volume contábil, o que faz sentido visto que também tem o maior volume corrigido em média.
- O neurônio 4 é o que menos recebeu objetos com 72, logo isso representa 0,24 porcento dos objetos, mas não tem o menor valor corrigido 6.41, mas por ser um grupo pequeno tem o menor faturamento com 44.3.

Para descrever o resultado foi feito apenas a análise dos grupos com maior tamanho e o com menor valor. Para analisar as demais distribuições estatísticas dos neurônios/vértices,

CS	Porcentagem
0	1.32
1	2.49
2	21.88
3	8.53
4	35.54
5	28.02
6	1.59
7	0.63

Tabela 12 – Distribuição da Classe Social, no neurônio 3.

pode ser observado na tabela 16. Outro fator importante estatístico a ser avaliado é referente ao atributo classe social e tipo de contrato. Essa análise se faz importante por alguns fatores: ambos os atributos são categóricos e por isso importante entender como se deu a distribuição dos objetos, outro fator importante é a classe social por ser o atributo usado hoje para categorizar os objetos e agora como se dá a distribuição depois da execução do algoritmo, por último essas informações serão importante justamente na aplicação das medidas de centralidades pois demonstraram quais atributos e qual a formação são importante de acordo com a análise da centralidade.

Observando os neurônios 3 e 4, que são o maior e o menor em objetos respectivamente, a distribuição por esses atributos categóricos ficou da seguinte forma:

A distribuição por classe social do neurônio 3 pode ser observado na tabela 12:

Veja que por essa distribuição esse grupo se divide bastante nas classes sociais 2, 4 e 5. As demais classes sociais apresentam resultados mas não são tão expressivas.

Na análise para tipo de contrato, a tabela 13 pode ser observada, veja que a categoria 1 é a que tem mais relevância sobre as demais.

A distribuição do neurônio 4, se dá da seguinte forma:

Demonstra que a classe social 5 é a que mais popula objetos neste neurônio. Na análise de tipo de contrato, como feito para o neurônio 3, ficou da seguinte forma:

Uma vez que a análise estatística da distribuição dos objetos, inicia-se a avaliação do algoritmo SOM. Na figura 22 é possível observar os principais resultados da execução do agrupamento pelo SOM. Na figura 22 pode se observar a quantidade de neurônios e suas

TP	Porcentagem
1	70.37
2	16.99
4	1.32
7	5.89
15	0.69
17	0.96
18	1.72
19	2.06

Tabela 13 – Distribuição do Tipo de Contrato, no neurônio 3.

TP	Porcentagem
4	16.67
5	83.33

Tabela 14 – Distribuição do Tipo Contrato, no neurônio 4.

CS	Porcentagem
1	50
7	47.22
18	2.78

Tabela 15 – Distribuição da Classe Social, no neurônio 4.

identificações (1 a 9) e a relação topologia dos neurônios adjacentes a partir da medida de distância (Euclidiana) de um neurônio e seus adjacentes. Assim, valores menores e em tons azuis escuros significam neurônios próximos no espaço de vetores de pesos. Cores com intensidades mais claras significam neurônios mais distantes entre si. Assim, nota-se que neurônio 7 é o mais distante de seus vizinhos 2, 5 e 7. Por sua vez, o neurônio 4 é o que tem o menor valor de distância e, por sua vez, está mais próximo dos seus vizinhos 2, 5 e 6. Percebe-se que até este momento, a análise feita no resultado é similar aos experimentos anteriores até aqui neste momento sem o estudo de redes complexas.

Assim como citado nos experimentos anteriores, a análise da **??**a é valiosa na descoberta de quantidade de grupos. Porém, ela não permite realizar a análise das relações entre os grupos e sua formação.

Por isso, há um mapeamento dos neurônios em um grafo, conforme 23, em que os neurônios viram vértices e a aresta é definida pelo inverso da distância como indicado na

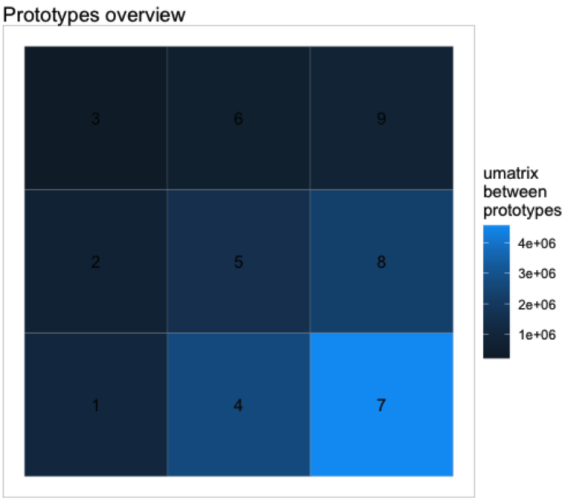


Figura 22 – U-Matrix cuja intensidade de cores indica a distância média entre vetores de pesos do neurônios adjacentes.

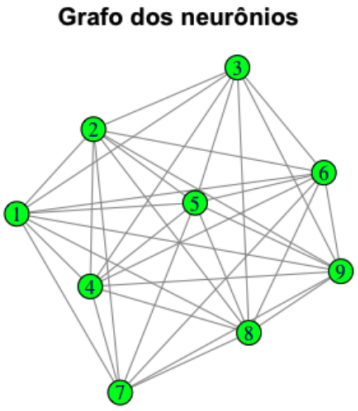


Figura 23 – Grafo tendo com vértices os neurônios do mapa SOM, cuja numeração vem da figura 22.

Vértices	1	2	3	4	5	6	7	8	9
1	0	0.04239	0.02817	0.00336	0.11668	0.03880	0.00161	0.00521	0.06904
2	0.04239	0	0.08400	0.00312	0.03110	0.45728	0.00155	0.00464	0.10985
3	0.02817	0.08400	0	0.00301	0.02269	0.10290	0	0.00440	0.04760
4	0.00336	0.00312	0.00301	0	0.00346	0.00310	0.00307	0.00950	0.00321
5	0.11668	0.03110	0.02269	0.00346	0	0.02912	0.00163	0.00545	0.04338
6	0.03880	0.45728	0.10290	0.00310	0.02912	0	0	0.00459	0.08857
7	0.00161	0.00155	0	0.00307	0.00163	0	0	0.00232	0.00157
8	0.00521	0.00464	0.00440	0.00950	0.00545	0.00459	0.00157	0	0.00484
9	0.06904	0.10985	0.04760	0.00321	0.04338	0.08857	0.00157	0.00484	0

Figura 24 – Resultado experimental com a base animais apresentado pelo reticulado da rede SOM junto com as cores da U-Matrix, a geração do grafo e, por fim, o mapa de cores (Heatmap) com as arestas do grafo.

Eq. 5. Ou seja, a Figura 24 ilustra uma matriz simétrica relacionando a aresta de um vértice com outros.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
tamanho grupo	***1786***	4745	10174	72	***2147***	4523	2428	468	3656
volume corrigido	10.58	10.78	9.73	6.41	4.82	9.45	7.91	7.39	10.01
faturado	**74.15**	73.79	69.68	44.3	145.21	68.1	**57.17**	72.49	68.01
margem	***42.65***	38.25	35.68	28.06	***54.16***	42.79	39.26	26.34	39.67
volume contábil	11.69	12.13	11.05	7.13	5.37	10.93	9.32	6.88	10.77

Tabela 16 – Características dos neurônios.

1		2		3		4		5		6		7		8		9	
CS	%	CS	%	CS	%	CS	%	CS	%	CS	%	CS	%	CS	%	CS	%
1	0.90	0	1.60	0	1.32	4	16.67	2	8.10	0	0.40	1	2.72	2	51.92	0	3.39
2	33.59	1	1.35	1	2.49	5	83.33	3	6.94	1	0.75	2	26.52	3	21.79	1	3.61
3	3.92	2	19.89	2	21.88			4	72.47	2	20.87	3	9.23	4	11.32	2	12.64
4	17.58	3	7.84	3	8.53			5	12.48	3	9.86	4	40.57	5	14.96	3	11.00
5	34.15	4	53.51	4	35.54					4	38.62	5	20.96			4	41.66
6	4.26	5	15.17	5	28.02					5	27.86					5	26.07
7	5.60	7	0.63	6	1.59					7	1.64					7	
				7	0.63												1.64

Tabela 17 – Feature Classe Social(CS) e sua distribuição nos neurônios. Importante observar que as classes sociais iniciam com O.

A partir do grafo da figura 23 é possível fazer algumas associações com a matriz-U como é o caso do neurônio/vértice 3 como sendo o mais afastado. Isso para o SOM indica a maior distância e para o grafo o menor peso. Contudo, a vantagem que se destaca na proposta deste trabalho é justamente a relação dos vértices com os demais, fazendo uma metáfora como sendo uma comunidade de grupos e como são as interações.

Uma vez que toda a análise das características do agrupamento dos neurônios foi realizada, pode ser analisada agora a formação dos neurônios pela análise do grafo e das centralidades. Para seguir na mesma linha dos experimentos anteriores, para a análise do experimento dessa base também foi utilizado um corte com quarties de 25 por cento e para esta base o corte manteve uma densidade alta no grafo de 0.972 mostrando assim uma alta conexão entre todos os neurônios. Os valores de centralidades para este experimento podem ser observados na tabela 12.

Outros dois atributos também foram analisadas para demonstrar a formação do agrupamento: Classe Social 17 e Tipo de Contrato 18.

Uma vez que toda análise das características do agrupamento dos neurônios foi realizado, pode ser analisada agora a formação dos neurônios pela análise do grafo e das centralidades. Para seguir na mesma linha dos experimentos anteriores, para a análise do

1		2		3		4		5		6		7		8		9	
TP	%	TP	%	TP	%	TP	%	TP	%	TP	%	TP	%	TP	%	TP	%
1	43.90	1	70.24	1	70.37	1	50	1	47.32	1	72.08	1	69,19	1	63,25	1	47.43
3	20.04	2	16.90	2	16.99	7	47.22	2	10.71	2	19.28	2	12,56	7	14,93	2	23
4	13.44	4	3.62	4	1.32	18	2.78	4	1.58	4	1.77	7	13,22	18	21,79	4	2.13
7	14.89	7	2.40	7	5.89			7	4.38	7	4.20	15	0,66			7	15.21
17	1.34	17	2.36	15	0.69			15	9.50	17	0.35	18	0,49			15	1.48
18	5.38	18	2.44	17	0.96			17	0.75	18	1.72	19	3,87			17	1.29
19	1.01	19	2.02	18	1.72			18	0.14	19	0.60					18	9.11
				19	2.06			19	25.62								

Tabela 18 – Feature Tipo Contrato(TP) e sua distribuição nos neurônios.

Neurônio	Grau	Intermediação	Proximidade
1	8	0	1.363
2	7	0	3.346
3	8	0	31.417
4	8	0	3.945
5	8	0	3.945
6	7	16	8.494
7	8	0	24.420
8	8	0	2.717
9	8	0	3.276

Tabela 19 – Medidas de Centralidades.

experimento dessa base também foi utilizado um corte com quartis de 25 porcento e para esta base o corte manteve uma densidade alta no grafo de 0.972. Logo para análise das centralidades é possível observar na tabela 19.

Pelo grafo apresentar um valor alto de densidade, representando assim um grafo fortemente conectado. A aplicação e centralidade de grau não se apresenta como uma boa métrica para este experimento. Como observado na tabela [19 apenas os neurônios 2 e 6 apresentam valores baixos, porém não tão baixos assim. Logo fica difícil concluir quais características seriam importantes para a rede modelada. Nesta demonstração apresenta que todos os atributos são de extrema importância.

Com a centralidade de intermediação já é possível realizar uma análise mais clara, isto porque ela demonstra que o neurônio 6 apresenta o maior valor e com isso demonstra que as características desse neurônio estão mais no caminho da formação dos demais neurônios. Ao observar as tabelas 16, 17, 18, as características desses grupos que mais se destacam são: o valor de margem representando 42.79 e valor contábil com 10.93 de média. Quando analisado as classes sociais é analisado que as classes 2, 4 e 5 mais se

destacam. No caso do tipo de contrato é observado que o tipo de contrato 1 é o com maior valor representativo com 72.08 por cento. Logo pode ser destacado como as características que estiveram mais presentes na formação dos demais neurônios.

Na centralidade de proximidade é possível trazer a afirmação da análise da matriz-U onde apontou que o neurônio 3 é o que está mais distante dos demais e isso pode ser visto também nesta análise onde ele apresenta o maior valor de proximidade de 21.417, essa análise também pode ser observada pelo neurônio 7 que também apresenta alto valor com 24.420, isso mostra que as características desses neurônios se afastam mais dos demais durante sua formação diferente dos que apresentaram valores mais baixos demonstrando assim maior proximidades.

Foi trabalhando neste experimento a possibilidade de gerar um contraste de agrupamento para a base de clientes diferente da existente hoje por classe social. Categorização pela classe social é feita puramente por uma metodologia da empresa em questão, já com o agrupamento pelo algoritmo SOM busca-se trabalhar a relação dos atributos em si a fim de se encontrar a melhor formação deste agrupamento posteriormente com a utilização de sistemas complexos busca-se entender se essa formação faz sentido ao especialista do domínio do negócio. O ponto principal desta análise foi principalmente olhar se os valores de centralidades trariam resultados coerentes, para uma base real e isto se mostrou satisfatório. O que pode ser mais explorado é justamente a modelagem do grafo ponderado. Isto porque o corte realizado para este experimento, que seguiu os demais de 25 por cento, mostrou-se um corte não interessante para a medida de centralidade de grau e com isso não permitiu uma avaliação de quais características poderiam ser mais importantes para esta rede. Em trabalhos futuros esse é um ponto importante a ser explorado.

Por fim, para este experimento uma outra análise que faria sentido, assim como nos demais experimentos é a de avaliar as ligações criadas pela modelagem do grafo. Porém esse ponto ficou prejudicado assim como na análise de centralidade de grau por ter uma alta densidade, mostrando assim alta conexão entre todos os vértices, não se pode tirar uma conclusão tão clara dessas relações.

7 CONCLUSÃO

Durante os estudos deste trabalho fica evidente o volume de dados que é gerado tanto pelo meio corporativo como pelo meio acadêmico. Com isso o estudo de técnicas avançadas que procuram entender a relação em os dados e assim tirar mais *insgiths* se torna cada vez mais utilizado no mundo.

Neste trabalho usou-se o algoritmo de Mapa Auto-Organizável, que trabalha bem com bases de dados que não se tem a necessidade de remover os *outliner*, e que seu resultado é a formação de uma matriz adjacente onde os neurônios são colocados. Combinado com Mapa Auto-Organizável, este trabalho trouxe a possibilidade da utilização de redes complexas, grafos, com medidas de centralidades, com o objetivo de estudar a relação entre os grupos e assim promover novos *insgiths* a quem esta usando as técnicas.

Na literatura é encontrado diversos estudos que buscam trazer métodos que respondam melhor ao problema de definição do número ideal de grupos a serem usados na formação do agrupamento. Este não foi o foco deste trabalho mas sim, estudar a formação dos grupos e também entender como funciona as relações no agrupamento.

Como visto durante o desenvolvimento deste trabalho, o estudo da formação dos grupos foi possível pelo fato do Mapa Auto-Organizável ter uma matriz adjacente na saída, e com existe a possibilidade de transformar a saída em uma rede, grafo. A partir deste ponto as medidas de centralidades foram usadas e assim, pode-se estudar a formação dos grupos. A analogia neste ponto pode ser feita quanto ao estudo de formações de grupos em redes sociais, onde se pode encontrar na literatura artigos que usam redes complexas e medidas de centralidades para explicar, para explicar por exemplo o comportamento social entre pessoas de diferentes classes sociais, gosto por atividades físicas e assim por diante.

Um ponto importante durante o desenvolvimento deste trabalho foi a transformação da matriz adjacente do espaço de saída em um grafo. Foram aplicados dois passos, primeiro a conversão dos pesos do espaço de saída em pesos das arestas, para este ponto foi utilizado o inverso da distância Euclidiana dos pesos a fim de transformar grandes

valores em pequenos valores. Posteriormente, aplicado uma heurística de remover arestas com valores muito pequenos, valores que poderiam representar os 25 por cento menores de todas as arestas.

Para o desenvolvimento do trabalho, foram utilizadas três conjuntos de dados, como explicado no capitulo de matérias e métodos. No conjunto de dados de animal, o entediamento da formação dos grupos seria mais simples de ser feito. O ponto que ficou claro neste experimento, foi que alguns grupos tiveram animais muito distintos sendo agrupados juntos, neste caso pode ser entendido como um ruído e com isso em trabalhos futuros pretende-se aplicar uma heurística que remova esses ruídos. No conjunto de dados de vinhos, *Wines*, existia o conhecimento prévio das classes que foram separadas conforme a região de cultivo. Neste experimento pode ser observado as características mais marcantes do vinhos que podem influenciar nós demais, como por exemplo o teor alcoílico de cada grupo de vinho. Como ultimo conjunto de dados, foi usado uma base de uma empresa de energia. Neste caso não se tinha um conhecimento prévio dos grupos apenas um atributo de classe social que foi utilizado que foi usado como guia. Por não se ter o conhecimento prévio dos grupos, este experimento foi importante para analisar o comportamento das técnicas aplicadas. Nele pode ser visto, por exemplo, quais clientes tinham uma maior utilização da energia e avaliar com quais grupos ele se relacionava, isso da uma oportunidade da empresa trabalhar melhor por exemplo suas campanhas de marketing.

Uma fator importante na conclusão deste trabalho é quanto a heurística proposta para remoção de possíveis ruídos nos grupos. Para o conjunto de animais isso mais simples de entender o ruído mas para os demais conjuntos de dados isso não foi possível, pois seria necessário a avaliação de um especialista do domínio e negocio para validar os ruídos, porém entende-se que a heurística também pode ser aplicada nos demais experimentos, isso porque se observado as tabelas de estatística dos grupos podemos observar valores pequenos de características que foram agrupadas e a ideia da heurística é justamente remover esses valores.

Como hipótese deste trabalho, foi concluído que o uso das medidas de centralidades pode ser usada como viabilizador para se entender a formação do agrupamento. Essa possibilidade se dá, pelo fato do resultado ser uma matriz adjacente, uma vez que essa matriz

se transforma em grafos, as medidas de centralidades aplicadas. Com os experimentos que foram propostos fica claro que as medidas de centralidades podem ser utilizadas e conseguem responder às hipóteses levantadas nesta pesquisa. Com este ponto a teoria de grafos com medidas de centralidades mostram que traz uma abordagem interessante ao avaliar os resultados se tornando assim uma possibilidade para utilização.

8 TRABALHOS FUTUROS

Com a conclusão deste trabalho, entende-se que em trabalhos futuros podem ser evoluídos alguns pontos:

- A remoção das arestas inconsistentes, onde neste trabalho foi usado o valor 25 por cento, é interessante aplicar outros valores pelos quarties ou até mesmo estudar outra técnica que possa ser aplicada.
- A dimensão do Mapa Auto-Organizável aplicada neste trabalho foi de 3 x 3, usar outros conjuntos de dados com dimensões maiores fará com que se tenha redes maiores e isso é importante.
- A implementação de um modelo de heurística para remoção de ruídos nós grupos. O exemplo claro fica no experimento, com o conjunto de dados de Animais, que acabou misturando características de mamíferos com características de aves. Quando analisado esse exemplo, foi percebido que a porcentagem de características de animais era bem pequena.
- Por último entende-se que a aplicação de outras medidas de centralidades aplicadas em redes sociais, pode ser interessante também como, por exemplo, centralidade de **eigenvector** e com isso não só estudar a importância do vértice mas também estudar as arestas importantes.

REFERÊNCIAS

ANKERST MARKUS M. BREUNIG, H.-P. K. J. S. M. Optics: Ordering points to identify the clustering structure. *Int. Conf. on Management of Data*, v. 28, p. 49–60, 1999.

ARRIETA, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, Elsevier, v. 58, p. 82–115, 2020.

ASSIS, E. C. d. Algoritmos de particionamento aplicados à análise estatística de formas. *Universidade Federal de Pernambuco*, 2018.

BARABÂSI A.-L.; JEONG, H. N.-Z. R. E. S. A. V. T. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, v. 311, p. p.590–614, 2002.

BERKHIN, P. Survey of Clustering Data Mining techniques. *Scientific Research*, p. 25–71, 2011.

BORGATTI S. P., E. M. G. The centrality of groups and classes. *Journal of Mathematical Sociology*, v. 23, p. 181–201, 1999.

CASTRO, L. d. F. *Introdução à Minerção de Dados: Conceitos Básicos, Algoritmos e Aplicações.* Elsevier: Saraiva, 2016.

COSTA, J. A. F. Classificação automática e análise de dados por redes neurais auto-organizávei. *Unicamp, SP*, 1999.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. *InterJournal*, Complex Systems, p. 1695, 2006. Disponível em: <http://igraph.org>.

DAVIES D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v, 2, p. p.224–227, 1979.

DUA, D.; GRAFF, C. *UCI Machine Learning Repository.* 2017. Disponível em: <http://archive.ics.uci.edu/ml>.

DUBES, R. C. *Algorithms for Clustering Data (Prentice Hall Advanced Reference Series : Computer Science)*. [S.l.]: Pearson College Div; First Edition edition, 1988.

ESTER, M. A density-based algorithm for discovering clusters in large spatial databases with noise. *Journal of Intelligent Learning Systems and Applications*, v. 10, p. 1, 1996.

FORINA M. ET AL, P. *UCI Machine Learning Repository*. 2003. Disponível em: <http://archive.ics.uci.edu/ml>.

GRANDO, F. On the analysis of centrality measures for complex and social networks. http://hdl.handle.net/10183/122516, 2015.

GRASSI R., S. S. T. A. Centrality in organizational networks. *International Journal of Intelligent Systems*, v. 25, p. 253–265, 2009.

HAYKIN. *Redes Neurais: princípios e prática. 2. ed.* [S.l.]: Bookman, 2001. 893 p.

HRUSCHKA E. R. EBECKEN, N. F. F. A genetic algorithm for cluster analysis. intelligent data analysis. *IEEE Transactions on Evolutionary Computation*, v. 1, p. 15–25, 2001.

HRUSCHKA E. R., E. N. F. F. A genetic algorithm for cluster analysis. *IEEE Transactions on Evolutionary Computation*, 2001.

JAIN. Data clustering: 50 years beyond K-Means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651–666, 2010.

JR. JOSEPH F.; ANDERSON, R. E. T. R. L. B. W. C. H. Multivariate data analysis. *Prentice Hall,*, v, 1, 1998.

KEIM D. A., M. F. S. J. e. Z. H. Challenges in visual data analysis. In Information Visualization. *Tenth International Conference on*, p. 9–16, 2006.

KITANI, E. C.; DEL-MORAL-HERNANDEZ, E.; SILVA, L. A. Somm–self-organized manifold mapping. Springer, p. 355–362, 2012.

KITANI, E. C.; HERNANDEZ, E. D. M.; SILVA, L. A. Learning embedded data structure with self-organizing maps. Springer, p. 225–234, 2013.

KOHONEN, T. Self-Organizing Maps. Third extended. Springer, Heidelberg, 2001.

KOHONEN, T. *Self-organizing maps.* [S.l.]: Springer Science & Business Media, 2012.

KOHONEN, T. Essentials of the self-organizing map. *Neural networks*, Elsevier, v. 37, p. 52–65, 2013.

MARIETTE, J. et al. Accelerating stochastic kernel som. In: M., V. (Ed.). *Proceedings of XXVth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017).* [S.l.]: i6doc, 2017. p. 269–274.

MONTEIRO, L. H. A. *Sistemas Dinâmicos Complexos.* [S.l.]: Editora Livraria da Física, 2014.

MURTAGH, F. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, v. 26, p. 354–359, 1983.

NETTO, P. O. B. *Grafos: Teoria, Modelos, Algoritmos.* [S.l.]: Editora Blucher, 2001.

NEWMAN, M. E. J. A measure of betweennes centrality based on random walks. *Social Networks*, v. 27, p. 39–54, 2005.

OLTEANU, M.; VILLA-VIALANEIX, N. On-line relational and multiple relational som. *Neurocomputing*, v. 147, p. 15–30, 2015.

OLTEANU, M.; VILLA-VIALANEIX, N.; COTTRELL, M. On-line relational som for dissimilarity data. In: P., E. et al. (Ed.). *Advances in Self-Organizing Maps (Proceedings of WSOM 2012, Santiago, Chili, 12-14 decembre 2012).* Berlin/Heidelberg: Springer Verlag, 2012. (Advances in Intelligent Systems and Computing series, v. 198), p. 13–22.

PEI, J. H. M. K. J. *Data Mining: Concepts and Techniques.* Waltham, MA, USA: Elsevier, 2011.

PETROVIć, S. A comparison between the silhouette index and the davies-bouldin index in labeling ids clusters. *Proceedings of the 11th Nordic Workshop on Secure IT-systems*, NORDSEC, v. 8, p. 53–64, 2006.

SAXENA, A. et al. A review of clustering techniques and developments. *Neurocomputing*, Elsevier, v. 267, p. 664–681, 2017.

SILVA, L. A.; COSTA, J. A. F. A graph partitioning approach to som clustering. In: *International Conference on Intelligent Data Engineering and Automated Learning*. [S.l.]: Springer, 2011. p. 152–159.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à Mineração de Dados: Com Aplicações em R.* Elsevier: Saraiva, 2016.

STROGATZ, S. H. Exploring complex networks. nature. *Nature Publishing Group*, v. 410, p. p.268–276, 2001.

ULTSCH, A. U *-matrix : a tool to visualize clusters in high dimensional data. In: . [S.l.: s.n.], 2004.

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. *IEEE Transactions on neural networks*, IEEE, v. 11, n. 3, p. 586–600, 2000.

VESANTO J.; ALHONIEMI, E. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, v, 11, p. p.586–600, 2000.

WARD, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, [S.l.]*, v, 58, p. p.236–244, 1963.

WU, R. L. Z. An optimal graph theoretic approach to data clustering: theory ans its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 15, n. 11, p. 1101–1113, 1993.