

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO



Marcos Neves do Vale

**Agrupamentos de Dados: Avaliação de Métodos e
Desenvolvimento de Aplicativo para Análise de Grupos**

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para
obtenção do título de Mestre pelo Programa de Pós-
Graduação em Engenharia Elétrica da PUC-Rio.

Orientadores: Marley M. B. R. Vellasco
Ricardo Tanscheit

Rio de Janeiro, agosto de 2005



Marcos Neves do Vale

**Agrupamentos de Dados: Avaliação de Métodos e
Desenvolvimento de Aplicativo para Análise de Grupos**

Dissertação apresentada como requisito parcial para
obtenção do título de Mestre pelo Programa de Pós-
Graduação em Engenharia Elétrica da PUC-Rio.
Aprovada pela Comissão Examinadora abaixo assinada.

Marley M. B. R. Vellasco
Orientadora
PUC-RIO

Ricardo Tanscheit
Orientador
PUC-RIO

Juan Guillermo Lazo Lazo
PUC-RIO

Flávio Joaquim de Souza
UERJ

José Eugenio Leal
Coordenador Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 29 de agosto de 2005

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Marcos Neves do Vale

Graduou-se em Engenharia Elétrica com Ênfase em Sistemas e Computação na UERJ em 2001 e concluiu o curso de pós-graduação em Análise, Projeto e Gerência de Sistemas na PUC-RJ em 2003.

Ficha Catalográfica

Vale, Marcos Neves do

Agrupamentos de dados : avaliação de métodos e desenvolvimento de aplicativo para análise de grupos / Marcos Neves do Vale ; orientadores: Marley M. B. R. Vellasco, Ricardo Tanscheit. – Rio de Janeiro : PUC, Departamento de Engenharia Elétrica, 2005.

120 f. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica

Inclui referências bibliográficas.

1. Engenharia elétrica – Teses. 2. Agrupamento. 3. Análise de agrupamentos. 4. Aplicativo. 5. Lógica fuzzy. 6. Extração de informação. I. Vellasco, Marley M. B. R. II. Tanscheit, Ricardo. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Aos meus pais, principalmente a minha esposa e a minha filha.

Agradecimentos

Aos meus orientadores professora Marley Vellasco e professor Ricardo Tanscheit pela confiança, estímulo e dedicação para a realização deste trabalho.

Aos meus amigos Douglas e Paulo Motta por todo apoio, paciência e pelas importantes contribuições.

Ao Rodrigo Simões por ter me ajudado, mesmo de longe.

A todos os amigos e familiares que de uma forma ou de outra me estimularam ou me ajudaram.

Resumo

Vale, Marcos Neves do. **Agrupamentos de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos.** Rio de Janeiro, 2005. 120p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A enorme massa de dados que é gerada pelas diversas empresas diariamente pode conter informações importantes que não são fáceis de serem extraídas. Com isso advém a necessidade de analisá-los automaticamente, de forma adequada, extraindo informação útil que pode agregar algum tipo de conhecimento. Uma das formas de se analisar os dados automaticamente é através da análise de agrupamentos. Ela procura encontrar grupos de dados semelhantes entre si. As técnicas de análise de agrupamentos revelam como os dados estão estruturados e resultam em um melhor entendimento sobre o negócio. Existe ainda hoje uma escassez de ferramentas para esse fim. Em um problema real de agrupamento de dados convém analisar os dados através da utilização de diferentes métodos, a fim de buscar aquele que melhor se adapte ao problema. Porém, as ferramentas existentes hoje em dia não são integradas, onde cada ferramenta possui um subconjunto dos métodos existentes de agrupamento. Dessa forma o usuário fica limitado à utilização de uma ferramenta específica ou é obrigado a conhecer diversas ferramentas diferentes, de forma a melhor analisar os dados de sua empresa. Esta dissertação apresenta uma revisão detalhada de todo o processo de análise de agrupamentos e o desenvolvimento de um aplicativo que visa não apenas a atender as deficiências presentes na maioria das ferramentas com esse fim, mas também a auxiliar, de forma mais completa, todo o processo de análise dos grupos. O aplicativo desenvolvido é de fácil utilização e permite que a ele sejam incorporados outros métodos eventualmente desenvolvidos pelo usuário. O aplicativo foi avaliado em três estudos de casos, os quais visam demonstrar a facilidade de uso do aplicativo, assim como avaliar as vantagens do uso de métodos de natureza fuzzy em uma base de dados real.

Palavras-chave

Agrupamento, análise de agrupamentos, aplicativo, lógica fuzzy, extração de informação.

Abstract

Vale, Marcos Neves do. **Data Clustering: Analysis of Methods and Development of Application for Cluster Analysis**. Rio de Janeiro, 2005. 120p. MSc. Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The enormous data mass that is daily generated by several companies can contain critical information that might not be easily retrieved, considering that the amount of data is generally huge and/or the target information might be spread through different data bases. Taking that into consideration, it might be necessary to properly analyze the data in an automatic way, so useful and valuable information can be extracted. One way of automatically analyzing data is through cluster analysis. This type of analysis searches for related similar data. These clusters settle a data structure model and with proper analysis can reveal important information. The techniques used in cluster analysis disclose how data is structured and allow a better knowledge of the business. Still today there is a lack of tools for this purpose. On a real situation with a data cluster problem it is wise to analyze the data through different methods, so we can find the one that better fits the problem. However, today the existing tools are not integrated, and each tool has a subgroup of existing cluster methods. This way the user stays limited to use only one specific tool or is forced to be aware of a number of different tools, so he would be able to better analyze the company data. This study presents a detailed review of the whole group analysis process and develops an application that not only suggests how to cover the currently lack of tools for this purpose, but also to help the complete cluster analysis process in a more extended way. The application developed is user friendly and allows other methods developed by users to be incorporated. The application has been evaluated into three case studies with the purpose of demonstrating its user friendly, as well as evaluating the advantages of using fuzzy methods on a true data base.

Keywords

clustering, clustering software, fuzzy clustering, information retrieval

Sumário

1	Introdução	15
1.1.	Motivação	15
1.2.	Objetivos	17
1.3.	Organização da Dissertação	18
2	Processo de Agrupamentos	20
2.1.	Seleção e Tratamento de dados	21
2.1.1.	Tratamento de Atributos	21
2.1.2.	Normalização dos Atributos	22
2.2.	Agrupamento de dados	24
2.2.1.	Medidas de Proximidade	24
2.2.1.1.	Dissimilaridade	25
2.2.1.2.	Similaridade	26
2.2.2.	Métricas Comuns em Medidas de Proximidade	27
2.3.	Análise dos resultados	29
2.3.1.	Gráfico da Silhueta	29
3	Métodos de Agrupamento de Dados	32
3.1.	Métodos Hierárquicos	32
3.1.1.	Métodos Aglomerativos	33
3.1.1.1.	Dendograma	35
3.1.1.2.	Coeficiente Aglomerativo (CA)	36
3.1.1.3.	Banner de Dissimilaridade	36
3.1.2.	Métodos Divisivos	38
3.1.2.1.	Dendograma	38
3.1.2.2.	Coeficiente Divisivo (CD)	39
3.1.2.3.	Banner de Dissimilaridade	40
3.1.3.	Métodos de Distância entre Grupos	41
3.1.4.	Métodos Hierárquicos Conhecidos	45
3.1.4.1.	Agglomerative Nesting (AGNES)	45

3.1.4.2. Divisive Analysis (DIANA)	46
3.1.4.3. Monothetic Analysis (MONA)	47
3.2. Métodos Particionais	48
3.2.1. Métodos Não-Exclusivos	49
3.2.1.1. Coeficiente de DUNN	50
3.2.2. Métodos Particionais Conhecidos	51
3.2.2.1. K-Means	51
3.2.2.2. Fuzzy C-Means	52
3.2.2.3. Fuzzy Analysis (FANNY)	55
3.2.2.4. Gustafson-Kessel	57
3.2.2.5. Gath-Geva	58
3.2.2.6. Mistura de Densidades	60
3.2.2.7. Partitioning Around Medoids (PAM)	62
3.2.2.8. Clustering Large Applications (CLARA)	64
4 Aplicativo para Análise de Agrupamentos	65
4.1. Modelagem do Aplicativo	68
4.1.1. Seleção e Tratamento de Dados	69
4.1.2. Parametrização do Aplicativo	70
4.1.3. Agrupamento de Dados e Geração de Resultados	72
4.1.4. Apresentação dos Resultados	74
4.1.5. Modelagem Detalhada do Aplicativo	79
4.2. Disponibilizando outros Métodos	80
4.2.1. Métodos Externos	80
4.2.1.1. Métodos no Matlab®	81
4.2.1.2. Métodos no R®	82
4.2.2. Métodos Internos	83
5 Estudos de Caso	84
5.1. Base de Dados Exemplo do Matlab®	84
5.2. Base de Dados Íris	94
5.3. Base de Dados de Hipertensão na Iha do Governador	101
6 Conclusões e Trabalhos Futuros	110

6.1. Conclusões	110
6.2. Trabalhos Futuros	111
7 Referências Bibliográficas	113
ANEXO - Guia de Instalação e Uso do Aplicativo	116
Guia de Instalação	116
Guia de Utilização	116

Lista de figuras

Figura 1:Gráfico ilustrativo de dados agrupados em quatro grupos.	20
Figura 2: Superfícies observadas pelas distâncias Euclidiana, Mahalanobis e Manhattan.	27
Figura 3: Gráfico da Silhueta.	31
Figura 4: Método Hierárquico Aglomerativo – Dendograma.	35
Figura 5: Banner de Dissimilaridade (Métodos Hierárquicos Aglomerativos).	36
Figura 6: Métodos Hierárquicos Aglomerativos: dendograma e banner.	37
Figura 7: Método Hierárquico Divisivo - Dendograma	39
Figura 8: Banner de Dissimilaridade (Métodos Hierárquicos Divisivos).	40
Figura 9: Métodos Hierárquicos Divisivos: dendograma e banner.	41
Figura 10: Exemplo de agrupamento de dados binário usando MONA.	47
Figura 11: Exemplo de um conjunto de 22 dados.	49
Figura 12: Gráfico Dissimilaridade entre os grupos X Valor de m.	55
Figura 13: Diagrama de Seqüência - Interface com Matlab® e R®.	65
Figura 14: Diagrama de Blocos: Visão geral do processo e aplicativo para análise de grupos.	68
Figura 15: Diagrama de Blocos: Visão detalhada do processo de Seleção e Tratamento de Dados.	69
Figura 16: Aplicativo de Análise de Grupos: Seleção e Tratamento de Dados.	70
Figura 17: Diagrama de Blocos: Visão detalhada do processo de Parametrização do Aplicativo.	70
Figura 18: Aplicativo de Análise de Grupos: Geração de Agrupamentos	71
Figura 19: Aplicativo de Análise de Grupos: Geração de Resultados.	72
Figura 20: Diagrama de Blocos: Visão detalhada do processo de Agrupamento de Dados e Geração de Resultados	72
Figura 21: Diagrama de Blocos: Visão detalhada do processo de Apresentação dos Resultados.	74
Figura 22: Aplicativo de Análise de Grupos: Gráfico da Silhueta.	75
Figura 23: Aplicativo de Análise de Grupos: Tabela de Pertinências.	76

Figura 24: Aplicativo de Análise de Grupos: Tabela de Médias.	76
Figura 25: Aplicativo de Análise de Grupos:Gráfico Comparativo.	77
Figura 26: Aplicativo de Análise de Grupos: Dendograma.	78
Figura 27: Diagrama de blocos detalhado do aplicativo desenvolvido	79
Figura 28: Base de Dados fcndata do Matlab®.	84
Figura 29: fcndata – carregando base de dados	85
Figura 30: fcndata – Geração de Agrupamentos (AGNES – Ligação Simples).	86
Figura 31: fcndata – Dendograma (AGNES – Ligação Simples).	87
Figura 32: fcndata – Gráfico da Silhueta (LINKAGE – Média das Ligações).	87
Figura 33: fcndata – Dendograma (LINKAGE – Média das Ligações).	88
Figura 34: fcndata – Gráfico da Silhueta (DIANA).	89
Figura 35: fcndata – Dendograma (DIANA).	89
Figura 36: fcndata – Gráfico da Silhueta (PAM).	90
Figura 37: fcndata – Gráfico da Silhueta (K-Means).	91
Figura 38: fcndata – Tabela de Pertinências / Filtro (FANNY).	91
Figura 39: fcndata – Gráfico da Silhueta (FANNY).	92
Figura 40: fcndata – Tabela de Pertinências / Filtro (Fuzzy C-Means).	93
Figura 41: fcndata – Gráfico da Silhueta (Fuzzy C-Means).	93
Figura 42: íris – Dendograma (AGNES – Ward).	95
Figura 43: íris – Dendograma (DIANA).	96
Figura 44: íris – Gráfico da Silhueta (AGNES - Ward).	97
Figura 45: íris – Gráfico da Silhueta (DIANA).	97
Figura 46: íris – Gráfico da Silhueta (K-Means).	98
Figura 47: íris – Gráfico da Silhueta (PAM).	99
Figura 48: íris – Gráfico da Silhueta (Fuzzy C-Means).	99
Figura 49: íris – Gráfico da Silhueta (FANNY).	100
Figura 50: Base de Dados da Ilha do Governador – Seleção de Variáveis.	102
Figura 51: Base de Dados da Ilha do Governador – Gráfico da Silhueta (LINKAGE – Média das Ligações).	102
Figura 52: Base de Dados da Ilha do Governador – Dendograma (LINKAGE – Média das Ligações).	103
Figura 53: Base de Dados da Ilha do Governador – Gráfico da Silhueta (LINKAGE – Ward).	103

Figura 54: Base de Dados da Ilha do Governador – Dendograma (LINKAGE – Ward).	104
Figura 55: Base de Dados da Ilha do Governador – Gráfico da Silhueta (DIANA).	104
Figura 56: Base de Dados da Ilha do Governador – Dendograma (DIANA).	105
Figura 57: Base de Dados da Ilha do Governador – Gráfico da Silhueta (K-Means).	106
Figura 58: Base de Dados da Ilha do Governador – Gráfico da Silhueta (Fuzzy C-Means).	106
Figura 59: Distribuição de indivíduos nos agrupamentos	108
Figura 60: Modelo de arquivo de base de dados	116
Figura 61: Métodos de Agrupamento de Dados	118
Figura 62: Filtro	119

Lista de tabelas

Tabela 1: Valores da Silhueta.	30
Tabela 2: Tabela ilustrativa da Matriz de Similaridades entre Grupos.	33
Tabela 3: Arquivo: dados_mean.txt – Formato dos dados.	80
Tabela 4: Arquivo: dados_idx.txt – Formato dos dados.	80
Tabela 5: Arquivo: dados_comparativo.txt – Formato dos dados.	81
Tabela 6: fcndata – Tabela de Médias (AGNES – Ligação Simples).	86
Tabela 7: fcndata – Tabela de Médias (LINKAGE – Média das Ligações).	87
Tabela 8: fcndata – Tabela de Médias (DIANA).	88
Tabela 9: fcndata – Tabela de Médias (PAM).	90
Tabela 10: fcndata – Tabela de Médias (K-Means).	90
Tabela 11: fcndata – Tabela de Médias (FANNY).	92
Tabela 12: fcndata – Tabela de Médias (FANNY).	93
Tabela 13: íris - Avaliação dos métodos de agrupamento de dados.	100
Tabela 14: K-Means: médias para cada agrupamento	107
Tabela 15: FCM com m=2: médias para cada agrupamento	107
Tabela 16: FCM com m=1,5: médias para cada agrupamento	107
Tabela 17: Pertinências de indivíduos com graus de pertinência altos e semelhantes	108
Tabela 18: Características dos indivíduos com graus de pertinência altos e semelhantes	109
Tabela 19: Média dos indivíduos com graus de pertinência altos e semelhantes.	109