Programa Interdisciplinar de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Glauber Cini

VISUAL ANALYTICS COMO FERRAMENTA DE AUXÍLIO AO PROCESSO DE KK: Um estudo voltado ao pré-processamento

São Leopoldo, 2017

Glauber Cini

VISUAL ANALYTICS COMO FERRAMENTA DE AUXÍLIO AO PROCESSO DE KDD:

Um estudo voltado ao pré-processamento

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre, pelo Programa Interdisciplinar de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos – UNISINOS

Orientador: Prof. Dr. João F. Valiati

São Leopoldo

2017

C575v Cini, Glauber.
Visual Analytics como ferramenta de auxílio ao processo de KDD : um estudo voltado ao pré-processamento / Glauber Cini. – 2017.
69 f. : il. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa Interdisciplinar de Pós-Graduação em Computação Aplicada, 2017.
"Orientador: Prof. Dr. João F. Valiati."

1. Visual Analytics. 2. KDD. 3. Visualização de informação. 4. Coordenadas paralelas. 5. WEKA. I. Título.

CDU 004

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecário: Flávio Nunes – CRB 10/1298)

Glauber Cini

Visual Analytics como ferramenta de auxílio ao processo de KDD : um estudo voltado ao pré-processamento

Dissertação apresentada à Universidade do Vale do Rio dos Sinos – Unisinos, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Aprovado em 29 de março de 2017

BANCA EXAMINADORA

Profª. Drª. Carine Geltrudes Webber – Universidade de Caxias do Sul – UCS

Profª. Drª. Marta Becker Villamil – Universidade do Vale do Rio dos Sinos – UNISINOS

Prof. Dr. João Francisco Valiati

Visto e permitida a impressão
São Leopoldo,

Prof. Dr. Sandro José Rigo
Coordenador PPG em Computação Aplicada

“Deves aprender as regras do jogo. E depois deves jogar melhor que todo mundo”.

(Albert Einstein)

AGRADECIMENTOS

Gostaria de agradecer a todos que se envolveram das mais diversas formas, contribuindo para a realização do presente trabalho.

A minha família pelo incentivo ao estudo durante todas as etapas da minha vida, em especial minha irmã Gláucia pelo alicerce e exemplo que ela representa. A Natália, minha namorada, agradeço pelo apoio, sendo uma das principais responsáveis pelo início desta jornada.

Agradeço também ao professor e orientador Dr. João F. Valiati, pela dedicação, conhecimento, tempo e orientação.

Por fim, agradeço a todos os demais colegas e docentes do PIPCA.

RESUMO

O *Visual Analytics* consiste na combinação de métodos inteligentes e automáticos com a capacidade de percepção visual do ser humano visando a extração do conhecimento de conjuntos de dados. Esta capacidade visual é apoiada por interfaces interativas como, sendo a de maior importância para este trabalho, a visualização por Coordenadas Paralelas. Todavia, ferramentas que disponham de ambos os métodos automáticos (KDD) e visuais (Coordenadas Paralelas) de forma genérica e integrada mostra-se primordial. Deste modo, este trabalho apresenta um modelo integrado entre o processo de KDD e o de Visualização de Informação utilizando as Coordenadas Paralelas com ênfase no *make sense of data*, ao ampliar a possibilidade de exploração dos dados ainda na etapa de pré-processamento. Para demonstrar o funcionamento deste modelo, um *plugin* foi desenvolvido sobre a ferramenta WEKA. Este módulo é responsável por ampliar as possibilidades de utilização da ferramenta escolhida ao expandir suas funcionalidades a ponto de conceitua-la como uma ferramenta *Visual Analytics*. Junto a visualização de Coordenadas Paralelas disponibilizada, também se viabiliza a interação por permutação das dimensões (eixos), interação por seleção de amostras (*brushing*) e possibilidade de detalhamento das mesmas na própria visualização.

Palavras-Chave: *Visual Analytics*. KDD. Visualização de Informação. Coordenadas Paralelas. WEKA.

ABSTRACT

Visual Analytics is the combination of intelligent and automatic methods with the ability of human visual perception aiming to extract knowledge from data sets. This visual capability is supported by interactive interfaces, considering the most important for this work, the Parallel Coordinates visualization. However, tools that have both automatic methods (KDD) and visual (Parallel Coordinates) in a generic and integrated way is inherent. Thus, this work presents an integrated model between the KDD process and the Information Visualization using the Parallel Coordinates with emphasis on the make sense of data, by increasing the possibility of data exploration in the preprocessing stage. To demonstrate the operation of this model, a plugin was developed on the WEKA tool. This module is responsible for expanding the possibilities of chosen tool by expanding its functionality to the point of conceptualizing it as a Visual Analytics tool. In addition to the delivered visualization of Parallel Coordinate, it is also possible to interact by permutation of the dimensions (axes), interaction by selection of samples (brushing) and possibility of detailing them in the visualization itself.

Keywords: Visual Analytics. KDD. Information Visualization. Parallel Coordinates. WEKA.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

2D	Duas dimensões
3D	Três dimensões
CLM4	*Community Land Model Version 4*
DM	*Data Mining*
EDEN	*Exploratory Data analysis Environment*
KDD	*Knowledge Discovery in Databases*
LES	*Large-Eddy Simulation*
ParCAT	*Parallel Climate Analysis Tools*
PCP	*Principal Component Analysis*
ROC	*Receiver Operating Characteristics*
SVM	*Support Vector Machines*
USA	*United States of America*
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Técnicas modernas de *Visual Analytics*, em sua conceituação geral, valem-se da integração de duas principais áreas da descoberta de conhecimento: Visualização de Informação e Descoberta de Conhecimento em Bancos de Dados (KDD, do inglês *Knowledge Discovery in Databases*) (KEIM; THOMAS, 2006), (THOMAS; COOK, 2005), (KEIM et al., 2008). A Visualização de Informação representa a exploração dinâmica de dados, interativamente, construindo uma compreensão visual entre os dados que estão sendo investigados e os analistas. Ela pode ajudar a descobrir, visualmente, diferentes tipos de padrões, tais como clusters, relacionamentos e associações, permitindo assim uma exploração visual de dados (KEIM, 2002), (WEGMAN, 2003). Por outro lado, o KDD consiste em vários algoritmos que pertencem tanto à área de mineração de dados quanto à estatística, onde técnicas algorítmicas extraem, das fontes de dados informações significativas, sob a forma de padrões, com o intuito de minimizar a lacuna entre o registro e a compreensão dos dados capturados.

No entanto, mais importante do que a definição do conceito de *Visual Analytics*, é abordar assertivamente as etapas da descoberta de conhecimento em que as áreas podem colaborar entre si de forma eficiente. Uma das etapas mais importantes do KDD é o pré-processamento, já conhecido por requerer mais esforço e tempo para ser realizado (GARCÍA; LUENGO; HERRERA, 2016), (ADRIANNS; ZATINGE, 1996), apesar de ser, também, o momento em que os analistas, de fato, aprenderão mais sobre a natureza dos dados. Aplicar algoritmos como normalização, discretização, reamostragem, detecção de valores atípicos (discrepâncias, *outliers*), entre outros, pode ser necessário para possibilitar a melhor realização do KDD, mas algoritmos e parâmetros podem limitar o real entendimento e as relações intrínsecas que os dados podem oferecer, sinalizando que o desempenho dos algoritmos pode ser estendido. Neste ponto, todos os benefícios da visualização podem ser utilizados para entender como o conjunto de dados está relacionado (*make sense of data*), quais amostras podem ser *outliers*, se há algum tipo de tendência, agrupamentos ou associações (HEINRICH; WEISKOPF, 2013).

Dentre as várias formas de Visualização de Informação encontradas na literatura (FAYYAD; WIERSE; GRINSTEIN, 2002), as Coordenadas Paralelas (INSELBERG, 1985) se destacam por permitir uma visualização adequada para dados multidimensionais (ZHOU et al., 2008), (HEINRICH; WEISKOPF, 2013), uma característica inerente ao KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

As Coordenadas Paralelas empregadas como um meio auxiliar ao pré-processamento de dados podem amplificar o sentido dos dados que estão sendo analisados. No entanto, a falta de um processo definido e a dificuldade de se encontrar ferramentas que possam trabalhar com ambas tecnologias em conjunto, dificultam o aprofundamento de estudos na área, uma vez que os estudos ocorrem de forma independente (SACHA et al., 2014).

Fayyad, Wierse e Grinstein (2002) reportam um extenso estudo das áreas de KDD e Visualização de Informação de forma paralela, de um lado há o KDD que utiliza algoritmos para auxiliar na compreensão do conjunto de dados por meio de várias ferramentas já

consolidadas, podendo-se citar: WEKA[1], RapidMiner[2], KNIME[3], Orange[4], entre outras. Porém não disponibiliza nenhuma ferramenta gráfica interativa para que os dados sejam relacionados e explorados. Do outro lado há a área de Visualização de Informação, que se apoia sobre algoritmos que melhoram a eficiência ao desenhar-se uma visualização sem tratar ou se preocupar com o conjunto de dados de entrada, quando este cenário poderia ser melhorado utilizando etapas do KDD, como por exemplo o pré-processamento e a mineração de dados.

Para apresentar os benefícios do *Visual Analytics* no pré-processamento, o presente trabalho estende sua abordagem ao expandir as funcionalidades de uma ferramenta KDD, adicionando a visualização de Coordenadas Paralelas. Consolidando, em uma concepção genérica, uma ferramenta de *Visual Analytics*, ao incorporar tais funções sobre uma ferramenta de mineração de dados. Não obstante, a aplicação resultante sugere um processo genérico de *Visual Analytics* anexo a uma ferramenta amplamente conhecida, com a finalidade de aumentar a popularidade da área e construir uma maneira mais fácil de evoluir, expandir e estudar o *Visual Analytics*.

A intenção deste trabalho não é criar uma ferramenta universal para este tipo de análise, mas sim adicionar a uma existente as premissas requeridas por tal área. Assim, tornando-se um caminho popular para outros estudos que derivem deste primeiro passo, com foco principalmente na didática e na transformação da ideia de tornar o *make sense of data* mais possível no pré-processamento do KDD.

1.1 Objetivo

O objetivo deste trabalho é investigar, consolidar e demonstrar a integração do processo de Descoberta do Conhecimento em Bases de Dados (KDD) com a Visualização de Informação de uma forma genérica, no sentido de permitir o uso efetivo do método de visualização, neste estudo representado pelas Coordenadas Paralelas, incorporado especificamente à etapa de pré-processamento, de forma a enriquecer o KDD com o emprego do *Visual Analytics*.

Tem-se ainda, como objetivos específicos:

- Expandir o uso da Visualização da Informação, com a técnica de Coordenadas Paralelas, ao processo de KDD, deste modo criando um ponto de partida comum para o estudo da análise visual;
- Efetivar o estudo de caso na ferramenta WEKA, com a incorporação do módulo de visualização interativa em conjunto com as funcionalidades de mineração de dados já existentes;

[1] WEKA: *Software* escrito em Java que contém uma coleção de algoritmos de aprendizado de máquina utilizados para tarefas de mineração de dados, encontrado em: http://www.cs.waikato.ac.nz/ml/weka/.

[2] RapidMiner: Ambiente integrado para a mineração de dados, possuindo limitações em sua versão gratuita. (https://rapidminer.com/).

[3] KNIME: Plataforma que integra diversos componentes do aprendizado de máquina e da mineração de dados através de um conceito modular de dados. (https://www.knime.org/).

[4] Orange: Ferramenta de visualização, mineração e análise de dados que utiliza programação visual para a formulação de fluxogramas que processam os dados. (http://orange.biolab.si/).

- Testar e avaliar o módulo desenvolvido sobre bases de dados clássicas para demonstrar efetividade e limitações de funcionalidades desenvolvidas;
- Disponibilizar o módulo desenvolvido em uma versão beta e disponibilizado em um repositório on-line para amplo uso da comunidade científica.

Não faz parte do escopo do presente trabalho, possivelmente recomendado como uma extensão desse trabalho, a análise de usabilidade. É importante salientar que o resultado desse trabalho não visa medir a assertividade da mineração de dados, mas permitir mecanismos para auxiliar no processo de exploração inicial dos dados referente a domínios de interesse.

1.2 Estrutura

O Capítulo seguinte apresenta os fundamentos teóricos referente ao *Visual Analytics* citando sua formulação a partir dos processos de KDD a Visualização de Informação. No Capítulo 3 são apresentados os trabalhos relacionados ao *Visual Analytics*, Coordenadas Paralelas e integração de técnicas de visualização a técnicas automáticas de extração de conhecimento. O processo de *Visual Analytics* no pré-processamento é apresentado no Capítulo 4. Na sequência, o Capítulo 5 demonstra a aplicabilidade do *plugin* proposto. Por último, no Capítulo 6 encontra-se a Conclusão.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os principais conceitos abordados neste trabalho. A seção 2.1 apresenta o processo de *Visual Analytics*, que busca traçar um paralelo entre a formulação clássica do processo de KDD junto ao processo de Visualização de Informação. A seção seguinte aborda a formulação, importância e análise das Coordenadas Paralelas, técnica de visualização amplamente utilizada em processos de extração de conhecimento de forma visual. Na seção 2.3 são apresentadas as considerações.

2.1 *Visual Analytics*

O raciocínio analítico, sobretudo utilizando KDD e apoiado pela Visualização de Informação, constitui o que se chama de Análise Visual ou *Visual Analytics*. O conceito de *Visual Analytics* surgiu primeiramente focando na segurança interna dos Estados Unidos da América. Porém, ganhou um contexto mais amplo descrevendo um campo multidisciplinar que combina várias áreas investigativas, incluindo a Visualização de Informação, Interação Humano-Computador, análise de dados e gerenciamento de dados geo-espaciais, temporais e de processamento estatístico (KEIM; THOMAS, 2006).

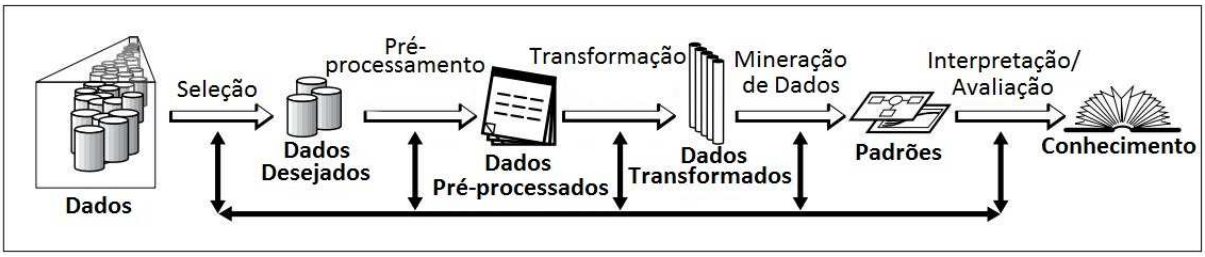
Este conceito cunhado por Thomas e Cook (2005) propõe que a subjetividade, conhecimento e capacidade de reconhecer padrões – habilidades intrínsecas do ser humano – devam ser utilizadas em conjunto com métodos automáticos para análise de dados. Além da capacidade humana, amplamente reconhecida por Ware (2012), Keim et al. (2008) sugerem que o grande volume de dados existentes – favorecidos por um crescimento na produção, coleta e armazenamento dos dados – também pode-se beneficiar da Análise Visual, tanto pelos benefícios cognitivos já citados como também auxiliando processos automáticos que apresentarem ineficiências ou falhas.

Desenvolvidas independentemente, as técnicas automáticas de análise, como as contidas nas etapas do KDD, e as visuais, contidas na Visualização de Informação, contextualizaram importantes discussões (KEIM et al., 2008). Estas discussões levaram a uma mudança do escopo, até então bastante limitado de ambos os campos, para o que hoje é chamado de pesquisa de Análise Visual. Um dos passos mais importantes neste sentido foi a necessidade de se mover a análise confirmatória de dados (resultados em formato de relatório) para a análise exploratória de dados (resultados interativos), o que foi afirmado pela primeira vez na comunidade de pesquisas em estatísticas por Tukey (1977).

Keim et al. (2008) citam que, mais tarde, com a disponibilidade de interfaces gráficas avançadas e com dispositivos de interação adequados, uma comunidade de pesquisa inteira dedicou seus esforços à Visualização de Informação, cita-se como exemplo: Card, Mackinlay e Shneiderman (1999), Chen (2006), Robert (2006) e Ware (2012). Em algum momento, esta comunidade reconheceu o potencial de integrar o usuário no processo de KDD através de técnicas de visualização. Esta integração expandiu consideravelmente o âmbito de aplicações tanto na Visualização de Informação quanto nos campos de KDD, resultando em novas técnicas, além de inúmeras e interessantes novas oportunidades de investigação.

Diante da proposta de um processo unificado, faz-se necessário abordar a origem da formulação do chamado *Visual Analytics*. Este processo originou-se ao traçar um paralelo e conduzir uma integração entre duas áreas com o objetivo de extrair conhecimentos em conjuntos de dados. De um lado, há o processo de KDD proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996) e exemplificado na figura abaixo.

Figura 1: Processo clássico de extração do conhecimento.



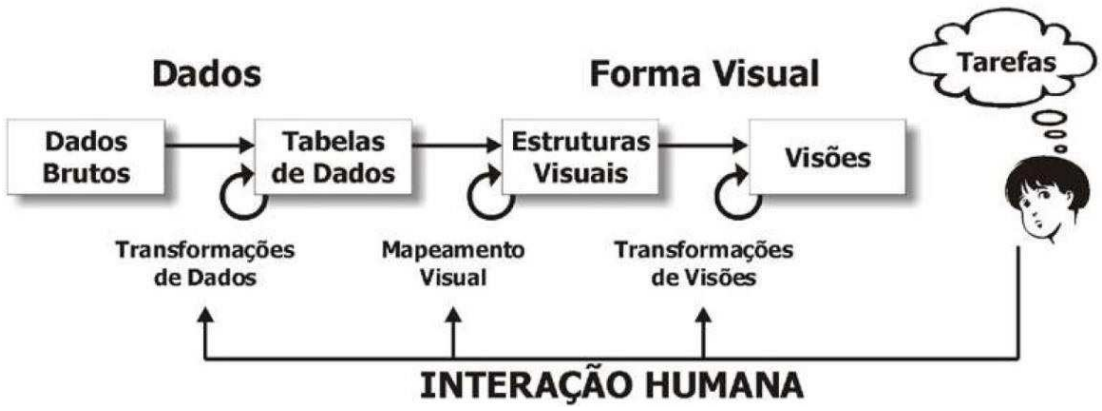
Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

O processo clássico é caracterizado pelas suas cinco etapas. A primeira refere-se à Seleção, onde dados de diversas origens são consultados e avaliados de acordo com a sua contribuição. Já o pré-processamento é considerada a etapa de maior esforço (GARCÍA; LUENGO; HERRERA, 2016), por exemplo, Adrianns e Zatinge (1996) estimavam que do esforço investido em todo o KDD, algo entre 60% e 80%, era demandado pelo pré-processamento, dentre suas principais atribuições incluem-se: a limpeza e seleção dos dados, transformações de forma geral e ampliação do conhecimento do domínio da informação. A Transformação, por sua vez, tem a função de preparar os dados para um formato mais adequado como, por exemplo, utilizando agregações ou sumarizações. Na etapa seguinte, chamada de mineração de dados, verifica-se a busca propriamente dita pelo conhecimento e identificação de padrões. Por fim, tem-se a etapa de Interpretação (ou Avaliação) onde valida-se a importância do conhecimento adquirido.

Este processo apresenta um caminho aprofundado na preparação dos dados, observa-se que (na Figura 1) das cinco etapas existentes, três – Seleção, Pré-processamento e Transformação – são pertinentes à consolidação de conjuntos de dados, onde os mesmos são utilizados na etapa seguinte, a mineração de dados, posto que esta estrutura-se sobre modelos e parâmetros pertinentes a sua aplicação.

Do outro lado há o processo Visualização de Informação proposto por Card et al. (1999) conforme visualizado abaixo.

Figura 2: Processo de Visualização de Informação.



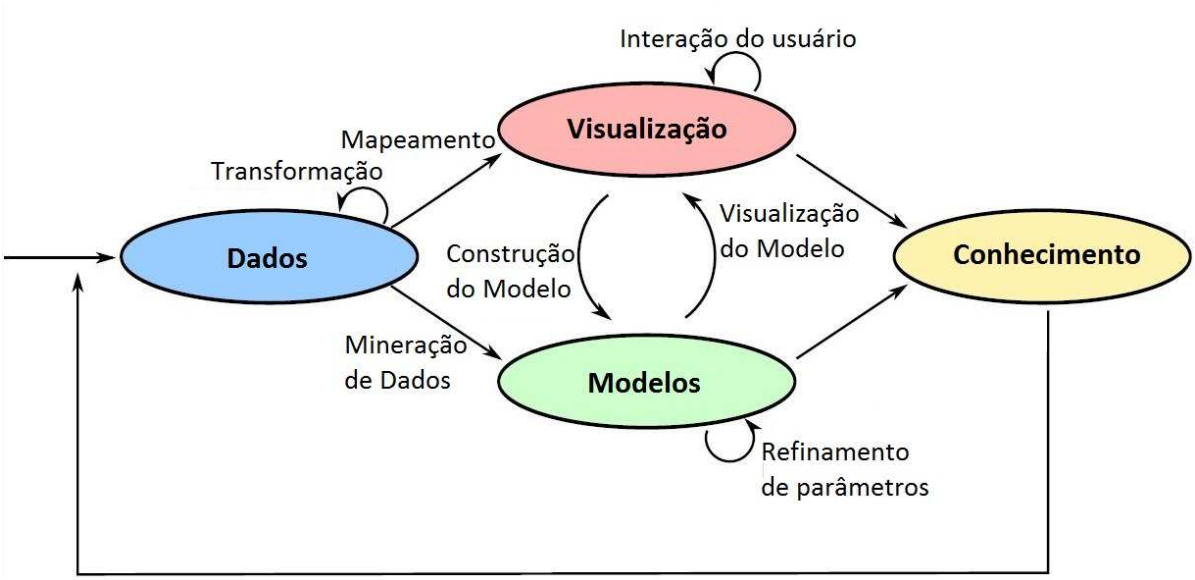
Fonte: Adaptado de Card et al. (1999).

Este processo é, essencialmente, composto por quatro etapas. A primeira é a Transformação de Dados, responsável por compilar diversas fontes brutas de dados em uma estrutura homogênea. Após ter os dados organizados, a etapa de Mapeamento Visual visa traduzir as amostras de dados em artefatos visuais. Diante destes artefatos, a etapa de Transformações de Visões busca transformar, de forma dinâmica, a interface gerada, possibilitando assim a interatividade. Ressalta-se também o momento chamado “Tarefas”, que pode ser lido como: “as diversas ações realizadas por um ser humano no processo de Visualização de Informação ao buscar o conhecimento”.

Desta forma, o conceito contido na Figura 2 apresenta-se semelhante ao modelo de preparação de dados sugerido por Fayyad, Piatetsky-Shapiro e Smyth (1996), sendo que Card et al. (1999) descrevem as etapas de forma análoga e unificada, visto que as Tabelas de Dados (Figura 2) são responsáveis por armazenar dados previamente processados, sendo por sumarizações, agrupamentos ou discretizações, assim gerando um modelo dos dados capaz de ser representado por técnicas visuais.

A fim de representar o processo unificado, Keim et al. (2008) sugerem um modelo integrado de ambos os campos: KDD (Figura 1) e Visualização de Informação (Figura 2). Este modelo está contido na Figura 3, onde observa-se os diferentes estágios (Dados, Visualização, Modelos e Conhecimento) e suas transições (flechas) no processo de *Visual Analytics*.

Figura 3: Processo de *Visual Analytics*.



Fonte: Adaptado de Keim et al. (2008).

Nesta figura pode-se concluir que Keim et al. (2008) agruparam os métodos de preparação de dados contidos nos processos de KDD e Visualização de Informação no estágio chamado Dados, sendo que o resultado deste estágio tem como objetivo a formulação de um conjunto de dados consolidado.

Os estágios subsequentes, munidos destes conjuntos, são responsáveis pela apresentação da informação a fim de prover padrões que, após analisados devem apresentar, cada qual em seu formato, o conhecimento. Primeiramente cita-se o estágio chamado de Modelos, pode-se associar este estágio aos modelos construídos na mineração de dados junto de seus algoritmos e parâmetros utilizados. O segundo, chamado de Visualização, é responsável por apresentar a informação de forma gráfica e interativa.

Vale-se ressaltar que, para a formulação do estágio de Visualização, Keim et al. (2008) citam a transição chamada de Mapeamento, que pode ser lida como Mapeamento Visual, abordado na subseção 2.2.1.

Depois de mapear os dados para uma visualização, é possível obter-se o conhecimento desejado diretamente, mas o mais provável é que somente uma primeira visualização não seja suficiente para que se efetue uma análise detalhada.

Neste sentido, a interação do usuário com a visualização vem com o objetivo de revelar informações de forma mais minuciosa. Esta interação é descrita por Shneiderman (1996) como o *Information Seeking Mantra*, que sugere que toda a visualização ofereça os três artefatos citados abaixo:

- ***Overview first***: a visualização deve ser analisada como um todo, sendo que desta perspectiva pode-se perceber possíveis padrões existentes ou sugerir, de modo geral, como os atributos estão relacionados no conjunto de dados (SHNEIDERMAN, 1996);
- ***Zoom* e *filter***: ambos envolvem a redução da complexidade dos dados apresentados, removendo informações supérfluas da visualização e permitindo uma nova organização dos dados. Diferente de seu conceito genérico, que limita-se a dizer que uma forma gráfica pode ser aproximada ou distanciada, o conceito de *zoom* proposto por Shneiderman (1996) apoia-se na cognição humana, ressaltando dois importantes aspectos, o primeiro é a capacidade de atentar-se a um ponto específico e analisá-lo de forma individual (*zooming-in*), o segundo diz respeito a habilidade de, munido da informação adquirida no *zooming-in*, se distânciar deste ponto específico da análise e ser capaz de inserir este conhecimento nas demais informações periféricas (CRAFT; CAIRNS, 2005). O *filter*, no que lhe concerne, busca reduzir a complexidade dos dados, porém sem modificar sua representação no ponto de vista do usuário. Para que isso seja possível, a ferramenta que disponibiliza o *filter* deve fornecer meios para que determinadas amostras de dados sejam ressaltadas ou ocultadas (SHNEIDERMAN, 1996), (KEIM, 2002) e (CRAFT; CAIRNS, 2005).
- ***Details-on-demand***: tipicamente na Visualização de Informação, inúmeras amostras de dados são plotadas sobre uma visualização, esta caracterísca que, muitas vezes, salta de dezenas de amostras para milhões, pode enfrentar dificuldades na sua representação, que consistem em visualizações sobrecarregadas e limitadas pelo tamanho dos monitores de computadores. Para que possa-se encontrar novas informações em situações como essa, busca-se adicionar informações sobre demanda, ou seja, em eventos provenientes do computador, como por exemplo, clicar ou passar com o *mouse* sobre a amostra, esta deve exibir detalhes (como sua representação numérico/nominal) (CRAFT; CAIRNS, 2005).

Em contraste com a Visualização da Informação tradicional, descobertas a partir da visualização podem ser reutilizadas para construir um modelo de análise automática. É possível que esses modelos visuais também sejam construídos a partir dos dados originais usando métodos de mineração de dados. Uma vez criado o modelo, tem-se a capacidade de interagir com os métodos automáticos a fim de modificar os parâmetros ou escolher outros algoritmos de análise. O modelo visual pode então, ser utilizado para verificar ou validar as descobertas destes modelos. Esta interação entre os estágios de Visualização e Modelos, de forma cíclica é definida pelas transições Construção do Modelo e Visualização do Modelo.

Os métodos visuais e automáticos, quando intercalados, caracterizam o processo de análise visual, que conduz a um contínuo refinamento das informações e provê uma verificação constante dos resultados. No *Visual Analytics*, o conhecimento pode ser adquirido a partir de visualização, análise automática, bem como das interações anteriores entre visualizações, modelos automáticos e analistas humanos, conceituando assim o último estágio chamado Conhecimento.

Além dos processos de KDD e de Visualização de Informação, o *Visual Analytics* também propõe um funcionamento de forma cíclica, denominado por Keim e Thomas (2006) como *feedback loops*, onde o conhecimento adquirido pode ser reutilizado em outros conjuntos de dados no intuíto de contribuir para que resultados melhores sejam encontrados de forma mais rápida e eficiente no futuro.

2.2 Coordenadas Paralelas

Métodos clássicos de visualização e análise de dados como, por exemplo, diagramas de dispersão, gráficos de coordenadas x-y e, com um esforço computacional maior, gráficos em três dimensões são ferramentas indispensáveis na construção inicial de um modelo visual dos dados. Porém, quando há um conjunto de dados que ultrapassa as três dimensões, tais métodos não oferecem uma estrutura gráfica viável para a formulação de uma visualização que englobe todas as dimensões de forma unificada (ZHOU et al., 2008).

Cenários multidimensionais são comuns no processo de extração de conhecimento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), assim se faz necessário a utilização de ferramentas visuais capazes de apresentar e formular modelos além das três dimensões.

Dentro das várias técnicas desenvolvidas, a de Coordenas Paralelas vem sendo amplamente adotada para a visualização de conjuntos de dados altamente dimensionados e multivariados. Utilizando eixos paralelos para a construção das dimensões, a técnica de Coordenadas Paralelas pode representar um dado *n*-dimensional em um espaço bidimensional (ZHOU et al., 2008), (HEINRICH; WEISKOPF, 2013).

Nas subseções seguintes estão descritos o mapeamento visual, importância geométrica, construção e análise das Coordenadas Paralelas.

2.2.1 Mapeamento Visual

A Visualização de Informação, contida no processo de *Visual Analytics*, necessita para a criação das visualizações a etapa chamada Mapeamento Visual, a qual tem como função mapear os dados em estruturas visuais, estas estruturas por sua vez, tem o intuito de comunicar informações do computador para o ser humano, utilizando uma representação visual como meio de comunicação.

Para isso um conjunto de dados é computacionalmente mapeado de forma visual por alguma função *F*, que leva o conjunto de dados como entrada e gera a representação visual como saída.

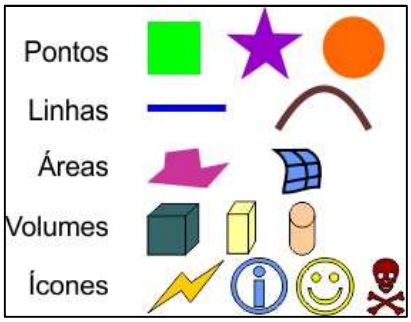
North (2005) cita quatro características importantes da função *F*:

- **Computável**: *F* é uma função matemática que pode ser executada por um algoritmo. Embora haja um espaço significativo para a criatividade na concepção dessas funções, a execução delas deve ser algorítmica;
- **Invertível**: Deve ser possível a utilização de F^{-1}, o inverso da função *F*, para reconstruir os dados da representação visual a um grau de precisão desejado. Se isso não for possível, a visualização será ambígua, enganosa, ou não interpretável;
- **Comunicável**: *F* (ou, de preferência, F^{-1}) deve ser conhecida pelo usuário para que ele possa interpretar a representação visual. Esta função pode ser comunicada junto à visualização, ou já conhecida por ele através de experiências anteriores. Em termos de usabilidade, esta é uma questão de capacidade de aprendizado;
- **Cognoscível**: F^{-1} deve minimizar a carga cognitiva para decodificar a representação visual. Esta é uma questão de percepção e desempenho humano.

Além da função *F*, o Mapeamento Visual é composto por espaços visuais (áreas que serão utilizadas para a plotagem), glifos e propriedades visuais.

Os glifos, também comumente chamados de marcas, mapeiam cada entidade de dado em uma entidade visual, seguindo um dicionário conforme mostra a Figura 4.

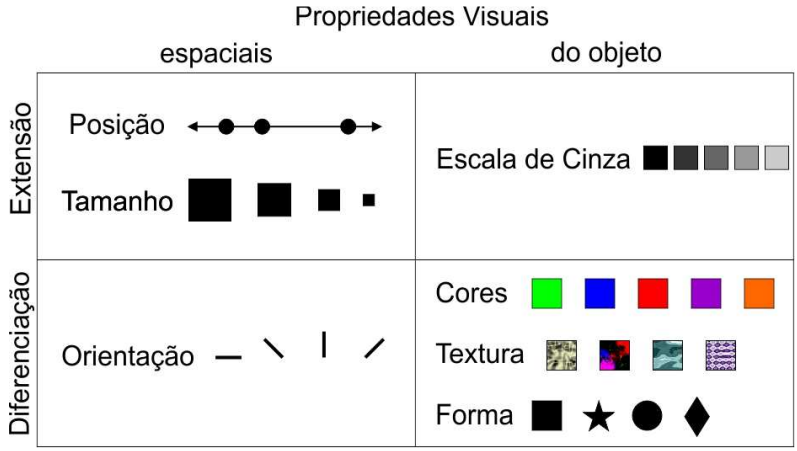
Figura 4: Adaptação do dicionário de glifos.



Fonte: Adaptado de North (2005).

Após mapear as entidades em glifos, deve-se definir o comportamento delas no espaço, para isso são utilizadas propriedades visuais. Essas propriedades incluem posição espacial, tamanho, cor, orientação e forma, as quais irão mapear os valores dos atributos de cada entidade, seguindo os exemplos da figura abaixo.

Figura 5: Adaptação das propriedades visuais.



Fonte: Adaptado de Card, Mackinlay e Shneiderman (1999) e North (2005).

Após o Mapeamento Visual, ainda é possível que se manipule dinamicamente a visão gerada. Como já mencionado, essa possibilidade de transformação é chamada de *Information Seeking Mantra* (SHNEIDERMAN, 1996).

2.2.2 Importância Geométrica

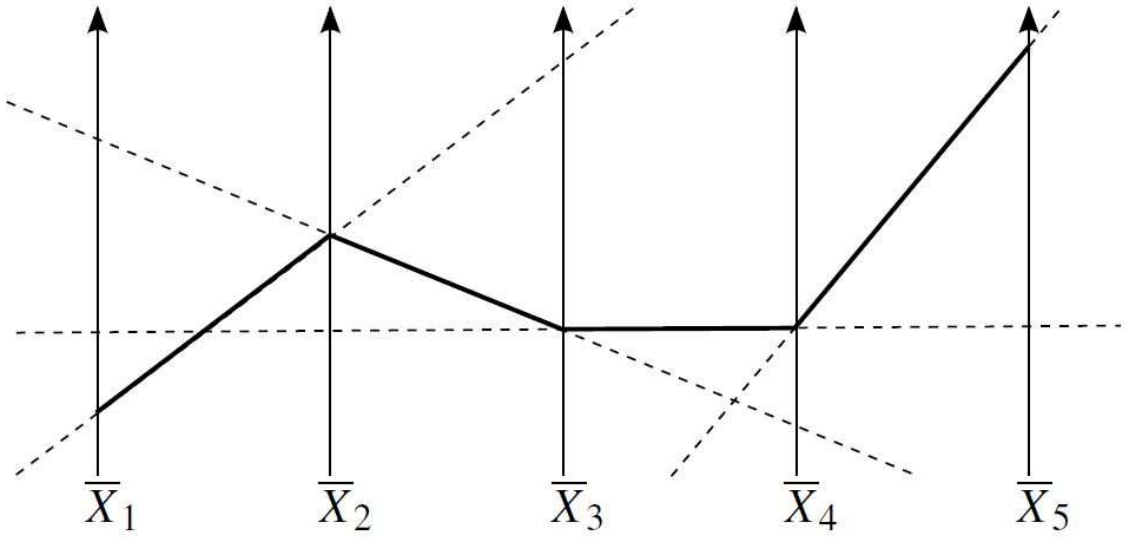
Um sistema de coordenadas, segundo Heinrich e Weiskopf (2013), proporciona um esquema para a localização de pontos uma vez informada as suas coordenadas e vice-versa. A escolha do sistema de coordenadas é, por conseguinte, um passo importante na visualização de dados, uma vez que transforma os dados que serão visualizados em uma representação geométrica (localização de uma amostra sobre um sistema de coordenadas). Com transformações nas coordenadas, as linhas retas (por exemplo, em coordenadas cartesianas) podem ser mapeadas para curvas (por exemplo, em coordenadas polares) ou pontos (por exemplo, em Coordenadas Paralelas), caracterizando assim o seu Mapeamento Visual, uma vez que as Coordenadas Paralelas se utilizam de linhas para representar cada um dos pontos de dados.

A escolha do sistema de Coordenadas Paralelas é, em grande parte, determinante para os padrões que serão exibidos por uma visualização e, portanto, é importante saber como analisa-la. Depois de introduzir a construção de Coordenadas Paralelas, é apresentado brevemente um modelo que pode ser utilizado e como um sistema de coordenadas cartesianas pode ser transformado em uma visualização de Coordenadas Paralelas.

2.2.3 Construção das Coordenadas Paralelas

Essencialmente, as Coordenadas Paralelas transformam padrões multidimensionais em padrões bidimensionais. Cada ponto do dado é representado por uma linha que atravessa *n* eixos paralelos, estes eixos representam cada uma das dimensões originais do espaço. Basicamente, as Coordenadas Paralelas são uma representação de *n* eixos *y* colocados um a um de forma paralela e subsequente. A distância de cada um desses eixos adjacentes é assumidamente igual. Um ponto no espaço *n*-dimensional se transforma em uma série de *n*-1 linhas conectadas, nas coordenadas paralelas, que intersectam cada um dos eixos na altura apropriada para o seu valor (BERTHOLD; HALL, 2003), (INSELBERG, 1985). Um exemplo de Coordenadas Paralelas com 5 dimensões é apresentado na Figura 6 onde, tipicamente, somente os segmentos entre as dimensões são desenhados (linha mais escura).

Figura 6: Construção de Coordenadas Paralelas com 5 dimensões.



Fonte: Adaptado de Heinrich e Weiskopf (2013).

2.2.4 Interação

Diferente de gráficos estáticos, as visualizações utilizadas no *Visual Analytics* podem ser dinâmicas. Esta característica possibilita um certo nível de interação fornecendo uma nova maneira para descobrir novas informações, detalhes e padrões. As formas mais clássicas de interação, segundo Shneiderman (1996), são descritas a seguir, bem como elas podem ser aplicadas sobre as Coordenadas Paralelas:

- **Localização**: possibilidade de verificar a qual amostra determinado glifo está representando, ou seja, no caso das Coordenadas Paralelas, descobrir a qual amostra do conjunto de dados determinada linha está representada;
- **Controles de ponto de vista**: a aplicação de *zoom* sobre a visualização e o reposicionamento das dimensões são opções frequentes para este tipo de controle. Nas Coordenadas Paralelas pode-se citar a possibilidade de reordenar as dimensões ou aproximar a visão em determinados pontos que podem se apresentar de forma mais densa;
- **Distorção**: aplicar diferentes colorações em determinadas amostras, assim como a possiblidade de ocultá-las. Aqui cita-se a técnica de *brush* que consiste em selecionar um subconjunto de amostras, este então ganha uma nova coloração de forma automática. Esta técnica é útil nas Coordenadas Paralelas, uma vez que possibilita a seleção de amostras em uma dimensão *i* e se consiga ver onde este ponto atravessa em todas as demais dimensões existentes.

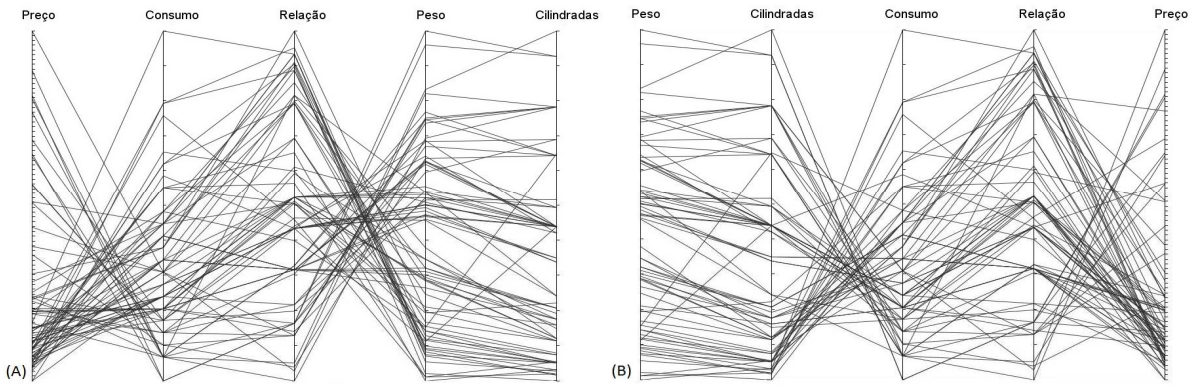
2.2.5 Análise

A análise que pode ser feita sobre Coordenadas Paralelas depende da qualidade da visualização gerada, do conhecimento do contexto e também, da capacidade cognitiva do

analista. Para exemplificar uma possível análise sobre Coordenadas Paralelas, Wegman (1990), sendo um dos pioneiros a empregar tal método sobre conjuntos de dados conhecidos, apresenta uma exemplificação sobre o conjunto de dados Carros (CHAMBERS et al., 1983). Este conjunto de dados possui 74 amostras de carros e, dentre seus 14 atributos, para os exemplos são utilizadas somente 5: Preço, Consumo, Relação (entende-se aqui o conjunto de engrenagens que compõem o sistema de marchas, este sistema é classificado de acordo com o tamanho da engrenagem, menores engrenagens maior o torque e menor a velocidade, maiores engrenagens menor o torque e maiores as velocidades), Peso e Cilindradas.

Analisando a Figura 7A, possivelmente a característica mais impressionante é o número de cruzamentos entre Relação (tamanho da engrenagem) e Peso. Este cruzamento sugere uma correlação negativa, o que faz sentido, uma vez que carros mais pesados tendem a ter motores de maior tamanho, que por sua vez proveem torques consideráveis, assim requerendo uma Relação menor. Reciprocamente, um carro mais leve tende a ter motores menores provendo um torque menor, requerendo assim uma Relação maior.

Figura 7: Conjunto Carros utilizando 5 dimensões com ordenações diferentes.



Fonte: Adaptado de Wegman (1990).

Considerando também (na Figura 7A) o vínculo entre Peso e Cilindrada, existe um número considerável de linhas paralelas aproximadas (uma quantidade relativamente baixa de cruzamentos), o que sugere uma correlação positiva. Esta visualização representa um fato que a maioria das pessoas tem como óbvio: que carros grandes (Peso) possuem grandes motores (Cilindradas). Bastante surpreendente, todavia, é o declive negativo que vai dos menores pesos às cilindradas moderadas. Isto claramente é uma discrepância que é incomum não somente para ambas as variáveis, mas também para o conjunto de dependências existentes como um todo. A mesma observação é enfatizada na Figura 7B.

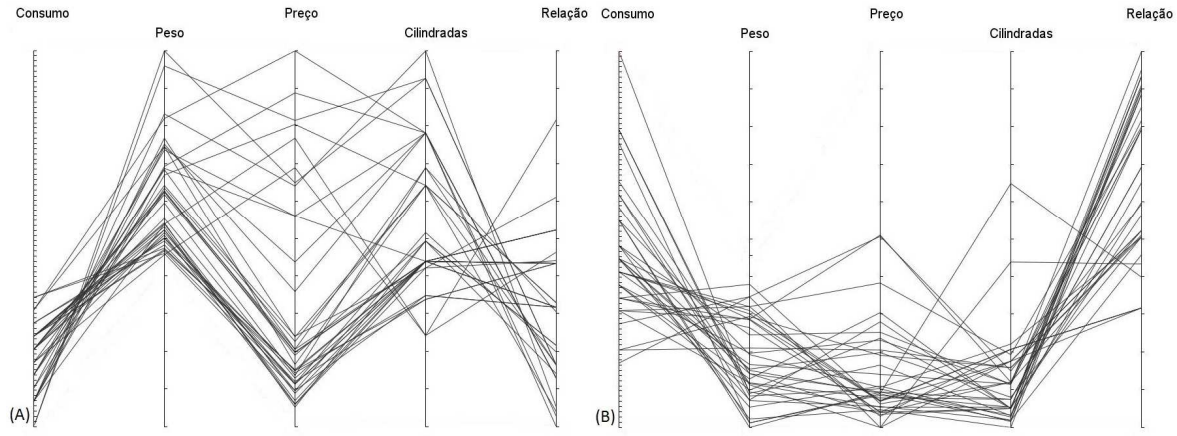
O vínculo entre Consumo e Preço também é de suma importância, uma vez que (na Figura 7A) os menores preços apresentam um limite hiperbólico aproximado e os maiores preços, claramente, ilustram cruzamentos. Isto sugere que carros mais baratos ou com consumos não muito eficientes não retratam uma grande correlação, todavia, para carros de valores maiores, quase sempre há um consumo ineficiente, outro ponto a se observar é que os melhores consumos estão localizados nos menores preços.

Permutada sobre a Figura 7A, a Figura 7B apresenta uma ligação instrutiva entre Relação e Consumo. Esta visualização sugere duas classes. Nota-se que existe um grande número de amostras que apresentam um pequeno aclive entre essas duas dimensões, mas há também um número considerável de amostras que apresentam um declive. Dentro dessas duas classes, existe um paralelismo aproximado, sugerindo assim, que a associação entre Relação e Consumo é aproximadamente linear. Isto é uma afirmação plausível, uma vez que: relações

menores implicam em grandes motores que implicam em um consumo elevado, porém relações maiores implicam em motores menores que, por sua vez, implicam em um consumo mais eficiente.

As Figura 8A e Figura 8B representam um subconjunto dos dados, ambas apresentam também uma nova permutação das dimensões. A Figura 8A descreve a classe de veículos que tem um consumo ineficiente, também apresenta veículos relativamente pesados e baratos, têm relativamente grandes motores e menores relações. Já na Figura 8B demonstra a classe de veículos que tem relativamente um consumo mais eficiente, são relativamente leves e baratos, têm relativamente motores menores e maiores relações. Isto, em 1970, caracterizava carros domésticos e carros importados respectivamente.

Figura 8: Conjunto Carros, novamente permutado e com subconjuntos selecionados.



Fonte: Adaptado de Wegman (1990).

2.3 Considerações

Este capítulo, na seção 2.1, ao abordar a definição de *Visual Analytics*, procura traçar um paralelo sobre os conceitos de KDD e Visualização de Informação que implicam na formulação da área da análise visual. Entretanto, não se restringe somente a conceituação, sendo que a mesma seção descreve o distanciamento que ambas as áreas (KDD e Visualização de Informação) apresentam na literatura. Áreas às quais são tratadas de forma paralela, porém não unificada, onde as tentativas de integração então citadas, procuram expor uma nova maneira de enxergar a análise visual diante de uma única estrutura. Mesmo com o objetivo definido, é notado que na mesma literatura, esta necessidade esbarra na carência por ferramentas que disponham desta funcionalidade.

Como técnica de Visualização de Informação foi demonstrada as Coordenadas Paralelas. Esta técnica é tida como central neste trabalho pela sua vasta popularidade e reconhecimento. Nesta constatação soma-se os benefícios que o método propicia ao KDD, uma vez que as Coordenadas Paralelas podem representar dados com alta dimensionalidade em um formato facilitado pela sua representação geométrica bidimensional, construída sobre linhas e eixos. Não obstante, esta técnica é capaz de fornecer, como resultado, padrões existentes no conjunto de dados, os quais são possíveis de serem validados pela etapa de mineração de dados.

Não se limitando as constatações mencionadas, as Coordenadas Paralelas são utilizadas nos trabalhos relacionados (Capítulo 3) onde auxiliam e ampliam a compreensão dos dados e

como eles estão relacionados dentro dos contextos apresentados. Não se restringindo somente ao seu uso efetivo, mas sim sendo aconselhadas como forma integrante no processo de extração de conhecimento.

Todavia, abordagens que integram ambas as áreas são encontradas na literatura e limitadas, somente, pelo fato de trabalharem com dados restritos a contextos específicos. Visto no capítulo seguinte, os trabalhos que buscam efetuar a união entre essas áreas limitam-se a ferramentas paralelas e individuais, onde a integração é realizada por vias não automatizadas (geram-se arquivos que são importados de uma ferramenta a outra), ou por ferramentas criadas especificamente para a área em que os dados estão inseridos.

Deste modo a falta de integração entre estas duas áreas, e a inexistência de um ferramental adequado são abordados nos trabalhos relacionados apresentados no Capítulo 3 e formulam a base para os Capítulo 4 e 5, que demonstram o processo proposto e sua aplicação respectivamente.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados três trabalhos que propõem artefatos unificados para o *Visual Analystics*, em comum, além da análise visual, todos os três utilizam as Coordenadas Paralelas para encontrar padrões em um grande volume de dados. O primeiro explana o desenvolvimento de um *software* chamado EDEN (***E****xploratory* ***D****ata analysis* ***EN****vironment*) (STEED et al., 2012) e sua aplicação na análise de simulações complexas sobre o clima na Terra, uma vez que análises convencionais são inadequadas diante da complexidade dos dados atuais (STEED et al., 2013). O segundo, proposto por Blaas, Botha e Post (2008), agrega melhorias na construção básica das Coordenadas Paralelas para grandes volumes de dados, no contexto dessa aplicação deve-se considerar as simulações matemáticas responsáveis por gerar as milhões de amostras que compõem o conjunto. Por último, o trabalho de Hasenauer et al. (2012) busca encontrar padrões sobre processos celulares, sendo este responsável por definir o destino de uma célula, ou seja, se ela irá sobreviver ou não. A última seção apresenta considerações sobre os três trabalhos relacionados traçando paralelos importantes para o posicionamento desse trabalho frente à literatura.

3.1 Exploratory Data analysis ENvironment (EDEN)

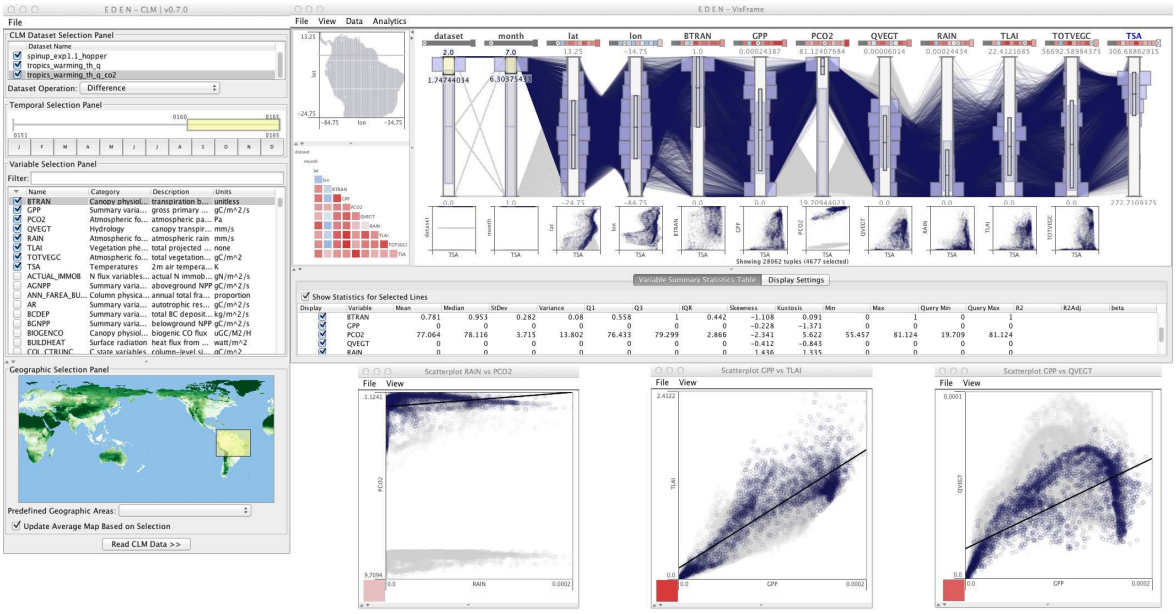
Diante de um meio ambiente altamente variável e em constante mudança, Steed et al. (2013) citam a necessidade de avaliar o clima do passado a fim de prever o futuro. Com capacidades computacionais cada vez mais elevadas e um conhecimento do planeta Terra mais avançado, simulações demasiadamente mais complexas puderam ser criadas com o intuito de reproduzir, de forma numérica e com maior fidelidade, o clima. Estas simulações geraram grandes massas de dados (*big data*), as quais apresentaram dificuldades para serem analisadas com as ferramentas convencionais utilizadas para a análise climática em escala global.

Tipicamente, avalia-se conexões entre as variáveis com visualizações básicas como, por exemplo, gráficos de pontos, gráficos de linhas e histogramas. Estas visualizações além de se limitarem a no máximo três dimensões, não trabalham de forma conjunta com os dados, sendo uma abordagem pouco recomendada por conta da limitada memória humana (RENSINK, 2002). Steed et al. (2013) citam que métodos estatísticos adicionais não são integrados com as ferramentas que geram as visualizações, dificultando a extração do conhecimento.

Para contornar tal cenário, propõem-se uma ferramenta integrada que dispõe de inúmeras visualizações e outras funcionalidades como análise de *big data*. Esta foi chamada de EDEN, com um forte apelo à alta capacidade visual humana de descobrir padrões interagindo com as Coordenadas Paralelas e outras visões baseadas em coordenadas (STEED et al., 2012).

O EDEN, conforme pode ser visto na Figura 9, é composto por diversos painéis, dentre os mais importantes citam-se os painéis de filtro, sendo eles: o painel para a seleção do conjunto de dados que será analisado, painel para a seleção temporal dos dados climáticos, painel de seleções das variáveis que serão analisadas e o painel de seleção geográfica. Além dos painéis de filtro, há os painéis de visualização, que englobam as Coordenadas Paralelas, matriz de dispersão e matriz de correlação.

Figura 9: Visão geral da ferramenta EDEN.



Fonte: Adaptado de Steed at al. (2013).

Modelos de simulações climáticas complexos e o aumento da massa de dados inerente a este cenário implicam em um contexto desafiador para a análise dos dados. Neste sentido o EDEN pode utilizar-se, como é proposto por Smith et al. (2013), de uma ferramenta auxiliar chamada ParCAT (***Par****allel* ***C****limate* ***A****nalysis* ***T****ools*), dado que esta tem como objetivo diminuir a densidade dos conjuntos de dados resultantes das simulações. Para isto, o ParCAT foi desenvolvido aplicando, de forma eficiente, a computação paralela, tanto de computadores domésticos como também *grids* computacionais de alto desempenho. Isto tem como finalidade processar os conjuntos a fim de criar agregações, médias, somas ou compilações dos dados, de *n* formas (perspectivas) possíveis e de forma paralela. Os conjuntos resultantes podem então ser analisados utilizando histogramas ou originarem arquivos que possam ser carregados em outras aplicações. Smith et al. (2013) ainda citam que o ParCAT, por funcionar por linha de comando, propicia uma maneira fácil de integra-lo, tanto no modo como foi introduzido ao EDEN como também ao criar a possibilidade de expandir qualquer outra ferramenta.

Para comprovar a eficácia do EDEN, dois conjuntos de dados resultantes de simulações reais foram utilizados. O primeiro é o CLM4 por pontos agrupados e o segundo é o CLM4 global. Ambos conjuntos de dados possuem 360 variáveis, sendo possível analisar algumas delas de forma bidimensional ou tridimensional. O volume de dados gerados em cada simulação para cada mês pode chegar aos 415 megabytes. Assim, o ParCAT é responsável por agrupar esses dados de acordo com o contexto da análise e apresentá-los nas visualizações existentes no EDEN.

A ferramenta EDEN, no exemplo da Figura 9, tem como seleção três conjuntos simultâneos, 8 variáveis e dados situados em um intervalo de 3 anos. Ainda nesta mesma figura, é possível verificar que somente uma parte das linhas está na cor azul, estas linhas estão selecionadas por distorção (ou *brushing)* e foram escolhidas, neste exemplo, por tratarem somente de “latitudes superiores”, esta seleção por sua vez ressalta uma característica interessante, a variável TOTVEGC (Total de Carbono Vegetal) apresenta a formação de três agrupamentos. Mas além disso, é possível verificar-se que as linhas não selecionadas (as de cor cinza) se sobressaem na variável RAIN (Chuva), que apresenta valores mais elevados do que

as linhas atualmente selecionadas. Tais observações são relevantes, uma vez que apresentam a relação que os dados possuem entre si.

3.2 Coordenadas Paralelas interativas aplicadas a vastos conjuntos temporais

As Coordenadas Paralelas têm recebido uma grande aceitação nas áreas de estatística e Visualização de Informação como método arbitrário para a análise de conjuntos de dados altamente dimensionados (WONG; BERGERON, 1994). Em parte, esta aceitação acontece uma vez que cada amostra dos dados é representada unicamente por uma linha e que, por causa de sua projeção, não ocasiona a perda de informações, situação que não acontece em outras visualizações, como por exemplo, nas matrizes de dispersão (CLEVELAND, 1993).

Blaas, Botha e Post (2008) afirmam que a limitação das Coordenadas Paralelas não se encontra no número de dimensões por amostra, mas sim pelo número de amostras associadas às linhas, sendo que a escalabilidade das Coordenadas Paralelas é limitada por dois fatores principais: poluição visual e redução do desempenho. Fatores os quais dificultam a utilização desta técnica para a exploração interativa dos dados.

Diante destas dificuldades, Blaas, Botha e Post (2008) propõem uma técnica que permite a interação exploratória de dados volumétricos temporais que, segundo os autores, vão de um milhão até dezenas de milhões de amostras. Esta técnica utiliza uma combinação entre quantização e compressão dos dados, utilizando estruturas de rápida computação e renderização gráfica. Obtendo como resultado a apresentação das amostras de forma densa e "quase" contínua entre os pares de eixos paralelos. Dentro das interações suportadas por essa abordagem, cita-se: distorção (*brushing*), permutação dos atributos e a possibilidade de referir-se ao dado original ao interagir com a linha da amostra.

De modo inicial deve-se considerar que a aplicação seja eficiente em termos de renderização das Coordenadas Paralelas, principalmente ao se utilizar uma interação. Para isso, processos paralelos (pré-processamento e renderização) foram adotados.

O pré-processamento necessita ser utilizado somente uma vez e é responsável por armazenar os dados *on-the-fly* (de forma dinâmica e manipulável sem que haja qualquer parada da aplicação). Sempre que um intervalo temporal é selecionado, esta forma de pré-processamento carrega os dados do disco e os compacta utilizando LZO[5] (Lempel-Ziv-Oberhumer) como método de compressão. Após comprimidos os dados são guardados em arquivos separados, essa separação é definida pelo atributo e intervalo de tempo que segundo Blaas, Botha e Post (2008), acelera o processo de procura/carregamento da informação. Porém, antes dos dados serem armazenados duas otimizações são realizadas. Primeiramente o dado é normalizado em um intervalo de 0 a 1, esta normalização é baseada na distribuição dos dados, que é analisada pelo seu histograma de modo que os valores tendam a prover uma visualização contínua no domínio de destino. Por último, há a otimização chamada de quantização, que se refere a estabelecer um número de casas decimais fixo para as amostras, evitando assim dados com representações numéricas altamente precisas, isto além de reduzir o tamanho da informação armazenada, atua como uma forma de normalização, que os autores concluíram beneficiar o processo de análise visual.

[5] LZO: Tecnologia de compressão dos dados que oferece a descompressão em tempo real segundo a página do autor da tecnologia (http://www.oberhumer.com/opensource/lzo/).

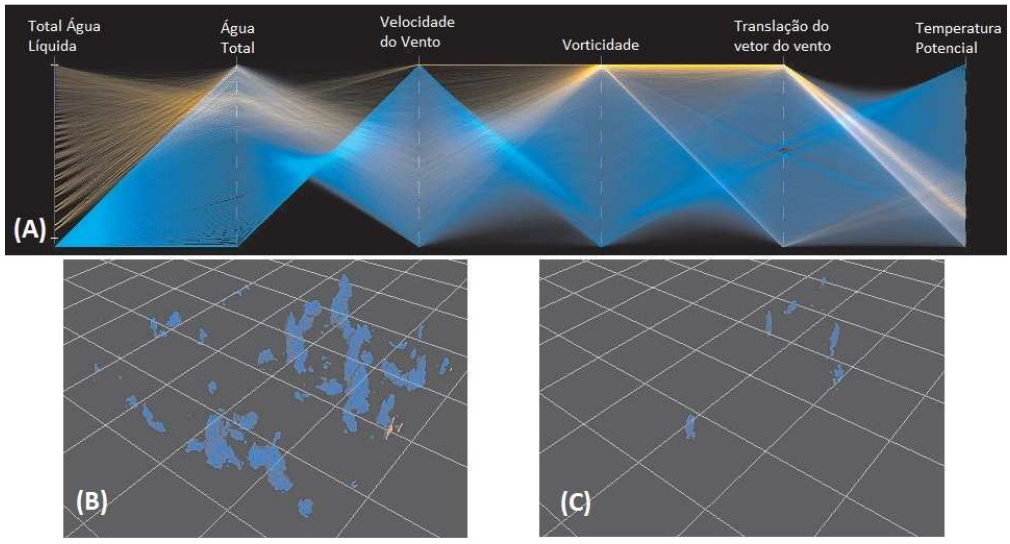
O processo de renderização, executado em paralelo ao pré-processamento, acontece a cada interação do usuário. Para que se evite a renderização de uma linha para cada amostra dos dados, sugere-se a criação de um histograma conjunto (2D) de acordo com a proposta de Artero, Oliveira e Levkowitz (2004). Este histograma é construído entre cada par de eixos (que representam cada qual seu atributo). Após criados estes histogramas conjuntos, utiliza-se a técnica de Muigg et al. (2008) que aborda uma forma de desenhar itens primitivos (linhas, pontos ou triângulos) de acordo com os *bins* de cada histograma conjunto. Se tratando de volumosas bases de dados, os *bins*, mesmo em número inferior ao de amostras, também apresentam uma grande concentração representada por uma mesclagem de cores, onde a intensidade de uma cor até outra é controlada por uma escala logarítmica.

As possibilidades interativas de seleção, ordenação e localização, implementadas por Blaas, Botha e Post (2008), utilizam-se dos processos paralelos descritos, sendo que a seleção necessita somente da etapa de renderização, uma vez que os dados já foram carregados, a visualização de Coordenadas Paralelas necessita somente reprocessar a intensidade das cores selecionadas (sendo que o usuário pode escolher qual cor determinada seleção deve assumir). Já a possibilidade de ordenação necessita que os histogramas conjuntos sejam criados para cada novo par de eixos formado, logo, esta criação requer que novos dados sejam carregados do disco, utilizando assim, a etapa de descompressão abordada no pré-processamento. Por último, a técnica de localização tem suporte em outras visualizações disponibilizadas pela ferramenta, onde o conjunto de visualizações disponíveis compartilha das mesmas amostras, ao se selecionar uma amostra em qualquer visualização, a mesma é destacada também nas outras visualizações.

Para comprovar a eficácia de seu trabalho, Blaas, Botha e Post (2008) utilizaram a simulação de Large-Eddy (LES), proposta por Smagorinsky (1963). Dessa simulação, um conjunto de dados de 4 atributos foi gerado: total de água líquida, temperatura potencial, vetor do vento e a quantidade de água total. Além destes 4 atributos, dois novos foram criados por derivações na etapa de pré-processamento, sendo eles: velocidade do vento e vorticidade.

A Figura 10 demonstra o resultado dessa simulação sobre as Coordenadas Paralelas (A), e duas replicações físicas: (B) apresentação de todas as nuvens do conjunto e (C) nuvens com maior velocidade. Ao todo o conjunto contém 1.300.000 amostras, resultando em total de 17.58GB de tamanho.

Figura 10: Adaptação da simulação Large-Eddy para 6 atributos.



Fonte: Adaptado de Blaas, Botha e Post (2008).

Nesta figura, verifica-se que a seleção de linhas nas Coordenadas Paralelas (Figura 10A – em laranja) assinala na Figura 10B os respectivos pontos (amostras), e que mesmo não demonstrando inicialmente uma relação entre os atributos, pode-se avaliar que a concentração de pontos que contém água está aproximada geograficamente, o que representa a formação de novas nuvens. Uma segunda seleção é realizada na Figura 10C, onde Blaas, Botha e Post (2008) ocultam todas as linhas que não apresentam a velocidade máxima, resultando assim, novamente, em pontos aproximados geograficamente que, conclui-se, representarem as nuvens de maior velocidade.

3.3 Análise Visual para modelos heterogêneos de populações celulares

As populações celulares são heterogêneas em termos de idade, ciclo e abundância de proteína. Esta heterogeneidade influencia nas decisões pertinentes ao destino da célula, como a morte ou a proliferação. Assim, para conhecer e controlar o comportamento populacional, os principais aspectos de variabilidade célula-a-célula devem ser elucidados. Porém, esta elucidação apresenta-se desafiadora, uma vez que restrições são aplicadas aos experimentos (HASENAUER et al., 2012). A maioria dos sistemas experimentais e dispositivos de medição permitem somente a avaliação simultânea de algumas propriedades de uma única célula. Com isto, a análise puramente experimental de processos que dependem de diferentes propriedades celulares é limitada.

Spencer et al. (2009) sugerem que a limitação dos experimentos pode ser contornada parcialmente utilizando modelos matemáticos. Estes modelos descrevem uma população heterogênea, frequentemente representando-a com o uso de agentes. Cada agente fornece uma descrição do mecanismo de transdução de sinal (quando uma célula converte um sinal ou estímulo em outro) individual para cada célula, demonstrando assim o seu comportamento. Nesta estrutura, a diferença de variabilidade pode ser tanto modelada de forma estocástica como determinística.

Hasenauer et al. (2012) focam-se nas diferenças determinísticas entre as células contidas em populações não interativas. Estas diferenças são comumente modeladas usando parâmetros diferenciais e condições iniciais (SPENCER et al., 2009). Diversos modelos existentes na literatura, como os propostos em Hasenauer et al. (2011a), Hasenauer et al. (2011b) e Koeppl et al. (2012), podem ser aplicados para inferir a distribuição dos parâmetros e as condições iniciais dos dados experimentais, a fim de obter modelos mecanicistas para populações de células. De todo modo, os modelos baseados em agentes resultantes são, na maioria dos casos, extremamente complexos.

Esta complexidade impede a análise desses modelos usando ferramentas para sistemas dinâmicos (GUCKENHEIMER; HOLMES, 1983), como por exemplo a sensibilidade ou a análise de bifurcação. Desse modo, para Hasenauer et al. (2012) não há nenhuma abordagem estruturada para a análise de modelos celulares populacionais heterogêneos disponível.

Diante disto, Hasenauer et al. (2012) propõem dois métodos para preencher esta lacuna e prover uma análise facilitada, são eles: Coordenadas Paralelas e SVM – *Support Vector Machines* – (HEARST et al., 1998). Sendo que ambos são amplamente utilizados para a análise de conjuntos de dados multidimensionais e foram endereçados para responder: “Quais parâmetros causam a heterogeneidade da resposta da população?”.

Para explorar esses métodos de análise, dois conjuntos de dados foram gerados por meio de simulação. O intuito contextualizado nestes dados gerados é verificar possíveis dependências entre os parâmetros com a seleção de marcadores moleculares.

Análises em conjuntos de dados simulados para modelos populacionais celulares heterogêneos, na grande maioria dos casos consiste na redução da complexidade ao passo em que informações importantes se mantém preservadas. Neste sentido, as técnicas de visualização podem ajudar a determinar quais são os parâmetros importantes evitando a perda de informação. Em Hasenauer et al. (2012), a técnica de Coordenadas Paralelas é utilizada para obter um maior conhecimento sobre as dependências entre os atributos, que por sua vez apresentam uma alta dimensionalidade. Especificamente nesta situação, as dimensões de maior interesse são as que claramente separam o conjunto de dados em classes, portanto apresentam-se como boas candidatas para a seleção de potenciais marcadores.

Em um segundo momento, os marcadores selecionados são utilizados para treinar as SVMs. Estas SVMs permitem uma evolução quantitativa da qualidade dos marcadores. Mesmo que as SVMs sejam úteis por si só, testar todas as combinações possíveis de marcadores resultaria em uma explosão combinatória. Assim, combinando SVMs com as Coordenadas Paralelas, o número de avaliações necessárias nas SVMs é reduzido drasticamente, resultando em uma alta redução da complexidade computacional.

Além de uma melhor compreensão do modelo, os resultados obtidos durante a análise podem ser utilizados tanto para adaptar o modelo populacional, como para induzir ou selecionar novos experimentos. A estrutura sugerida por Hasenauer et al. (2012) integra visualizações interativas com métodos automáticos de análise, além de permitir simultaneamente que o conhecimento obtido seja utilizado de forma crítica sobre o atual modelo, incorporando assim um importante aspecto do *Visual Analytics* (THOMAS; COOK, 2006).

Inserido em um contexto em que o número de amostras é conhecidamente vasto, Hasenauer et al. (2012) sugerem a utilização de melhorias nas Coordenadas Paralelas. Segundo os autores, o grande número de amostras resulta em um grande número de linhas que irão transpassar os eixos, que, ao se sobreporem podem esconder informações sobre o conjunto de dados, ocultando assim possíveis padrões que poderiam ser visualizados. Sugerido como solução, optou-se por estimar a densidade das linhas e desenhá-las de forma que exista uma variação da intensidade da cor até a transparência (*alpha blending*). Caso a variável em questão seja contínua, mapas de cores também podem ser utilizados, onde uma cor é gradativamente transformada em outra.

Já as SVMs permitem a derivação de preditores para propriedades qualitativas e quantitativas. Estes indicadores podem ser usados para avaliar o conteúdo informativo de um subconjunto dos parâmetros sobre as respectivas propriedades, facilitando assim a análise quantitativa de um marcador.

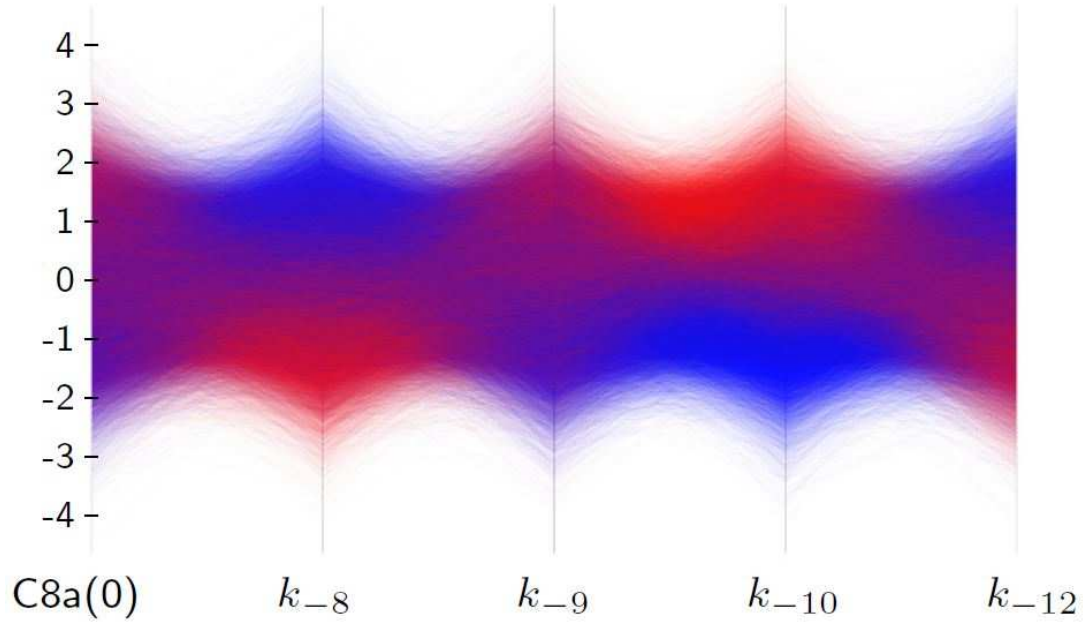
Para ilustrar o processo de *Visual Analytics* proposto, Hasenauer et al. (2012) utilizam um modelo de sinalização pró-apoptótica gerado matematicamente para análise. Sinalizações pró-apoptóticas estão envolvidas no processo de apoptose (WAJANT; PFIZENMAIER; SCHEURICH, 2003), que também é denominado como morte celular programada. A apoptose é um importante processo fisiológico para a remoção de células infectadas, que apresentam um mal funcionamento ou que não são mais necessárias em organismos multicelulares. O caminho de sinalização apoptótica converge para um efeito em cascata de caspases, onde caspases iniciadoras e caspases efetoras são ativadas. Caso a atividade das caspases efetoras ultrapassarem um certo limiar, induz-se a apoptose (SPENCER; SORGER, 2011).

A partir de experiências de citrometria de fluxo, é sabido que a quantidade de caspase 8 (C8) e caspase 3 (C3), caspases 8 e 10 associadas com proteínas RING (CARP) e proteína inibidora de apoptose (IAP) é diferente entre células individuais. Esta diferença é modelada por diferentes taxas de síntese (k_{-8}, k_{-9}, k_{-10} e k_{-12}) dentre as células individuais, estas taxas dão origem as dimensões analisadas pelo estudo realizado em Hasenauer et al. (2012).

Diante deste modelo celular populacional heterogêneo, analisou-se: (a) como acontece a decisão de uma célula sofrer ou não um apoptose dentro das 12 primeiras horas e (b) como o tempo da morte é influenciado pelos parâmetros escolhidos. Assim, duas variáveis de interesse surgem, a primeira define se a célula sobrevive ou não – provendo o resultado do processo de decisão – e a segunda é o tempo, que provê o prazo para o acontecimento do apoptose. Como indicador para o apoptose, a quantidade de C3 ativa (C3a) é utilizada. Caso exista mais de 5.000 cópias do C3a em uma célula, assume-se que esta célula irá sofrer um apoptose dentro de 10 minutos, definindo o tempo da morte da célula.

Para estudar o destino de vida ou morte de uma célula, um conjunto com 100.000 amostras foi plotado em uma visualização de Coordenadas Paralelas (Figura 11). Como somente duas classes são consideradas (morta ou viva), o *alpha blending* pode ser utilizado para visualizar a densidade de ambas as classes onde existe regiões de sobreposição, sendo que a intensidade da cor vermelha representa as células mortas e a intensidade da cor azul representa as células vivas. Quanto mais intensa a cor, maior o número de sobreposições e maior a densidade da região.

Figura 11: Parâmetros considerados e subgrupos.



Fonte: Adaptado de Hasenauer et al. (2012).

Analisando a Figura 11, onde a primeira dimensão representa a parametrização inicial da caspase 8 ativa – C8a(0) – verifica-se de forma aparente que o segundo e o quarto parâmetro (k_{-8} e k_{-10}) apresentam uma separação razoável entre as classes (vermelho = morta, azul = viva). Sendo assim, a maioria das células sobreviventes apresentam valores elevados de k_{-8} e baixos valores de k_{-10}, o que corresponde a valores expressivos de IAP e não expressivos de C3 respectivamente. Contudo, os outros parâmetros também influenciam no processo, porém de forma menos significante.

Partindo desta análise, Hasenauer et al. (2012) consideram os parâmetros k_{-8} e k_{-10} para computar a qualidade da classificação utilizando as SVMs. Sendo que como resultado, os autores citam que, em conjunto, o desempenho da classificação das variáveis é razoável. E que o resultado da combinação destes dois marcadores ultrapassa as demais combinações, uma vez que as demais classificações atingem níveis de falsos positivos maiores que 50%.

Por último, Hasenauer et al. (2012) buscam encontrar uma dependência entre o os parâmetros (k_{-8}, k_{-9}, k_{-10} e k_{-12}) e o momento da morte de uma célula. Porém, ao se apresentar as amostras sobre uma visualização de Coordenadas Paralelas, percebeu-se uma grande sobreposição, indicando que no caso do tempo, vários parâmetros podem ser determinantes.

3.4 Considerações

Neste capítulo foram apresentados três trabalhos que buscam demonstrar aplicações de técnicas visuais conjuntas com técnicas automáticas. Constatou-se que o conceito de *Visual Analytics* é utilizado em diversas áreas da análise de dados, porém somente dentro de contextos específicos. Soma-se a isto o fato que os trabalhos contam somente com ferramentas construídas com o intuito de suprir a necessidade do contexto ao qual os seus respectivos dados estão inseridos, não sendo apresentada uma solução genérica ou padronizada do processo de análise visual.

Esta lacuna causada pela falta de ferramental adequado provê uma nova forma de pesquisar meios para a extração do conhecimento e reconhecimento de padrões. Contemplar ambas as áreas, automáticas e visuais, sugere a criação ou ampliação de novas ferramentas a fim de suprir o distanciamento ocasionado pelo estudo paralelo de ambas as áreas.

Ao analisar-se o trabalho proposto por Steed et al. (2012), que se refere à ferramenta de análise EDEN, é possível avaliar o esforço empregado na construção de uma aplicação capaz de trabalhar com conjuntos de dados multidimensionais e apresentá-los de forma gráfica. Conjuntamente, este trabalho emprega técnicas de agrupamento e sumarização de dados, sinalizando de forma clara a necessidade de utilizar outros modelos que contenham a informação organizada de forma diferenciada, podendo assim ser apresentada nas visualizações. A combinação destas possibilidades é ampliada ao passo que o EDEN dispõe, assim como a proposta pelo presente trabalho, da visualização de Coordenadas Paralelas interativa, a qual proporciona uma maior facilidade na exploração e entendimento das relações entre os dados analisados.

Sabe-se que, ao observar-se de forma mais detalhada, o EDEN aceita como entrada de dados conjuntos que atendam a estrutura nele utilizada. O que faz com que esta ferramenta não se limite a dados do clima. Esta constatação é válida ao passo que indica uma possível generalização de sua utilização, mas que cria ressalvas ao observar-se que, utilizar o EDEN neste formato implica na criação de arquivos que serão importados para esta ferramenta, e que *feedbacks*, interações ou filtros criados em suas visualizações não possam ser retransmitidos para um novo conjunto de dados. De fato, os autores não elucidam, nem ao menos, a possibilidade de retransmissão dos dados ao utilizar-se a ferramenta complementar ParCAT, impossibilitando assim uma análise circular dos dados.

Ainda ao se referir ao ParCAT, verifica-se que a funcionalidade de redução de massa de dados nele contida não é reproduzida no presente estudo. Este fato é importante, uma vez que as Coordenadas Paralelas em seu formato clássico (INSELBERG, 1985), o qual é tido como principal por este trabalho, limita-se a desenhar cada amostra como uma linha que irá cortar

todos os eixos sendo que, computacionalmente, é limitado pelas tecnologias utilizadas tanto para a renderização como para o desenvolvimento da aplicação.

Já em (BLAAS; BOTHA; POST, 2008), verifica-se um extenso trabalho relacionado ao desempenho das Coordenadas Paralelas, este desempenho refere-se ao número de amostras de dados quando estas superam a casa dos milhões. Neste caso, a preocupação dos autores está voltada totalmente para os conjuntos de dados em sua formulação original, ou seja, que não tenham sido pré-processados e nem minerados, deste modo, o número de amostras permanece intacto, porém estas são traduzidas para um formato que possibilite a construção das Coordenadas Paralelas.

Este esforço empregado por Blaas, Botha e Post (2008) é restringido pela possibilidade de que, ao se deparar com uma massa volumétrica de dados, as amostras sejam apresentadas de maneira “quase” contínua, possibilitando assim que os recursos gráficos computacionais suportem tanto a exibição das inúmeras amostras quanto as possiblidades de interação. Essa tratativa dos dados nada mais é do que uma abordagem automática para a diminuição da densidade do conjunto, sendo que ao desenhá-lo de forma “quase” contínua, o computador não necessita redesenhar a cada nova interação, todas as milhões de amostras existentes.

Constata-se que nos trabalhos de Blaas, Botha e Post (2008) e Steed et al. (2012) há uma semelhança no esforço investido para diminuir o volume resultante de amostras (ou traduzi-las para formatos menos densos) a fim de possibilitar que a visualização de Coordenadas Paralelas seja renderizada de forma eficiente. Todavia deve-se verificar que, no momento em que a renderização ocorre, em ambos os casos, a estrutura visual utilizada nas Coordenadas Paralelas é a clássica, concluindo assim que os dados foram trabalhados previamente a renderização. Isto se reflete no contexto abordado pelo presente trabalho, onde em uma possível inviabilidade de renderização da visualização, algoritmos de redução de densidade dos dados possam vir a ser utilizados.

Por fim tem-se (HASENAUER et al., 2012), onde ressalta-se que a combinação das Coordenadas Paralelas com as SVMs propiciou não só a melhora do desempenho das SVMs como também serviu para homologar os resultados obtidos, uma vez que as saídas das SVMs condizem com as propriedades ressalvadas na visualização. Porém, os autores não utilizam nenhuma ferramenta que possibilite a utilização de tais técnicas de forma integrada, sugerindo assim, que o conhecimento obtido foi realizado por ferramentas distintas em momentos distintos. O mesmo trabalho também sugere uma melhoria nas Coordenadas Paralelas, sendo que as amostras são renderizadas utilizando cores (com aplicação de transparência) diferentes de acordo com agrupamentos previamente criados. Esta forma de coloração também é considerada uma técnica de distorção, porém, como está contido no Capítulo 4, a única possibilidade de distorção abordada no presente trabalho é por *brushing* (onde amostras selecionadas ganham cores diferentes).

Analisando-se os três trabalhos relacionados, verifica-se o empenho investido, principalmente, na criação e utilização de ferramentas que possam prover o conhecimento esperado. Isto sugere que, mesmo diante de várias tecnologias em seu estado da arte, necessita-se de um ferramental que tanto possibilite integrações como que ofereça, mesmo que de forma inicial, o básico de ambas técnicas (visuais e automáticas). Define-se aqui o básico, como a possibilidade de integração e intercomunicação das técnicas, ou seja, que ações tomadas no processamento dos dados reflitam em tempo real nas visualizações, sem a necessidade de mover-se os dados de um lado a outro através dos mais diversos formatos existentes, situação que nos trabalhos relacionados não acontece.

Verifica-se também uma popularidade na utilização das Coordenadas Paralelas como técnica utilizada, a mesma é constante em inúmeros trabalhos contidos na literatura, fazendo dela, a técnica considerada de maior importância para este trabalho. Conjuntamente, sugere-se que a criação de todas as visualizações de Coordenadas Paralelas citadas nos trabalhos foi construída de forma independente, isto é, não houve a menção de nenhuma construção prévia (como uma biblioteca, códigos fontes ou *framework*), que pôde (ou optou-se por) ser utilizada.

Outro ponto que se destaca nestes trabalhos relacionados, diz respeito a questão de dados volumétricos, ou seja, conjuntos que possam conter uma grande quantidade de amostras. Em (STEED et al., 2012) nota-se um grande apontamento a uma ferramenta secundária chamada ParCAT, cujo objetivo é reduzir a densidade dos dados. Em (BLAAS; BOTHA; POST, 2008) tratativas tecnológicas são utilizadas para contornar a problemática da apresentação de um vasto número de amostras em uma interface gráfica, situação semelhante a constatada em (HASENAUER et al., 2012). Sabendo-se desta limitação de dados apresentados nas Coordenadas Paralelas, este trabalho não busca criar nenhuma otimização como as apresentadas, limitando-se a formulação inicial proposta por Inselberg (1985).

Não obstante, os trabalhos relacionados contemplam, de forma unânime, as principais características do *Information Seeking Mantra* (SHNEIDERMAN, 1996), sendo existente a possibilidade de interação com a visualização ao se: a) permutar os eixos; b) aplicar alguma distorção, sendo as utilizadas *brushing* (amostras com cores diferentes) e transparência; c) saber a qual (ou quais) amostra (amostras) determinada linha representa.

Pontua-se que, o presente trabalho limita a interatividade das Coordenadas Paralelas ao propor desenvolver dois formatos de distorção. O primeiro diz respeito a permutação dos eixos e o segundo refere-se a técnica de *brushing*. A escolha destas duas distorções baseia-se no fato que ambas são utilizadas tanto na formulação clássica como também nas abordadas nos trabalhos aqui relacionados. Vale-se esclarecer que, na formulação clássica, os eixos já eram permutados para facilitar a exploração dos dados, todavia essa permutação acontecia sem a possibilidade de interação do usuário, sendo necessária intervenções na formulação do gráfico para que isso fosse possível.

Já as Coordenadas Paralelas abordadas pelos trabalhos relacionados não especificam de forma clara a possibilidade e facilidade de permutação de seus eixos, fato que no presente trabalho acontece de forma natural, sendo que o usuário tem pleno controle na reordenação das dimensões. O *brushing*, por sua vez, é inerente à análise exploratória dos dados, no sentido em que a limitação ou diferenciação das amostras desenhadas, foi parte fundamental para observar o comportamento dos dados contidos nos trabalhos relacionados, adiciona-se a este fato que, reduzir o número de amostras a serem visualizadas foi utilizada por Wegman (1990), mesmo que de forma manual, como sugestão de análise para as Coordenadas Paralelas (subseção 2.2.5, Figura 8).

Além das distorções citadas, a possibilidade de verificar qual amostra (dado original) determinada linha representa, também é permitida neste trabalho.

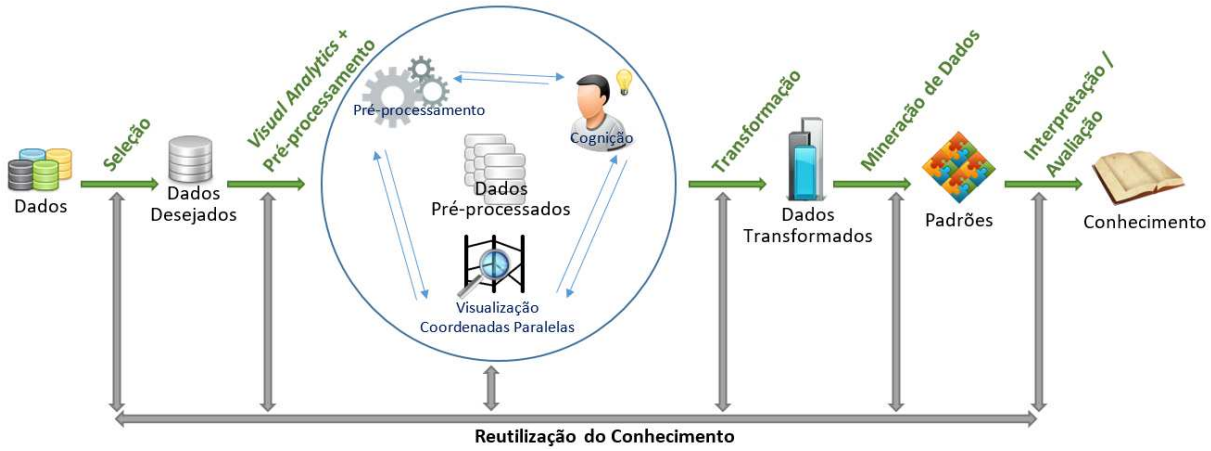
4 *VISUAL ANALYTICS* NO PRÉ-PROCESSAMENTO

A proposição de trabalho visa apresentar a integração entre as áreas de Descoberta de Conhecimento em Bases de Dados e Visualização de Informação para permitir que ferramentas de mineração de dados integrem o *Visual Analytics* a seu processo, com foco na exploração inicial aos dados e suporte à etapa de pré-processamento dos dados.

4.1 Uma abordagem envolvendo Coordenadas Paralelas

Conforme mencionado na Seção 2.1 o pré-processamento é a etapa do KDD que demanda o maior esforço. Assim, a proposição de trabalho visa incluir, diante de um processo de KDD consolidado, o *Visual Analytics* por meio da técnica de Coordenadas Paralelas na etapa de pré-processamento, que por usa vez, tem o intuito de auxiliar na exploração inicial dos dados, no conhecimento de domínio e no processo de pré-processamento dos dados. Como conhecimento entende-se discrepâncias, agrupamentos, tendências e relações entre os atributos. A técnica de Coordenadas Paralelas apresentada de forma interativa torna-se uma maneira amigável de se pré-processar e minerar dados, sendo que visualizações geralmente propiciam uma melhor compreensão do que está acontecendo com os dados, distinguindo-se de uma simples execução algorítmica. A Figura 12 apresenta o modelo proposto de integração do processo de KDD para que este inclua a análise visual

Figura 12: Processo de KDD com *Visual Analytics* integrado ao pré-processamento.



Fonte: Elaborado pelo autor.

Diferentemente da ideia inicial oferecida pela área de KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), onde o processo de visualização é dedicado à consolidação dos padrões descobertos ao final do processo de KDD, a ideia proposta quer efetivamente demonstrar as possibilidades da incorporação do *Visual Analytics* tanto para exploração e limpeza dos dados, como para aumentar a compreensão do domínio da informação, tornando a visualização um mecanismo de suporte à etapa de pré-processamento dos dados.

Ao contrário do que a área de visualização já oferece, ou seja, uma gama de técnicas e métodos de análise visual disponíveis em ferramentas onde o único propósito é visualizar os dados do domínio de interesse, este trabalho traz o *Visual Analytics*, por meio das Coordenadas Paralelas, integrado ao KDD permitindo o uso, a tomada de decisão e o suporte ao contato

inicial aos dados, tanto para fins de conhecimento de domínio como auxílio no pré-processamento, podendo atuar para fins didáticos e promocionais da visualização de informação (tratada aqui como o *Visual Analytics*) à mineração de dados, diferente unicamente da visualização de informação dedicada a domínios específicos.

Na figura acima é demonstrado como a formulação do modelo proposto atua, onde as duas primeiras etapas do modelo inicial não são alteradas, sendo elas responsáveis por consolidar um conjunto de dados único, como fonte de dados na etapa de pré-processamento. Esta etapa, que tem sua reformulação no presente trabalho, é alterada ao incluir fortemente um ser humano no processo de exploração e conhecimento do domínio, sendo que este interage com a visualização de Coordenadas Paralelas. Ressalta-se que os itens constantes na etapa de *Visual Analytics* sofrem apontamentos (setas) em ambos os sentidos, isto representa a capacidade colaborativa que as áreas podem oferecer mutuamente. Neste ponto, os dados são apresentados em uma visualização interativa que é disponibilizada para uma análise exploratória, o usuário, diante desta visualização, interage de forma dinâmica com os dados. Esta interação, auxiliada pelo conhecimento do contexto, faz com que se guie o pré-processamento de forma mais eficiente, sendo que a cognição pode apresentar uma sensibilidade subjetiva – conhecimento do contexto – em relação a algoritmos, isso amplia a detecção de padrões, mesmo que por discrepâncias, que poderiam ser indetectáveis de forma automática. Com os dados pré-processados, pode-se alimentar novamente a visualização e assim sucessivamente, este ciclo tem o intuito de enfatizar o conhecimento sobre os dados e refinar, a cada nova iteração, o conjunto de dados que seguirá para as etapas seguintes, também não alteradas perante ao modelo inicial.

O KDD também pode ser trabalhado de forma cíclica, onde um conhecimento extraído ao final de um processo pode ser reutilizado em outros conjuntos de dados ou a fim de melhorar os resultados de um processo já realizado. Esta análise circular foi nomeada de Reutilização do Conhecimento em referência ao *Feedback Loop* (KEIM; THOMAS, 2006) que inclui, além da reutilização, a comprovação dos resultados obtidos numericamente de forma visual.

4.2 WEKA

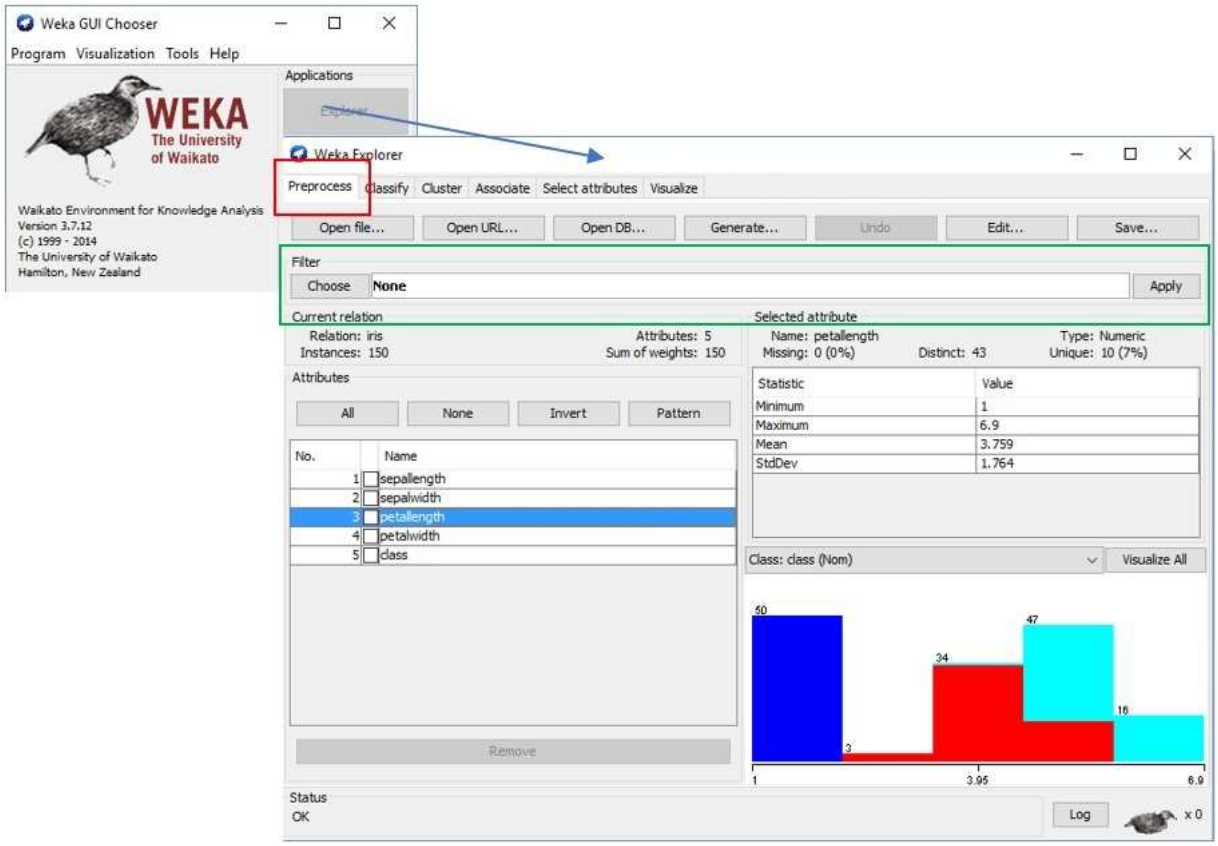
A escolha do WEKA como ferramenta de estudo de caso deve-se ao fato desta aplicação possuir uma vasta aceitação tanto no meio acadêmico como no meio corporativo, sua ativa comunidade, suas inúmeras citações e pelos milhões de *downloads* realizados, sendo contabilizado 1,4 milhão (HALL et al., 2009) e atualmente possuindo 7.631.187[6]. Além de sua intrínseca natureza integrada, permitindo a disponibilização de um ambiente unificado à extração de conhecimento e aprendizado de máquina para consolidação do WEKA.

Datada em 1992, esta ideia fundou-se sobre uma realidade onde os algoritmos de aprendizado encontravam-se escritos em diferentes linguagens de programação, disponíveis em diferentes plataformas e operados sobre inúmeros formatos de dados. Assim, foi visionado que o WEKA não só funcionasse como uma compilação de algoritmos, mas sim como uma base de código, ideia muito comum nos dias atuais, onde pesquisadores pudessem implementar novos algoritmos sem se preocupar com problemas periféricos como, por exemplo, a manipulação dos dados e validações de esquemas.

[6] Total de *downloads* até a data de 29/08/2016 coletados da página https://sourceforge.net/projects/weka, atual repositório do executável da aplicação.

O WEKA, segundo Hall et al. (2009), contém diversas interfaces gráficas que disponibilizam um fácil acesso as suas funcionalidades. Dentre as existentes, a de maior importância é o *Explorer* (Figura 13, seta azul) que é composto por diversos painéis (entende-se interfaces subordinadas ao *Explorer* organizadas por meio de abas) onde cada qual corresponde a uma diferente tarefa da mineração de dados. O primeiro painel (Figura 13, retângulo vermelho), de nome *Preprocess*, fornece um meio de carregar e transformar os dados, sendo que estes são manipulados pelos filtros (*filters*, Figura 13, retângulo verde).

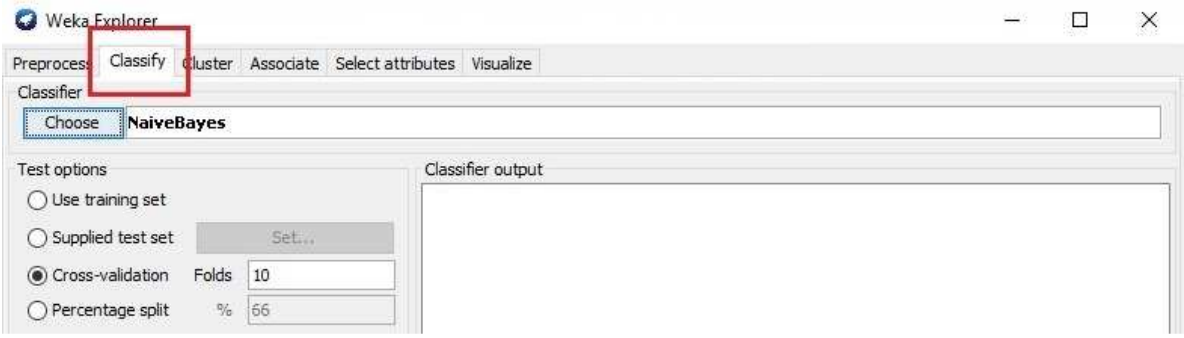
Figura 13: WEKA *Explorer*, aba *Preprocess* e opção *Filter*.



Fonte: Elaborado pelo autor com base na ferramenta WEKA versão 3.8.

Como segundo painel tem-se o *Classify* (Figura 14), composto por algoritmos de classificação e regressão. Por padrão, esta aba executa um *cross-validation* para o algoritmo que foi selecionado e utilizado sobre os dados preparados no painel anterior. Da mesma forma, este painel possibilita que subconjuntos de teste e treino sejam utilizados, habilitando assim a representação gráfica deste modelo como, por exemplo, utilizando árvores de decisão. Ademais é possível visualizar os erros de predição em uma matriz de dispersão como também possibilita a avaliação por meio de curvas ROC – *Receiver Operating Characteristic* – (FAWCETT, 2006) e outras curvas de *threshold*. Modelos também podem ser salvos e carregados neste painel.

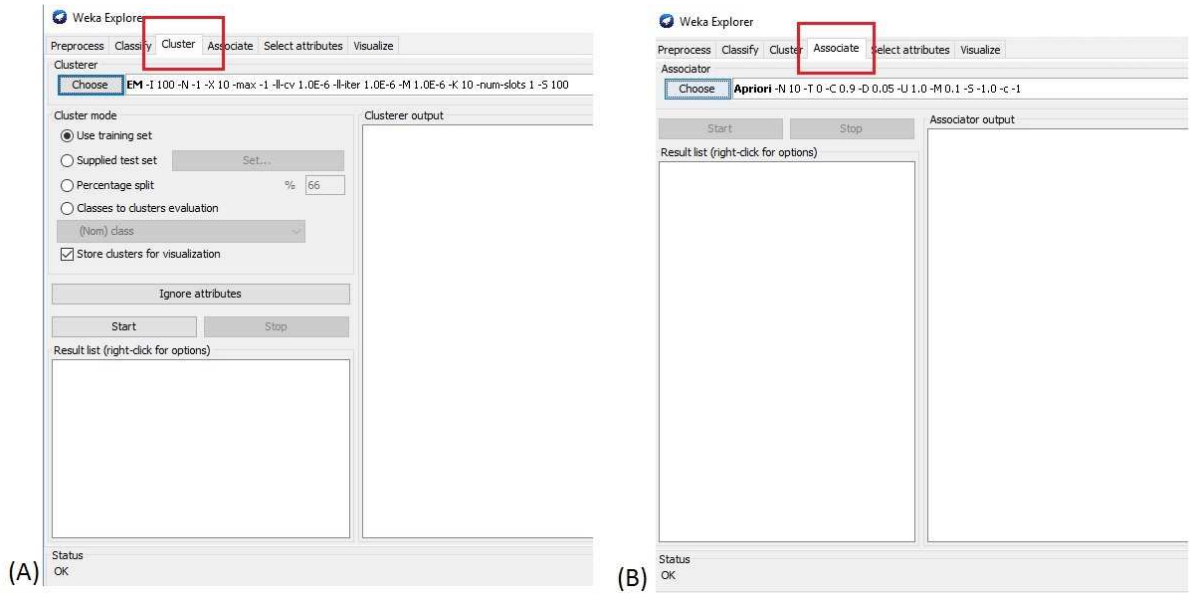
Figura 14: Aba *Classify*



Fonte: Elaborado pelo autor com base na ferramenta WEKA versão 3.8.

Algoritmos de aprendizado supervisionado e não supervisionado também são suportados pelo WEKA, sendo nomeados respectivamente como algoritmos de clusterização e métodos de mineração por regras associativas. Estes estão disponíveis nos painéis subsequentes. O primeiro método se encontra no terceiro painel de nome *Cluster* (Figura 15A), que permite que se use algoritmos de clusterização sobre os dados carregados no *Preprocess*. No quarto painel, chamado *Associate* (Figura 15B), encontram-se os algoritmos clássicos de mineração de regras associativas como o Apriori.

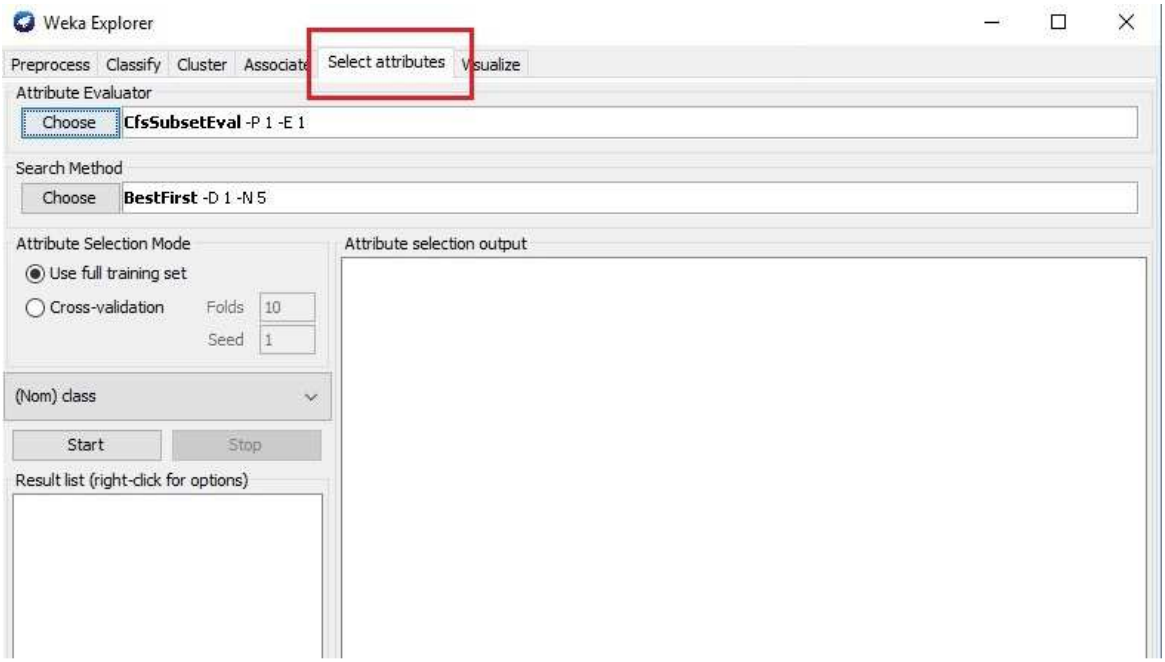
Figura 15: Abas *Cluster* e *Associate*



Fonte: Elaborado pelo autor com base na ferramenta WEKA versão 3.8.

Porém uma das etapas práticas mais importantes na mineração de dados é a tarefa de se identificar quais atributos contidos nos dados apresentam maior poder preditivo. Para isto, Hall et al. (2009) citam o quinto painel existente no WEKA, sendo este inteiramente dedicado para esta finalidade. Nomeado como *Select attributes* (Figura 16), este painel possui uma ampla variedade de algoritmos para a identificação dos atributos mais significativos de um conjunto de dados, possibilitando também que diferentes métodos de procura sejam combinados com diferentes critérios de avaliação.

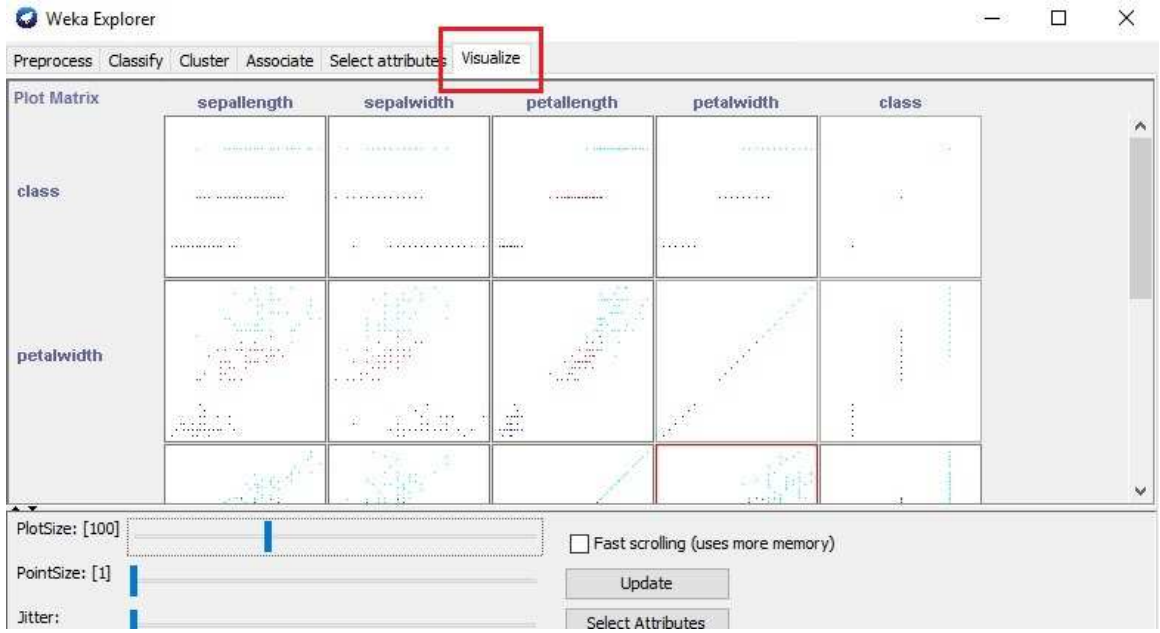
Figura 16: Aba *Select attributes*



Fonte: Elaborado pelo autor com base na ferramenta WEKA versão 3.8.

Por último o WEKA disponibiliza o painel chamado *Visualize* (Figura 17), que provê uma visualização da matriz de dispersão, que é colorida de acordo com o último atributo do conjunto de dados, geralmente reservado para representar as classes de interesse. Esta matriz possibilita algumas interações como por exemplo: restringir o número de amostras apresentadas, selecionar quais atributos serão apresentados na matriz e possibilidade de alterar o diâmetro dos círculos que representam as amostras no gráfico.

Figura 17: Aba *Visualize*



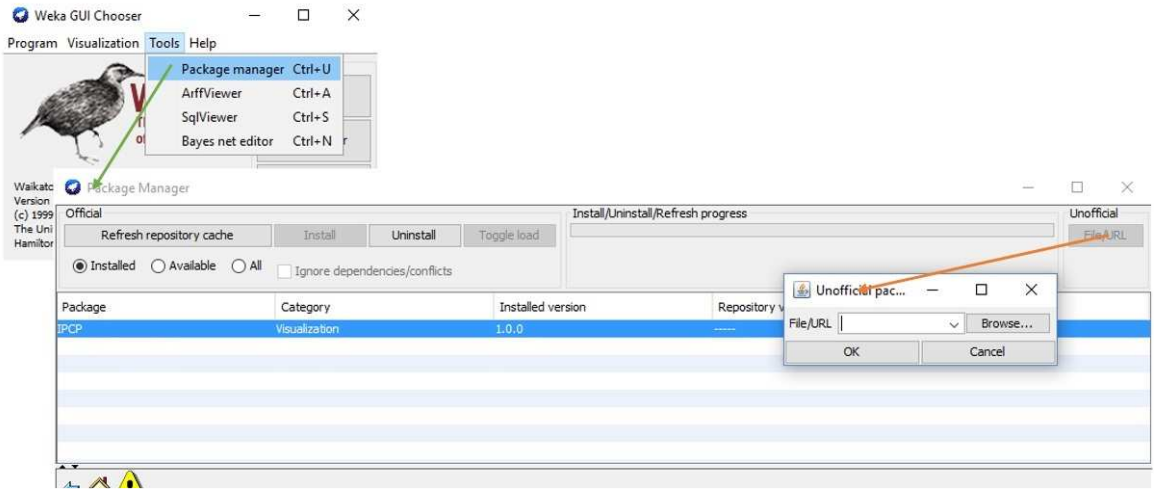
Fonte: Elaborado pelo autor com base na ferramenta WEKA versão 3.8.

Além de possuir todos os painéis citados, a ferramenta os trabalha de forma integrada, sendo assim, ações realizadas em cada uma das abas podem facilmente ser refletidas nas demais sem a necessidade de se utilizar outras ferramentas, conversores de dados ou algoritmos de terceiros. Como reflexo da integração das abas pode-se citar o exemplo de que: toda a alteração do conjunto de dados por meio da aba de *Preprocess* transforma o atual conjunto, sendo que o mesmo pode ser utilizado no instante seguinte nas abas subsequentes. Esta possibilidade torna o WEKA uma ferramenta nativamente interativa no contexto de trânsito de informações entre as diversas etapas do KDD. Neste sentido, é possível configurar um modelo automático de mineração que pode ser reutilizado em outros conjuntos de dados.

Por ser uma ferramenta de código aberto, o WEKA expande vastamente, em termos de desenvolvimento de *software*, as possibilidades de extensão e agregação de novas funcionalidades. Esta extensão já era explorada em sua versão 3.6, porém não havia nada que possibilitasse a integração de novas funcionalidades de forma facilitada, ou seja, na grande maioria das vezes, adicionar novas funções (ou ampliar as existentes) ao WEKA implicava na recompilação parcial ou total de seu código fonte.

A versão 3.8 estável do WEKA, ao contrário de sua antecessora (3.6), mudou este cenário na medida em que um gerenciador de pacotes (pode-se interpretar como gerenciador de *plug-ins*, Figura 18, seta verde) foi adicionado.

Figura 18: WEKA 3.8 e seu gerenciador de pacotes



Fonte: Elaborado pelo autor com base na ferramenta WEKA versão 3.8.

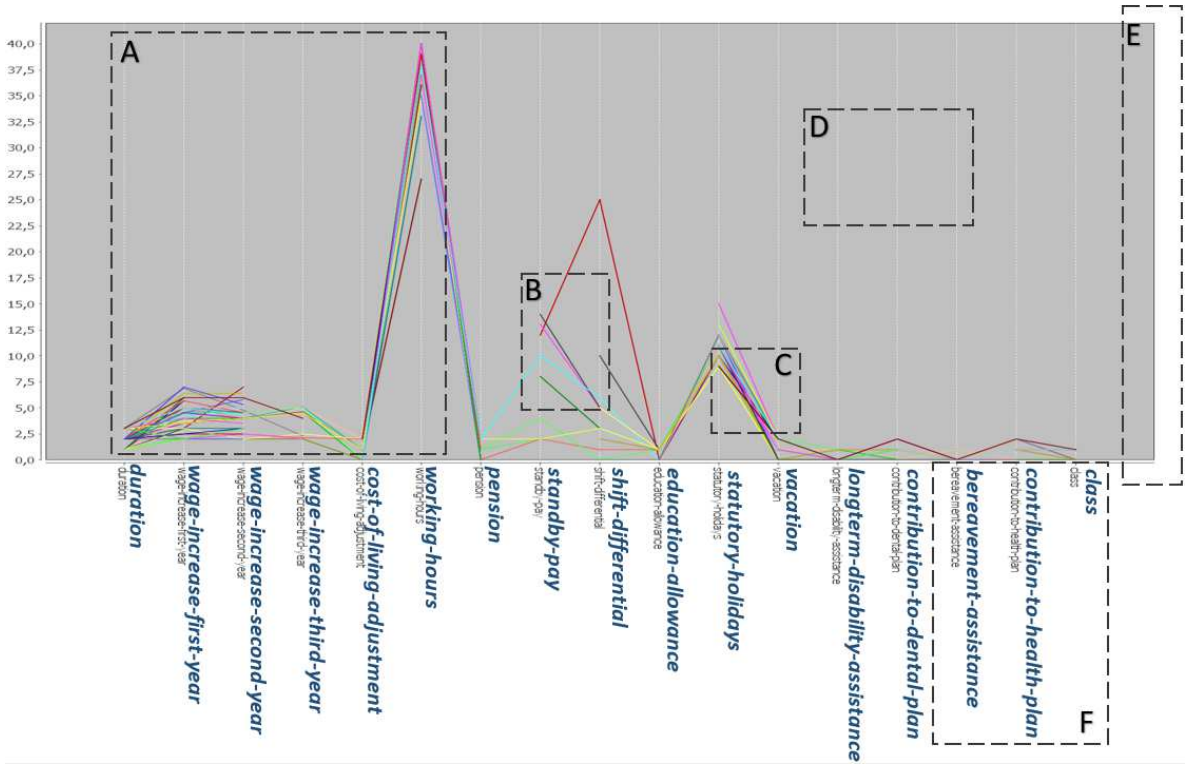
Este gerenciador possibilita a extensão da ferramenta WEKA por meio de pacotes (aplicações) de terceiros, isto é, torna-se realizável a instalação de novas funcionalidades a partir de um repositório. Não obstante, pode-se utilizar neste mesmo gerenciador a instalação de pacotes "não oficiais" (Figura 18, seta laranja), que consistem em aplicações não homologadas pelo WEKA, logo, não existente em seu repositório oficial.

Atualmente há disponível uma extensão que se propõe a apresentar a visualização das Coordenadas Paralelas[7], porém a versão disponibilizada no repositório do WEKA se demonstra limitada em todos os aspectos pertinentes ao *Visual Analytics*, sendo que a versão oferecida é

[7] PCP: *Parallel Coordinates Plot*, *plugin* já oferecido para *download* no repositório do WEKA. Fonte: https://sourceforge.net/projects/wekapcp/. Este *plugin* é o criticado na seção 4.2 do presente trabalho.

formada por um único gráfico estático, sem a possibilidade de interação. A figura abaixo pontua as limitações observadas.

Figura 19: Extensão de Coordenadas Paralelas encontrada no repositório do WEKA.



Fonte: Elaborado pelo autor com base na ferramenta WEKA versão 3.8.

Lista-se como limitações: inexistência de normalização das dimensões (Figura 19A), onde-se vê claramente discrepâncias entre máximos e mínimos; linhas “cortadas”, situação não concebida à visualização de Coordenadas Paralelas (Figura 19B); inúmeras cores que não permitem que se perceba o fluxo e a densidade do número de amostras (Figura 19C); impossibilidade de permutação dos eixos, por ser uma ação do usuário é representado de forma ilustrativa pela Figura 19D; interface que comprime a distância entre os eixos, não se auto ajustando a conjuntos de dados com mais ou menos atributos, sendo uma ação da interface é representada de forma ilustrativa pela Figura 19E; o nome dos atributos também são renderizados na imagem, porém em uma resolução muito baixa, onde na Figura 19F foram reescritos ao lado das originais, com o intuito de melhorar sua apresentação no presente trabalho e por apresentarem resolução semelhante na própria ferramenta.

Reforça-se o item A mencionado, que é a inexistência de normalização das dimensões, ou seja, graficamente eixos têm diferentes valores de máximo e mínimo. Em B verifica-se linhas “cortadas”, isso não existe nas Coordenadas Paralelas. Seguido pelo C que ressalta as inúmeras cores utilizadas, evitando que se perceba o fluxo e a densidade do número de amostras. D é pertinente à interface, pois a janela se comprime ao trabalhar com mais eixos ao invés de oferecer uma barra de rolagem. E por último, a letra E sinaliza a impossibilidade de permutação dos eixos.

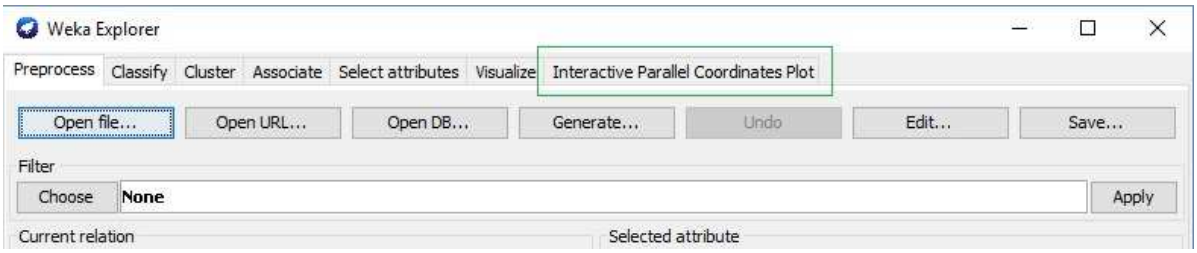
4.3 Integração das Coordenadas Paralelas com a ferramenta WEKA

A adição de novas funcionalidades no WEKA é factível por dois fatores: a) o WEKA é uma ferramenta de código aberto, facilitando assim o entendimento de seu funcionamento, construção e extensão; b) a partir da versão 3.7 o WEKA consta com um gerenciador de pacotes (uma espécie de gerenciados de *plugins*).

No gerenciador de pacotes do WEKA (Figura 18) foi disponibilizado de forma oficial o plugin apresentado pelo presente trabalho, ele pode ser localizado pela sigla *IPCP*, onde é possível instalá-lo de forma automática, não necessitando das funcionalidades de instalação não oficiais. Uma vez adicionado, este pacote será integrado ao WEKA, sendo listado no mesmo local como "instalado", podendo ser removido ou atualizado caso necessário.

O resultado dessa integração tem como objetivo adicionar uma nova aba as demais já existentes no WEKA *Explorer*, conforme figura abaixo.

Figura 20: Abas que compõem o WEKA *Explorer*.



Fonte: Elaborado pelo autor.

Esta aba (Figura 20, retângulo verde), por estar interconectada às funcionalidades do WEKA, terá seu funcionamento dinâmico, ou seja, todas as manipulações que afetarem os dados de entrada serão refletidas para a visualização de Coordenadas Paralelas existente nesta nova aba.

4.4 Desenvolvimento das Coordenadas Paralelas

Para que a construção das Coordenadas Paralelas como um *plugin* do WEKA fosse possível, uma série de tecnologias foram empregadas no seu desenvolvimento. Como ferramentas de suporte à construção citam-se o Java[8] e o NetBeans[9]. E como fonte de consulta, ideias e meios de se estruturar uma aplicação de visualização utilizou-se o XDAT[10], ParVis[11] e PCP[7]. Como resultado foi produzia uma aplicação com código fonte constituído por 21 classes utilizando-se fundamentalmente das bibliotecas *Swing* e *AWT*, ambas nativas do Java, para a criação da interface. A primeira utilizada para dispor os painéis laterais e inferiores, sendo

[8] Java: Linguagem de programação orientada a objetas, *open source*, compilada para *bytecodes* e interpretada por uma máquina virtual. Fonte: https://www.java.com/pt_BR/.

[9] NetBeans: IDE de desenvolvimento de aplicações multiplataforma e *open source*. Fonte: http://netbeans.org/.

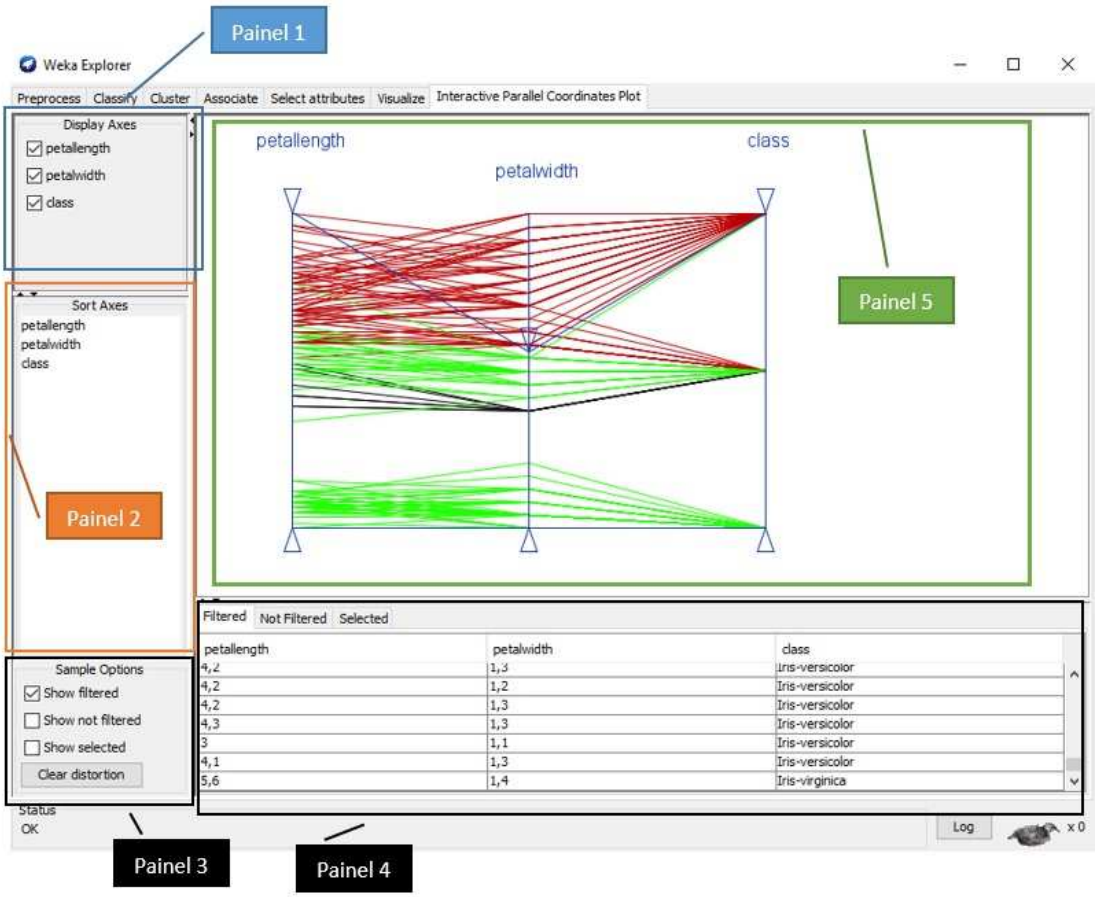
[10] XDAT: Aplicação que disponibiliza, dentre algumas visualizações, a de Coordenadas Paralelas no formato *open source* e escrita em Java. Fonte: http://www.xdat.org/.

[11] ParVis: Aplicação exclusiva para Coordenas Paralelas. Fonte: https://github.com/eagereyes/ParVis.

aplicada também para a apresentação da informação em forma de listas ou tabelas, a segunda para a criação da visualização – como um gráfico – em si. Esta, sendo posteriormente disponibilizada de forma pública e como código livre, em um repositório online *https://github.com/glaubercini/ipcp*.

O resultado deste desenvolvimento pode ser visto na figura abaixo, com a aba nomeada como *Interactive Parallel Coordinates Plot*, onde todos os seus painéis são detalhados posteriormente.

Figura 21: Painéis existentes no *plugin* de Coordenadas Paralelas



Fonte: Elaborado pelo autor.

Os 5 painéis, como mostra a Figura 21, existentes no *plugin* desenvolvido têm como funções:

- Painel 1: Controlar quais atributos serão apresentados na visualização de Coordenadas Paralelas, ao marcar ou desmarcar as caixas de seleção, automaticamente a visualização é renderizada;
- Painel 2: Responsável por possibilitar a permutação dos eixos, este painel oferece esta função por meio de “arrastar e soltar”, toda interação neste painel também é refletida de forma automática para a visualização;
- Painel 3 e 4: Estes painéis são complementares, ou seja, o Painel 3 é responsável por habilitar ou não a construção das tabelas que serão desenhadas no Painel 4. A construção do Painel 3 foi motivada pelo fato de tentar melhorar o desempenho da aplicação como um todo, evitando que a qualquer interação do usuário as tabelas fossem recriadas. Desta forma, elas somente são recriadas quando necessário. Além disto, no Painel 4 é possível

clicar com o botão direito do mouse e exportar a tabela que está sendo visualizada para o formato ARFF (formato nativo do WEKA), criando assim um novo conjunto de dados fruto da análise visual. No Painel 3 ainda se encontra a opção de limpar as distorções realizadas;

- Painel 5: A visualização em si consta neste painel, nele é possível arrastar os triângulos existentes nas extremidades de cada eixo para filtrar as amostras (linhas), além de possibilitar ressaltar amostras ao passar o mouse “por cima” das linhas. Para cada uma das situações uma cor foi atribuída. Amostras em verde estão dentro da área filtrada, amostras em vermelho estão fora da área filtrada e as em preto são as selecionadas ao clicar (ou passar o mouse “por cima”). Os eixos foram coloridos de azul e o fundo de branco. A escolha das cores são, em sua essência, as três cores básicas (vermelho, verde e azul) e as cores opostas preto e branco. A escolha é motivada pelo fato de que desta forma não há confusão visual, as cores são diferentes, evitando que o utilizador possa se confundir (por exemplo dois tons de cores parecidos que significariam coisas diferentes).

5 DEMONSTRAÇÃO DA APLICABILIDADE

Ao longo do presente trabalho foram apresentados os conceitos de KDD e Visualização de Informação que, integrados, deram origem ao *Visual Analytics*. De forma a transcender esta conceituação integrada, buscou-se não somente torná-la realidade, mas também que ela fosse de fácil acesso e usabilidade a todos os interessados nesta área de pesquisa. Ferramentas de KDD foram citadas ao longo do trabalho, outras de visualização, e outras ainda focadas nas Coordenadas Paralelas. Todos estes artefatos são importantes à área da pesquisa da Análise Visual, porém não a tornavam acessível por não oferecer um ambiente integrado e conhecido na literatura.

Ao se disponibilizar uma visualização (dentre as inúmeras possíveis e existentes) sobre uma ferramenta de KDD amplamente conhecida, abre-se uma nova possibilidade, tanto para a disseminação do conceito de *Visual Analytics* quanto para o surgimento de novas ideias a partir deste primeiro passo.

Não suficiente à disseminação, o *plugin* proposto vem como um alicerce para que o *make sense of data* seja cada vez mais possível no WEKA. Onde agora é praticável carregar conjuntos de dados e conhece-los por outras perspectivas, interagir com eles e, quando necessário, alterar a sua estrutura inicial (permutando eixos, removendo eixos ou amostras). Conhecer os dados e alterar sua formulação traz novas visões de como pode-se extrair conhecimento deles por meio do KDD, sendo que, por muitas vezes, somente a interatividade disponibilizada já contribui para este aspecto.

Embora cenários mais complexos de utilização se vislumbrem, com domínios de dados interessantes, sua utilização foi descartada por necessitar de um conhecimento especializado de domínio mais refinado. Desta forma, para demonstrar algumas funcionalidades deste *plugin* dois cenários de dados clássicos (Auto MPG e Iris) foram utilizados.

Destes cenários foi criado um momento para cada um. Como momento entende-se uma possível exploração dos dados e de como pode-se conhecê-los melhor. Os momentos são demonstrados nas subseções seguintes.

5.1 Conjunto de Dados Auto MPG

O conjunto Auto MPG (QUINLAN, 1993), possui dados sobre carros, que se referem ao consumo, em milhas por galão, de combustível em cidades. Em sua composição original, este conjunto possui 9 atributos (sendo o primeiro a classe) e tem 406 amostras, onde 8 delas não possuem o valor da classe informado.

Para o presente trabalho somente os atributos listados na Tabela 1 foram utilizados, onde o atributo original “car name” foi removido por possuir valores discretos e únicos, não colaborando para o momento formulado, além de também desconsiderar as 8 amostras que não possuem classe.

Tabela 1: Atributos utilizados no conjunto de dados Auto MPG.

Atributo	Descrição	Amostras com valores inexistentes	Contínuo/Discreto
mpg	Milhas por galão (consumo)	0 (zero)	Contínuo
cylinders	Cilindros	0 (zero)	Discreto
displacement	Cilindradas	0 (zero)	Contínuo
horsepower	Cavalos de potência	6 (seis)	Contínuo
weight	Peso	0 (zero)	Contínuo
acceleration	Aceleração	0 (zero)	Contínuo
model	Modelo (Ano)	0 (zero)	Discreto
origin	Origem	0 (zero)	Discreto

Fonte: Elaborado pelo autor.

Este momento, contido na Figura 22, foi gerado para demonstrar uma possibilidade inicial de análise dos dados. Primeiramente utilizou-se o algoritmo não supervisionado “ReplaceMissingValues” para atualizar com a média os valores numéricos não existentes nas amostras. Após uma análise visual sobre todo o conjunto, com o intuito de notar possíveis relações entre os atributos, optou-se por não exibir o atributo “model” e por permutar o atributo “horsepower” para que este ficasse ao lado de “acceleration”. Após, filtrou-se através do eixo de “horsepower” até a altura indicada na Figura 22 onde uma leve troca no comportamento da relação entre os atributos foi visualmente notada.

Figura 22: Conjunto Auto MPG sendo analisado no *plugin* proposto

Fonte: Elaborado pelo autor.

Este primeiro momento demonstrou, em linhas gerais, que carros com um número maior de cavalos de potência (amostras vermelhas) também são os mais pesados e tem em sua grande maioria a origem o número 1 (um, USA), constatação obtida ao filtrar o atributo “horsepower"

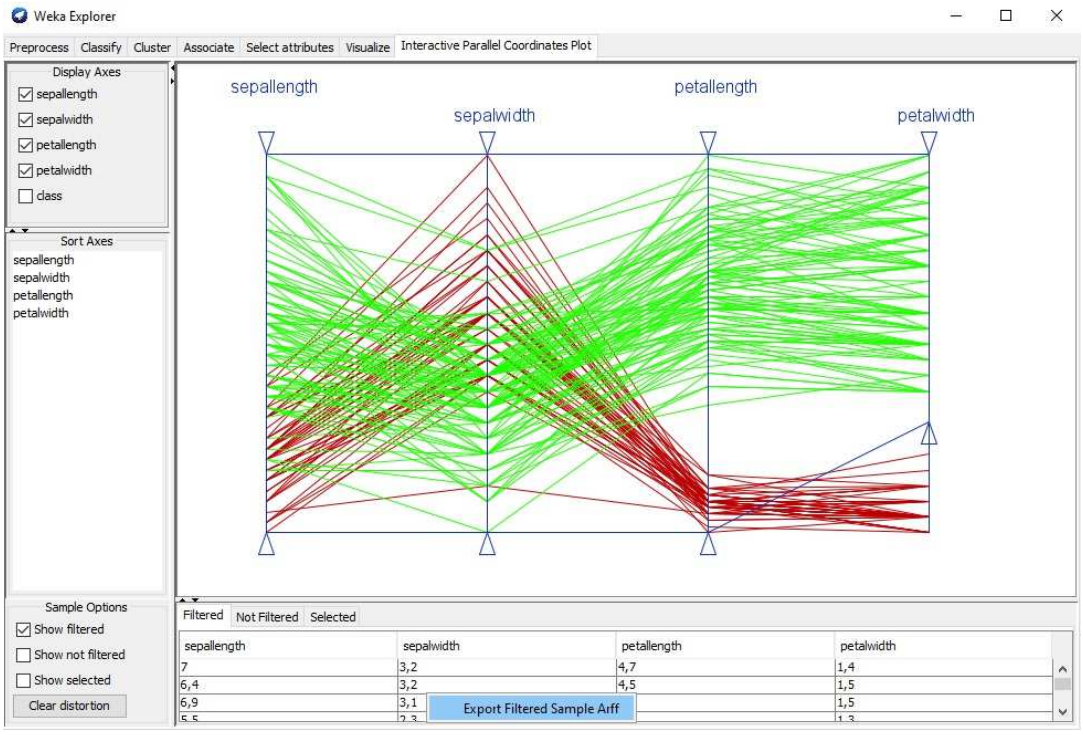
até a indicação da flecha azul existente no eixo correspondente (a flecha foi posta com o intuito de enfatizar até que altura utilizou-se o filtro, que originalmente é representado por um triângulo azul, que é difícil de ser reconhecido na proporção da imagem utilizada). As 3 linhas em preto (também enfatizadas) representam sugestões de algumas possíveis descobertas que uma simples perspectiva da análise visual pode guiar a investigação, elas referem-se as amostras que apresentaram um comportamento diferente das demais instâncias, o que pode sinalizar possíveis discrepâncias.

Constata-se que a amostra selecionada de letra A, apresenta um declive acentuado entre os atributos “displacement” e “acceleration”, situação que não é aferida visualmente em outros itens deste conjunto. Na de letra B, é possível verificar que esta amostra, pertencente ao agrupamento de não filtradas (em vermelho), apresenta a mais elevada relação entre “weight” e “acceleration”, tem o mesmo valor no atributo “cylinders”, com exceção da letra C, e possui o aclive menos acentuado entre “acceleration” e “horsepower", diferenciando-se das demais amostras do mesmo grupo. Por fim, a de letra C, também pertencente ao grupo das não filtradas, teve o atributo “cylinders” com valor diferente dentre as amostras do mesmo agrupamento. É possível, neste *plugin*, aprofundar-se nos detalhes que cada uma destas amostras pode oferecer de forma interativa, sendo que também são apresentados os valores reais de cada instância na tabela de valores abaixo da visualização.

5.2 Conjunto de Dados Iris

Para o segundo momento, o conjunto de dados escolhido foi o Iris (FISHER, 1938), o qual remete-se a botânica sendo constituído por 150 amostras e 2 pares de atributos: comprimento e largura da sépala e comprimento e largura da pétala. Um quinto atributo completa o conjunto, sendo este a classe à qual a planta pertence.

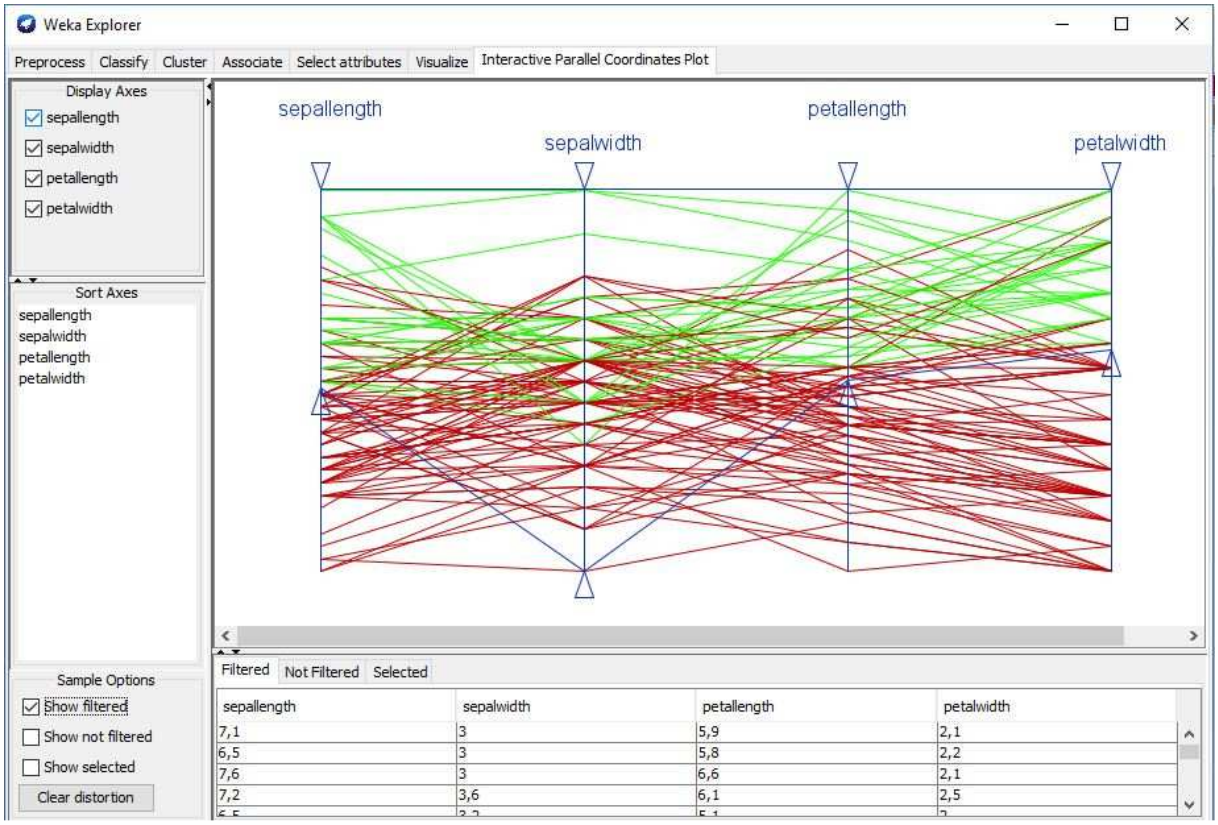
Figura 23: Conjunto de dados Iris com primeiro filtro aplicado



Fonte: Elaborado pelo autor.

Os demais atributos da Figura 23 não foram permutados. As amostras filtradas (em vermelho) representam a classe que é claramente separável linearmente (classe “Iris-versicolor”). Desta forma, optou-se por explorar somente as demais amostras que não pertencem a esta classe, para isto utilizou-se a opção de exportar os dados filtrados (amostras em verde) pelo menu *Export Filtered Sample Arff*. Esta exportação de dados originou uma análise mais aprofundada deste momento que pode ser visualizada na Figura 24.

Figura 24: Análise aprofundada do conjunto de dados Iris somente com as classes “Iris-setosa” e “Iris-virginica” exportado pelo *plugin* proposto



Fonte: Elaborado pelo autor.

Este segundo momento, agora com a exploração focada somente nos dados que não puderam ser separados linearmente, apresenta uma tendência: os maiores valores de comprimento da sépala e pétala, “sepallength” e “petallength” respectivamente, e a largura da pétala “petalwidth” tendem a formar um grupo de valores maiores. Na Figura 24, pode-se observar que filtrando estes três atributos até determinada altura é possível dividir, em sua maioria, as amostras que tendem a ficar no grupo verde, apresentando valores maiores, e no grupo vermelho, de valores menores.

Verifica-se também, que o atributo largura da sépala “sepalwidth”, visualmente, não colabora para esta análise inicial, podendo ser suprimido da visualização gerada mediante a escolha do analista.

5.3 Considerações

Os dois momentos apresentados são pontuais e limitados, visto que para apresentá-los o trabalho dispõe de apenas figuras. Todavia é suficiente para demonstrar o potencial integrado do *plugin* sendo que, ao dispor de tal artefato na etapa de pré-processamento pode-se, como demonstrado no primeiro momento (Auto MPG, Figura 22), utilizar a visualização após resolver os valores não existentes com algoritmos já existentes no WEKA e após trabalhar de forma exploratória o *make sense of data*.

Não se limitando a isto, o *plugin* ainda possibilita uma maneira facilitada de permutar os atributos, permitindo também que os dados possam ser separados em três grupos: a) filtrados; b) não filtrados; c) selecionados. A permutação, conjuntamente com os três grupos de amostras pode originar novos conjuntos de dados permitindo que estes possam ser utilizados como um novo conjunto de dados na ferramenta WEKA, situação verificada na Figura 23. Esta característica habilita uma análise mais aprofundada das características que o especialista, no momento da análise, necessita avaliar.

Tratando-se de uma extensão interativa, genérica e que tem como um dos principais limitadores a subjetividade humana, outras análises visuais poderiam ser realizadas sobre os conjuntos utilizados nas demonstrações deste capítulo. Conclusões poderiam ser formuladas caso, por exemplo, especialistas – um conhecedor de automóveis e um biólogo –, se munissem de tal ferramenta. Colaborando, possivelmente, com outras perspectivas que, no entanto, estão fora do escopo deste trabalho.

6 CONCLUSÕES

O *Visual Analytics* é retratado na literatura, por muitas vezes, como uma área paralela à extração de conhecimento. Esta utilização desconexa entre as áreas restringe seus estudos a contextos específicos, distanciamento este que acaba por não propiciar um ambiente que, de fato, efetive a utilização da Visualização de Informação conjuntamente com o KDD.

Verifica-se também na literatura, que quando tais áreas são integradas, as ferramentas analisadas não compatibilizam o formato de dados aceitos por cada uma. Impedindo que uma Análise Visual conjunta com qualquer análise de dados automática seja feita de maneira natural, em um mesmo ambiente computacional, obrigando que os dados sejam convertidos e reconvertidos a cada nova necessidade de análise, o que pode causar resistência por parte dos usuários sobre a utilização de ambas tecnologias.

A fim de especificar um processo completo, que abranja ambas as áreas (visuais e automáticas), o presente trabalhou apresentou na subseção 4.1 uma abordagem que padroniza a inclusão de técnicas visuais na etapa de pré-processamento do KDD, com o intuito de ampliar o conhecimento do contexto no qual as informações dos conjuntos de dados estão inseridas, expandindo as possibilidades do *make sense of data*. Apresentar um processo unificado, sugere um modelo que possa ser utilizado como ponto de partida para outras pesquisas relacionadas ao *Visual Analytics*, evitando que desconexões entre processos possam se perpetuar, somando, de fato, à literatura.

Não obstante, este trabalho entrega, simultaneamente ao modelo, um *plugin* para a ferramenta WEKA, onde busca-se não somente propor e demonstrar tal modelo exclusivamente de forma teórica, mas efetivá-la na forma de um módulo de *Visual Analytics* ligado ao KDD. A entrega deste *plugin* para a comunidade facilita o conhecimento da área da Análise Visual, amplia sua popularização e proporciona um ponto de partida genérico e comum sobre uma ferramenta amplamente conhecida e utilizada para o KDD. Pontua-se que este *plugin* está no repositório oficial do WEKA, utilizando o nome de IPCP, tornando sua instalação e utilização praticável por qualquer usuário do WEKA 3.7 em diante, e mais, ele é disponibilizado no formato de código aberto.

Os dois artefatos apresentados (modelo e *plugin*) corroboram para a eficácia que a subjetividade humana pode proporcionar, característica pouco explorada no KDD. De forma complementar, o Capítulo 5 demonstra, diante de dois conjuntos de dados clássicos, que é possível traçar características visualizando sua distribuição sobre a visualização interativa de Coordenadas Paralelas. A colaboração é ressaltada ao passo que algoritmos de correção de dados inexistentes (contidos de forma nativa no WEKA) são utilizados previamente à renderização das Coordenadas Paralelas. Esta mutualidade faz com que não somente o processo se demonstre, mas que a ferramenta, de fato, cumpre os propósitos de: a) integrar as tecnologias; b) facilitar análises circulares que transitem entre algoritmos automáticos (característicos do KDD) e visualizações interativas; c) disponibilizar tal tecnologia de forma nativa, por meio de um gerenciador de pacotes.

Apesar das facilidades e benefícios mencionados que o presente trabalho trás na análise de dados, muito há que evoluir nesse sentido. Assim, como trabalhos futuros são propostos, como a elaboração de uma metodologia para auferir possíveis variações de desempenho, ocasionadas pela subjetividade humana, por meio da análise visual, ao KDD resultando assim em uma métrica.

Outra possível extensão proposta trata da resolução da ineficiência à renderização de conjuntos de dados densos, ocasionada pelas bibliotecas 2D oferecidas de forma nativa pelo Java. Assim, a construção das visualizações sobre tecnologias que utilizem recursos gráficos mais avançados, utilizando, por exemplo, a computação gráfica beneficiada por *hardwares* específicos poderia colaborar no sentido de agilizar a projeção visual dos dados.

Por fim, a possibilidade de agregar outras formas de Visualização de Informações, além das Coordenadas Paralelas, para a realização do Visual Analytics. Abrindo novos horizontes por meio de extensões ao *plugin* apresentado, como, por exemplo, a adição de outras técnicas visuais exploratórias, uma vez que seu código é aberto.

REFERÊNCIAS

ADRIAANS, P.; ZANTINGE, D. **Data Mining**. Addison Wesley Longman, Harlow, Inglaterra, 1996.

ANDRIENKO, G.; ANDRIENKO, N.; BAK, P.; KEIM, D.; WROBEL, S. **Visual Analytics of Movement**. Springer Science & Business Media, 2013.

ANDRIENKO, G.; ANDRIENKO, N.; JANKOWSKI, P.; KEIM, D.; KRAAK, M. J.; MACEACHREN, A.; WROBEL, S. **GeoVisual Analytics, Time to Focus on Time.** Information Visualization, v. 13, n. 3, p. 187-189, 2014.

ARTERO, A. O.; DE OLIVEIRA, M. C. F.; LEVKOWITZ, H. **Uncovering Clusters in Crowded Parallel Coordinates Visualizations**. Information Visualization, 2004. INFOVIS 2004. IEEE Symposium. IEEE, 2004. p. 81-88.

BERTHOLD, M. R.; HALL, L. O. **Visualizing Fuzzy Points in Parallel Coordinates**. Fuzzy Systems, IEEE Transactions, v. 11, n. 3, p. 369-374, 2003.

BLAAS, J.; BOTHA, C. P.; POST, F. H. **Extensions of Parallel Coordinates for Interactive Exploration of Large Multi-timepoint Data Sets**. Visualization and Computer Graphics, IEEE Transactions, v. 14, n. 6, p. 1436-1451, 2008.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Information Visualization. Readings in Information Visualization: Using Vision to Think**, Morgan Kaufmann Publishers, San Francisco, California, USA, p. 1-34, 1999.

CARPENDALE, S. **Evaluating Information Visualizations.** Information Visualization. Springer Berlin Heidelberg. p. 19-45, 2008.

CHAMBERS, J. M.; CLEVELAND, W. S.; KLEINER, B.; TUKEY, P. A. **Graphical Methods for Data Analysis**, Wadsworth Statistics/Probability Series, Monterey, CA, 1983.

CHEN, C. **Information Visualization: Beyond the Horizon**. Springer Science & Business Media, 2006.

CLEVELAND, W. S. **Visualizing Data**. Hobart Press, Summit, New Jersey, 1993.

CRAFT, B.; CAIRNS, P. **Beyond Guidelines: What Can We Learn From The Visual Information Seeking Mantra?**. Information Visualisation, 2005. Proceedings. Ninth International Conference. IEEE, p. 110-118, 2005.

FAWCETT, T. **An Introduction to ROC Analysis**. Pattern Recognition Letters, v. 27, n. 8, p. 861-874, 2006.

FAYYAD, U. M.; WIERSE, A.; GRINSTEIN, G. G. **Information Visualization in Data Mining and Knowledge Discovery**. Morgan Kaufmann, 2002.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. AI Magazine, v. 17, n. 3, p. 37-54, 1996.

FISHER, R. A. **The Use of Multiple Measurements in Taxonomic Problems**. Annals of Eugenics, v. 7, n. 2, p. 179-188, 1936.

GARCÍA, S.; LUENGO, J.; HERRERA, F. **Tutorial on Practical Tips of the Most Influential Data Preprocessing Algorithms in Data Mining**. Knowledge-Based Systems, 98, p. 1-29, 2016.

GUCKENHEIMER, J.; HOLMES, P. **Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields**. Applied Mathematical Sciences, v. 42, 1983.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **The WEKA Data Mining Software: An Update**. ACM SIGKDD Explorations Newsletter, v. 11, n. 1, p. 10-18, 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques 3nd**. Elsevier, 2011.

HAO, M. C.; MARWAH, M.; JANETZKO, H.; DAYAL, U.; KEIM, D. A.; PATNAIK, D.; RAMAKRISHNAN, N.; SHARMA, R. K. **Visual Exploration of Frequent Patterns in Multivariate Time Series**. Information Visualization, v. 11, n. 1, p. 71-83, 2012.

HASENAUER, J.; HEINRICH, J.; DOSZCZAK, M.; SCHEURICH, P.; WEISKOPF, D. **A Visual Analytics Approach for Models of Heterogeneous Cell Populations**. EURASIP Journal on Bioinformatics and Systems Biology, v. 2012, n. 1, p. 1-13, 2012.

HASENAUER, J.; WALDHERR, S.; DOSZCZAK, M.; RADDE, N.; SCHEURICH, P.; ALLGÖWER, F. **Identification of Models of Heterogeneous Cell Populations From Population Snapshot Data**. BMC Bioinformatics, v. 12, n. 1, 2011.

HASENAUER, J.; WALDHERR, S.; DOSZCZAK, M.; SCHEURICH, P.; RADDE, N.; ALLGÖWER, F. **Analysis of Heterogeneous Cell Populations: A Density-based Modeling and Identification Framework**. Journal of Process Control, v. 21, n. 10, p. 1417-1425, 2011.

HEARST, M. A.; DUMAIS, S. T.; OSMAN, E.; PLATT, J.; SCHOLKOPF, B. **Support Vector Machines**. Intelligent Systems and their Applications, IEEE, v. 13, n. 4, p. 18-28, 1998.

HEINRICH, J.; WEISKOPF, D. **State of the Art of Parallel Coordinates**. STAR Proceedings of Eurographics, v. 2013, p. 95-116, 2013.

INSELBERG, A. **The Plane With Parallel Coordinates**. The Visual Computer, v. 1, n. 2, p. 69-91, 1985.

KEIM, D. A. **Information Visualization and Visual Data Mining**. Visualization and Computer Graphics, IEEE Transactions, v. 8, n. 1, p. 1-8, 2002.

KEIM, D. A. **Visual Exploration of Large Data Sets**. Communications of the ACM, v. 44, n. 8, p. 38-44, 2001.

KEIM, D. A.; KOHLHAMMER, J.; ELLIS, G.; MANSMANN, F. (Ed.). **Mastering the Information Age-solving Problems with Visual Analytics**. Florian Mansmann, 2010.

KEIM, D. A.; MANSMANN, F.; SCHNEIDEWIND, J.; ZIEGLER, H. **Challenges in Visual Data Analysis**. IEEE Tenth International Conference on Information Visualization, London, UK, p. 9-16, jul. 2006.

KEIM, Daniel A.; THOMAS, Jim. **Scope and Challenges of Visual Analytics**. Tenth International Conference on Information Visualisation, p. 9-16, 2006.

KEIM, D.; ANDRIENKO, G.; FEKETE, J. D.; GÖRG, C.; KOHLHAMMER, J.; MELANÇON, G. **Visual Analytics: Definition, Process, and Challenges**. Springer Berlin Heidelberg, 2008.

KOEPPL, H.; ZECHNER, C.; GANGULY, A.; PELET, S.; PETER, M. **Accounting for Extrinsic Variability in the Estimation of Stochastic Rate Constants**. International Journal of Robust and Nonlinear Control, v. 22, n. 10, p. 1103-1119, 2012.

KOHLHAMMER, J.; KEIM, D.; POHL, M.; SANTUCCI, G.; ANDRIENKO, G. **Solving Problems with Visual Analytics**. Procedia Computer Science, v. 7, p. 117-120, 2011.

LEE, B.; ISENBERG, P.; RICHE, N. H.; CARPENDALE, S. **Beyond Mouse and Keyboard: Expanding Design Considerations for Information Visualization Interactions**. IEEE Transactions on Visualization and Computer Graphics, v. 18, n. 12, p. 2689-2698, 2012.

MUIGG, P.; KEHRER, J.; OELTZE, S.; PIRINGER, H.; DOLEISCH, H.; PREIM, B.; HAUSER, H. **A Four-level Focus+ Context Approach to Interactive Visual Analysis of Temporal Features in Large Scientific Data**. Computer Graphics Forum, Blackwell Publishing Ltd, p. 775-782, 2008.

NORTH, C. **Information Visualization**. Handbook of Human Factors and Ergonomics, Fourth Edition, John Wiley & Sons, p. 1209-1236, 2005.

QUINLAN, J. R. **Combining Instance-Based and Model-Based Learning**. Proceedings of the Tenth International Conference on Machine Learning, p. 236-243, 1993.

RENSINK, R. A. **Change Detection**. Annual Review of Psychology, v. 53, n. 1, p. 245-277, 2002.

ROBERT, S. **Information Visualization - Design for Interaction**. Pearson Education Limited, 2ª edição, 2006.

SACHA, D.; SENARATNE, H.; KWON, B. C.; ELLIS, G.; KEIM, D. A. **The Role of Uncertainty, Awareness, and Trust in Visual Analytics**. IEEE transactions on visualization and computer graphics, v. 22, n. 1, p. 240-249, 2016.

SACHA, D.; SENARATNE, H.; KWON, B. C.; KEIM, D. A. **Uncertainty Propagation and Trust Building in Visual Analytics**. IEEE VIS 2014. Paris, 2014.

SCHWABER, K. **SCRUM Development Process**. Business Object Design and Implementation. Springer London, p. 117-134, 1997.

SHNEIDERMAN, B. **Inventing Discovery Tools: Combining Information Visualization with Data Mining**. Discovery Science. Springer Berlin Heidelberg, v. 1, p. 17-28, 2001.

SHNEIDERMAN, B. **The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations**. Visual Languages, IEEE Symposium Proceedings, p. 336-343, 1996.

SMAGORINSKY, J. **General Circulation Experiments with the Primitive Equations: I. the Basic Experiment**. Monthly Weather Review, v. 91, n. 3, p. 99-164, 1963.

SONNINEN, M.; GOERLANDT, F. **Exploring the Context of Maritime SAR Missions Using Visual Data Mining Techniques**. 43 Scientific Journals of the Maritime University of Szczecin, n. 43, p. 79-88, 2015.

SPENCER, S. L.; GAUDET, S.; ALBECK, J. G.; BURKE, J. M.; SORGER, P. K. **Non-genetic Origins of Cell-to-cell Variability in TRAIL-induced Apoptosis**. Nature, v. 459, n. 7245, p. 428-432, 2009.

SPENCER, S. L.; SORGER, P. K. **Measuring and Modeling Apoptosis in Single Cells**. Cell, v. 144, n. 6, p. 926-939, 2011.

STEED, C. A.; RICCIUTO, D. M.; SHIPMAN, G.; SMITH, B.; THORNTON, P. E.; WANG, D.; SHI, X.; WILLIAMS, D. N. **Big Data Visual Analytics for Exploratory Earth System Simulation Analysis**. Computers & Geosciences, v. 61, p. 71-82, 2013.

STEED, C. A.; SHIPMAN, G.; THORNTON, P.; RICCIUTO, D.; ERICKSON, D.; BRANSTETTER, M. **Practical Application of Parallel Coordinates for Climate Model Analysis**. Procedia Computer Science, v. 9, p. 877-886, 2012.

THOMAS, J. J.; COOK, K. A. **A Visual Analytics Agenda**. Computer Graphics and Applications, IEEE, v. 26, n. 1, p. 10-13, 2006.

THOMAS, J.; COOK, K.: **Illuminating the Path: Research and Development Agenda for Visual Analytics**. IEEE-Press, 2005.

TUKEY, J. W. **Exploratory Data Analysis**. Addison-Wesley, Reading MA, 1977.

WAJANT, H.; PFIZENMAIER, K.; SCHEURICH, P. **Tumor Necrosis Factor Signaling**. Cell Death & Differentiation, v. 10, n. 1, p. 45-65, 2003.

WARE, C. **Information Visualization: Perception for Design**. Elsevier, 3ª edição, 2012.

WEGMAN, E. J. **Hyperdimensional Data Analysis Using Parallel Coordinates**. Journal of the American Statistical Association, v. 85, n. 411, p. 664-675, 1990.

WEGMAN, E. J. **Visual Data Mining**. Statistics in Medicine, v. 22, n. 9, p. 1383-1397, 2003.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann, 2011.

WU, W.; XU, J.; ZENG, H.; ZHENG, Y.; QU, H.; NI, B.; YUAN, M.; NI, L. M. TelCoVis: **Visual Exploration of Co-occurrence in Urban Human Mobility Based on Telco Data**. Visualization and Computer Graphics, IEEE Transactions, v. 22, n. 1, p. 935-944, 2016.

ZHOU, H.; YUAN, X.; QU, H.; CUI, W.; CHEN, B. **Visual Clustering in Parallel Coordinates**. Computer Graphics Forum. Blackwell Publishing Ltd., v. 27, n. 3, p. 1047-1054, 2008.