Um Estudo da Aplicação de *Clustering* de Séries Temporais em Dados Médicos

Newton Spolaôr[1], Huei Diana Lee[1,2], Carlos Andres Ferrero[2], Cláudio Saddy Rodrigues Coy[2], João José Fagundes[2], Feng Chung Wu[1,2,3]

[1]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85869-970 – Foz do Iguaçu, PR, Brasil

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

[3]Faculdade de Ciências Médicas – Universidade Estadual de Campinas
Serviço de Coloproctologia
Caixa Postal 6111, 13083-970 – Campinas, SP, Brasil

newtonspolaor@hotmail.com, hueidianalee@gmail.com

Abstract. *The current technological progress allows the development of technologies to store a great amount of information, from which is possible to extract useful knowledge. Accordingly, processes that support Time Series Analysis, as clustering, may help the medical area on the study of incontinence of feces and diseases from Ano-rectal Manometry domain. In this work, it's presented a study of Time Series clustering appliance in Ano-rectal Manometry exams. The clusters built during the process may be analyzed, along with specialists, to identify and to consolidate concepts which may give support to the decision making process regarding this domain.*

Resumo. *O progresso tecnológico atual permite o desenvolvimento de tecnologias para o armazenamento de uma quantidade ampla de informações, a partir da qual é possível extrair conhecimentos úteis. Nesse sentido, processos que auxiliam na Análise de Séries Temporais, como o clustering, podem auxiliar a área médica no estudo da Incontinência Fecal e de doenças do domínio de Manometria Ano-retal. Neste trabalho é apresentado um estudo da aplicação de clustering de Séries Temporais em exames de Manometria Ano-retal. Os clusters construídos durante o processo podem ser analisados, com a contribuição de especialistas, para identificar e consolidar conceitos que posteriormente auxiliem em processos de tomada de decisão desse domínio.*

1. Introdução

O avanço tecnológico atual permite a aplicação de sistemas de gerenciamento de dados para o armazenamento de conjuntos de dados cada vez maiores, a partir dos quais é possível extrair conhecimentos úteis. A aquisição manual desse conhecimento é uma tarefa complexa devido ao alto custo de tempo envolvido, o que motiva a utilização de

processos computacionais para realizar uma análise mais completa dos dados [Lee 2005]. Na área de Análise de Séries Temporais — AST — são propostos métodos que possibilitam a interpretação de informações provenientes da seqüencialidade de um tipo de dado denominado Série Temporal — ST.

O *clustering* é uma das tarefas que podem contribuir com a AST. O objetivo dessa tarefa é auxiliar na identificação de conceitos a partir de *clusters* (agrupamentos), os quais geralmente compreendem elementos, sem classe ou rótulo associado, próximos entre si e distantes em relação aos elementos de outros *clusters* [Everitt et al. 2001]. Esse processo é realizado, geralmente, a partir das etapas de Representação dos Dados, Seleção da Medida de Similaridade e Aplicação do Agrupamento [Jain and Dubes 1988]. A primeira etapa envolve tarefas como a redução de problemas provenientes de ruídos e dados faltantes. A representação dos exemplos do conjunto de dados sob processamento deve ser feita de forma cautelosa, pois pode influenciar na compreensibilidade dos *clusters* [Jain et al. 1999]. Na segunda etapa geralmente é construída uma Matriz de Similaridade — MS — para descrever a semelhança entre os *clusters*, conforme alguma medida de similaridade. Essa medida deve ser selecionada de acordo com as propriedades dos dados do domínio considerado. A Aplicação do Agrupamento envolve a junção dos elementos representados na MS em *clusters*, com o auxílio de um determinado algoritmo de *clustering*. Após essa etapa, a qualidade dos *clusters* construídos pode ser avaliada com o auxílio de critérios e medidas de avaliação e essa análise pode indicar a necessidade de aplicação de outras técnicas nas etapas do processo de *clustering*, de modo a possibilitar a construção de agrupamentos mais representativos.

O *clustering* pode ser aplicado em áreas como a medicina, em exames médicos de diversas especialidades, com o intuito de auxiliar em processos de tomada de decisão. O exame de Manometria Ano-retal — MA — é imprescindível no diagnóstico da Incontinência Fecal — IF —, a qual se caracteriza pela incapacidade de controle da passagem de fezes ou de gases em lugares socialmente aceitáveis [Saad 2002]. Alguns fatores considerados nesse diagnóstico são a pressão média de repouso e a Pressão de Contração Voluntária — PCV —, as quais podem ser coletadas a medida que o tempo varia. Esses dados podem ser modelados computacionalmente como ST, o que permite a aplicação de *clustering* para auxiliar na aquisição de conhecimento.

O objetivo deste trabalho, em andamento, é apresentar um estudo da aplicação do processo de *clustering* de ST em dados provenientes de exames de MA. O trabalho constitui parte do Projeto de Análise Inteligente de Dados, o qual é desenvolvido pelo Laboratório de Bioinformática — LABI — em parceira com o Centro de Estudos Avançados em Segurança de Barragens — CEASB —, o Laboratório de Inteligência Computacional — LABIC — da Universidade de São Paulo — USP/São Carlos — e o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas — UNICAMP.

Este trabalho está organizado da seguinte maneira: na Seção 2 são apresentadas abordagens de Representação dos Dados, Seleção da Medida de Similaridade e Aplicação do Agrupamento que podem ser utilizadas no conjunto de dados considerado. Na Seção 3, o delineamento experimental realizado é descrito e discutido. A conclusão e os trabalhos futuros são apresentados na Seção 4.

2. Materiais e Métodos

A MA é um dos exames fisiológicos mais aplicados na medicina para auxiliar no diagnóstico de IF e doenças ano-retais como a doença de Hirschsprung, as quais são relacionadas às funções de continência e evacuação do paciente [Morais et al. 2005, Freys et al. 1998]. Alguns estudos relacionados à IF estimam uma maior incidência dessa condição em idosos e em mulheres, as quais apresentam uma probabilidade maior de desenvolver a enfermidade se realizam parto normal. No entanto, fatores epidemiológicos determinantes associados à doença ainda não foram encontrados. A intensidade da IF apresentada por um paciente pode ser analisada de acordo com diversas classificações, dentre as quais uma das mais utilizadas divide-se em três graus [Pinho 2000]. Um paciente com IF em Grau I apresenta incontinência apenas para gases. A IF em Grau II acrescenta a incapacidade de conter fezes líquidas às deficiências observadas no Grau I, enquanto que o Grau III associa todos esses distúrbios à incontinência para fezes sólidas [Saad 2002].

Neste trabalho foram considerados 20 exames de MA, realizados no período de Maio/1995 a Novembro/1996 pelo Serviço de Coloproctologia da UNICAMP, dos quais oito referem-se a pacientes com IF em Grau III e doze correspondem a pacientes em condições normais. Esses exames são compostos por informações como a história clínica dos pacientes e os valores de PCV em milímetros de mercúrio, os quais foram coletados em uma freqüência de oito medições por segundo com o auxílio de um cateter composto por oito sensores. Cada um desses sensores foi modelado como uma ST, totalizando oito ST por exame de MA. A soma dos valores de PCV dessas oito ST foi utilizada para definir uma única ST representativa de cada exame. A análise das 20 ST resultantes, com o auxílio de especialistas, possibilitou identificar para cada ST, três seções de PCV, com duração individual de aproximadamente 40 segundos. A delimitação das seções foi realizada de forma automática em todas as ST, a partir da seleção dos três intervalos de 40 segundos com as maiores médias de PCV, conforme é exemplificado na Figura 1. As três seções de cada exame foram concatenadas para compor o conjunto de 20 ST utilizado no trabalho.

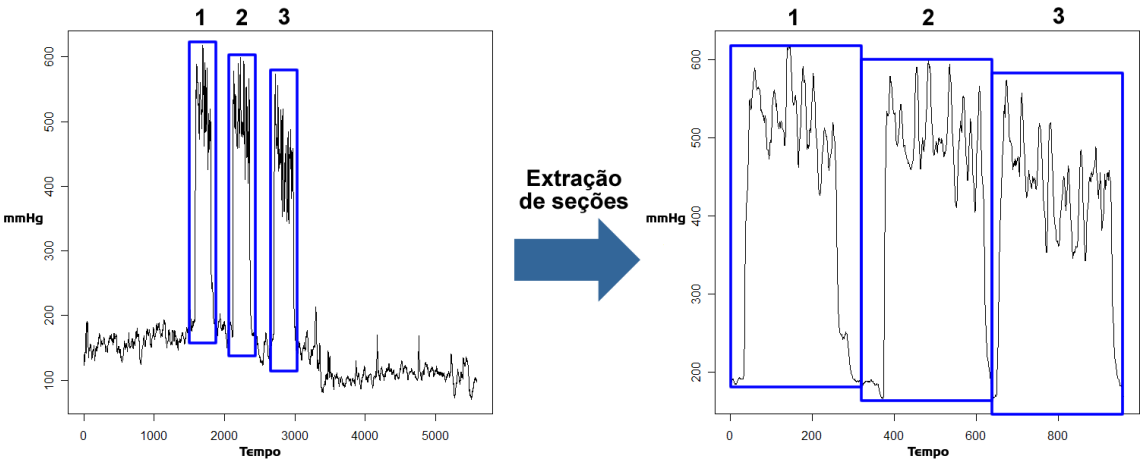

Figura 1: Extração automática de seções nas ST de MA

O processo de *clustering* pode ser aplicado nesse conjunto de dados para auxiliar na identificação de conceitos intrínsecos aos dados. Esse processo é constituído de

três etapas: (1) Representação dos Dados, (2) Seleção da Medida de Similaridade e (3) Aplicação do Agrupamento, conforme é ilustrado na Figura 2.

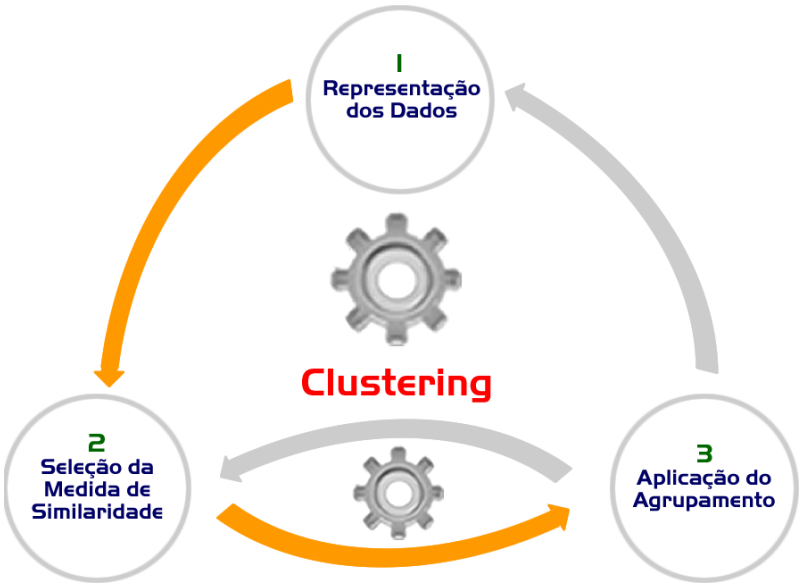

Figura 2: Processo de *clustering*

Cada etapa pode ser realizada com o auxílio de diversas técnicas, as quais devem ser avaliadas e selecionadas de acordo com o domínio considerado, visto que não existe uma abordagem adequada para todos os casos. Opcionalmente, critérios e medidas de avaliação de *clustering* podem ser aplicados após a Etapa 3 para averiguar a qualidade dos *clusters* construídos. A partir do resultado dessa análise é possível observar a necessidade de utilização de outras técnicas em cada etapa do processo. Algumas técnicas propostas para cada etapa do *clustering* de ST e para a avaliação dos *clusters* são descritas a seguir.

2.1. Etapa 1: Representação dos Dados

A seleção do modo de descrição dos dados envolve, dentre outras tarefas, a eliminação ou a redução de problemas de escala e de dados faltantes [Pyle 1999]. Uma das abordagens utilizadas nessa etapa corresponde à representação de ST a partir dos dados originais, na qual não são realizadas transformações nos dados. Esse modo de representação pode apresentar problemas, como a maior sensibilidade a ruídos e à alta dimensionalidade, os quais comprometeriam a qualidade das demais etapas do processo de *clustering* [Wang et al. 2006].

Outras abordagens propostas para a descrição de ST correspondem às representações por discretização e por características. A discretização permite representar ST a partir de um alfabeto simbólico, o que as torna mais adequadas à manipulação e ao processamento posterior e possibilita, dependendo da técnica utilizada, a redução da dimensionalidade do conjunto de dados [Antunes and Oliveira 2001]. A utilização de características na descrição de ST pode proporcionar melhorias ao processo de *clustering* provenientes da redução de problemas relacionados à alta dimensionalidade e da eliminação da diferença de comprimento entre as séries, o que possibilita a aplicação de determinadas medidas de similaridade [Wang et al. 2006].

2.2. Etapa 2: Seleção da Medida de Similaridade

As medidas de similaridade são utilizadas pelo processo de *clustering* para construir uma MS que representa a semelhança entre os *clusters* representativos de um determinado conjunto de dados. Algumas métricas que podem ser aplicadas nesse sentido são as normas L_p [Jain et al. 1999], como a distância Euclidiana, e o *Dynamic Time Warping* — DTW [Berndt and Clifford 1994].

2.2.1. Normas L_p

As normas L_p correspondem às medidas de distância Euclidiana, Manhattan e Minkowski. Dentre essas métricas a Euclidiana, ou norma L_2, é a mais utilizada em *clustering*, visto que permite a determinação da distância entre os elementos no espaço multidimensional, principalmente na determinação da distância entre objetos em duas e três dimensões [Jain et al. 1999]. No entanto, a distância Euclidiana apresenta algumas restrições, como a sensibilidade a ruídos e a *outliers*[1] [Mörchen 2006]. A distância Manhattan, também conhecida como norma L_1 ou *city-block*, corresponde à distância entre dois pontos a partir da soma dos segmentos de reta que os unem em um espaço bidimensional. A generalização das medidas Euclidiana e Manhattan é definida a partir da medida de distância Minkowski, ou L_p. Na Equação 1 é apresentada a distância Minkowski entre duas ST X e Y de tamanho n, as quais são definidas, respectivamente, a partir das notações $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_n$.

$$d(X,Y) = \left(\sum_{a=1}^{n} |X_a - Y_a|^p\right)^{1/p} (p \geq 1) \quad (1)$$

A distância Minkowski entre as ST X e Y, ilustrada na Equação 1, considera o parâmetro p, o qual é o elemento que permite a generalização comentada anteriormente. Se p = 1, a medida representada é a Manhattan. A medida Euclidiana é descrita pela Equação 1 com p = 2. Se o p considerado for infinito, a distância Minkowski é determinada pela maior distância entre os pontos que compõem X e Y [Berkhin 2002]. Um aspecto que pode limitar a aplicação das normas L_p corresponde à inviabilidade do cálculo de similaridade entre ST de tamanhos diferentes.

2.2.2. *Dynamic Time Warping*

A distância DTW foi proposta para, entre outros motivos, solucionar algumas limitações das métricas L_p, como a não aplicabilidade em ST de tamanhos diferentes e a sensibilidade a pequenas distorções no eixo temporal [Berndt and Clifford 1994]. A técnica DTW permite identificar o *warping path*, o qual corresponde à combinação de pontos que minimiza a dissimilaridade e alinha virtualmente as ST consideradas no eixo do tempo [Mörchen 2006]. Para exemplificar essa técnica, duas ST denominadas X e Y de tamanhos n e m, respectivamente, são consideradas. O cálculo do DTW entre ambas é realizado a partir de uma matriz de distâncias M de ordem n x m que representa

[1]Observação que apresenta um comportamento anormal ou diferente em relação ao restante da população

a dissimilaridade entre cada ponto x_i e y_j. Desse modo, o *warping path* W é determinado como a seqüência de k distâncias em M que representa a menor distância entre cada ponto de X e Y. A distância utilizada internamente pelo DTW geralmente corresponde a uma norma L_p, como a distância Euclidiana ou a distância Manhattan [Mörchen 2006]. O custo computacional quadrático é uma desvantagem do DTW que pode ser reduzida a partir de determinadas otimizações [Ratanamahatana and Keogh 2004].

2.3. Etapa 3: Aplicação do Agrupamento

O agrupamento de ST pode ser realizado de várias maneiras, a partir de abordagens como a hierárquica, a qual se caracteriza por organizar os *clusters* gerados em uma árvore denominada dendograma. Essa representação corresponde a um gráfico que exibe os agrupamentos construídos conforme a variação do valor da medida de similaridade, o que possibilita a visualização dos *clusters* em diferentes níveis de glanularidade [Berkhin 2002]. Um ponto de corte pode ser definido no dendograma para auxiliar na visualização, em um determinado valor de similaridade, dos *clusters* construídos de acordo com uma proximidade menor que a considerada nesse ponto.

Uma das vantagens do *clustering* hierárquico corresponde a não necessidade da definição inicial da quantidade de *clusters* que devem ser gerados [Wang et al. 2006]. Entretanto, a aplicação desse método a grandes conjuntos de dados pode ser limitada devido às complexidades de espaço e tempo, respectivamente, de ordem quadrática e cúbica que esse algoritmo pode atingir no pior caso [Tan et al. 2005].

O *clustering* hierárquico pode ser realizado de modo divisivo ou aglomerativo. O modo aglomerativo é a abordagem hierárquica mais utilizada, visto que o modo divisivo pode apresentar um alto custo computacional. O agrupamento aglomerativo define, inicialmente, um *cluster* para cada exemplo e a MS que representa a semelhança entre todos os *clusters*. Posteriormente, a MS é analisada para identificar os dois *clusters* com a maior proximidade entre si, os quais são agrupados em um novo *cluster*. Esse *cluster* é adicionado à MS, de modo que a semelhança entre todos os *clusters* é recalculada de acordo com um determinado algoritmo. Esse processo é realizado, geralmente, até a construção de um *cluster* que contenha todos os exemplos. Alguns algoritmos propostos para a determinação da semelhança entre os *clusters*, a partir de pares de exemplos nos quais cada elemento pertence a um *cluster* distinto, são o *Single-link* [Sneath 1957], o *Complete-link* [King 1967] e o *Average-link* [Jain et al. 1999].

***Single-Link*.** Corresponde ao método de *clustering* hierárquico mais simples e utilizado. Caracteriza-se pela determinação da distância entre os *clusters* a partir do par de exemplos de menor dissimilaridade, o que possibilita a detecção de *clusters* concêntricos, alongados ou não-compactos. Esse comportamento simula uma árvore geradora mínima, a qual é descrita na teoria dos grafos. O *Single-link* pode ser afetado pelo *chaining effect*, o qual corresponde à existência de exemplos, entre dois *clusters* com pouca afinidade, que podem formar uma ponte que resulte no agrupamento de ambos em um mesmo *cluster* [Jain et al. 1999];

***Complete-Link*.** Esse método considera o par de exemplos de maior distância para definir a similaridade entre os *clusters*, no qual cada exemplo pertence a um *cluster* distinto. Os *clusters* resultantes tendem a ser compactos, o que justifica a utilização dessa abordagem em várias aplicações de *clustering* [Jain and Dubes 1988];

***Average-Link*.** Uma abordagem intermediária às demais apresentadas é o *Average-link*, o qual define a similaridade entre *clusters* a partir da média de todas as distâncias entre os pares de exemplos considerados. Esse método pode reduzir problemas relacionados à sensibilidade a *outliers* [Tan et al. 2005].

2.4. Avaliação de *Clustering*

A avaliação de *clustering* tem como objetivo analisar a qualidade dos *clusters* construídos após a Etapa 3. Os critérios de avaliação internos, externos e relativos auxiliam nessa análise, pois podem indicar objetivamente se os *clusters* construídos representam conceitos inerentes ao conjunto de dados [Jain and Dubes 1988]. As estratégias de avaliação de *clustering*, determinadas a partir do critério selecionado, são aplicadas com o auxílio de medidas de avaliação, as quais permitem testar estatisticamente a qualidade do agrupamento. O Coeficiente de Correlação *Cophenetic* — CCC — é um exemplo de medida que pode ser aplicada na avaliação de algoritmos de *clustering* hierárquico. Esse coeficiente indica a correlação linear existente entre a MS original e a *cophenetic matrix*, a qual representa os valores de similaridade que compuseram cada *cluster* no dendograma. Quanto mais próximo de 1 for o coeficiente, maior será a probabilidade da hierarquia representar adequadamente os conceitos presentes nos dados, de acordo com a medida de similaridade utilizada [Halkidi et al. 2001].

3. Estudo de Caso

O processo de *clustering* de ST apresentado neste trabalho foi aplicado sobre o conjunto de 20 ST de MA descrito anteriormente. O delineamento experimental foi constituído pela avaliação de uma técnica de representação dos dados, quatro medidas de similaridade e três abordagens de *clustering* hierárquico aglomerativo, o que resulta em 12 distintas configurações do processo de *clustering* apresentadas a seguir:

Etapa 1: Representação dos Dados. Representação de ST a partir dos dados originais;
Etapa 2: Seleção da Medida de Similaridade. Normas L_1, L_2, L_3 e DTW;
Etapa 3: Aplicação do Agrupamento. *Single-link*, *Complete-link* e *Average-link*.

Cada configuração foi implementada com o auxílio do ambiente computacional R© versão 2.6.2[2], o qual também permitiu a construção dos dendogramas que representam os *clusters* construídos em cada experimento. A avaliação da qualidade do *clustering*, realizada a partir do CCC, é exibida na Tabela 1, na qual são dispostas as abordagens de *clustering* nas colunas e as medidas de similaridade nas linhas.

Tabela 1: Resultados da aplicação do CCC nos experimentos

CCC	*Complete-link*	*Single-link*	*Average-link*
L_1	0,7460	0,8637	0,8701
L_2	0,7595	0,8828	0,8814
L_3	0,7673	0,8913	0,8883
DTW	0,5970	0,6033	0,7031

[2]www.r-project.org

A análise dos valores dos CCC, em relação às medidas de similaridade, indicou que a medida L_3 auxiliou na construção dos *clusters* de melhor representatividade do conjunto de dados, independentemente do algoritmo de *clustering* considerado. No sentido oposto, a métrica DTW contribuiu para atribuir a todas as abordagens de *clustering* os respectivos menores valores do CCC. Ao se analisar os algoritmos de *clustering* especificamente, as abordagens *Single-link* e *Average-link* apresentaram os maiores valores de CCC nas quatro medidas de similaridade avaliadas. O *Complete-link* foi superado em todas as medidas de similaridade pelos demais algoritmos, no que se refere à correlação *cophenetic*. A abordagem *Single-link* combinada à medida de similaridade L_3 foi a mais representativa do conjunto de dados, apresentando uma correlação *cophenetic* de 0,8913. O dendograma resultante desse experimento e as ST correspondentes aos pacientes com IF em Grau III (vermelhas) e em condições normais (azuis) são apresentados na Figura 3.

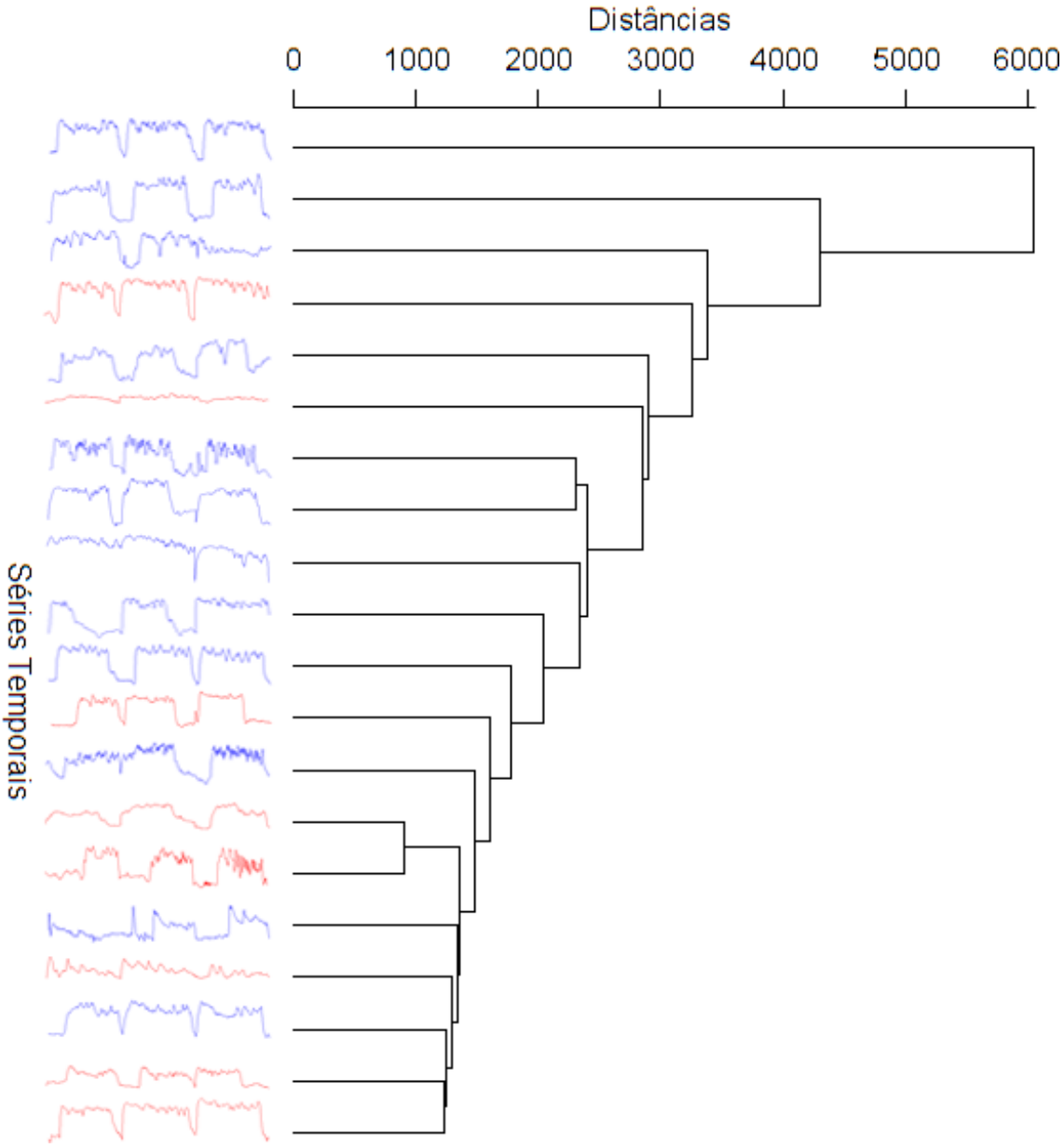

Figura 3: Dendograma correspondente ao *Single-link* associado à norma L_3

Os agrupamentos construídos nesses experimentos podem ser analisados com o auxílio dos dendogramas gerados no processo de *clustering*. Essas ferramentas descrevem graficamente a construção de cada *cluster*, em diferentes níveis de glanularidade, de acordo com a variação do valor da medida de similaridade. A associação de um ponto de corte a essas estruturas permite a determinação de um nível de glanularidade, o qual compreende *clusters* que podem representar conceitos úteis a algum propósito.

O processo de *clustering* é prioritariamente aplicado na análise exploratória de conjuntos de exemplos não rotulados e de conjuntos de exemplos cuja rotulação é muito custosa. Os agrupamentos gerados por esse processo devem ser cuidadosamente analisados, com o auxílio de especialistas, para eliminar os *clusters* que não revelam nenhum conceito ou que levam à visualização errônea das relações entre os exemplos. A definição de pontos de corte nos dendogramas para seleção dos *clusters* com uma determinada glanularidade pode contribuir nessa análise, de modo que seja possível identificar e consolidar conhecimentos que auxiliem em processos de tomada de decisão relacionados à MA.

4. Conclusão

Neste trabalho em andamento foi apresentado um estudo da aplicação do processo de *clustering* de ST em exames de MA. O delineamento experimental permitiu avaliar o desempenho de uma abordagem de representação de dados associada a três algoritmos de *clustering* e a quatro medidas de similaridade, totalizando 12 experimentos. O algoritmo *Single-link*, associado à medida de similaridade L_3, construiu o agrupamento de melhor qualidade, de acordo com o CCC.

Como trabalhos futuros se propõem a análise dos dendogramas com o auxílio de especialistas da área médica. Essa análise pode contribuir para a identificação e a consolidação de conhecimentos que auxiliem em processos de tomada de decisão relacionados ao diagnóstico de IF e de doenças do domínio de MA. Serão aplicadas conjuntamente outras medidas de avaliação objetiva de *clustering*, as quais poderão indicar aos especialistas os dendogramas mais representativos dos conceitos presentes nos dados.

Agradecimentos

Ao Programa de Desenvolvimento Tecnológico Avançado — PDTA/FPTI-BR — pelo auxílio por meio da linha de financiamento de bolsas.

Referências

Antunes, C. M. and Oliveira, A. L. (2001). Temporal data mining: an overview. In *Knowledge Discovery and Data Mining*, pages 1–15.

Berkhin, P. (2002). Survey of clustering data mining techniques. Technical report, Accrue Software, California, USA.

Berndt, D. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Workshop on Knowledge Discovery in Databases*, pages 229–248.

Everitt, B. S., Landau, S., and Leese, M. (2001). *Cluster analysis*. A Hodder Arnold, 4 edition.

Freys, S. M., Fuchs, K. H., Fein, M., Heimbucher, J., Sailer, M., and Thiede, A. (1998). Inter- and intraindividual reproducibility of anorectal manometry. *Langenbeck's Archives of Surgery*.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., New Jersey, USA.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, 69:86–101.

Lee, H. D. (2005). *Seleção de atributos importantes para a extração de conhecimento de bases de dados.* Tese de doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos - Brasil.

Morais, M. B., Sdepanian, V. L., Tahan, S., Goshima, S., Soares, A. C. F., Motta, M. E. F. A., and Neto, U. F. (2005). A manometria anorretal (método do balão) no diagnóstico diferencial da doença de hirschsprung. *Revista da Associação Médica Brasileira*, 51:313 – 317.

Mörchen, F. (2006). *Time series knowledger mining*. Tese de doutorado, Philipps-Universität Marburg, Marburg, Germany.

Pinho, M. (2000). *Incontinência fecal*. Revinter, Rio de Janeiro, Brasil.

Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann, California, USA.

Ratanamahatana, C. A. and Keogh, E. (2004). Making time-series classification more accurate using learned constraints. In *SIAM International Conference on Data Mining*, pages 11–22.

Saad, L. H. C. (2002). *Quantificação da função esfincteriana pela medida da capacidade de sustentação da pressão de contração voluntária do canal anal*. Tese de doutorado, Universidade Estadual de Campinas.

Sneath, P. H. A. (1957). The applications of computers to taxonomy. *General Microbiology*, pages 201–226.

Tan, P., Steinbach, M., and Kumar, V. (2005). *Introduction to data mining*. Addison Wesley, 1 edition.

Wang, X., Smith, K., and Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 13(3):335–364.