Dual Projections:
Uma Nova Abordagem para Redução de Dimensionalidade e Exploração do Espaço de Atributos.

Francisco Morgani Fatore

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC/USP, para o Exame de Qualificação, como parte dos requisitos para a obtenção do título de Mestre na Área de Ciências de Computação e Matemática Computacional.

Orientador: Prof. Dr. Fernando Vieira Paulovich

USP – São Carlos/SP
Abril/2013

Resumo

Em aplicações de análise de dados, raramente se conhece a dimensionalidade intrínseca dos dados, isto é, o conjunto de variáveis observadas que são realmente relevantes para a compreensão do fenômeno estudado. Métodos automáticos são frequentemente utilizados para compor subconjuntos de atributos, ou combinações entre eles, que agreguem a maior parte da informação contida nos dados. Entretanto, esses métodos evitam ao máximo a interação do usuário, o que além de tornar o processo pouco intuitivo, impede que o usuário modifique os resultados de acordo com a sua experiência na área. Técnicas de visualização computacional têm sido utilizadas com sucesso na análise exploratória de conjuntos de dados, pois permitem que o usuário utilize sua percepção visual para detectar padrões e seu conhecimento sobre o domínio para interagir com os dados e orientar as análises. Este projeto de mestrado propõe o uso de projeções multidimensionais coordenadas entre itens e dimensões para a execução da tarefa de redução de dimensionalidade de forma mais intuitiva, ágil e confiável. Além disso, propõe-se um método que permite ao usuário modificar os resultados obtidos, quando necessário, por meio da transformação do espaço de atributos.

Abstract

Sumário

1 Introdução **1**
1.1 Contextualização e Motivação 1
1.2 Organização da Monografia 1

2 Revisão Bibliográfica **3**
2.1 Considerações Iniciais 3
2.2 Redução de Dimensionalidade Automática 4
2.2.1 Seleção de Características 4
2.2.2 Extração de Características 4
2.3 Redução de Dimensionalidade Interativa 5
2.3.1 Value and Relation Display 5
2.3.2 Brushing Dimensions 7
2.4 Considerações Finais 7

3 Proposta de Trabalho **9**
3.1 Considerações Iniciais 9
3.2 Objetivos 9
3.3 Metodologia 10
3.3.1 Similaridade entre Dimensões 11
3.3.2 Mapeamento no Espaço Bidimensional 12
3.3.3 Visualização da Incerteza 12
3.3.4 Mecanismos de Interação 12
3.3.5 Forma de Avaliação 12
3.4 Resultados Esperados 13
3.5 Resultados Preliminares 13
3.6 Plano de Atividades e Cronograma Previsto 13

Lista de Figuras

2.1 VaR: Value and Relation 5

3.1 VaR: Value and Relation 10
3.2 VaR: Value and Relation 12

Lista de Tabelas

3.1 Cronograma de atividades 11
3.2 Cronograma de atividades 14

Lista de Siglas

KDD	Knowledge Discovery in Databases
MDS	Multidimensional Scaling
PCA	Principal Component Analysis
SOM	Self-Organizing Maps
SVD	Singular Value Decomposition
TFIDF	Term-Frequency Inverse Document-Frequency
???	????

Capítulo

1

Introdução

1.1 Contextualização e Motivação

1.2 Organização da Monografia

O restante desta monografia está estruturado da seguinte maneira:

- No Capítulo 2 apresenta-se um levantamento sobre os trabalhos que buscam de algum modo fazer uso de representações visuais para a execução da tarefa de redução de dimensionalidade.
- No Capítulo 3 discute-se com mais detalhes a proposta de trabalho e a metodologia adotada. Apresenta-se também um cronograma das atividades necessárias para a conclusão do trabalho.

Capítulo

2

Revisão Bibliográfica

2.1 Considerações Iniciais

Os trabalhos contidos na literatura que mais se assemelham ao o aqui proposto aparecem sobre o nome de métodos de redução de dimensionalidade. A redução de dimensionalidade é o processo realizado para se representar dados de alta dimensionalidade em um espaço de menor dimensionalidade, onde, idealmente, o espaço reduzido corresponde à dimensionalidade intrínseca dos dados. A dimensionalidade intrínseca dos dados é o número mínimo de parâmetros necessários para descrever as propriedades dos dados (Fukunaga, 1990).

Diversas aplicações se beneficiam de métodos de redução de dimensionalidade. Konig (Konig, 2000), por exemplo, apresenta melhorias na precisão de sistemas de classificação e no desempenho de sistemas de reconhecimento automático ao preceder os procedimentos com o processo de redução de dimensionalidade. Até mesmo outras melhorias não tão diretas podem ser alcançadas por meio do uso de técnicas de redução. Trata-se do caso do mesmo trabalho apresentado por Konig, onde métodos de redução de dimensionalidade são utilizados para reduzir a complexidade de designs de circuitos integrados, resultando em uma redução na área e no consumo de energia dos circuitos.

Mais formalmente, este o problema de se reduzir a dimensionalidade dos conjuntos de dados pode ser descrito da seguinte forma: Dado um conjunto de dados representado por uma matriz $\mathbf{X}$ composta por n vetores $\mathbf{x}_i$ $(i \in \{1, 2, ..., n\})$ $m-$dimensonas. Uma técnica de redução de dimensionalidade é uma transformação $t : \mathbf{X} \rightarrow \mathbf{Y}$, onde $\mathbf{Y}$ trata-se de uma matriz composta por n vetores $\mathbf{y}_i$ $(i \in \{1, 2, ...n\}$ de dimensionalidade p $(p < m)$. Normalmente $p \ll m$ e,

idealmente, p equivale à dimensionalidade intrínseca dos dados, fazendo com que t mantenha em $\mathbf{Y}$ o máximo das propriedades de $\mathbf{X}$ quanto for possível.

A literatura em redução de dimensionalidade é extensa e os métodos desenvolvidos apresentam grande diversidade em relação a aspectos matemáticos e computacionais. Buscando uma melhor organização, este capítulo foi dividido em duas seções. A primeira busca descrever sucintamente os métodos automáticos e apresentar suas limitações, principalmente evidenciar que a falta da participação do usuário no processo faz com que muitas vezes os resultados obtidos não sejam facilmente compreendidos. A segunda seção apresenta os métodos que buscam de algum modo utilizar representações visuais para a execução da tarefa e que diante das limitações dos métodos automáticos, se mostram como uma alternativa interessante, pois permitem a interação do usuário. No entanto, a pesquisa de técnicas visuais para o problema redução de dimensionalidade ainda se encontra em um estágio inicial e os métodos desenvolvidos apresentam grandes limitações.

2.2 Redução de Dimensionalidade Automática

A redução de dimensionalidade automática pode ser realizada seguindo duas abordagens. A primeira, dita seleção de características (*feature selection*), busca selecionar quais dos atributos do conjunto de dados são realmente relevantes para a análise segundo algum critério. Já a segunda abordagem, parte da combinação entre atributos para criar um novo conjunto de dimensões que busca conservar as propriedades e relacionamentos do conjunto original. Por construir um novo conjunto de atributos, a segunda abordagem recebe o nome de extração de características (*feature extraction*).

2.2.1 Seleção de Características

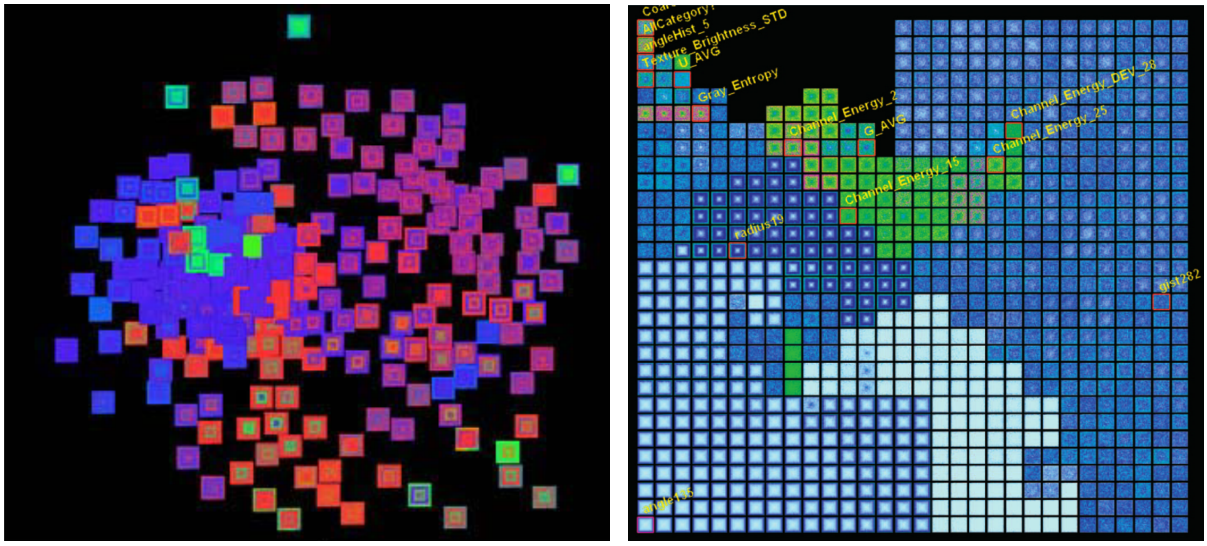
2.2.2 Extração de Características

Existem três principais abordagens para se reduzir a dimensionalidade dos conjuntos de dados a partir da combinação dos atributos. Análise de Componentes Principais (*Principal Component Analysis)* ou simplesmente PCA, realiza combinações lineares sobre os atributos de modo que o novo espaço agregue a maior parte da variância dos dados. Para análises onde relações não lineares devem ser consideradas, *Multimensional Scaling* (MDS) é uma alternativa interessante, pois trata-se de um algoritmo de otimização iterativo não linear, que busca minimizar as distâncias entre os elementos no espaço projeto e no espaço original. A área de aprendizado de máquina contribuiu com o método não supervisionado *Self Organizing Maps* (SOM) para transformar conjuntos de dados em mapas bidimensionais.

2.3 Redução de Dimensionalidade Interativa

2.3.1 Value and Relation Display

A técnica VaR (Value and Relation) (Yang et al., 2004) une os conceitos de MDS e glifos para representar as dependências entre as dimensões de uma base de dados. Como mostra a Figura 3.2a, cada glifo representa uma dimensão e de acordo com seus posicionamentos no plano o usuário pode compreender quais dimensões se relacionam entre si. O usuário é capaz de construir espaços dimensionais reduzidos que conservam certas características dos dados por meio de seleções sobre os dados ou pelo uso de um método automático que a partir de uma dimensão de referência e um *threshold* definido pelo usuário retorna as dimensões mais semelhantes.



(a) **(b)**

Figura 2.1: (a) Exemplo da técnica VaR para um conjunto de 50.000 itens e 361 dimensões. Cada dimensão é representada por um glifo e seus posicionamentos refletem a similaridade entre as dimensões, de modo que glifos que se encontram próximos indicam atributos que apresentam alguma relação entre si. É possível notar certas sobreposições entre os glifos, condição que pode dificultar as análises realizadas pelo usuário. (b) Exemplo de representação alternativa proposta como extensão da técnica VaR para um conjunto de 11.413 itens e 838 dimensões. O principal objetivo da representação é evitar a sobreposições de glifos ocorrente na versão anterior da técnica.

O procedimento para a construção desta visualização inicia pela construção de uma matriz de distâncias que é responsável por capturar os relacionamentos entre pares de dimensões do conjunto de dados (como a correlação). Sobre esta matriz de distâncias aplica-se uma técnica de MDS para mapear cada dimensão em uma posição de um espaço bidimensional. Finalmente, cria-se um glifo orientado a pixels para cada dimensão que será utilizado para representar cada dimensão no plano.

Observando a Figura 3.2a é possível notar que o uso de glifos faz com que ocorram sobreposições, pois cada glifo requer um espaço relativamente grande para que seja observado adequadamente. As sobreposições dificultam as análises de regiões de interesse e podem fazer com que o usuário alcance conclusões inválidas, devido a oclusão de algum elemento importante. Buscando tratar este problema Yang et al. desenvolveram a extensão (Yang et al., 2007) ilustrada na Figura 3.2b para a técnica VaR, onde apresentaram alternativas para a projeção de glifos no plano. No entanto, diferentemente das projeções, as alternativas propostas não são capazes de transmitir os relacionamentos entre as dimensões tão bem quanto o resultado obtido pelo MDS.

A representação explícita das dimensões do conjunto de dados serve como inspiração para este trabalho de mestrado. Já o uso de glifos orientado a pixels se mostrou inadequado para situações com um elevado número de elementos, causando indesejadas sobreposições entre elementos. Um outro aspecto importante que os próprios autores mencionam em relação ao uso deste tipo de glifos é que os usuários têm dificuldade em comparar glifos que se encontram afastados entre si.

Uma etapa fundamental da técnica VaR que merece uma maior atenção é o método utilizado para a criação da matriz de distâncias. Os autores desenvolveram um novo método para cálculo da correlação entre dimensões que busca encontrar a maior discriminação entre os atributos. A seguir apresenta-se uma breve descrição deste método:

1. Dado um conjunto de dados com m dimensões;
2. Normaliza-se os valores em respeito às colunas (invariância contra escala e translação);
3. Para cada par de dimensões $Par(i, j)$ com $1 \leq i \leq m$ e $i < j \leq m$, constrói-se um histograma da diferença entre os valores $Hist(i, j)$. O número de *bins* (classes) do histograma $numBins$ é uma constante definida pelo usuário;
4. Para $k = 1$ até $numBins$ calcula-se Var_k:
   (a) Constrói-se a matriz M_k. A posição $M_k(i, j)$ da matriz será dada pelo valor 1 subtraído da razão entre a população contida em $k-$classes mais frequentes de $Hist(i, j)$ e o total de elementos;
   (b) Var_k corresponde à variância dos elementos não diagonais da matriz;
5. Retorna-se M_k que apresenta maior Var_k.

Este método proposto foi comparado à Distância Euclidiana entre os elementos e mostrou-se que o novo método apresenta um aumento na discriminação entre os atributos de 45% a 95% maior. Ou seja, utilizando este método, os autores conseguiram separar melhor dimensões diferentes e agrupar melhor as que apresentam certa semelhança. No entanto, não foram realizadas comparações com outras medidas de correlação entre variáveis bem estabelecidas na literatura (para uma melhor discussão sobre essas medidas favor consultar a Subseção 3.3.1. Apesar dos autores mencionarem que o cálculo de uma medida de correlação não está vinculado ao processo da técnica de visualização, trata-se de uma etapa diretamente relacionada com a projeção dos dados, consequentemente está fortemente atrelada à qualidade do *layout* apresentado.

2.3.2 Brushing Dimensions

A exploração das relações entre as dimensões não precisa estar vinculada somente a representações visuais dos atributos do conjunto de dados. Turkay et al. (Turkay et al., 2011) propuseram um método de múltiplas visões que permite que o usuário interaja tanto com as dimensões quanto com os itens da base de dados. O principal mecanismo de interação é a seleção que se reflete em outras visões e permite que se visualize, por exemplo, as dimensões que melhor representam subconjuntos dos dados. Uma das limitações deste trabalho é falta de medidas que consideram pares de dimensões, como medidas de correlação, o que dificulta a observação de dependências entre os atributos.

2.4 Considerações Finais

Capítulo

3

Proposta de Trabalho

3.1 Considerações Iniciais

A proposta deste trabalho gira em torno do conceito de *projeções multidimensionais coordenadas entre itens e atributos*.

A metodologia proposta por este trabalho é inspirada em uma combinação entre as projeções de atributos de (Yang et al., 2004) e às visões duplamente coordenadas entre itens e atributos de (Turkay et al., 2011).

O processo de projeção dos atributos inicia pela criação de uma matriz de distâncias que busca capturar os relacionamentos entre as dimensões. Em seguida, com base nessa matriz, aplica-se um método semelhante a MDS para posicionar os atributos em um plano bidimensional. Diferentemente de (Yang et al., 2004), aqui as dimensões não serão representadas por glifos, mas sim por pontos. Apesar de glifos serem capazes de fornecer informações adicionais sobre os dados, no caso de projeções eles acentuam o problema de sobreposição de elementos.

A etapa de coordenação entre visões dos itens e atributos será realizada

3.2 Objetivos

Dentro do contexto apresentado anteriormente, o seguinte parágrafo representa a declaração dos objetivos deste trabalho de mestrado:

> *"Este projeto de mestrado tem como objetivo desenvolver mecanismos interativos sobre projeções multidimensionais coordenadas entre atributos e itens que auxiliem*

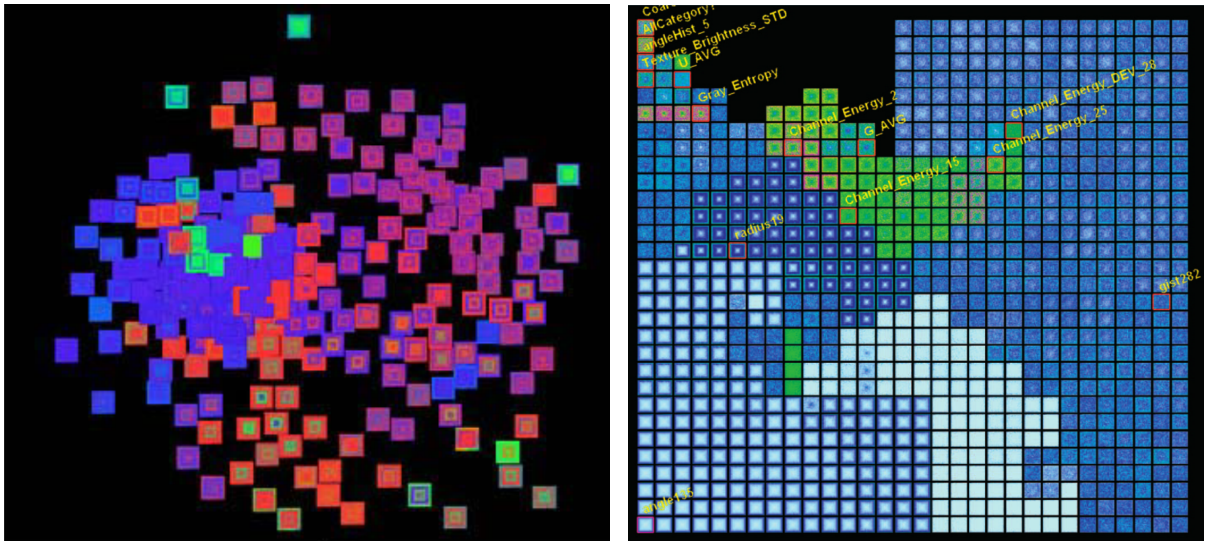
o usuário na tarefa de redução de dimensionalidade. Os mecanismos devem permitir tanto a seleção quanto a combinação de atributos. Caso os resultados não reflitam o conhecimento do usuário, este poderá manipular as projeções para transformar o espaço de atributos de modo a criar um modelo mais representativo."

As seguir apresenta-se os passos necessários para atingir esses objetivos.

3.3 Metodologia

A metodologia proposta por este trabalho é ilustrada pela Figura 3.1. O processo inicia pelo cálculo da similaridade entre as dimensões. Com base neste cálculo, cria-se uma matriz de distâncias entre os atributos que é utilizada para projetá-los em um espaço bidimensional.

inicia pela escolha de uma medida para o cálculo da similaridade entre pares de dimensões. Com base nessas distâncias projeta-se as dimensões utilizando alguma técnica de projeção multidimensional, como MDS. Neste ponto permite-se que o usuário utilize os mecanismos interativos de redução de dimensionalidade e de transformação do espaço de atributos. Ele pode, por exemplo, construir um espaço dimensional reduzido que é prontamente apresentado em visualizações coordenadas. Finalmente, se o resultado obtido não for satisfatório, o usuário pode iniciar novamente o ciclo partindo do novo espaço dimensional construído, ou pode realizar novas manipulações sobre os dados.



(a) **(b)**

Figura 3.1: asd

3.3.1 Similaridade entre Dimensões

Uma tarefa fundamental para a execução deste trabalho é maneira como define-se a similaridade entre as dimensões. Este problema pode ser formulado da seguinte maneira (Ankerst et

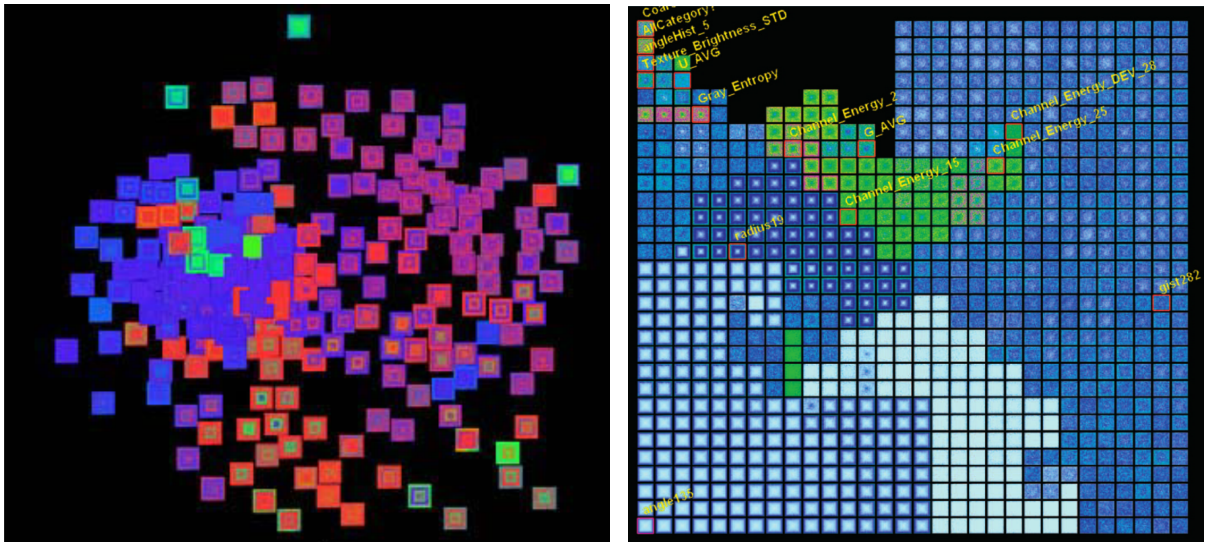
al., 1998): Dado um conjunto de dados contendo n elementos com $m-$dimensões, suas colunas podem ser descritas por m vetores A_i $(0 \leq i < m)$, cada uma contendo n números reais $a_{i,k}, (0 \leq k < n)$. Deseja-se definir uma medida de similaridade S que dado dois vetores retorne um número real $(S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R})$ que satisfaça as seguintes propriedades:

1. Positividade: $\forall A_i, A_j \in \mathbb{R}^n : S(A_i, A_j) \geq 0$
2. Reflexividade: $\forall A_i, A_j \in \mathbb{R}^n : (A_i = A_j) \Leftrightarrow S(A_i, A_j) = 0$
3. Simetria: $\forall A_i, A_j \in \mathbb{R}^n : S(A_i, A_j) = S(A_j, A_i)$, onde $(0 \leq i, j < d)$.

Uma medida de similaridade entre duas variáveis x e y muito utilizada é o coeficiente de correlação linear de Pearson p, dado por:

$$p(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}, \tag{3.1}$$

onde *var* corresponde à variância e *cov* à covariância. Quando x e y são completamente dependentes entre si, o valor de $p(x, y)$ é 1 ou -1. Caso não exista nenhuma dependência linear entre as variáveis, o valor obtido é 0. Porém, nos conjuntos de dados as relações entre as variáveis nem sempre são lineares e daí que surge a maior limitação deste coeficiente. Como pode ser observado pela Figura 3.2 e pelos valores apresentados na Tabela 3.1, a medida não é capaz de capturar dependências não lineares entre as variáveis.



(a) **(b)**

Figura 3.2: asd

Na literatura encontram-se diversos outros métodos que poderiam ser utilizados para o cálculo de similaridade. Métodos de regressão (Friedman et al., 2001; Cleveland et al., 1988; Stone, 1977) têm bom desempenho quando as relações podem ser descritas por uma função,

Tabela 3.1: Cronograma de Atividades. As marcações em preto indicam atividades que são priorizadas no período.

mas além desta situação falham em encontrar até os mais simples relacionamentos. Os métodos baseados em curvas principais (Hastie et al., 1989; Tibshirani, 1992; Delicado et al., 2008) e outros de correlação (Rényi, 1959; Breiman et al., 1985; Kosorok, 2009) são aplicáveis a um domínio mais abrangente de dados, porém não conseguem capturar os relacionamentos tão eficientemente quanto a medida MIC (Reshef et al., 2011) mesmo para casos simples, como pode ser observado pela Tabela 3.1.

A qualidade de uma medida de similaridade costuma variar de acordo com o domínio em que é aplicada. Normalmente, uma medida é dita adequada quando há uma concordância entre o valor obtido e a opinião de um especialista da área. No entanto, mesmo um especialista sobre um assunto pode ter dificuldade em determinar com precisão a semelhança entre dois objetos. Assim, é muito difícil definir um modelo matemático que meça a similaridade entre dois atributos com precisão para todas as aplicações de forma genérica. A MIC é uma medida que se propõe a executar esta difícil tarefa e por isso foi escolhida como a medida de similaridade entre dimensões a ser utilizada neste trabalho.

3.3.2 Mapeamento no Espaço Bidimensional

Uma vez definido o cálculo de similaridade entre as dimensões, cria-se uma matriz de distâncias

3.3.3 Visualização da Incerteza

3.3.4 Mecanismos de Interação

3.3.5 Forma de Avaliação

A forma mais adotada na literatura para a avaliação de métodos de redução de dimensionalidade é a comparação dos erros obtidos em tarefas de classificação ao utilizar diferentes técnicas.

Ao reduzir o número de atributos irrelevantes ou redundantes, pode-se melhorar o desempenho computacional e a precisão das técnicas operando sobre os dados, como agrupadores e classificadores de dados. Pretende-se avaliar as contribuições deste trabalho justamente pela quantificação do desempenho de tais métodos ao utilizar as técnicas desenvolvidas, seguida de uma comparação com técnicas já estabelecidas na literatura.

3.4 Resultados Esperados

3.5 Resultados Preliminares

3.6 Plano de Atividades e Cronograma Previsto

As principais atividades deste trabalho de mestrado são as seguintes:

1. Cumprimento dos créditos das disciplinas exigidos pelo programa;
2. Exame de proficiência em língua inglesa;
3. Levantamento bibliográfico sobre técnicas de visualização computacional e redução de dimensionalidade;
4. Levantamento bibliográfico sobre técnicas visuais interativas para redução de dimensionalidade;
5. Adoção e implementação de uma metodologia para o cálculo da similaridade entre dimensões;
6. Escrita da monografia de qualificação e sua apresentação para uma banca avaliadora;
7. Implementação de um modelo visual que transmita simultaneamente informações sobre os itens e dimensões de uma base de dados;
8. Desenvolvimento de mecanismos de seleção e combinação para redução interativa de dimensionalidade;
9. Desenvolvimento de um mecanismo interativo para transformação do espaço de atributos;
10. Avaliação dos Resultados;
11. Redação de artigos científicos e participação em congressos e eventos;
12. Escrita da dissertação de mestrado bem como sua apresentação para uma banca avaliadora;

O cronograma de execução das atividades é apresentado na Tabela 3.2, assumindo um projeto de duração de vinte e quatro meses.

Tabela 3.2: Cronograma de Atividades. As marcações em preto indicam atividades que são priorizadas no período.

	2012		2013		2014	
Atividade	1º S.	2º S.	1º S.	2º S.	1º S.	2º S.
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						

Bibliografia

Ankerst, M., S. Berchtold e D. Keim (1998). "Similarity clustering of dimensions for an enhanced visualization of multidimensional data". Em: *Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258)*. IEEE Comput. Soc, pp. 52–60, ISBN: 0-8186-9093-3. DOI: 10.1109/INFVIS.1998.729559. ENDEREÇO: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=729559.

Breiman, L. e J. H. Friedman (1985). "Estimating Optimal Transformations for Multiple Regression and Correlation". Em: *Journal of the American Statistical Association* 80.391, pp. 580–598. DOI: 10.1080/01621459.1985.10478157. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1985.10478157. ENDEREÇO: http://www.tandfonline.com/doi/abs/10.1080/01621459.1985.10478157.

Cleveland, W. S. e S. J. Devlin (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting". Em: *Journal of the American Statistical Association* 83.403, pp. 596–610. DOI: 10.1080/01621459.1988.10478639. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478639. ENDEREÇO: http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478639.

Delicado, P. e M. Smrekar (set. de 2008). "Measuring non-linear dependence for two random variables distributed along a curve". Em: *Statistics and Computing* 19.3, pp. 255–269. ISSN: 0960-3174. DOI: 10.1007/s11222-008-9090-y. ENDEREÇO: http://www.springerlink.com/index/10.1007/s11222-008-9090-y.

Friedman, J, T Hastie e R Tibshirani (2001). *The elements of statistical learning*. ENDEREÇO: http://www-stat.stanford.edu/~tibs/book/preface.ps.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition (2nd ed.)* San Diego, CA, USA: Academic Press Professional, Inc. ISBN: 0-12-269851-7.

Hastie, T. e W. Stuetzle (1989). "Principal Curves". Em: *Journal of the American Statistical Association* 84.406, pp. 502–516. DOI: 10.1080/01621459.1989.10478797. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1989.10478797. ENDEREÇO: http://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478797.

Konig, A. (2000). "Dimensionality reduction techniques for multivariate data classification, interactive visualization, and analysis-systematic feature selection vs. extraction". Em: *. . . -Based Intelligent Engineering Systems and Allied . . .* ENDEREÇO: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=885757.

Kosorok, M. R. (jan. de 2009). "On Brownian Distance Covariance and High Dimensional Data." Em: *The annals of applied statistics* 3.4, pp. 1266–1269. ISSN: 1932-6157. DOI: 10.1214/09-AOAS312. ENDEREÇO: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2889501\&tool=pmcentrez\&rendertype=abstract.

Reshef, D. N. et al. (dez. de 2011). "Detecting novel associations in large data sets." Em: *Science (New York, N.Y.)* 334.6062, pp. 1518–24. ISSN: 1095-9203. DOI: 10.1126/science.1205438. ENDEREÇO: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3325791\&tool=pmcentrez\&rendertype=abstract.

Rényi, A. (1959). "On measures of dependence". English. Em: *Acta Mathematica Academiae Scientiarum Hungarica* 10 (3-4), pp. 441–451. ISSN: 0001-5954. DOI: 10.1007/BF02024507. ENDEREÇO: http://dx.doi.org/10.1007/BF02024507.

Stone, C. J. (jul. de 1977). "Consistent Nonparametric Regression". Em: *The Annals of Statistics* 5.4, pp. 595–620. ISSN: 0090-5364. DOI: 10.1214/aos/1176343886. ENDEREÇO: http://projecteuclid.org/euclid.aos/1176343886.

Tibshirani, R. (dez. de 1992). "Principal curves revisited". Em: *Statistics and Computing* 2.4, pp. 183–190. ISSN: 0960-3174. DOI: 10.1007/BF01889678. ENDEREÇO: http://link.springer.com/10.1007/BF01889678.

Turkay, C., P. Filzmoser e H. Hauser (2011). "Brushing Dimensions; A Dual Visual Analysis Model for High-Dimensional Data". Em: *Visualization and Computer . . .* 17.12. ENDEREÇO: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6065027.

Yang, J. et al. (2004). "Value and Relation Display for Interactive Exploration of High Dimensional Datasets". Em: *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pp. 73 –80. DOI: 10.1109/INFVIS.2004.71.

Yang, J. et al. (2007). "Value and relation display: interactive visual exploration of large data sets with hundreds of dimensions." Em: *IEEE transactions on visualization and computer graphics* 13.3, pp. 494–507. ISSN: 1077-2626. DOI: 10.1109/TVCG.2007.1010. ENDEREÇO: http://www.ncbi.nlm.nih.gov/pubmed/17356216.