Lancet: Um sistema para *clustering* hierárquico aplicado à biodeterioração do concreto

Jean Metz[1*], **Huei D. Lee**[1] , **Leonilda C. dos Santos**[2] ,
Renato B. Machado[1,2] , **Feng C. Wu**[1,3,4]

[1]Laboratório de Bioinformática - LABI
Universidade Estadual do Oeste do Paraná - UNIOESTE
Caixa Postal 961, Foz do Iguaçu, Paraná, 85870-650

[2]Usina Hidrelétrica Binacional de Itaipu

[3]Serviço de Coloproctologia da Faculdade de Ciências Médicas
Universidade Estadual de Campinas - UNICAMP

[4]Instituto de Tecnologia em Automação e Informática - ITAI
Campus da Universidade Estadual do Oeste do Paraná - UNIOESTE
Caixa Postal 1511, Foz do Iguaçu, Paraná, 85856-000

labi@unioeste.br

Abstract. *Machine Learning systems have been used to extract knowledge from several areas, such as: marketing, medical diagnosis and biological analysis. This paper presents a non-supervised Machine Learning system, called Lancet, developed as part of the Intelligent Data Analysis Applied to Concrete Biodeterioraction project in a partnership of the Bioinformatics Laboratory with the Environment Laboratory of Itaipu. This system allows to apply three diferent approaches of agglomerative hierarchical clustering algorithms: Single Link, Centroid and CURE. Moreover, this paper presents a preliminary case study towards the analysis of information collected from Itaipu Dam, that might help to explore the biological phenomenon of concrete biodeterioraction.*

Resumo. *Sistemas de Aprendizado de Máquina têm sido utilizados para extração de conhecimento nas mais diversas áreas, tais como: marketing, diagnósticos e análises biológicas. Este trabalho apresenta o Lancet, um sistema de AM não supervisionado desenvolvido pelo Laboratório de Bioinformática em parceria com o Laboratório Ambiental de Itaipu, que permite a realização do clustering hierárquico aglomerativo considerando três abordagens distintas: Single Link, Centroid e CURE. Além disso, é apresentado um estudo de caso preliminar direcionado à análise de informações coletadas na barragem da Usina Hidrelétrica Binacional de Itaipu, as quais podem auxiliar na exploração do fenômeno biológico da biodeterioração do concreto.*

1. Introdução

O crescente avanço tecnológico e a diminuição de limitações computacionais permitem que instituições das mais diversas áreas produzam grande quantidade de dados. Surge então, a

[*]Bolsista do ITAI - Instituto de Tecnologia em Automação e Informática

necessidade de técnicas cada vez mais eficientes para manipulação desses volumes de dados e extração de informações que possam ser úteis no processo de tomada de decisão. Desse modo, diversos métodos têm sido desenvolvidos para analisar grandes conjuntos de informações com a finalidade de descobrir padrões ou inferir conhecimento. O Aprendizado de Máquina – AM, uma subárea da Inteligência Artificial – IA, tem sido utilizado para auxiliar na construção de sistemas computacionais capazes de adquirir conhecimento de maneira automática. Entre as diversas aplicações dos Sistemas de AM estão, por exemplo, o reconhecimento facial, o controle de robôs, o diagnóstico médico e as análises de informações biológicas [Lee et al., 2000, Ferro et al., 2002, Monard and Lee, 2003].

Este trabalho faz parte do projeto de Análise Inteligente de Dados Aplicado à Biodeterioração do Concreto desenvolvido pelo LABI - Laboratório de Bioinformática[1] da Universidade Estadual do Oeste do Paraná – UNIOESTE – em parceria com o Laboratório Ambiental de Itaipu, e tem como principal objetivo estudar e aplicar técnicas de Inteligência Artificial para auxiliar a detecção e exploração da biodeterioração do concreto na barragem da Usina Hidrelétrica Binacional de Itaipu [Dalcin, 2002, Metz, 2004].

As estruturas de concreto armado são, freqüentemente, comprometidas devido a ação de fatores como biológicos (bactérias e fungos), físicos (variação de temperatura) e mecânicos (vibrações na estrutura). A combinação desses fatores pode causar degradação na estrutura do concreto e conseqüentemente, diminuição do tempo de vida útil desse material, aumentando os custos e esforços para sua manutenção [Santos, 2002]. Esse fenômeno biológico é denominado biodeterioração do concreto. Sob esse prisma, a aquisição de conhecimento e exploração desse fenômeno é indispensável para minimizar os custos com manutenção nas estruturas de concreto e prolongar seu tempo de vida útil. Essas necessidades motivaram o desenvolvimento do projeto de Análise Inteligente de Dados Aplicada à Biodeterioração do Concreto.

Diversos estudos sobre a biodeterioração do concreto têm sido desenvolvidos direcionando-se às análises de estruturas como fundações de edifícios e tubulações de esgoto. Desse modo, pouco se sabe sobre a ocorrência desse fenômeno em barragens de usinas hidrelétricas.

Sob o ponto de vista computacional, o Aprendizado de Máquina é uma das áreas que podem ser utilizadas na exploração dos dados e na detecção de padrões para auxiliar no processo de aprendizado desse fenômeno, ainda pouco estudado. Uma das principais tarefas associadas a esse projeto é o agrupamento ou *clustering* dos dados. Essa tarefa é de grande importância, pois os exemplos do conjunto de dados não estão previamente rotulados, ou seja, são dados não supervisionados [Monard and Baranauskas, 2003]. Para a realização do *clustering* sobre esses dados, foi necessária a implementação do sistema *Lancet*, o qual permite a execução de três algoritmos de *clustering* hierárquico aglomerativo: *Single Link*, *Centroid* e uma variação do algoritmo *CURE* [Guha et al., 1998, Metz, 2004].

Este trabalho está organizado da seguinte maneira: na Seção 2 são apresentados brevemente alguns conceitos relacionados ao *clustering*. Na Seção 3 é apresentado o sistema *Lancet*, assim como os experimentos realizados para sua avaliação. O domínio do estudo de caso preliminar e seus experimentos são discutidos brevemente na Seção 4 e finalmente, na Seção 5 são apresentadas considerações finais.

[1] http://www.foz.unioeste.br/labi

2. *Clustering*

Usualmente, as técnicas de *clustering* são aplicadas sobre grandes conjuntos de dados não supervisionados com objetivo de identificar padrões que possam descrever algum conhecimento. Desse modo, a tarefa consiste no agrupamento desses dados, o qual pode ser realizado de diversos modos: probabilístico, evolutivo, particionamento iterativo, hierárquico entre outros. Neste trabalho utilizou-se a abordagem hierárquica para o *clustering*.

Para a realização do *clustering* são necessárias algumas atividades:

- Representar os dados: a preparação e transformação dos dados para o *clustering* devem ser realizadas nessa atividade. Para isso, algumas técnicas para seleção e construção de atributos são freqüentemente utilizadas.

- Determinar a medida de similaridade: nessa atividade devem ser determinadas as métricas de similaridade que serão utilizadas para o agrupamento dos dados.

- Agrupar os dados: essa atividade resume-se em submeter o conjunto de exemplos a algum algoritmo para realizar o agrupamento dos dados.

- Avaliar os resultados obtidos: nesse estágio do *clustering* o grau de significância dos resultados obtidos pelo algoritmo de *clustering* deve ser analisado. Desse modo, a participação do especialista é fundamental para a correta avaliação dos resultados.

2.1. *Clustering* hierárquico

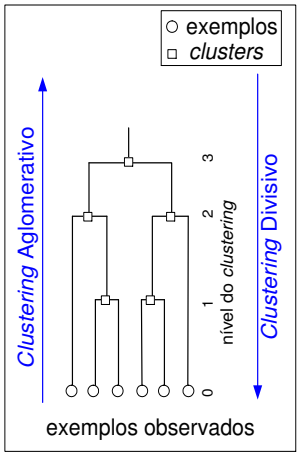

Figura 1: Resultado do *clustering* hierárquico.

No aprendizado não supervisionado, o conjunto de dados é composto por exemplos não classificados. Assim, esse modo de aprendizado consiste em agrupar uma coleção de exemplos não rotulados segundo alguma medida de similaridade. O *clustering* hierárquico constrói hierarquicamente esses agrupamentos em uma estrutura conhecida como dendograma (Figura 1), na qual os nós "pais" agrupam os exemplos pertencentes aos nós "filhos". Essa técnica permite analisar os *clusters* em diferentes níveis de granularidade, pois cada nível do dendograma descreve um conjunto diferente de agrupamentos. Existem duas abordagens que podem ser derivadas do *clustering* hierárquico: aglomerativo e divisivo. Na primeira, inicialmente os dados são distribuídos de modo que cada exemplo represente um *cluster* e então, esses *clusters* são, recursivamente, agrupados por meio de alguma medida de similaridade, até que todos os exemplos pertençam a apenas um *cluster*. O *clustering* divisivo, inicia-se com apenas um agrupamento contendo todos os dados e então divide, recursivamente, o *cluster* mais apropriado até que alcance algum critério de parada, que freqüentemente é o número de *clusters* desejados [Berkhin, 2002].

Uma das maneiras usuais para representar o conjunto de exemplos é o formato atributo-valor (Tabela 1), no qual os exemplos são representados por $e_r = (a_{r1}, a_{r2}, ..., a_{rm})$ e os a_{rq} representam os atributos do exemplo r para $r = 1, 2, ..., n$ e $q = 1, 2, ..., m$, onde n é o

Tabela 1: Conjunto de dados no formato atributo-valor.

Exemplos	a_1	a_2	...	a_m
e_1	a_{11}	a_{12}	...	a_{1m}
e_2	a_{21}	a_{22}	...	a_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
e_n	a_{n1}	a_{n2}	...	a_{nm}

número de exemplos contidos no conjunto de dados e m
é o número de atributos que representam os exemplos.

2.2. Medidas de similaridade

No processo de *clustering*, os agrupamentos são determinados segundo uma medida de similaridade que deve ser, cuidadosamente, selecionada de acordo com as características do conjunto de dados, tais como tipo e escala dos atributos [Martins, 2003].

Para a formação dos agrupamentos devem ser calculadas as medidas de similaridade *intra-cluster* e *inter-cluster*. A primeira medida determina a semelhança entre exemplos pertencentes a um mesmo *cluster*, ao passo que as medidas *inter-cluster* determinam a similaridade entre agrupamentos. As métricas utilizadas, geralmente, para o cálculo da similaridade *intra-cluster* são: Euclidiana, Manhattan e Minkowsky.

As métricas de distância *inter-cluster* são usualmente divididas em duas classes: métricas de grafos e métricas geométricas. A principal diferença entre essas métricas está no modo como os *clusters* são representados. Na primeira, todos os exemplos são considerados para representação do *cluster*, o que permite capturar *clusters* de formatos arbitrários. Essas métricas consideram o conjunto de dados como sendo um grafo completamente conectado, no qual os vértices correspondem aos exemplos e as arestas representam o custo da função de similaridade. As métricas de grafo mais comuns são: *Single Link*, *Average Link* e *Complete Link*. As métricas geométricas, por outro lado, utilizam apenas um valor como representante do agrupamento, geralmente o centro do *cluster*, tornando-se mais propensas a formarem *clusters* esféricos. Entre as métricas geométricas, a mais utilizada é *Centroid* [Olson, 1995, Dash and Liu, 2001, Jain et al., 1999].

3. Sistema *Lancet*

O sistema *Lancet* foi desenvolvido em virtude da necessidade de uma ferramenta computacional para a realização do *clustering*. Para tanto, foram identificados os seguintes requisitos: boa representação dos resultados; flexibilidade para analisar os agrupamentos finais; baixo custo computacional e capacidade de identificar *clusters* de diversos formatos e tamanhos. Também foram observadas algumas características de *software* que seriam desejáveis, tais como a portabilidade e reutilização de código, as quais puderam ser obtidas por meio do paradigma orientado a objetos e da linguagem de programação *JAVA*[2].

Uma boa representação dos agrupamentos finais é fundamental, pois facilita o entendimento dos padrões obtidos. Desse modo, optou-se pela abordagem de *clustering* hierárquico que representa os agrupamentos por meio de um dendograma e com isso, permite que o *clustering* seja analisado em diversos níveis, tornando assim, a análise dos resultados mais flexível [Metz, 2004].

O *clustering* hierárquico tem sido utilizado em busca de padrões em grandes conjuntos de dados. Nessas aplicações, os algoritmos de *clustering* implementam, usualmente, técnicas baseadas em grafos ou geométricas. Outros algoritmos propostos, tal como o *CURE* [Guha et al., 1998], utilizam técnicas intermediárias entre as de grafos e as geométricas, e com isso, podem ser capazes de identificar formatos diversos dos *clusters* com menor custo computacional. Além disso, o *CURE* permite variar alguns parâmetros para aproximar seus resultados

[2]http://www.java.sun.com.

aos obtidos pelas técnicas *Single Link* e *Centroid*. Devido a grande popularidade desses métodos e aos resultados que eles têm apresentado, o sistema proposto neste trabalho é uma variação do *CURE*, e permite ainda, que as técnicas *Single Link* e *Centroid* sejam também utilizadas para o cálculo da similaridade *inter-cluster*. Nesse sistema a similaridade *intra-cluster* é calculada por meio da medida de distância Manhattan.

Dessa maneira, com o sistema *Lancet* é possível escolher entre três abordagens: *Single Link*, *Centroid* e *CURE*. Quando a abordagem *CURE* é utilizada, é necessário ainda definir dois parâmetros de entrada: o número (c) de elementos que serão eleitos representantes do *cluster* e um fator de encolhimento (α) que aproxima os representantes para o centro do *cluster*. Após a aplicação de α, os representantes serão utilizados para o cálculo da distância *inter-cluster*. Nessa abordagem, os valores de α devem estar no intervalo $[0, 1]$ e c um inteiro no intervalo $[1, n]$, onde n é o número de exemplos no conjunto de dados. A variação desses parâmetros permite modelar os agrupamentos e possivelmente identificar, de maneira mais fiel, diversos formatos dos *clusters*.

Para a execução do sistema *Lancet*, o conjunto de dados deve estar representado no formato atributo-valor. Inicialmente, todos os exemplos são considerados *clusters* independentes, assim, o primeiro passo é a construção de uma *Heap* que armazenará esses *clusters*. Depois da construção da *Heap*, a similaridade entre os *clusters* é calculada e em seguida, a *Heap* é ordenada de acordo com a menor distância entre os pares de *clusters* mais próximos. O laço principal do algoritmo é responsável pelo agrupamento dos *clusters*. A cada iteração, o par de *clusters* que possui a menor distância é extraído da *Heap* e agrupado em um novo *cluster*. Após a criação do novo *cluster* e sua inserção na *Heap*, é iniciada uma nova iteração até que reste apenas um agrupamento na *Heap*. O fluxograma de execução desse algoritmo é apresentado na Figura 2.

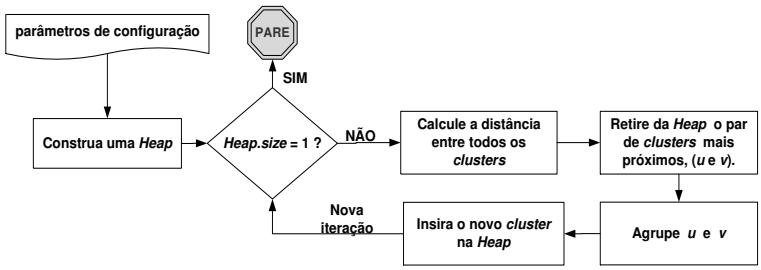

Figura 2: Fluxograma de execução do algoritmo de *clustering*.

- Entradas do sistema: a entrada padrão do sistema *Lancet* é um arquivo *.dat* que armazena o conjunto de dados não supervisionados. No cabeçalho desse arquivo deve estar especificado o número de exemplos, número de atributos e se os exemplos possuem um identificador. Caso os exemplos estejam identificados, a primeira coluna do arquivo deve informar os respectivos identificadores. O sistema permite ainda que dados supervisionados sejam submetidos ao *clustering*. Porém, esses dados são utilizados somente para avaliação do sistema. Nesse caso, o cabeçalho do arquivo deve ser modificado e deverá apresentar, além das informações descritas anteriormente, a quantidade de classes e a distribuição dos exemplos nessas classes, assim como um *flag* que indica que o conjunto de dados será utilizado para teste de avaliação do sistema.

- Saídas do sistema: as saídas geradas pelo sistema dependem do tipo de experimento que está sendo realizado. Quando o conjunto de dados é não supervisionado os resultados obtidos são

o dendograma, que descreve os agrupamentos em cada iteração do *clustering*, a distribuição dos exemplos nos agrupamentos encontrados, as informações sobre o conjunto de dados e um arquivo contendo os exemplos rotulados com os *clusters* encontrados. Além disso, é gerada outra saída que apresenta os resultados de maneira resumida. Por outro lado, quando o conjunto de dados é supervisionado, o resultado apresenta também a performance do algoritmo por meio de uma matriz de confusão que indica a precisão obtida para cada agrupamento.

3.1. Avaliação do sistema

Com o objetivo de avaliar a qualidade dos agrupamentos encontrados pelo sistema *Lancet*, foram realizados experimentos utilizando conjuntos de dados supervisionados. Esses experimentos foram conduzidos sobre quatro conjuntos de dados artificiais compostos por exemplos que representam pontos no espaço em duas dimensões. Adicionalmente, utilizou-se o conjunto de dados *iris* proveniente do repositório de dados da *University of California at Irvine* – UCI[3] [Blake and Merz, 1998] . Na Tabela 2 são descritos, para cada conjunto de dados, número de exemplos, número de atributos, número de classes, classe majoritária[4] – cm – e o erro da classe majoritária. Na Figura 3 é exibida a representação geométrica dos agrupamentos nos conjuntos de dados bidimensionais.

Tabela 2: Resumo dos conjuntos de dados.

conjunto de dados	# exemplos	# atributos	# classes	cm	erro cm
dataset 1	1990	2	5	5	75,20%
dataset 2	1171	2	8	8	76,25%
dataset 3	1184	2	3	3	47,30%
dataset 4	1867	2	5	5	32,20%
iris	150	4	3	-	66,66%

A realização dos experimentos utilizando as bases de dados bidimensionais permite analisar as principais características dos algoritmos *Single Link*, *Centroid* e *CURE*. Espera-se que a abordagem *Single Link* tenha maior facilidade para identificar os agrupamentos de formatos não circulares, ao passo que a abordagem *Centroid* deve identificar corretamente os *clusters* circulares podendo, porém, particioná-los se houver grande diferença entre os diâmetros dos agrupamentos vizinhos. Uma das características da abordagem *CURE* é a possibilidade de simular o comportamento dos algoritmos *Single Link* ou *Centroid*. Desse modo, é esperado que os resultados dessa abordagem sejam semelhantes aos produzidos pelo algoritmo *Single Link* e *Centroid* quando α for configurado com valores próximos de 0 ou próximos de 1, respectivamente. Além disso, devido ao menor número de representantes, espera-se que a abordagem *CURE* apresente melhor performance, em relação ao tempo de execução, que a abordagem *Single Link*.

O conjunto de dados *iris* foi selecionado para o *clustering* devido à distribuição de seus agrupamentos. Esse conjunto contém 150 exemplos representados por quatro atributos (comprimento da pétala, largura da pétala, comprimento da sépala e largura da sépala) e três classes (*Setosa*, *Versicolor* e *Virginica*) igualmente distribuídas, contendo 50 exemplos cada. Nesse conjunto de dados, apenas a classe *Setosa* é linearmente separável das demais. Dessa maneira, os resultados dos experimentos realizados sobre esses dados permitem avaliar o comportamento dos algoritmos diante de agrupamentos mais complexos.

[3] http://lib.stat.cmu.edu/datasets/csb.

[4] Erro cometido sempre que um novo exemplo é classificado como pertencente à classe mais freqüênte.

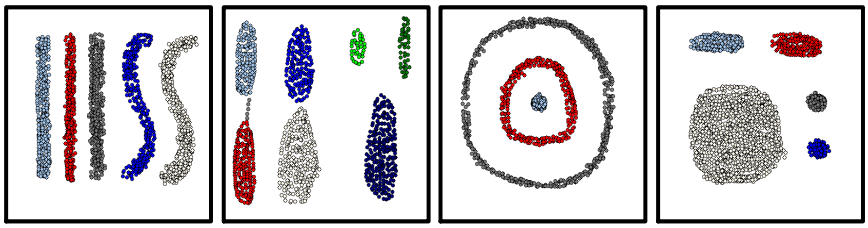

Figura 3: Representação geométrica dos conjuntos de dados. a) *dataset* 1, b) *dataset* 2, c) *dataset* 3 e d) *dataset* 4.

Para cada conjunto de dados foram realizados 11 experimentos, dos quais, dois utilizaram as abordagens *Single Link* e *Centroid* e os demais utilizaram a abordagem *CURE* variando o fator de encolhimento em 0,3, 0,5 e 0,8 e o número de representantes em 5, 10 e 15.

A análise dos resultados alcançados na realização dos experimentos sobre cada conjunto de dados é apresentada a seguir:

- dataset 1: esse conjunto de dados contém cinco *clusters* de formato alongado e linearmente separáveis (Figura 3-a). Desse modo, como era esperado a abordagem *Single Link* identificou corretamente todos os agrupamentos, enquanto os algoritmos *CURE* com α igual a 0,5 e 0,8 e *Centroid* particionaram os agrupamentos. Por outro lado, também era esperado que o algoritmo *CURE* com α = 0,3 encontrasse *clusters* semelhantes aos da abordagem *Single Link*, entretanto, devido ao baixo número de representantes não foi possível identificá-los corretamente.

- dataset 2: esse conjunto de dados possui sete agrupamentos em forma de elipse e também uma pequena conexão entre dois desses agrupamentos (Figura 3-b). Os experimentos realizados considerando esses dados mostraram que todas as abordagens apresentaram bons resultados e que alguns agrupamentos foram corretamente mapeados em quase todos os testes. Entretanto, em alguns experimentos, diversos elementos foram indevidamente agrupados. Apenas os algoritmos *Centroid* e *CURE* com α = 0,8 e c = 15 foram capazes de identificar a conexão como um *cluster* único, enquanto as outras abordagens agruparam a conexão com um dos *clusters* que ela conecta.

- dataset 3: esse conjunto de dados descreve três agrupamentos em forma de circunferências concêntricas, conforme apresentado na Figura 3-c. A aplicação do *clustering* sobre esse conjunto de exemplos foi realizada para analisar como os algoritmos detectam pequenos agrupamentos que estão inseridos em agrupamentos maiores. Observou-se que apenas a abordagem *Single Link* foi capaz de identificar corretamente os agrupamentos. Isso ocorreu devido a facilidade que essa abordagem possui em identificar agrupamentos alongados, e desse modo permitiu percorrer a circunferência dos *clusters* e diferenciá-los. As demais abordagens dividiram o *cluster* externo em duas ou três partições e algumas vezes, uma das partições foi agrupada com os *clusters* internos.

- dataset 4: os testes realizados sobre esses dados permitiram avaliar o comportamento dos algoritmos diante de exemplos que descrevem agrupamentos de tamanhos diferentes (Figura 3-d). Nesse caso, os resultados gerados pelos algoritmos foram próximos do esperado, pois a abordagem *Centroid* particionou o círculo maior e agrupou os círculos menores. O mesmo ocorreu com o algoritmo *CURE* (α = 0,8, c = 5 e c = 15). Por outro lado, como os agrupamentos estão bem distribuídos e separados, os algoritmos *Single Link* e *CURE* com α = 0,3 e α = 0,5 foram capazes de identificá-los corretamente.

- iris: esse conjunto de dados contém três agrupamentos com 50 exemplos cada, dos quais um é linearmente separável. Essa propriedade permitiu que todas as abordagens identificassem, corretamente, pelo menos um dos *clusters* contidos nos dados. Entretanto, os algoritmos não foram capazes de separar os outros agrupamentos. Os métodos *Single Link* e *CURE* com α igual a 0,3 e 0,5 não apresentaram bons resultados, pois diversos exemplos presentes em um dos *clusters* foram inseridos indevidamente em outro *cluster*. As abordagens *Centroid* e *CURE* com α igual a 0,8 apresentaram resultados melhores.

Os experimentos indicaram que, em geral, o algoritmo *Single Link* apresentou os melhores resultados. Das quatro bases de dados bidimensionais utilizadas nos experimentos, esse algoritmo mapeou corretamente três (*dataset* 1, *dataset* 3 e *dataset* 4) e apresentou bons resultados com o conjunto de dados *dataset* 2. Entretanto, esse algoritmo não foi capaz de identificar corretamente os agrupamentos do conjunto de dados *iris*. Desse modo, nota-se que a abordagem *Single Link* é capaz de gerar bons *clusters* sobre conjunto de dados com agrupamentos bem separados. Por outro lado, o algoritmo *Centroid* conseguiu mapear corretamente apenas um dos conjunto de dados (*dataset* 2) e apresentou resultado razoável com a base de dados *iris*.

O algoritmo *CURE*, em alguns casos gerou agrupamentos semelhantes aos encontrados pelas demais abordagens. Entretanto, com os parâmetros de configuração utilizados não foi capaz de reproduzir o comportamento do algoritmo *Single Link* para todos os conjuntos de dados, pois em alguns casos a aplicação do fator de encolhimento acarretou em resultados similares aos da abordagem *Centroid*. Assim, com valores de α menores e com um número maior de representantes, provavelmente, os *clusters* finais seriam semelhantes aos produzidos pelo algoritmo *Single Link* a um custo computacional menor.

Desse modo, quando não há informação prévia sobre as características dos agrupamentos, em geral, as abordagens *Single Link* ou *CURE* considerando maior número de representantes e valores de α menores, devem ser capazes de identificar melhor os agrupamentos. Usualmente, para aplicações reais não se tem conhecimento prévio sobre o conjunto de dados, assim, no estudo de caso realizado neste trabalho foi utilizado o sistema *Lancet* com as configurações *CURE* com α = 0,3 e c = 15.

4. Estudo de caso

As estruturas de concreto são projetadas e construídas para manter condições máximas de segurança, estabilidade e funcionalidade durante sua utilização. Entretanto, a ação de bactérias, fungos e outros organismos promovem a degradação dessas estruturas e conseqüentemente, a diminuição do tempo de vida útil desse material. Esse fenômeno biológico, conhecido como biodeterioração do concreto, envolve a participação de micro e macrorganismos contribuindo para a degradação de materiais expostos a condições ambientais específicas. Ele pode ocorrer pela assimilação de compostos do próprio material ou pela excreção de substâncias durante a reprodução dos microrganismos, que em conjunto com as atividades metabólicas desses seres vivos podem produzir ácidos, proporcionando a dissolução de compostos hidratados do cimento e danificando a estrutura de concreto [Shirakawa et al., 1998].

Alguns especialistas acreditam que a corrosão induzida por microrganismos é, atualmente, o fenômeno que mais causa destruição e prejuízos na área da Engenharia Civil [Shirakawa et al., 1998]. Assim, os estudos relacionados à biodeterioração são, principalmente, motivados pelo custo de manutenção e recuperação de estruturas de concreto.

Devido a sua relevância, a biodeterioração tem sido tema de diversos estudos, entretanto, esses trabalhos são, freqüentemente, direcionados para análises de estruturas subterrâneas de edificações, pontes ou tubulações de esgoto. Assim, poucos trabalhos são realizados com objetivo de analisar a presença da biodeterioração do concreto em barragens de usinas hidrelétricas, o que acarreta na falta de informação sobre a ação desse fenômeno em estruturas de concreto. Dessa maneira, este trabalho tem como um de seus objetivos realizar um estudo de caso preliminar direcionado à análise de informações coletadas na barragem da Usina Hidrelétrica Binacional de Itaipu que poderá auxiliar no melhor entendimento desse fenômeno.

A identificação desse fenômeno é possível, muitas vezes, apenas por meio de observação visual [Santos, 2002]. Um forte indicador da presença de biodeterioração é conhecido como biofilme, caracterizado como uma película de consistência gelatinosa e coloração variável de acordo com a presença de luz e oxigênio. Esse indicador é produzido pela excreção de substâncias extracelulares e produtos ácidos dos microrganismos presentes no meio combinados a outros fatores ambientais, por exemplo, a água [Santos, 2002]. Entretanto, quando esse indicador não é suficiente, outros fatores devem ser analisados. A quantidade de bactérias e fungos, fontes de nutrientes e valores de pH são algumas variáveis que podem ser analisadas para auxiliar na identificação da biodeterioração do concreto.

Para a extração de conhecimento e exploração do processo de biodeterioração foram realizadas as seguintes etapas: coleta dos dados, pré-processamento, experimentos e análise dos resultados. Além disso, foi necessário definir as ferramentas utilizadas para a extração de conhecimento das bases de dados da biodeterioração. As ferramentas utilizadas e as etapas realizadas são apresentadas a seguir.

- Ferramentas utilizadas: duas ferramentas computacionais foram utilizadas para a análise dos dados da biodeterioração do concreto. Inicialmente, empregou-se o sistema *Lancet* na identificação de agrupamentos na base de dados. Em seguida, utilizou-se o sistema See5[5] na construção de regras que pudessem auxiliar na análise do conhecimento extraído pelo *clustering*.

- Coleta dos dados: o especialista do Laboratório Ambiental de Itaipu é o responsável pela coleta das amostras de água, as quais são adquiridas, diretamente, nos drenos por onde escoa a água da barragem. Esses drenos estão localizados em diversos túneis distribuídos em diferentes pontos da barragem e identificados de acordo com a sua altura em relação ao nível do mar. Após a coleta, as amostras são levadas ao laboratório para que sejam analisados os valores do pH, quantidade de bactérias e fungos filamentosos. Foram coletadas 1412 amostras de quatro túneis da barragem: El.20, El.55, El.60 e El.125.

- Pré-processamento: nessa etapa os dados foram formatados segundo as especificações do sistema *Lancet*. Inicialmente, os dados estavam armazenados em planilhas eletrônicas, apresentando alguns atributos categóricos e exemplos com atributos não informados. Para atender às exigências do sistema, os atributos categóricos foram mapeados para atributos numéricos e os exemplos que continham atributos faltantes foram removidos. O conjunto de dados resultante foi composto por 1386 amostras sendo cada uma representada por sete atributos: mês, estação, túnel, contagem de heterotróficas – CH, contagem de fungos – CF, *Escherichia coli* – Ecoli e pH.

- Experimentos e análise dos resultados: os experimentos com o conjunto de dados da biodeterioração foram executados em duas fases. Na primeira, os dados foram submetidos ao *cluster-*

[5]http://www.rulequest.com

ing com objetivo de identificar agrupamentos de exemplos similares. Para isso, o sistema *Lancet* foi configurado para utilizar a abordagem *CURE* com parâmetros α = 0,3 e c = 15. Para esse estudo preliminar, a análise dos resultados iniciou-se no décimo nível do dendograma, e desse modo, foram considerados os resultados obtidos para 10 *clusters*. Após a execução do *clustering*, um novo atributo foi adicionado ao conjunto de dados, que representa o *cluster* ao qual o exemplo foi atribuído. Os agrupamentos finais identificados pelo algoritmo e a quantidade de exemplos neles agrupados são apresentados na Tabela 3.

Tabela 3: Resultado do *clustering* sobre o conjunto de dados da biodeterioração do concreto.

cluster	1	2	3	4	5	6	7	8	9	10
número de exemplos	23	19	4	10	3	219	561	2	542	1

Na segunda fase do experimento, os exemplos classificados pelo algoritmo de *clustering* foram submetidos à ferramenta See5 para geração de regras, as quais puderam auxiliar na tarefa de interpretação dos resultados obtidos pelo *clustering*.

Para auxiliar na identificação de padrões nos dados, foram gerados dois conjuntos de regras. Um desses conjuntos considerou todos os atributos (conjunto A), enquanto o segundo conjunto não considerou os atributos referentes ao mês e estação de coleta (conjunto B).

Observando os conjuntos de regras A e B, foi possível notar padrões de *clusters* presentes nos túneis. Na Tabela 4 são apresentados os *clusters* identificados em cada um dos túneis considerando o conjunto de regras A. Nessa Tabela, os *clusters* são representados por C_i, onde i é o número do *cluster*. É interessante notar que o túnel mais alto, El.125, possui um conjunto de *clusters*, totalmente, distinto dos túneis mais baixos. Ainda, nota-se que os túneis El.20, El.55 e El.60 (túneis mas baixos e próximos entre si) possuem um *cluster* em comum (*cluster* 6). Pode também ser observado que as amostras desses três túneis foram distribuídas em três *clusters* diferentes, o que sugere alguma diferenciação de características nesses agrupamentos.

Tabela 4: *Clusters* identificados a partir do conjunto de regras A.

Túnel	C_0	C_1	C_3	C_4	C_5	C_6
El.20		x		x		x
El.55		x				x
El.60						x
El.125	x		x		x	

Tabela 5: Resultados para o conjunto de regras A.

Túnel		CH	CF	pH
El.20	$\triangle$	672000	7700000	12,20
	∇	0	0	6,99
	μ	22705,95	55197,64	9,55
	σ	102423,20	542388,90	0,70
El.55	$\triangle$	672000	7700000	12,20
	∇	0	0	6,68
	μ	10623,81	27289,89	9,33
	σ	66707,53	396893,30	0,74
El.60	$\triangle$	672000	7700000	12,20
	∇	0	0	6,68
	μ	8479,77	32823,23	9,12
	σ	58676,03	419051,10	0,78
El.125	$\triangle$	74730	73125	9,56
	∇	0	0	7,49
	μ	958,43	1533,58	8,56
	σ	6286,13	8357,52	0,53

Para a análise da possível existência de relações entre os conjuntos de *clusters* encontrados em cada túnel e a quantidade de fungos, bactérias e valores de pH da água, foram calculados os valores máximo, mínimo, média e desvio padrão ($\triangle, \nabla, \mu, \sigma$) para cada conjunto de dados presentes nos túneis. Os resultados para o conjunto de regras A, que considera todos os atributos, são apresentados na Tabela 5. Observando essa tabela nota-se que os valores máximos de bactérias e fungos se mantém nos túneis mais baixos e diminui, consideravelmente, no túnel El.125.

Analisando o segundo conjunto de regras, o qual não considera os atributos mês e estação, observa-se que há um *cluster* com características comuns a todos os túneis (*cluster* 8). Por outro lado, existem dois *clusters* que agrupam somente amostras coletadas no túnel El.125,

o que reforça a hipótese de que esses exemplos apresentam um comportamento distinto dos demais. Além disso, nota-se, novamente, que os túneis El.20, El.55 e El.60 possuem *clusters* em comum. Isso demonstra que, possivelmente, esses túneis apresentam algumas características semelhantes. Os *clusters* (C_i) identificados a partir do conjunto de regras B são apresentados na Tabela 6. e os valores de máximo, mínimo, média e desvio padrão ($\triangle, \nabla, \mu, \sigma$) dos atributos são apresentados na Tabela 7.

A partir da análise dos valores apresentados na Tabela 7, observou-se que em geral, os valores de pH e a quantidade de bactérias e fungos diminuem a medida que a altura do local de coleta aumenta. Entretanto, há uma exceção no túnel El.60, no qual os valores máximo dos atributos CF e CH, não seguiram esse padrão. Essa exceção aos padrões identificados ocorreu devido a grande quantidade de fungos encontrados em algumas poucas amostras retiradas do túnel El.60.

Tabela 6: *Clusters* identificados a partir do conjunto de regras B.

Túnel	C_1	C_3	C_5	C_6	C_8
El.20	x				x
El.55	x			x	x
El.60	x			x	x
El.125		x	x		x

Por meio das análises dos resultados obtidos, a partir desses experimentos, foi possível a identificação de padrões considerados interessantes pelo especialista do domínio e que sugerem a necessidade de estudos mais aprofundados.

Tabela 7: Resultados para o conjunto de regras B.

Túnel		CH	CF	pH
El.20	$\triangle$	638600	218180	10,35
	∇	0	0	7,27
	μ	13186,95	2332,21	8,94
	σ	69974,75	18398,14	0,64
El.55	$\triangle$	504000	54400	11,24
	∇	0	0	6,68
	μ	2771,06	897,23	8,80
	σ	28235,91	4639,81	0,73
El.60	$\triangle$	25280	5490000	9,76
	∇	0	0	7,08
	μ	772,26	30135,45	8,34
	σ	2667,24	373359,70	0,41
El.125	$\triangle$	74730	73125	9,56
	∇	0	0	7,28
	μ	528,32	978,79	8,39
	σ	4843,44	6702,52	0,54

5. Considerações finais

A grande quantidade de dados produzidos e armazenados tem se tornado uma das principais dificuldades relacionadas à análise de informações presentes nesses dados. Em contrapartida, diversos métodos, tais como métodos de Aprendizado de Máquina, que visam a aquisição de conhecimento de modo automático, têm sido aplicados para auxiliar a minimizar esse problema.

O sistema de AM apresentado neste trabalho, *Lancet*, foi avaliado por meio de experimentos, nos quais foram utilizadas diversas configurações e conjuntos de dados supervisionados. Os resultados obtidos por meio dos experimentos permitiram analisar o comportamento dos algoritmos diante de conjuntos de dados que descrevem agrupamentos com formatos e tamanhos variados. Também foi apresentado um estudo de caso que realizou uma análise preliminar do fenômeno biológico da biodeterioração do concreto, cujos experimentos permitiram identificar alguns padrões interessantes, segundo avaliação do especialista do domínio, e que serão considerados em novos estudos durante o desenvolvimento do projeto de Análise Inteligente de Dados Aplicada à Biodeterioração do Concreto.

Atualmente, o *Lancet* é aplicável somente à conjuntos de dados numéricos. Desse modo, algumas melhorias poderão ser adicionadas a esse sistema como a capacidade de manipular atributos categóricos, exemplos com atributos faltantes e tratamento de *outliers*, assim como o desenvolvimento de uma ferramenta de visualização para facilitar a análise do *clustering*. Outros trabalhos futuros incluem a aplicação do *clustering* a outros conjuntos de dados da biode-

terioração que acrescentam algumas variáveis ainda não consideradas neste estudo. Com isso, outros padrões possivelmente poderão ser identificados, ampliando o conhecimento adquirido sobre a biodeterioração do concreto em barragens de usinas hidrelétricas.

Referências

Berkhin, P. (2002). *Survey of clustering data mining techniques*. Relatório técnico, Accrue Software, San Jose, CA.

Blake, C. and Merz, C. (1998). UCI repository of machine learning databases.

Dalcin, A. P. (2002). Aplicando técnicas de aprendizado de máquina para extração de conhecimento de bases de dados ambientais/biológicas. Monografia de conclusão do curso de Ciência da Computação.

Dash, M. and Liu, H. (2001). *Efficient hierarchical clustering algorithms using partially overlapping partitions*. *Lecture Notes in Computer Science*, 2035:495–506.

Ferro, M., Lee, H. D., and Esteves, S. C. (2002). Intelligent data analysis: A case study of the diagnostic sperm processing. In *Proceedings of the ACIS - CSITeA02*, páginas 352–356.

Guha, S., Rastogi, R., and Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, páginas 73–84.

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

Lee, H. D., Monard, M. C., and Esteves, S. C. (2000). Indução construtiva guiada pelo conhecimento: um estudo de caso do processamento sêmen diagnóstico. In *Proceedings of the IBERAMIA/SBIA*, páginas 157–166, Atibaia, SP.

Martins, C. A. (2003). *Uma Abordagem para Pré-processamento de Dados Textuais em Algoritmos de Aprendizado*. Tese de Doutoramento, ICMC-USP.

Metz, J. (2004). Técnicas de clustering hierárquico aplicadas à extração de conhecimento de bases de dados: Um estudo de caso da biodeterioração do concreto. Monografia de conclusão do curso de Ciência da Computação.

Monard, M. C. and Baranauskas, J. A. (2003). *Conceitos sobre Aprendizado de Máquina*, páginas 87–114. In [Rezende, 2003]. Parte I, Capítulo 5, ISBN 85-204-1683-7.

Monard, M. C. and Lee, H. D. (2003). *Processamento de Sêmen Diagnóstico*, páginas 461–463. In [Rezende, 2003]. Parte II, Aplicação V, ISBN 85-204-1683-7.

Olson, C. F. (1995). *Parallel algorithms for hierarchical clustering*. *Parallel Computing*, 21(8):1313–1325.

Rezende, S. O. (2003). *Sistemas Inteligentes: Fundamentos e Aplicações*. Editora Manole, Barueri, SP, Brasil.

Santos, L. C. dos (2002). Estudo quantitativo automatizado no monitoramento de microrganismos ambientais em drenos da barragem de concreto de itaipu. Tese de Mestrado, Centro Federal de Educação Tecnológica do Paraná - CEFET/PR, Curitiba/PR.

Shirakawa, M., John, V., Cincotto, M. A., and Gambale, W. (1998). A biodeterioração de materiais de construção civil. *Téchne*, (33):36–39.