CDD: 005

O USO INTEGRADO DE K-NN E SCATTER PLOTS 2D NA MINERAÇÃO VISUAL DE DADOS

THE INTEGRATED USE OF K-NN AND SCATTER PLOTS 2D IN VISUAL DATA MINING

Clodis Boscarioli, Fernando Yukio Tabusadani, Jorge Bidarra

Universidade Estadual do Oeste do Paraná – UNIOESTE

Recebido para publicação em 20/05/2008
Aceito para publicação em 28/11/2008

RESUMO

Com o aumento da quantidade de dados disponível, torna-se necessário o uso de ferramentas que os transforme em informações significativas, ou seja, em conhecimento capaz de ser utilizado de forma direta. Embora as técnicas de Mineração de Dados sejam uma boa solução, os resultados obtidos podem ser prejudicados, ou mesmo inviabilizados, se a elas não forem associadas técnicas de visualização que facilitem sua interpretação, por parte dos interessados. Este artigo apresenta o uso integrado dos algoritmos K-NN e Scatter Plots 2D, buscando integrar mineração de dados e visualização de dados.

Palavras-chave: Mineração visual de dados. Análise exploratória de dados. Visualização de dados. Aprendizagem supervisionada.

ABSTRACT

With the increasing amount of data available, it is necessary the use of tools that transform this information in knowledge capable of being used in direct way. Although the techniques of Data Mining are good solutions, the results may be no clear if they are not associated with the techniques of visualization to facilitate their interpretation by users. This paper presents the integrated use of K-NN algorithms and Scatter Plots 2D, seeking to integrate data mining and visualization of data.

Keywords: Visual data mining. Exploratory data analysis. Data visualization. Supervised learning.

1 Introdução

O estudo das técnicas de visualização em mineração de dados (*Data Mining*) é de grande relevância, uma vez que oferecem maior poder de interpretação ao usuário final, proporcionando-lhe melhor absorção de conhecimentos implícitos nos dados. Além disso, os usuários do domínio da aplicação, responsáveis pela tomada de decisão, desejam gráficos e dados sumariados, que, de forma rápida, lhes dê respaldo às suas ações.

O processo de mineração visual de dados (*Visual Data Mining*) engloba o processo de mineração de dados integrado a técnicas de visualização. A interpretação e aplicação dos padrões identificados em um processo de mineração de dados são muito dependentes do usuário especialista na aplicação, sendo, portanto, interessante oferecer-lhe ferramentas que facilitem a análise visual dos dados (TABUSADANI; BOSCARIOLI, 2007).

Visualização é um processo que realiza uma transformação da informação, da forma numérica ou textual para representações visuais, permitindo a observação gráfica dessa informação de forma a facilitar a percepção de características ocultas nos dados. Segundo (WONG, 1999 *apud* RODRIGUES JÚNIOR, 2003), um sistema deste tipo deve englobar as melhores características tanto do homem, quanto da máquina, que são, respectivamente, a capacidade do ser humano de perceber padrões, exceções, tendências e relacionamentos, e o enorme poder de processamento dos computadores, para se ter uma ferramenta analítica eficiente.

Um sistema de análise visual de dados deve se basear em requisitos como simplicidade, autonomia do usuário e confiabilidade, e deve ter a capacidade de guiar o usuário durante a análise dos dados e na geração de conclusões acerca dos resultados obtidos.

A Figura 1 apresenta as fases de um processo de VDM. Inicialmente, têm-se os dados brutos, como foram armazenados na(s) base(s) de dados. Contudo, é conveniente que os dados sejam previamente preparados (etapa de pré-processamento) para a retirada de possíveis distorções[1] ou tratamento de dados ausentes, a fim de gerar uma representação conveniente aos algoritmos de mineração (CARVALHO et al, 2003). Após essa preparação, o processo de mineração de dados propriamente dito é aplicado, que efetuará a busca por padrões, estruturas e tendências ocultos nos dados. Por fim, inicia-se o processo de apresentação das informações resultantes, que são mapeadas para estruturas visuais, utilizando-se das mais variadas técnicas de visualização, no intuito de oferecer melhores condições de interpretação ao usuário final.

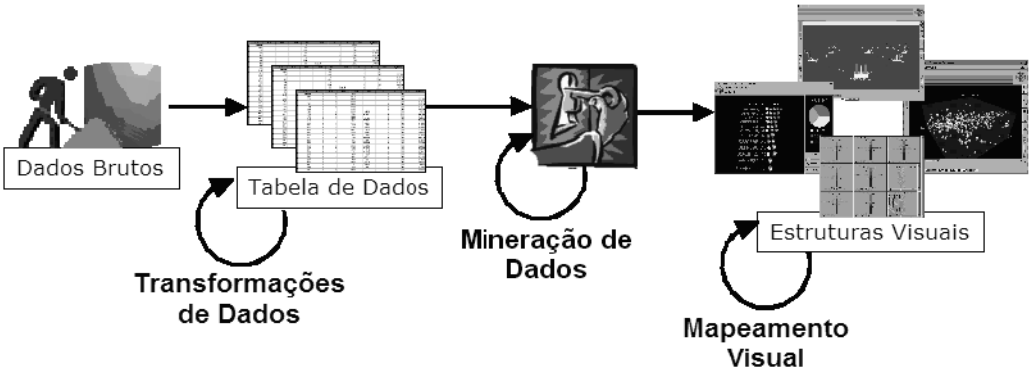

Figura 1 - O processo de VDM. Extraída de Rezende (2005).

Este artigo apresenta a integração de mineração e visualização de dados, a partir da interação das técnicas K-NN e *Scatter Plots*, propiciando recursos de análise visual ao usuário final. Está organizado como segue: a Seção 2 aborda os elementos teóricos de sustentação ao estudo e necessários à sua compreensão. Na Seção 3, a ferramenta desenvolvida é apresentada. Uma avaliação experimental é detalhada na Seção 4. Por fim, a Seção 5 traz as principais conclusões e propostas de trabalhos futuros desse trabalho.

2 Materiais e métodos

Essa seção destina-se à apresentação dos elementos teóricos de contextualização desse artigo, no tocante à descrição das técnicas utilizadas tanto para mineração quanto para a análise visual de dados.

[1] Maiores detalhes sobre a etapa de pré-processamento podem ser encontrados em Batista (2003).

2.1 Mineração de Dados

Um processo de mineração de dados dá-se por aprendizado, que pode ser supervisionado ou não-supervisionado (NORVIG; RUSSELL, 2004), (BOSCARIOLI, 2008):

- **Aprendizado Supervisionado** é aquele no qual são fornecidos classes e exemplos de cada classe ao sistema, que precisa encontrar a descrição (propriedades comuns nos exemplos) de cada classe. Neste tipo de aprendizado, existe um "professor" que guia esse processo de aprendizado. O professor, na realidade, é o conhecimento prévio das classes que estão sendo descritas pelo conjunto de exemplos de treinamento.
- **Aprendizado Não-Supervisionado** diz respeito ao tipo de aprendizado onde o sistema precisa descobrir a classe dos objetos pelas propriedades que os mesmos têm em comum. É realizada a análise dos exemplos fornecidos e tenta-se determinar se alguns deles podem ser agrupados. Após a criação dos grupos, normalmente é necessária uma análise mais refinada para determinar o que cada agrupamento significa no contexto do problema analisado.

A mineração de dados pode ser dividida por tarefas, como Classificação, Regras de Associação, Análise de Agrupamentos e Predição (HAN; KAMBER, 2001). Para cada uma dessas tarefas, há vários métodos propostos.

O foco deste trabalho é a tarefa de classificação, cuja função de aprendizado mapeia dados de entrada, ou conjuntos de dados de entrada, para um número finito de classes. Cada exemplo pertence a uma classe entre um conjunto pré-definido de classes. O objetivo de um algoritmo de classificação é encontrar algum relacionamento entre os atributos e uma classe, de modo que o processo de classificação possa usar esse relacionamento para descobrir a classe de um exemplo desconhecido. A técnica de classificação implementada neste estudo é o algoritmo K-NN (AHA et al, 1991 apud RIBEIRO et al, 2006), (MITCHELL, 1997), que armazena instâncias de treinamento na memória como pontos no espaço *n*-dimensional, definido pelos *n* atributos que os descrevem. Quando uma nova instância precisa ser classificada, a classe mais frequente dentre os *K* vizinhos mais próximos é escolhida.

K-NN procura *K* elementos do conjunto de treinamento que estejam mais próximos de um determinado elemento desconhecido, ou seja, que tenham a menor distância. Estes *K* elementos são chamados de *K*-vizinhos mais próximos. A seguir, verifica-se quais são as classes a que pertencem esses *K* vizinhos, de modo que a classe mais frequente venha a ser a classe do novo elemento (DUDA et al, 2001).

Embora existam várias medidas para cálculo de distância entre dois pontos (EVERITT; RABE-HESKETH, 1997), a Distância de Manhattan é uma métrica bastante utilizada para tal fim, e adotada nesse estudo, cuja definição segue abaixo.

Sejam dois pontos, $X = (x_1, x_2, \ldots, x_n)$ e $Y = (y_1, y_2, \ldots, y_n)$, a distância de Manhattan entre X e Y é dada pela Equação 1. Nesta, equação a distância (d (x,y)) é dada pela soma dos módulos das diferenças entre cada um dos elementos x_n e y_n dos pontos X e Y, em que *n* varia entre 1 e o número de elementos de cada ponto.

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n| \quad (4.1)$$

K-NN é um classificador que possui apenas um parâmetro livre (o número de *K*-vizinhos), que é controlado pelo usuário com o objetivo de obter uma melhor classificação.

A Figura 2 apresenta um exemplo de classificação KNN, com três classes presentes, e dois pontos a serem classificados. Neste caso, foram usados os 11 pontos mais próximos de cada ponto desconhecido. Observa-se que um ponto será classificado na classe B e o outro ponto na classe C, a partir das distâncias calculadas. Uma desvantagem deste algoritmo é que ele pode ser computacionalmente exaustivo para grandes bases de dados.

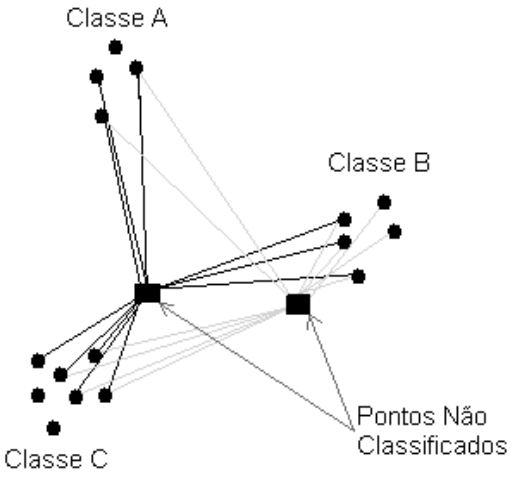

Figura 2 - Exemplo de Classificação usando o algoritmo KNN.

2.2 Visualização de dados

Existe um grande número de técnicas de visualização que podem ser usadas para visualizar dados. Em adição às técnicas 2D/3D convencionais como gráfico de barras, gráfico de linhas etc., existem técnicas de visualização mais sofisticadas, como as descritas por Oliveira e Levkowitz (2003).

As técnicas orientadas a Pixels, em que este trabalho se insere, baseiam-se no mapeamento de um elemento de informação para um pixel colorido do dispositivo de exibição, proporcionando a análise simultânea do maior número de dados possível. Para cada atributo dos elementos de dados, uma imagem pode ser construída e exibida em uma janela separada. A técnica se preocupa em como os pixels serão organizados na tela, otimizando a percepção do usuário em relação ao conjunto de dados (RODRIGUES JÚNIOR, 2003). A técnica mais conhecida dessa abordagem e implementada neste estudo é a Scatter Plots**,** utilizada na projeção de dados de alta dimensionalidade para duas dimensões, em que as dimensões são combinadas duas a duas e exibidas uma em função da outra.

O algoritmo Scatter Plots é utilizado para a visualização de dados, de modo que cada item de dado é representado por um ponto e a sua classe é representada por uma cor. Essa técnica é especialmente capaz de expressar correlações entre as dimensões da base de dados. A Figura 3 apresenta um exemplo de utilização desta técnica, composta por itens de carros de origem japonesa (cor verde) e europeia (cor azul). No exemplo, as características dos carros são combinadas duas a duas e plotadas pela sua correlação.

O maior problema desta técnica se dá quando existem muitos atributos a serem representados, pois o espaço delimitado para plotar os pontos que representam os itens de dados fica bastante reduzido, e como consequência, têm-se um amontoado de pontos que pouco contribui para a observação e para a análise dos dados pelos usuários finais. Outro problema ocorre quando os valores dos atributos são idênticos – ou muito próximos, pois, no momento de plotagem desses pontos, haverá sobreposição e, portanto, alguns pontos não poderão ser observados pelo usuário final.

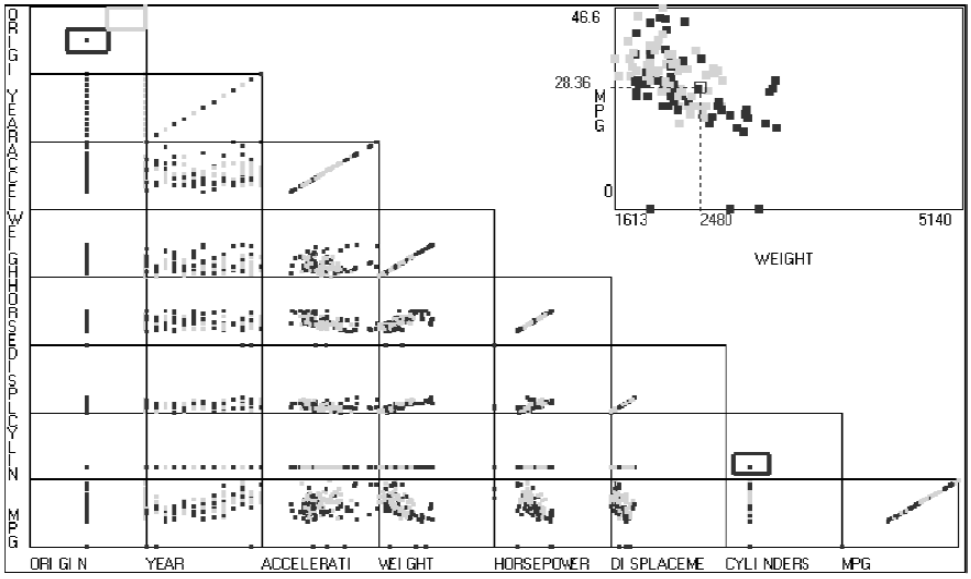

Figura 3 - Ilustração da Técnica Scatter Plots. (Extraída de Rodrigues Júnior, 2003)

O objetivo do algoritmo de mineração de dados K-NN é analisar novos itens de dados e classificá-los nos grupos em que suas características mais se aproximam. Normalmente, a apresentação dos resultados obtidos é feita em tabelas, porém, esta não é a melhor forma de realizar esta exibição, devido às dificuldades de perceber relação entre os dados.

3 Resultados e discussão

Um protótipo foi implementado em linguagem JAVA com Sistema Gerenciador de Banco de Dados PostGreSQL. Na versão atual, há uma restrição quanto ao posicionamento dos atributos nas tabelas a serem utilizadas, em que o último atributo à direita deve conter sempre o atributo classe, ou seja, a função objetivo da classificação por K-NN. Caso o atributo classe não esteja nesta posição, o software não funcionará ou apresentará resultados incorretos.

Testes foram realizados nessa ferramenta, com o objetivo de verificar as taxas de acerto do algoritmo K-NN e de averiguar a qualidade da apresentação dos resultados por parte da técnica Scatter Plots 2D, bem como o de observar visualmente os benefícios de se ter os dois algoritmos atuando de forma integrada.

Para melhor compreensão da proposta, a descrição da ferramenta é dividida em duas partes. A primeira descreve aspectos teóricos e decisões de projeto e a segunda apresenta a ferramenta para utilização do usuário final.

3.1 Aspectos teóricos da ferramenta proposta

Na ferramenta desenvolvida, também é possível observar os resultados obtidos com o uso da técnica K-NN, através de uma tabela. Para isto, o usuário deve escolher as duas tabelas necessárias à classificação de dados e o valor de *K* a ser utilizado pelo algoritmo. Porém, o mais aconselhável é realizar a análise através do uso, em conjunto com a técnica Scatter Plots 2D.

O principal objetivo da técnica Scatter Plots 2D é fornecer ao usuário uma representação gráfica de cada um dos itens de dados em forma de um ponto no dispositivo de exibição, para que, através das características de abstração do ser humano, seja possível perceber padrões e tendências nos dados. Para que isto seja possível, esta técnica deve buscar meios para que os usuários possam identificar e diferenciar os agrupamentos existentes.

Cada classe existente é representada por uma cor diferente para que o usuário possa visua-lizar a qual classe cada ponto (item de dado) pertence. Com esta possibilidade, é possível analisar relações entre os itens de dados e fazer inferências sobre eles. Para realizar a visualização, é necessária apenas a escolha de uma tabela com os dados já classificados.

Para que a integração seja possível, os resultados obtidos pelo algoritmo K-NN devem ser passados para o algoritmo Scatter Plots 2D, para que este possa exibir, além dos itens com classes originalmente conhecidas, também os itens recentemente classificados. Uma estrutura de lista é utilizada para passar os resultados ao Scatter Plots. Nesta lista, estão os valores dos atributos de cada item de dado, seguidos pelas suas "novas" classes.

Os usuários devem ter um meio de interagir com os resultados apresentados pelo algoritmo K-NN, para então visualizar todas as alterações realizadas na apresentação global. Para que isto seja possível, deve-se utilizar a técnica Scatter Plots 2D, em que um esquema de cores foi utilizado para diferenciar os grupos originais (já existentes) dos grupos de dados que foram classificados.

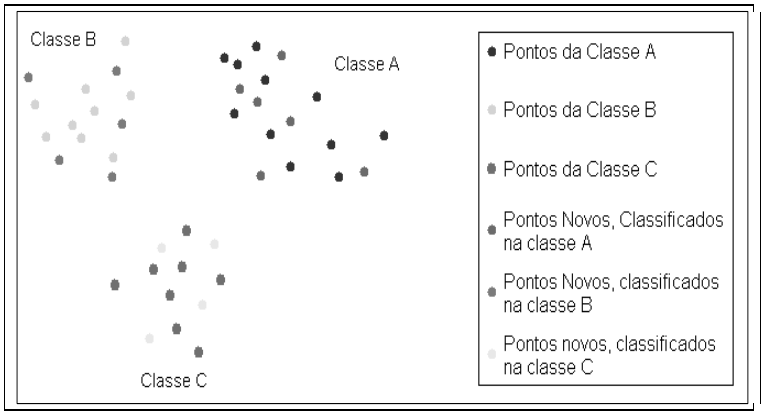

Figura 4 - Exemplo de Uso do Algoritmo K-NN.

A Figura 4 apresenta um exemplo da aplicação desse esquema de cores, em que os pontos em azul representam aqueles que já estavam classificados na classe A; os pontos em verde, os que já estavam classificados na classe B; e os pontos em cinza, os que já estavam classificados na classe C. Os pontos em vermelho, roxo e anil são os novos pontos que foram classificados pelo algoritmo K-NN, nas classes A, B e C, respectivamente. Desta forma, o usuário tem clareza sobre o que foi modificado pela inserção dos novos dados.

Para a utilização das técnicas de mineração e de visualização de forma integrada são necessárias duas tabelas (de treinamento e de testes). Após ser realizada a classificação pelo algoritmo K-NN, os resultados obtidos serão passados para o algoritmo de visualização, que então irá exibi-los, utilizando o esquema de cores apresentado acima.

3.2 Apresentação da ferramenta desenvolvida

O sistema de mineração visual de dados ora apresentado permite ao usuário a utilização das técnicas Scatter Plots 2D e K-NN de forma isolada ou integradas (Figura 5).

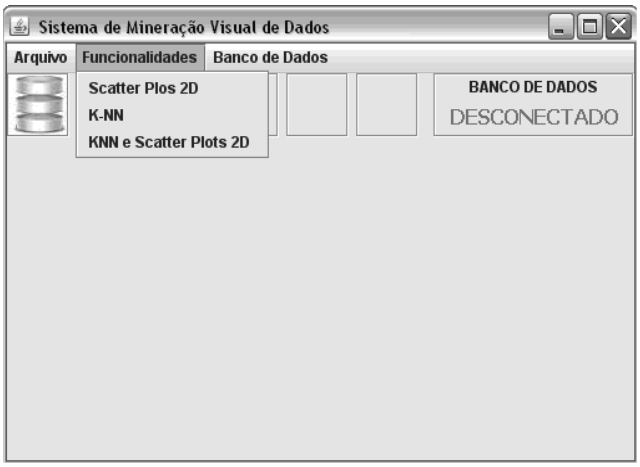

Figura 5 - Menu das funcionalidades de análise disponíveis.

Para utilizar qualquer das funcionalidades da ferramenta, o usuário deve se conectar ao SGBD PostGreSQL. Para se conectar, o usuário deve determinar o SERVER (localhost ou IP), o nome da BASE de dados ao qual se deseja conectar, a PORTA em que o PostGreSQL está aceitando conexões e o USUÁRIO e a SENHA de acesso ao banco, como ilustrado na Figura 6.

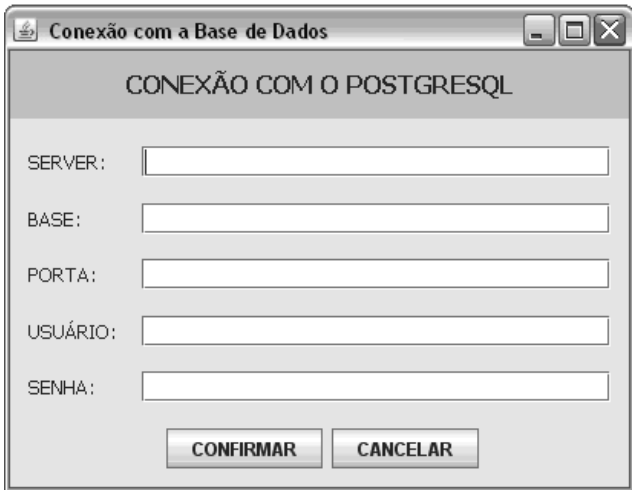

Figura 6 - Tela de Conexão com o Banco PostGreSQL.

Depois de efetuada a conexão com o PostGreSQL e escolhida alguma das funcionalidades da ferramenta, será exibida uma nova janela, onde serão listadas todas as tabelas presentes no banco de dados escolhido (Figura 7). Para o uso das funcionalidades K-NN e K-NN e Scatter Plots 2D são necessárias duas tabelas. A primeira ("TABELA DE EXEMPLOS") é a tabela que possui dados já classificados (de origem) e que servirão de exemplo para poder classificar, através do algoritmo K-NN, os novos dados presentes na segunda tabela ("TABELA DE ITENS P/ CLASSIFICAR"). Apenas para a funcionalidade Scatter Plots 2D será requerida somente uma tabela, contendo os itens já classificados.

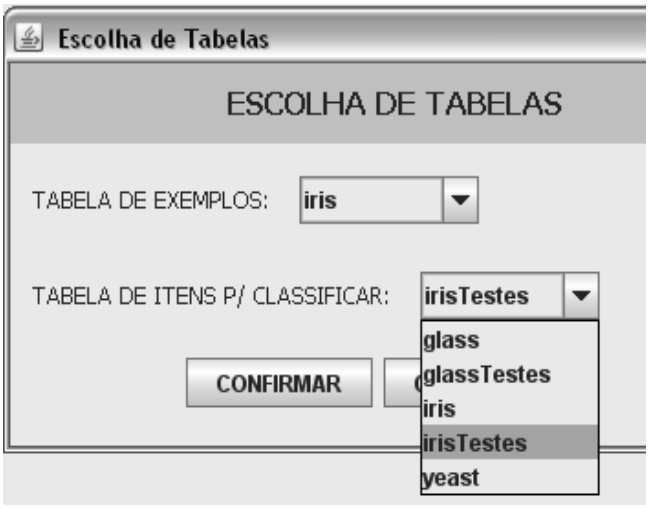

Figura 7 - Tela de Escolha de Tabelas.

3.3 Um Exemplo de Análise Experimental

Com o objetivo de verificar a precisão com que o algoritmo K-NN implementado no protótipo realiza a classificação dos novos itens de dados, realizaram-se alguns testes de validação. Existem vários métodos para avaliação de acuidade de classificadores como Cross-Validation, Leave-on-Out e Um Terço Dois Terços. Para os testes realizados, utilizou-se a técnica Cross-Validation, bastante citada na literatura. Nesta, o conjunto de dados (com as classes dos itens de dado conhecidas) é dividido igualmente em *n* subconjuntos, de forma aleatória. O treinamento efetua-se concatenando *n*-1 subconjuntos, e a validação conta com o uso do subconjunto restante. As fases de treinamento e teste são depois repetidas *n* vezes, permutando circularmente os subconjuntos.

O erro final para cada valor de *K* é calculado usando a média dos erros de cada fase.

Embora a ferramenta tenha sido testada em outras bases de dados (TABUSADANI, 2007) pertencentes à UCI[2] (NEWMAN et al., 1998), apenas a análise da base de dados ÍRIS – criada por R. A. Fisher em 1988 – é apresentada nesse artigo.

A base de dados Iris contém dados de três tipos de uma flor conhecida por Iris, Setosa, Versicolour e Virginica. Além do atributo de classificação, possui outros quatro atributos que trazem as informações de tamanho e espessura de sépala e tamanho e espessura de pétala, com todas as dimensões dadas em centímetros.

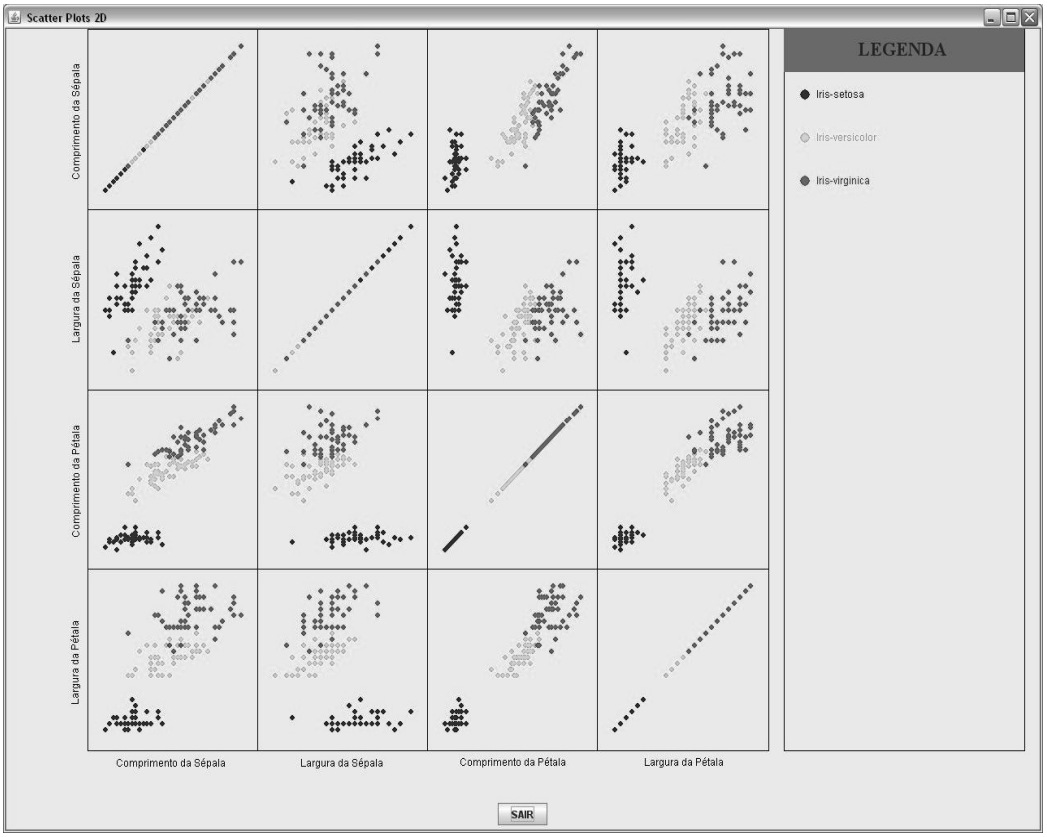

Figura 8 - Técnica de Scatter Plots 2D aplicada à base de dados IRIS.

É composta por 150 amostras, sem nenhum valor ausente para seus atributos e com distribuição igualitária dos dados em classes. Sabe-se também que a classe Setosa é linearmente separável das classes Versicolour e Virginica, porém estas não o são entre si.

Essa base de dados foi dividida em dez partes, cada uma com a mesma quantidade de itens de dados. Como a quantidade de itens em cada classe é a mesma (50 cada classe), dividiu-se cada classe em dez partes de cinco itens cada, e depois juntou-se uma parte de cada classe para formar cada conjunto de teste.

Caso a escolha seja pelo uso apenas de Scatter Plots 2D, será exibido um resultado como o da Figura 8, em que cada quadrado representa a correlação entre dois atributos da base de dados e a legenda, apresentada mais à direita, traz a distinção de cada tipo de classe.

A técnica Scatter Plots 2D apresenta um problema quando o número de dimensões é grande, pois o tamanho de cada quadrado pode se tornar muito pequeno, ocasionando a falta de espaço para plotar os pontos. Na tentativa de amenizar este problema, a ferramenta oferece a oportunidade ao usuário de realizar um zoom no quadrado que desejar. A Figura 9 apresenta o zoom da combinação entre os atributos "Largura da Sepala" e "Comprimento da Petala" da base de dados IRIS.

[2] Este é um repositório de bases de dados de acesso público que pode ser acessado em: http://mlearn.ics.uci.edu/MLRepository.html

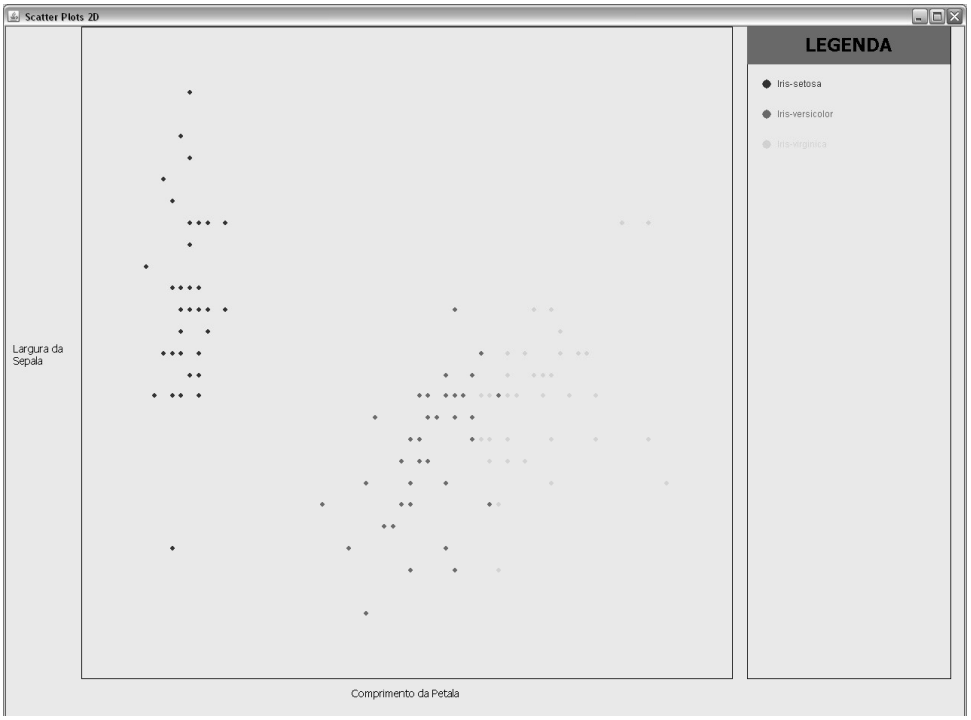

Figura 9 - Exemplo de zoom em um quadrado de Scatter Plots 2D.

Caso seja feita a escolha por executar apenas o algoritmo de mineração de dados, ou seja, sem a exibição gráfica da Scatter Plots 2D, os resultados obtidos serão apresentados em uma tabela, como demonstrado na Figura 10.

Resultado da Classificação pelo Algoritmo K-NN

Comprimento da Sépala	Largura da Sépala	Comprimento da Pétala	Largura da Pétala	CLASSE
4.8	3.0	1.4	0.3	Iris-setosa
5.1	3.8	1.6	0.2	Iris-setosa
4.6	3.2	1.4	0.2	Iris-setosa
5.3	3.7	1.5	0.2	Iris-setosa
5.0	3.3	1.4	0.2	Iris-setosa
5.7	3.0	4.2	1.2	Iris-versicolor
5.7	2.9	4.2	1.3	Iris-versicolor
6.2	2.9	4.3	1.3	Iris-versicolor
5.1	2.5	3.0	1.1	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
6.7	3.0	5.2	2.3	Iris-virginica
6.3	2.5	5.0	1.9	Iris-virginica
6.5	3.0	5.2	2.0	Iris-virginica
6.2	3.4	5.4	2.3	Iris-virginica
5.9	3.0	5.1	1.8	Iris-virginica

FECHAR

Figura 10 - Resultado Tabular de Classificação pelo K-NN.

Se forem escolhidas as duas técnicas em conjunto, "K-NN e Scatter Plots 2D", o resultado obtido será semelhante ao apresentado na Figura 11, em que são exibidos os itens de dados com classes originalmente definidas e também os itens de dados que foram classificados pelo K-NN, todos representados por cores diferentes, como já dito. Para obtenção desse resultado, 90% do conjunto de dados original foi utilizado como conjunto de treinamento e 10% como conjunto de testes, de acordo com a definição da técnica de *Cross-Validation,* já explicada.

Na Figura 11, os pontos de cores azul, vermelho e verde representam os itens de dados que já estavam classificados originalmente e que foram utilizados como conjunto de treinamento para que o algoritmo K-NN pudesse classificar os novos itens, representados na figura pelas cores anil, amarelo e roxo. Os pontos em anil representam aqueles que foram classificados na classe "Setosa"; os pontos em amarelo, aqueles que foram classificados na classe "Versicolor"; e os pontos em roxo, os que foram classificados na classe "Íris-Virginica".

A partir da observação da Figura 11, o usuário pode obter maior conhecimento do comportamento da base de dados e também pode ver os relacionamentos entre os itens de dados que já estavam classificados e os novos itens de dados classificados pelo K-NN.

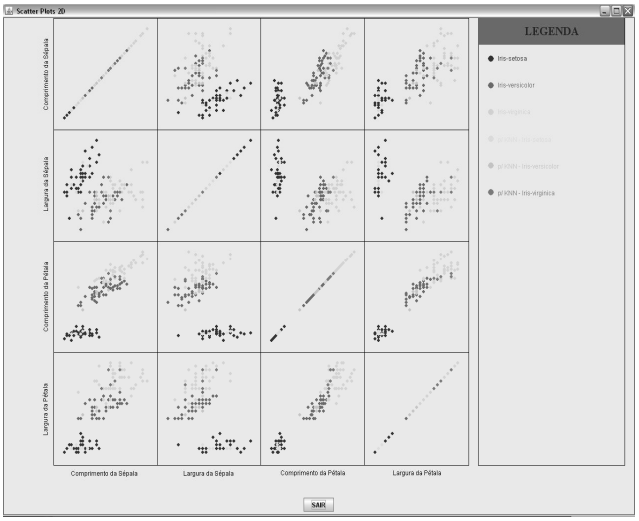

Figura 11 - Técnica de K-NN e Scatter Plots 2D para a base de dados ÍRIS.

Com os conjuntos de testes formados, cada um destes teve seus itens classificados pelo algoritmo K-NN, utilizando-se diferentes valores de *K* (número de vizinhos mais próximos), e os resultados obtidos foram comparados, item a item, com as classes originais da base de dados (Tabela 1).

O percentual de acertos foi sendo acumulado e ao final foi realizada a média simples dos mesmos, apresentados na Tabela 2. Pode ser interessante observar as taxas médias de acerto para cada valor de *K*. Em negrito, estão os maiores valores médios obtidos para cada base de dados. Com estas informações, é possível aperfeiçoar os resultados obtidos pelo K-NN, para cada base de dados, no momento de uma nova classificação, utilizando-se do valor *K* que obteve a maior média.

Tabela 1 - Taxas de Acertos de Classificação do K-NN para diferentes conjuntos de teste e valores de K.

CONJUNTO	TAXA DE ACERTO		MÉDIA GERAL DE ACERTO
1º	K = 3	100%	95,92%
	K = 5	100%	
	K = 7	100%	
	K = 10	100%	
	K = 15	100%	
2º	K = 3	93%	
	K = 5	93%	
	K = 7	93%	
	K = 10	93%	
	K = 15	93%	
3º	K = 3	100%	
	K = 5	100%	
	K = 7	100%	
	K = 10	100%	
	K = 15	100%	
4º	K = 3	93%	
	K = 5	100%	
	K = 7	93%	
	K = 10	93%	
	K = 15	100%	
5º	K = 3	87%	
	K = 5	87%	
	K = 7	87%	
	K = 10	93%	
	K = 15	100%	
6º	K = 3	100%	
	K = 5	93%	
	K = 7	87%	
	K = 10	87%	
	K = 15	93%	
7º	K = 3	87%	
	K = 5	93%	
	K = 7	93%	
	K = 10	93%	
	K = 15	93%	
8º	K = 3	100%	
	K = 5	100%	
	K = 7	93%	
	K = 10	93%	
	K = 15	93%	
9º	K = 3	100%	
	K = 5	100%	
	K = 7	100%	
	K = 10	100%	
	K = 15	100%	
10º	K = 3	100%	
	K = 5	100%	
	K = 7	100%	
	K = 10	100%	
	K = 15	100%	

Tabela 2 - Média de Acertos de Classificação da base IRIS pelo K-NN para diferentes valores de K.

K	MÉDIA DE ACERTOS
K = 3	96%
K = 5	96%
K = 7	94%
K = 10	95%
K = 15	**97%**

4 Conclusões

As técnicas de mineração de dados e de visualização, se utilizadas de forma conjunta, oferecem, sem dúvida, um grande benefício, resultando em software que proporciona ao usuário uma melhor visão das correlações e do comportamento dos dados e tornando o processo de tomada de decisão muito mais interativo e pautado nas informações. A junção dessas técnicas enriquece a análise exploratória dos dados.

O diferencial da ferramenta apresentada está na integração dos algoritmos de mineração (K-NN) e de visualização (Scatter Plots 2D), proporcionando a visualização dos resultados encontrados pelo algoritmo K-NN pela técnica Scatter Plots 2D, com a utilização do esquema de cores apresentado. Com esta integração, o usuário tem a possibilidade de observar o comportamento dos dados, tanto dos que já possuem classes de origem (provenientes da base de dados), quanto dos que foram classificados pelo algoritmo K-NN.

Ao analisar os resultados dos testes realizados, percebeu-se que a forma de apresentação dos resultados realmente influencia na facilidade de interpretação dos mesmos. Na apresentação tradicional, feita em tabelas, é mais difícil observar e entender o comportamento e os relacionamentos existentes entre os dados, do que quando essa apresentação é realizada por estruturas gráficas.

Apesar dos problemas típicos da técnica Scatter Plots 2D citados neste artigo, a possibilidade de o usuário realizar a classificação de dados e analisá-los de forma gráfica aumenta a interpretabilidade dos resultados, possibilitando maior absorção de conhecimento a partir dos dados.

Como trabalhos a serem desenvolvidos futuramente, podem-se citar a implementação de outras técnicas de visualização e classificação na ferramenta; a definição de heurísticas para a determinação da melhor solução-problema de classificação, bem como a realização de uma avaliação da interface.

REFERÊNCIAS

BATISTA, G. E. de A.P.A. **Pré-processamento de dados em aprendizado supervisionado**. Tese. São Carlos-SP: Universidade de São Paulo, 2003.

BOSCARIOLI, C. **Análise de agrupamentos baseada na topologia dos dados e em mapas auto-organizáveis**. Tese. São Paulo-SP. Escola Politécnica. Universidade de São Paulo, 2008.

CARVALHO, F. P. de; FAGUNDES JUNIOR, A.; SILVEIRA; J. G. KDD-NMS: um sistema de descoberta de conhecimento e mineração em bases de dados de sistemas de gerência de redes. In: WORKSHOP WRNP2, 4. SBRC 2003. **Anais...** Natal/RN, Maio, 2003.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification.** 2.ed. New York: John Wiley & Sons, 2001.

EVERITT, B. S.; RABE-HESKETH, S. **The analysis of proximity data**. Londres: Hodder Arnold Publishers, 1997.

HAN, J.; KAMBER, M. **Data mining:** concepts and techniques. San Francisco: Morgan Kaufmann, 2000.

MITCHELL, T. **Machine learning**. New York: McGraw Hill, 1997.

NEWMAN, D. J. et al. **UCI:** repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine, CA, 1998. Disponível em: <http://www.ics.uci.edu/~mlearn/MLRepository.html>

NORVIG, P.; RUSSELL, S. J. **Inteligência artificial.** 3.ed. Editora Campus, 2004.

OLIVEIRA, M. C. F. de; LEVKOWITZ, H. From visual data exploration to visual data mining: a survey**.** In: **IEEE transactions on visualization and computer graphics.** Piscataway – USA: IEEE Educational Activities Department, 2003, p. 378-394.

REZENDE, S. O. Mineração de dados. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 15. São Leopoldo-RS, 2005. **Anais...** p. 397-433.

RIBEIRO, R.; KOERICH, A. L.; ENEMBRECK, F. Uma nova metodologia para avaliação de desempenho de algoritmos baseados em aprendizagem por reforço. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 26, Porto Alegre, 2006, **Anais...** p. 433-446.

RODRIGUES JÚNIOR, J. F. **Desenvolvimento de um framework para análise visual de informações suportando data mining.** Dissertação. São Carlos-SP: Universidade de São Paulo, 2003.

TABUSADANI, F. Y. **Mineração visual de dados:** um estudo e prototipação preliminares. Monografia. Cascavel-PR: Universidade Estadual do Oeste do Paraná, Dezembro, 2007.

TABUSADANI, F. Y.; BOSCARIOLI, C. Um experimento de avaliação de ferramentas para análise visual de dados. In: CONGED - CONGRESSO DE TECNOLOGIAS PARA GESTÃO DE DADOS E METADADOS DO CONE SUL, 5. Cascavel, 2007. **Anais...** p. 79-89.