Pós-Graduação em Ciência da Computação

Framework para Detecção de Anomalias em Bases de Folha de Pagamento Baseado em Mapas Auto-Organizáveis

Por

Anderson de Souza Andrade

Dissertação de Mestrado Profissional

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, ABRIL/2013

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Anderson de Souza Andrade

"*Framework* para Detecção de Anomalias em Bases de Folha de Pagamento Baseado em Mapas Auto-Organizáveis"

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO.

ORIENTADOR: Prof. Dr. ADRIANO LORENA INÁCIO DE OLIVEIRA

RECIFE, ABRIL/2013

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

Andrade, Anderson de Souza
Framework para detecção de anomalias em bases de folha de pagamento baseado em mapas auto-organizáveis / Anderson de Souza Andrade. - Recife: O Autor, 2013.
x, 77 f.: il., fig., tab.

Orientador: Adriano Lorena Inácio de Oliveira.
Dissertação (mestrado) - Universidade Federal de Pernambuco. CIn, Ciência da Computação, 2013.

Inclui referências.

1. Ciência da Computação. 2. Inteligência Artificial. I. Oliveira, Adriano Lorena Inácio de (orientador). II. Título.

004 CDD (23. ed.) MEI2013 – 147

Dissertação de Mestrado Profissional apresentada por **Anderson de Souza Andrade** à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título**, "Framework para detecção de anomalias em bases de folha de pagamento baseado em mapas auto-organizáveis",** orientada pelo **Professor Adriano Lorena Inácio de Oliveira** e aprovada pela Banca Examinadora formada pelos professores:

_____

Prof. Francisco de Assis Tenório de Carvalho
Centro de Informática / UFPE

_____

Prof. Renato Fernandes Corrêa
Centro de Artes e Comunicação / UFPE

_____

Prof. Adriano Lorena Inácio de Oliveira
Centro de Informática / UFPE

Visto e permitida a impressão.
Recife, 12 de abril de 2013.

_____

Profª. EDNA NATIVIDADE DA SILVA BARROS
CoordenadorA da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Agradecimentos

Agradeço Àquele que está em todas as coisas, é a energia que nos anima e permeia nosso viver.

A minha mãe Sônia exemplo de força, serenidade, caráter e perseverança, que ensinou que desafios existem para serem superados com trabalho e dedicação.

A minha esposa Janine que com carinho me deu suporte e apoio nos momentos difíceis, compreendendo minha ausência em busca do objetivo maior. Seu companheirismo e amor fazem da minha caminhada nesse mundo algo de prazeroso e recompensador tornando-me cada dia uma pessoa melhor.

Aos meus irmãos André, Adriana e Aline, ao amor incondicional e à confiança que sempre depositaram em mim, nossa união me deixa próximo a vocês mesmo quando o tempo e a distância nos afastam.

Ao meu Grande Amigo Bergson, irmão que a vida me presenteou, sua presença, bom-humor e bondade nos momentos em que mais precisei me deram forças para seguir adiante com a batalha do dia-a-dia.

Aos amigos de jornada André, Hilson e Wanderson por todo apoio nesse caminhar.

A Confraria Carlos Pena Filho pelos momentos de lucidez etílica.

Ao Profº Adriano Lorena por me guiar nos momentos mais tortuosos desse trabalho e acreditar que era possível.

A CHESF, empresa que foi minha segunda casa por dez anos e aos amigos que lá fiz em especial a Marleide pelo apoio dado neste trabalho.

A todos que contribuíram direta ou indiretamente para este trabalho.

Resumo

O aumento na complexidade do ambiente de negócios e o acirramento da competição implicam a necessidade de informações para tomada de decisão em um espaço de tempo cada vez menor. Por outro lado, sistemas de informação mais abrangentes e complexos geram cada vez mais dados, tornando inviável a atividade de auditoria não assistida por métodos computacionais. As técnicas de inteligência artificial, particularmente aprendizagem de máquina, estão entre as mais apropriadas para lidar com esse tipo de problema. Dentre as técnicas de aprendizagem de máquina, as redes neurais artificiais vêm desempenhando um papel comprovadamente eficaz como ferramenta de apoio a atividade de auditoria. Diante desse cenário e alinhado ao estado da arte no uso da tecnologia da informação na atividade de auditoria, essa dissertação propõe a construção de um framework para detecção de anomalias em bases de dados baseado na rede neural artificial Mapas auto-organizáveis - Self-Organizing Maps (SOM). Utilizando as propriedades de mapeamento da Rede SOM, o framework consiste em: (i) demonstrar que dados visualmente distantes da área de influência da rede SOM são anomalias, e (ii) estabelecer um critério, baseado em intervalo de percentil, para classificação dos dados como possíveis anomalias independentemente da região do mapa SOM em que se encontrem. Ademais, este trabalho usa a análise de trajetória SOM na função de classificador de anomalia, a fim de comparar o limiar fixo baseado na vizinhança do neurônio com o limiar baseado em intervalo de percentil. O framework proposto foi aplicado em uma base de dados real de folha de pagamento. Os resultados apresentados na dissertação mostraram que o framework conseguiu obter bons resultados neste problema.

Palavras-chave: Mapas auto-organizáveis, Auditoria assistida por software, SOM, Mineração de Dados, Análise de Trajetória SOM, Folha de pagamento.

Abstract

The increasing of complexity and competition in the business environment implies on the needs of faster decision making information. In the other hand information systems more complex and wider create even more data making auditing without computer assistance impossible. Artificial intelligence techniques, specially machine learning, are one of the most appropriate to handle this specific problem. Among machine learning techniques, artificial neural networks have proving their effectiveness as an information technology supporting tool in auditing. Facing this and aligned with the state of art in the auditing use of information technology, this master thesis proposes a framework for anomaly detection in payroll databases based on the artificial neural network Self-Organizing Maps (SOM). Using the mapping properties of SOM, the framework consists of: (i) demonstrate that data visually far from the SOM network influence are anomalies and (ii) establish criteria based on confidence interval to classify data as possible anomalies apart of SOM's map region they were found. Moreover, this work uses trajectories SOM as anomalous data classifier to compare a fix threshold based on neuron neighborhood with a threshold based on confidence interval. The proposed framework is applied in a real payroll database. The outcomes presented shows that the framework achieved good results with this problem.

Keywords: Self-Organizing maps, Computer assisted auditing, Data Mining, SOM Trajectories, Payroll.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo

1

1. Introdução

Este capítulo contextualiza o estágio atual da atividade de auditoria diante do uso de novas tecnologias para suporte a essa atividade, discute a motivação para o desenvolvimento do trabalho proposto nesta dissertação, relata os objetivos gerais e específicos almejados e descreve como esta dissertação está organizada.

1.1. Contextualização

Ao longo dos anos, a complexidade no ambiente de negócios e consequentemente do ambiente de auditoria vêm crescendo. O acirramento da competição e a necessidade de informações para tomada de decisão em um espaço de tempo cada vez menor marcam o atual ambiente de negócios. Os sistemas computacionais estão mais abrangentes e complexos, gerando cada vez mais dados. Naturalmente estes dados disponíveis, muitas vezes, de forma exclusivamente eletrônica precisam ser auditados visando certificar que as saídas dos processos que geram esses dados estão corretas. Entretanto, diante da quantidade de dados a serem auditados tornou-se inviável a atividade de auditoria não assistida por métodos computacionais. Nesse sentido, as técnicas de inteligência artificial, particularmente aprendizagem de máquina, se caracterizam por fornecer um conjunto de métodos que podem suportar de forma comprovadamente eficiente à atividade de auditoria. As áreas de auditoria de erros materiais e de fraudes de gestão, que consistem em identificar divergências de valores nas contas auditadas por erro no processo ou por erro do gestor, são exemplos de utilização das técnicas de aprendizagem de máquina em auditoria[1].

A maturidade e evolução do uso da tecnologia de informação em atividades de auditoria, buscando ferramentas de suporte baseadas em sistemas computacionais é objeto de estudo em diversos trabalhos científicos. A tabela 1 apresenta os cinco estágios de utilização de tecnologia da informação em auditoria [1].

No primeiro estágio suítes de escritório são utilizadas. No segundo estágio alguns bancos de dados, e-mail e gráficos também são adaptados. No terceiro estágio bancos de dados internos e externos, software de auditoria e modelos corporativos são usados. No quarto estágio sistemas especialistas, sistemas de suporte a decisão e softwares de auditoria especiais para auditoria contínua são utilizados. Ferramentas baseadas em redes neurais artificiais enquadram-se neste estágio e também no próximo estágio. No quinto estágio, utiliza-se métodos avançados como sistemas baseados em redes neurais artificiais a fim de aumentar a confiabilidade nos serviços de auditoria [1].

Estágio	Softwares utilizados	Aplicações em Auditoria
I	Processador de texto, planilha eletrônica.	Documentação, relatório de auditoria, cálculos e análises financeiras.
II	Gráficos, bancos de dados, e-mail.	Planejamento da auditoria, comparação de informações financeiras, análise da empresa.
III	Softwares de auditoria, bancos de dados históricos e modelos corporativos.	Testes dos sistemas de informação, averiguação dos dados.
IV	Sistemas especialistas, sistemas para tomada de decisão, sistemas para auditoria contínua[2].	Análise de especialistas para encontrar pontos para auditoria.
V	Métodos avançados, sistemas baseados em redes neurais artificiais.	Confiabilidade dos serviços de auditoria.

Tabela 1 – Os cinco estágios de utilização de TI em Auditoria

Redes neurais artificiais vêm desempenhando um papel bastante eficaz como ferramenta de apoio a atividade de auditoria, existem pesquisas que mostram que basicamente todas as áreas de auditoria podem beneficiar-se do seu uso, desde a área de investigação de erros materiais à avaliação de riscos [3], [4]. Neste contexto muito já foi produzido com métodos de aprendizado supervisionado.

Num contexto intraempresarial defende-se que a atividade de auditoria deve ser realizada em todas as bases de dados que sejam relevantes para o resultado da empresa e com cobertura total das operações realizadas nessas bases de dados [2]. Entre essas bases de dados destaca-se, nesse trabalho, a base gerada pelo processamento da folha de pagamento. Como citado anteriormente atividades de auditoria como auditoria em erros materiais e fraudes de gestão são beneficiadas pelo uso de técnicas de inteligência artificial. Além delas, pesquisas com dados oficiais de domínio público demonstram que a confiabilidade dos serviços de auditoria pode ser melhorada com a aplicação de redes neurais artificiais[1]. Vale ressaltar a área de auditoria que explora a capacidade de uma empresa manter sua saúde financeira nominada de *going concern*, a utilização de redes neurais artificiais nesta área aumenta bastante a confiabilidade da opinião do auditor [5], [6].

1.2. Motivação

A maior parte dos trabalhos relacionados à aplicação de aprendizagem de máquina em auditoria financeira utiliza técnicas de aprendizagem supervisionada, com o objetivo de predizer o comportamento de uma determinada rubrica[1], com base no

[1] Artigo de orçamento[54].

treinamento da rede neural prever um valor esperado,ou classificar a informação, tal classificação vai depender da área de auditoria que está sendo explorada, pode ser um erro ou possível anomalia se for uma auditoria para detectar erros materiais, ou um ponto relevante para auditoria se estiver sendo tratado de procedimentos de revisão analítica da auditoria [1], [3]. Esta abordagem tem produzido bons resultados, estando seu desempenho relacionado: ao algoritmo de aprendizagem utilizado [7], a arquitetura da rede neural [1] e a formação do repositório de dados criados para investigação[8].

Entretanto, alguns problemas reais não permitem obter uma base com dados rotulados como é exigido pelas técnicas de aprendizagem supervisionadas. Nesses casos verifica-se a necessidade de utilizar métodos de aprendizagem não supervisionada de descoberta de conhecimento. Dentre os métodos de aprendizagem não supervisionada existentes, um método de *clustering* foi escolhido por prover uma boa capacidade de visualização dos resultados e bom desempenho para se trabalhar com repositórios n-dimensionais, a rede neural artificial mapas auto-organizáveis, *self-organizing maps* (SOM) [9]. Ademais já foi demonstrado que a rede neural SOM pode ser utilizada em problemas de auditoria [10] bem como alguns *frameworks* para detecção *outliers* e de anomalias, que são objetos cujas características desviam significantemente da maior parte dos dados da base [11], baseados na rede neural artificial SOM já foram propostos [12], [13], [14].

O *framework* proposto foi aplicado em uma empresa pública do setor elétrico, com aproximadamente 5.700 funcionários, que responde a legislações nacionais e internacionais e utiliza um sistema de recursos humanos com base de dados relacional. Foi selecionado para compor a base de dados consolidada um período histórico de 72 meses de Janeiro de 2005 à Dezembro de 2011. O processo de folha de pagamento é definido como critico pela Alta Gestão, sendo necessário garantir a corretude na percepção dos valores devidos aos seus funcionários bem como, que os recolhimentos de encargos previstos na legislação sejam realizados nas datas definidas por cada órgão governamental, sob pena de incorrência de multas caso isso não seja obedecido.

1.3. Objetivos

Esta dissertação tem como objetivo geral definir um *framework* para detecção de anomalias em um repositório de dados de folha de pagamento, através da rede neural artificial de aprendizado não supervisionado SOM, que utiliza a percepção do mundo real através da abstração de dados n-dimensionais para uma aproximação bidimensional do conjunto de dados investigados. Para este fim serão investigadas métricas de qualidade dos mapas SOM gerados e soluções propostas para visualização dos dados em um mapa SOM e para identificação de possíveis anomalias no conjunto de dados investigados.

Ao passo que o repositório dos dados investigados deve refletir a necessidade do algoritmo de aprendizagem utilizado, além de representar informações representativas quanto à variação de remuneração dos funcionários. Descrever os critérios para escolha dos atributos do repositório junto aos especialistas de negócio bem como os próprios atributos torna-se um dos objetivos desse trabalho.

A qualidade dos resultados produzidos pela rede neural artificial SOM está diretamente associada aos parâmetros de treinamento utilizados como o tamanho da rede e normalização dos dados, com os quais a rede é configurada. Portanto, determinar corretamente os parâmetros de configuração da rede é fundamental para um bom desempenho do *framework*.

1.4. Organização do Trabalho

Esta dissertação está estrutura em 5 capítulos, sendo que este capítulo discutiu as considerações iniciais e motivação do trabalho bem como os objetivos que essa dissertação pretende alcançar. Os demais capítulos são:

Capítulo 2: Fundamentação Teórica

Este capítulo tem como objetivo descrever a fundamentação teórica revisando os conceitos de descoberta de conhecimento em bancos de dados e redes neurais artificiais, com ênfase na rede neural artificial SOM: seu algoritmo, suas propriedades, métricas de desempenho e a abordagem de análise de trajetórias SOM. Tais conceitos são utilizados para a elaboração do trabalho detalhado neste

documento. Neste capítulo também é abordada aplicações de redes neurais artificiais na atividade de auditoria.

Capítulo 3: Framework de Rede SOM para detecção de anomalias em bases de dados de folha de pagamento

Este capítulo descreve o *framework* proposto pelo trabalho, como foi construída a base de dados, a seleção dos dados de treinamento e validação, as *toolboxes* utilizadas, os parâmetros de treinamento e as métricas de avaliação de desempenho da rede.

Capítulo 4: Aplicação do Framework em folha de pagamento

Este capítulo mostra a aplicação do *framework* proposto em um problema real e as principais conclusões acerca dos dados, seus reflexos no mapa SOM e como o mapa SOM se adapta ao espaço de entrada. Aqui é explorada uma característica do mapa SOM que é a análise das trajetórias de um determinado conjunto de dados do espaço de entrada no espaço de saída.

Capítulo 5: Conclusão

Este capítulo discute as principais contribuições da dissertação, e sugere trabalhos futuros que contribuirão para o avanço do estudo realizado.

Capítulo

2

2. Fundamentação Teórica

Este capítulo contém a fundamentação teórica utilizada no desenvolvimento do trabalho de dissertação descrito neste documento, iniciando com a definição dos principais conceitos de descoberta de conhecimento em banco de dados (Seção 2.1), passando por uma revisão em redes neurais artificiais com ênfase na rede neural artificial mapas auto-organizáveis, SOM (*Self-organizing maps*), (Seção 2.2) e finalizando com a descrição dos trabalhos correlatos (Seção 2.3), incluindo um relato sobre a aplicação de redes neurais artificiais na atividade de auditoria.

2.1. Descoberta de Conhecimento em Banco de Dados

O conceito Descoberta de Conhecimento em Banco de Dados KDD (*knowledge discovery in databases*), refere-se a todo o processo de identificação de padrões válidos, não triviais e potencialmente úteis, perceptíveis a partir dos dados [15], KDD possui três fases: pré-processamento, mineração de dados e pós-processamento. Nesta seção veremos as fases do processo de KDD com um aprofundamento na fase de mineração de dados.

O processo de KDD não pode ser resumido à fase de mineração de dados, esta última é a aplicação de algoritmos específicos para extração de padrões nos dados. É importante enfatizar que KDD se refere a todo o processo de descoberta de conhecimento incluindo: como os dados são armazenados e acessados, como algoritmos podem ser escalados e ainda executados eficientemente, como resultados podem ser interpretados e visualizados, e como a interação homem-máquina pode ser modelada e suportada de forma proveitosa [16].

A descoberta de conhecimento em banco de dados é uma área interdisciplinar que relaciona procedimentos científicos distintos, envolvendo técnicas estatísticas, aprendizagem de máquina, reconhecimento de padrões, visualização de dados, inteligência artificial e alto desempenho computacional. O objetivo da unificação é a extração de conhecimento de alto nível, de conjunto de dados de baixo nível, no contexto de grandes bancos de dados [15].

Fases do Processo de KDD

O processo de KDD é iterativo e interativo, possuindo um conjunto de atividades ilustradas na Figura 1. O processo é iterativo por possuir etapas sequenciais que podem ser revisitadas, pois as descobertas realizadas (ou a falta delas) podem levar a novas hipóteses. A natureza interativa do processo reside no fato de que, em todas as atividades, são tomadas decisões que dependem de conhecimento do domínio e da utilização que será feita do conhecimento descoberto.

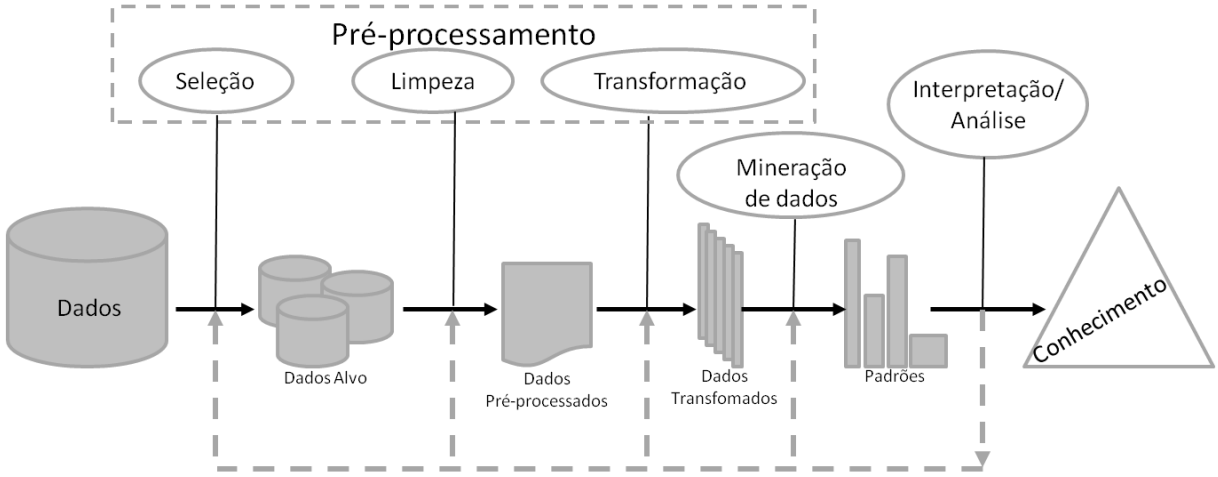


Figura 1 - Fluxo de tarefas no processo de KDD (adaptada de [16]**).**

As atividades de seleção, pré-processamento, transformação, mineração de dados e interpretação e análise dos dados, ilustradas na Figura 1, estão distribuídas nas três fases do processo de KDD: pré-processamento, mineração de dados e pós-processamento[16].

Pré-processamento

Os dados do mundo real em seu estado bruto podem estar em bases diversas e não consolidadas, os dados podem estar incompletos e com ruídos. Portanto, a fase de pré-processamento é crucial para uma maior acurácia e eficiência dos resultados da fase de mineração de dados. Apesar da importância dada à fase de mineração de dados é a fase de pré-processamento que demanda maior tempo, em torno de 80% do tempo utilizado em todo o processo de KDD[8].

A seguir serão descritas as atividades que fazem parte da fase de pré-processamento:

- Identificação do Problema – consiste em entender o domínio da aplicação e definir os principais objetivos a serem alcançados na aplicação do processo de KDD, bem como o que é conhecimento relevante do ponto de vista do usuário final.
- Seleção dos Dados – é a seleção do conjunto de dados a ser minerado propriamente dito, seja ele um subconjunto de variáveis ou uma amostra dos dados a serem minerados. Para realizar esta seleção a compreensão do domínio é indispensável, uma vez que os resultados da fase de mineração de dados serão obtidos a partir dos dados aqui selecionados.

- Limpeza dos Dados – como citado anteriormente, dados incompletos, inconsistentes e ruídos são características comuns em grandes conjuntos de dados do mundo real. O não tratamento dessas falhas nos dados pode ser um problema significativo na tarefa de descoberta de padrões, gerando dúvidas sobre os resultados. A limpeza dos dados é realizada através do pré-processamento dos dados. Isso se faz através da integração de dados heterogêneos, tratamento de dados nulos ou ausentes, eliminação de dados incompletos, repetição de registros, problemas de tipagem e tratamento de ruídos [17].
- Transformação dos Dados – nesta atividade, os dados são transformados ou consolidados em formas apropriadas ao problema. Raramente um projeto inicia-se com a hipótese já definida. Em muitos casos, a população inteira pode ser muito diversa para compreensão, mas detalhes dos subconjuntos da população que se comportem com o foco da análise, podem ser trabalhados [17].

Mineração de Dados

É a principal fase do processo de descoberta de conhecimento em banco de dados. Corresponde a extração de padrões dos dados e não se trata de apenas uma técnica, mas de um conjunto de técnicas heterogêneas e distintas entre si. A aplicação de cada técnica depende do problema, da mesma forma que duas ou mais técnicas podem ser utilizadas em conjunto de forma complementar.

Pós-processamento

Esta fase engloba as atividades de interpretação e análise dos dados minerados e padrões gerados e ações a serem tomadas no conhecimento descoberto.

- Interpretação e análise dos dados minerados – corresponde a analisar os padrões gerados, é possível que seja necessário o retorno a uma das atividades anteriores para mais uma iteração. Pode envolver a visualização dos modelos e padrões extraídos, bem como a visualização dos dados nos modelos extraídos.

- Ações sobre o conhecimento descoberto – consiste em usar o conhecimento diretamente, incorporá-lo a outro sistema ou simplesmente documentá-lo e reportar as partes interessadas [15].

Mineração de Dados

A etapa de mineração de dados desempenha um importante papel dentro do processo de KDD. Estima-se que 80% da informação relevante contida em um banco de dados pode ser descoberta utilizando alguma linguagem de consulta estruturada [8]. Entretanto, 20% dessa informação, que pode conter um conhecimento vital para a organização, permanece escondida e precisa de outros métodos para ser alcançada.

Uma vez que a base de dados esteja preparada, qualquer técnica que ajude a extrair algum conhecimento escondido dos dados pode ser considerada mineração de dados. Técnicas de mineração de dados formam, assim, um grupo bastante heterogêneo. Dentre as técnicas existentes podemos citar: ferramentas de consulta, técnicas estatísticas, técnicas de visualização, classificação, regras de associação, sumarização, análise de sequência, regressão e agrupamento [8], [16]. Nesta dissertação, será dada ênfase, devido à natureza do problema e dos dados encontrados, a técnica de agrupamento.

Em [8], [17], a fase de mineração de dados é definida ela própria como um processo e como todo processo de desenvolvimento de software o processo de mineração de dados também pode ser auxiliado com o uso de metodologias. “Uma metodologia de engenharia de software é um processo para a produção organizada de software, com utilização de uma coleção de técnicas predefinidas e convenções de notação. Uma metodologia costuma ser apresentada como uma série de etapas, com técnicas e notação associadas a cada etapa.”[17].

Abaixo é apresentada a metodologia CRISP-DM, *Cross-Industry Standard Process for Data Mining*, desenvolvida pela indústria com o objetivo de lançar um processo padrão para o desenvolvimento de projetos de mineração de dados. Não é objetivo deste trabalho realizar uma revisão detalhada desta metodologia, entretanto faz-se necessário, para melhor entendimento, apresentar suas fases e tarefas.

Cross-Industry Standard Process for Data Mining – CRISP-DM

A metodologia CRISP-DM foi lançada em 1996 por um consórcio de empresas, com o objetivo de ser um modelo de processo, bem como uma metodologia abrangente para projetos de mineração de dados documentada, não-proprietária e disponível gratuitamente.

O modelo de referência CRISP-DM divide o ciclo de vida de um projeto de mineração de dados em seis fases distintas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e entrega [18]. A Figura 2 - Modelo de referência CRISP-DM (adaptado de **[18]**). ilustra o fluxo do processo CRISP-DM.

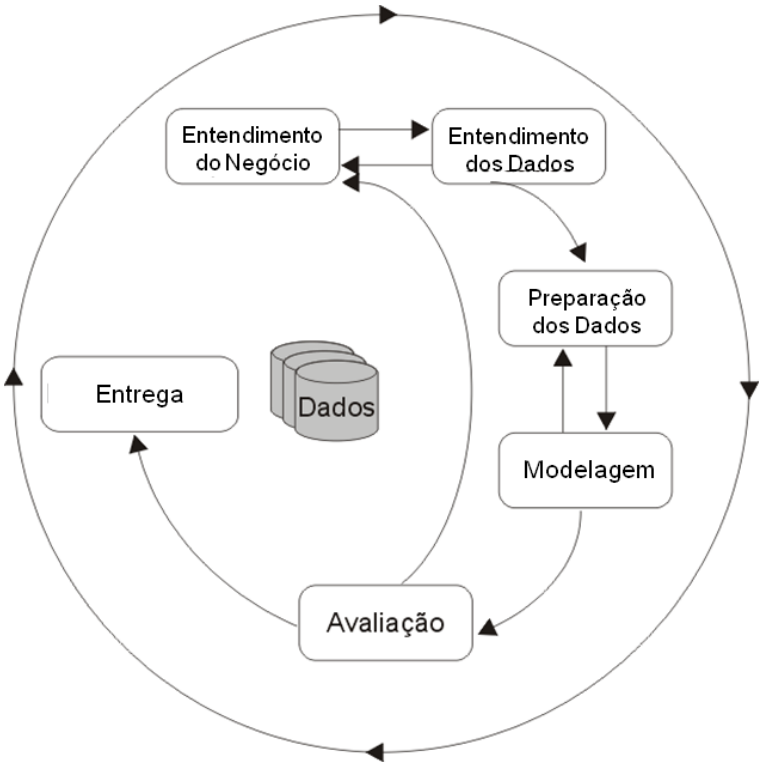


Figura 2 - Modelo de referência CRISP-DM (adaptado de [18]**).**

Uma breve descrição de cada fase e suas atividades é feita em seguida:

1. Entendimento do Negócio – é a fase mais importante em um projeto de mineração de dados, o entendimento inicial do negócio foca em entender os objetivos do projeto de uma perspectiva do negócio, convertendo este conhecimento em uma definição sucinta do problema a ser atacado do ponto de vista da mineração de dados e então desenvolver um planejamento preliminar para alcançar os objetivos desejados. Visando entender quais

dados serão analisados e como, é vital conhecer completamente o negócio para o qual está sendo modelada a solução.

Atividades da fase:

- Determinar os objetivos do negócio
- Avaliar a situação
- Determinar as metas da mineração de dados
- Produzir um plano de projeto

2. Entendimento dos Dados – inicia-se com a seleção dos dados. Nesta fase é preciso que a formação da base de dados seja dominada, para que sejam identificados tanto problemas de qualidade nos dados quanto subconjuntos interessantes para formação de hipóteses sobre informações escondidas.

   Atividades da fase:

   - Coletar dados iniciais
   - Descrever os dados
   - Explorar os dados
   - Verificar a qualidade dos dados

3. Preparação dos Dados – cobre todas as atividades para construir um conjunto de dados que alimentará a(s) ferramenta(s) de modelagem. Atividades tais como a seleção de tabelas, registros e atributos bem como a transformação e limpeza dos dados estão incluídas nessa fase.

   Atividades da fase:

   - Seleção dos dados
   - Limpeza dos dados
   - Reconstrução dos dados
   - Integração dos dados
   - Formatação dos dados

4. Modelagem – nessa fase é escolhida e aplicada a técnica, ou técnicas, de mineração de dados adequada a resolução do problema definido na fase de entendimento do negócio e geração de modelos úteis ao processo de KDD. Uma vez que técnicas diferentes podem requerer conjuntos de dados com formatações diferentes a iteração com a fase de preparação de dados pode ocorrer.

   Atividades da fase:

   - Selecionar a técnica de mineração de dados
   - Construir o modelo

- Testar o modelo
- Evoluir o modelo

5. Avaliação – uma vez pronto, o modelo e precisa ser avaliado e sua construção revisada para certificar que os requisitos do projeto foram todos atendidos e os objetivos do negócio alcançados. Nesta fase o projetista define onde aplicar resultados obtidos, se o modelo será entregue ou passará por novas iterações.

   Atividades da fase:
   - Avaliar os resultados
   - Revisar o processo
   - Definir aplicação dos resultados

6. Entrega – os resultados obtidos, o conhecimento descoberto, devem ser organizados e apresentados de forma que o usuário possa usá-lo. A forma de apresentação difere de projeto para projeto, podendo variar de um relatório final a personalização das páginas web da organização.

   Atividades da fase:
   - Planejamento da entrega
   - Planejamento do monitoramento e manutenção
   - Produzir o relatório final
   - Revisar o projeto

O processo de KDD pode assumir o papel principal em diversas áreas de negócio, entretanto para que o mesmo possa agregar valor é de suma importância que a fase de mineração de dados seja bem planejada e que a técnica correta, ou um conjunto delas, seja utilizada.

2.2. Mapas Auto-Organizáveis – *Self Organizing Maps* (SOM)

Após elencar algumas técnicas de mineração dados, vamos detalhar a rede neural artificial Mapas Auto-Organizáveis (SOM), iniciando com os conceitos de redes neurais e suas aplicações.

Redes Neurais Artificiais

As redes neurais artificiais constituem um conjunto de técnicas de aprendizagem de máquina que modelam a forma com que o cérebro executa uma tarefa em particular. Isto é realizado pela interconexão de unidades de processamento permitindo que a rede neural aprenda com os exemplos que lhe são apresentados[19]. Tais unidades de processamento, ou neurônios, são as unidades fundamentais da rede neural. Cada neurônio recebe e processa valores de entrada externos à rede e emite um valor de saída [20], este valor é calculado por uma função de transferência e será transmitido para outro neurônio, que pode estar situado em uma camada mais interna, ou para a saída da rede neural[1].

Enquanto o neurônio processa o valor de entrada recebido, através da função de transferência e dos pesos, o processo de aprendizado também é iniciado. O neurônio executa uma operação de atualização em uma memória local, ajustando os pesos sinápticos de maneira ordenada, adaptando a rede neural ao seu ambiente. Esta é a abordagem tradicional de um algoritmo de aprendizagem.

A estrutura de processamento paralelo e distribuído e a capacidade de aprendizado são as características mais importantes nas redes neurais artificiais. As redes neurais artificiais aprendem com os exemplos que são apresentados durante sua fase de treinamento, a cada exemplo apresentado os pesos da rede podem ser alterados, procurando assim, um conjunto de pesos que permita à rede executar determinada tarefa. Da capacidade de aprendizado deriva-se a generalização. Generalização é a capacidade de uma rede neural produzir saídas plausíveis para entradas não apresentadas durante o treinamento[19].

Entretanto, o uso dessas capacidades está associado à escolha do paradigma de aprendizagem correto, pois problemas diferentes, possivelmente, requerem abordagens diferentes para sua resolução. Cada paradigma de aprendizagem possui algoritmos de aprendizagem que satisfazem as condições do paradigma. A Figura 3 apresenta uma taxonomia dos processos de aprendizagem. Os principais paradigmas de aprendizagem são: aprendizado supervisionado e aprendizado não-supervisionado.

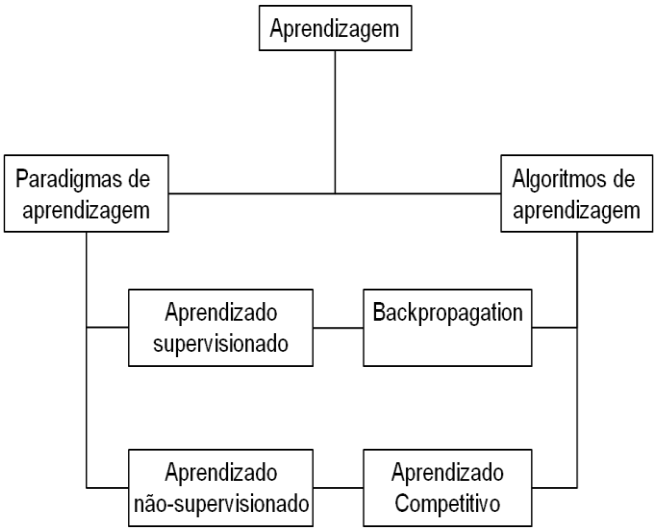


Figura 3 - Taxonomia dos processos de aprendizagem (adaptado de [1]**).**

2.2.1.1 Paradigmas de Aprendizagem

Como foi dito anteriormente, cada paradigma de aprendizagem possui características próprias. A escolha de uma rede neural que se adeque a fase de mineração de dados dependerá do problema que está sendo resolvido, bem como da modelagem da base de dados criada nas fases iniciais do processo de KDD.

Abaixo são descritos os principais paradigmas de aprendizagem encontrados na literatura sobre redes neurais artificiais.

Aprendizado Supervisionado

No aprendizado supervisionado deve existir um conhecimento a priori do ambiente para o qual a rede neural está sendo modelada. Este conhecimento é representado através de um conjunto de exemplos entrada-saída, ou vetores de treinamento. Conceitualmente, o treinamento se dá seguinte forma: o vetor de treinamento é apresentado à rede, que computa a resposta; a reposta computada pela rede é então comparada à resposta ideal previamente conhecida; os parâmetros da rede são ajustados combinando a influência do vetor de treinamento e do erro. O erro é a diferença entre a resposta da rede e a resposta ideal conhecida a priori. Este ajuste segue então de forma iterativa com o objetivo de transferir o conhecimento a priori, fornecido pelo ambiente, para a rede neural através do treinamento. Quando esta transferência for o mais completa possível, a rede neural estará pronta para ser

apresentada ao ambiente sem supervisão. A Figura 4 ilustra esta forma de aprendizado.

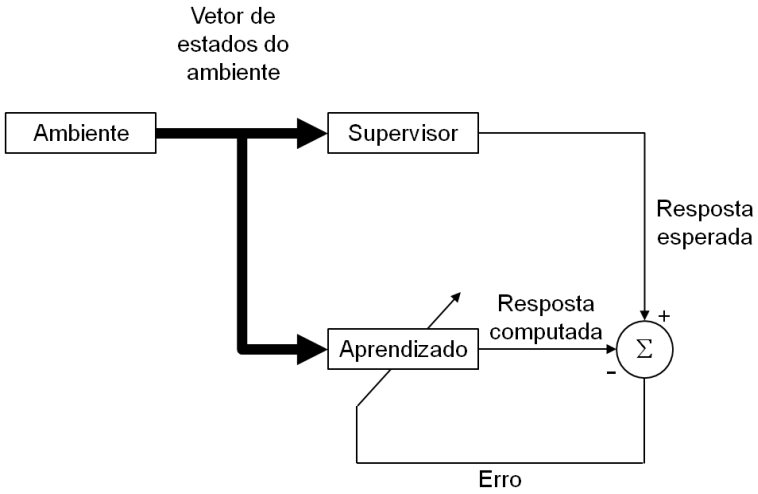


Figura 4 - Aprendizado Supervisionado (adaptado de [19]**).**

Aprendizado Não-Supervisionado

No aprendizado não-supervisionado, não há nenhum conhecimento a priori do ambiente para o qual a rede neural está sendo modelada para auxiliar o processo de aprendizagem. A rede neural é apresentada aos dados de entrada e cria uma representação desses estados como saída da rede. O ajuste da rede neural leva em conta alguma medida de qualidade da representação criada e os parâmetros da rede devem ser otimizados para que a representação fique mais próxima do desejado. Uma vez ajustada, a rede desenvolve a capacidade de formar representações dos dados de entrada, bem como de criar novas classes automaticamente. A Figura 5 ilustra o processo de aprendizado não-supervisionado.

Figura 5 - Aprendizado Não-Supervisionado (adaptado de [19]**).**

Apesar da aparente simplicidade do processo ilustrado, o fato de não haver conhecimento a priori do ambiente ao qual a rede neural está sendo modelada, insere um grau de dificuldade maior na interpretação dos dados de saída, ficando o

usuário da rede, responsável pela interpretação das representações geradas no processo de treinamento.

2.2.1.2 Algoritmos de Aprendizagem

Um algoritmo de aprendizagem é um conjunto de regras bem definidas para a solução de um paradigma de aprendizado [19]. Diversos algoritmos podem adequar-se a um determinado paradigma aqui citado, a diferença entre eles está nos parâmetros de ajustes. Para exemplificar os algoritmos existentes, cita-se aqui dois algoritmos um para cada tipo de paradigma elencado anteriormente.

Backpropagation

Utilizado para resolver o paradigma de aprendizado supervisionado. Para um dado vetor de entrada ele gera um vetor de saída, então a diferença entre o vetor de saída gerado pela rede e o vetor esperado pelo supervisor, o erro, é transmitido de volta através da rede neural para modificar os pesos da mesma.

Aprendizado Competitivo

Utilizado para resolver o paradigma de aprendizado não-supervisionado, neste algoritmo todos os neurônios de saída competem entre si, de modo a formarem uma camada competitiva. Um dos neurônios é escolhido o vencedor, seus pesos são ajustados para que ele fique mais próximo do vetor de entrada que ativou o neurônio vencedor.

Mapas Auto-Organizáveis (SOM)

A utilização de mapas constitui um poderoso mecanismo de representação da informação. Eles projetam um padrão, ou característica particular, do espaço de entrada em uma posição específica, codificando assim, a localização deste padrão no espaço de saída[21].

A SOM – *Self-Organizing Map (mapa auto-organizável)*– é uma arquitetura de rede neural artificial, baseada em um processo de aprendizado não-supervisionado, que realiza um mapeamento de um espaço de entrada de dados multivariados, n-dimensionais, para um espaço de saída, geralmente, bidimensional [22]. Para tanto,

SOM utiliza um algoritmo de aprendizagem competitiva, que procura preservar a topologia do espaço de entrada[23].

Portanto, a rede SOM caracteriza-se pela formação de um mapa topográfico ordenado no qual as localizações espaciais, ou coordenadas, dos neurônios da grade são indicativas das características estatísticas contidas nos padrões de entrada[9].

Esta característica demonstra duas propriedades desejáveis na rede SOM:

- Quantização vetorial, onde dados semelhantes ou que estejam próximos no espaço de entrada serão mapeados para neurônios vizinhos ou regiões adjacentes nos neurônios de saída[24], permitindo que similaridades e dissimilaridades nos vetores de entrada sejam representadas nos neurônios do espaço de saída[9].
- Projeção vetorial, ao fim do processo cada vetor do espaço de entrada estará mapeado em um neurônio no espaço de saída[24].

A

Figura 6 ilustra o mapeamento realizado pelo algoritmo SOM.

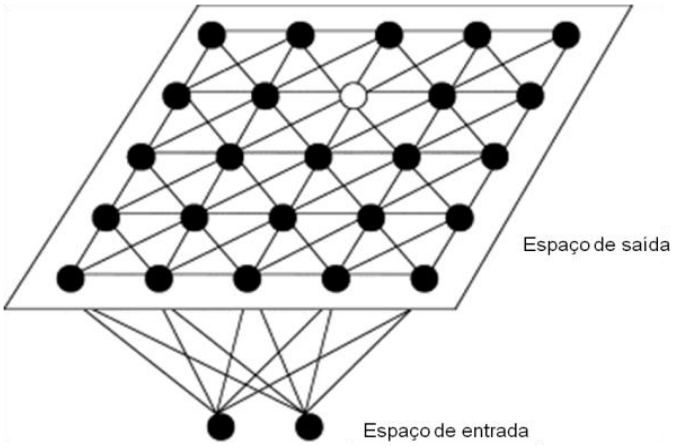


Figura 6 - Mapeamento SOM (adaptado de [25]**).**

2.2.1.3 Algoritmo SOM

Em sua definição original, o algoritmo para treinamento de redes SOM realiza uma transformação de um espaço de entrada n-dimensional para um espaço de saída bidimensional, estabelecendo um relacionamento não-linear entre eles. Para tanto, o

algoritmo responsável pela formação do mapa executa três processos, mas antes é necessário inicializar os pesos sinápticos da rede de forma randômica. Esses processos são: Competição, Cooperação e Adaptação Sináptica. Abaixo os processos são detalhados.

O Processo Competitivo

O processo competitivo ao qual a rede SOM é submetida visa encontrar o melhor neurônio (*BMU – Best Matching Unit*), ou o neurônio vencedor, para cada vetor do espaço de entrada. O melhor neurônio de um determinado vetor de entrada é o neurônio que preserva a menor distância euclidiana entre o vetor de entrada e o vetor de pesos sinápticos. Pode-se demonstrar esse enunciado da seguinte maneira:

Considerando n a dimensão do espaço de entrada, um vetor de entrada pode ser denotado por:

$$\mathrm{x} = [x_1, x_2, \dots, x_m]$$

Equação 1 - Vetor de entrada

O vetor de pesos sinápticos de cada vetor de entrada possui a mesma dimensão do espaço de entrada. O vetor de pesos sinápticos pode ser representado por:

$$\mathrm{p}_k = [p_{k1}, p_{k2}, \dots, p_{km}], \qquad k = 1, 2, \dots, m$$

Equação 2 - Vetor de Pesos Sinápticos

Onde *m* é o número de neurônios da rede, logo calculando a mínima distância entre o vetor entrada e o vetor de pesos sinápticos tem-se:

$$i(\mathrm{x}) = \arg\min_k ||\mathrm{x} - \mathrm{p}_k||, \qquad k = 1, 2, \dots, m$$

Equação 3 - Neurônio Vencedor

O índice i(x) pode ser usado para denotar o neurônio vencedor do processo de competição. O processo de competição mapeia um espaço de entrada contínuo em um espaço de saída discreto[19].

O Processo Cooperativo

Após encontrar o melhor neurônio, a rede passa a aprender de forma cooperativa, com o vetor de entrada mapeado para o neurônio vencedor. A cooperação se dá através de interações laterais entre o neurônio vencedor e seus vizinhos. A

intensidade do estímulo do melhor neurônio em sua vizinhança topológica decai em razão da distância entre o melhor neurônio e o neurônio estimulado e do tempo.

Para definir a região de abrangência, vizinhança topológica, do estímulo do melhor neurônio dentro do mapa, é necessário definir uma função de vizinhança. Do ponto de vista da distância, a função Gaussiana resolve as necessidades das interações laterais. Entretanto, faz-se necessário uma alteração na função de vizinhança para que a mesma possa atender ao requisito de diminuição do estímulo que o melhor neurônio produz em sua vizinhança topológica em função do tempo na convergência de aprendizado do algoritmo SOM. Abaixo, se apresenta as funções:

$$h_{k,i(x)} = \exp\left(-\frac{d_{i,k}^2}{2\sigma^2}\right)$$

Equação 4 - Função Gaussiana de Vizinhança Topológica

$$h_{k,i(x)}(t) = \exp\left(-\frac{d_{i,k}^2}{2\sigma^2(t)}\right), \qquad t = 1, 2, \ldots,$$

Equação 5 - Função de Vizinhança algoritmo SOM

O processo cooperativo é finalizado quando o número *t* é atingido, por exemplo, o número de interações desejadas.

O Processo Adaptativo

O processo adaptativo pode ser dividido em duas fases: Ordenação e Convergência. A ordenação topológica do mapa SOM é alcançada por meio da adaptação sináptica à qual a rede é submetida. Esse processo consiste em adaptar o vetor de pesos sinápticos de um neurônio ao seu vetor de entrada correspondente. Entretanto, o incremento simples do vetor de pesos sinápticos levaria a saturação interferindo no aprendizado.

A atualização do vetor de pesos sinápticos deverá ser, assim como a função de vizinhança, dependente do tempo. De tal modo que o processo de adaptação sináptica pode ser representado pela função abaixo:

$$p_k(t+1) = p_k(t) + \eta(t)h_{k,i(x)}(t)(x - p_k(t))$$

Equação 6 – Função de adaptação sináptica

Onde, $p_k(t)$ é o vetor de pesos sinápticos do neurônio *k* no tempo *t*, $\eta(t)$ é a taxa de aprendizado em função do tempo *t* e $h_{k,i(x)}(t)$ é a função de vizinhança utilizada no processo cooperativo.

A convergência do mapa SOM decorre do fato de que a função de adaptação sináptica atualiza o vetor de pesos sinápticos *p*, do neurônio vencedor *i* na direção do vetor de entrada *x* em função do tempo. Através de apresentações sucessivas aos dados de treinamento, os vetores de pesos sinápticos tendem a seguir a distribuição dos vetores de entrada ordenando o mapa de acordo com o espaço de entrada fornecido. Assim, neurônios vizinhos tendem a possuir vetores de pesos sinápticos similares.

Versão em Lote do Algoritmo SOM

O processo incremental definido pelo processo adaptativo mostrado acima pode ser substituído pela versão em lote do algoritmo SOM, a qual é significativamente mais rápida [26] e não sofre de problemas de convergência por não requerer a especificação da taxa de aprendizado $\eta(t)$ [27].

A versao em lote do algoritmo SOM também é incremental, mas no lugar de apresentar ao mapa um vetor de entrada a cada iteração, todo o conjunto de dados do espaço de entrada é apresentado ao mapa antes de quaisquer ajustes serem realizados. Em cada etapa do treinamento, os dados são particionados de acordo com as regiões de Voronoi e cada vetor de entrada pertence ao conjunto de dados cujo neurônio está mais próximo. Os vetores de pesos sinápticos dos neurônios são calculados pela Equação 7:

$$p_k = \frac{\sum_t h_{k,i(x)} x(t)}{\sum_t h_{k,i(x)}}$$

Equação 7 – Função do vetor de pesos sinápticos do algoritmo de treinamento em lote

Onde $i(x)$ é o melhor neurônio do vetor de entrada *x* [28].

2.2.1.4 Propriedades do mapa SOM

O mapa SOM pode ser definido como uma transformação não-linear de um espaço de entrada contínuo, representado pelos dados, em um espaço de saída discreto, representado pelos neurônios. Dado um vetor de entrada *x*, o algoritmo SOM

encontra um neurônio vencedor *i(x)* no espaço de saída. O vetor de pesos sinápticos p_i do neurônio *i(x)* pode ser visto como um ponteiro deste neurônio para o espaço de entrada. De fato, o vetor de pesos sinápticos p_i pode ser visto como as coordenadas do neurônio *i(x)* projetadas no espaço de entrada. A Figura 7 ilustra o relacionamento entre o mapa SOM e o vetor p_i.

O mapa SOM possui propriedades importantes que precisam ser citadas:

1. Aproximação do Espaço de Entrada

   O algoritmo SOM tem a capacidade de reduzir a dimensionalidade e comprimir dados do espaço de entrada, baseado na teoria de quantização vetorial. O objetivo dessa teoria é manter uma boa aproximação entre o espaço de saída e o espaço de entrada, utilizando como ponteiro entre eles o vetor de pesos sinápticos dos vetores de entrada.
2. Ordenação Topológica

   Consequência direta do processo adaptativo do algoritmo SOM, o neurônio vencedor é influenciado pelos dados do seu vetor de entrada correspondente, ao mesmo tempo em que influencia os neurônios vizinhos, moldando o mapa de acordo com o estimulo recebido dos padrões de entrada.
3. Sensibilidade a Densidade

   O mapeamento no espaço de saída é sensível ao número de ocorrências no espaço de entrada. Regiões com maior probabilidade de ocorrência no espaço de entrada serão representadas por uma região maior no espaço de saída.
4. Seleção de características

   Considerando uma distribuição não-linear como espaço de entrada, o mapa SOM é capaz de selecionar um conjunto com as melhores características que se aproximam desta distribuição[19].

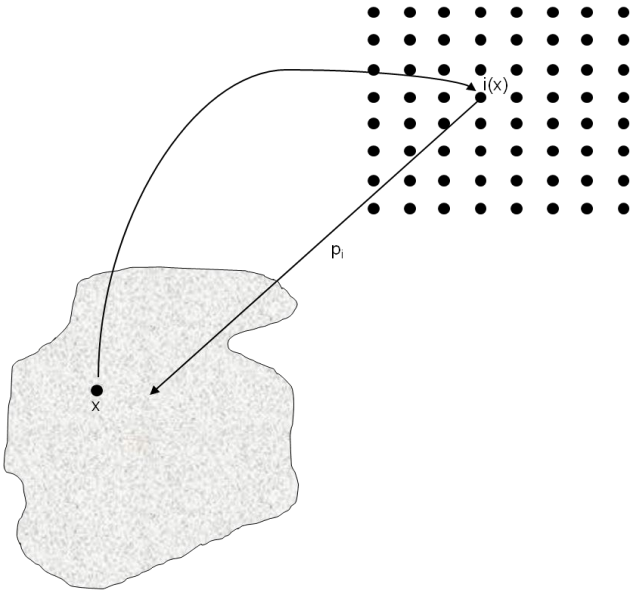


Figura 7 – Relacionamento entre o mapa SOM e o vetor de pesos sinápticos do neurônio vencedor (adaptado de [19]**).**

2.2.1.5 Métricas de qualidade

A qualidade do mapeamento realizado por um algoritmo SOM pode ser avaliada usando os seguintes critérios: grau de continuidade do mapeamento é quando vetores que estão próximos no espaço de entrada são mapeados próximos no espaço de saída; resolução do mapeamento, vetores que estão distantes no espaço de entrada não serão vizinhos no espaço de saída; e como o mapeamento está refletindo a distribuição de probabilidade do espaço de entrada[23].

Para medir esses critérios, bem como a acurácia em preservar a topologia do espaço de entrada projetado pelo mapeamento, diversas métricas de qualidade foram propostas[29]. Porém, para fins dessa dissertação serão discutidas três métricas: Erro de quantização (*quantization error – QE*), Erro topográfico (*topographic error – TE*) e o Produto topográfico (*topographic product – TP*).

Erro de Quantização

Esta métrica está relacionada a todos os algoritmos que realizam quantização vetorial e agrupamento, desconsidera a topologia do mapa em seu cálculo. O erro de quantização (QE) é medido calculando a média das distâncias dos vetores de entrada aos seus melhores neurônios no espaço de saída[30].

O erro de quantização decai em razão do incremento de neurônios no mapa, isso acontece porque os vetores de entrada ficam distribuídos de forma mais esparsa no mapa[31].

Erro Topográfico

É uma métrica de preservação de topologia que leva em consideração tanto o conjunto de dados que está sendo utilizado quanto a topologia do mapa que está sendo gerado pelo algoritmo SOM. O erro topográfico (TE) é calculado da seguinte maneira: os melhores e os segundo melhores neurônios dos vetores de entrada são encontrados, para cada ocorrência onde o melhor neurônio e o segundo melhor neurônio não sejam adjacentes no mapa, é considerado um erro. O erro total encontrado é normalizado para ficar em um intervalo entre 0 e 1, no qual 0 significa uma preservação total da topologia.

Existe um *trade-off* entre as métricas erro de quantização e erro topográfico, quanto menor o erro de quantização maior o erro topográfico, isso acontece porque, para se obter um erro de quantização menor basta apenas aumentar o número de neurônios no mapa, entretanto quanto maior o mapa, maior a probabilidade que o melhor neurônio e o segundo melhor neurônio não sejam adjacentes, aumentando o erro topográfico.

Produto Topográfico

O produto topográfico (P) é uma métrica de preservação das relações de vizinhança em mapeamentos entre espaços de diferentes dimensionalidades. Adaptado para o algoritmo SOM por [21]. O produto topográfico indica se o tamanho do mapa está apropriado ao conjunto de dados de entrada e para seu cálculo é considerado apenas os vetores de pesos sinápticos de cada neurônio do espaço de saída [32].

O algoritmo do produto topográfico realiza uma comparação entre as ordens de classificação dos k-vizinhos mais próximos nos espaços de entrada e saída, investigando [33]. Baseado em três variáveis P1, P2 e P3, onde P1 mede a distorção no espaço de entrada e é dada pela Equação 8, P2 a distorção no espaço de saída e é dada pela Equação 9.

$$P_1(j,k) = \frac{dist(m_j, m_{n_k^A(j)})}{dist(m_j, m_{n_k^V(j)})}$$

Equação 8 – Distorção no espaço de entrada

$$P_2(j,k) = \frac{dist(r_j, r_{n_k^A(j)})}{dist(r_j, r_{n_k^V(j)})}$$

Equação 9 – Distorção no espaço de saída

Onde:

j refere-se ao neurônio;

m_j refere-se ao vetor referência do neurônio;

r_j refere-se ao vetor posição do neurônio;

$n_k^V(j)$ refere-se ao k-vizinho mais próximo à j no espaço de entrada V;

$n_k^A(j)$ refere-se ao k-vizinho mais próximo à j no mapa SOM A;

dist é uma medida de distância qualquer como a distância Euclidiana.

A variável P3, dada pela Equação 10, combina os resultados das Equações Equação 8 e Equação 9 calculando a média geométrica de P1 e P2, obtendo o relacionamento topológico entre o neurônio j e seus k-vizinhos mais próximos. Se as ordens de classificação dos k-vizinhos mais próximos forem iguais então P1 e P2 são iguais a 1, caso contrário, P1 é maior que 1 e P2 está entre 0 e 1.

$$P_3(j,k) = \left(\prod_{l=1}^{k} P_1(j,l) \times P_2(j,l) \right)^{\frac{1}{2k}}$$

Equação 10 – Produtório das variáveis P1 e P2

Apesar de P1, P2 e P3 serem calculados para cada neurônio, o produto topográfico de todo o mapa pode ser estendido para toda a rede SOM e para todas as possíveis ordens de vizinhança pela Equação 11 que define o Produto Topográfico (P) [32], [34]. Este valor pode ser interpretado da seguinte maneira: se P < 0, então o mapa está pequeno para o conjunto de dados; se P > 0, então o mapa está grande para o conjunto de dados utilizados. Valores do produto topográfico próximos a zero são considerados bons [35].

$$P = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{k=1}^{N-1} log(P_3(j,k))$$

Equação 11 – Produto Topográfico de uma rede SOM

2.2.1.6 Análise de Trajetórias SOM

Na análise de trajetórias SOM a cadeia de melhores neurônios dos vetores de entrada sequenciados no tempo representa a trajetória desses vetores no mapa da rede SOM [36]. Esta técnica surgiu derivada de uma proposta de modificação na teoria de mapas auto-organizáveis para incluir a esta um sentido de temporalidade [37]. Dados temporais podem estar dispostos como uma sequência ordenada de eventos ou transações, mesmo não possuindo um indicador de tempo, infere-se, nesses casos, que um evento só ocorre após o término do evento anterior, ou os dados podem conter a informação temporal de forma explícita como um rótulo ou atributo do vetor de entrada, nesses casos é possível encontrar a distância temporal exata entre cada vetor de entrada, um exemplo são as séries temporais [38].

Como técnica de mineração de dados a análise de trajetórias SOM tem-se mostrado bastante versátil em suas aplicações como reconhecimento de sinais [36], predição de ações em bolsas de valores, detecção e predição de falhas [39], monitoramento e controle de máquinas e na atividade de auditoria [10]. A relação entre esses campos de aplicação está na natureza dos dados tempos-dependentes. Apesar das áreas de aplicação diversas a análise de trajetórias SOM possui, basicamente, dois papeis como técnica de mineração de dados: predição e visualização.

Quando introduzida como ferramenta de visualização a análise de trajetórias SOM fornece uma visão dinâmica do comportamento temporal dos dados analisados, permitindo uma interação imediata entre o especialista e o mapeamento dos vetores de entrada e seus melhores neurônios na rede. Neste sentido, mesmo que os *clusters* do mapa SOM não sejam conhecidos é possível visualizar com clareza qualquer variação nos neurônios do mapa. Associada a outras características do mapa SOM como o mapeamento bidirecional entre o espaço de entrada e os neurônios-saída, a análise de trajetórias SOM permite identificar a posição de cada vetor de entrada no mapa em cada ponto do tempo. A Figura 8 ilustra exemplos de trajetórias no mapa SOM.

Figura 8 – Exemplos de visualização de trajetórias de dois indivíduos em função do tempo. (a) indivíduo que tem seus melhores neurônios modificados. (b) indivíduo sem alteração na trajetória (adaptado de [10]**).**

Em problemas de predição mais geral ou detecção de falhas o mapa SOM deve estar dividido em *clusters* bem conhecidos, existem vários mecanismos para realizar essa divisão não sendo objetivo dessa seção detalhá-los. Esses *clusters* precisam ser avaliados por especialistas do negócio. Fazendo uma analogia é como se as regiões do mapa SOM estivessem sendo rotuladas. Assim, mesmo que os dados de treinamento sejam não-rotulados e o método de aprendizagem seja não-supervisionado o mapa SOM precisa ter suas regiões ou neurônios rotulados. A partir da divisão do mapa SOM é possível visualizar o comportamento da trajetória e através dos exemplos apresentados a rede consegue predizer o fim da trajetória, pois o processo de treinamento da rede SOM delimita as fronteiras das regiões distintas do mapa [40]. A Figura 9 ilustra exemplos de trajetórias SOM utilizadas para predição e detecção de falhas.

A Figura 9 é interpretada da seguinte maneira. No Mapa SOM é exibida a trajetória dos melhores neurônios e nos Rótulos do mapa SOM é mostrada a mudança da situação. Pode-se perceber que o melhor neurônio inicia sua trajetória em um neurônio rotulado como funcionamento normal passando por fases de deterioração e termina em um neurônio que indica falha no funcionamento do dispositivo. A análise de trajetórias SOM consegue detectar a falha do dispositivo, de forma análoga se esta análise fosse realizada momentos antes da falha ocorrer seria possível predizer sua ocorrência devido às áreas do mapa que o melhor neurônio estava percorrendo em sua trajetória.

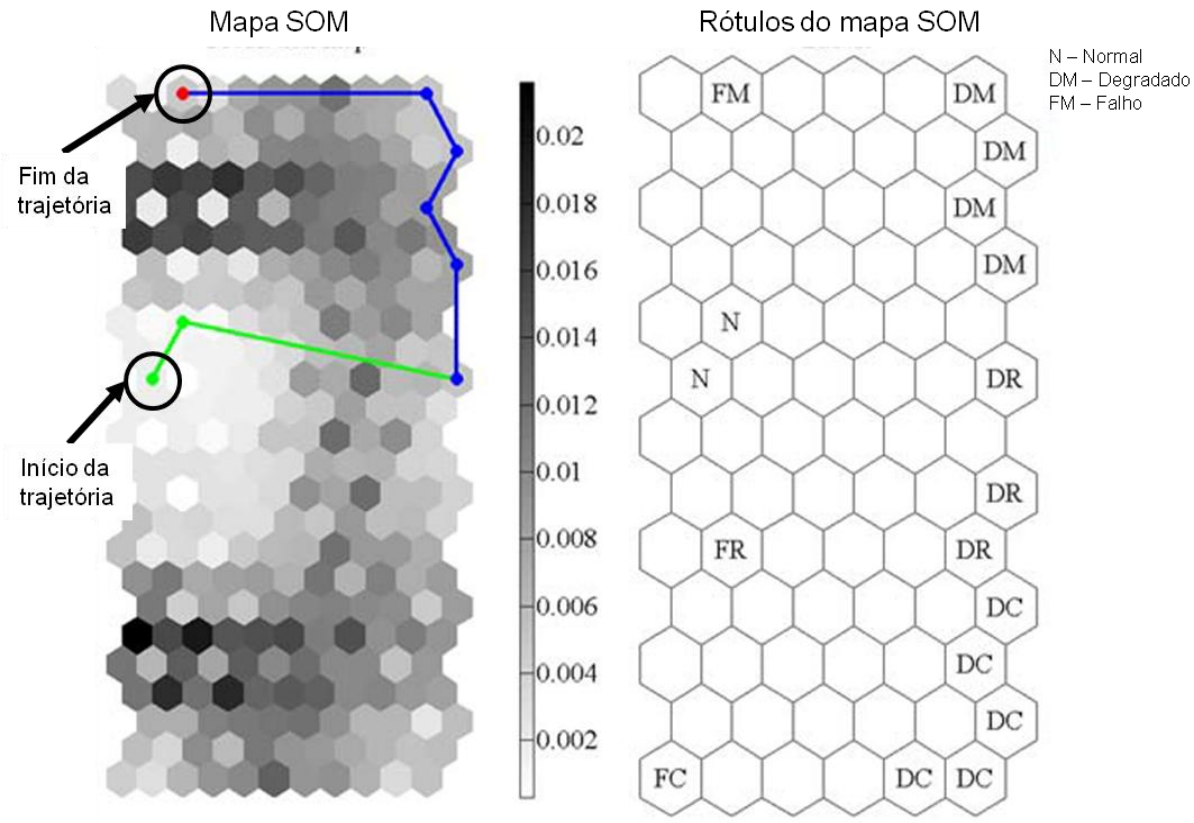


Figura 9 – Exemplo da aplicação da análise de trajetórias SOM em predição de falhas (adaptado de [39]**).**

A análise de trajetórias SOM se mostra uma poderosa ferramenta de análise individual dentro do espaço de entrada total. É ainda mais interessante quando o mapa SOM possui grupos bem conhecidos, sendo possível não apenas acompanhar a atividade do indivíduo no mapa SOM, mas também predizer quando o mesmo mudará de grupo.

2.3. Redes Neurais Artificiais em Auditoria

Esta seção descreve áreas em que a atividade de auditoria pode ser ajudada pela aplicação de redes neurais artificiais. A atividade de auditoria requer capacidade de predição, controle e classificação. A literatura demonstra que redes neurais artificiais já tiveram essas capacidades comprovadas em outros campos, e assinala que elas são poderosas ferramentas para modelar e entender o comportamento cognitivo humano e possuem propriedades de reconhecimento de padrões [41], [42].

A atividade de auditoria depende fortemente da capacidade de decisão do auditor, a compreensão de quais conjuntos de evidências são relevantes para construção de uma auditoria eficiente, que indique pontos fortes e fracos da organização, ou ainda,

que consiga responder se os dados e informações divulgadas são verdadeiras, é essencial [43]. Tal compreensão, ou coleta de evidências, é conhecida como procedimento de revisão analítica, que visa melhorar a eficiência da auditoria. Basicamente, trata-se de comparar os resultados esperados com os resultados observados[1]. As redes neurais artificiais, então, servem para suportar essa decisão selecionando conjuntos de dados a serem investigados ou indicando relações não esperadas.

Entre as áreas de auditoria que obtiveram bons resultados ao utilizar redes neurais artificiais nos procedimentos de revisão analítica destacam-se: erros materiais; fraudes de gestão; continuidade de negócio, dificuldades financeiras e falência; avaliação de controle de riscos e custos de auditoria. O objetivo dessa seção não é aprofundar-se no conceito ou prática de execução de auditoria e sim elencar áreas de auditoria e citar os modelos de redes neurais artificiais que foram utilizados por elas.

Erros Materiais

A aplicação de redes neurais artificiais na área de detecção de erros materiais tem como objetivo direcionar o auditor para aquelas contas onde os relacionamentos observados não estão consistentes com os relacionamentos esperados. É responsabilidade do auditor decidir se há necessidade de alguma investigação adicional. Os modelos de redes neurais artificiais utilizados para esse fim devem possuir capacidade tanto de predição de valores futuros quanto de classificação de dados. Devido a essa necessidade a arquitetura mais referenciada na literatura é a rede neural artificial MLP utilizando o algoritmo de aprendizagem *backpropagation*[3].

Fraudes de Gestão

Fraudes de gestão podem ser conceituadas como a fraude cometida pelo gestor de forma proposital, que prejudica investidores e credores, através de informações financeiras enganosas. Portanto, auditores devem ser imparciais quanto à atuação da gestão, considerando a possibilidade de apresentação de informações erradas pela gestão no processo de auditoria. Novamente, nesta área de auditoria, a

arquitetura de rede neural artificial mais referenciada é a MLP. Entretanto aqui, podemos observar também o algoritmo de aprendizagem AutoNet[1].

Continuidade do Negócio, Dificuldades Financeiras e Falência

Estas três áreas de auditoria estão interligadas e relacionam-se por serem avaliações distintas acerca do mesmo conceito. Continuidade do negócio, do inglês *going concern*, é a capacidade de uma organização estar financeiramente bem 01 (um) ano após a divulgação do resultado de seu balanço. Caso a organização demonstre em seu balanço sinais de falência ou outras dificuldades que ameacem a continuidade do negócio, o auditor pode emitir uma opinião negativa em seu relatório de auditoria indicando que tal organização encontra-se com dificuldades para manter-se no negócio. Entretanto, se não há possibilidade de a organização manter-se no negócio o relatório deve indicar a falência da mesma[3].

Apesar de [3] apontar para uma predominância da arquitetura MLP com algoritmo de aprendizagem *backpropagation* nos trabalhos produzidos nessa área, nesse mesmo estudo mostra-se que abordagens híbridas podem ser mais eficientes do que as redes neurais artificiais MLP com *backpropagation*. Estudos mais recentes trazem uma abordagem que utiliza a análise de trajetória de uma rede SOM e mostrou-se efetiva em predizer se uma organização está demonstrando sinais de dificuldades financeiras ou falência [10], [40], [44], [45].

Avaliação de Riscos de Controles Internos e Custos de Auditoria

Avaliação de riscos dos controles internos de uma organização é um processo sistemático, que relaciona a opinião dos auditores sobre fatores de riscos relevantes e sua importância a condições ou eventos adversos que podem ocorrer. O resultado desse relacionamento identifica atividades passíveis de auditoria [3]. Neste tipo de avaliação existe a necessidade de analisar grandes volumes de dados, impossibilitando, muitas vezes, os auditores de detectarem distorções relevantes nos demonstrativos financeiros da organização. Os relacionamentos entre as variáveis dos controles internos devem ser identificados e analisados sempre, tornando a avaliação de risco dos controles internos uma tarefa de difícil execução.

Quanto às arquiteturas de redes neurais artificiais mais utilizadas vê-se que a arquitetura MLP destaca-se, entretanto, combinada com outras técnicas de

inteligência artificial, dentre as quais se pode destacar a aplicação de sistemas especialistas [1].

2.4. Considerações Finais

Este capítulo apresentou os principais conceitos utilizados nessa dissertação. Desde a concepção de um projeto de KDD e todas as nuances que devem ser analisadas como o entendimento do problema, a modelagem apropriada da base de dados ao problema existente e a importância de uma metodologia que auxilie a fase de mineração de dados. Dentre as possíveis técnicas para mineração de dados foi escolhida a técnica de aprendizagem de máquina de redes neurais artificiais e uma revisão nos seus conceitos, nos paradigmas e algoritmos de aprendizagem de forma geral pode ser encontrada nesse capítulo, com um aprofundamento no algoritmo SOM mostrando seu processo de aprendizagem, suas propriedades e métricas de qualidade.

Por fim, uma revisão no estado da arte da aplicação de redes neurais artificiais na atividade de auditoria foi realizada. É visto que as pesquisas nessa área vêm tendo mais atenção. Entretanto, pôde-se observar nesta revisão que: a arquitetura de rede neural mais utilizada é a MLP de aprendizado supervisionado e as bases de dados utilizadas são, em sua maioria, bases de dados públicas, rotuladas, de pequeno ou médio porte e referentes a eventos passados. Essas características nem sempre podem ser encontradas em um ambiente real de produção, onde os dados estão sendo criados naquele momento pelos sistemas operacionais é com esse desafio que seguimos para o próximo capítulo onde é definido um *framework* para detecção de anomalias em bases de dados de folha de pagamento.

Capítulo

3

3. *Framework* para detecção de anomalias em bases de dados de folha de pagamento

Este capítulo apresenta as fases para construção do *framework* proposto neste trabalho, desde o entendimento do problema de detecção de anomalias em bases de folha de pagamento. Ele descreve a construção da base de dados consolidada para a etapa de mineração de dados, apresenta as ferramentas que serão usadas para criação e treinamento da rede SOM e discute a arquitetura de rede escolhida: parametrizações e formas de treinamento. Por fim, descreve-se o *framework* para detecção de anomalias em bases de dados de folha de pagamento propriamente dito.

3.1. Entendimento do Problema

Um dos grandes problemas enfrentados pelas equipes de auditoria é definir o conjunto de dados que será investigado. As evidências coletadas de forma manual durante a fase de preparação da auditoria, nem sempre são suficientes para direcionar o auditor para possíveis fraudes. Este problema cresce exponencialmente quando trata de grandes bases de dados como as que suportam sistemas de folha de pagamento.

Sistemas de folha de pagamento precisam ser auditados, pois, normalmente, as despesas de pessoal representam uma parcela importante dos custos de uma instituição, além da importância no cumprimento da legislação vigente no país.

A detecção de anomalias em uma base de folha de pagamento deve encontrar padrões dentro da base de dados, que indiquem desvio dos resultados esperados, atuando de forma automática e fornecendo aos especialistas do negócio uma amostragem de funcionários que devem ser investigados de forma mais criteriosa, agindo preventivamente, a fim de evitar o pagamento de valores errados aos funcionários e, consequentemente, o recolhimento indevido ou a sonegação de algum imposto.

3.2. Definição da Base de Dados

Como todo projeto de KDD, a definição da base de dados é essencial para que os objetivos sejam alcançados. Os dados foram coletados do sistema integrado de recursos humanos da instituição em que o experimento foi realizado. O sistema está em produção e é utilizado por diversas áreas de negócio da instituição, entre eles o setor responsável pela folha de pagamento. O sistema utiliza o modelo relacional e possui cerca de 1.750 tabelas.

Inicialmente, foi definido junto aos especialistas do negócio o subconjunto das tabelas que continham informações relacionadas ao processo da folha de pagamento. Após o primeiro levantamento, foi verificado um grande número de tabelas e dados relacionados. Portanto, uma especificação mais detalhada do problema precisou ser realizada, de modo que o número de tabelas e campos selecionados ficassem restritos aqueles que indicassem diferença de valor líquido a

ser recebido pelos funcionários ou fossem sensíveis a essas variações. As tabelas utilizadas estão elencadas abaixo com uma breve descrição dos campos que foram utilizados:

- REG_EMPREGOS – possui informações básicas do funcionário, os campos selecionados nesta tabela foram utilizados como identificadores nos vetores do espaço de entrada. Colunas utilizadas: Empresa e Matrícula do funcionário.
- REG_CARGOS – possui informações do histórico de cargos do funcionário, a mudança de cargo implica em um aumento salarial. Colunas utilizadas: Código do cargo.
- REG_HORARIOS – possui informações do histórico de horários do empregado, mudança no horário pode implicar em aumento ou redução salarial. Colunas utilizadas: Código do horário.
- REG_AFASTAMENTOS – possui informações de afastamento do empregado tais como licenças. Esta tabela foi utilizada para criar o indicador AFASTADO, quando o funcionário está afastado o indicador é setado para 1, quando o funcionário está ativo o indicador é setado para 0.
- MOVIMENTOS_CALCULADOS – esta tabela armazena as rubricas calculadas na folha de pagamento. Colunas utilizadas: Código do evento, valor do evento e data de competência.

A partir das tabelas acima foram criados programas na linguagem PL/SQL para executar: a seleção, limpeza e uma parte da transformação dos dados. Esta última consistiu na criação do indicador AFASTADO e, principalmente, na transformação dos dados da tabela MOVIMENTOS_CALCULADOS.

Originalmente, cada rubrica paga ao funcionário estava codificada como um registro na tabela MOVIMENTOS_CALCULADOS. Entretanto, para construção da base de dados consolidada foi necessário selecionar as rubricas que resultavam em variações no valor líquido do funcionário e transformá-las em colunas de modo que cada funcionário possuísse apenas um registro por data de competência.

A base de dados consolidada é composta pelas colunas abaixo discriminadas, sendo três colunas usadas como rótulos (data de competência, empresa e matrícula) e as demais como dimensões totalizando dez dimensões:

- Data de competência – represente o mês do registro de informações do funcionário. Tipo Numérico.
- Empresa – empresa na qual o funcionário trabalha. Tipo Numérico.
- Matrícula – código que identifica unicamente o funcionário. Tipo Numérico.
- Código do cargo – cargo no qual o funcionário está alocado naquela data de competência. Tipo Numérico.
- Código do horário – horário cargo no qual o funcionário está alocado naquela data de competência. Tipo Numérico.
- Indicador de afastamento – indica se o funcionário está ativo ou afastado naquela data de competência. Tipo Numérico.
- Função Gratificada – valor percebido por um funcionário investido em função de comissão. Tipo Numérico.
- Hora Extra Normal – hora extra realizada em dias normais ou fins de semana. Tipo Numérico.
- Hora Extra Normal Noturna – hora extra realizada entre 22h e 05h. Tipo Numérico.
- Complemento Gratificação Função DGF – valor percebido por funcionários que tem função em comissão incorporada ao salário. Tipo Numérico.
- Função Gratificada Substituição – valor percebido por um funcionário investido em função de comissão interinamente. Tipo Numérico.
- Adiantamento de Saldo Negativo – valor pago a funcionário de forma adiantada e que está sendo cobrado. Tipo Numérico.
- Líquido a Receber – valor final percebido pelo funcionário. Tipo Numérico.

Contudo, para o processamento pelo algoritmo SOM implementado, a base de dados consolidada precisa estar formatada como um arquivo texto separado por espaços. A última transformação a qual os dados são submetidos é a normalização, este procedimento é necessário pela natureza dos dados. O método usado para normalização dos dados é o da variância. Este método realiza uma transformação linear que dimensiona os valores de cada atributo de tal modo que sua variância seja igual a 1 e a média dos valores do atributo normalizado seja igual a zero. Este método foi escolhido em razão de que alguns atributos possuem valores contínuos e de grande variação.

3.3. Ferramentas Utilizadas

Na fase de definição da pesquisa houve a necessidade de decisão entre a utilização de uma ferramenta que implementasse o algoritmo SOM ou a implementação do mesmo. O requisito principal era que a solução escolhida fosse confiável do ponto de vista do resultado gerado pela rede. Após uma breve pesquisa, duas ferramentas foram selecionadas para testes, ambas utilizam como plataforma de desenvolvimento o *software* Matlab: o pacote *Neural Network Clustering Tool* nativo do Matlab e o pacote SOM *Toolbox* 2.0[2] [46], [47], [48] do Laboratório de Ciências da Computação e Informação da Universidade de Helsinki.

Mantendo o requisito de confiabilidade e adicionando o requisito usabilidade foi escolhido o pacote SOM *Toolbox* 2.0, pois oferece uma boa usabilidade, permitindo o acesso a todas as suas funções, as variáveis geradas durante o processo de treinamento e aos parâmetros de treinamento da rede.

Do ponto de vista da confiabilidade o pacote SOM *Toolbox* 2.0 foi utilizado por trabalhos em diversas áreas, como o estudo de métricas de qualidade para o algoritmo SOM [30], [29], [31], visualização de agrupamentos [24], [49] e predição em séries temporais [50], [39].

Para a visualização dos mapas, criação de gráficos com as projeções do mapa SOM no espaço de entrada e extração de algumas métricas de qualidade foi utilizado o pacote SOMVIS *Toolbox*[3], utilizado em [51], desenvolvido pelo Instituto de Tecnologia de Software e Sistemas Interativos da Universidade de Tecnologia de Viena. Este pacote utiliza os programas, estruturas de dados e variáveis do SOM *Toolbox* 2.0.

3.4. Parametrização da Rede

Após a fase de pré-processamento dos dados, inicia-se a fase de mineração de dados propriamente dita. A escolha dos parâmetros de treinamento da rede neural artificial SOM é parte vital desse processo, pois uma rede mal ajustada pode levar a resultados inócuos. Nessa fase várias arquiteturas devem ser testadas com o

[2] http://www.cis.hut.fi/projects/somtoolbox/
[3] http://www.ifs.tuwien.ac.at/dm/somvis-matlab/index.html

objetivo de minimizar o retrabalho, pois, como foi visto o processo de KDD é iterativo, entretanto uma rede bem ajustada nessa fase gera melhores resultados para a fase de análise dos resultados.

Topologia da Rede

No mapa SOM existem duas propriedades espaciais que precisam ser definidas. A primeira diz respeito à forma como os neurônios interligam-se aos seus vizinhos. Como ilustra a Figura 10, os neurônios podem possuir: uma vizinhança retangular, interligando-se a quatro neurônios vizinhos, exceto nas bordas do mapa; ou uma vizinhança hexagonal, interligando-se a seis neurônios vizinhos, exceto nas bordas do mapa.

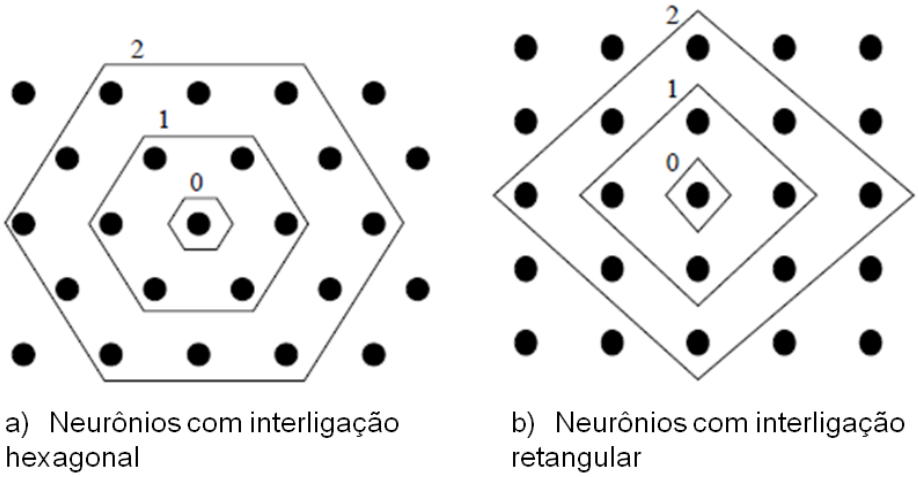


Figura 10 – Tipos de Vizinhança: a) Vizinhança com interligação hexagonal, b) Vizinhança com interligação retangular (adaptado de [48]**).**

A segunda propriedade é a forma da grade em si. O algoritmo SOM utilizado no framework, permite o uso de três tipos de grade para o mapa SOM: folha, cilíndrica ou toroidal, as duas últimas mais apropriadas para representar dados circulares[48]. A Figura 11 apresenta os tipos de grade citados aqui.

O *framework* proposto utiliza o tipo de vizinhança hexagonal e o formato folha, por se tratarem de arquiteturas que representam com maior simplicidade e acurácia dados não cilíndricos [52].

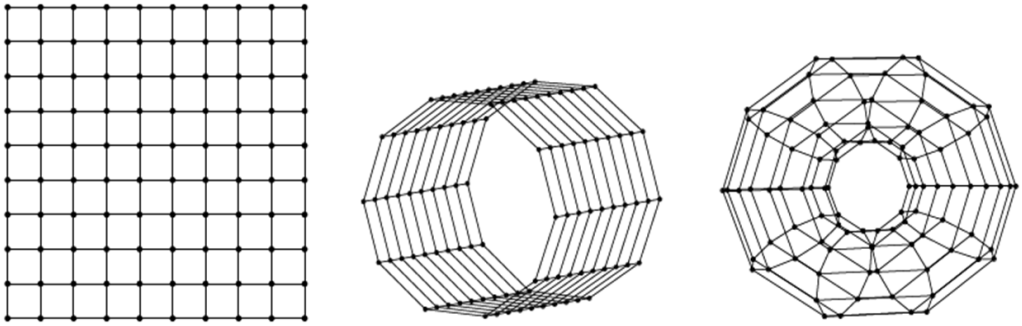


Figura 11 – Tipos de Grid (adaptado de [48]**).**

Tamanho da Rede

A decisão sobre o tamanho da rede neural SOM, isto é, quantos neurônios um mapa SOM terá no seu espaço de saída, é um problema muito discutido e ainda não resolvido na literatura. Muitos fatores influenciam na escolha do tamanho da rede e um fator bastante importante é o tamanho do espaço de entrada.

As medidas erro de quantização, erro topográfico e produto topográfico podem auxiliar nessa escolha. Será considerado o melhor mapa aquele que minimizar os valores das duas primeiras medidas e manter o valor do produto topográfico próximo de 0 [21]. Entretanto, vale ressaltar o *trade-off* existentes entre o erro de quantização e o erro topográfico. A Figura 12 ilustra o produto topográfico para uma das topologias de rede testadas, com a evolução das variáveis P1, P2 e P3 conforme o aumento do número dos k-vizinhos mais próximos e no topo o resultado do Produto Topográfico, P, estendido para todo o mapa SOM.

A ferramenta utilizada fornece uma heurística que determina o tamanho do mapa SOM em função do conjunto de dados de entrada, utilizando a seguinte fórmula.

$$N \cong 5 \times \sqrt{x}$$

Equação 12 – Heurística para determinar tamanho do mapa SOM

Onde, *N* é o número de neurônios e x é a quantidade de vetores do espaço de entrada.

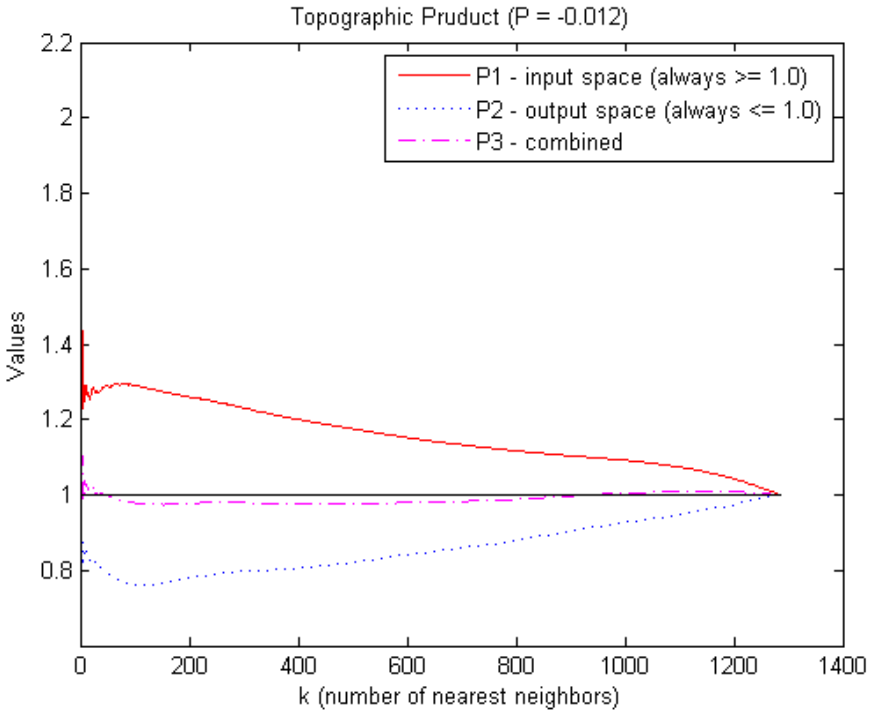


Figura 12 – Produto Topográfico para um mapa SOM com grade 39x33 neurônios.

Parâmetros de Treinamento da Rede

Até agora foram mostrados os parâmetros de rede que definem as propriedades espaciais do mapa, tais propriedades são muito importantes, pois definem como o estímulo do neurônio vencedor será propagado e como o espaço de entrada será representado no mapa SOM. Entretanto, ainda não foram definidos os parâmetros de treinamento, tais parâmetros ajustam como o algoritmo SOM irá realizar o treinamento, a função de vizinhança utilizada, quantas épocas terá o treinamento e a forma de inicialização da rede.

O pacote SOM *Toolbox* 2.0 divide o treinamento da rede em três etapas: inicialização, treinamento abrangente e treinamento para sintonia fina.

Na inicialização, os pesos sinápticos da rede são inicializados. É possível escolher a função de inicialização da rede que pode ser linear em função dos dois maiores autovetores da matriz de covariância dos dados de treinamento, ou randômica em função dos valores máximo e mínimo de cada componente dos dados de treinamento.

Traçando um paralelo com o algoritmo SOM visto na seção 2.2.2.1 deste trabalho, no treinamento abrangente ocorrem os processos competitivo e cooperativo, que

compreende a escolha do neurônio vencedor com base no vetor de entrada escolhido como exemplo, e a parte do processo adaptativo que diz respeito à ordenação topológica. O treinamento para sintonia fina pode ser comparado à fase de convergência do processo adaptativo. Na prática a diferença entre as duas etapas está no parâmetro raio, tal parâmetro possui valor mais alto na fase de treinamento abrangente e decai ao longo das épocas dessa fase de treinamento.

A função de vizinhança e a quantidade de épocas de treinamento são parâmetros que podem ser definidos, independente, do algoritmo de treinamento escolhido. Os algoritmos disponíveis são: treinamento sequencial e treinamento *batch*. Ambos treinam a rede de forma iterativa, a diferença entre os dois é que, a cada época, o algoritmo sequencial escolhe aleatoriamente o vetor de entrada que será apresentado à rede, enquanto que o algoritmo *batch* apresenta todo o espaço de entrada à rede SOM.

Para obter a melhor parametrização da rede do *framework* foi conduzida uma pesquisa empírica, sendo utilizadas para aferição do desempenho da rede as métricas já citadas no capítulo de fundamentação teórica: erro de quantização, erro topográfico e produto topográfico. Após testar algumas parametrizações para rede do *framework* aqui apresentado, a pesquisa conduziu a melhor parametrização da rede SOM. Abaixo um quadro resumo da melhor parametrização encontrada para a rede SOM.

Parâmetro	Valor
Vizinhança dos Neurônios	Hexagonal
Grade do Mapa	Folha
Quantidade de Neurônios	1.287
Topologia dos Neurônios	39x33
Algoritmo de Inicialização	Linear
Função Vizinhança	Gaussiana
Algoritmo de Treinamento	*Batch*
Épocas Treinamento Abrangente	50
Épocas Treinamento Sintonia Fina	15
Raio Inicial Treinamento Abrangente	5
Raio Final Treinamento Abrangente	1,25
Raio Inicial Treinamento Sintonia Fina	1,25
Raio Final Treinamento Sintonia Fina	1

Tabela 2 – Quadro Resumo parametrização Rede SOM

3.5. Treinamento da Rede

Para treinar e validar as redes, foram usados os dados da base de dados construída nos moldes da seção 3.2. Os dados de uma base de folha de pagamento têm uma natureza temporal, cada rubrica ou um conjunto de rubricas são pagos uma única vez por ciclo de folha de pagamento, a periodicidade deste ciclo varia com as características de cada instituição. Nos experimentos realizados uma periodicidade mensal foi considerada para o ciclo de folha de pagamento.

A natureza temporal dos dados é importante em dois momentos da aplicação do framework proposto: (i) na ordenação dos dados para visualização e (ii) no treinamento da rede, uma vez que os dados demonstram um comportamento progressivo. Os dados da base de folha de pagamento podem ser classificados como tempos-dependentes possuindo um rótulo de tempo de forma que, considerando todo o conjunto de dados, um vetor de entrada escolhido de forma aleatória representa informações válidas para um ponto no tempo.

A divisão dos conjuntos de treinamento e validação seguiu respeitando uma ordenação temporal, de tal forma que para analisar a folha de pagamento de um determinado mês utilizando o framework, são selecionados os meses imediatamente anteriores ao mês analisado. Assim como em uma série temporal uma linha de atraso também deveria ser escolhida. A definição da linha de atraso foi tomada considerando um período de meses que abrangesse as situações mais cotidianas na evolução das rubricas consideradas para a criação da base de dados. Entre elas estão: promoção por mérito do funcionário, reajuste salarial, férias e 13º salário. Todas essas situações ocorrem uma vez ao ano em períodos distintos, porém bem definidos. Ao fim, a linha de atraso foi oportunamente definida em 12 meses.

A regra para divisão dos conjuntos de validação e treinamento ficou assim definida:

- Base de validação – a base de dados pré-processados da folha de pagamento referente ao mês que se deseja verificar a existência ou não de anomalias.

$$\chi(v) = m$$

Equação 13 – Definição do conjunto de validação da rede

- Base de treinamento – a base de dados pré-processados da folha de pagamento dos doze meses anteriores ao mês que se deseja verificar a existência ou não de anomalias. Esta linha de atraso foi escolhida, pois os eventos de uma folha de pagamento podem ser observados ao longo de um ano.

$$\chi(t) = \{m-1, m-2, \dots, m-11, m-12\}$$

Equação 14 – Definição do conjunto de treinamento da rede

As equações Equação 13 e Equação 14 descrevem os conjuntos de validação e treinamento da rede, onde $\chi(v)$ é o espaço de entrada de validação, $\chi(t)$ é o espaço de entrada de treinamento e m é o mês investigado.

A rede SOM é criada com os parâmetros supracitados e treinada com os dados da base de treinamento. Após o treinamento, a validação é feita apresentando a base de validação à rede.

3.6. Modelo para detecção de anomalias em bases de folha de pagamento

O *framework* para a detecção de anomalias em uma base de dados de folha de pagamento baseado em mapas auto-organizáveis proposto nessa dissertação tem o objetivo de selecionar um conjunto de funcionários que podem ter anomalias em sua folha de pagamento. Para tanto, a abordagem em dois passos descrita em [12] é utilizada:

1. Análise visual da projeção do espaço de entrada no mapa SOM.
2. Detecção de possíveis anomalias em dados que estejam inseridos em grupos de neurônios.

No entanto, em [12] seu objetivo era a detecção de *outliers* em uma base de dados, sem preocupação adicional sobre o significado do dado que estava sendo considerado *outlier*. As bases de dados utilizadas em seus experimentos são, em sua maioria, bases geradas a partir de modelos matemáticos que não exigem uma redução significativa de dimensionalidade ao ser representada por um mapa SOM [12]. Em nosso estudo de caso, entretanto, lida-se com um problema real e o dado considerado anomalia precisa ser investigado por um especialista a fim de

comprovar se há erro no cálculo da folha de pagamento. A Figura 13 ilustra o fluxo de detecção de anomalias que o *framework* proposto deve executar.

Em linhas gerais o *framework* define que após o conjunto de dados estar definido para o mês a ser explorado uma rede SOM deve ser treinada, a partir dos resultados obtidos os dois passos do *framework* podem ser executados. A análise visual utilizará as projeções para identificar possíveis anomalias e a análise dos erros de quantização vai usar o conjunto de erros de quantização gerado para toda rede SOM e discriminar aqueles neurônios que ficarem fora do intervalo de confiança definido.

Ao fim, os especialistas de negócio devem validar quais vetores de dados detectados como possíveis anomalias são erros propriamente ditos e quais vetores de dados estão apenas com um comportamento que foge do comum, mas que não possuem erros.

Análise visual da projeção do espaço de entrada no mapa SOM

Dado uma rede SOM treinada, um neurônio é considerado *outlier* quando seu vetor de pesos sinápticos encontra-se relativamente distante dos demais [12]. Os neurônios da rede SOM são uma representação discreta de um espaço de entrada contínuo, ou seja, para visualizar neurônios considerados *outliers* é necessário projetá-los sobre os dados do espaço de entrada. A ferramenta SOMVIS *Toolbox* fornece diversas projeções tanto para os dados quanto para o mapa SOM. Foi escolhido o método análise das componentes principais, PCA, por sua capacidade de redução de dimensionalidade na projeção de dados [53]. A projeção PCA é realizada para três dimensões, entretanto para melhor visualização do mapa e do espaço de entrada, as projeções serão exibidas em duas dimensões. Uma vez projetados, neurônios distantes dos demais representam dados ou subconjuntos de dados que também estão distantes, como ilustrado na Figura 14.

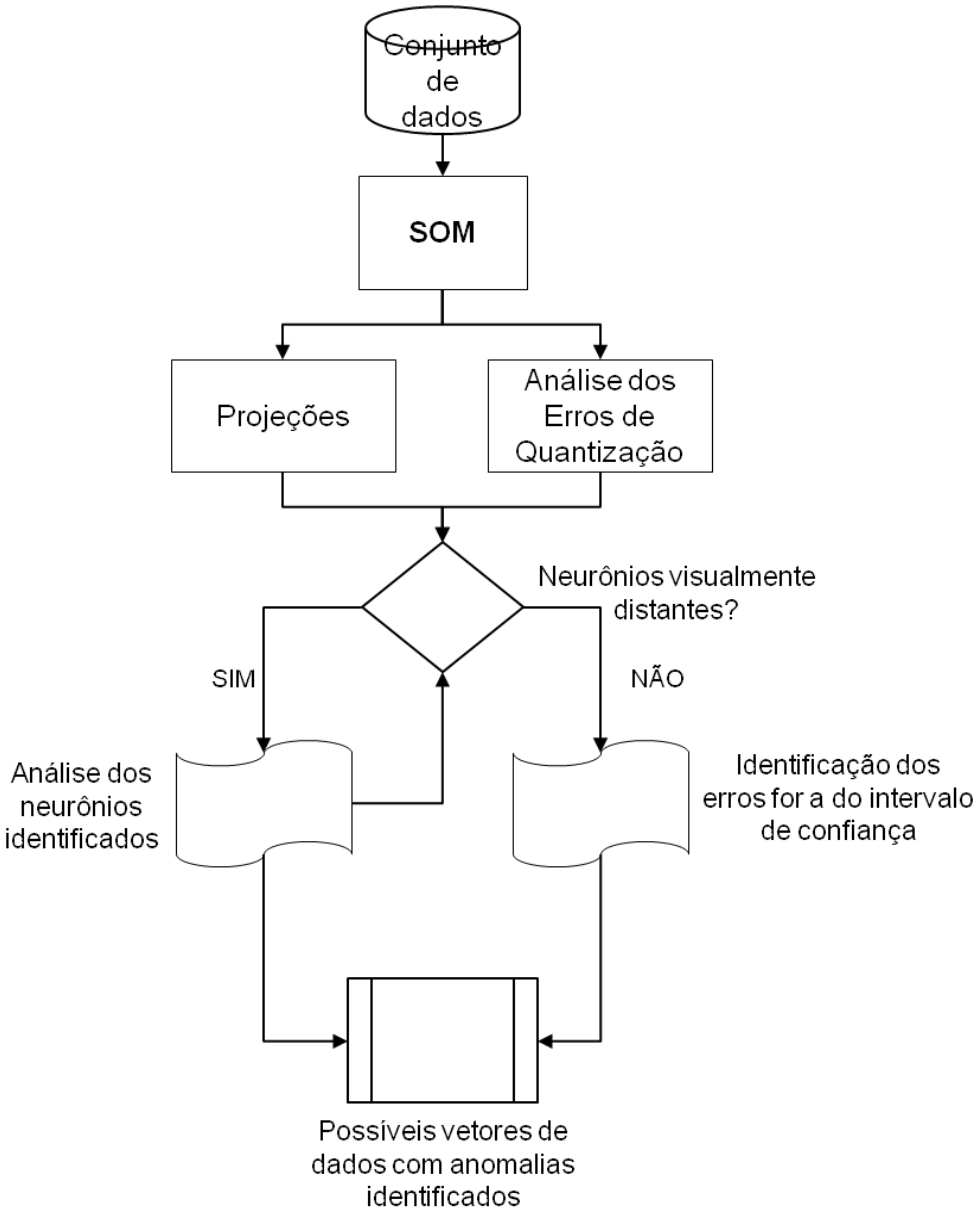
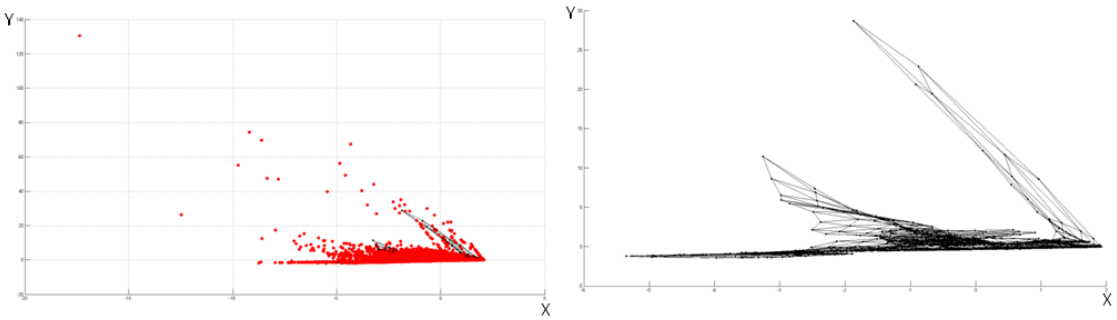


Figura 13 – Fluxo do framework para detecção de anomalias em bases de dados de folha de pagamento



a) Projeção da Rede SOM e do Espaço de Entrada

b) Projeção da Rede SOM

Figura 14 – Exemplo de uma Rede SOM projetada em seu Espaço de Entrada

A análise visual pode identificar tanto vetores de entrada quanto neurônios que estejam destacados do grupo. Uma característica do pacote SOM *Toolbox* 2.0 é que mesmo após as transformações sofridas nos dados e na grade original é possível identificar cada vetor de entrada e cada neurônio através de rótulos inseridos durante as fases de carga dos dados e de treinamento da rede. Dessa forma, o especialista poderia identificar visualmente os funcionários que estão em uma área mais afastada do gráfico.

Detecção de possíveis anomalias em dados que estejam inseridos em grupos de neurônios.

O segundo passo na detecção de anomalias em bases de folha de pagamento, consiste em obter o conjunto de erros de quantização do espaço de entrada do conjunto de dados de validação e classificar, por meio de algum parâmetro, quais dados do espaço de entrada serão considerados anomalias, logo, necessitando da avaliação dos especialistas.

Como determinar quais erros são significativamente grandes é uma questão que precisa ser definida. Alguns métodos baseados em intervalo de confiança foram propostos na literatura. Em [12], uma abordagem estatística completa busca no conjunto de erros de quantização de uma rede SOM quais erros podem ser classificados como *outliers* e demonstra isso através de alguns métodos de visualização estatísticos como o histograma e o *box-plot*.

No entanto, essa abordagem é de difícil aplicação ao problema de auditoria em bases de folha de pagamento, pois, como em situações reais as bases de folha de pagamento possuem milhares de registros, é possível que o número de vetores de entrada indicados como *outliers* seja demasiadamente grande a ponto de inviabilizar a análise dos especialistas.

Outro método visto em [13] e revisado posteriormente em [14], propõe a criação de um intervalo de confiança, aqui chamado de intervalo de percentil por não possuir as características necessárias para ser classificado como intervalo de confiança, baseado em um percentual qualquer p, onde $p = 1-\alpha$, e considerando que a distribuição do conjunto de erros de quantização possui um intervalo de

normalidade. Os limites superior e inferior desse intervalo são calculados pelo método estatístico percentil, tal como segue [13]:

- Limite Inferior (e_p^-): é o $100(\frac{\propto}{2})^O$ percentil da distribuição dos erros de quantização associados aos vetores de entrada.
- Limite Superior (e_p^+): é o $100(1-\frac{\propto}{2})^O$ percentil da distribuição dos erros de quantização associados aos vetores de entrada.

O intervalo de normalidade pode ser usado para classificar tanto um novo vetor de entrada **x** como todo o espaço de entrada χ como normal/anormal, de acordo com seu erro de quantização, pela simples condição:

$$\text{SE } e^{novo} \in [e_p^-, e_p^+]$$
$$\text{ENTÃO } \boldsymbol{x}^{novo} \text{ é NORMAL}$$
$$\text{SENÃO } \boldsymbol{x}^{novo} \text{ é ANORMAL}$$

Esta condição estabelece que se o erro de quantização do vetor de entrada apresentado à rede $\boldsymbol{x}^{novo}$ está dentro do intervalo de normalidade $[e_p^-, e_p^+]$, então o cálculo da folha de pagamento deste funcionário está normal com $100p$% de confiança.

Assim, os especialistas podem definir o conjunto de análise de acordo com os limites superior e inferior do intervalo de normalidade, ou em uma situação ideal analisar todo conjunto de dados que está sendo validado para encontrar o α ideal. Os experimentos relatados no próximo capítulo, utilizando o *framework* definido neste trabalho e ilustrado na Figura 13, conduziram a pesquisa para um α ideal de 1% do conjunto total de dados, esse percentual foi o suficiente para que se identificassem os funcionários com comportamentos que fogem da normalidade e poderiam indicar possíveis erros na folha de pagamento.

3.7. Considerações Finais

As fases do processo de KDD foram objeto de estudo nesse capítulo. A definição e entendimento do problema, onde foi explicado o quão sensível é o processo de cálculo de folha de pagamento para uma organização. A criação da base de dados,

sendo explicada a origem dos dados a fase de pré-processamento até a base encontrar-se no formato adequado para a fase de mineração de dados. A justificativa da escolha do pacote SOM *Toolbox* 2.0 como ferramenta de apoio para criação da rede SOM e geração dos mapas e gráficos. A parametrização da rede e a escolha dos subconjuntos de dados para realização do treinamento da rede.

Por fim, foi especificado um *framework* para detecção de anomalias em bases de folha de pagamento baseado em mapas auto-organizáveis, considerando: a distância entre os neurônios projetados no espaço de entrada [12] e a classificação do erro de quantização de determinado vetor de entrada em relação a um intervalo de confiança [13].

Capítulo

4

4. Aplicação do *Framework* para detecção de anomalias em bases de dados de folha de pagamento

---

Este capítulo apresenta os experimentos realizados usando o *framework* proposto neste trabalho. Iniciamos com uma descrição detalhada sobre a seleção dos dados para os experimentos e apresentamos os resultados dos experimentos nas duas fases do *framework*. O capítulo também faz uma discussão sobre a aplicação da análise de trajetórias SOM para o problema de detecção de anomalias em bases de dados de folha de pagamento, incluindo uma comparação entre a análise de erros de quantização e a análise de trajetórias SOM na solução deste problema.

4.1. Seleção dos dados dos experimentos

Ao todo, três pares de bases treinamento-validação foram criados para realização dos experimentos. Os períodos de cada par treinamento-validação foram escolhidos por um conhecimento a priori da base de dados, sabia-se que nos períodos selecionados existem alguns casos que o *framework* deveria detectar como possíveis anomalias e outros que demonstram a capacidade da rede SOM aprender com os exemplos apresentados a ela.

Abaixo segue uma breve descrição das bases:

- Base1: Conjunto de dados de validação referente ao mês de Abril/2011, com espaço de entrada de 5.310 vetores e conjunto de dados de treinamento referente aos meses de Abril/2010 à Março/2011, com espaço de entrada de 63.378 vetores. O conjunto de dados de validação dessa base possui casos que o *framework* deveria detectar como possível anomalia da base. A Figura 15 – Produto Topográfico do mapa SOM para a base de dados Base1 mostra o produto topográfico da rede SOM para a Base1, o valor do produto topográfico, P = 0.007, demonstra que a rede SOM possui um bom tamanho para a Base1.

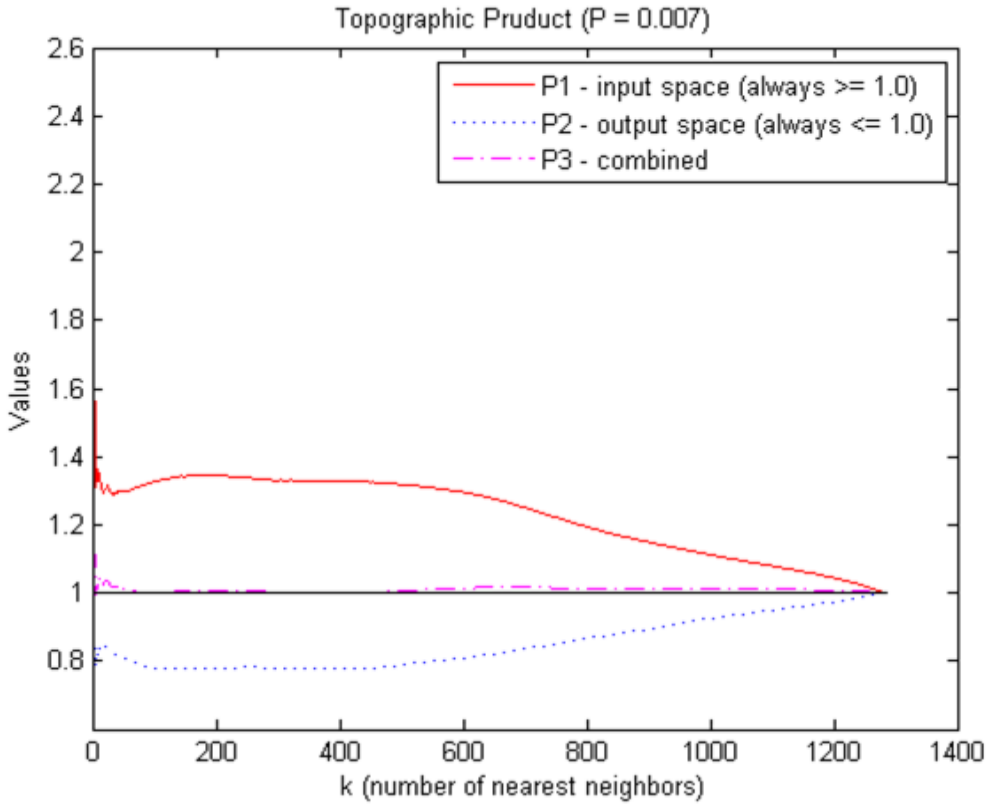


Figura 15 – Produto Topográfico do mapa SOM para a base de dados Base1

- Base2: Conjunto de dados de validação referente ao mês de Maio/2011, com espaço de entrada de 5.144 vetores e conjunto de dados de treinamento referente aos meses de Maio/2010 à Abril/2011, com espaço de entrada de 63.176 vetores. Apesar de o processo de aprendizado ser inerente à rede SOM, o conhecimento a priori sobre esta base permitiu comprovar a capacidade de aprendizado da mesma. A Figura 16 mostra o produto topográfico da rede SOM para a Base2, o valor do produto topográfico, *P* = 0.002, demonstra que a rede SOM possui um bom tamanho para a Base2.

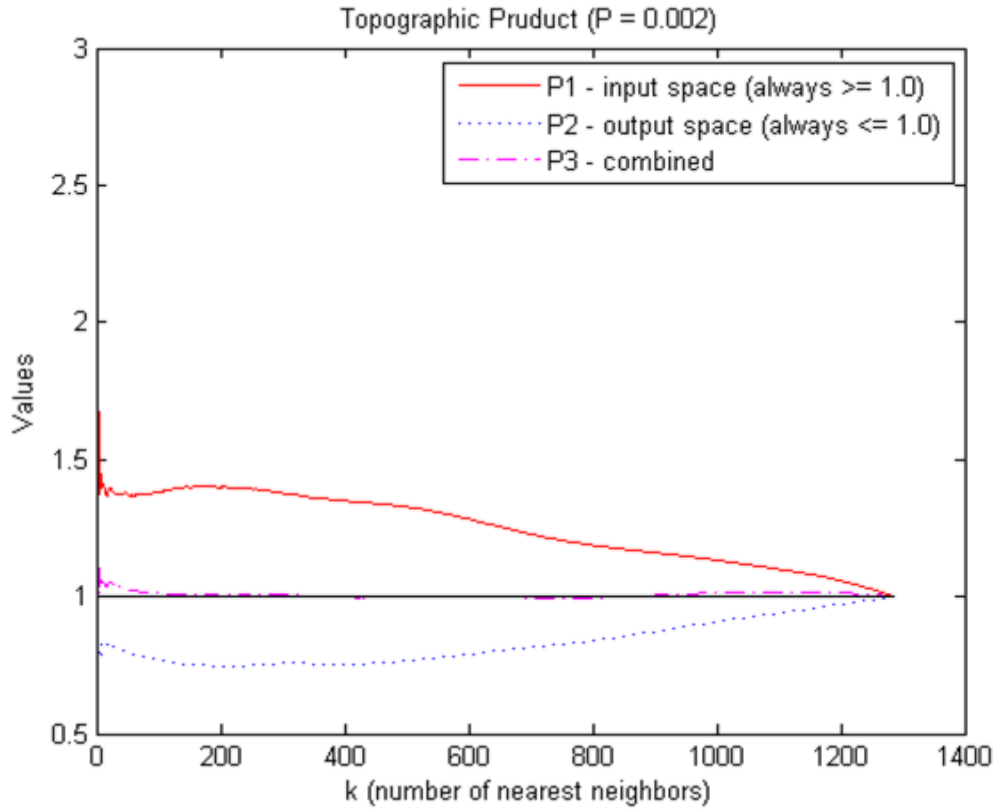


Figura 16 – Produto Topográfico do mapa SOM para a base de dados Base2

- Base3: Conjunto de dados de validação referente ao mês de Janeiro/2011, com espaço de entrada de 63.755 vetores e conjunto de dados de treinamento referente aos meses de Janeiro/2010 à Dezembro/2010, com espaço de entrada de 4.515 vetores. O conjunto de dados de validação dessa base possui um erro inserido no processo de pré-processamento dos dados. A Figura 17 mostra o produto topográfico da rede SOM para a Base3, o valor do produto topográfico, *P* = -0.006, demonstra que a rede SOM possui um bom tamanho para a Base3.

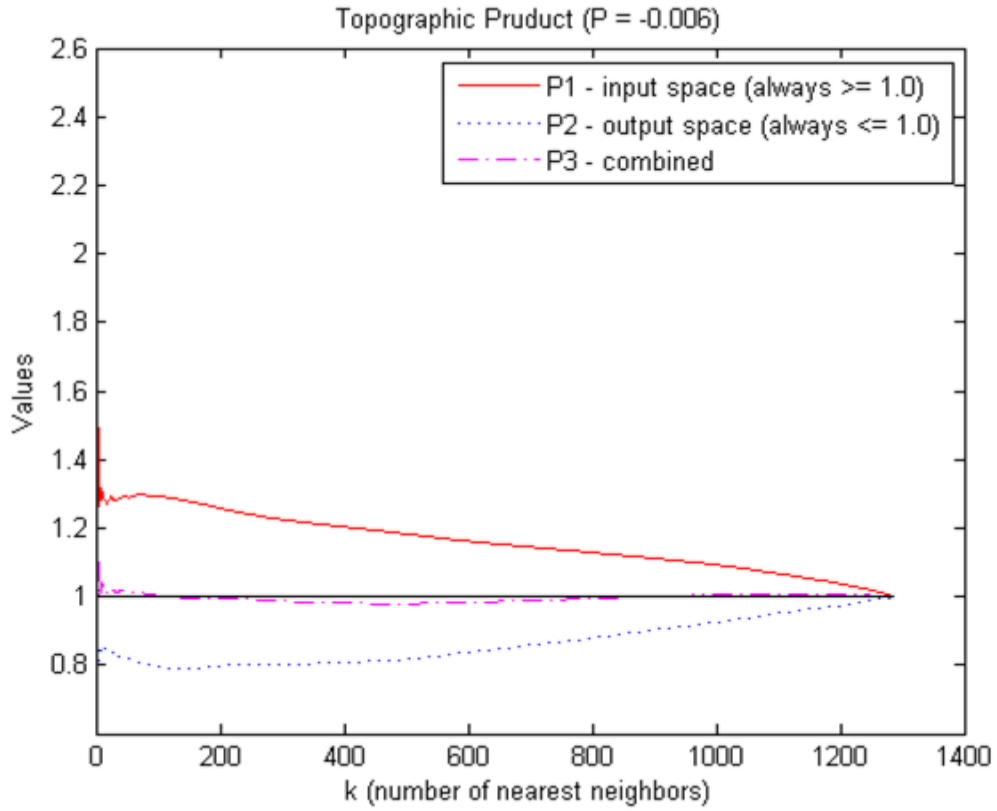


Figura 17 – Produto Topográfico do mapa SOM para a base de dados Base3

4.2. Análise visual da projeção do espaço de entrada no mapa SOM.

A análise visual da projeção do espaço de entrada no mapa SOM explora uma das principais características da rede SOM, a sua capacidade de visualização. Para demonstrar esse passo do *framework* será utilizada a Base3 definida na seção 4.1. A Base3 possui um erro que foi inserido durante a fase de pré-processamento dos dados. Alguns funcionários tiveram a coluna Líquido a Receber preenchida com o valor da coluna Data de competência. O desvio gerado por esse erro pode ser detectado visualmente na projeção do mapa SOM sobre o espaço de entrada.

Um mapa SOM é apenas um conjunto de neurônios ligados entre si respeitando sua topologia e formato já descritos na seção 3.4.1, conforme ilustrado na Figura 18. Durante o processo de treinamento da rede os vetores de pesos sinápticos são influenciados pela distribuição do espaço de entrada, o mapa SOM então assume uma forma que tenta adequar-se ao espaço de entrada, o resultado dessa deformação provocada pelo processo de treinamento pode ser visualizado através

da projeção dos vetores de pesos sinápticos da rede utilizando o método de análise das componentes principais, como pode ser visto na Figura 19.

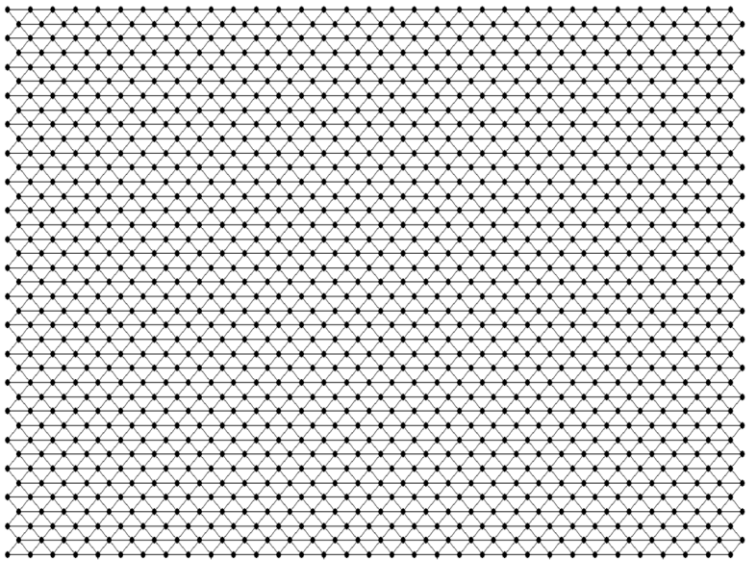


Figura 18 – Mapa SOM Topologia Hexa e Formato de Folha

Analisando a projeção do mapa SOM é possível perceber que um vetor de entrada projetado em uma área remota do espaço pode influenciar mais de um neurônio no espaço de saída. A Figura 20 utiliza a projeção das componentes principais tanto dos vetores de pesos sinápticos dos neurônios da rede quanto dos vetores de entrada para ilustrar a influência dos vetores de entrada, destacados em verde na Figura 20, na distribuição dos neurônios no mapa SOM. Outra característica é que algumas áreas do mapa tem uma densidade maior de neurônios e em outras áreas os neurônios estão distribuídos de forma mais esparsa. Isso acontece por conta da distribuição dos dados do espaço de entrada.

O *framework* proposto indica que uma análise mais criteriosa seja realizada para os vetores de entrada que estão nas áreas mais distantes da projeção do mapa. O mapeamento da rede SOM baseia-se no cálculo da distância Euclidiana entre os vetores de entrada e os vetores de pesos sinápticos, vetores de entrada mais distantes no gráfico representam uma distância maior do vetor de pesos sinápticos, significando que a rede ainda não possui o conhecimento apresentado por esses vetores de entrada, podendo significar uma anomalia.

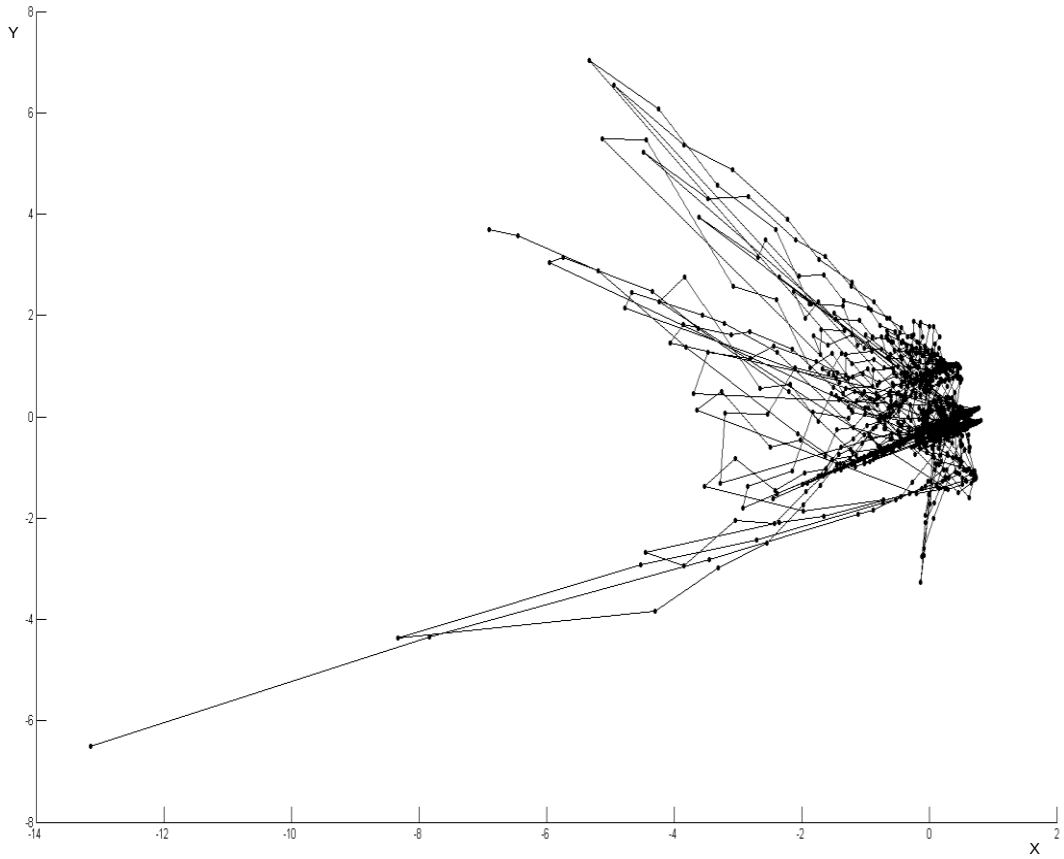


Figura 19 – Projeção do Mapa SOM após treinamento da rede

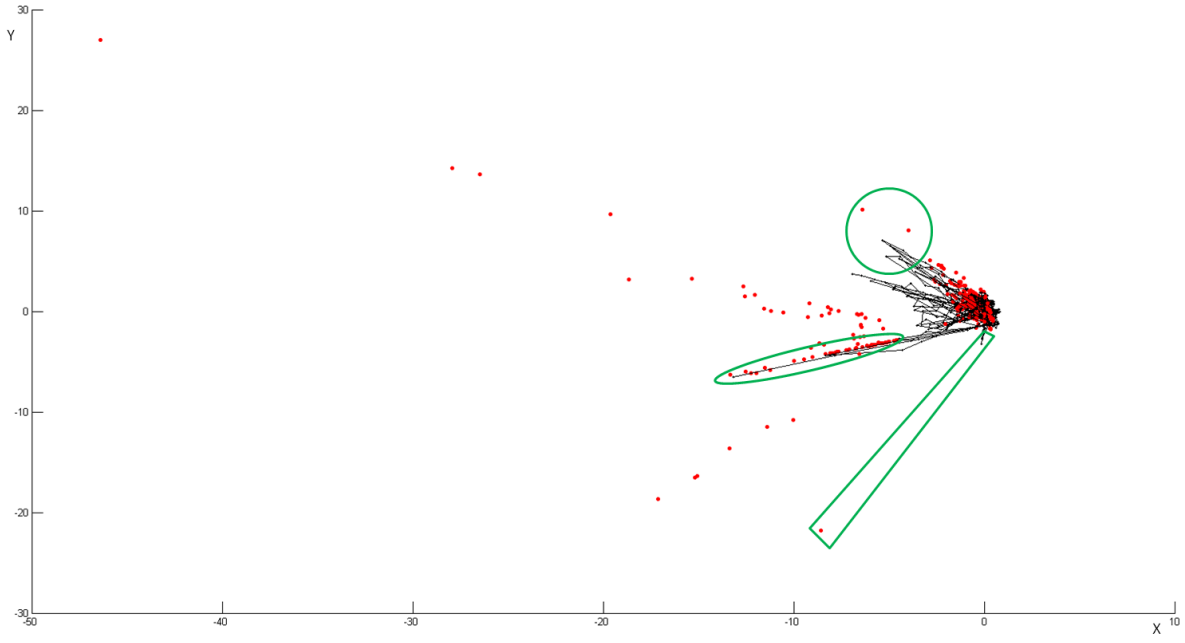


Figura 20 – Influência dos vetores de entrada sobre a distribuição dos neurônios da rede SOM

As projeções exibidas nesta seção foram obtidas da seguinte maneira:

1. A rede foi treinada com o conjunto de dados de treinamento da Base3;
2. O conjunto de dados de validação da Base3 foi apresentado à rede SOM já treinada. Este conjunto de dados possui um erro inserido como explicado anteriormente;
3. Foram encontrados os melhores neurônios para cada vetor entrada do conjunto de validação;
4. Utilizando as funções de projeção e visualização das componentes principais da ferramenta SOM *Toolbox* 2.0 foram criados os mapas aqui exibidos.

O erro inserido no conjunto de validação atinge 86 registros do mês de competência Janeiro/2011. Na Figura 21, que mostra os dados projetados sobre o mapa da rede, foram identificados com círculos vermelhos os funcionários que tiveram registros com erros e foram utilizados para análise da distância entre o vetor de entrada desse funcionário e seu respectivo melhor neurônio. Para melhor visualização, optou-se por exibir apenas alguns casos. Na mesma figura também se visualiza que os vetores de entrada com o desvio gerado pelo erro da base, rotulados com as matrículas dos funcionários, estão dispersos pelo gráfico influenciando a disposição dos neurônios da rede.

Abaixo estão demonstradas as análises de quatro funcionários, três funcionários com erros na base de dados e um com a base de dados correta, para exemplificar como a distância entre as componentes principais do vetor de entrada e as componentes principais do melhor neurônio pode ser um bom indicativo na detecção de anomalias.

	Matrícula	95.486		Data de Competência	Janeiro/2011
	Vetor de Entrada	822		Melhor Neurônio	39
Coordenadas na projeção	x	-10,0388		x'	-4,2994
	y	-10,8044		y'	-3,8371
	z	-0,4169		z'	-0,0656
	9,0337	Distância do vetor de entrada para o melhor neurônio.			

Tabela 3 – Análise da distância entre o vetor de entrada e o melhor neurônio funcionário 95.846

	Matrícula	225.860		Data de Competência	Janeiro/2011
	Vetor de Entrada	3946		Melhor Neurônio	5
Coordenadas na projeção	x	-8,5792		x’	-0,8794
	y	-21,8179		y’	-1,8441
	z	25,0598		z’	1,1363
	32,1026	Distância do vetor de entrada para o melhor neurônio.			

Tabela 4 – Análise da distância entre o vetor de entrada e o melhor neurônio funcionário 225.860

	Matrícula	108.227		Data de Competência	Janeiro/2011
	Vetor de Entrada	1224		Melhor Neurônio	273
Coordenadas na projeção	x	-46,4106		x’	-5,3232
	y	26,9889		y’	7,0298
	z	-0,5241		z’	0,3783
	45,6876	Distância do vetor de entrada para o melhor neurônio			

Tabela 5 – Análise da distância entre o vetor de entrada e o melhor neurônio funcionário 108.227

	Matrícula	214.620		Data de Competência	Janeiro/2011
	Vetor de Entrada	3396		Melhor Neurônio	874
Coordenadas na projeção	x	0,2830		x’	0,2926
	y	-0,1128		y’	-0,0740
	z	-0,3822		z’	-0,4747
	0,1008	Distância do vetor de entrada para o melhor neurônio			

Tabela 6 – Análise da distância entre o vetor de entrada e o melhor neurônio funcionário 214.620

Como pode ser observado nas tabelas Tabela 3, Tabela 4 e Tabela 5 os vetores de entrada dos funcionários que possuem erro na base de dados estão mais distantes dos seus respectivos melhores neurônios quando comparados ao vetor de entrada do funcionário cuja base de dados não contém erros para o mês de referência, Tabela 6. A Figura 21 é uma representação da Figura 20 com destaque para os vetores de entrada analisados acima e que possuem o erro inserido no conjunto de dados de validação da Base3.

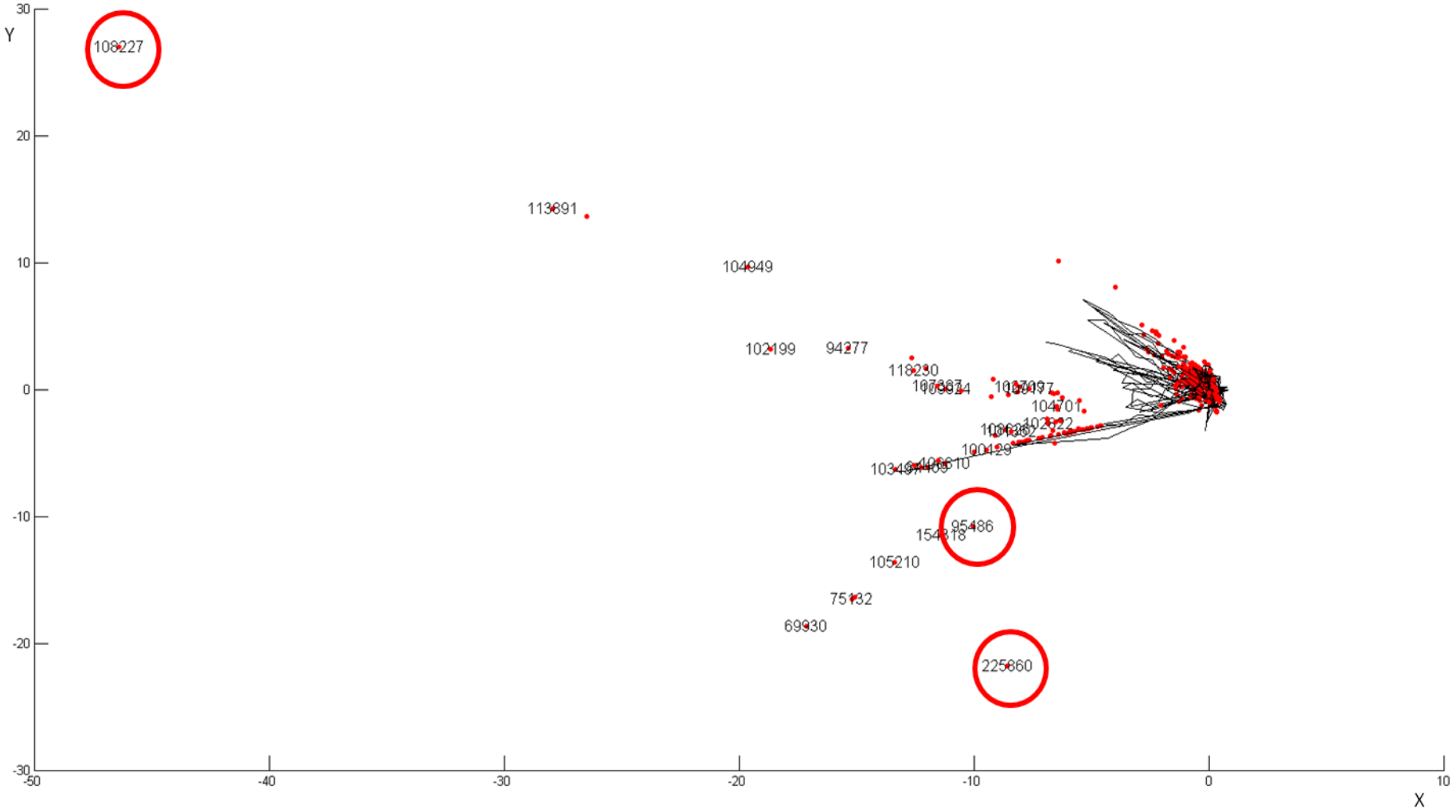


Figura 21 – Projeção dos dados na Rede SOM

4.3. Detecção de possíveis anomalias em dados que estejam inseridos em grupos de neurônios.

A seção anterior apresentou experimentos que comprovam a utilidade do primeiro passo do *framework* proposto nessa dissertação. Como foi visto, vetores de entrada distantes de seus respectivos melhores neurônios apontam para possíveis anomalias. Entretanto, a abordagem supracitada limita-se a análise dos vetores de entrada periféricos ao mapa SOM, não é possível, por exemplo, detectar visualmente vetores de entrada distantes de seus respectivos melhores neurônios nas áreas mais densas do mapa. As possíveis anomalias não detectadas visualmente, onde a densidade de neurônios é alta, precisam de um método complementar para realizar essa classificação. O segundo passo do *framework* propõe a solução para este problema.

O experimento aqui desenhado visa demonstrar o uso do método de análise dos erros de quantização para detecção de possíveis anomalias em uma base de dados de folha de pagamento. Para este experimento foram criadas duas bases de dados Base1 e Base2, detalhadas na seção 4.1. Ambas as bases são conjuntos de dados reais correspondentes a ciclos de folha passados e não possuem registro de erros detectados. Porém, os ciclos de folha não foram escolhidos aleatoriamente, com o objetivo de comprovar a eficácia do método proposto foram escolhidos dois ciclos nos quais já era sabido existir casos de comportamentos fora do usual para um funcionário.

A rede SOM, inicialmente, é treinada com o conjunto de dados de treinamento da Base1. Após o treinamento, o conjunto de dados de validação da Base1 é apresentado à rede SOM. Utilizando as funções de qualidade do pacote SOM *Toolbox* 2.0, os erros de quantização de cada vetor de entrada do espaço de entrada de validação são medidos.

Para analisar o conjunto de erros de quantização e classificar quais vetores de entrada serão considerados como possíveis anomalias é preciso definir o intervalo de percentil da rede. Seguindo os critérios citados na seção 3.6, os especialistas devem definir o valor α para encontrar os limites superior e inferior (e_p^-, e_p^+) do intervalo de percentil. Para α = 0,01 então e_p^- = 0,5 e e_p^+ = 99,5, logo o intervalo de

percentil será [0,5; 99,5], ou seja, qualquer vetor de entrada cujo erro de quantização faça parte dos 0,5% menores erros de quantização da rede ou dos 99,5% maiores erros de quantização da rede precisará ser investigado.

Aplicou-se o método estatístico percentil para ordenar o conjunto de erros de quantização dos vetores de entrada do conjunto de validação. De um total de 5.310 registros no conjunto de validação, foram identificadas 22 ocorrências acima do limite superior de 99,5% e 27 ocorrências abaixo do limite inferior de 0,5%, enquanto que 5.261 registros foram classificados dentro do intervalo de normalidade. Segundo o *framework* proposto as 49 ocorrências precisam ser investigadas por especialistas, pois são possíveis anomalias nos dados de validação.

A análise foi realizada nas 49 ocorrências detectadas fora do intervalo de percentil é exibida na Tabela 7. Em seguida, um quadro resumo com os tipos de atividade que poderiam ter majorado o valor do erro de quantização desses vetores de entrada será exibido.

Matrícula	Data de Competência	Vetor de Entrada	Melhor Neurônio	Erro de quantização	Percentil	Justificativa
191.485	Abril/2011	3457	159	35,1213	100,00%	Afastamento
70.700	Abril/2011	368	7	14,5166	99,90%	Cargo em comissão substituição
162.272	Abril/2011	2260	121	14,4398	99,90%	Afastamento Líquido Zero Adto Salarial
46.485	Abril/2011	108	121	12,8443	99,90%	Afastamento Líquido Zero Adto Salarial
214.620	**Abril/2011**	**4028**	**7**	**12,4926**	**99,90%**	**Cargo em comissão substituição**
98.213	Abril/2011	990	35	11,8127	99,90%	Salário menor
224.111	Abril/2011	4578	7	11,4688	99,80%	Cargo em comissão substituição
231.401	Abril/2011	4962	7	11,1916	99,80%	Cargo em comissão substituição
60.283	Abril/2011	228	7	9,8723	99,80%	Cargo em comissão substituição
171.212	Abril/2011	2617	10	9,5429	99,80%	Horas extras em valor acima do

						normal
167.711	Abril/2011	2496	121	9,2905	99,80%	Afastamento Líquido Zero Adto Salarial
154.997	Abril/2011	2015	663	9,1085	99,70%	Complemento de cargo em comissão
103.454	Abril/2011	1202	208	9,0750	99,70%	Horas extras em valor acima do normal Cargo em comissão substituição
181.617	Abril/2011	2863	121	8,7656	99,70%	Afastamento Líquido Zero Adto Salarial
193.968	Abril/2011	3590	208	8,6375	99,70%	Cargo em comissão substituição
98.167	Abril/2011	987	39	8,5871	99,70%	Horas extras em valor acima do normal Complemento de cargo em comissão
218.855	Abril/2011	4276	208	8,0765	99,60%	Horas extras em valor acima do normal Cargo em comissão substituição
221.333	Abril/2011	4418	208	8,0372	99,60%	Cargo em comissão substituição
231.738	Abril/2011	4986	208	8,0027	99,60%	Cargo em comissão substituição
236.187	Abril/2011	5172	1162	7,9997	99,60%	Cargo em comissão substituição
59.315	Abril/2011	216	208	7,9970	99,60%	Cargo em comissão substituição
217.913	Abril/2011	4226	1041	0,2341	0,40%	Normal
41.807	Abril/2011	75	1025	0,2340	0,40%	Normal
234.427	Abril/2011	5108	256	0,2340	0,40%	Normal

213.861	Abril/2011	3984	1002	0,2340	0,40%	Normal
230.642	Abril/2011	4909	1079	0,2339	0,40%	Normal
213.780	Abril/2011	3982	1079	0,2339	0,30%	Normal
184.608	Abril/2011	3049	965	0,2339	0,30%	Normal
186.147	Abril/2011	3163	1025	0,2338	0,30%	Normal
232.866	Abril/2011	4997	1232	0,2337	0,30%	Normal
79.839	Abril/2011	519	1025	0,2337	0,30%	Normal
227.005	Abril/2011	4735	1195	0,2336	0,30%	Normal
188.085	Abril/2011	3260	1118	0,2336	0,20%	Normal
216.828	Abril/2011	4156	1272	0,2336	0,20%	Normal
98.264	Abril/2011	995	1200	0,2336	0,20%	Normal
62.448	Abril/2011	260	1025	0,2335	0,20%	Normal
217.069	Abril/2011	4173	1232	0,2335	0,20%	Normal
192.724	Abril/2011	3527	1276	0,2335	0,10%	Normal
213.284	Abril/2011	3958	1278	0,2335	0,10%	Normal
205.230	Abril/2011	3880	1278	0,2335	0,10%	Normal
76.414	Abril/2011	464	1025	0,2335	0,10%	Normal
96.482	Abril/2011	917	1025	0,2335	0,10%	Normal
216.887	Abril/2011	4160	1194	0,2335	0,00%	Normal
217.077	Abril/2011	4174	676	0,2335	0,00%	Normal
216.917	Abril/2011	4161	1270	0,2335	0,00%	Normal
225.517	Abril/2011	4645	675	0,2331	0,00%	Normal
153.427	Abril/2011	1972	1161	0,2329	0,00%	Normal
160.725	Abril/2011	2164	870	0,2321	0,00%	Normal

Tabela 7 – Ocorrências fora do intervalo de percentil da rede SOM

A linha realçada na Tabela 7 refere-se ao exemplo que estava sendo acompanhado durante os experimentos para comprovar que a análise dos erros de quantização estava detectando possíveis anomalias na base de dados. O funcionário em destaque, matrícula 214.620, recebeu no mês de Abril/2011 valores referentes a um cargo em comissão por motivo de substituição de chefia, este comportamento é uma novidade quando comparado ao histórico do funcionário e era esperado que após o treinamento fosse possível detectá-lo, isto foi exatamente o que aconteceu, confirmando que o método de análise dos erros de quantização da rede SOM pode ser utilizado para detecção de anomalias.

Ao fim da análise das ocorrências fora do intervalo de percentil ficou claro para os especialistas que o limite inferior do intervalo de percentil não era relevante, uma vez que todas as ocorrências levantadas estavam dentro de sua normalidade, sem alterações em relação aos ciclos de folha passados desses funcionários. Entretanto,

como é observado na Tabela 7, o limite superior do intervalo de percentil trouxe bastante informação aos especialistas.

Diante do exposto acerca da não importância das informações trazidas pelo limite inferior do intervalo de percentil, os especialistas decidiram manter o valor α = 0,01, mas alterando o intervalo para [0; 99]. Assim, aplicamos novamente o método estatístico percentil para ordenar o conjunto de erros de quantização dos vetores de entrada do conjunto de validação. O resultado foi de 43 ocorrências acima do limite superior de 99%. Abaixo a Tabela 8 exibe um quadro resumo com os fatores de alteração no comportamento dos 43 vetores de entrada do conjunto de validação.

Principais fatores encontrados para alteração de comportamento	%Ocorrências
Afastamento do funcionário de suas atividades	46
Percepção de valores não recebidos normalmente ou acima do normal	54

Tabela 8 – Quadro Resumo dos fatores responsáveis por desvios no erro de quantização

Os resultados obtidos já indicavam um bom desempenho do método de análise de erros de quantização como classificador de um determinado vetor de entrada como normal ou anomalia, uma vez que todas as 43 ocorrências fora do intervalo de percentil naquele ciclo de folha apontaram para um comportamento diferente dos funcionários investigados. No entanto, é de grande importância para o problema estudado, que pudesse ser verificado que a rede SOM tem a capacidade de aprender com a evolução temporal dos dados do funcionário.

Para tanto, uma nova base foi criada, Base2. O ciclo de folha do conjunto de dados de validação desta base de dados é Maio/2011, logo o conjunto de treinamento engloba o mês de Abril/2011. Esse detalhe é importante por uma razão, o comportamento dos dados de folha de pagamento do funcionário 214.620 é bastante semelhante nas duas bases utilizadas para validação e completamente diferente dos doze meses anteriores que compõem conjunto de dados de treinamento da Base1, porém esse comportamento faz parte do conjunto de dados de treinamento da Base2, uma vez que este conjunto engloba o mês de validação da Base1.

Assim como foi feito com o conjunto de dados da Base1, a rede SOM é treinada com os dados de treinamento e depois a base de validação é apresentada para computação dos erros de quantização de cada vetor de entrada em relação ao seu respectivo melhor neurônio. Computado o conjunto de erros de validação e

estabelecido o intervalo de percentil entre 0% e 99%, [0,2329; 4,0026], verifica-se que o erro de quantização do funcionário 214.620, com 1,0137, fica dentro do intervalo de percentil e é classificado como normal. Isso significa que a rede SOM foi capaz de aprender com os dados apresentados durante a fase de treinamento.

4.4. Análise da Trajetória no mapa SOM

A cadeia de melhores neurônios de vetores de entrada sequenciados no tempo representa a trajetória desses vetores no mapa da rede SOM [36]. A análise de trajetórias SOM se mostra uma poderosa ferramenta de análise individual dentro do espaço de entrada total. É ainda mais interessante quando o mapa SOM possui grupos bem conhecidos, sendo possível não apenas acompanhar a atividade do indivíduo no mapa SOM, mas também predizer quando o mesmo mudará de grupo [10].

Nessa dissertação, devido à impossibilidade de estabelecer grupos bem delimitados, a análise de trajetórias atuará como um método de classificação para definir se o melhor neurônio é uma possível anomalia, esta classificação será realizada com base na distância, em número de vizinhos, entre os melhores neurônios de um funcionário. A Figura 22 ilustra as trajetórias dos melhores neurônios de dois funcionários com comportamentos distintos na base de dados Base3.

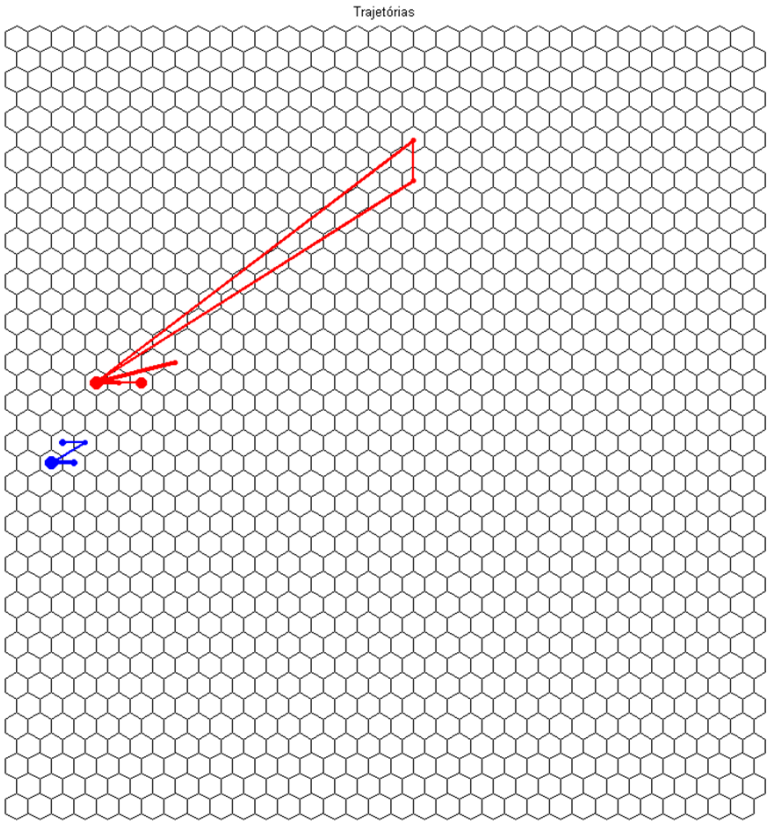


Figura 22 – Trajetórias de dois funcionários com comportamentos distintos.

Pode-se notar que na Figura 23 há uma grande oscilação na vizinhança dos melhores neurônios nos meses de Abril e Maio de 2011, enquanto que na Figura 24 a vizinhança dos melhores neurônios mantém-se praticamente estável durante todo período apresentado.

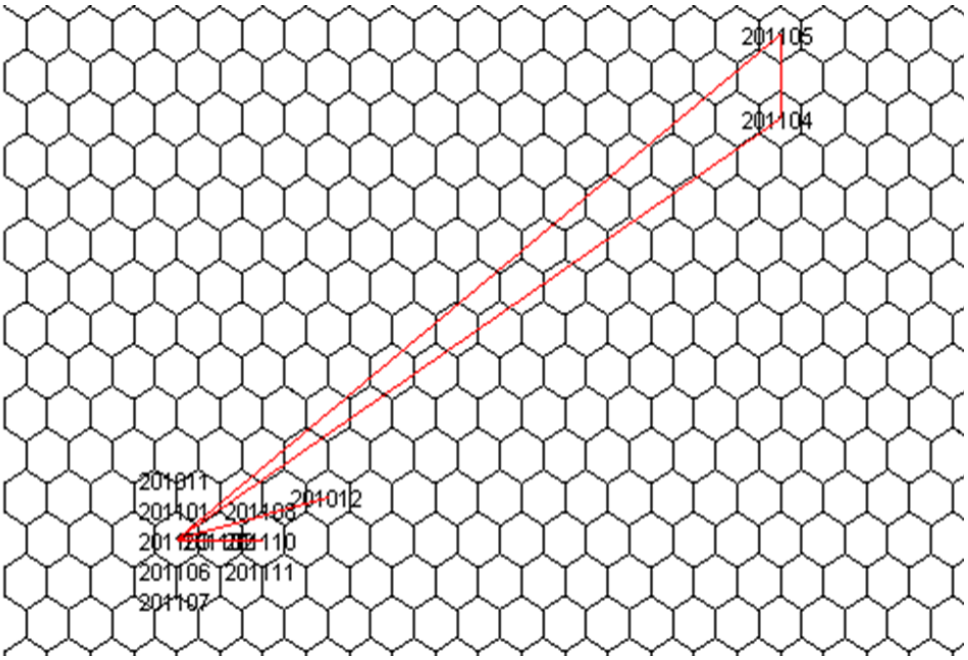


Figura 23 – Trajetória funcionário 01

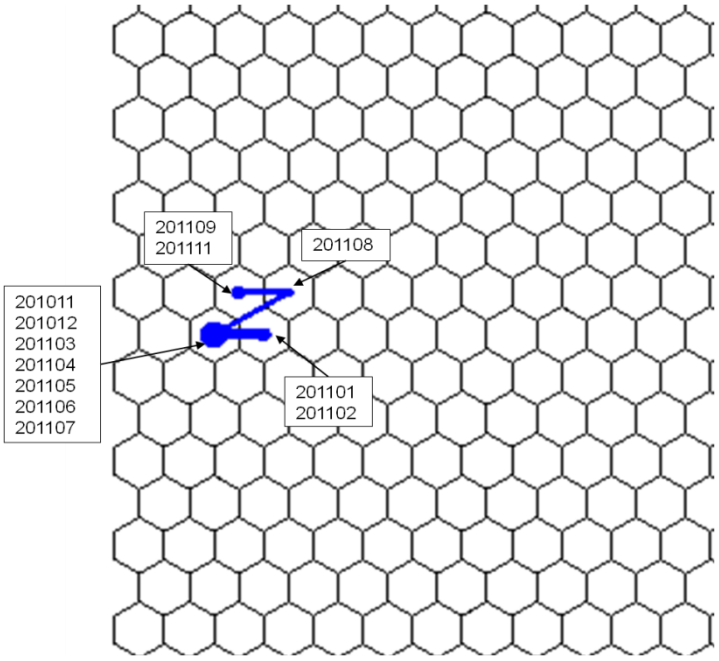


Figura 24 – Trajetória funcionário 02

No entanto, para que a análise de trajetórias SOM fosse efetivamente utilizada, um mecanismo de descoberta semiautomática de possíveis anomalias na base de folha de pagamento foi construído. Com o auxílio da ferramenta SOM *Toolbox* 2.0, foram geradas: a lista dos melhores neurônios, ordenados de forma crescente pela identificação do funcionário e data de competência do ciclo de folha, e a matriz de vizinhança por neurônio da rede. A lista e a matriz foram carregadas em tabelas de banco de dados.

O cruzamento das informações de melhores neurônios e distância entre neurônios via programação permite reconstruir, de forma automática, o quão próximo no espaço de saída estão dois vetores de entrada sequenciais do ponto de vista temporal. Entretanto, para classificar determinado vetor de entrada como anomalia faz-se necessário definir um limiar de vizinhança.

O limiar de vizinhança define a distância aceitável entre neurônios para dois vetores adjacentes no espaço de entrada não serem considerados anomalia no espaço de saída. Residem aqui os seguintes desafios:

- O tamanho ótimo para o limiar. A definição do limiar não pode ser pequena o suficiente para que qualquer variação seja interpretada como anomalia, nem grande o bastante a ponto de ignorar grandes variações de vizinhança.

- A falta de formalismo na definição deste limiar. Apesar de algumas pesquisas contemplarem o uso de trajetórias no mapa SOM como ferramenta para predição e controle, ainda falta um aprofundamento na definição de um limiar fixo.

Nesse experimento os limiares utilizados foram definidos de forma empírica. Os experimentos forma executados com três limiares diferentes para: 3 (três), 5 (cinco) e 7 (sete) neurônios vizinhos. Foi utilizada a Base1 definida na seção 4.1. A Tabela 9 demonstra os resultados obtidos após a fase de descoberta semiautomática das possíveis anomalias.

Limiar de Vizinhança	Possíveis Anomalias
3	3.905
5	2.764
7	1.719

Tabela 9 – Quantidade de possíveis anomalias na Base1 por limiar de vizinhança.

Os experimentos utilizando a análise de trajetória SOM com limiar fixo de vizinhança não obtiveram bons resultados. Vale ressaltar que, dependendo do limiar utilizado, a quantidade de possíveis anomalias chega a 70% do tamanho da base utilizada no experimento para o mês que está sendo validado. Atualmente, os especialistas responsáveis pelo processo de validação dos dados de folha de pagamento estimam que o grau de assertividade do processo esteja em torno de 98%, ou seja, o número de erros encontrados é muito menor do que o indicado pelo experimento.

4.5. Análise da Trajetória versus Análise dos Erros de Quantização

Como visto acima, em um conjunto de dados ordenado no tempo a análise de trajetória em uma rede SOM permite acompanhar a evolução do comportamento dos melhores neurônios em função do tempo, sendo possível visualizar perturbações no comportamento usual de um funcionário. Tais perturbações podem significar anomalias que precisam ser investigadas pelos especialistas. Para realização da análise de trajetória na rede SOM, o conjunto de dados para o treinamento da rede não precisa estar dividido em conjuntos de treinamento e validação, uma vez que o mês alvo para detecção de anomalias precisa ser comparado com o histórico dos melhores neurônios.

A análise dos erros de quantização por sua vez precisa de conjuntos de treinamento e validação bem distintos, a rede deve ser treinada utilizando o conjunto de treinamento e o conjunto de validação – mês alvo – é apresentado à rede já treinada para detecção das possíveis anomalias, diferente da análise de trajetórias a análise dos erros de quantização não considera o histórico de vetores de entrada de um determinado funcionário na classificação daquele padrão como normal ou anormal, a análise de erros de quantização considera unicamente o erro de quantização daquele padrão de entrada em relação à rede já treinada.

Entretanto, a maior diferença entre os tipos de detecção de anomalias aqui comparados está no classificador utilizado por cada uma das abordagens. Enquanto na análise de trajetória SOM é necessário definir um limiar de vizinhança fixo, na análise de erros de quantização é definido um intervalo de percentil onde, os vetores de entrada com erros de quantização dentro desse intervalo são classificados como normais e os vetores de entrada com erros de quantização fora desse intervalo são classificados como anomalias.

A definição de um limiar fixo traz consigo a vantagem de ser um critério objetivo no momento da classificação do vetor de entrada como possível anomalia, porém a escolha do valor do limiar não é nada simples, pois, um limiar pequeno resulta em um conjunto de investigação muito grande com alarmes falsos, inversamente um limiar grande pode resultar em um conjunto de investigação pequeno e pode não estar detectando erros reais da base de dados. Por sua vez, utilizar um intervalo de percentil permite o controle sobre o conjunto de vetores de entrada que será classificado como possível anomalia.

Os experimentos realizados demonstraram que a análise de trajetórias SOM é uma poderosa ferramenta de visualização que permite, inclusive, acompanhar a evolução do comportamento de um funcionário em função do tempo. Entretanto, para o problema de classificador de padrões, apontam uma maior eficácia para análise de erros de quantização da rede SOM. Esse mecanismo dispõe de uma regra simples para classificação de padrões como normais ou possíveis anomalias, além de maior controle do grupo a ser investigado pelos especialistas de negócio.

4.6. Considerações Finais

Os resultados apresentados neste capítulo comprovam que a utilização de um *framework* baseado em mapas auto-organizáveis pode ser uma boa técnica para resolução do problema de detecção de anomalias em bases de folha de pagamento. Os experimentos conduzidos com dados reais conseguiram demonstrar algumas características dos mapas SOM como a distância entre o vetor de entrada e o seu melhor neurônio está relacionada à distribuição dos dados no espaço projetado, a influência de vetores de entrada *outliers* na formação do mapa SOM e a capacidade de aprendizado da rede por exemplos, mesmo em uma base de dados não rotulada.

A análise de erros de quantização juntamente com a definição de um intervalo de percentil utilizando técnicas estatísticas demonstrou-se uma técnica eficaz para classificar vetores de entrada como possíveis anomalias, conseguindo elencar casos que realmente necessitam da análise dos especialistas de negócio. Apesar, de ter se mostrado uma poderosa ferramenta de visualização a análise de trajetórias SOM precisa de um aprofundamento maior para ser utilizada como uma técnica de classificação de padrões para o problema definido nessa dissertação. A quantidade de casos classificados como possíveis anomalias por essa técnica nos experimentos ficou muito maior do que o erro apontado pela experiência dos especialistas.

Capítulo

5

5. Conclusão

Este capítulo tem como objetivo apresentar as considerações finais sobre os principais tópicos abordados nesta dissertação, incluindo as principais contribuições alcançadas e as indicações de trabalhos futuros.

5.1. Considerações Finais

Diversos trabalhos abordando a aplicação de redes neurais artificiais baseadas em mapas auto-organizáveis para detecção de anomalias têm concentrado seus estudos em mapear séries temporais para fins de controle e predição. Devido a algumas características dos mapas auto-organizáveis, eles tem mostrado melhor performance para executar esse tipo de tarefa do que as tradicionais redes neurais artificiais *Multi-Layer Perceptron*. Esses trabalhos, normalmente, utilizam dados rotulados e de baixa dimensionalidade. No entanto existem poucos relatos de trabalhos na indústria que abordem a utilização de mapas auto-organizáveis no mapeamento de dados temporais de alta dimensionalidade e não rotulados com o objetivo da detecção de anomalias.

Esta dissertação propôs a criação de um *framework* para detecção de anomalias em bases de folha de pagamento baseado em mapas auto-organizáveis que visa auxiliar o processo de auditoria e testes de folha de pagamento de uma instituição. O modelo classifica os vetores de entrada de um conjunto de dados como normal ou possível anomalia. Para tanto, é utilizado o conceito de intervalo de percentil como limiar de classificação. Os experimentos comprovaram a eficácia do modelo com dados reais, também foi mostrada a capacidade da rede neural em aprender a partir dos exemplos apresentados e a visualização de trajetórias SOM como uma técnica de acompanhamento dinâmico do comportamento dos melhores neurônios dos vetores de entrada de cada funcionário.

O restante do capítulo está organizado como segue. Na seção 5.2 estão listadas a principais contribuições obtidas pelo trabalho e na seção 5.3 as limitações e os trabalhos futuros que podem ser iniciados a partir das contribuições dessa pesquisa.

5.2. Principais Contribuições

Esta pesquisa oportunizou um trabalho pioneiro na instituição na qual ela foi realizada, pois apesar da imensa quantidade de dados disponíveis nos sistemas administrativos, as informações recuperadas dessas bases utilizam apenas métodos tradicionais como linguagens de consulta a banco de dados. De tal modo que a pesquisa produziu contribuições tanto para o meio acadêmico como para a indústria.

Abaixo estão elencadas as principais contribuições do trabalho na ordem em que o ele foi desenvolvido:

- *Framework* para detecção de anomalias em bases de dados de folha de pagamento

  A definição do de um *framework* genérico que pode ser utilizado também em outras bases financeiras não rotuladas é a principal contribuição dessa dissertação.

- Definição de base de dados consolidada

  Apesar da criação da base e seleção dos dados de um projeto de KDD ser uma atividade muito específica do problema e da natureza dos dados explorados. A base de dados especificada neste trabalho traz consigo o conhecimento de especialistas e pode ser referência para outros trabalhos que tentem resolver o problema de descoberta de conhecimento em bases de folha de pagamento.

- Ferramentas utilizadas

  A revisão dos pacotes SOM *Toolbox* 2.0 e SOMVIS *Toolbox* e a comprovação de que são ferramentas confiáveis servem para que outros trabalhos utilizem as mesmas ferramentas.

- Treinamento da rede

  A modelagem das bases para os experimentos que ordena de forma temporal os dados respeitando sua natureza. Além de prover maior facilidade no momento de investigação dos dados para garantir que são ou não anomalias.

- Utilização da análise de erros de quantização como ferramenta de classificação de vetores de entrada de alta dimensionalidade.

  Os registros encontrados na literatura mostram a aplicação da análise dos erros de quantização para séries temporais ou em bases geradas por modelos matemáticos ambas de baixa dimensionalidade. A aplicação desta técnica em uma base de alta dimensionalidade com bons resultados é uma

contribuição importante para a literatura de redes SOM aplicadas na atividade de auditoria.

- Utilização da análise de trajetórias SOM como ferramenta de mineração de dados.
  A utilização de análise de trajetórias SOM neste trabalho demonstra que esta técnica também é útil quando aplicada em bases de dados não rotuladas na descoberta de conhecimento não supervisionada. Entretanto, sua aplicação é mais efetiva quando os *clusters* do mapa são bem conhecidos, mesmo assim a técnica pode ser bastante útil para visualização de variações nas trajetórias dos melhores neurônios da rede SOM.

5.3. Trabalhos Futuros

Muito ainda precisa ser feito para viabilizar o *framework* como ferramenta de uso operacional para os especialistas de negócio e auditores, bem como para consolidar o uso da análise de trajetórias como um instrumento de predição do comportamento da folha de pagamento dos funcionários. Ficam aqui sugestões de pesquisa e desenvolvimento para continuidade da pesquisa que resultou nesta dissertação:

- Implementação de uma interface amigável ao usuário para realizar os testes de folha de pagamento.
- Realizar *clustering* da rede SOM a fim de investigar melhor o uso da análise de trajetórias SOM.
- Investigar melhor o problema de limiar fixo de vizinhança na análise de trajetórias SOM.

Referências

[1] E. Koskivaara, Artificial Neural Networks in Auditing : State of the Art, *The ICFAI Journal of Audit Practice*, vol. 1, pp. 12–33, 2004.

[2] R. Fragoso, Indicadores para Auditoria Contínua, *CLAI - Congresso Latino Americano de Auditoria Interna*, 2010.

[3] E. Koskivaara, Artificial neural networks in analytical review procedures, *Managerial Auditing Journal*, vol. 19, no. 2, pp. 191–223, 2004.

[4] E. Koskivaara, Artificial Neural Networks for Analytical Review in Auditing, University of Turku, 2004.

[5] K. Omoteso, The application of artificial intelligence in auditing: Looking back to the future, *Expert Systems with Applications*, vol. 39, no. 9, pp. 8490–8495, 2012.

[6] H. C. Koh, C. K. Low, Going concern prediction using data mining techniques, *Managerial Auditing Journal*, vol. 19, no. 3, pp. 462–476, 2004.

[7] A. L. I. Oliveira, G. Azevedo, A. Barros, and A. L. M. Santos, Sistema de Suporte à Auditoria de Folhas de Pagamento Baseado em Redes Neurais, *IV ENIA - Encontro Nacional de Inteligência Artificial*, 2003, pp. 487–496.

[8] P. Adriaans, D. Zantinge, *Data Mining*, 1a. ed. 1996, pp. 1–158.

[9] T. Kohonen, The self-organizing map, *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[10] N. Chen, B. Ribeiro, A. Vieira, A. Chen, Clustering and visualization of bankruptcy trajectory using self-organizing map, *Expert Systems With Applications*, vol. 40, no. 1, pp. 385–393, 2013.

[11] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, J. Han, On community outliers and their efficient detection in information networks, *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’10*, p. 813, 2010.

[12] A. Muñoz, J. Muruzábal, Self-organizing maps for outlier detection, *Neurocomputing*, vol. 18, no. 1–3, pp. 33–60, Jan. 1998.

[13] G. a Barreto, J. C. M. Mota, L. G. M. Souza, R. a Frota, L. Aguayo, Condition monitoring of 3G cellular networks through competitive neural models, *IEEE transactions on neural networks*, vol. 16, no. 5, pp. 1064–1075, 2005.

[14] L. Aguayo, G. a. Barreto, Novelty Detection in Time Series Through Self-Organizing Networks: An Empirical Evaluation of Two Different Paradigms, *10th Brazilian Symposium on Neural Networks*, pp. 129–134, 2008.

[15] U. Fayyad, G. Piatetsky-shapiro, P. Smyth, From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, vol. 17, pp. 37–54, 1996.

[16] J. de A. Loureiro, Técnicas de Agrupamento de Dados na Mineração de Dados Químicos, Unversidade Federal de Pernambuco - UFPE, 2005.

[17] A. C. D. A. Simões, Mineração de Dados baseada em Árvores de Decisão para Análise do Perfil de Contribuintes, Universidade Federal de Pernambuco - UFPE, 2008.

[18] C. Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[19] S. Haykin, *Neural networks*, 2a. ed., vol. 8, no. 3. Hamilton: Prentice Hall International, Inc., pp. 1–897, 1999.

[20] E. J. Spinosa, Detecção de novidade com aplicação a fluxos contínuos de dados, USP, 2008.

[21] H. U. Bauer, K. R. Pawelzik, Quantifying the neighborhood preservation of self-organizing feature maps, *IEEE transactions on neural networks*, vol. 3, no. 4, pp. 570–579, 1992.

[22] J. Venna, S. Kaski, Neighborhood Preservation in Nonlinear Projection Methods: An Experimental Study, *International Conference on Artificial Neural Networks — ICANN 2001*, vol. 2130, pp. 485–491, 2001.

[23] K. Kiviluoto, Topology preservation in self-organizing maps, *Proceedings of International Conference on Neural Networks (ICNN'96)*, vol. 1, pp. 294–299, 1996.

[24] J. M. L. Gorricha, Visualization of Clusters in Geo-referenced Data Using Three-dimensional Self-Organizing Maps by Visualization of Clusters in Geo-referenced Data Using Three-dimensional Self-Organizing Maps, 2010.

[25] R. J. Kuo, L. M. Ho, C. M. Hu, Integration of self-organizing feature map and K-means algorithm for market segmentation, *Computers & Operations Research*, vol. 29, no. 11, pp. 1475–1493, 2002.

[26] T. Kohonen, Analysis of processes and large data sets by a self-organizing method, *Proceedings of the Second International Conference on Intelligent Processing and Manufacturing of Materials. IPMM'99*, vol. 1, pp. 27–36, 1999.

[27] C. Ding, J. C. Patra, User Modeling for Personalized Web Search With Self-Organizing Map, *Journal of the American Society for Information Science and Technology*, vol. 58, no. 4, pp. 494–507, 2007.

[28] G. Craessaerts, J. De Baerdemaeker, W. Saeys, Fault diagnostic systems for agricultural machinery, *Biosystems Engineering*, vol. 106, no. 1, pp. 26–36, 2010.

[29] E. A. Uriarte, F. D. Martín, Topology Preservation in SOM, *International Journal of Applied Mathematics and Computer Sciences*, vol. 1, no. 1, pp. 19–22, 2005.

[30] G. Cabanes, Y. Bennani, Learning topological constraints in Self-Organizing Map, *ICONIP'10 Proceedings of the 17th international conference on Neural information processing: models and applications*, pp. 367–374, 2010.

[31] G. Pölzlbauer, Survey and Comparison of Quality Measures for Self-Organizing Maps, *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*, pp. 67–82, 2004.

[32] G. Spanakis, G. Siolas, A. Stafylopatis, DoSO: a document self-organizer, *Journal of Intelligent Information Systems*, vol. 39, no. 3, pp. 577–610, 2012.

[33] G. J. Goodhill, S. Finch, T. J. Sejnowski, Quantifying neighbourhood preservation in topographic mappings, pp. 1–17, 1996.

[34] T. Villmann, R. Der, M. Herrmann, T. M. Martinetz, Topology preservation in self-organizing feature maps: exact definition and measurement, *IEEE transactions on neural networks*, vol. 8, no. 2, pp. 256–266, 1997.

[35] H.-U. Bauer, M. Herrmann, T. Villmann, Neural maps and topographic vector quantization, *Neural networks*, vol. 12, no. 4–5, pp. 659–676, 1999.

[36] M. Varsta, J. Heikkonen, J. Lampinen, J. D. R. Millán, Temporal Kohonen Map and the Recurrent Self-Organizing Map: Analytical and Experimental Comparison, *Neural Processing Letters*, vol. 13, no. 3, pp. 237–251, 2001.

[37] G. J. Chappell, J. G. Taylor, The Temporal Kohonen Map, *Neural Networks*, vol. 6, no. 3, pp. 441–445, 1993.

[38] M. Shahnawaz, A. Ranjan, M. Danish, Temporal Data Mining : An Overview, *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 1, no. 1, pp. 20–24, 2011.

[39] L. F. Gonçalves, E. L. Schneider, V. B. Henriques, M. Lubaszewski, Fault Prediction in Electrical Valves Using Temporal Kohonen Maps, *11th Latin American Test Workshop (LATW),* pp. 1–6, 2010.

[40] P. du Jardin, E. Séverin, Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model, *Decision Support Systems*, vol. 51, no. 3, pp. 701–711, 2011.

[41] E. Koskivaara, Artificial Neural Network Models for Predicting Patterns in Auditing Monthly Balances, *Journal of Operational Research Society*, vol. 51, no. 67, pp. 1060–1069, 1996.

[42] W. Zhou, G. Kapoor, Detecting evolutionary financial statement fraud, *Decision Support Systems*, vol. 50, no. 3, pp. 570–575, 2011.

[43] P. Castellón González, J. D. Velásquez, Characterization and detection of taxpayers with false invoices using data mining techniques, *Expert Systems with Applications*, vol. 40, no. 5, pp. 1427–1436, 2013.

[44] K. Kiviluoto, P. Bergius, Exploring corporate bankruptcy with two-level self-organizing maps, *Fifth International Conference on Computational Finance*, pp. 373–380, 1998.

[45] S. Kaski, J. Sinkkonen, J. Peltonen, Bankruptcy Analysis with Self-Organizing Maps in Learning Metrics, *IEEE Transactions on Neural Networks*, vol. 12, pp. 936–947, 2001.

[46] J. Vesanto, Self-Organizing Map for Data Mining in MATLAB : the SOM Toolbox, p. 1, 1999.

[47] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, Self-organizing map in Matlab : the SOM Toolbox, *Proceedings of the Matlab DSP Conference*, pp. 35–40, 1999.

[48] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, SOM Toolbox for Matlab 5, pp. 1–60, 2000.

[49] J. Vesanto, M. Siponen, O. Simula, P. Vasara, An approach to automated interpretation of SOM, *Advances in Self-Organizing Maps*, pp. 89–94, 2001.

[50] J. Vesanto, Using the SOM and Local Models in Time-Series Prediction, no. 1, 1997.

[51] E. Pampalk, A. Rauber, D. Merkl, Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps, *International Conference on Artifical Neural Networks (ICANN'02)*, pp. 871 – 876, 2002.

[52] T. Kohonen, Essentials of the self-organizing map, *Neural Networks*, vol. 37, pp. 52–65, 2013.

[53] V. J. Hodge, J. I. M. Austin, A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, no. 1969, pp. 85–126, 2004.

[54] Dicionário Online de Português. Disponível em: http://www.dicio.com.br/rubrica_2/, último acesso em Março 2013.