Avaliação da Redução de Dimensionalidade e da Discretização Aplicados na Recuperação de Conteúdo de Dados Médicos

Everton A. Cherman[1], Huei Diana Lee[1], Carlos A. Ferrero[1,2], André G. Maletzke[1,2]
Cláudio Saddy Rodrigues Coy[3], João José Fagundes[3], Feng Chung Wu[1,3]

[1]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85856-970 – 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

{evertoncherman,hueidianalee,anfer86,andregustavom}@gmail.com
wufengchung@gmail.com, ccoy@terra.com.br,jjfagundes@mpcnet.com

Abstract. *The content retrieval of time series databases is one of data mining tasks which can be used to support experts in decision-making process. However, the implementation of this task in large datasets may demand a high computational effort. In order to cope with this problem pre-processing methods can be used. In this work we evaluate the influence of dimensionality reduction and discretization methods in content retrieval task using data from anorectal manometry and electrocardiogram. In both datasets, pre-processing methods have presented a positive influence in the content retrieval performance.*

Resumo. *A recuperação de conteúdo em bases de dados de séries temporais é uma das tarefas em mineração de dados que pode auxiliar especialistas no processo de tomada de decisão. No entanto, a aplicação dessa tarefa em grandes conjuntos de dados pode apresentar alto custo computacional. Métodos de pré-processamento podem ser aplicados com o intuito de auxiliar nesse problema. Neste trabalho é avaliada a influência dos métodos de redução de dimensionalidade e de discretização na recuperação de conteúdo. Nessa avaliação foram utilizados dados de exames de manometria ano-retal e de eletrocardiograma. Em ambos os conjuntos de dados, os métodos de pré-processamento influenciaram positivamente no desempenho da recuperação de conteúdo.*

1. Introdução

O avanço da tecnologia tem permitido a aquisição e o armazenamento de grandes quantidades de dados em diversas áreas do conhecimento. Na área médica são registradas informações referentes ao estado de saúde de pacientes por meio de exames como o Eletrocardiograma – ECG – e a Manometria Ano-retal – MA. A análise desses dados pode

proporcionar informações úteis para apoiar especialistas no processo de tomada de decisão. No entanto, a realização de uma análise manual é cada vez mais difícil, devido ao crescente volume de dados armazenados [Lee, 2005, Rezende, 2003].

Nesse contexto, e com a finalidade de otimizar as avaliações, ferramentas computacionais podem auxiliar na análise semi-automática desses elementos, por meio da aplicação de métodos de mineração de dados. Esses métodos são comumente aplicados em dados não-temporais, sendo que, para a aplicação em dados temporais, como dados contidos em exames de MA e de ECG, técnicas específicas devem ser empregadas. A análise desses dados, denominada Mineração de Séries Temporais – MST –, possibilita a extração de informações de acordo com diversas tarefas de interesse, como predição, classificação, detecção de padrões, recuperação de conteúdo, entre outras.

A recuperação de conteúdo em Séries Temporais – ST – é uma aplicação de grande interesse que tem como intuito recuperar, a partir de uma base de dados e de uma ST fornecida como consulta, os exemplares mais semelhantes à essa ST [Mörchen, 2006]. Essa tarefa disponibiliza após análise, informações úteis para especialistas. Na medicina, por exemplo, podem ser disponibilizados todos os registros de pacientes que realizaram exames com características semelhantes a um novo exame.

Um problema decorrente dessa análise está relacionada a quantidade de atributos, também denominada de dimensão, que constitui cada registro, ST, pois influencia diretamente no custo computacional. Desse modo, métodos de pré-processamento podem ser utilizados com o intuito de tornar os dados mais adequados e reduzir o custo computacional da análise de grandes conjuntos de dados [Lin et al., 2003].

A redução de dimensionalidade propõe reduzir uma ST S de dimensão N para uma ST S' de dimensão M onde $M << N$. Para tanto, têm sido propostos métodos, como o *Piecewise Aggregate Approximation* – PAA [Lin et al., 2003].

Por outro lado, existe a discretização, que consiste na transformação dos valores do domínio dos números reais para valores discretos, o que traz benefícios como a simplificação da representação de uma ST e a ênfase em determinadas características, como o comportamento global da ST [Han & Kamber, 2006]. Diversos métodos têm sido propostos [Han & Kamber, 2006, Lin et al., 2003], dentre esses Particionamento Uniforme – PU –, Particionamento por Entropia Máxima – PEM – e *Symbolic Aggregate approXimation* – SAX.

Assim sendo, o objetivo deste trabalho consiste em avaliar o desempenho da tarefa de recuperação de conteúdo em bases de dados de ECG e de MA e verificar a influência da redução de dimensionalidade e da discretização nesse desempenho.

Este trabalho está inserido no projeto de Análise Inteligente de Dados, o qual é desenvolvido por meio de uma parceria entre o Laboratório de Bioinformática – LABI – da Universidade Estadual do Oeste do Paraná – UNIOESTE/Foz do Iguaçu –, o Laboratório de Inteligência Computacional – LABIC – da Universidade de São Paulo – USP/São Carlos – e o Serv. de Coloproctologia da Universidade Estadual de Campinas – UNICAMP.

O restante deste trabalho está organizado da seguinte maneira: na Seção 2 é apresentada a descrição dos conjuntos de dados utilizados, o método e o protótipo de sistema utilizados para a realização dos experimentos; na Seção 3 são apresentados e discutidos

os resultados e na Seção 4 são descritas as conclusões e os trabalhos futuros.

2. Materiais e Métodos

Nesta seção os conjuntos de dados utilizados são apresentados, bem como o método utilizado para realizar a avaliação do desempenho da recuperação de conteúdo quando submetida à redução de dimensionalidade e à discretização. Ao final, é descrito o protótipo de sistema computacional que apoiou na aplicação do método.

2.1. Descrição dos Dados

Dois conjuntos de dados são utilizados, referentes a exames de ECG e a exames de MA. O primeiro conjunto está relacionado ao ECG, no qual os sinais são obtidos por eletrodos pré-posicionados no corpo. Um ECG completo utiliza doze eletrodos, mas apenas alguns são frequentemente acionados para a realização de um diagnóstico simples [Olszewski, 2001]. A base de dados de ECG usada neste trabalho foi obtida no repositório *online* PhysioNet[1] [Goldberger et al., 2000] e é constituída de 200 registros. Cada registro é referente ao comportamento do coração do paciente durante determinado período de tempo estipulado para o exame. Esse registro é relacionado a um paciente anormal, podendo apresentar o sinal clínico taquicardia supra-ventricular ou não supra-ventricular. Na Figura 1 são apresentados exemplos utilizados neste trabalho.

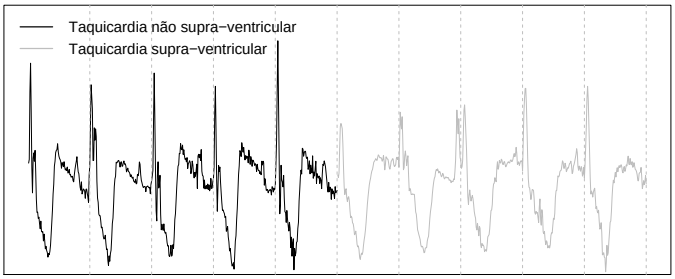

Figura 1. Exemplos de ST concatenadas contidas na base de dados de ECG

O segundo conjunto está relacionado à MA, o qual é um exame importante no diagnóstico de pacientes quanto à incontinência fecal. Essa condição é caracterizada pela perda da habilidade e da capacidade do paciente em controlar a passagem de fezes e gases em tempo e lugar adequados e socialmente aceitáveis e é constituído por graus variados. Os dados desse exame são capturados por meio de oito sensores dispostos radialmente no esfíncter anal [Saad, 2002]. Na Figura 2 é apresentado um exemplo de uma ST representativa dos dados capturados por um dos oito sensores do exame de MA.

A base de dados de MA é composta por 20 exames, os quais foram realizados pelo Serviço de Coloproctologia da Faculdade de Ciências Médicas da UNICAMP no período de Maio/1995 a Novembro/1996. Desses exames, doze representam pacientes normais e oito representam pacientes anormais com incontinência fecal de Grau III.

2.2. Método Experimental

A avaliação da redução de dimensionalidade e da discretização sobre a recuperação de conteúdo foi realizada de acordo com quatro etapas descritas a seguir:

[1] http://www.physionet.org/

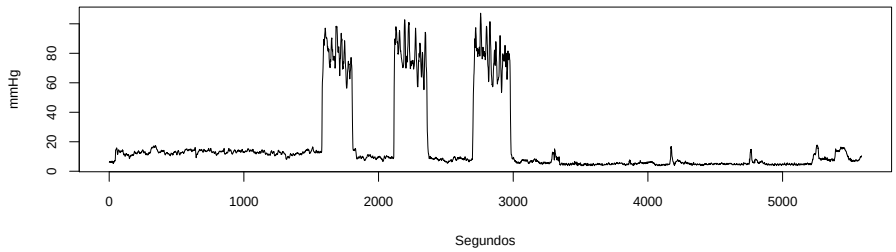

Figura 2. Exemplo de ST obtida de um dos oito sensores do exame de MA

1. Preparação dos Dados;
2. Transformação dos Dados;
3. Aplicação de Recuperação de Conteúdo;
4. Avaliação dos Resultados.

2.2.1. Etapa 1: Preparação dos Dados

Essa etapa tem como objetivo realizar a limpeza, a seleção e a preparação dos dados para formatos adequados às Etapas 2 e 3. Desse modo, são selecionados, nas ST, somente os segmentos de interesse para análise.

No caso dos exames de ECG, é realizado um pré-processamento com o intuito de identificar e segmentar o ciclo de início e o fim de cada batimento cardíaco do paciente e posteriormente selecionar um batimento cardíaco para representar o exame. Além disso, também é realizada uma normalização dos dados, para comparar os batimentos cardíacos com base somente na morfologia de cada ST [Olszewski, 2001].

Para os exames de MA, a preparação dos dados teve como finalidade constituir uma única série temporal que represente os dados dos oito sensores do exame de um paciente de maneira adequada para a diferenciação de pacientes com e sem incontinência fecal [Ferrero et al., 2007]. Para tanto, foram realizados três procedimentos: 1) soma das oito ST obtidas a partir dos sensores; 2) segmentação dos momentos de contração voluntária do paciente; e 3) concatenação das séries temporais representativas dos três intervalos de tempo de contração voluntária. O Procedimento 1 consiste em construir uma ST resultante da soma das oito ST do exame. Essa soma é dada pela Equação 1.

$$Sr_t = S1_t + \ldots + S8_t \tag{1}$$

onde Sr é a ST resultando, e Si_t é o valor da ST do sensor i no instante t.

Referente aos Procedimentos 2 e 3, na Figura 3 é ilustrada a identificação, a extração e a concatenação das seções que representam os intervalos de tempo de contrações voluntárias do paciente. Assim, é possível obter uma representação das três seções de contração voluntária do exame em uma única série temporal.

2.2.2. Etapa 2: Transformação dos Dados

Nessa etapa, são aplicadas as técnicas de transformação dos dados com base nos elementos resultantes da Etapa 1. Essa transformação consiste em gerar novas bases de dados de ST aplicando, nas bases de dados originais, diferentes configurações da técnica de

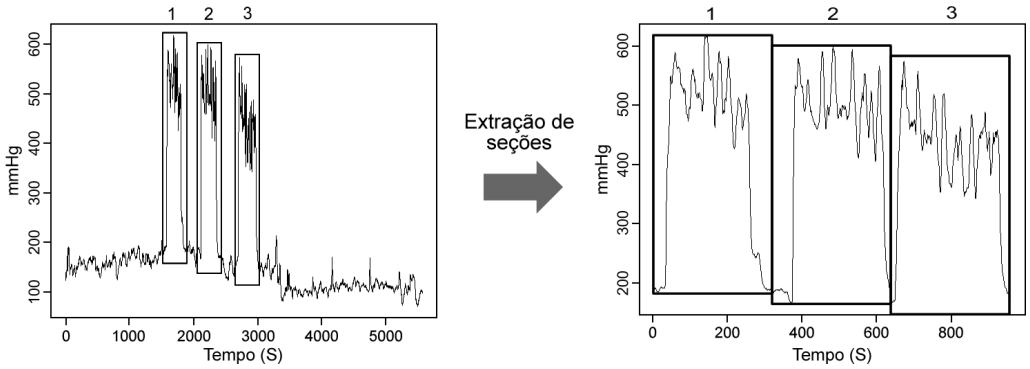

Figura 3. Identificação, extração e concatenação dos momentos de contração voluntária do exame de MA [Spolaôr et al., 2008]

redução de dimensionalidade PAA e das técnicas de discretização PU, PEM e SAX.

A redução de dimensionalidade consiste na diminuição da dimensão da série, isto é, da quantidade de pontos que a representa. Por exemplo, no caso do método PAA, para uma ST de tamanho N, se definida uma redução para dimensionalidade de 50%, esta poderia ser representada por $\frac{N}{2}$ pontos, calculados por médias locais de dois pontos.

A discretização permite que valores contínuos de cada ponto, sejam transformados para um espaço discretizado. Assim, é possível representar uma ST por uma série de símbolos. O alfabeto, indica o número de partes em que foi dividido o espaço contínuo. Diversos critérios de particionamento desse espaço têm sido propostos, entre esses, PU, PEM e SAX [Han & Kamber, 2006, Lin et al., 2003]. Para o PU, todas as partes do espaço dividido contêm o mesmo tamanho. No método PEM, o critério para a divisão do espaço contínuo é definido como todas as partes dessa divisão contenham a mesma quantidade de informações, enquanto o método SAX realiza a divisão do espaço baseado na distribuição gaussiana dos dados.

As configurações consideradas para a discretização e para a redução de dimensionalidade são apresentadas nas Tabelas 1 e 2.

Tabela 1. Configurações dos métodos de discretização

	Tamanho do alfabeto									
Método	2	3	4	5	6	7	8	9	10	Total
PU	•	•	•	•	•	•	•	•	•	9
PEM	•	•	•	•	•	•	•	•	•	9
SAX		•	•	•	•	•	•	•	•	8

Tabela 2. Bases de dados geradas a partir da combinação de discretização e redução de dimensionalidade

	Método de discretização				
Dimensão relativa ao conjunto de dados original	PU	PEM	SAX	Sem discretização	Total
100%	9	9	8	1	27
50%	9	9	8	1	27
12,5%	9	9	8	1	27
Total geral					81

A base de dados sem discretização considerando 100% de dimensão relativa em

relação ao conjunto de dados original, refere-se ao conjunto de dados controle. Desse modo, é possível avaliar a influência do pré-processamento para a tarefa de recuperação de conteúdo e identificar configurações que apresentam desempenho melhor que o controle.

2.2.3. Etapa 3: Aplicação de Recuperação de Conteúdo

Consiste na aplicação da tarefa de recuperação de conteúdo sobre as bases de dados de MA e de ECG, geradas na etapa anterior. No entanto, como a avaliação da aplicação de recuperação de conteúdo é uma tarefa custosa em relação à avaliação de outras tarefas, como a classificação, foram selecionadas algumas bases de dados específicas, geradas na etapa anterior. O critério utilizado para a seleção foi baseado em trabalho anterior [Cherman, 2008], no qual foi avaliado o desempenho da tarefa de classificação, utilizando o algoritmo k-NN, sobre as 81 bases de dados e as que apresentaram melhora com diferença estatisticamente significativa na precisão, em relação ao desempenho do conjunto de dados controle, foram selecionadas para a recuperação de conteúdo.

2.2.4. Etapa 4: Avaliação dos Resultados

Essa avaliação é efetuada para evidenciar a qualidade das recuperações realizadas, no intuito de comparar diferentes configurações de recuperação. Uma das técnicas de análise de resultados consiste na construção do gráfico $Precision \times Recall$ – PR –, o qual permite avaliar a relevância do conteúdo recuperado [Davis & Goadrich, 2006]. Para isso, são fornecidos: um exemplar (registro) como argumento de consulta e um valor de k para o método k-NN [Mörchen, 2006]. De acordo com este algoritmo, são recuperados os k registros mais próximos ao registro fornecido como consulta e em seguida, torna-se possível o cálculo do valor de $Precision$ e de $Recall$ referentes à essa recuperação. Esse processo é realizado por meio das Equações 2 e 3, respectivamente.

$$Precision = \frac{E_{RR}}{T_{ER}} \qquad (2) \qquad\qquad Recall = \frac{E_{RR}}{T_E} \qquad (3)$$

onde, E_{RR} corresponde ao número de exemplos relevantes recuperados, isto é, da mesma classe do registro fornecido como consulta; T_{ER} corresponde ao total de exemplos recuperados; e T_E refere-se ao total de registros contidos na base de dados que contém a mesma classe do registro fornecido como consulta. O valor de k sempre é incrementado em 1, iniciado com 1 e finalizando quando o $Recall$ for igual a 1 (100%).

Para realizar a comparação entre diferentes gráficos PR obtidos do processo de recuperação de conteúdo, é possível utilizar a área sob esse gráfico, a qual é diretamente proporcional à precisão da recuperação, isto é, quando a precisão de recuperação de conteúdo é máxima, o valor da área abaixo do gráfico também será máximo [Davis & Goadrich, 2006]. Com o intuito de generalizar a avaliação, a técnica *Leave-One-Out Cross-validation* é utilizada, a qual consiste em segmentar o conjunto de dados em i partes iguais, onde i representa a quantidade de registros contidos na base de dados. Assim, i iterações são realizadas, sendo o i-ésimo exemplar selecionado para teste e o restante para treinamento [Rezende, 2003].

Sendo assim, para uma base de dados com x exemplares, são gerados x gráficos PR. Consequentemente, x valores de áreas são calculados e a precisão da recuperação

de conteúdo para um determinado conjunto de dados é representada pela média e pelo Desvio-Padrão – DP – dessas x áreas calculadas.

2.3. TimeSSys

Para auxiliar na aplicação desse método, foi desenvolvido um protótipo de sistema denominado *Time Series System – TimeSSys*. Esse sistema permite apoiar pesquisas relacionadas ao tema de ST e é desenvolvido e mantido, em parceria entre o Laboratório de Bioinformática da UNIOESTE e o Laboratório de Inteligência Computacional da USP. O objetivo desse sistema é disponibilizar ferramentas que auxiliem em todas as etapas do processo de análise de séries temporais e também simplificar a integração de ferramentas desenvolvidas por pesquisadores parceiros em suas linhas de pesquisa. O sistema deve possibilitar sua utilização por instituições geograficamente distantes, isto é, o *TimeSSys* deve ser utilizado em um ambiente WEB e que contemple padrões de dados e ferramentas que facilitem a integração e a utilização de funcionalidades desenvolvidas por diversos pesquisadores, sem a necessidade de recursos sofisticados de *hardware*.

O desenvolvimento do sistema é baseado em ferramentas livres. A camada de negócio é implementada em linguagem R[2], a qual disponibiliza uma grande diversidade de funcionalidades relacionadas à análise estatística, análise de séries temporais e visualização. A camada de apresentação tem como tecnologia principal o Rpad[3]. Essa ferramenta fornece um ambiente simples e flexível para o desenvolvimento das interfaces WEB, além de estar diretamente associada com a linguagem R. A linguagem *eXtensible Markup Language* – XML[4] é utilizada para representação dos dados persistentes.

Para a realização dos estudos de caso foram utilizados os módulos de pré-processamento, relacionados à discretização e à redução de dimensionalidade, e o módulo de recuperação de conteúdo. O módulo de discretização permite ao usuário a escolha do método e da configuração do tamanho do alfabeto para a discretização, bem como a possibilidade de selecionar o conjunto de dados para realizar essa operação. O resultado da operação pode ser disponibilizado de maneira gráfica ou por meio da exportação dos resultados. No módulo de redução de dimensionalidade, o usuário também deve selecionar o método e a proporção de redução da dimensão, bem como o conjunto de dados a ser aplicada a operação. O resultado, da mesma maneira que o módulo de discretização, é disponibilizado de maneira gráfica ou por meio da exportação dos dados. Para a recuperação de conteúdo o módulo consiste em disponibilizar classes de *scripts*[5] para realizar a avaliação da tarefa de recuperação de conteúdo sobre um conjunto de dados. O resultado desse módulo é acessado por meio da exportação de resultados.

3. Resultados e Discussão

O método descrito na Seção 2 foi aplicado aos dados e os resultados referentes aos estudos de caso de ECG e de MA são apresentados nas Tabela 3 e 4, respectivamente. Nessas tabelas, a primeira coluna refere-se à identificação da base de dados e a segunda e terceira colunas estão relacionadas, respectivamente, ao método e ao tamanho do alfabeto

[2]http://www.r-project.org/

[3]http://www.rpad.org/Rpad/

[4]http://www.w3c.org/XML/

[5]Pequeno e coeso procedimento/método computacional com o objetivo de realizar uma tarefa específica

utilizados para a discretização do conjunto de dados. Na quarta e quinta colunas são apresentadas a média e o desvio-padrão – DP – das áreas das curvas PR representativas do desempenho da recuperação de conteúdo. A última coluna refere-se à existência de diferença estatisticamente significativa – d.e.s. – na comparação dos desempenhos utilizando o teste estatístico Kruskal-Wallis com nível de significância de 95%. Esse teste foi utilizado para comparar o desemepenho das bases de dados controle, ecg_1 e ma_1, com as demais bases de dados, as quais os métodos de pré-processamento foram aplicados.

3.1. Eletrocardiograma

Para o domínio de ECG, foram avaliadas nove bases de dados, incluindo as de controle (ecg_1) e sem discretização (ecg_2 e ecg_3). Como mencionado, as bases de dados com discretização foram selecionadas por meio de trabalho anterior baseado nos desempenhos obtidos na tarefa de classificação quando comparados ao desempenho da base de dados ecg_1. Desse modo, foram selecionadas as bases de dados ecg_4, ecg_5, ecg_6, ecg_7, sem redução de dimensionalidade, e ecg_8 e ecg_9, com redução para 50% da dimensão original. O método de discretização aplicado nesses conjuntos foi o PU. Vale ressaltar que as bases de dados discretizadas pelos outros métodos e também com redução de dimensão mais acentuada (12,5%) não apresentaram desempenho suficiente para serem selecionadas.

Tabela 3. Resultados relacionados ao conjunto de dados de ECG

i	Dimensão	Método de discretização	Tamanho do alfabeto	Média das áreas	DP	d.e.s.
ecg_1	100%	–	–	0,666	0,228	
ecg_2	50%	–	–	0,661	0,221	
ecg_3	12,5%	–	–	0,654	0,221	
ecg_4	100%	PU	6	0,701	0,235	$\surd$
ecg_5	100%	PU	8	0,701	0,236	$\surd$
ecg_6	100%	PU	9	0,703	0,233	$\surd$
ecg_7	100%	PU	10	0,702	0,236	$\surd$
ecg_8	50%	PU	6	0,697	0,234	
ecg_9	50%	PU	8	0,698	0,233	

Desse modo, as bases de dados somente com redução de dimensionalidade, ecg_2 e ecg_3, não apresentaram diferença estatisticamente significativa em relação à ecg_1. Do ponto de vista prático, é possível utilizar essas bases dados com menor dimensão para realizar a tarefa de recuperação de conteúdo, sem que ocorra degeneração da precisão da tarefa e com o benefício de menor custo computacional relacionado ao tempo e ao espaço.

Para os desempenhos das bases de dados ecg_4, ecg_5, ecg_6 e ecg_7, observou-se diferença estatisticamente significativa em relação ao desempenho de ecg_1. Essas bases de dados contêm apenas discretização utilizando PU, mostrando que o desempenho da recuperação de conteúdo melhorou ao aplicar esse método. Para os conjuntos ecg_8 e ecg_9, os quais contêm discretização e também redução de dimensionalidade, não foi identificada diferença estatisticamente significativa apesar de apresentarem médias de áreas maiores que a média de ecg_1. Esse fato mostra que a influência positiva da discretização nas bases de dados de ECG não foi suficiente para compensar a degeneração causada pela redução de dimensionalidade a ponto de ainda manter uma diferença estatisticamente significativa.

3.2. Manometria Ano-retal

Quanto ao domínio de MA, doze bases de dados foram avaliadas. No caso das bases de dados somente com redução de dimensionalidade, ma_2 e ma_3, não foi observada diferença

estatisticamente significativa em relação à base de dados controle ma_1. Sendo assim, da mesma maneira que para os conjuntos de ECG, é possível utilizar as bases de dados com menor dimensão sem perdas no desempenho da tarefa de recuperação de conteúdo.

Em relação às bases de dados ma_i, para $i = 4, \ldots, 12$, as quais são bases de dados discretizadas, também foi constatada diferença estatisticamente significativa em relação ao conjunto ma_1. Esse fato mostra que a aplicação da discretização não adicionou maior precisão na tarefa de recuperação de conteúdo para os dados de MA.

Tabela 4. Resultados relacionados ao conjunto de dados de MA

i	Dimensão	Método de discretização	Tamanho do alfabeto	Média das áreas PR	DP	d.e.s.
ma_1	100%	–	–	0,504	0,125	
ma_2	50%	–	–	0,505	0,125	
ma_3	12,5%	–	–	0,507	0,127	
ma_4	100%	PEM	4	0,462	0,131	
ma_5	100%	PEM	7	0,451	0,133	
ma_6	100%	PEM	9	0,455	0,133	
ma_7	50%	PEM	4	0,458	0,127	
ma_8	50%	PEM	7	0,451	0,131	
ma_9	12,5%	PEM	4	0,463	0,131	
ma_{10}	12,5%	PEM	7	0,450	0,133	
ma_{11}	12,5%	PEM	9	0,453	0,129	
ma_{12}	12,5%	PU	2	0,477	0,125	

Um aspecto interessante observado nesses resultados foi a predominância de bases de dados discretizadas pelo método PEM, diferentemente do domínio de ECG, o qual teve predominância de bases de dados discretizadas com o método PU. Também pode ser observada a predominância de determinados tamanhos de alfabetos. Nos conjuntos de ECG, os alfabetos de tamanho 6 e 8 apresentaram maior quantidade de bases de dados. No domínio de MA, foi observada maior frequência para os alfabetos de tamanho 4 e 7.

4. Conclusão

Neste trabalho foi apresentada um método para avaliação objetiva da influência do pré-processamento sobre tarefas de MST e também foram apresentados dois estudos de caso envolvendo dados de medicina, mais especificamente dados relacionados à ECG e à MA. Nesses estudos de caso foi possível observar que a redução de dimensionalidade, tanto para ECG quanto para MA, influenciou positivamente na aplicação da tarefa de recuperação de conteúdo, pois, ao utilizar as bases com dimensões reduzidas, foi possível obter precisões semelhantes às precisões obtidas pela utilização das bases de dados com dimensão total. Esse fato representa um ganho computacional relacionado ao tempo e ao espaço necessários para a realização dessa tarefa, principalmente quando aplicada à grandes conjuntos de dados. No caso de ECG, a discretização também influenciou positivamente na qualidade de realização da tarefa de recuperação de conteúdo, obtendo precisões maiores, com diferença estatisticamente significativa, quando comparada aos dados sem redução de dimensionalidade. Para a base de dados de MA essa característica não foi observada.

Neste trabalho foi também apresentado o protótipo de sistema *TimeSSys*, o qual auxiliou na realização do estudo de caso e que tem como o principal objetivo dar suporte ao desenvolvimento de pesquisas inter-institucionais relacionadas à MST.

Como trabalhos futuros pretende-se expandir essa avaliação a uma quantidade

maior de exames de MA e também a outros conjuntos de dados da área médica, bem como realizar uma avaliação mais específica da influência dos diferentes tamanhos de alfabeto de discretização sobre as tarefas de MST.

Agradecimentos: Trabalho realizado com o auxílio do Programa de Desenvolvimento Tecnológico Avançado – PDTA da Fundação Parque Tecnológico Itaipu – FPTI-BR e do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

Referências

Cherman, E. A. (2008). Representação de séries temporais: Aplicação em dados de medicina. Monografia de conclusão de curso. Universidade Estadual do Oeste do Paraná.

Davis, J. & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pág. 233–240, New York, USA. ACM.

Ferrero, C. A., Lee, H. D., Monard, M. C., Wu, F. C., Coy, C. S. R., Fagundes, J. J., & Góes, J. R. N. (2007). Aplicação de métodos de séries temporais para a identificação de seções em exames de manometria anorretal. In *II Congresso da Academia Trinacional de Ciências*, pág. 1–10, Foz do Iguaçu, Brasil.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C., & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*.

Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Elsevier, San Francisco, EUA, 2 edição.

Lee, H. D. (2005). *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. PhD thesis, Universidade de São Paulo.

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th Workshop on Research Issues in Data Mining and Knowledge Discovery*, pág. 2–11, New York, USA. ACM.

Mörchen, F. (2006). Time series knowledger mining. Master's thesis, Philipps-Universität Marburg, Marburg, Germany.

Olszewski, R. T. (2001). *Generalized Featue Extraction for Structural Pattern Recognition in Time-Series Data*. PhD thesis, School of Computer Science Carnegie Mellon University.

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Manole.

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

Spolaôr, N., Lee, H. D., Ferrero, C. A., Coy, C. S. R., Fagundes, J. J., & Wu, F. C. (2008). Um estudo da aplicação de clustering de séries temporais em dados médicos. In *Anais do III Congresso da Academia Trinacional de Ciências*, pág. 1–10, Foz do Iguaçu, Brasil.