UNIVERSIDADE FEDERAL DO PAMPA

CIÊNCIA DA COMPUTAÇÃO

TONI FERREIRA MONTENEGRO

PADRONIZAÇÃO E CLUSTERIZAÇÃO DA INFORMAÇÃO DE SAÚDE: UMA PROPOSTA PARA DESCOBERTA DE CONHECIMENTO NO AUXÍLIO À COLABORAÇÃO MÉDICA

Trabalho de Conclusão de Curso

Alegrete

2011

TONI FERREIRA MONTENEGRO

PADRONIZAÇÃO E CLUSTERIZAÇÃO DA INFORMAÇÃO DE SAÚDE: UMA PROPOSTA PARA DESCOBERTA DE CONHECIMENTO NO AUXÍLIO À COLABORAÇÃO MÉDICA

Trabalho de Conclusão de Curso apresentado como parte das atividades para obtenção do título de bacharel em Ciência da Computação na Universidade Federal do Pampa.

Orientador: Prof. Dr. Cleo Zanella Billa

Alegrete

2011

TONI FERREIRA MONTENEGRO

PADRONIZAÇÃO E CLUSTERIZAÇÃO DA INFORMAÇÃO DE SAÚDE: UMA PROPOSTA PARA DESCOBERTA DE CONHECIMENTO NO AUXÍLIO À COLABORAÇÃO MÉDICA

Trabalho de Conclusão de Curso apresentado como parte das atividades para obtenção do título de bacharel em Ciência da Computação na Universidade Federal do Pampa.

Orientador: Prof. Dr. Cleo Zanella Billa

Alegrete

2011

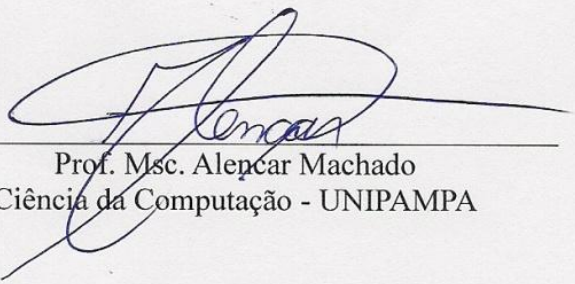
TONI FERREIRA MONTENEGRO

PADRONIZAÇÃO E CLUSTERIZAÇÃO DA INFORMAÇÃO DE SAÚDE: UMA PROPOSTA PARA DESCOBERTA DE CONHECIMENTO NO AUXÍLIO À COLABORAÇÃO MÉDICA

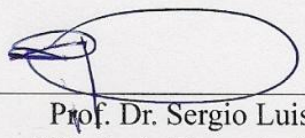
Trabalho de Conclusão de Curso apresentado como parte das atividades para obtenção do título de bacharel em Ciência da Computação na Universidade Federal do Pampa.

Trabalho apresentado e aprovado em: 02 de Janeiro de 2012.
Banca Examinadora:

Prof. Dr. Cleo Zanella Billa
Orientador
Ciência da Computação - UNIPAMPA



Prof. Msc. Alencar Machado
Ciência da Computação - UNIPAMPA



Prof. Dr. Sergio Luis Sardi Mergen
Ciência da Computação - UNIPAMPA

A grande mãezona da família e a pessoa que sempre acreditou que veria seus netos formados. Tenho certeza que mesmo do céu estás feliz por ter seu desejo concedido. Saudades, vó Aracy!

AGRADECIMENTOS

Primeiramente, agradeço a Deus por guiar meus caminhos e fazer crer que tudo dá certo quando se acredita e se tem fé.

Aos meus familiares, que tanto me deram apoio nos momentos difíceis, minha mãe e meu pai, que mesmo de longe, deram força para continuar lutando e não desistir jamais de um sonho. Vocês sempre foram exemplo de carinho, trabalho e dedicação. Obrigado pelos ensinamentos e conselhos que servirão para a vida toda. É graças a vocês que hoje sou acima de tudo, uma pessoa de bem.

Meu irmão, que me ensinou a trilhar seus passos e seguir em frente em busca de uma formação. Você foi um espelho de dedicação e empenho nos estudos. Que possamos estar juntos em nossas caminhadas nos ajudando como sempre fizemos. Mano, você é único, porém o melhor!

Charlise, minha companheira, amiga e namorada, obrigado pela compreensão e por estar ao meu lado em todos os momentos que ficamos distantes. Você me deu forças para seguir em frente, sempre compreendendo minhas ausências. Tinha a certeza de que naquele dia 2 de junho, há 7 anos atrás, tinha encontrado o amor da minha vida e a mulher que estará comigo para sempre. Nega, eu te amo!

Aos colegas, pelos momentos de descontração e para aqueles companheiros da turma de 2007/2, pelas madrugadas afora nas quais dividíamos tarefas e realizávamos trabalhos em conjunto. Aos professores, pelo incentivo e dedicação. Agradeço especialmente ao meu orientador Cleo, obrigado pela paciência interminável, por ter me apresentado à área de informática médica e me incentivado na produção deste trabalho. Sem sua atenção e dedicação nas aulas e no projeto este trabalho não teria tomado forma.

Finalmente, a todos os amigos e familiares que me incentivaram durante a graduação. Cada um de vocês será lembrado porque em algum momento foram importantes na minha vida.

Organizar a informação é acender uma luz no caminho daquele que busca o conhecimento.

Manoel Britto

RESUMO

O uso de Tecnologias da Informação e Comunicação em instituições de atendimento de saúde vem aumentando consideravelmente nos últimos anos. Estimulados pela evolução tecnológica e pela necessidade de se prover informação sobre os dados clínicos de pacientes de uma forma organizada e completa, sempre buscando atender as mais variadas necessidades dos profissionais da área de saúde. Nesse contexto, esta pesquisa apresenta subsídios necessários para organização e padronização da informação médica, bem como, apresenta uma análise das informações contidas em uma base de dados médica através do processo de descoberta de conhecimento no que se refere ao método de *data mining*. Inicialmente, com base nos dados oriundos de um Prontuário Eletrônico do Paciente, realizou-se a padronização das informações pertinentes a um sistema de informação de saúde, de modo que se apresentassem de forma organizada e completa. Tal padronização foi realizada através de ferramentas de representação de dados em computadores, como a linguagem *eXtensible Markup Language* (XML) estruturada por um *Document Type Definition* (DTD). Posteriormente, buscou-se realizar uma análise dos dados através da tarefa de clusterização, por meio do algoritmo de agrupamento *k-means* objetivando a descoberta de conhecimento de modo a auxiliar médicos no que se refere ao trabalho colaborativo que desempenham. No processo de *data mining* utilizou-se a ferramenta *Waikato Environment for Knowledge Analysis* (WEKA). Ferramenta esta que possui código aberto, foi desenvolvida em linguagem Java, sendo possível o total uso de suas implementações de forma gratuita em qualquer outro sistema. Através da análise de clusters geraram-se conhecimentos específicos sobre diagnósticos e medicamentos, no que se refere aos agrupamentos criados para cada um deles e as relações encontradas entre alguns desses atributos pela clusterização e análise dos grupos gerados. A relação entre esses grupos servirá como regras de associação, visando sua utilização em um sistema de prontuário eletrônico que potencialize a colaboração entre médicos através da plena utilização do próprio sistema de Prontuário Eletrônico do Paciente.

Palavras-chave: Informação médica. Organização. Padronização. *Data Mining*. Clusterização. Algoritmo *k-means*. WEKA. Colaboração entre médicos.

ABSTRACT

The use of Information and Communication Technology in health care institutions has increased considerably in recent years. Spurred by technological evolution, and the need to provide organized and complete information of patient on clinical data, this research presents subsidies to organization and standardization of medical information, and presents an analysis of information contained in a medical database through the process of knowledge discovery by data mining process. Initially, based on data from an Electronic Patient Record, there was the standardization of relevant information to health information systems, so it can be presented neatly and completely. This standardization was performed using tools of data representation in computers, such as eXtensible Markup Language (XML) and Document Type Definition (DTD). Subsequently, we analyze data using k-means clustering algorithm, aiming to discover knowledge to help physicians in collaborative work. In the process of data mining, we used the Waikato Environment for Knowledge Analysis (WEKA) tool. This tool has open source and it was developed in Java, being possible to use it for free on any other system. Through cluster analysis, specific knowledge of diagnostics and drugs were generated, as well as some peculiarities of the groups of patients. The relationship between these groups should be association rules, in order to use an electronic medical records system that leverages collaboration among physicians.

Keywords: Medical information. Standardization. Organization. Clustering. Data Mining. K-means algorithm. WEKA. Collaboration among physicians.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

TIC – Tecnologia da Informação e Comunicação

SUS – Sistema Único de Saúde

PNIIS – Política Nacional de Informação em Saúde

PEP – Prontuário Eletrônico do Paciente

EHR – *Eletronic Health Record*

HL7 – *Health Level Seven*

XML – *eXtensible Markup Language*

HTML – *HiperText Markup Language*

XHTML – *eXtensible HiperText Markup Language*

SGML – *Standard Generalized Markup Language*

IM – Informática Médica

CI – Ciência da Informação

CC – Ciência da Computação

AM – Assistência Médica

RES – Registro Eletrônico de Saúde

IOM – *Institute of Medicine*

CPRI – *Computer-based Patient Record Institute*

DM – *Data Mining*

OSI – *Open Systems Interconnection*

ISO – *International Organization for Standardization*

KDD – *Knowledge Discovery in Databases*

DCBD – Descoberta de Conhecimento em Bases de Dados

WWW – *World Wide Web*

XSL – *EXtensible Stylesheet Language*

DTD – *Document Type Definition*

XSD – *XML Schema Definition*

DOM – *Document Object Model*

SAX – *Simple Api for XML*

MM – *Medical Middleware*

SGBD – Sistema de Gerenciamento de Bancos de Dados

ODM – *Oracle Data Mining*

IMD – *Intelligent Miner for Data*

IBM – *International Business Machines*

WEKA – *Wakaito Environment for Knowledge Alalysis*

DB – *Data Base*

EM – *Expectation Maximization*

W3C – *World Wide Web Consortium*

CSV – *Comma-Separated Values*

CLI – *Command-Line Interface*

ARFF – *Attribute-Relation File Format*

UNIFESP – Universidade Federal de São Paulo

API – *Application Programming Interface*

CID – Classificação Internacional de Doenças

OMS – Organização Mundial de Saúde

SUMÁRIO

1. INTRODUÇÃO

Desde o advento das novas Tecnologias de Informação e Comunicação - TIC em sistemas baseados em software, do alto crescimento da medicina e dos conhecimentos médicos no que diz respeito a saúde e o bem estar do ser humano, que vem ascendendo o alto grau de importância de se ter formas de estudar a informação na prestação de cuidados de saúde, bem como, na forma pela qual o conhecimento médico é criado, modelado, compartilhado e aplicado dentro de organizações de prestação de serviços de saúde.

Nas mais diferenciadas áreas médicas que tratam da saúde humana como uma forma de informação e conhecimento disseminado existe a utilização dos mais variados recursos de tecnologias de informação, como aparelhos de diagnósticos por imagem, cadastro de pacientes em prontuários digitais, tratamento de pacientes através de aparelhos controlados por computadores, culminando com a rapidez em diagnosticar doenças e antecipar tratamentos.

Devido ao grande volume de informação referente aos dados clínicos de pacientes e a necessidade de se ter acesso a essas informações em sistemas distintos localizados em diferentes instituições de saúde espalhadas pelo planeta, surgiram discussões sobre formas de organização e integração da informação de saúde do paciente.

Segundo documento do Departamento de Informação e Informática do SUS – Política Nacional de Informação em Saúde (PNIIS, 2004), a informação de saúde no Brasil deve ser organizada e integrada de modo a otimizar o sistema de gestão e promover a disseminação da informação clínica do paciente para isso, estipulou algumas diretrizes a serem adotadas. A seguir destacam-se algumas delas:

a) Fortalecer as áreas de informação e informática nas três esferas do governo, apoiando a sua organização e desenvolvimento através de:
   - criação de mecanismos de articulação, com vistas a integração dos sistemas de informação em saúde; (...).
b) Estabelecer um Registro Eletrônico de Saúde que permita recuperar, por meios eletrônicos, as informações de saúde do indivíduo em seus diversos contatos com o sistema de saúde, com o objetivo de melhorar a qualidade dos processos de trabalho em saúde incluindo a disponibilidade local de informações para a atenção à saúde.

Ao estabelecer um registro eletrônico sobre a saúde de um paciente, a informação contida nele, além de ser muito bem definida e segura, deve ser apresentada de modo que estimule a colaboração entre profissionais da área médica visando uma maior praticidade e rapidez na compreensão das diferentes enfermidades de pacientes, principalmente em cuidados de longa duração de pacientes crônicos ou com doenças sindrômicas.

No estudo de caso realizado por Barsottini (2005), destacam-se três aspectos importantes sobre a colaboração médica por meio de prontuário do paciente ao longo do tempo:

1. Falta de informação sobre a lógica de diagnóstico.
2. Falta de informação sobre a terapia e medicação utilizada.
3. Coleta de dados inadequada e apresentação.

A descoberta de conhecimento, a mineração e o agrupamento em bases de dados são técnicas para seleção de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis, aplicando-as em novos subconjuntos de dados. Tais processos, em conformidade com a colaboração médica centrada no registro eletrônico do paciente são de extrema importância para estimular a cooperação de médicos ao longo de um tratamento de saúde, promovendo a troca de conhecimentos e acelerando processos de estabilização das manifestações de doenças crônicas ou em descoberta de ações preventivas para tratamento de doenças sindrômicas ou similares.

Este trabalho busca a descoberta de padrões significativos à colaboração médica através do agrupamento de dados habitualmente encontrados em registros eletrônicos do paciente. Os padrões serão inferidos através do uso de algoritmos de classificação de dados adotados para fins de definir similaridades entre variáveis e aproximá-las segundo uma medida.

1.1. Objetivos

Para este trabalho o objetivo é a descoberta de conhecimento médico através de regras inferidas pelo agrupamento não supervisionado de dados oriundos de um sistema de Prontuário Eletrônico do Paciente (PEP), no qual deverá obter os resultados de um conjunto de testes realizados utilizando algoritmos de *clustering*. Tais regras de conhecimento deverão ser futuramente gravadas em um formato no padrão XML visando sua utilização e transferência dentro de um sistema de prontuário eletrônico, facilitando a colaboração médica, tanto na visualização gráfica das regras de associação, quanto na cooperação entre

profissionais da área no que diz respeito a troca de conhecimento para tratamentos de enfermidades, diagnósticos ou afins.

Para tanto, primeiramente será elaborada uma pesquisa conceitual contextualizando a importância do uso da informática na medicina e a informação médica, bem como as definições do Prontuário Eletrônico do Paciente. Serão apresentados também, conceitos específicos sobre mineração de dados, análise de *clusters*, e em relação ao algoritmo utilizado nos testes.

Após, serão selecionadas as informações relevantes a serem analisadas no agrupamento, realizado um conjunto de teste específico e apresentados os resultados e as regras obtidas a partir dos mesmos.

1.2. Conceitos

Nesta seção serão apresentados os conceitos básicos sobre Prontuário Eletrônico do Paciente e agrupamento de informações e em seguida será descrito o algoritmo de agrupamento k-means.

1.2.1. A informática no contexto da medicina

A evolução tecnológica das últimas décadas foi acompanhada pela grande utilização das mais variadas tecnologias em muitas áreas além das áreas de informática, como é o caso da área de saúde, proporcionando evoluções na medicina, seja na obtenção, análise e execução de exames, na organização de sistemas de registro eletrônico do paciente, ou até na representação gráfica de evoluções de quadros clínicos e tratamentos.

Porém, não é só nesse contexto que a informática contribui para o avanço da medicina, mas também no que diz respeito ao armazenamento e organização de grandes volumes de informações que são geradas atualmente nos centros de saúde espalhados pelo mundo.

Como forma de auxílio a todos os profissionais da área médica e de saúde, visando atender a todas as necessidades recorrentes da expansão das tecnologias em conjunto com o aumento das demandas de informações disseminadas diariamente, que surgiram necessidades de adotar as Tecnologias de Informação e Comunicação para a prática da boa medicina, com uma maior rapidez nas tomadas de decisão, do mesmo modo em que as agências bancárias, por exemplo, não conseguem operar de forma completa e rápida sem o apoio da informática e

da computação, a medicina, cada vez mais, atua em conjunto com a informática buscando atender de forma apropriada as necessidades da população.

A informação referente a saúde de um indivíduo está devidamente ligada a forma de como vamos trabalhar com dados extremamente importantes para a manutenção da vida de cada ser humano. Levando em conta os avanços do mundo moderno no surgimento das grandes tecnologias como a internet e os recursos da web, o tratamento e a disseminação dessas informações por meio de computadores tornaram-se objeto de estudos nas mais variadas áreas tanto de informática como na medicina.

Surge, então, a Informática Médica, campo de atuação que objetiva a utilização de recursos, dispositivos e métodos necessários para aperfeiçoar a aquisição, armazenamento, recuperação e utilização das informações em saúde e biomedicina. É uma disciplina que engloba conceitos e faz intersecção direta com áreas como a Ciência da Informação, Ciência da Computação e Assistência Médica.

Para Sigulem *et. al* (1998), "o objetivo fundamental da Informática Médica é o de colocar a disposição do Médico a informação, onde e quando ela for necessária". Portanto, quanto mais informação disponibilizar ao médico, seja ela fonte de uma base de dados do paciente, colhida através da colaboração com outros profissionais ou até mesmo, inferida através da análise e classificação dos dados disponíveis, mais rapidamente ele terá subsídios para sugerir tratamentos e prescrever medicamentos.

1.2.2. Prontuário Eletrônico do Paciente

Na literatura, várias são as definições para o registro eletrônico de um paciente, vamos sintetizar e descrever algumas delas. A primeira definição destacada foi a do *Institute of Medicine* (IOM, 1997), definindo que o prontuário eletrônico do paciente é um "registro eletrônico que reside em um sistema especificamente projetado para apoiar os usuários fornecendo acesso a um completo conjunto de dados corretos, alertas, sistemas de apoio à decisão e outros recursos, como links para bases de conhecimento médico".

Por sua vez, o *Computer-based Patient Record Institute* (apud MURPHY, HANKEN e WATERS, 1999) ressalta que o prontuário eletrônico é um registro computadorizado de paciente cuja informação é "mantida eletronicamente sobre o *status* e cuidados de saúde de um indivíduo durante toda a sua vida".

Visto que as novas criações computacionais tendem a substituir vários documentos ou ferramentas físicas, como no caso do prontuário de papel, que se torna obsoleto quando

comparado com o registro eletrônico. Em sua pesquisa, Massad (2003), lista várias vantagens do Prontuário Eletrônico do Paciente (PEP) sobre o prontuário escrito em papel, dentre elas destacam-se:

> **Legibilidade:** registros feitos à mão são difíceis de ler, na maioria das vezes. Os dados na tela ou mesmo impressos são muito mais fáceis de ler.
> **Segurança de dados:** a preocupação com os dados é frequente, principalmente no que se refere a perda desses dados por mau funcionamento do sistema, porém um sistema bem projetado com recursos de "backup" seguros e planos de desastres, pode garantir melhor e de forma mais confiável os dados contra danos e perdas.
> **Confidencialidade dos dados do paciente:** o acesso ao prontuário pode ser dado por níveis de direitos dos usuários e este acesso ser monitorado continuamente. Auditorias podem ser feitas para identificar acessos não autorizados.
> **Integração com outros sistemas de informação:** uma vez em formato eletrônico, os dados do paciente podem ser integrados a outros sistemas de informação e bases de conhecimento, sendo armazenados localmente ou a distância.
> **Assistência à pesquisa:** o dado estruturado pode facilitar os estudos epidemiológicos. Os dados em texto-livre podem ser estudados por meio de uso de palavras-chave.(Sittig, 1999 apud MASSAD *et al*, 2003, p. 7,8).

O uso de prontuários eletrônicos desenvolveu a importância de se manter a consistência das informações que deverão estar contidas nos mesmos, estipulando regras padronizadas para adequação dos meios digitais no registro das informações do paciente, no uso dessas informações em prol de seu bem-estar e em busca de um melhor desenvolvimento do conhecimento médico sobre os dados que se apresentam a ele através desses sistemas.

2. PADRONIZAÇÃO DA INFORMAÇÃO EM PRONTUÁRIOS ELETRÔNICOS

Trabalhar com a integração de sistemas de informação médicos nada mais é do que formatar esses sistemas de forma a serem compreendidos globalmente por distintos sistemas, localizados em distintas áreas, sejam elas físicas ou de aplicação, podendo acessar a mesma informação de forma concisa e ao mesmo tempo, otimizando a troca de informação e os processos administrativos e de assistência médico-hospitalar.

Estar com todas as informações do paciente disponíveis eletronicamente de modo automatizado e organizado é algo que vem sendo desejado pelas instituições de saúde há muito tempo. Os sistemas de prontuários eletrônicos não são triviais se levarmos em consideração que um paciente é atendido em diversos locais situados em diversos lugares do mundo sendo que cada instituição armazena uma parte das informações do paciente em seus sistemas locais.

A padronização da informação contida em um prontuário eletrônico do paciente é indispensável para criação, modelagem e compartilhamento de conhecimentos médicos, tanto no enfoque do bem estar do paciente, quanto na melhoria da capacidade de colaboração e cooperação entre os profissionais da saúde buscando anteder as necessidades dos profissionais e os anseios da sociedade em geral, pois todas as pessoas em alguma etapa de suas vidas necessitam de algum tipo de atendimento médico, gerando informação aos sistemas de saúde.

> A maioria dos dados de saúde é um texto narrativo que muitas vezes não é acessível fácil de encontrar na estação de trabalho clínica. Padrões relacionados com XML (XMLSchema, Xforms, XSL e Topic Maps, etc) oferecem uma infraestrutura que pode mudar essa situação. Em sua proposta, o padrão XML atua como um padrão com uma ideia de “plug-and-play XML”, ou seja, desenvolver novas aplicações que utilizem padrões XML, para que o mesmo torne-se mais abrangente (SCHWEIGER et al, 2004).

Assumindo o cenário de análise obtido através da base de dados médica do tipo *Microsoft Database* (MDB) extraída junto a Unifesp, houve a percepção de que a mesma não possuía os requisitos necessários para a execução dos testes porque apresentava inconsistência, duplicação de dados e tabelas não normalizadas. Assumindo esse contexto, visou-se a criação de uma espécie de documento padronizado para armazenar os dados médicos do prontuário do paciente juntamente com os resultados dos agrupamentos e testes realizados.

Para facilitar a elaboração da padronização dos dados coletados na base, algumas tecnologias possibilitam a organização de documentos com dados através de estruturas bem elaboradas. Essas ferramentas aliadas com a rede mundial de computadores (*internet)* permitem uma distribuição cada vez mais rápida, fácil e segura da informação entre sistemas distintos e complexos.

Dentre estas tecnologias, neste trabalho destacam-se os padrões referenciais em dados de saúde HL7, OpenEHR, também as tecnologias de formatação e apresentação de dados XML e DTD, ferramentas de análise da sintática de documentos XML (DOM e SAX). XML e DTD constituem a base para formação e proposta do padrão de dados extraídos do banco de dados e farão parte de um sistema de prontuário eletrônico. DOM e SAX servirão para extrair os atributos relevantes da base de dados XML criada, bem como para ser a base de informações para os testes de agrupamento e, também, para fomentar o uso de padronização em informações de prontuários eletrônicos.

A seguir, são descritas as tecnologias e ferramentas citadas acima, que podem ser utilizadas para tratamento da informação médica em PEP retirada da literatura analisada.

2.1. Padrão HL7

A *Health Level Seven International*[1] (HL7) é uma organização sem fins lucrativos fundada em 1987 e foi designada a elaboração e manutenção de padrões para a interoperabilidade da tecnologia da informação de saúde. Possui membros em mais de 55 países do mundo que fornecem um framework para intercâmbio, integração, compartilhamento e recuperação das informações de saúde por meio eletrônico. Essas normas definem como as informações de saúde são empacotadas e transmitidas de uma parte a outra, define a linguagem, estrutura e os tipos de dados necessários para uma perfeita integração entre sistemas.

> O HL7 tem o objetivo de definir normas para a transmissão de dados como, por exemplo, dados sobre registros de pacientes, admissão, transferências de pacientes, seguros, taxas e contas a pagar, pedidos e testes de laboratório, exames de imagem, observações médicas e de enfermagem, prescrições de dieta, pedidos à farmácia, pedidos de suprimentos e arquivos; enfim, o padrão HL7 tem a capacidade de comunicar sistemas considerados heterogêneos como o sistema administrativo de

[1] Disponível em: < http://www.hl7.org/>

> um hospital, o sistema financeiro e o sistema de informações clínicas do paciente. (PETRY, LOPES, 2005).

Os padrões HL7 suportam a prática clínica e de gestão, entrega e avaliação dos serviços de saúde. A denominação "Level Seven" refere-se ao nível de aplicação do modelo de sete camadas de Comunicação entre Sistemas Abertos denominado *Open Systems Interconnection* (OSI) mantidos pela Organização Internacional de Padronização (ISO). Tal padrão é reconhecido como um dos padrões mais utilizados no mundo.

O padrão HL7 tem por unidade básica de informação que é trocada entre os sistemas designada através de mensagens. Especificando características de troca de mensagens entre sistemas distintos, as mais variadas formas e tipos de mensagens e do que cada mensagem é constituída. (PETRY, 2005).

Portanto, o padrão trabalha geralmente com protocolos de troca de mensagens, mas também elabora outras iniciativas que são incluídas nas definições e atualizações de seus padrões. Destacam-se algumas dessas definições:

- Sintaxe de Ardem (para representação de conhecimento médico)
- Padronização de estruturas de documentos XML
- Especificação para definição de vocabulário para uso em mensagens e documentações médicas
- Especificações de modelagens de Prontuários Eletrônicos do Paciente
- Especificações de segurança, privacidade e confidencialidade das informações médicas, entre outras.

Este padrão foi empregado como principal referencial teórico para elaboração da proposta de organização a partir de documentos XML da informação médica presente na base de dados estudada.

2.2. Padrão OpenEHR

O OpenEHR[2] é um padrão aberto que define um conjunto de especificações para a arquitetura de um Registro Eletrônico de Saúde, do inglês *Eletronic Health Record* (EHR), ou simplesmente, Prontuário Eletrônico do Paciente. O padrão não é considerado um aplicativo ou elemento de software, pois sua finalidade do projeto é permitir a interoperabilidade

[2] Disponível em: <http://www.openehr.org>

semântica da informação de saúde entre e dentro de sistemas de EHR em um formato não proprietário.

A fundação OpenEHR que possui a propriedade intelectual das especificações da arquitetura do padrão é uma empresa sem fins lucrativos, sendo os sócios fundadores a *University College London* e a empresa australiana *Ocean Informatics*. A fundação possui mais de 15 anos de pesquisas e tem a colaboração de uma comunidade internacional de pessoas que compartilham o objetivo de apoiar o atendimento de saúde do paciente melhorando a qualidade na criação e interoperabilidade em sistemas de prontuários eletrônicos.

No artigo de Leslie (2007) define-se que o OpenEHR é uma especificação para proteger as informações de saúde partilháveis e fornecer uma base sobre a qual se deve construir interoperáveis e modulares aplicações de software de modo a suportar o fluxo de trabalho clínico distribuído, como é o caso dos cuidados de saúde.

Leslie (2007) também cita algumas maneiras nas quais podem ser implementadas as especificações do padrão OpenEHR, destacam-se algumas delas:

- EHR escalável: a partir de um registro pessoal de saúde para organizações de pequeno/médio/grande porte ou para sistemas de registros clínicos regionais, estaduais ou até em programas nacionais de saúde.
- Baseado em mensagens, *web-service*, aplicações de middleware e integrado a sistemas clínicos existentes.
- Visando a interoperabilidade semântica, buscando ter um Prontuário Eletrônico do Paciente verdadeiramente compartilhável.

2.3. XML

O XML (*eXtensible Markup Language*), é uma linguagem derivada do SGML (*Standard Generalized Markup Language*), regulada pelo W3C (*World Wide Web Consortium*) é um padrão de representação de dados em computador. O XML provê um formato para descrever dados estruturados o que facilita declarações mais precisas do conteúdo (Zavalik, 2004).

Um documento XML é estruturado de forma hierárquica através de *tags* definidas de acordo com a necessidade do usuário. Facilita a criação de estruturas-base para documentos e padrões de dados altamente estruturados de modo a facilitar a flexibilidade e fornecer a

aplicação de semântica a textos ou dados com composições complexas de forma simples e organizada.

O princípio do XML é, na verdade, muito simples: qualquer um pode criar marcações que adicionam semântica a um texto, dizendo o que significa um determinado pedaço de informação (Nardon, 2000).

Essas funcionalidades tem despertado interesse da área de informática médica no que diz respeito ao uso de padrões e pelo fato de ter a característica de poder ser utilizado por qualquer plataforma de software ou hardware, também pelo fato de ser uma linguagem simples de manipular e estruturar dados textuais, como os que encontramos em prontuários, tanto escritos quanto eletrônicos.

2.4. DTD

Segundo a W3C[3], o DTD (*Document Type Definition*) é a especificação de um conjunto de regras para definir a estrutura de um documento. Define os blocos de construção de arquivos XML. Cria uma estrutura do documento através de uma lista de elementos e atributos como se fosse um vocabulário da linguagem XML, onde as definições feitas no DTD surtem efeito no XML.

As principais utilidades do uso de DTD em XML são destacadas como segue: com um DTD, cada documento XML pode transportar consigo uma descrição de sua própria estrutura e grupos independentes de pessoas podem usar um padrão DTD para intercâmbio de dados e também pode verificar se um dado vindo do exterior é válido ou não.

2.5. XML SCHEMA

O XML *Schema*[4] é uma alternativa ao DTD baseado no próprio XML. Utilizado para descrever a estrutura de um documento XML. Também é referido como XML *Schema Definition* (XSD).

O modelo tornou-se um padrão de recomendação da W3C em 2001, e tem por objetivo definir os blocos de construção legal de um arquivo XML, da mesma forma que um DTD. Pode definir elementos e atributos que aparecem em um XML, bem como, a ordem e o

[3] Disponível em: <http://w3schools.com>
[4] Disponível em: < http://www.w3schools.com/Schema/schema_intro.asp>

número de elementos filhos, verificar se um elemento está vazio ou não, os tipos de dados para os elementos e atributos e os valores fixos para os mesmos.

É o sucessor do DTD, pois será altamente utilizado na maioria das aplicações web pelo fato de ser extensível e escrito com a própria linguagem XML, também pelo fato de suportarem tipos de dados e *namespaces*[5] em sua estrutura.

Lapeyre (2005) destaca alguns motivos para usar gramáticas DTD ou *Schemas* em processos de validação, documentação e automação. Dentre esses motivos destacam-se:

- é uma máquina de validação da estrutura do documento;
- é um contrato entre produtores e consumidores (tanto para validar o que foi produzido/enviado quanto para verificar se foi corretamente recebido/consumido);
- por ser uma especificação formal de tipos de informação, determinando quais tipos de dados serão aceitos ou não em uma gramática;
- para assegurar que a informação está de acordo com o modelo (validação).

Com o uso de estruturas como DTD, XMLSchema e XML é possível promover a elaboração de um padrão para um sistema de prontuário eletrônico, onde sua estrutura geral poderá ser validada e os dados num documento XML sejam bem formados.

2.6. DOM E SAX

O DOM (*Document Object Model*) é uma especificação de interface de programação neutra com a finalidade de manipular, interpretar, extrair, gerar dados e tratar eventos em arquivos HTML, XHTML e XML. O objetivo do DOM é servir como uma interface uniforme de programação para aplicações diversas que manipulem XML sem considerar plataforma e linguagem utilizada.

O SAX (*Simple* API *for* XML) é uma segunda proposta de API para Java que manipula dados XML. Possui uma definição de busca mais simplificada, que varre um documento XML a partir do início e produz eventos com a leitura, notificando abertura, fechamento e alterações de elementos de uma *tag*. Proposta ideal para percorrer estruturas XML que possuam poucos registros.

A figura 3 demonstra as estruturas tanto do DOM quanto do SAX, e como elas se comportam para percorrer um mesmo documento XML de exemplo.

[5] *Namespaces* XML proveem um método para evitar conflitos com nomes de elementos.

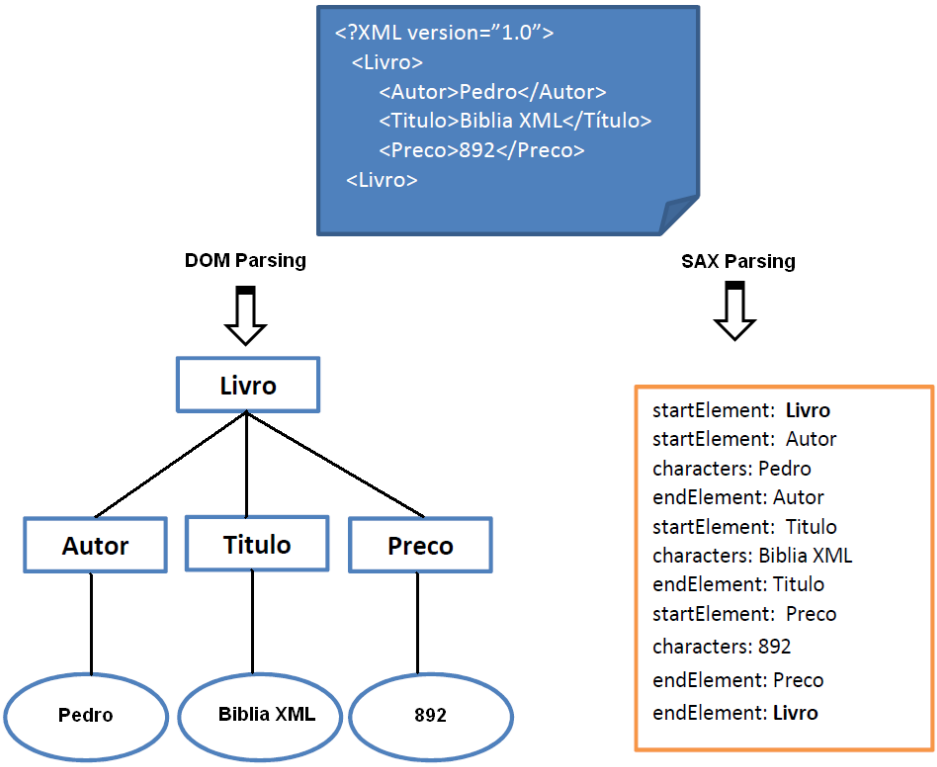


FIGURA 1 – Comparação da estrutura entre DOM e SAX

Geralmente, o DOM monta uma estrutura hierárquica de objetos, tal estrutura tem o formato de uma árvore que permite a navegação nas extremidades do documento criado de modo a garantir a validação das entradas dos dados de uma estrutura XML, por exemplo.

Já o SAX é uma estrutura baseada em eventos que são disparados durante o processamento do documento. Os eventos podem ser capturados por ouvintes cadastrados e diversas ações podem ser tomadas em cada situação. Alguns exemplos de eventos da API SAX podem ser: início e fim do documento; início e fim de um elemento; início e fim de uma *tag*; nó que contenha algum caractere; nó que contenha comentários; etc.

Estas tecnologias foram exploradas neste projeto buscando organizar as informações de modo que fosse possível a criação de um método padrão de se criar, editar e extrair e gravar os dados com informações médicas, e que o mesmo documento padrão pudesse ser, futuramente utilizado no projeto e implementação de um sistema completo de Prontuário Eletrônico do Paciente.

Este projeto adotou uma estrutura padronizada para representação das informações de um prontuário do paciente baseada em uma definição DTD (Apêndice A) que especifica os atributos relevantes para a construção dos registros através de documentos XML (Apêndice B). Esta representação padronizada servirá, tanto para coleta e arquivamento dos dados da clusterização, quanto para a futura implementação de um sistema de PEP.

3. CLUSTERIZAÇÃO OU AGRUPAMENTO DE DADOS

O processo de análise estatística de dados desempenha um importante papel na área de Mineração de Dados - *Data Mining* (DM), vários são os métodos de se observar os dados oriundos de bases de dados e descobrir conhecimento objetivando uma busca efetiva por padrões de interesse dentro de um conjunto de informações.

Frente a essa área, surgiu uma inovadora subárea que busca atender a necessidade de analisar essas informações, envolvendo áreas afins como banco de dados, inteligência artificial e estatística, denominada *Knowledge Discovery in Databases* (KDD) ou Descoberta de Conhecimento em Bases de Dados (DCBD), termo que se refere ao conceito de buscar e descobrir conhecimento sobre dados presentes em bancos de dados.

> Algoritmos de Clusterização dividem os dados em grupos úteis ou significativos, chamados *cluster*, nos quais a similaridade intracluster é maximizada e a similaridade inter-*cluster* é minimizada. Estes *cluster* descobertos podem ser usados para explicar as características da distribuição dos dados subjacentes e assim servir como base para várias técnicas de análise e mineração de dados. As aplicações de clusterização incluem caracterização de diferentes grupos de clientes baseado nos padrões de compra, categorização de documentos na World Wide Web, agrupamento de genes e proteínas que possuem funcionalidades similares, agrupamento de localizações geográficas propensas a terremotos através de dados sismológicos, etc. (KARYPIS, 2002,p. 4).

Como parte dessa abordagem, a análise de *clusters* ou clusterização é definida como a classificação não supervisionada de dados, formando agrupamentos ou *clusters*. Ela representa uma das principais etapas de processos de análise de dados (...) (Jain et al., 1999).

Ochi, Dias e Soares (2004, p.3) abordam que a distância entre dois dados é considerada como um importante critério para identificar sua similaridade, onde as diferenças dos valores entre cada atributo são trabalhadas, ou seja, maior é a similaridade entre o par dos dados quanto menor for a distância entre eles.

Para Ochi, Dias e Soares (2004), as medidas de distâncias geralmente utilizadas são:

a) **distância euclidiana:** considera a distância *d* entre dois dados Xi e Xj no espaço p-dimensional.

$$d(X_i, X_j) = \left[\sum_{l=1}^{p} (x_{il} - x_{jl})^2 \right]^{\frac{1}{2}} \quad (1)$$

b) distância city-block: corresponde a soma das diferenças entre todos os *p* atributos de dois dados Xi e Xj, não sendo indicada para os casos em que existem uma correlação entre tais atributos:

$$d(X_i, X_j) = \sum_{l=1}^{p} \left| x_{il} - x_{jl} \right| \quad (2)$$

> O problema de clusterização possui aplicações nas mais variadas áreas de pesquisa incluindo, por exemplo: computação visual e gráfica, computação médica, biologia computacional, engenharia de transportes, redes de computadores, entre outras. (OCHI, DIAS, SOARES, 2004).

Dentro do contexto da informação médica, a análise de cluster é uma forte aliada no que se refere a descoberta de grupos de similaridades que são encontrados pela execução de algoritmos de *data mining* em coleções de dados que contenham informações de pacientes. Com o auxílio de ferramentas apropriadas e baseando-se em tecnologias de ponta, como é o caso da clusterização, cada vez mais os profissionais da medicina terão em suas mãos os subsídios tecnológicos necessários para melhor desempenhar seu papel como profissional.

3.1. Algoritmo k-means

Os algoritmos de clusterização são utilizados frequentemente em sistemas que necessitam da busca por padrões ou similaridades entre os dados, como é o caso da mineração e análise de dados.

Um dos mais conhecidos algoritmos de clusterização ou agrupamento de dados é o k-means. Para Jain et al. (1999) sua popularidade se dá devido a sua facilidade de compreensão e implementação e sua ordem de complexidade O(n), onde n é o número de padrões. Utiliza o conceito de centroides como forma de representação dos grupos, onde o centroide equivale-se ao centro do grupo e é calculado pela média de todos os objetos do grupo.

Tal algoritmo é definido, segundo Hair *et al.*, 2005, como um método de agrupamento não hierárquico por repartição e sua ideia principal é calcular pontos que representem os "centros" de um número de *clusters* designado inicialmente. Os centros desses *clusters* são

espalhados homogeneamente sobre o conjunto de dados e movidos segundo a aplicação de uma heurística até alcançar um equilíbrio aceitável. Os centroides iniciais são formados através da inclusão iterativa de novos casos de teste ao *cluster* cujo seu centro esteja mais próximo. A média se altera, com a inclusão de cada caso, por consequência, altera o centroide. Este processo continua até que não haja mais alterações nas médias ou que se chegue a um número determinado de iterações.

Sendo assim, o critério de agrupamento do k-means pode ser definido como o seguinte:

$$E = \sum_{k=1}^{K} \sum_{x_i \in C_k} d(x_i, x_{0k}) \tag{3}$$

onde x_{0k} é o centroide do *cluster* C_k e $d(x_i, x_{0k})$ é a distância entre os pontos x_i e x_{0k}. O centroide pode ser a média ou a mediana de um grupo de pontos. Em outras palavras, o objetivo do k-means é minimizar a distância entre cada ponto e o seu respectivo centroide (HAIR *et. al.*, 2005).

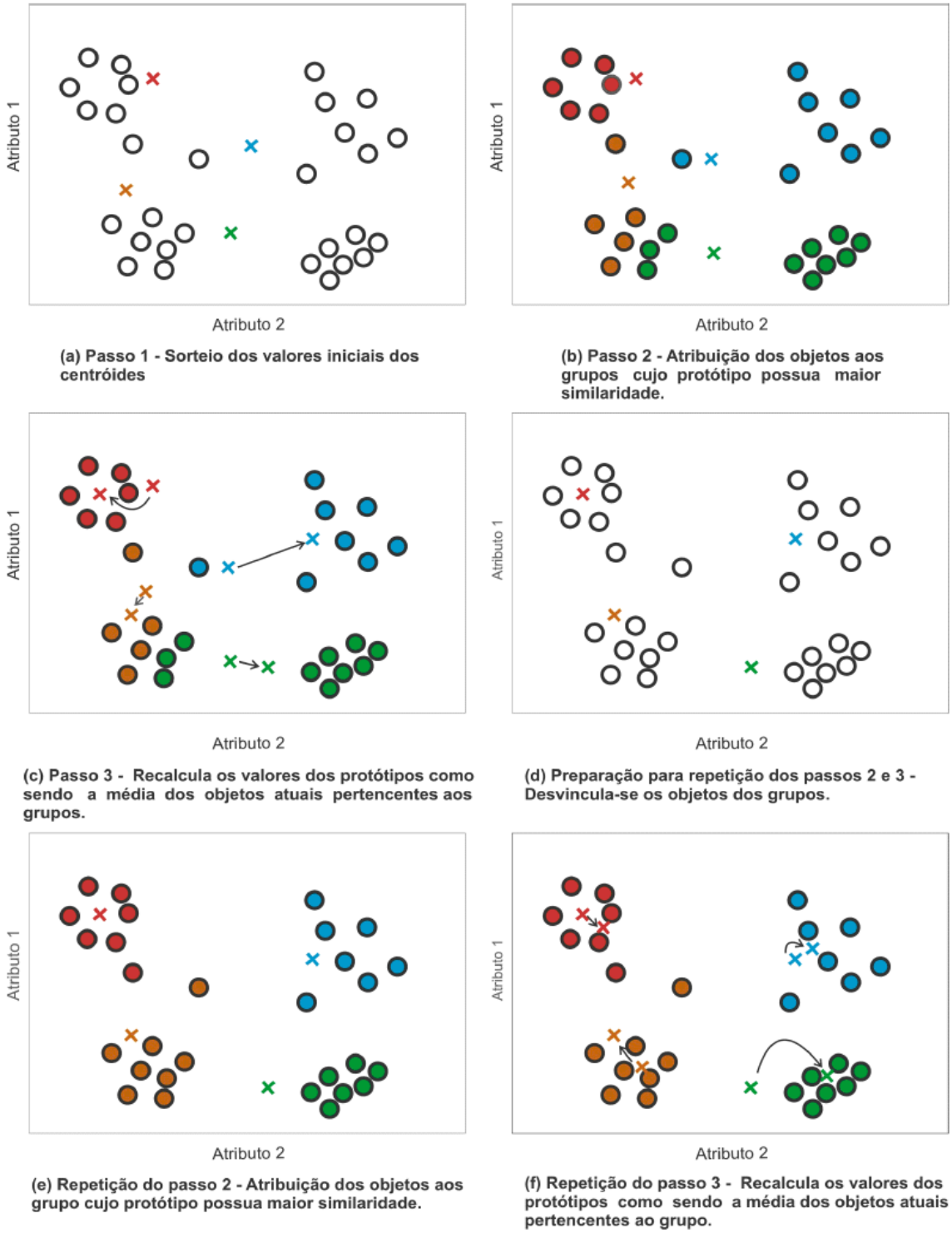
A descrição do algoritmo do k-means pode ser elucidada em 4 passos, de acordo com Fontana e Naldi (2009).

1. Atribuem-se valores iniciais para os protótipos seguindo algum critério, por exemplo, sorteio aleatório desses valores dentro dos limites de domínio de cada atributo;
2. Atribui-se cada objeto ao grupo cujo protótipo possua maior similaridade com o objeto;
3. Recalcula-se o valor do centroide (protótipo) de cada grupo, como sendo a média dos objetos atuais do grupo;
4. Repete-se os passos 2 e 3 até que os grupos se estabilizem;

A figura 1 ilustra a execução do algoritmo k-means. O passo 1 do algoritmo e a inicialização dos protótipos é verificado em 1 (a); em 1 (b) executa-se o passo 2, atribuição de cada objeto ao grupo cujo protótipo possuir uma maior similaridade ao objeto; em 1 (c) executa-se o passo 3 com o recálculo dos valores dos protótipos, como sendo a média dos objetos atualmente presentes em cada grupo; em 1 (d) prepara-se para repetição dos passos 2 e

3, desvinculando os objetos dos grupos aos quais pertenciam; finalmente, em 1 (e) e 1 (f) ocorre a repetição dos passos 2 e 3, respectivamente.

Os passos 2 e 3 do algoritmo repetem enquanto houver alterações nos grupos, ou seja, as médias variam ou enquanto não se atende o número de iterações propostas.



Fonte: Fontana e Naldi, 2009, p. 22.

FIGURA 2 – Exemplo de execução do algoritmo k-means

4. ARQUITETURA DE IMPLEMENTAÇÃO E TESTES

Neste capítulo serão previamente apresentadas: a arquitetura do projeto em geral, suas características adotadas neste trabalho, bem como o ambiente de testes e os testes realizados e, por fim, apresentam-se os resultados obtidos e os conhecimentos adquiridos nos testes.

4.1. Arquitetura

Devido a complexidade na estruturação de um sistema de prontuário eletrônico por possuir armazenagem de um grande volume de dados vindos de vários setores distintos dentro e fora de uma instituição de saúde, ainda precisa ser pedagógico, informativo e dar subsídios para a colaboração e tomada de decisão médica.

Para a extração de informações pertinentes aos testes de clusterização e para gravação dos resultados das regras obtidas através dos *clusters* encontrados, como também para padronizar as informações contidas na base de dados, elaborou-se um modelo arquitetural que gerenciou este trabalho.

No modelo da arquitetura, extraem-se os dados de uma base de dados médica e realizam-se as criações de documentos de marcação com semântica aplicada. Esses documentos são previamente validados, para posteriormente serem utilizados como formatação do prontuário com o qual se obtém informações de acompanhamento clínico, com a finalidade de criação de modelos ou regras que sirvam como apoio aos profissionais da área de saúde na tomada de decisão e colaboração.

A figura 4 descreve a ilustração da arquitetura deste projeto, demonstrando a extração, e tratamento dos dados da base, a geração do padrão em XML com aplicação de DTD e a criação de regras de conhecimento a partir da clusterização ou agrupamento. O modelo de reconhecimento de padrões agrupará os dados e dará subsídios para interpretação e casamento das informações obtidas.

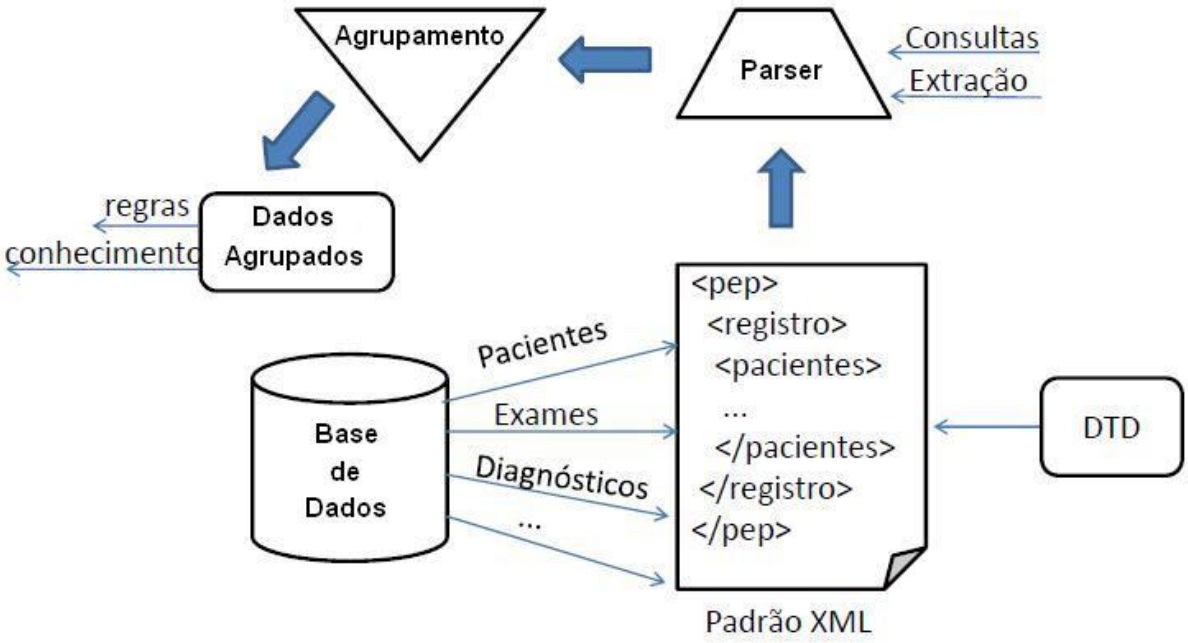


FIGURA 3 – Desenho arquitetural do projeto

Como é observado, extraem-se registros das tabelas do banco de dados médico, os dados textuais são adequados e distribuídos em *tags* XML. O documento XML é validado segundo uma DTD que apresenta quais dados são elementares para a formação do registro do paciente. Após a criação dos registros devidamente organizados segundo a DTD, é feita a manipulação dos mesmos através do *parser*[6], que analisa a gramática aplicada no XML e onde também é possível realizar consultas no documento. Após a extração dos dados textuais do prontuário, realiza-se seleção dos atributos que formarão a coleção de dados para os testes de agrupamento realizados com uma ferramenta de *data mining* buscando encontrar padrões significativos à colaboração médica.

Os resultados dos agrupamentos devem ser registrados no próprio documento XML que gerencia os dados do PEP de forma que apresente graficamente as regras descobertas, facilitando ao médico ou profissional de saúde a obter subsídios necessários ao comportamento colaborativo com outros profissionais e até mesmo descobrindo características importantes para um melhor desempenho de suas habilidades e de seus conhecimentos podendo compartilhá-los com outras pessoas de forma atrativa, sempre buscando uma melhor qualidade de vida para a população.

[6] Também conhecida como Análise Sintática: processo de analisar uma sequência de entrada de arquivo ou teclado e determinar sua estrutura gramatical segundo uma gramática formal pré-determinada.

4.2. Decisões de projeto

Nesta seção serão apresentadas as questões adotadas para o projeto dos testes realizados, o cenário de aplicação que foi destacado para os casos de teste, os dados selecionados para análise, a descrição das ferramentas utilizadas e o ambiente de implementação e testes realizados.

Com os dados devidamente estruturados e coletados no documento XML, faz-se a análise de agrupamento dos mesmos com o auxílio de uma ferramenta de mineração de dados a de algoritmos de classificação não supervisionada de maneira que sejam apresentadas ao usuário regras relevantes sobre a situação de saúde dos pacientes ou encontre padrões significativos para os dados retirados da base.

4.2.1. Ambiente e cenário de análise

Para analisar o comportamento dos dados após o agrupamento dos mesmos, foram destacados de uma base de dados médica informações sobre o código do diagnóstico e o medicamento indicado a cada paciente cadastrado. A base de dados utilizada foi concedida pelo Ambulatório Clínico de Ensino da Unifesp (2009) e é composta de 56 tabelas, com cerca de 10900 registros de pacientes. Estes dados estavam apenas registrados no SGBD, suas tabelas não estavam estruturadas e a base de dados não estava normalizada.

Para a construção do cenário de análise foi necessário destacar algumas tabelas da base de dados, as tabelas “Diagnósticos”, “Exames”, “Medicamentos” e “Histórias Clínicas” foram selecionadas para a retirada dos atributos para a análise de *clusters*. Destas tabelas, por sua vez, retiraram-se os seguintes atributos: os códigos CID – Classificação Internacional de Doenças adotado e publicado pela Organização Mundial de Saúde (OMS) e é globalmente utilizado para classificar doenças, sintomas e uma grande variedade de sinais e queixas de pacientes e a descrição dos medicamentos atribuídos para cada paciente cadastrado.

4.2.2. Ferramentas de mineração de dados e KDD

O presente trabalho utilizou ferramentas que auxiliaram no processo de produção, análise e mineração dos dados através de agrupamento dos mesmos em *clusters*.

Buscando encontrar uma ferramenta de *data mining*, foram analisados vários sistemas existentes no mercado e optou-se em adotar um sistema gratuito que atendesse os requisitos

da pesquisa. Os sistemas de mineração de dados disponíveis e que foram objetos comparativos nessa pesquisa são: *Oracle Data Mining*, *WizRule*, *Intelligent Miner for Data*, *Tanagra*, WEKA, dentre outros.

Oracle *Data Mining* (ODM): ferramenta integrada ao ambiente que gerencia o banco de dados da Oracle, onde ocorrem todos os processos de descoberta de conhecimento. Possuindo uma plataforma simples, segura e integrada. Características que destacam essa ferramenta em relação as demais, pois não há necessidade de extração dos dados que serão processados. Outras informações podem ser encontradas no *site* da Oracle (www.oracle.com).

WizRule: software desenvolvido com o objetivo de analisar e descrever grupos de dados, identificando possíveis erros nos mesmos. Revela todas as regras que modelam a base de dados e indica os casos de desvios. Criado pela empresa *WizSoft*, pode ser adquirido através de *download* a partir do *site* da empresa na *internet* (www.wizsoft.com).

Intelligent Miner for Data: também conhecida como *DB2 Intelligent Miner for Data*, esta ferramenta foi criada pela IBM, independente dos sistemas IBM, pode ser utilizada em plataformas Windows e rodar com outros SGBDs relacionais. O pacote do *DB2* combina o uso de algoritmos de mineração de dados com o objetivo de resolver problemas de KDD. Demais informações sobre a ferramenta podem ser encontradas no site da IBM (www.ibm.com).

Tanagra: Software livre de *Data Mining* utilizado para fins acadêmicos e de pesquisas científicas. O Tanagra propõem vários métodos de mineração a partir da análise exploratória dos dados, aprendizagem estatística e de máquina. Projeto sucessor do SIPINA e implementa vários algoritmos de aprendizado supervisionado, *clustering* e análises estatísticas paramétricas e não paramétricas tendo uma construção iterativa e possui elementos de visualização gráfica, como as árvores de decisão. Projeto vinculado a Universidade Lion, França e criado pelo professor Ricco Rakotomalala. Mais informações podem ser encontradas na página do projeto Tanagra (eric.univ-lyon2.fr/~ricco/tanagra).

WEKA: Ferramenta de KDD que implementa algoritmos para preparação, mineração e validação de dados. Permite a visualização gráfica dos resultados através de árvores de decisão e diagramas de dispersão. É um software de código aberto e possui algoritmos de clusterização e classificação de dados, também apresenta modelos para criação de redes neurais e bayesianas. Maiores informações no site do WEKA (www.cs.waikato.ac.nz).

A seguir, a tabela 1 mostra o comparativo sintetizando as ferramentas de *data mining* analisadas, destacando os métodos que cada uma implementa, as plataformas que dão suporte, bem como a sua disponibilidade comercial.

TABELA 1

Comparativo entre ferramentas de *data mining*

Ferramenta	Métodos de KDD	Plataformas	Pago
Oracle Data Mining	Classificação, Regressão, Associação, Clusterização e Mineração de Textos	Windows	Não
WizRule	Sumarização, Classificação, Detecção de Desvios	Windows	Apenas demo
Intelligent Miner for Data	Classificação, Regras de Associação, Sequenciais, Clusterização, Sumarização	Windows e Linux	Não
Tanagra	Classificação, Regras de Associação, Regressão, Clusterização	Windows	Sim
WEKA	Classificação, Regressão, Regras de Associação, Clusterização	Windows, Linux e Macintosh	Sim

Tal análise teve como objetivo a escolha de uma ferramenta que possuísse a disponibilidade gratuita de sua utilização e fosse possível utilizar suas implementações em outra aplicação ou sistema. Baseado na análise das ferramentas e foram descartadas aquelas cujos códigos-fonte não estavam disponíveis e era possível apenas realizar testes práticos de mineração através da utilização da própria aplicação.

Para o desenvolvimento desta pesquisa utilizou-se a ferramenta de mineração de dados WEKA, devido a facilidade de compreensão e organização de seu código-fonte através da API da ferramenta disponibilizada na web.

4.2.3. WEKA

Wakaito Environment for Knowledge Alalysis (WEKA) é uma ferramenta de código aberto composta por uma coleção de algoritmos de aprendizado de máquina para tarefas de *data mining.* Estruturado e implementado na linguagem de programação Java pelo grupo de pesquisadores em Aprendizado de Máquina do curso de Ciência da Computação da Universidade de Wakaito na Nova Zelândia. Possui implementada atualmente as seguintes tarefas e métodos ou algoritmos de mineração.

a) **Tarefas:** pré-processamento de dados, aplicação de filtros em atributos e instâncias, classificação, clusterização, associação, seleção de atributos e visualização de dados e resultados de análises.

b) **Métodos ou classes de algoritmos:** bayes, functions, lazzy, meta, mi, misc, rules, trees, cobweb, DBscan, EM, fathestfirst, hieralquical clusterer, make density based clusterer, simple k-means, a priori, filtered associator, entre outros.

O WEKA pode ser utilizado de várias maneiras e possui interfaces de usuário bastante amigáveis e interativas. A seguir descreve-se as quatro interfaces implementadas na versão 3.6.5 do software:

a) **Explorer:** interface mais simples que compõe todos os métodos e tarefas que são aplicadas nas bases de dados.

b) **Experimenter:** interface mais robusta que aplica um ou vários métodos de classificação em diferentes conjuntos de dados, podendo realizar comparações estatísticas sobre os esquemas criados.

c) **KnowledgeFlow:** interface alternativa ao *explorer*, porém tem a representação de forma gráfica e totalmente interativa, onde pode-se selecionar os componentes do WEKA e arrastá-los a uma tela de modelagem, conectando-os entre si de modo a formar um fluxo de conhecimento para processamento e análise dos dados.

d) **Simple Client (CLI):** interface baseada em linha de comando. Possui uma aparência simples, porém completa, onde o usuário é capaz de executar qualquer operação do software através da digitação de comandos.

Por sua vez, Morate (2004) elucida que a WEKA suporta os seguintes atributos:

e) **numeric:** representa números reais.

f) **integer:** representa números inteiros.

g) **date:** representa unidades de tempo (dd Dia; MM Mês; yyyy Ano, HH Horas; mm Minutos; ss Segundos).

h) **string:** representa uma cadeia de textos. Não é utilizado no processo de *data mining*, mas sim como identificador das instâncias.

i) **enumerado:** consistem em representar entre chaves, separando por vírgula, valores que podem ser tomados como atributos, como por exemplo: *@attribute periodo {M,V,N,I}.*

O WEKA suporta arquivos do tipo ARFF, CSV e C45 como entrada de dados, mas só consegue executar o processo de mineração a partir do arquivo *Attribute-Relation File Format* ARFF.

Tal sistema foi escolhido para ser utilizado nessa pesquisa pelo fato de atender as necessidades de implementação de algoritmos de mineração de dados e análise de *cluster*, permitindo o uso de seu código e de sua API Java em novas tendências de implementações de sistemas de prontuário com análise dos seus próprios dados, por exemplo, além de ser totalmente disponível para acesso comum.

4.2.4. Características dos testes

Os testes do projeto foram realizados levando em consideração algumas restrições encontradas durante o processo de engenharia. Para isso foram feitas algumas restrições dos atributos que seriam analisados, devido a limitações de software e hardware encontradas

durante os testes. As limitações foram em relação ao tamanho da pilha *heap* de alocação de memória na máquina virtual *Java* na qual executa o WEKA. Para o sistema no qual os casos de testes foram realizados, o máximo de memória aceitável foi de 512 MB.

Portanto, os testes foram elaborados apenas com os atributos: CID de diagnósticos e descrição de medicamentos de cada paciente registrado na base. Foram extraídas e analisadas com o WEKA 3264 instâncias com 1552 atributos.

Todos os testes foram realizados em plataforma *Windows* e a restrição do número de instâncias se deu devido insuficiência de memória para uso no WEKA. A seguir demonstram-se os testes realizados e seus respectivos resultados.

4.3. Testes

Nesta seção serão descritos os testes de agrupamento realizados com a ferramenta de mineração WEKA, versão 3.6.5. Aqui será abordado o algoritmo *Simple K-means*, implementação do algoritmo k-means, previamente descrito na seção 1.2.4.

4.3.1. Preparação dos dados

Para descrever a utilização do software usado no agrupamento dos dados, utilizou-se a interface *explorer* do sistema, onde foram carregados os dados extraídos para análise. Como os dados extraídos são classificados como dados categóricos[7], pois possuem informações descritivas, necessitou-se a filtragem das informações porque o WEKA não realiza agrupamento de dados categóricos, apenas com dados numéricos.

A filtragem dos dados foi realizada pela técnica *NominalToBinary*[8], que consistem em transformar os dados nominais em dados de ordem binária, necessários para a execução do agrupamento com o algoritmo k-means.

A figura 5 mostra a tela inicial da interface *explorer* do WEKA com os dados previamente filtrados carregados e prontos para análise.

[7] Conjunto de dados cuja escala de classificação é nominal ou uma *string* de caracteres
[8] Classe descrita na API do WEKA disponível em: <http://weka.sourceforge.net/doc.stable/>

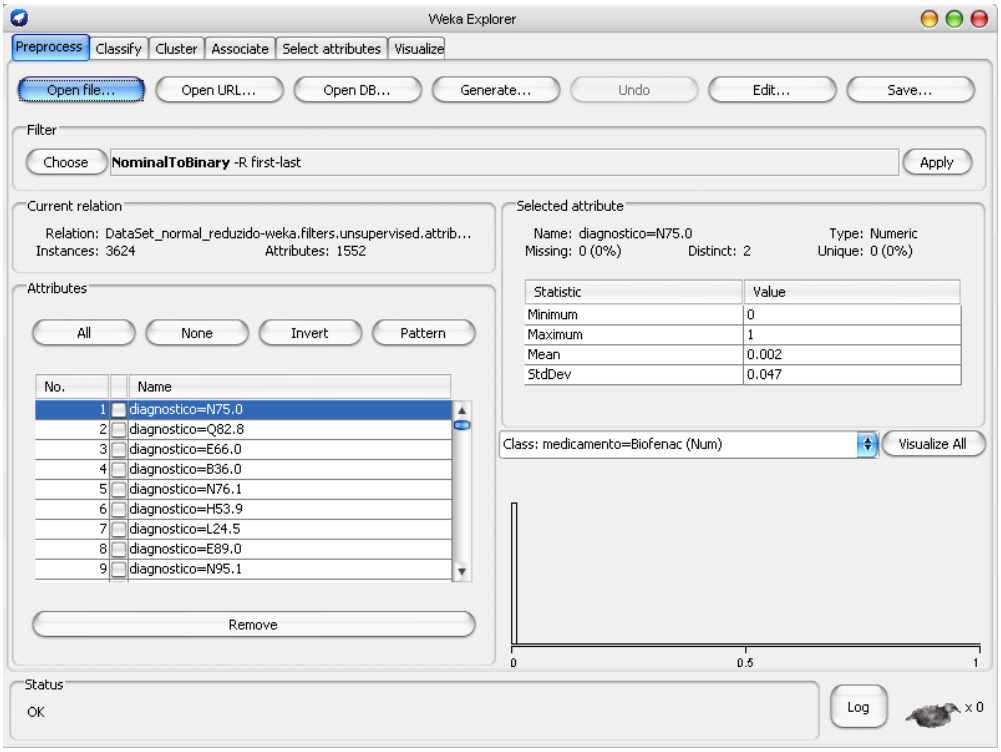


FIGURA 4 – Tela do WEKA: Filtragem dos dados

Os dados “binarizados” apresentam um valor máximo e outro mínimo, bem como a média e o desvio padrão de todas as ocorrências dessa instância (atributo) no arquivo de teste.

Após a preparação dos dados, realiza-se o processo de clusterização, explanado na próxima seção.

4.3.2. Clusterização

Nessa etapa, os dados passam pelo processo de clusterização, onde o algoritmo k-means é aplicado para geração de *clusters* e descoberta de conhecimento não supervisionado de informações referentes aos diagnósticos e medicamentos encontrados na base de dados.

Para esta etapa, foram elaboradas quatro casos de teste de acordo com o número de grupos (*clusters*) que se pretendia atender. Foram estipulados os valores de: 2, 10, 20 e 30 *clusters* e cada um deles foi analisado separadamente.

A figura 6 ilustra a seleção do algoritmo de clusterização implementado no WEKA *SimpleKMeans*, bem como a seleção de seus parâmetros, como: a distância utilizada (Euclidiana), o número máximo de iterações e o número de *clusters* desejado, neste caso, são criados 2 *clusters*.

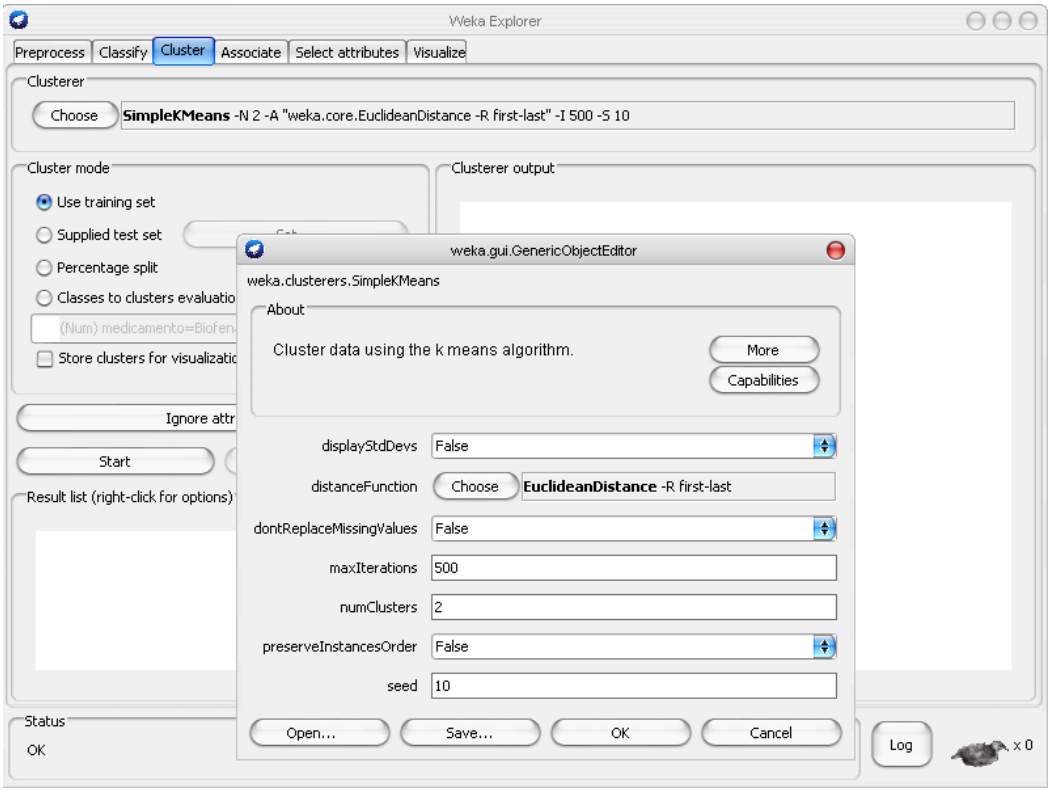


FIGURA 5 – Tela do WEKA: Seleção e atributos do algoritmo *SimpleKMeans*

Após configurar os atributos e parâmetros de execução do algoritmo, se dá início ao processo de análise de *cluster*. A sessão seguinte irá apresentar os resultados e as informações adquiridas com a aplicação dos 4 casos de testes realizados.

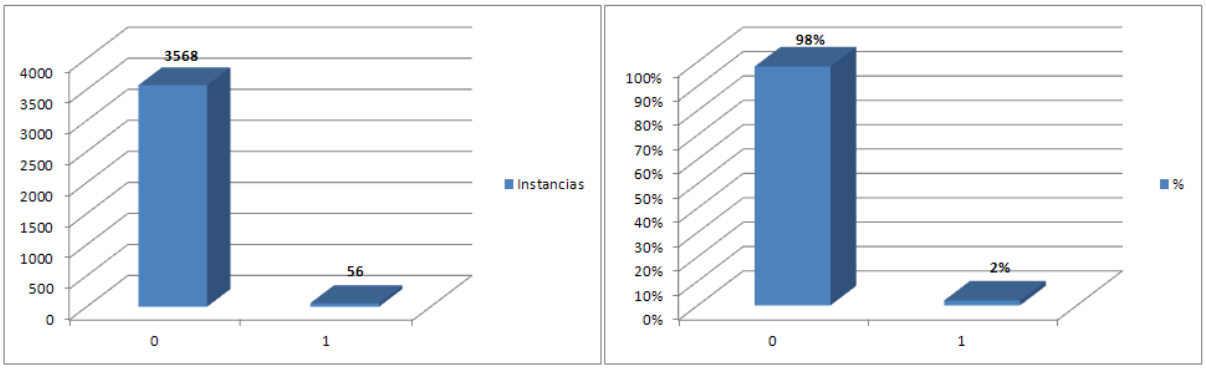
4.3.3. Análise dos resultados e geração de regras

Realizadas as etapas de preparação e análise dos dados, exploram-se os resultados obtidos de modo a encontrar similaridades e informações sobre os conjuntos gerados.

As tabelas 2, 3, 4 e 5 (Apêndice C) representam os resultados alocados em 2, 10, 20 e 30 *clusters*, respectivamente, identificando o número de instâncias agrupadas em cada *cluster* e a porcentagem de elementos que cada grupo recebeu.

As tabelas foram geradas em conformidade com os resultados encontrados a partir dos resultados extraídos da interface do WEKA (Apêndice D), que calcula os centroides para cada um dos atributos (CIDs e medicamentos) distribuindo-os entre os *clusters*. Os centroides cujos valores se destacam entre os demais são expostos na análise, bem como os atributos correspondentes a eles.

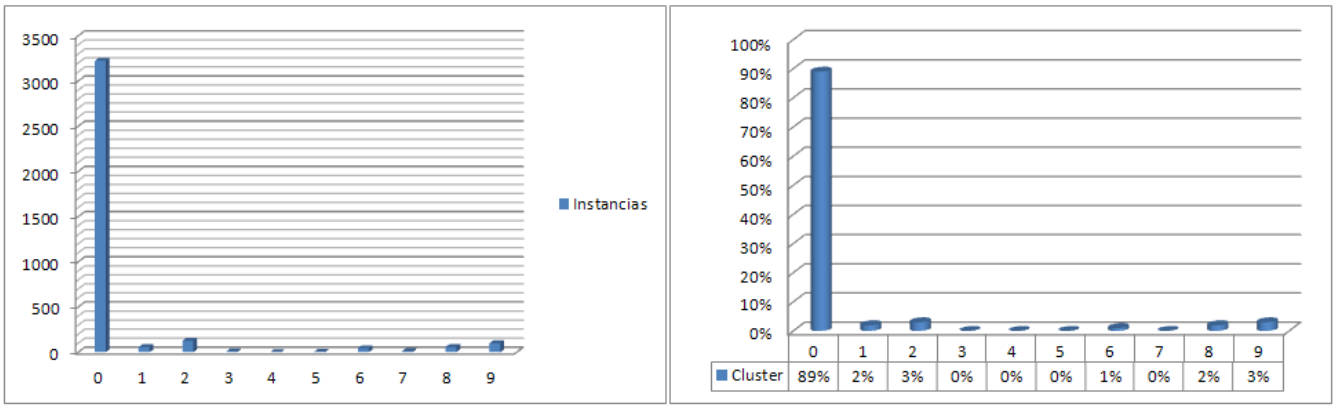
As representações gráficas de cada caso de teste de acordo com o número de *clusters* alocados podem ser analisadas nas figuras 6, 7, 8 e 9, respectivamente. Os gráficos determinam a alocação de instâncias para cada cluster e a porcentagem de atributos que cada cluster recebeu.



(a) número de instâncias alocadas por cluster

(b) porcentagem de alocação dos clusters

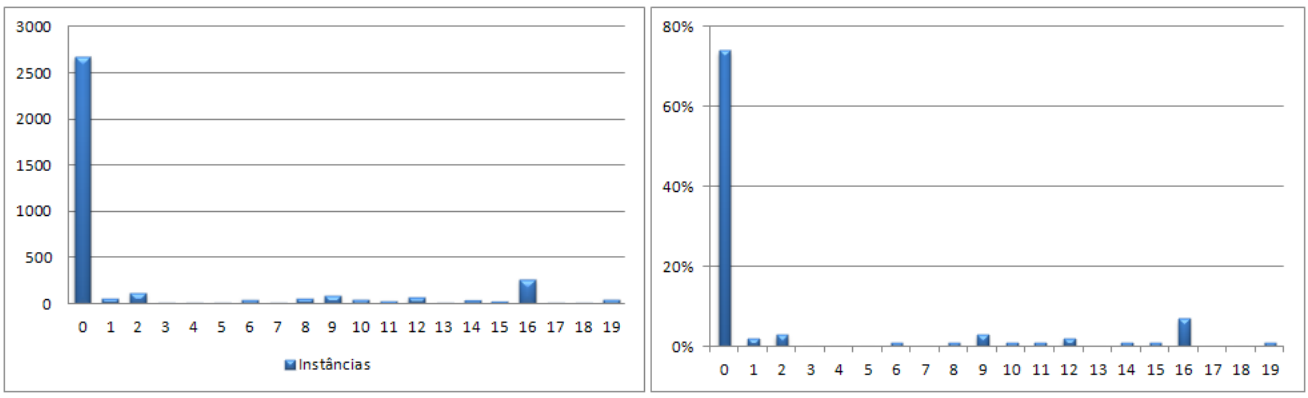
FIGURA 6 – Gráfico de resultados da clusterização em 2 clusters



(a) número de instâncias alocadas por cluster

(b) porcentagem de alocação dos clusters

FIGURA 7 – Gráfico de resultados da clusterização em 20 clusters



(a) número de instâncias alocadas por cluster

(b) porcentagem de alocação dos clusters

FIGURA 8 – Gráfico de resultados da clusterização em 20 clusters

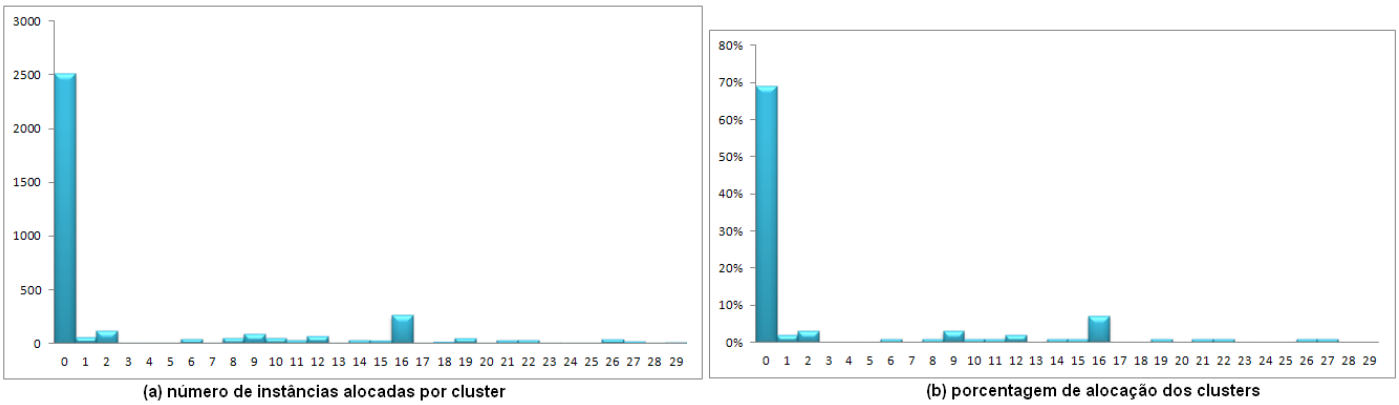


FIGURA 9 – Gráfico de resultados da clusterização em 30 clusters

Com base nos resultados obtidos, pode verificar que quanto maior for o número de *clusters*, maior será a distribuição dos centroides pelos pontos mais extremos da análise. Demonstrou-se também que o primeiro *cluster* engloba a maior concentração de instâncias para todos os casos testados, pois o primeiro *cluster* contém um grupo de pacientes que possuem um diagnóstico e um medicamento diretamente associado, o que ocorre em quase todos os casos. Essa associação se dá devido ao princípio de escolha dos centroides iniciais no WEKA que são feitos de maneira aleatória e tendo que os atributos são, em sua maioria, distintos, foram alocados inicialmente no primeiro *cluster* e depois convergidos aos demais.

Também é possível verificar que o *cluster* 2 apresenta, em todos os casos o mesmo valor de atributos, este grupo engloba as pessoas que foram diagnosticadas com doenças do sistema nervoso e do coração, tendo como centroides mais significativos os diagnósticos cujos códigos são I10 e F41.2, o primeiro se refere a “Hipertensão” e o segundo a “Transtorno Misto Ansioso e Depressivo”. O *cluster* 9 agrupa medicamentos receitados e possui uma relação com o grupo de CIDs do *cluster* 2, porque seus maiores centroides apontam para medicamentos para tratamento de depressões ou hipertensão, casos como os antidepressivos “Tryptanol” e “Citalopram” e calmantes como “Stilnox”.

O *cluster* 16, nos casos de teste com 20 e 30 *clusters* também apresenta uma divisão representativa nos centroides de medicamentos, sendo que os medicamentos mais representativos são: “AAS”, “Buscopan” e “Dipirona”, ou seja, um grupo com analgésicos, antitérmicos e anti-inflamatórios.

Os *clusters* 6 e 12 agrupam, nos casos de teste com mais de 10 instâncias, diagnósticos relevantes de diabetes e doenças do trato digestivo e metabólicas. Dentre essas se destacam: “Hipergliceridemia” e “Dispepsia”. Também alguns casos de hipotireoidismo e anemia.

Os demais *clusters* obtiveram uma porcentagem baixa e não relevante para análise, portanto, contém grupos de diagnósticos e medicamentos pouco utilizados na base.

5. CONSIDERAÇÕES FINAIS

Na área de atuação médica, onde vários profissionais necessitam da informação sobre pacientes, enfermidades, diagnósticos, tratamentos, medicamentos, etc. Dados que são altamente importantes para a manutenção da saúde do indivíduo. Para apresentação desses dados, torna-se indispensável o uso de sistemas de Prontuário Eletrônico do Paciente que contenha um grande conjunto de dados que seja, ao mesmo tempo, organizado, com informações relevantes para uso nas mais variadas instituições de saúde, adotando um padrão para apresentação dessas informações de modo que as mesmas sejam incorporadas em um sistema PEP e que possam contribuir para a colaboração e tomada de decisão médica.

O uso das Tecnologias de Informação e Comunicação no processo de mineração de dados, mais especificamente na descoberta de conhecimento em bases de dados é cada vez mais importante no que diz respeito a metodologia exploratória de dados presentes em um banco de dados.

Tal importância se dá devido ao grande crescimento na forma de como o mundo moderno trata da informação atualmente. Grandes volumes de dados passam diariamente por constantes atualizações e verificações, sendo sempre requisitados e devendo transmitir rapidamente conhecimento ao seu consumidor, obtendo as informações necessárias para que alcance sucesso em sua busca.

Nesta pesquisa também foi abordada a relevância do processo de clusterização em dados categóricos na descoberta de conhecimento médico no que se refere a avaliação dos resultados obtidos com a implementação de casos de teste realizados através do algoritmo k-means.

A partir dos resultados obtidos com a clusterização dos dados de diagnósticos e medicamentos extraídos da base XML, permitiram gerar juízo sobre algumas características e particularidades importantes que só foram descobertas através do uso da análise dos resultados dos centroides criados pela execução do algoritmo k-means. Percebeu-se, também, que pode-se extrair conhecimentos sobre diagnósticos e medicamentos, apontando quais os grupos de pacientes estão com determinado diagnóstico e receberam a prescrição de determinado medicamento, bem como a porcentagem de alocação desse grupo.

Essas informações, quando arquivadas e representadas em um sistema de prontuário eletrônico pode contribuir claramente com a tomada de decisão médica e, também na colaboração entre eles, tendo que facilitaria a comunicação e visualização de evoluções de

enfermidades sindrômicas ou crônicas onde um profissional pode não ser capaz de cuidar de todos os aspectos e sintomas do paciente, por exemplo.

Mostrando as relações entre os atributos médicos e de saúde do paciente obtidas através do processo de clusterização será possível, através da análise gráfica desses resultados, por exemplo, criar situações nas quais profissionais da área médica possam facilmente analisar e compartilhar os conhecimentos obtidos em prol da manutenção da saúde e do bem estar da população em geral.

5.1. Proposta de trabalhos futuros

A fim de dar continuidade à pesquisa, alguns trabalhos podem ser realizados, como a aplicação de testes com outros algoritmos de clusterização, gerando, futuramente um estudo comparativo.

Pretende-se, também, a criação de um sistema de Prontuário Eletrônico do Paciente, que atenda aos requisitos básicos que um prontuário deve conter, e, ainda possa utilizar-se das clusterizações, elaborando modelos gráficos com as regras de conhecimento adquiridas através da mineração feita pelos algoritmos da ferramenta WEKA.

Outra proposta seria armazenamento ou registro das informações obtidas com as clusterizações e a criação gráfica das relações criadas, como, por exemplo, do medicamento diurético propranolol com o diagnóstico de hipertensão. Tal relação pode ser feita através de grafos, pontos, gráficos, entre outros.

6. REFERÊNCIAS

DICK, R. S, STEEN, E. B, DETMER, D. E. Edts. **The Computer Based Patient Record – An essential Technology for Health Care**. Committee on Improving the Patient Record. Institute of Medicine. National Academy Press, Washington DC, 1997.

HALL, M, FRANK, E, HOLMES, G, PFAHRINGER, B, REUTEMANN, P, Ian H. WITTEN, I. H. **The WEKA Data Mining Software: An Update**. SIGKDD Explorations, Volume 11, Issue 1. (2009).

BARSOTTINI, C. N, WAINER, J. **Patterns of Collaboration and Non-collaboration Among Physicians**. H. Fuks, S. Lukosch, and A.C. Salgado (Eds.): CRIWG 2005, LNCS 3706, pp. 248–254, Springer, 2005.

SIGULEM, D, ANÇÃO, M. S, RAMOS,M. P, LEÃO, B. F. **Sistema de Apoio à Decisão em Medicina**. Atualização Terapêutica - Manual Prático de Diagnóstico e Tratamento", 1998. Disponível em: <www.virtual.epm.br/material/tis/curr-med/sad_html/sistema>. Acesso em: 19 out. 2011.

INSTITUTE OF MEDICINE. Committee on Improving the Patient Record. **The Computer-Based Patient-Record: An Essential Technology for Health Care** (2 nd Edition). Washington DC: National Academy Press, 1997.

RAKOTOMALALA, R. **TANAGRA : un logiciel gratuit pour l'enseignement et la recherche**, in Actes de EGC, 2005, RNTI-E-3, vol. 2, pp.697-702, 2005.

MASSAD, Eduardo et. al. **Prontuário Eletrônico do Paciente na Assistência, Informação e Conhecimento Médico**. Faculdade de Medicina da Universidade Federal de São Paulo. 2003. Disponível em: <www.sbis.org.br/site/arquivos/prontuario.pdf>. Acesso em: 10 ago. 2011.

MURPHY, G.F., HANKEN, M.A., WATERS, K.A. **Electronic Health Records**: Changing the Vision. Philadelphia : W.B. Saunders Company, 1999.

KARYPIS, G. **CLUTO – A Clustering Toolkit**. Department of Computer Science and Engineering. University of Minnesota. Minneapolis, 2002.

LAPEYRE, D. A. **Introduction to XML Schema Languages**. Mulberry Technologies, Inc. 2005. Disponível em: <http://www.xmlphilly.org/July-schema-talk.pdf>. Acesso em: 16 dez. 2011.

JAIN, A. K, MURTY, M. N, FLYNN, P. J. **Data Clustering: A Review**. The Ohio State University. In: ACM Computing Surveys, Vol. 31, No. 3, September 1999.

HASTIE, T, TIBSHIRANI, R, FRIEDMAN, J. **The elements of statistical learning - Data Mining, inference and prediction**. Ed. Springer. 2001.

MITCHELL, T. **Machine learning**. Ed. Mc Graw-Hill International Editions, 1997.

SCHWEIGER, R, BRUMHARD, M, HOELZER, S, DUDECK, J. **Implementing Health Care Systems Using XML Standards**. Institute for Medical Informatics, Justus-Liebig-University, Heinrich-Buff-Ring 44, 35392 Giessen, Alemanha. 2004.

HAIR, J. F. Jr.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise Multivariada de Dados**. 5 ed. Porto Alegre: Bookman, 2005.

NARDON, F. B, FURUIE, S, TACHINARDI, U. **Novas Tecnologias para Construção do Prontuário Eletrônico do Paciente**. Instituto do Coração do Hospital das Clínicas da Faculdade de Medicina da USP. São Paulo. 2000.

PETRY, K, LOPES, P. M. A. **Modelos para Interoperabilidade de Sistemas Hospitalares Utilizando o Padrão HL7**. Universidade Federal de Santa Catarina. 2005. Disponível em: <http://projetos.inf.ufsc.br/arquivos_projetos/projeto_377/Interoperabilidade%20de%20Sistemas%20Hospitalares%20Utilizando%20o%20Padr%E3o%20HL7.pdf>. Acesso em: 17 dez. 2011.

LESLIE, H. **OpenEHR – The World's Record**. PulseIT. Sydney. 2007. Disponível em: <http://www.openehr.org/301-OE.html>. Acesso em: 05 jan. 2012.

ZAVALIK,C. **Integração de Sistemas de Informação através de Web services**. UFRGS. Porto Alegre: Programa de Pós- Graduação em Computação, 2004.

GONÇALVES, L. W. **Prontuário Eletrônico do Paciente Adotando Padrões para a Telemedicina no Brasil**. Universidade Federal do Rio Grande do Sul. Porto Alegre. 2010. Disponível em: <http://www.lume.ufrgs.br/bitstream/handle/10183/26348/000757803.pdf?sequence=1>. Acesso em: 04 jan. 2012.

FONTANA, A., NALDI, M. C. **Estudo de Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados**. 2009. Universidade de São Paulo. ISSN - 0103-2569.

GRIRA, N, CRUCIANU, M. BOUJEMAA, N. **Unsupervised and Semi-supervised Clustering: a Brief Survey**. INRIA Rocquencourt, B.P. 105. 78153 Le Chesnay Cedex, France. 2005.

WITTEN, I, FRANK, E. **Data mining - Practical machine learning tools and Techniques with JAVA implementations**, Morgan Kaufmann Publishers, 2000.

MORATE. D G. **Manual de WEKA**. Versão em Espanhol. Disponível em: <http://metaemotion.com/diego.garcia.morate/download/weka.pdf>. Acesso em: 12 set 2011.

SCOSS, A. M. **A Clusterização e Classificação no Processo de *Data Mining* para Análise do Desempenho Docente no Ensino de Graduação**. Curso de Pós-Graduação Especialização em MBA Gerenciamento em Banco de Dados. Criciúma. 2006.

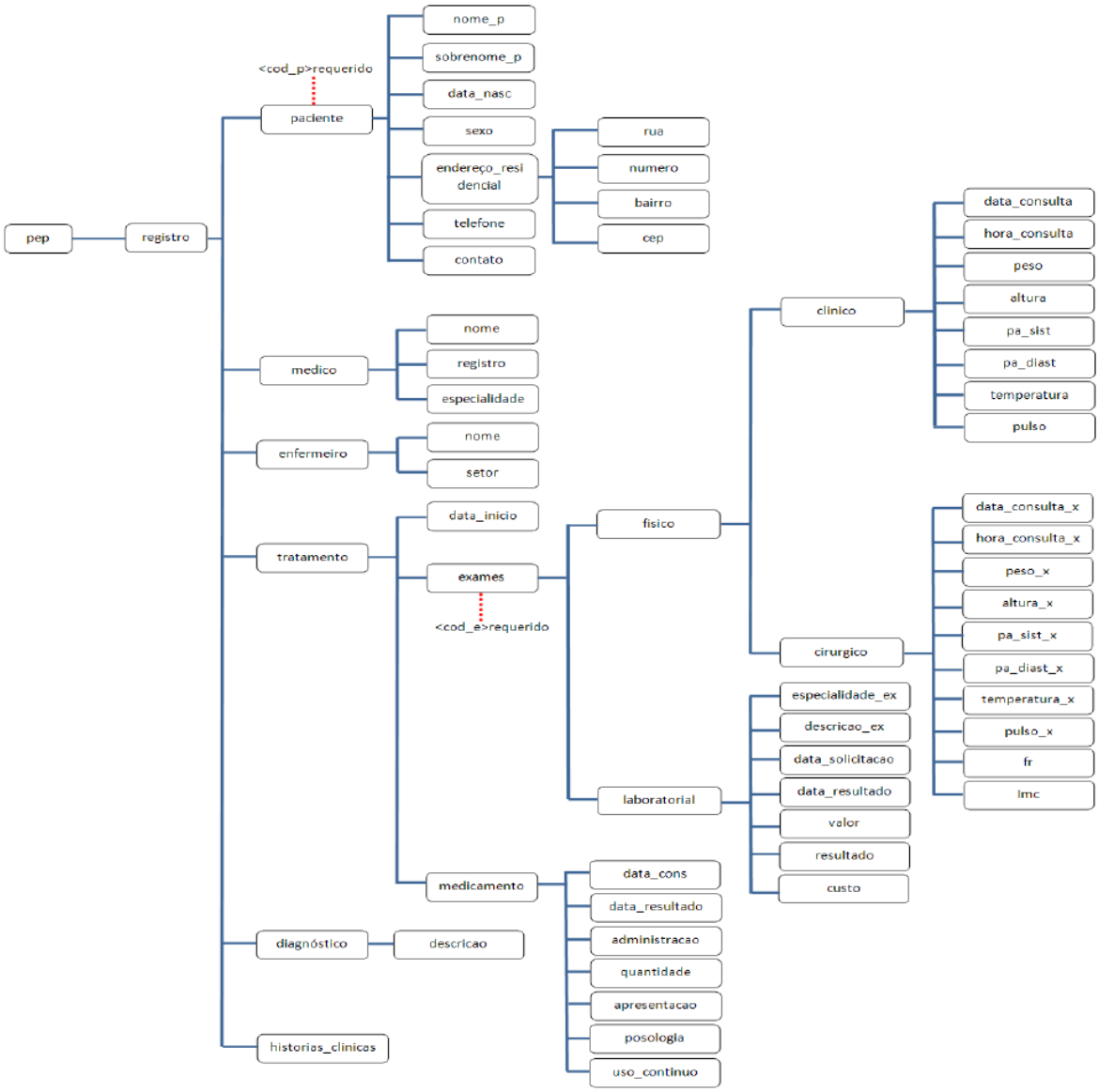
OCHI, L. S.; DIAS, C. R.; SOARES, S. S. F. **Clusterização em Mineração de Dados**. 2004. In: ERI RJ/ES - Escola Regional de Informática Rio de Janeiro - Espírito Santo - IV : 2004 nov. : Vitória - ES, Rio das Ostras.

DE ARAÚJO, C. R. L., MACIEL, C. P, MARQUES, D. C. **Manual Para Elaboração e Normalização de Trabalhos Acadêmicos – Conforme Normas da ABNT**. Sistema de Bibliotecas da Universidade Federal do Pampa. Alegrete. 2010.

LINDEN, R. **Técnicas de Agrupamento**. In: Revista de Sistemas de Informação da FSMA n. 4 (2009) pp. 18-36. Faculdade Salesiana Maria Auxiliadora. Rio de Janeiro. 2009.

BRASIL, 2004. Ministério da Saúde. Secretaria Executiva. Departamento de Informação e Informática do SUS. **Política Nacional de Informação e Informática em Saúde - PNIIS**. Inclui deliberações da 12ª Conferência Nacional de Saúde. Versão 2.0. Brasília, 29 de março de 2004. Disponível em: <http://www2.datasus.gov.br/DATASUS/APRESENTACAO/PoliticaInformacaoSaude29_03_2004.pdf> Acesso em: 13 de outubro de 2011.

APÊNDICE A – Estrutura DTD



APÊNDICE B – Padronização XML

```
<?xml version="1.0" encoding="UTF-8"?>
<dataroot xmlns:od="urn:schemas-microsoft-com:officedata" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="Diagnosticos.xsd"
generated="2010-11-10T15:33:48">
    <Diagnosticos>
        <codigo_paciente>8789</codigo_paciente>
        <nome_paciente>JOSE FAGUNDES</nome_paciente>
        <sobrenome_paciente>NEVES</sobrenome_paciente>
        <cid>I48</cid>
        <diagnostico>"Flutter" e fibrilação atrial</diagnostico>
        <situacao>Atual</situacao>
        <data_inicio>2006-06-21T00:00:00</data_inicio>
        <data_consulta>2006-06-21T00:00:00</data_consulta>
        <hora_consulta>08:34:00</hora_consulta>
        <registro_profissional>99910</registro_profissional>
        <nome_profissional>[A5-04]</nome_profissional>
        <sobrenome_profissional>Alessandra Stanquini Lopes</sobrenome_profissional>
        <profissao_profissional>Aluno</profissao_profissional>
        <especialidade>Clínica Médica</especialidade>
    </Diagnosticos>
    <Diagnosticos>
        <codigo_paciente>8898</codigo_paciente>
        <nome_paciente>CELIA MARIA</nome_paciente>
        <sobrenome_paciente>PERUSO</sobrenome_paciente>
        <cid>I48</cid>
        <diagnostico>"Flutter" e fibrilação atrial</diagnostico>
        <situacao>Atual</situacao>
        <data_inicio>2006-11-14T00:00:00</data_inicio>
        <data_consulta>2006-11-14T00:00:00</data_consulta>
        <hora_consulta>08:15:00</hora_consulta>
        <registro_profissional>99910</registro_profissional>
        <nome_profissional>[A5-04]</nome_profissional>
        <sobrenome_profissional>Alessandra Stanquini Lopes</sobrenome_profissional>
        <profissao_profissional>Aluno</profissao_profissional>
        <especialidade>Clínica Médica</especialidade>
```

APÊNDICE C – Tabelas de resultado das clusterizações

TABELA 2

Resultado da clusterização alocada em 2 *clusters*

Cluster	**Instâncias alocadas**	**Porcentagem do Grupo**
0	3568	98%
1	56	2%

TABELA 3

Resultado da clusterização alocada em 10 *clusters*

Cluster	**Instâncias alocadas**	**Porcentagem do Grupo**
0	3222	89%
1	56	2%
2	125	3%
3	8	0%
4	1	0%
5	4	0%
6	44	1%
7	10	0%
8	56	2%
9	98	3%

TABELA 4

Resultado da clusterização alocada em 20 *clusters*

Cluster	**Instâncias alocadas**	**Porcentagem do Grupo**
0	2678	74%
1	56	2%
2	120	3%
3	8	0%
4	1	0%

5	4	0%
6	44	1%
7	10	0%
8	54	1%
9	93	3%
10	50	1%
11	33	1%
12	66	2%
13	10	0%
14	36	1%
15	26	1%
16	264	7%
17	7	0%
18	16	0%
19	48	1%

TABELA 5

Resultado da clusterização alocada em 30 *clusters*

Cluster	**Instâncias alocadas**	**Porcentagem do Grupo**
0	2514	69%
1	56	2%
2	119	3%
3	8	0%
4	1	0%
5	4	0%
6	44	1%
7	10	0%
8	54	1%

9	91	3%
10	50	1%
11	33	1%
12	66	2%
13	10	0%
14	35	1%
15	26	1%
16	264	7%
17	7	0%
18	16	0%
19	48	1%
20	9	0%
21	31	1%
22	34	1%
23	2	0%
24	4	0%
25	5	0%
26	39	1%
27	21	1%
28	8	0%
29	15	0%

APÊNDICE D – Arquivo textual de resultado de clusterização no WEKA para 10 *clusters*

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 10 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     DataSet_normal_reduzido-weka.filters.unsupervised.attribute.NominalToBinary-A-Rfirst-last-weka.filters.unsupervised.attribute.NominalToBinary-A-
Instances:3624
Attributes:1552
[list of attributes omitted]
Test mode:evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 6782.12384940012
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data          0          1          2          3          4          5          6          7          8          9
                             (3624)     (3222)       (56)      (125)        (8)        (1)        (4)       (44)       (10)       (56)       (98)
=============================================================================================================================================
diagnostico=N75.0             0.0022     0.0025          0          0          0          0          0          0          0          0          0
diagnostico=Q82.8             0.0003     0.0003          0          0          0          0          0          0          0          0          0
diagnostico=E66.0             0.0174     0.0183     0.0179      0.008      0.125          0          0          0          0     0.0179          0
diagnostico=B36.0             0.0014     0.0016          0          0          0          0          0          0          0          0          0
diagnostico=N76.1             0.0091     0.0096     0.0179      0.008          0          0          0          0          0          0          0
diagnostico=H53.9             0.0006     0.0006          0          0          0          0          0          0          0          0          0
diagnostico=L24.5             0.0003     0.0003          0          0          0          0          0          0          0          0          0
diagnostico=E89.0             0.0017     0.0016          0          0          0          0       0.25          0          0          0          0
diagnostico=N95.1             0.0155     0.0164     0.0179          0      0.125          0          0          0          0     0.0179          0
diagnostico=M79.0             0.0063     0.0062     0.0179      0.008          0          0          0          0          0     0.0179          0
diagnostico=F33.9             0.0047     0.0053          0          0          0          0          0          0          0          0          0
diagnostico=E14.8             0.0003     0.0003          0          0          0          0          0          0          0          0          0
diagnostico=I10                0.077     0.0829     0.0714      0.032          0          0          0     0.0455          0     0.0357          0
diagnostico=R32               0.0044     0.0034          0      0.024          0          0          0          0          0     0.0357          0
diagnostico=Z72.0             0.0177      0.018     0.0357      0.016          0          0          0          0          0     0.0357          0
diagnostico=B82.9             0.0055     0.0056          0      0.016          0          0          0          0          0          0          0
diagnostico=F45.0             0.0033     0.0034          0          0          0          0          0          0          0     0.0179          0
diagnostico=K30               0.0295          0     0.0179      0.024          0          0          0     0.0682        0.1     0.0179          1
diagnostico=G99.0             0.0003     0.0003          0          0          0          0          0          0          0          0          0
diagnostico=J32.9             0.0033     0.0031     0.0179      0.008          0          0          0          0          0          0          0
diagnostico=I70.2             0.0025     0.0025          0      0.008          0          0          0          0          0          0          0
diagnostico=E11.5             0.0011     0.0012          0          0          0          0          0          0          0          0          0
diagnostico=I87.2              0.005     0.0056          0          0          0          0          0          0          0          0          0
diagnostico=Z00.6             0.0006     0.0006          0          0          0          0          0          0          0          0          0
diagnostico=F41.2             0.0105     0.0109     0.0536          0          0          0          0          0          0          0          0
diagnostico=A58               0.0003     0.0003          0          0          0          0          0          0          0          0          0
diagnostico=R20.8             0.0003     0.0003          0          0          0          0          0          0          0          0          0
diagnostico=F44.7             0.0014     0.0016          0          0          0          0          0          0          0          0          0
diagnostico=F44.6             0.0003     0.0003          0          0          0          0          0          0          0          0          0
diagnostico=Z92.0             0.0003     0.0003          0          0          0          0          0          0          0          0          0
```