Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Departamento de Engenharia Elétrica

Hugerles Sales Silva

Redução de dimensões utilizando esboços

Campina Grande, Paraíba

Julho, 2014

Hugerles Sales Silva

Redução de dimensões utilizando esboços

Relatório de Estágio Supervisionado submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Processamento Digital de Sinais

Orientador: Edmar Candeia Gurjão

Campina Grande, Paraíba

Julho, 2014

Hugerles Sales Silva

Redução de dimensões utilizando esboços

Relatório de Estágio Supervisionado submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências no Domínio da Engenharia Elétrica.

Edmar Candeia Gurjão
Orientador

Luciana Ribeiro Veloso
Avaliador

Campina Grande, Paraíba
Julho, 2014

Ao meu avô José Pinto (in memorian), a quem dedico todo meu amor, gratidão e eterna saudade.

Agradecimentos

Tal como um trem, que entre a estação de partida e a de chegada encontra estações intermediárias, assim foi a construção deste relatório, onde pessoas especiais contribuíram durante minha jornada, cada qual em uma estação.

Ao mundo e a Campina Grande, por terem me ensinado a viver e a sobreviver no meio da selva!

A Bonfim e Rosária, meus pais, por terem me ensinado a caminhar!

A Lucinha, a qual considero minha segunda mãe, pelo grande impulso dado no início da minha graduação.

Ao meu tio Gilberlânio e minha avó Francisca, por tudo!

Aos professores Wamberto Queiroz e Patrícia Leal por me fazerem despertar para a área de Telecomunicações. A José Ewerton por fornecer a base de quase tudo que aprendi na área. A Edmar Candeia Gurjão por sanar todas as minhas dúvidas, sejam pessoais ou profissionais.

Aos colegas de curso, em especial ao quinteto: Eduardo Souza, Rodrigo Almeida, Rafael de Melo, Raphael Borges e Herbet Filipe; por partilhar momentos de alegrias, de festas, de noites mal dormidas, de mal humor, de notas baixas e altas. Uma amizade que se propaga além da sala de aula e que eu levo pra minha vida toda.

"Não deve prometer andar na escuridão aquele que não viu o anoitecer."

- J. R. R. Tolkien

Resumo

Neste relatório será apresentado um estudo sobre métodos para redução de dimensão, especialmente o de projeção aleatória, baseado no lema de Johnson-Lindenstrauss para criar esboços de grandes conjuntos de dados, mas que preserve, em grande medida, as mesmas características dos dados originais. Uma revisão bibliográfica, além da teoria básica de métodos para extração de atributos em um conjunto de dados e de obtenção de esboços de matrizes são apresentados. Para concluir, algumas aplicações na análise de dados na área da saúde e meteorológicos são evidenciadas e determinados parâmetros como: tempo de processamento para encontrar determinadas tendências, erro médio absoluto, entre outros, são analisados e comparados.

Palavras-chaves: Esboço. Redução de dimensionalidade. Projeção aleatória. Lema de Johnson-Lindenstrauss.

Abstract

This report presents a study on methods for dimensionality reduction, specially the random projection, based on Johnson-Lindenstrauss lemma, to create sketches of big set of data, maintaing the same characteristics as the original data. One literature review, beyond the basic theory of methods to the extraction of attributes in one set of data and the obtainment of data sketches. Applications in the analyses of data in the area of health care and meteorology are evidenced and determined parameters like: processing time to find certain tendencies, absolute medium error, among others, are analysed and compared.

Keywords: Sketch. Dimensionality reduction. Randon projection. Johnson-Lindenstrauss lemma.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

LAPSI	Laboratório de Automação e Processamento de Sinais e Informação
FFT	Transformada Rápida de Fourier
PR	Projection Random
JL	Johnson-Lindenstrauss
PCA	Análise de Componentes Principais
ARE	Erro Absoluto Relativo
DWT	Transformada Wavelet Discreta
DCT	Transformada do Cosseno Discreta
SVD	Decomposição de Valor Singular
DSP	Processamento Digital de Sinais
DFT	Transformada de Fourier Discreta
WFT	Windowed Fourier Transform
UFCG	Universidade Federal de Campina Grande

Sumário

1 Introdução

Muitas fontes de dados são observações que evoluem ao longo do tempo gerando uma série de dados temporal. Em muitos casos, estas séries temporais possuem um número bastante elevado de atributos ou de observações, fazendo com que a análise completa destes dados possam ficar computacionalmente custosa se todo o conjunto for considerado. Alguns algoritmos de processamento podem não apresentar resultados satisfatórios dependendo do tamanho do conjunto de dados em análise. Uma solução para isto é fazer uma síntese de dados.

A redução de dados consiste em obter uma representação reduzida do conjunto de dados, que seja muito menor em volume, mas que produza os mesmos (ou quase os mesmos) resultados obtidos ao se analisar os dados originais. Existem várias estratégias para redução de dados, dentre as quais destacam-se na literatura: agregação, amostragem, sintetização de dados, discretização e hierarquia de conceito [17].

Inúmeras técnicas de mineração de dados não são eficientes para dados com alta dimensão devido à "maldição da dimensionalidade" e também porque a precisão e a eficiência de uma consulta degradam rapidamente na medida em que a dimensão aumenta. A redução de dimensionalidade mostra eficiência: no armazenamento e recuperação quando feita a compressão de dados, no desempenho dos algoritmos, na redução do custo computacional e no relacionamento entre os atributos.

Na redução de dimensões é necessário fazer a extração de atributos. A idéia desta extração é que dado um conjunto de pontos no espaço d-dimensional, projeta-se este conjunto de pontos em um espaço de dimensão menor, preservando ao máximo as informações dos dados originais. Em particular, escolhe-se uma projeção que minimize o erro médio quadrático na reconstrução dos dados originais. Existem muitos métodos de extração de características, dentre eles estão: a projeção aleatória - RP, a análise de componentes principais - PCA, a decomposição do valor singular - SVD, entre outros.

Este trabalho utiliza a projeção aleatória para retirar características dos conjuntos de dados utilizando esboços (do inglês, *sketches*). Este método baseia-se no lema de Johnson-Lindenstrauss [6] e tem como vantagens: o baixo custo computacional e a facilidade de implementação. Segundo Johnson-Lindenstrauss, é possível fazer uma redução de dados considerando um erro ϵ tolerável. A redução de dimensão de dados é aplicada em processamento textual, recuperação de informação em banco de imagens, análise de dados em *microarrays*, classificação de proteínas, reconhecimento facial, dados meteorológicos, química combinatorial, etc.

Neste contexto foi desenvolvido este estágio supervisionado, visando responder à

problemas encontrados com o uso de grandes matrizes de dados, e ajudar nos desafios de engenharia que são encontrados no dia-a-dia, além de obter uma experiência com o ambiente, organização e modo de funcionamento do setor do estágio.

1.1 O Laboratório

A atividade relatada neste trabalho foi desenvolvida durante o estágio realizado no Laboratório de Processamento de Sinais e Informação - LAPSI, localizado no bloco CJ do Departamento de Engenharia Elétrica da Universidade Federal de Campina Grande, cuja coordenadora é a professora Luciana Veloso. O estágio foi realizado durante o período de 09 de junho de 2014 à 25 de julho de 2014, com carga horária de 180 horas e atendendo os requisitos previstos na Resolução N° 01/2012 do Colegiado do Curso de Graduação de Engenharia Elétrica e em consonância com a Lei do Estágio (Lei N° 11.788/2008).

O laboratório conta com computadores com sistemas operacionais Linux Ubuntu e Windows, plataformas para desenvolvimentos de rádios definidos por software, DSPs da Texas Instruments e instrumentos de medição, como osciloscópios e analisadores de espectro.

Além das pesquisas de iniciação científica e tecnológica, trabalhos de conclusão de curso e estágios, são também desenvolvidas no LAPSI atividades relativas a pós-graduação. As atuais linhas de pesquisa do LAPSI se concentram nas áreas de Rádio Definido por Software, Processamento de Sinais e Amostragem Compressiva.

1.2 Objetivos do estágio

O objetivo deste estágio foi o de utilizar projeções aleatórias com base no lema de Johnson-Lindenstrauss para criar esboços de grandes matrizes de dados. Determinadas características são observadas nas matrizes originais e nos seus respectivos esboços e parâmetros como: tempo de processamento para encontrar determinadas tendências, erro absoluto, entre outros, são analisados e comparados.

2 Fundamentação Teórica

2.1 Redução de dimensão

Muitos serviços geram dados em muitas dimensões, por exemplo, leituras meteorológicas de um determinado estado brasileiro durante uma década, consumo mensal de residências em uma cidade realizados por Smart Grids, classificação de documentos, reconhecimentos de dígitos manuscristos, entre outros. Inúmeros algoritmos ou técnicas utilizadas para fazer o processamento de grandes dados funcionam muito bem em uma dimensão baixa, porém se deterioram quando escalados para um espaço dimensional elevado. Este fênomeno é chamado de "maldição da dimensionalidade" e estabelece que à medida que o número de atributos medidos cresce em relação ao número de amostras, os pontos de dados tendem a ser equidistantes e seu espaço subjacente torna-se cada vez mais esparso. Isto gera efeitos negativos na maioria das metodologias de análise de dados, como aquelas de reconhecimento de padrões e aprendizagem de máquina que são baseadas em distâncias.

A redução de dimensão atualmente é um tema de grande interesse, cujo objetivo é reduzir a quantidade de dimensões em uma estrutura que preserve, em grande medida, as mesmas características dos dados originais. Uma visão intuitiva deste processo é mostrado na Figura 1. A análise de dados com muitos atributos torna-se complexa e pode ficar muito cara computacionalmente se todo o conjunto de dados for considerado. Em muitas aplicações de mineração de dados, a alta dimensionalidade dos dados restringe a escolha de métodos de processamento de dados.

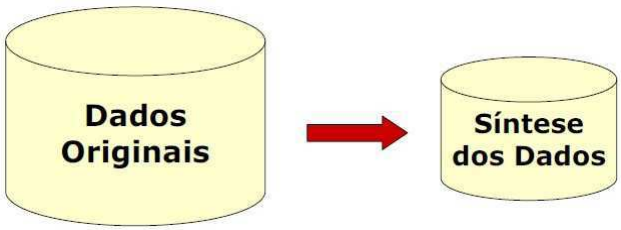

Figura 1 – Visão intuitiva da síntese de dados.

Fonte: [17].

Algumas estratégias para redução de dados tem sido destacadas na literatura, dentre as quais citam-se a

- Agregação - Tem como objetivo combinar dois ou mais atributos (ou objetos) em um atributo único. Esta estratégia tem como finalidade reduzir, fazer uma granu-

laridade (por exemplo, cidades agregadas em estados, regiões ou países) e analisar determinadas tendências dos dados.

- Amostragem - É geralmente usada em investigações preliminares de dados e também na análise final. O princípio chave da amostragem é que uma amostra produzirá resultados de qualidade semelhante aqueles produzidos pelo conjunto de dados completo se a amostra for representativa. Uma amostra é representativa se ela tem aproximadamente as mesmas propriedades (de interesse) do conjunto de dados original.

- Sintetização de dados - O conjunto de dados pode ser reduzido por meio de uma representação adequada para os dados. A sintetização pode ser feita por métodos paramétricos (regressão linear e regressão múltipla) e não-paramétricos (histogramas, clusterização, amostragem).

A principal abordagem de reduzir os dados é fazer uma extração de atributos. Este é um processo que define novas dimensões em função de todos os atributos do conjunto original, ou seja, os novos atributos são combinações lineares dos originais. A idéia consiste basicamente em projetar um conjunto de pontos em um espaço de menor dimensão, preservando ao máximo as informações do conjunto de dados originais. Existem inúmeros métodos propostos para esta tarefa. Neste trabalho, devido o fato da projeção aleatória (do inglês, *Randon Projection* - RP) ser computacionalmente simples e consistir de um mapeamento linear, este método será abordado.

2.1.1 Métodos para redução de dimensão

2.1.1.1 Análise das Componentes Principais - PCA

Na análise de componentes principais - PCA, a decomposição de autovalores da matriz de covariância dos dados é calculada como $E\{XX^T\} = A\Lambda A^T$, onde as colunas da matriz A são os autovetores da matriz de covariância dos dados $E\{XX^T\}$ e Λ é uma matriz diagonal contendo os respectivos autovalores.

Se a redução da dimensionalidade do conjunto de dados é desejada, os dados podem ser projetados em um subespaço gerado pelos autovetores mais importantes:

$$X^{PCA} = A_k^T X \tag{2.1}$$

onde a matriz A_k, de ordem $d \times k$ contém os k autovetores correspondentes aos k maiores autovalores. PCA é uma maneira ótima para projetar dados: o erro quadrado introduzido na projeção é minimizada sobre todas as projeções para um espaço k-dimensional. Infelizmente, a decomposição dos autovalores da matriz de covariância dos dados (cujo tamanho é $d \times d$ para dados d-dimensional) não é simples. A complexidade computacional

de analisar as componentes principais é $O(d^2N) + O(d^3)$. Existem métodos computacionalmente menos onerosos para encontrar apenas alguns autovetores e autovalores de uma grande matriz.

2.1.1.2 Transformação Linear Ortogonal

A Transformação Linear Ortogonal preserva distâncias L_p, onde p pode ser qualquer número inteiro positivo. A distância Euclidiana ($p = 2$) é a mais utilizada como medida de similaridade no esboço de matrizes. Esta transformação é satisfeita por qualquer transformação ortogonal. Entre este tipo de transformação estão a DFT, Wavelet e SVD. Tais técnicas de redução de dados seguem o esquema abaixo:

- Encontrar um conjunto de vetores V completos, normais e ortogonais do mesmo tamanho que as séries temporais;
- Transformar a série temporal para o espaço gerado por V;
- Manter as mais significativas coordenadas d ($d < n$).

As d primeiras coordenadas formam um vetor que é utilizado para aproximar a série temporal original. A escolha do limiar d definido pelo utilizador depende das características dos conjuntos de dados.

2.1.1.3 Tranformada de Fourier Discreta - DFT

A Transformada de Fourier Discreta foi utilizada pela primeira vez para redução da dimensão de séries temporais por Agrawal, Faloutsos e Swami [2]. Ela tem sido amplamente utilizada desde então na comunidade para extração de dados.

A DFT para essa aplicação tem seus prós e contras. No lado positivo, a DFT possui uma boa capacidade de comprimir a maioria dos sinais naturais, especialmente aqueles com tendências óbvias. O cálculo da DFT é rápido ($O(n^2)$). No entanto, a DFT suporta medidas ponderadas de distância.

2.1.1.4 Tranformada Wavelet Discreta - DWT

A Transformada de Fourier resume as características de frequências das séries temporais a partir de uma visão global. Como uma representação de séries temporais, a Wavelet é boa em compressão de sinais estacionários. A aproximação pode ser calculada de forma linear, mas a transformação Wavelet exige que os sinais devam ter um comprimento de $n = 2^{\text{inteiro}}$, caso contrário, a série temporal têm de ser preenchida, o que introduz um custo.

Acredita-se geralmente que DWT funciona para qualquer aplicação na qual DFT funciona. No entanto, Wu ao comparar DFT e DWT [11] afirmou que, embora a técnica baseada em DWT possua várias vantagens, DWT não reduz os erros de correspondência relativos ou aumenta a precisão em busca de similaridade. Técnicas baseadas em DFT e DWT produzem resultados comparáveis quando se busca similaridade em bancos de dados de séries temporais.

2.1.1.5 Decomposição de Valor Singular - SVD

A decomposição de valor singular - SVD é uma técnica de redução de dimensionalidade linear ótima. No entanto, é computacionalmente custosa. Ele requer um tempo $O(MN^2)$ e espaço $O(MN)$ em que M é o número de linhas de uma matriz, enquanto N é o número de colunas. Qualquer inserção no banco de dados requer o recálculo da transformação. SVD não suporta medidas de distância ponderadas ou medidas não euclidianas.

Usando SVD, a dimensão pode ser reduzida através da projeção dos dados para o espaço gerado pelos vetores singulares esquerdos correspondendo aos k maiores valores singulares:

$$X^{SVD} = U_k^T X \tag{2.2}$$

onde U_k é de tamanho $d \times k$ e contém esses k vetores singulares. $X = USV^T$ onde as matrizes ortogonais U e V contêm os vetores singulares a esquerda e à direita de X, respectivamente, e a diagonal de S contém os valores singulares de X.

Drinea e Huggins (2001) propuseram a abordagem de SVD aleatória. Eles afirmam que a amostragem das linhas ou colunas podem formar uma nova matriz com vetores singulares semelhantes às da matriz original.

As transformadas ortogonais diferem em suas propriedades. A DFT e a Transformada Wavelet são independentes dos dados, o que significa que a matriz de transformação é determinada a priori, enquanto transformadas dependentes de dados são utilizadas para aperfeiçoar dados específicos e, portanto poderá atingir um melhor desempenho, concentrando a energia em algumas características do vetor de características. Por outro lado, os algoritmos dependentes de dados sofrem com aumento no tempo de processamento. Devido à evolução de conjuntos de dados ao longo do tempo, um recálculo da matriz de transformação é necessário para evitar a degradação do desempenho.

Transformações independentes de dados (DFT e DWT) são utilizadas principalmente em algoritmos onde os dados mudam rapidamente, enquanto SVD encontra a sua aplicação onde os dados são atualizados de forma lenta.

2.1.1.6 Transformada do Cosseno Discreto - DCT

A Transformada do Cosseno Discreto - DCT é um método largamente utilizado para a compressão de imagens e, como tal, pode também ser usado na redução da dimensionalidade dos dados de imagem. DCT é computacionalmente menos onerosa do que PCA, porém apresenta desempenhos próximos. DCT pode ser realizada por simples operações matriciais: uma imagem é transformada no espaço DCT e a redução da dimensionalidade é realizada com uma transformação inversa, descartando os coeficientes de transformada correspondentes para as frequências mais altas. A complexidade computacional da transformada do cosseno discreto é da ordem de $O(dN \log_2(dN))$ para uma matriz de dados de tamanho $d \times N$.

2.1.2 Métodos para extração de atributos em um conjunto de dados: projeção aleatória

Em projeções aleatórias, os dados originais no espaço d-dimensional são projetados para um subespaço k-dimensional ($k << d$) utilizando uma matriz aleatória da ordem $k \times n$ com R colunas de magnitude 1. Usando a notação matricial onde $\mathbf{X}_{n \times q}$ é o conjunto original de q observações d-dimensional temos que

$$\mathbf{X}_{k \times q}^{RP} = \mathbf{R}_{k \times n} \mathbf{X}_{n \times q} \tag{2.3}$$

é a projeção de dados sobre um subespaço k-dimensional menor. A idéia chave do mapeamento aleatório decorre do lema de Johnson-Lindenstrauss: se os pontos em um espaço vetorial são projetados sobre um subespaço selecionado aleatoriamente de alta dimensão, então as distâncias entre os pontos são aproximadamente preservadas.

Projeção aleatória é computacionalmente muito simples: formar uma matriz aleatória $\mathbf{R}$ e projetar os $n \times q$ dados da matriz $\mathbf{X}$ em k dimensões é da ordem de $O(nkq)$, e se a matriz de dados $\mathbf{X}$ é esparsa com cerca de c entradas diferentes de zero por coluna, a complexidade é da ordem de $O(ckq)$.

Estritamente falando, a equação (2.3) não é uma projeção por que $\mathbf{R}$ não é geralmente ortogonal. Um mapeamento linear como em (2.3) pode causar distorções significativas no conjunto de dados se $\mathbf{R}$ não é ortogonal. Ortogonalizar $\mathbf{R}$ é computacionalmente caro. Em vez disso, podemos contar com um resultado apresentado por Hecht-Nielsen: em um espaço de alta dimensão, existe um número muito maior de direções quase ortogonais do que propriamente ortogonais. Assim, vetores podem ter direções aleatórias suficientemente próximas de ortogonais, e equivalentemente $\mathbf{R}^T\mathbf{R}$ seria aproximada por uma matriz identidade.

Ao comparar o desempenho da projeção aleatória ao de outros métodos de redução de dimensão, é instrutivo ver como a similaridade de dois vetores é distorcida na redução

de dimensionalidade. Medimos a similaridade de vetores de dados, quer como sua distância euclidiana ou como seu produto interno. No caso dos dados de imagens, a distância Euclidiana é uma medida utilizada de similaridade. Documentos de texto, por outro lado, são geralmente comparados de acordo com o cosseno do ângulo entre os vetores de dados.

A distância euclidiana entre dois vetores de dados $\mathbf{x}_1$ e $\mathbf{x}_2$ em um espaço de grande dimensão é dado como $||\mathbf{x}_1 - \mathbf{x}_2||$. Após a projeção aleatória, a distância é aproximada por uma distância Euclidiana escalada destes vetores no espaço reduzido por

$$\sqrt{d/k}||\mathbf{R}\mathbf{x}_1 - \mathbf{R}\mathbf{x}_2|| \tag{2.4}$$

onde d é a dimensão original, k é a dimensão reduzida do conjunto de dados, $\mathbf{R}$ é a matriz aleatória com distribuição normal $N(0,1)$ e $\mathbf{x}_1$ e $\mathbf{x}_2$ são dois vetores de dados. O termo escalado $\sqrt{d/k}$ leva em conta a diminuição do número de dimensões dos dados: de acordo com o lema de Johnson-Lindenstrauss, a norma esperada de uma projeção de um vetor unitário para um subespaço aleatório através da origem é $\sqrt{k/d}$.

A escolha da matriz aleatória R é um dos principais pontos de interesse. Neste relatório, adotamos R como tendo uma distribuição normal $N(0,1)$.

A projeção aleatória pode ser aplicada para mascarar dados, recuperar informação e fazer a redução de atributos representando os índices. Como um exemplo, vejamos a Tabela 1 que representa uma amostra do conjunto de dados de arritmia cardíaca da UCI *Machine Learninh Repository*. A Tabela 2 e a Tabela 3 são os dados originais transformados com base na distribuição normal e com base na distribuição mais simples, respectivamente. Estas tabelas evidenciam uma das aplicações de projeção aleatória, que é fazer a redução de atributos representando os índices.

ID	**Idade**	**Peso**	**Frequência Cardíaca**	**IntDef**	**QRS**	**PRint**
123	75	80	63	32	91	193
342	56	64	53	24	81	174
254	40	52	70	24	77	129
446	28	58	76	40	83	251
286	44	90	68	44	109	128

Tabela 1 – Amostra do conjunto de dados de arritmia cardíaca da UCI *Machine Learninh Repository*.

ID	Atr1	Atr2	Atr3
123	-50.40	17.33	12.31
342	-37.08	6.27	12.22
254	-55.86	20.69	-0.66
446	-37.61	-31.66	-17.58
286	-62.72	37.64	18.16

Tabela 2 – Dados originais transformados com base na distribuição normal.

ID	Atr1	Atr2	Atr3
123	-55.50	-95.26	-107.93
342	-51.00	-84.29	-83.13
254	-65.50	-70.43	-66.96
446	-85.50	-140.87	-72.74
286	-88.50	-50.22	-102.76

Tabela 3 – Dados originais transformados com base em uma distribuição diferente da normal.

2.2 Esboço de um vetor

Dado um vetor $\mathbf{t} = \{t[1], t[2] \cdots, t[l]\}$, o esboço (do inglês, *sketch*) será representado por $\mathbf{S}(t)$ e possui tamanho k. É escolhido um vetor aleatório $v_i[1, \cdots, l]$ onde cada componente $v_i[j]$ é uma variável aleatória independente com distribuição normal $N(0, 1)$ e v_i é normalizado com magnitude 1. Logo,

$$\mathbf{S}(t)[i] = \mathbf{t} \cdot \mathbf{v}_i = \sum_j t[j] \cdot v_i[j] \tag{2.5}$$

O esboço nada mais é do que o produto interno entre dois vetores.

Exemplo 2.1 - Seja $\mathbf{t} = (2, 1, 3, 1)$ e suponha que queiramos construir um esboço de tamanho dois. Os vetores aleatórios com distribuição normal $N(0, 1)$ são $\mathbf{v}_1$ e $\mathbf{v}_2$, onde

$$\mathbf{v}_1 = (-0.45, -0.09, 0.10, 0.87),$$

$$\mathbf{v}_2 = (-0.19, 0.73, -0.61, 0.21).$$

O esboço de $\mathbf{t}$ é $\mathbf{S}(t) = (0.18, -1.28)$.

Exemplo 2.2 - O esboço de um vetor é válido em muitas aplicações. Uma delas é na análise dos dados do consumo de água em residências. As linhas da Tabela 4 representam os meses do ano, enquanto as colunas a quantidade em litros do consumo mensal de cada casa. O esboço é de tamanho cinco, ou seja, há uma redução dos dados em 50% - ver Tabela 5. É escolhido o mês de fevereiro e calculado a média de consumo. A média dos

dados originais foi de 17,9 litros, enquanto que no esboço foi de 17,5 litros. O tempo de processamento foi 54% para o esboço e o resultado foi confiável, com um erro ϵ tolerável associado a média.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Jan	31	322	15	5	6	44	2	4	5	3
Fev	1	23	104	5	6	7	11	14	3	5
Mar	144	66	44	12	6	87	34	54	13	1
Abr	46	44	15	67	7	67	35	32	76	232
Mai	3	55	2	55	11	56	64	1	75	43
Jun	2	123	52	43	14	45	23	4	65	54
Jul	4	432	54	23	3	34	12	5	34	21
Ago	45	34	33	532	4	23	34	6	34	0
Set	76	62	31	11	44	12	1	34	23	100
Out	23	109	1	10	45	445	2	23	65	43
Nov	67	22	2	15	76	65	13	43	54	2
Dez	99	22	4	5	88	34	14	52	7	5

Tabela 4 – Consumo de água mensal em dez residências durante um ano.

	C1	C2	C3	C4	C5
Jan	2,9054	-119,5539	-352,7070	64,7758	-17,0609
Fev	2,2683	-59,9673	-86,9506	-28,2861	-2,0964
Mar	35,0372	-65,0204	-425,011	25,5675	8,7456
Abr	21,4514	-198,105	-427,007	41,7731	-3,8387
Mai	74,6445	-128,265	-397,849	-202,439	-37,0493
Jun	245.654	3,6586	30,0664	43,1726	79,8437
Jul	-2,0029	-7,529	-30,8787	11,6565	2,0942
Ago	89,8794	-152,646	-604,2	-115,006	-29,8587
Set	7,7227	-14,1725	-57,3678	-11,0521	-2,3724
Out	1,7676	-0,1159	-6,8728	4,2551	3,9785
Nov	-17,6151	-133,155	-429,567	7,2576	-46,7283
Dez	16,7979	66,4676	-32,2179	2,7023	29,2934

Tabela 5 – Esboço com redução de 50% dos dados.

O esboço de um vetor possui muitas propriedades. Muitas delas decorrem do lema de Johnson-Lindenstrauss.

Lema 2.1 (Johnson-Lindenstrauss) *Seja $\vec{v}_1$, $\vec{v}_2$,$\cdots$, $\vec{v}_m$ uma sequência de pontos no espaço d-dimensional sobre os reais e tenhamos ϵ, $F \in (0,1]$. Existe então um mapeamento linear f dos pontos do espaço d-dimensional para os pontos do espaço k-dimensional onde $k = O(log(1/F)/\epsilon^2)$ de tal modo que o número de vetores que aproximadamente preservam seu comprimento é pelo menos $(1-F)m$. Dizemos que um vetor $\vec{v}_i$ preserva*

aproximadamente o seu comprimento se:

$$||\vec{v}_i||^2 \leq ||f(\vec{v}_i)||^2 \leq (1+\epsilon)||\vec{v}_i||^2$$

Lema 2.2 (Johnson-Lindenstrauss modificado) *Para qualquer* $0 < \epsilon < 1$ *e qualquer inteiro* n, k *é um inteiro positivo tal que,*

$$k > 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \log n.$$

Então, para qualquer conjunto V *de* n *pontos pertencentes a* R^d, *existe um mapeamento* $f : R^d \to R^k$ *tal que para todo* $u, v \in V$

$$(1-\epsilon)||u-v||^2 \leq ||f(u)-f(v)||^2 \leq (1+\epsilon)||u-v||^2$$

com probabilidade 1/2.

Este teorema tem a propriedade adicional que aumentando k, pode-se aumentar a probabilidade de sucesso, conforme necessário.

Dimitris Achlioptas estendeu o lema de JL para distribuições diferentes da normal [2].

Lema 2.3 (Dimitris) *Seja* P *um conjunto arbitrário de* n *pontos pertencentes a* R^d *representado por uma matriz* A *da ordem* $n \times d$. *Dado* $\epsilon, \beta > 0$, *temos*

$$k_0 = \frac{4+2\beta}{\epsilon^2/2 - \epsilon^3/3} \log n.$$

Para um inteiro $k \geq k_0$, *temos* R *como sendo uma matriz aleatória de ordem* $d \times k$ *com* $R(i;j) = r_{ij}$, *onde* $\{r_{ij}\}$ *são variáveis aleatórias independentes a partir de qualquer uma das duas distribuições de probabilidades seguintes:*

$$r_{ij} = \begin{cases} +1 & \textit{com probabilidade } 1/2 \\ -1 & \textit{com probabilidade } 1/2 \end{cases}$$

ou

$$r_{ij} = \begin{cases} +\sqrt{3} & \textit{com probabilidade } 1/6 \\ 0 & \textit{com probabilidade } 2/3 \\ -\sqrt{3} & \textit{com probabilidade } 1/6 \end{cases}$$

Seja

$$E = \frac{1}{\sqrt{k}} AR.$$

Temos um mapeamento $f : R^d \to R^k$ *da* i*-ésima linha de* A *para a* i*-ésima linha de* E. *Com a probabilidade mínima de* $1 - n^{-\beta}$, *para todo* $u, v \in P$

$$(1-\epsilon)||u-v||^2 \leq ||f(u)-f(v)||^2 \leq (1+\epsilon)||u-v||^2.$$

2.2.1 Distância do esboço

A distância Euclidiana dos esboços de duas séries temporais é dada por

$$d_{sk} = ||\mathbf{x}_{sk} - \mathbf{y}_{sk}|| = \sqrt{(x_{sk1} - y_{sk1})^2 + (x_{sk2} - y_{sk2})^2 + \cdots + (x_{skk} - y_{skk})^2} \qquad (2.6)$$

onde $\mathbf{x}_{sk}$ e $\mathbf{y}_{sk}$ representam os respectivos esboços dos vetores de dados originais $\mathbf{x}_s$ e $\mathbf{y}_s$. As Figuras 2 e 3 mostram uma comparação entre a distância do esboço e a distância original sob um gráfico de curvas e barras. Nestas figuras, o tamanho do esboço é 64. Na Figura 2, a distância do esboço ocorre em uma faixa estreita em torno da distância real. Na Figura 3 é dada a relação entre a distância de esboço e a distância real. Esta relação se apresenta quase de forma simétrica, como um sino, e se parece com uma distribuição normal com centro no valor ideal. A Figura 4 indica a equivalência entre a distância de esboço e a distância real.

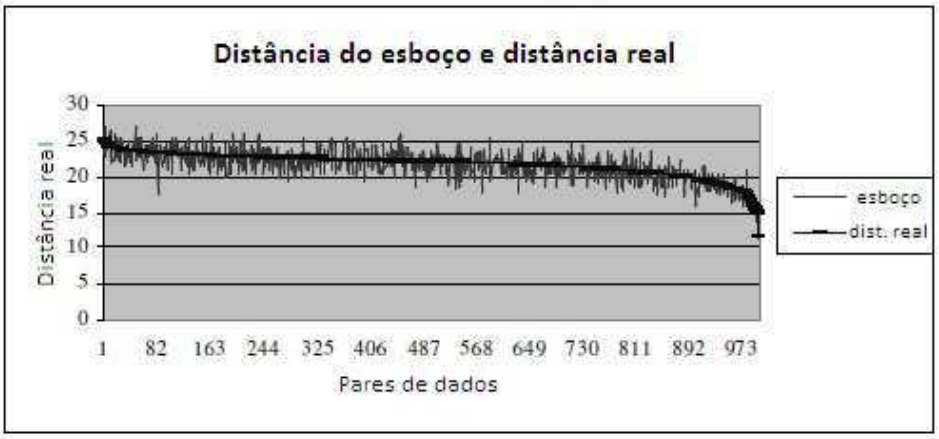

Figura 2 – Distância real *versus* Distância do esboço.

Fonte: [2].

2.2.2 Esboço com janela fixa

Nesta seção, focamos em computar o esboço para cada subvetor de um dado tamanho l em $T[1, \cdots, n]$. Isto é, nós precisamos calcular o esboço de $\lceil \frac{n}{l} \rceil$ vetores: $\mathbf{t}_1 = T[1, \cdots, l]$, $\mathbf{t}_2 = T[2, \cdots, l+1]$, $\cdots$, $\mathbf{t}_{n-l+1} = T[n-l+1, \cdots, n]$. Vamos fixar nossa atenção em uma das componentes, isto é, $S(t_i)[j]$ para cada vetor $\mathbf{t}_i$.

Inicialmente é gerado um vetor aleatório $\mathbf{v}_j$. Consideramos cada um dos vetores $\mathbf{t}_i$ e computamos $T[i, \cdots, i+l-1]$. Diretamente, $\mathbf{v}_j[1, \cdots, l]$. Isto requer um tempo $O(nl)$ já que existem $n-l+1$ vetores e que para cada um destes a complexidade de calcular o produto interno é $O(l)$. Nós agora repetimos o procedimento inteiro para todas as outras componentes do esboço de tamanho k, onde todo o procedimento leva um tempo total de $O(nlk)$.

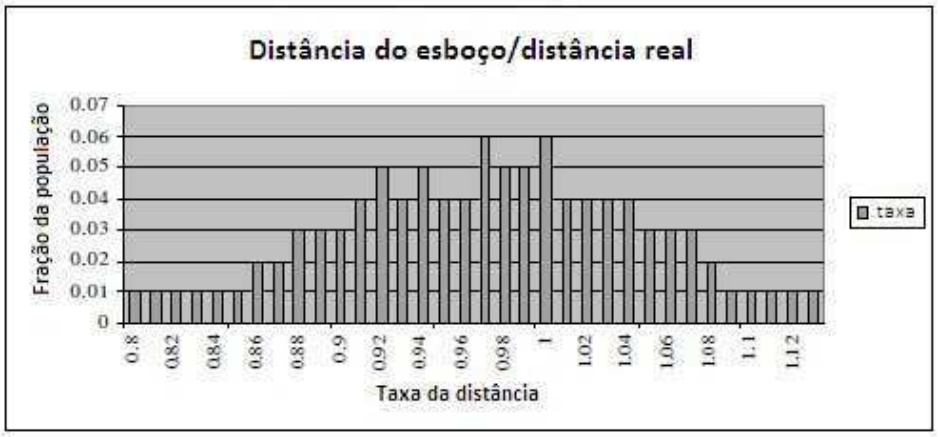

Figura 3 – Relação entre a distância do esboço/distância real.

Fonte: [2].

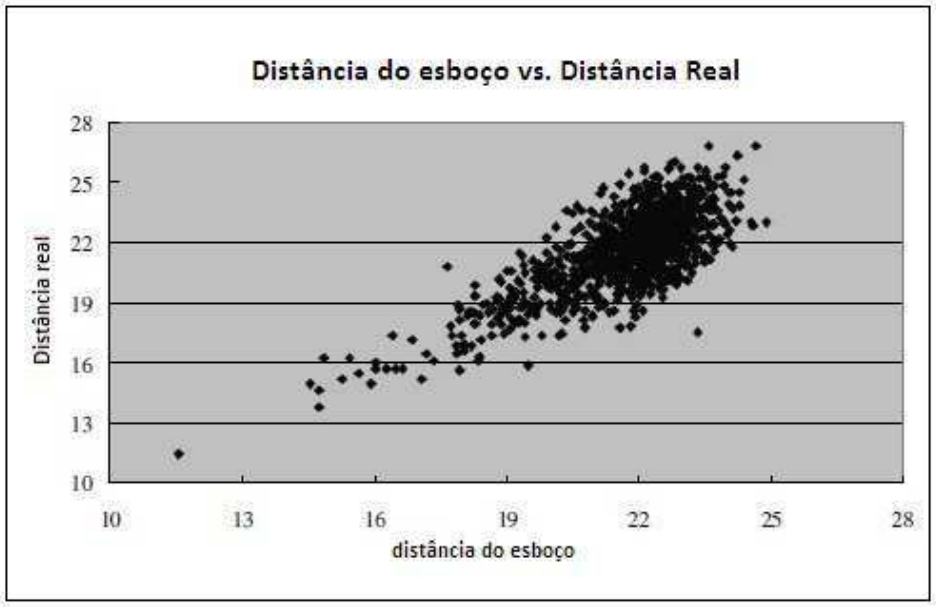

Figura 4 – GEMINI Framework.

Fonte: [2].

Uma observação importante é que podemos computar todos estes esboços utilizando a Transformada Rápida de Fourier - FFT. O problema de computar os esboços de todos os subvetores de tamanho l simultaneamente é precisamente o problema de calcular uma convolução polinomial de dois vetores $\mathbf{t}$ e $\mathbf{v}_j$. Esta observação é evidente quando consideramos a definição de convolução polinomial.

Teorema 2.1 (Convolução Polinomial) *Dados dois vetores* $A[1, \cdots, a]$ *e* $B[1, \cdots, b]$, $a \geq b$, *a convolução é o vetor* $C[1, \cdots, a+b]$ *onde* $C[k] = \sum_{1 \leq i \leq b} A[k-i] \times B[i]$ *para* $2 \leq k \leq a+b$, *com qualquer intervalo de referência considerado 0.*

Por exemplo, se $A = [1, 10, 2, 4]$ e $B = [7, 2]$, nós temos $C[7, 72, 34, 32, 8]$. Convolução polinomial de dois vetores pode ser computada em um tempo $O(a \log b)$ usando a FFT.

Lema 2.4 (Esboços de subvetores utilizando convolução polinomial) *Esboços de todos os subvetores de tamanho l podem ser computados em um tempo $O(nk \log l)$ usando convolução polinomial.*

A Figura 5 mostra um exemplo utilizando convolução polinomial para calcular os esboços. O vetor (2,1,3,1) é convoluído com dois vetores **t** com distribuição normal $N(0, 1)$ e norma 1. As mesmas coordenadas de todos os três esboços de comprimento dois são computados ao mesmo tempo.

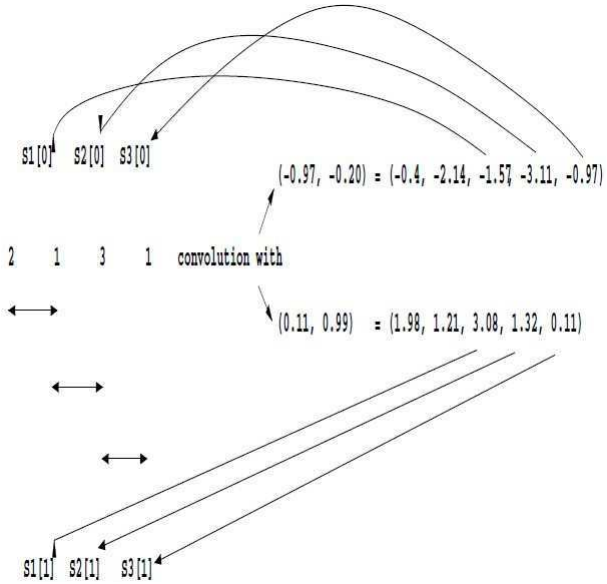

Figura 5 – Usando convolução polinomial para computar esboços.

Fonte: [4].

2.2.3 Computando esboços para uma faixa de subvetores

Nesta seção consideramos um problema mais geral de calcular esboços para qualquer subvetor de tamanho entre l e u de um grande vetor. Algumas abordagens consideram todos os possíveis subvetores e computa cada esboço diretamente. Tem-se $O(\sum_{l \leq i \leq u} \frac{n}{i})$ subvetores. No pior caso, isto é $O(n^2)$ subvetores e a maior parte deles são de tamanho $\theta(n)$, por isso, todo o algoritmo levará um tempo $O(n^3)$ que é proibitivo.

Para melhorar o desempenho estruturamos o seguinte algoritmo: iremos cuidadosamente construir um reservatório de esboços onde armazenaremos-os, onde este será

um pequeno subconjunto do conjunto de todos os esboços que precisamos. Após o pré-processamento, seremos capazes de determinar o esboço de qualquer subvetor do vetor original em um tempo O(1) bastante preciso.

Escolhemos $1 \leq l \leq u$ tal que l seja uma potencia de 2. Para cada l, computamos o esboço de todos os subvetores de $\vec{t}$. De fato, são computadas duas versões do esboço, cada uma usando diferentes variáveis aleatórias; que são chamadas de S^1 e S^2. O conjunto resultante dos esboços é o que nós chamamos de reservatório. Este possui tamanho total $O(n \log(u-l)k)$ e requer um tempo $O(n \log u \log(u-l)k)$ para computá-los; isto é $O(n \log^3 n)$ no pior caso.

Como segundo passo, determinamos o esboço para qualquer subvetor $\vec{t_i}$. Fixamos inicialmente uma componente de interesse, digamos $S(t_i)[j]$. Existem duas possibilidades:

1. $L = 2^r$, cujo caso nós temos os esboços para estes subvetores no reservatório, para procurar apenas o esboço de iteresse.
2. $2^r < L < 2^{r+1}$. Neste caso, computamos: $S'(T[i, \cdots, i+L-1])[j] = S^1(T[i, \cdots, i+2^r-1])[j] + S^2(T[i+L-2^r, \cdots, i+L-1])[j]$ - ver Figura 6.

É observado que S' satisfaz uma propriedade muito semelhante ao lema de JL.

Teorema 2.2 (Esboços de subvetores) *Para um dado conjunto de vetores L de tamanho l, ϵ fixo e $\epsilon < 0.5$, então $k = \frac{9 \log |L|}{\epsilon^2}$, entao para quaisquer pares de vetores $\vec{u}$, $\vec{w} \in L$*

$$(1-\epsilon)||u-v||^2 \leq ||f(u)-f(v)||^2 \leq 2(1+\epsilon)||u-v||^2$$

com probabilidade 1/2.

Note o fator 2 adicional na segunda desigualdade.

2.2.4 Avaliação da acurácia do esboço

Para comparar o desempenho dos resultados obtidos é utilizado o erro relativo absoluto.

Inicialmente comparamos a similaridade entre os resultados obtidos com a análise da matriz original e com o respectivo esboço. Os vetores de dados originais são chamados de D_1 e D_2, enquanto os seus esboços de s_1 e s_2, respectivamente. A medida de similaridade utilizada neste relatório é a norma L_2. Seja S_1 a medida de similaridade entre D_1 e D_2 e

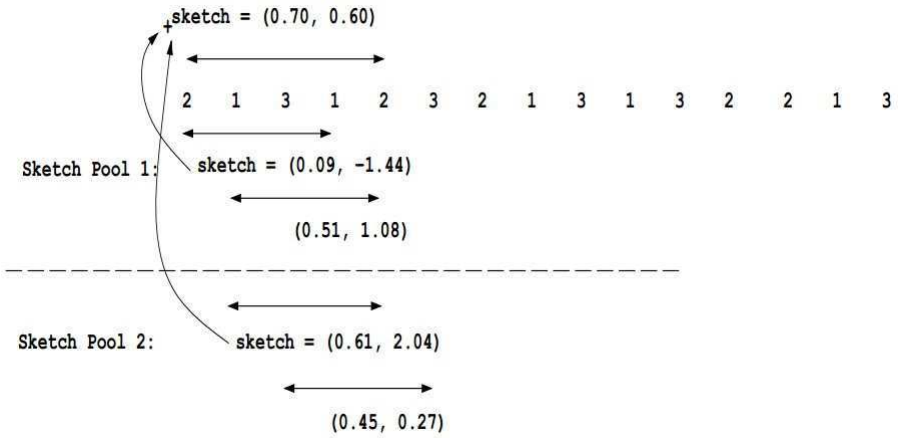

Figura 6 – Esboços para uma subsequência de tamanho 5 de um esboço de subsequência de tamanho 4.

Fonte: [4].

S_2 a medida de similaridade entre s_1 e s_2, então o erro relativo absoluto (ARE) é definido por

$$\text{ARE}(\%) = \frac{|S_1 - S_2|}{S_1} 100. \tag{2.7}$$

3 Resultados e Análises

3.1 Estudo de caso 1 - Aplicação do esboço na análise de dados meteorológicos

Com o crescente volume de informações e da capacidade computacional, apoiados pelo vertiginoso crescimento tecnológico da computação, a capacidade e disponibilidade de dados não é mais um empecilho para grande parte da comunidade científica. Um problema relevante é o uso de informações de modo eficiente e eficaz.

Na meteorologia o problema não é diferente, pois existem uma enorme quantidade de dados que servem de entrada para os modelos que fazem as previsões, sejam do tempo, de temperatura ou do clima, demandando grande quantidade de recursos computacionais e tempo.

A análise aqui requer encontrar determinadas tendências e processar dados meteorológicos utilizando esboços de matrizes. Os dados foram retirados da Estação Padrão de Belo Horizonte (Estação de Lourdes, $5^{\underline{o}}$ Distrito de Meteorologia), considerando dados a partir do ano de 1986 até o ano de 2004. Esta sequência de dados são apresentados na Tabela 6 e na Tabela 7.

Foi definida uma matriz D de ordem $m \times n$, onde m representa os anos, de 1986 à 2004 e n representa os meses de um ano. Adotou-se $m = 19$ e $n = 12$. Foi gerada uma matriz A aleatória com dimensões $n \times q$, com $q < 12$ e distribuição normal $N(0, 1)$, com média zero e variância um. O intuito é que ao realizar o produto $\mathbf{S}_{m\times q} = \mathbf{D}_{m\times n}\mathbf{A}_{n\times q}$ as informações contidas em $\mathbf{D}$ sejam repassadas para $\mathbf{S}$ e que a análise que seria feita na matriz original agora possa ser feita em uma matriz de menor dimensão. Nas figuras a seguir é interessante atentar para o fato de que foi plotado $\mathbf{S}^T$.

Inicialmente foi selecionado dois anos aleatoriamente, representados pelos vetores de dados D_1 e D_2 e calculado seus respectivos esboços, s_1 e s_2. O tamanho do esboço foi variado e para cada valor de m o esboço foi repetido 1000 vezes. A cada rodada uma nova matriz $\mathbf{A}$ foi gerada. A Figura 7 mostra os erros relativos médios em função do tamanho do esboço.

Foi definido dois conjuntos de similaridade. Um deles com os dados originais e outro com o esboço. Se ao comparar estes vetores fosse percebido que eles diferiam em 3 ou mais posições um erro seria declarado. A Figura 8 mostra a porcentagem de erros em função do tamanho do esboço. Como esperado, o erro decai quando m aumenta, e com $m = 11$ o erro é menor que 5% e apresenta um bom resultado tendo ocorrido uma

T(ºC)	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out
1986	25,86	24,92	24,07	25,56	21,37	19,07	18,33	20,51	20,36	23,01
1987	24,03	24,90	23,34	21,35	19,01	19,92	20,86	21,45	24,57	24,57
1988	24,44	23,50	24,66	22,75	21,70	18,59	17,48	17,47	22,70	22,12
1989	23,84	25,34	24,73	23,52	20,12	18,93	17,88	21,70	21,82	22,41
1990	24,63	23,87	24,33	23,75	20,19	19,39	19,41	18,81	21,82	22,41
1991	22,13	23,28	22,85	22,11	20,09	19,84	18,71	19,49	20,62	21,80
1992	21,92	23,39	23,02	23,22	22,05	21,99	19,03	19,85	22,20	21,57
1993	26,50	22,92	25,46	22,48	19,78	18,39	20,29	20,06	22,79	26,53
1994	22,79	25,61	22,29	21,93	21,41	19,20	19,62	19,77	23,13	23,34
1995	25,24	23,76	23,32	22,38	20,98	19,25	19,80	21,71	21,61	22,47
1996	24,00	24,31	24,12	23,61	19,77	18,78	18,82	19,39	21,60	26,42
1997	23,17	23,72	21,98	21,62	19,35	18,84	19,46	20,37	23,77	23,74
1998	24,25	25,37	24,64	23,52	20,82	18,69	24,44	26,00	26,96	22,20
1999	24,59	24,90	23,58	22,61	19,86	19,51	23,00	24,59	21,42	21,10
2000	23,37	23,40	22,94	23,36	20,29	19,27	23.15	21,14	22,29	21,93
2001	24,44	26,00	26,00	24,44	26,00	19,63	20,14	19,22	21,18	21,95
2002	23,00	24,59	24,59	23,00	24,59	23,39	22,92	25,61	26,00	21,44
2003	23,31	24,90	21,89	22,33	21,99	19,03	19,85	24,90	24,59	22,53
2004	23,35	23,00	25,00	22,55	19,85	20,71	19,35	22,14	21,89	24,90

Tabela 6 – Temperaturas médias mensais, em ºC, ao longo do período de anos.

T(ºC)	Nov	Dez
1986	22,84	22,59
1987	23,33	22,60
1988	22,76	22,88
1989	24,44	26,00
1990	24,44	26,00
1991	23,00	24,59
1992	23,38	21,89
1993	23,68	22,57
1994	22,72	23,25
1995	24,01	22,48
1996	21,89	22,48
1997	24,97	22,58
1998	23,55	24,22
1999	21,16	23,82
2000	21,41	22,57
2001	23,22	22,45
2002	22,48	24,59
2003	21,98	24,90
2004	21,35	23,40

Tabela 7 – Temperaturas médias mensais, em ºC, ao longo do período de anos (continuação da tabela anterior).

redução de dimensionalidade de 42,10%.

Em uma outra etapa de simulação foi retirada uma média de temperatura durante o ano de 1987 na matriz de dados original e depois feito uma média no esboço com redução de 50%. A média dos dados originais foi de 22,49º Celsius, enquanto que no esboço foi de 21,85º Celsius. O tempo de processamento foi 74,28% menor para o esboço e o resultado foi confiável, com um erro tolerável de 2,92% associado a média.

A Figura 9 mostra a variação do tempo em função do tamanho do esboço. Quanto maior o m, maior o tempo de processamento. Para $m = 11$, é medido um ganho de aproximadamente 64,7% no tempo de processamento da execução.

Em muitas aplicações nós podemos observar erros associados ao uso de esboços. No entanto, dependendo da aplicação estes erros podem ser desprezíveis.

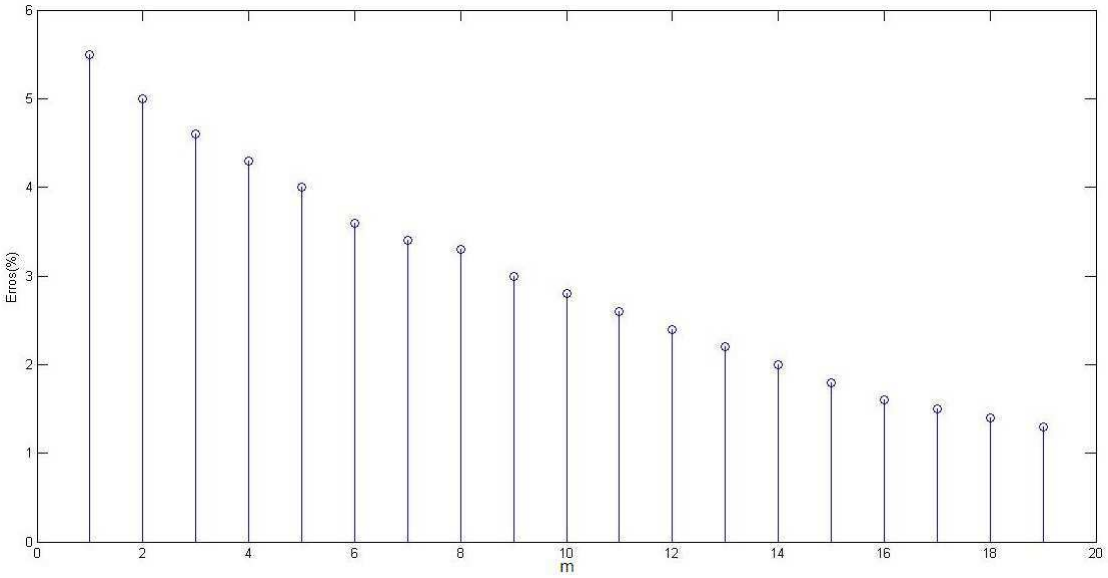

Figura 7 – Erro relativo em função do tamanho do esboço m.

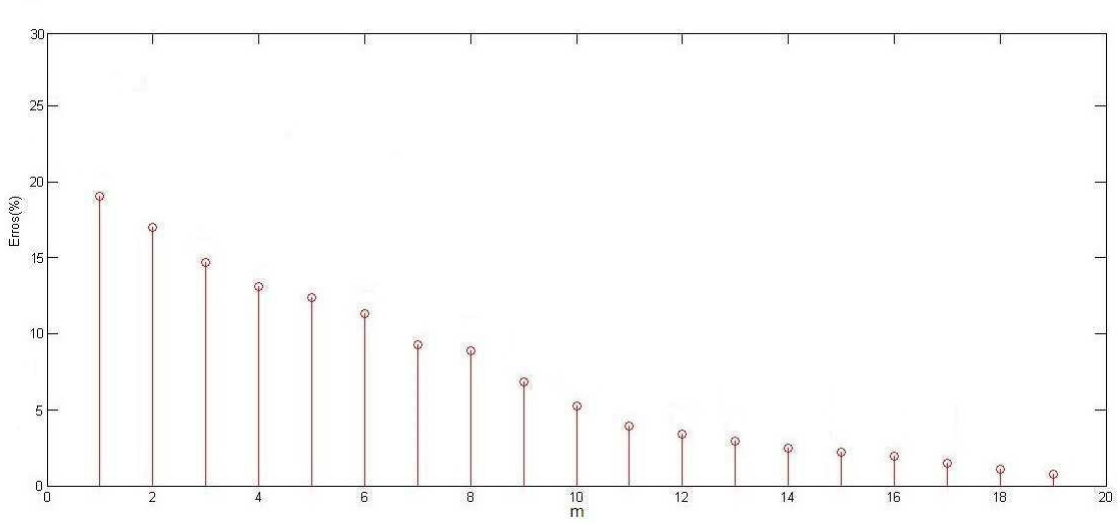

Figura 8 – Porcentagem de erros entre similaridades de conjuntos em função do tamanho do esboço m.

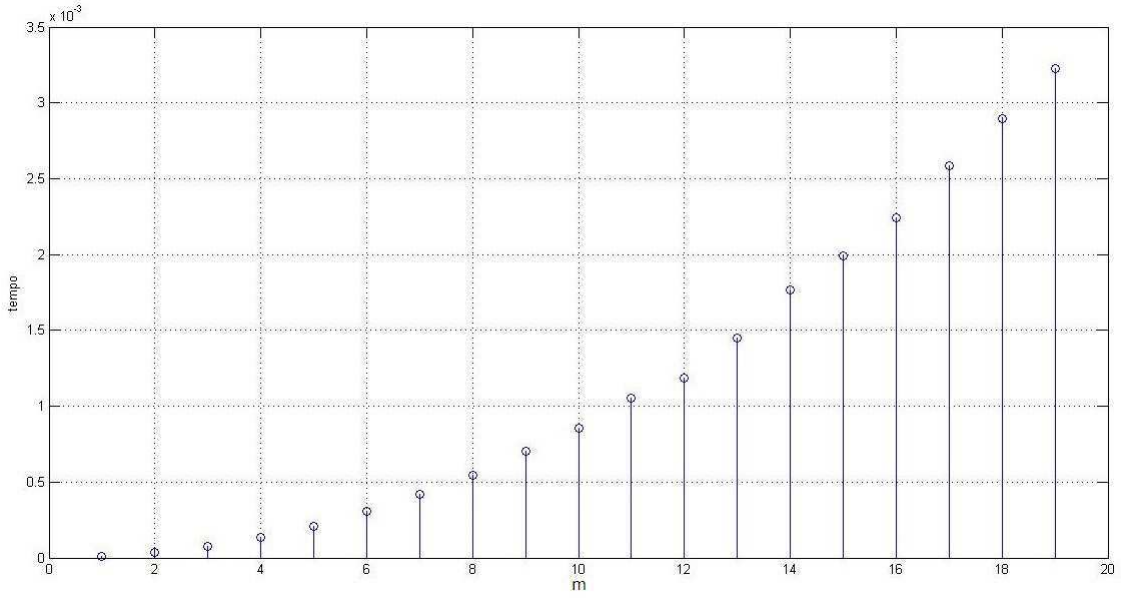

Figura 9 – Tempo em função do tamanho do esboço m.

3.2 Estudo de caso 2 - Aplicação do esboço na análise de dados na área da saúde.

O atendimento público de saúde - SUS realiza exames gratuitos para o diagnóstico de possíveis doenças. Um dos inúmeros exames laboratoriais que o SUS oferece é o hemograma, mais conhecido por exame de sangue, onde por dia a quantidade de exames realizados é muito alta para um único laboratório em uma determinada cidade. Processar os dados para determinar uma média de exames pagos pelo atendimento público em um

conjunto de cidades, estados ou até mesmo no país durante um intervalo de interesse é um problema desafiador devido a quantidade de dados. Neste contexto, o trabalho realizado neste estágio supervisionado consistiu em utilizar um esboço dos dados para analisar determinadas tendências e verificar se os resultados obtidos na matriz original e no *sketch* são aceitáveis.

Foi definida uma matriz D de ordem $m \times n$, onde m representa os meses de um ano e n representa as cidades de um estado brasileiro, que corresponde aos dados referentes a quantidade de exames realizados. Adotou-se $m = 12$ e $n = 223$. São analisados 223 municípios, que representa o estado da Paraíba. Foi gerada uma matriz A aleatória com dimensões $n \times q$, com $q < 223$ e distribuição normal $N(0, 1)$. O intuito é que ao realizar o produto $S_{m \times q} = D_{m \times n} A_{n \times q}$ as informações contidas em D sejam repassadas para S e que a análise que seria feita na matriz original agora possa ser feita em uma matriz de menor dimensão. As dimensões da matriz A são $p = 12$ e $q = 112$.

Ao projetarmos a informação sobre A obtemos um esboço com metade da dimensão da matriz original. Escolhemos o mês de março para calcular uma média no estado, devido ser o mês que mais há exames de sangues realizados (dados obtidos do HRNI - Icó/Ce). A média nos dados originais foi de 12.179 exames. No esboço obtivemos uma média de 11.845 exames realizados. O erro relativo absoluto é de 2,81%. Este erro ϵ é tolerável.

4 Conclusão

Este estágio atingiu a maior parte de seus objetivos, tanto no estudo teórico quanto na simulação do algoritmo pra analisar os esboços de matrizes. A quantidade de aplicações na qual é possível utilizar projeção aleatória para se fazer uma redução de dados é muito grande, mostrando então o destaque que este tema tem ganhado atualmente. O esboço de dados se mostra eficiente também devido à alta dimensionalidade restringir a escolha dos métodos de processamento dos dados a serem utilizados, fenômeno este ocasionado pela "maldição da dimensionalidade".

Com este projeto foi possível aplicar e aprofundar os conhecimentos adquiridos durante a graduação em um problema prático e usual ocorrido no dia-a-dia. Os resultados foram satisfatórios, com um erro ϵ tolerável, associado as simulações do algoritmo conforme previsto pelo lema de Johnson-Lindenstrauss.

5 Referências Bibliográficas

[1] E. Drinea, P. Drineas, and P. Huggins, "**A randomized singular value decomposition algorithm for image processing**", Panhellenic Conference on Informatics (PCI), IEEE Jornal Comm., vol. 20, nº 20, Set. 2001.

[2] X. Zhao "**High Perfomance Algorithms for Multiple Streaming Time Series**". New York University, New York, Jan. 2006.

[3] A. Dieb and E. C. Gurjão, "**Processing of Smart Meters Data Based on Random Projections**". Proceedings of IEEE PES Conference on Innovative Smart Grid Technologies, 2013, São Paulo.

[4] P. Indyk, N. Koudas and S. Muthukrishnan, "**Identifying Representative Trends in Massive Time Series Data Sets Using Sketches**". VLDB Conference, Cairo, Egito, 2009.

[5] P. Politopoulos, E. Markatos and S. Ioannidis, "**Evaluation of Compression of Remote Network Monitoring Data Streams**", IEEE Networks Operations and Management Symposium, April 2006.

[6] A. K. Menon, "**Random Projections ands Applications to Dimensionality Reduction**", The University of Sidney - Australia, March 2007.

[7] A. Gilbert and P. Indyk, "**Space Recovery Using Sparse Matrices**", Proceedings of the IEEE, 98(6):937, june 2010.

[9] A. Hyvarinen, J. Karhunen, and E. Oja. "**Independent Component Analysis**". Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, 2001.

[10] I. K. Fodor. "**A survey of dimension reduction techniques**". Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, June 2002.

[11] H, B. Borges. "**Redução de dimensionalidade em bases de dados de expressão gênica**", Dissertação de mestrado em Ciências da Computação. Pontifícia Universidade Católica do Paraná, Curitiba 2006.

[12] D. Achlioptas. "**Database-friendly random projections**". In Proc. ACM Symp. on the Principles of Database Systems, 2001, p. 274-281.

[13] A. S. A. Pessoa, J. D. S. da Silva e H. C. Júnior. "**Redução de dados meteorológicos aplicados a previsão climática por redes neurais**". Brasília, 2009.

[14] I. Pereira, T. Alves, R. Pinheiro e E. Assis. "**Metodologia de tratamento de dados climáticos para inserção em softwares de simulação energética de edifícios**". I

Conferência Latino-Americana de Construçao Sustentável. 18-21 julho 2004, São Paulo.

[15] E. Bingham and H. Mannila. "**Random projection in dimensionality reduction: Applications to image and text data**". Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 245-250). ACM Press.

[16] X. Z. Fern and C. E. Brodley. "**Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach**". Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[17] S. R. de M. Oliveira "**Métodos Usados para Redução e Sintetização de Dados**". Disponível em: <http://www.ime.unicamp.br/ wanderson/Aulas/MT803.pdf>. Acessado em 10 de julho de 2014.