UNIVERSIDADE ESTADUAL DO CEARÁ

CENTRO DE CIÊNCIAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MESTRADO ACADÊMICO EM CIÊNCIA DA COMPUTAÇÃO

FRANCISCO LEONARDO JALES MARTINS

UM SISTEMA E-HEALTH EM BIG-DATA PARA ANÁLISE E DETECÇÃO DE RISCO DE CHOQUE SÉPTICO EM PACIENTES ADULTOS.

FORTALEZA – CEARÁ

2017

FRANCISCO LEONARDO JALES MARTINS

UM SISTEMA E-HEALTH EM BIG-DATA PARA ANÁLISE E DETECÇÃO DE RISCO DE CHOQUE SÉPTICO EM PACIENTES ADULTOS.

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências e Tecnologia da Universidade Estadual do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação

Orientador: Prof. Dr. Joaquim Celestino Júnior

Co-Orientador: Prof. Rafael Lopes Gomes

FORTALEZA – CEARÁ

2017

Dados Internacionais de Catalogação na Publicação

Universidade Estadual do Ceará

Sistema de Bibliotecas

Martins, Francisco Leonardo Jales.
UM SISTEMA E-HEALTH EM BIG-DATA PARA ANÁLISE E DETECÇÃO DERISCO DE CHOQUE SÉPTICO EM PACIENTES ADULTOS. [recurso eletrônico] / Francisco Leonardo Jales Martins. - 2017.
1 CD-ROM: il.; 4 ¾ pol.

CD-ROM contendo o arquivo no formato PDF do trabalho acadêmico com 58 folhas, acondicionado em caixa de DVD Slim (19 x 14 cm x 7 mm).

Dissertação (mestrado acadêmico) - Universidade Estadual do Ceará, Centro de Ciências e Tecnologia, Mestrado Acadêmico em Ciência da Computação, Fortaleza, 2017.
Área de concentração: Ciências da Computação.
Orientação: Prof. Dr. Joaquim Celestino Júnior.
Coorientação: Prof. Dr. Rafael Lopes Gomes.

1. E-HEALTH. 2. CHOQUE SÉPTICO. 3. CLUSTERIZAÇÃO. 4. BIGDATA. 5. ANÁLISE DE DADOS. I. Título.

ATA DA CENTÉSIMA SÉTIMA DEFESA PÚBLICA DE DISSERTAÇÃO DE MESTRADO

Aos 31 e um dia do mês de agosto de dois mil e dezessete, no miniauditório do prédio de Pesquisa e Pós-Graduação em Computação, do Mestrado Acadêmico em Ciência da Computação – MACC, realizou-se a sessão pública de defesa da dissertação de **Francisco Leonardo Jales Martins,** aluno regularmente matriculada no Mestrado Acadêmico em Ciência da Computação–MACC, Intitulada: **"UM SISTEMA E-HEALTH EM BIG-DATA PARA ANÁLISE E DETECÇÃO DE RISCO DE CHOQUE SÉPTICO EM PACIENTES ADULTOS".** A Banca Examinadora reuniu-se no horário de 15:00h às 17:30 horas, sendo constituída pelos Professores Doutores: Joaquim Celestino Júnior (Orientador-UECE), Rafael Lopes Gomes (Coorientador/UECE), Marcial Porto Fernandez (UECE), Leonardo Sampaio Rocha (UECE) e Francisca Elisângela Teixeira Lima (UFC). Inicialmente o mestrando expôs seu trabalho e a seguir foi submetido à arguição pelos membros da Banca, dispondo cada membro de tempo para tal. Finalmente a Banca reuniu-se em separado e concluiu por considerar o mestrando APROVADO, por sua dissertação e sua defesa pública. Eu, Prof. Dr. Joaquim Celestino Júnior, orientador e presidente da Banca, lavrei a presente ata que será assinada por mim e os demais membros da Banca. Fortaleza, 31 de agosto de 2017.

Prof. Dr. Joaquim Celestino Júnior
(Orientador/UECE)

Rafael Lopes Gomes
(Coorientador/UECE)

Marcial Porto Fernandez
(UECE)

Leonardo Sampaio Rocha
(UECE)

Francisca Elisângela Teixeira Lima
(UFC)

À minha família, por sua capacidade de acreditar em mim e investir em mim. Mãe, seu cuidado e dedicação foi que deram, em alguns momentos, a esperança para seguir. Pai, sua presença significou segurança e certeza de que não estou sozinho nessa caminhada.

AGRADECIMENTOS

Inicialmente à minha família, meus pais Marcondes e Graça, meus irmãos Leandro e Letícia, meus tios e tias Gerúsia, Chagas, Francinildo, Jeane, Joaquim que estiveram presentes durante toda minha vida acadêmica. Em especial ao meu avô João Taveira e minha avó Raimunda que sempre me guiaram pelos melhores caminhos da minha vida.

Aos meus amigos e amigas que estiveram presentes durante esse período acadêmico.

Ao professor e amigo especial, Joaquim Celestino Júnior, pela oportunidade e confiança que depositou em mim durante todos esses anos, me dando apoio durante todos os momentos dificeis vividos. Obrigado meu amigo!!

Ao professor e amigo André Cardoso, por todo apoio e conselhos durante a minha vida acadêmica, não esqueço de nenhuma palavra de sabedoria.

A todos os professores que estiveram nessa caminhada, e colaboraram de alguma forma: Mariela, Valdísio, Jorge Luiz, Glauber Cintra, Leonardo Sampaio, Marcial, Ana Luiza, Jeandro Bezerra, Gustavo, etc ...

Em especial a Caroline, minha princesa, que esteve presente em todos os momentos me dando apoio. Seu amor me deu tranquilidade para concluir mais essa etapa. Agradeço ao senhor por colocar você na minha vida.

Obrigado a todos.

“Apenas o estudo liberta o homem.”

(Edson Pessoa)

RESUMO

A internet das coisas(IoT) é definido como um sistema de dispositivos computacionais que se inter-relacionam e transferem dados através de uma rede sem a interferência humana. Uma das áreas de atuação em IoT é a E-health, que é um termo utilizado para práticas de cuidados do paciente através de dispositivos eletrônicos, nos quais sensores podem monitorar o sinais vitais de pacientes e analisá-los em tempo real. A análise das informações geradas em tempo real por esses sensores é um grande problema de dados, uma vez que a quantidade de dados é muito grande, as aplicações são sensíveis ao tempo e o formato de dados é heterogêneo. Entre as doenças existentes, a sepse é uma síndrome clínica fatal, resultante de infecção. A identificação de estratégias de prevenção específicas para sepsis é uma prioridade de saúde pública, salvando vidas e despesas públicas. Nesse contexto, a dissertação propõe um sistema de análise utilizando a infraestrutura em Big Data para detectar pacientes que tem alto risco de sofrer um choque séptico, já que a descoberta e intervenção precoce reduz a alta taxa de mortalidade associada a sepse grave, ou choque séptico. Devido a grande similaridade de dados capturados pelos sensores, foi proposto um algoritmo baseado na clusterização iterativo k-means, que reduz significativamente a quantidade de dados analisados, visando ter diagnósticos mais precisos dos dados que estão fora da normalidade. Os resultados mostram que a proposta é eficiente e tem análises bem precisas, ótima sensibilidade e especificidade, atingindo 93% a taxa de falsos positivos com 91% de precisão.

Palavras-chave: E-health. Choque Séptico. Clusterização. BigData. Análise de Dados.

ABSTRACT

The Internet of Things (IoT) is defined as a system of computational devices that interrelate and transfer data over a network without human interference. One of the areas of action in IoT is E-health, which is a term used for patient care practices through electronic devices, in which sensors can monitor the vital signs of patients and analyze them in real time. An analysis of the information generated in real time by these sensors is a major data problem, since the amount of data is very large, as applications are time sensitive and data format is heterogeneous. Among existing infections, sepsis is a fatal clinical syndrome resulting from infection. Identification of a specific prevention strategy for sepsis and a public health priority, saving public lives and expenditures. In this context, the dissertation proposes a system of analysis, use a Big Data infrastructure to detect patients at high risk for septic shock, since early detection and intervention reduces a high mortality rate associated with sepsis or septic shock. Due to the great similarity of data captured by the sensors, an algorithm based on iterative k-means clustering was proposed, which reduces the amount of data analyzed, aiming to have more accurate diagnoses of data that are out of normality. The results show that the proposal is efficient and has very precise analyzes, excellent sensitivity and specificity, adhering 93% to the false positive rate with 91% accuracy.

Keywords: E-health. Septic Shock. Clustering. BigData. Data Analysis.

LISTA DE ILUSTRAÇÕES

Figura 1 – Parâmetros de *Big Data* (SABIA; ARORA, 2014) 19
Figura 2 – Pipeline Análise Bigdata 24
Figura 3 – Funcionamento K-means. 32
Figura 4 – Fluxograma de comunicação e geração de dados. 34
Figura 5 – Fluxograma de detecção de quedas. 34
Figura 6 – Arquitetura de análise de dados. 35
Figura 7 – Modelagem Fuzzy. 35
Figura 8 – Framework 38
Figura 9 – Fases de processamento de dados em *Big Data* 38
Figura 10 – Arquitetura 39
Figura 11 – Funcionamento K-means Proposto. 43
Figura 12 – Arquitetura do Cenário de teste. 46
Figura 13 – Formação dos Clusters 50
Figura 14 – Formação dos clusters em todas as dimensões utilizadas 51

LISTA DE TABELAS

Tabela 2 – Resultados do sistema proposto comparando com o proposto em (NGUYEN *et al.*, 2014) 52

LISTA DE QUADROS

Quadro 1 – Critérios de Diagnósticos de Sepse (DELLINGER *et al.*, 2013). 25
Quadro 2 – Classes correspondentes. 44
Quadro 3 – Sumário dos parâmetros de simulação. 47
Quadro 4 – Medidas de concordância observada para dados categóricos (LANDIS; KOCH, 1977) 49
Quadro 5 – Resultados da análise Kappa e dos índices de concordância 52

LISTA DE ALGORITMOS

Algoritmo 1 – Pseudo Algoritmo K-means. 33

Algoritmo 2 – Algoritmo Kmeans with clustering modificado. K-means-C(X, α) . 43

Algoritmo 3 – Algoritmo do Analisador 44

LISTA DE ABREVIATURAS E SIGLAS

5V's	Variedade, Volume, Valor, Veracidade e Velocidade
CPU	Unidade Central de Processamento
DE	Departamento de Emergência
EHR	Electronic Health Record
EMR	Electronic Medical Record
ETL	Extract, Transforming e Loading
HDFS	Hadoop Distributed File
MIMIC II	Multiparameter Intelligent Monitoring in Intensive Care
SATA	Serial AT Attachment
SIRS	Síndrome de Resposta Inflamatória
SQL	Structured Query Language
UTI	Unidade de Terapia Itensiva

LISTA DE SÍMBOLOS

D_i	Distorção do cluster i
EP	Erro Padrão
F_k	Função de avaliação
FN	Falso Negativo
FP	Falso Positivo
IC	Intervalo de confiança
$IC_{95\%}$	Intervalo de confiança de 95%
K_a	Índice Kappa
k_it	t-ésimo objeto pertencente ao cluster i
K	Número de Clusters
N_i	Quantidade de objetos pertencentes ao cluster i
P_0	Taxa de aceitação relativa
P_e	Taxa hipotética de aceitação
P_{avg}	Índice de concordância negativa
P_{pos}	Índice de concordância positiva
S_k	Soma das distorções dos k clusters
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
w_i	Centróide do cluster i
α_k	Fator de peso

SUMÁRIO

1 INTRODUÇÃO 17
1.1 OBJETIVOS 19
1.1.1 Objetivo Geral 19
1.1.2 Objetivo Específico 19
1.1.3 Contribuições 20
1.2 ORGANIZAÇÃO DOS CAPÍTULOS 20
2 FUNDAMENTAÇÃO TEÓRICA 21
2.1 BIG DATA 21
2.1.1 Volume 21
2.1.2 Variedade 22
2.1.3 Valor 22
2.1.4 Veracidade 23
2.1.5 Velocidade 23
2.1.6 Análise de dados em BigData 23
2.2 SEPSE E CHOQUES 25
2.2.1 Choques 25
2.2.1.1 Choque Séptico 26
2.3 CLUSTERIZAÇÃO 26
2.3.1 Métodos de particionamento 27
2.3.2 Métodos hierárquicos 27
2.3.3 Métodos baseados em densidade 28
2.3.4 Métodos baseados em Grid 28
2.4 CORRELAÇÃO DE PEARSON 29
2.5 SÍNTESE DO CAPÍTULO 29
3 TRABALHOS RELACIONADOS 30
3.1 SELECTION OF K IN K-MEANS CLUSTERING 30
3.1.1 Clusterização K-meios 31
3.2 SMART VITAL SIGNS MONITORING AND NOVEL FALLS PREDICTION SYSTEM FOR OLDER ADULTS 33
3.3 AUTOMATED ELECTRONIC MEDICAL RECORD SEPSIS DETECTION IN THE EMERGENCY DEPARTMENT 36

4 UM SISTEMA E-HEALTH EM BIG-DATA PARA ANÁLISE E DETECÇÃO DE CHOQUE SÉPTICO EM PACIENTES ADULTOS. . . . 37
4.1 PRESSUPOSIÇÕES 37
4.2 VISÃO GERAL 37
4.3 ARQUITETURA 39
4.3.1 Representação dos Dados 40
4.3.2 Clusterização 40
4.3.2.1 Funcionamento 42
4.3.3 Analisador de Dados 43
4.4 SÍNTESE DO CAPÍTULO 45
5 SIMULAÇÕES E RESULTADOS 46
5.1 AMBIENTE 46
5.2 CENÁRIO 46
5.3 PARÂMETROS DOS TESTES 47
5.4 MÉTRICAS 47
5.5 RESULTADOS 49
5.6 SÍNTESE DO CAPÍTULO 53
6 CONCLUSÕES 54
6.1 TRABALHOS FUTUROS 54
REFERÊNCIAS 55
GLOSSÁRIO 58

1 INTRODUÇÃO

O setor da saúde do paciente cresceu rapidamente nos últimos 30 anos, gerando grandes quantidades de dados, impulsionado pela manutenção de registros, conformidade e requisitos regulatórios.

A importância no monitoramento de pacientes tem crescido substancialmente nos últimos anos. Esse cenário provoca uma alta demanda, exigindo cada vez mais da atual infraestrutura de serviços de saúde. O uso da tecnologia de computação ubíqua, pode representar uma solução para este problema, onde o paciente pode ser monitorado a todo momento por meio de sensores utilizados no ambiente domiciliar ou hospitalar que coletam dados biológicos (pressão arterial, frequência cardíaca, etc.).

Recentemente, a grande quantidade de dados gerados nos cuidados da saúde do paciente referem-se aos *Electronic Health Record (EHR)*, nos quais são grandes, complexos e difíceis de gerenciar utilizando os sistemas atuais. Além disso, estes sistemas não são facilmente controlados com as ferramentas e métodos tradicionais. Logo, o conceito de *E-health* (EYSENBACH, 2001), emergiu como uma solução para o gerenciamento do EHR.

Um dos campos da medicina que requer um cuidado especial através do monitoramento é a identificação e tratamento precoce de pacientes que sofrem de sepse grave ou choque séptico. A sepse é uma causa importante de hospitalização e a principal causa de mortes em Unidade de Terapia Itensiva (UTI) (ENGEL *et al.*, 2007; MARSHALL *et al.*, 2005; ALBERTI *et al.*, 2003). No Brasil, a taxa de mortalidade de pacientes com sepse grave(hipoperfusão) e choque séptico(falência circulatória) é de 46,9% e 52,2%, (SILVA *et al.*, 2004), respectivamente. Pesquisas (WESTPHAL; LINO, 2015) comprovam que a identificação precoce, ou suspeita de sepse, reduz a taxa de mortalidade significativamente.

De acordo com os Centros para o Controle e Prevenção de Doenças (CDC), a sepse é a 11ª principal causa de morte nos Estados Unidos sendo o motivo mais caro e terceiro para hospitalizações, com alta morbidade e mortalidade, e US$ 20,3 bilhões em custos agregados hospitalares (TORIO; ANDREWS, 2006). O Sistema de Saúde da Universidade do Kansas através de monitoramento de pacientes por pacientes não evasivos foi capaz de reduzir a permanência da UTI em quase 3 dias, reduzindo o custo hospitalar em US$ 14 milhões e ajudando cerca de 1000 pacientes em um período de 6 meses (LATHAM *et al.*, 2017).

Nos últimos anos, o crescimento no volume de dados (não só médicos, mas em diversas áreas) aumentou exponencialmente. Criando assim a ideia de *Big Data*, que é usado

para descrever o crescimento, disponibilidade e utilização de informações, estruturadas ou não, a partir de fontes e domínios diferentes. Essa massa de dados gerada é de difícil processamento, devido a grande quantidade de registros gerados todos os dias por redes sociais, sinais de GPS, dispositivos móveis, vídeos etc.

A análise em *Big Data* em Saúde é fundamentalmente um conjunto de metodologias, procedimentos, estruturas e tecnologias que são aplicados para transformar dados brutos em dados significativos. Essas informações são usadas para tornar as tarefas de tomada de decisão mais efetivas, sejam eles estratégicos, táticos ou operacionais. Analisar essa massa de dados é um processo complexo e difere entre os demais dados estruturados em 5 parâmetros - Variedade, Volume, Valor, Veracidade e Velocidade (5V's). São os desafios de gerenciamento em *Big Data* (DEMCHENKO, 2013):

- Variedade: As fontes de dados são heterogêneas. Os arquivos são de vários formatos e tipos diferentes, podendo ser estruturado, ou não, tais como texto, áudio, vídeos, históricos e outros. Além disso, estes arquivos podem ser estruturados ou não, e quando estruturados, possuem formatação específica.
- Volume: A quantidade de dados gerados está em pleno crescimento, este fato resulta em tamanho de arquivos cada vez maiores, e consequentemente gera um volume de dados excessivo a ser armazenado. Então as pesquisas são direcionadas a reduzir o custo de armazenamento, tendo em vista que nos próximos 4 anos, o volume desses dados irá crescer em torno de 50 vezes (ANSHARI; ALAS, 2015).
- Valor: A valoração dos dados da empresa foi um dos mais recentes V's incluído nas pesquisas em *Big Data*. Principalmente o valor agregado de todo trabalho desenvolvido, coleta, armazenamento e análise dos dados para compensar os custos envolvidos.
- Veracidade: Quando lidamos com grande volume e variedade de dados haverá uma grande quantidade de dados que não interessam. *Big Data* e tecnologias de análise trabalham com estes tipos de dados.
- Velocidade: Os dados gerados são em alta velocidade. Às vezes, um minuto é muito tarde em *Big Data* pra dados que são sensíveis ao tempo, fato que limita o tempo de processamento de alguns cenários. Para algumas organizações, a velocidade dos dados é o principal desafio. As mensagens de mídia social e transações de cartão de crédito são feitas em milissegundos e os dados gerados por este, são colocando em bancos de dados.

A clusterização fornece técnicas e medidas para reduzir o tamanho dos dados a ser analisado, através do agrupamento dos nós em clusters de tamanho gerenciável. Esta estratégia é

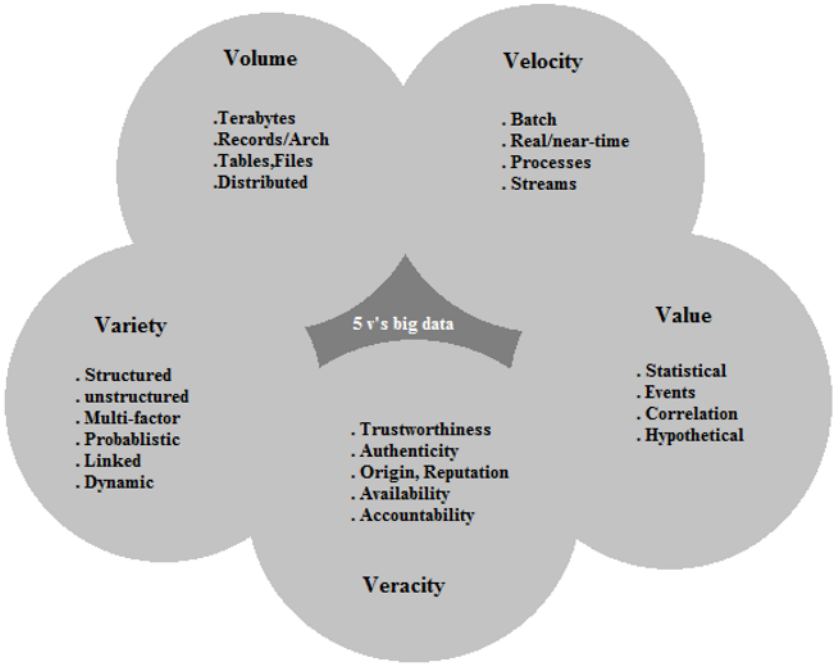

Figura 1 – Parâmetros de *Big Data* (SABIA; ARORA, 2014)

aplicada em vários campos de pesquisa. Em uma massa de dados clusterizada, cada grupo de nós tem um líder, chamado de cluster-head. O cluster-head eleito determina e funde as informações do agrupamento e transmite no sistema.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Este trabalho propõe um sistema *e-health* em *Big Data* para analisar e detectar se um paciente está evoluindo para o estado de sepse grave, podendo levar a um estado de choque séptico em pacientes adultos, utilizando clusterização K-meios de dados similares gerados pelos sensores coletores.

1.1.2 Objetivo Específico

- Aumentar a precisão da análise dos dados;
- Diminuir a ocorrência de falso positivo (Especificidade);
- Diminuir a ocorrência de falso negativo (Sensibilidade);
- Detectar risco de sepse e choque séptico em pacientes adultos.

1.1.3 Contribuições

Este trabalho tem as seguintes contribuições:

- Utilização da clusterização K-meios para agrupar dados similares em *Big Data* para melhorar a performance de análise;
- Diminuir a ocorrência de falsos positivos e falsos negativos, para ter uma análise mais precisa;
- Eficiência na análise dos dados de pacientes para realizar o tratamento precoce, assim reduzindo a taxa de mortalidade causada pela sepse grave, ou choque séptico.

1.2 ORGANIZAÇÃO DOS CAPÍTULOS

A dissertação está organizada em seis capítulos: O capítulo 1 apresenta a introdução do trabalho; O capítulo 2 apresenta fundamentação teórica do trabalho. Portanto especifica uma visão geral sobre *Big Data*, clusterização, análise de dados em *Big Data*, sepse e choques, e Similaridade de Dados; O capítulo 3 apresenta os trabalhos relacionados que serviram de inspiração para a pesquisa; O capítulo 4 detalha o projeto, onde será mostrado a arquitetura, as pressuposições e a solução apresentada; O capítulo 5 detalha a arquitetura dos testes, resultados obtidos; O capítulo 6 expõe a conclusão obtida do trabalho e detalha os possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta um estudo detalhado sobre *Big Data*, componentes de análise em *Big Data*, clusterização, similaridade de dados, sepse e choques. Na Seção 2.1, está descrito o que é *Big Data*, suas características e desafios de pesquisa. Na Seção 2.2, está descrito o que é sepse e choques, em seguida será falado sobre choque séptico. Na Seção 2.3, estão descritos os tipos e abordagens de clusterização. Na Seção 2.4, está descrito o que é a similaridade de dados, e a correlação de Pearson.

2.1 BIG DATA

Big data é usado para descrever o crescimento, disponibilidade e utilização de informações, estruturados ou não, a partir de fontes e domínios diferentes. Ferramentas avançadas, software e sistemas são necessários para capturar, armazenar, gerenciar e analisar os conjuntos de dados, em um período de tempo que preserva o valor intrínseco dos dados.

O conceito de *Big Data* era originalmente o domínio exclusivo da pesquisa universitária avançada e laboratórios. Entretanto, este passou a ser aplicado de forma mais ampla para cobrir ambientes comerciais, devido os avanços tecnológicos maciços, incluindo: poder de processamento mais barato, computação em cluster, menor custo de armazenamento Serial AT Attachment (SATA) e melhorias de desempenho de rede que permitem às empresas realizar tarefas de computação que anteriormente exigiam sistemas altamente sofisticados e caros.

Como dito anteriormente, dos conceitos em *Big Data*, o mais comum é o volume. Devido ao crescimento do volume de dados (TANKARD, 2012). Entretanto, de acordo com (WU *et al.*, 2014), desafios em *Big Data* devem considerar as relações entre cinco dimensões: Volume, Variedade, Valor, Veracidade e Velocidade dos dados.

2.1.1 Volume

Gerenciamento de volumes elevados e crescentes de dados tem sido uma questão difícil para muitas décadas. No passado, este desafio foi atenuado por processadores cada vez mais rápidos para nos fornecer os recursos necessários para lidar com o aumento do volume de dados. Mas, há uma mudança fundamental: volume de dados está aumentando mais rápido do que recursos de computação, e velocidades da Unidade Central de Processamento (CPU) são limitados.

No passado, os sistemas de processamento de dados de grandes dimensões tinham que se preocupar com o paralelismo entre os nós em um ponto; agora, é preciso lidar com o paralelismo dentro de um único nó. Infelizmente, técnicas de processamento de dados paralelos aplicadas no passado para os dados entre nós não podem ser aplicadas diretamente para intra-paralelismo. Estas mudanças levam a repensar como podemos projetar, construir e operar os componentes de processamento de dados.

Nesse contexto, há algumas soluções baseadas em um sistema de arquivos distribuídos, conhecido como *Hadoop Distributed File (HDFS)*, que viabilizam a disponibilidade desses dados em múltiplos nós.

2.1.2 Variedade

Os dados podem ser obtidos de diversas fontes(Redes sociais, dados de imagem, dados de sensores, etc...) e não estão prontos para integração com outros sistemas de software. O processamento desses grandes volumes de dados não estruturados é uma característica comum, e permite extrair informações significantes que podem ser consumidas por uma aplicação. Dependendo da aplicação, alguns tipos de armazenamento tornam-se mais eficiente que outros (DUMBILL, 2012).

Algoritmos de análise de máquina esperam dados homogêneos, e não consegue compreender a heterogeneidade dos elementos. Em consequência, os dados devem ser cuidadosamente estruturado, como uma primeira etapa, ou antes, a análise de dados.

No entanto, os sistemas computacionais funcionam mais eficientemente se eles podem armazenar vários itens que são todos idênticos em tamanho e estrutura. A representação eficiente de acesso e análise dos dados semi-estruturados requerem mais trabalho.

2.1.3 Valor

Em *Big Data*, o valor agregado está relacionado ao custo da coleta, armazenamento e processamento dos dados (WEBER; OTTO; ÖSTERLE, 2009). A qualidade da informação exige exatidão, integridade, consistência e relevância. Em (MANYIKA *et al.*, 2011), cita a confiança e experiência para a extração dos dados em *Big Data*.

Através da mineração e análise eficaz de dados, expõe informações de negócios valiosas de dados (estruturados, ou não) de streaming e warehouses. Essa percepção pode ser usada para ajudar a renovar as cadeias de abastecimento, melhorar o planeamento do sistema,

vendas e atividades de marketing. Medir o desempenho em todos os canais pode transformar o sistema em uma atividade on-demand. A estratégia *Big Data* dá às empresas a capacidade de analisar melhor esses dados com o objetivo de acelerar o crescimento rentável.

2.1.4 Veracidade

A privacidade de dados é uma grande preocupação, no contexto de *Big Data*. Para registos de saúde, existem leis rígidas que regem o que pode ou não ser feito. No entanto, há um grande temor público sobre do uso inadequado de dados pessoais, através da ligação de dados de várias fontes. A qualidade da informação, garantindo a integridade, segurança e a relevância, deve ser o foco.

2.1.5 Velocidade

A frequência com que os dados são gerados ou entregues também são relevantes à definição de Big Data. O grande fluxo dos dados é praticamente em tempo real e as janelas de atualização tendem a ser reduzidas à frações de segundos. Dessa forma, vários dispositivos interconectados propiciam o crescimento na taxa de produção desses dados, o que resulta na caracterização dos stream de dados (DUMBILL, 2012).

Dado um grande conjunto de dados, muitas vezes é necessário encontrar elementos em que respondam a um critério especifico. Este tipo de pesquisa é suscetível de ocorrer várias vezes. A digitalização de todos os dados definir para encontrar elementos adequados é obviamente impraticável. Em vez disso, estruturas de índice são criadas com antecedência a fim de encontrar elementos qualificados rapidamente. Entretanto cada estrutura de índice é criada apenas para algumas classes de critérios que podem atender a novos tipos especificados, sendo necessário conceber novas estruturas de índice para suportar tais critérios. Tarefa difícil devido o volume de dados crescer rapidamente e as consultas terem prazos de resposta cada vez mais curtos.

2.1.6 Análise de dados em BigData

Uma área atual de pesquisa em Big data é a análise dos dados. Os desafios, segundo (WU *et al.*, 2014), são:

- Heterogeneidade: Algoritmos de análise devem trabalhar em dados heterogêneos, os quais devem ser cuidadosamente estruturados.

- Escala: Gerenciamento eficiente de uma grande quantidade de dados.
- Timeliness: Rapidez no processamento de um conjunto de dados.
- Privacidade: Garantir a privacidade dos dados, devidos a contratos, leis entre outros tópicos.
- Colaboração Humana: Aplicar padrões que os humanos podem facilmente detectar.

Para fazer com que todo o processo de descoberta de conhecimento em bases de dados seja mais claro, (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) resume em: Seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação. Como mostrado na Figura 2:

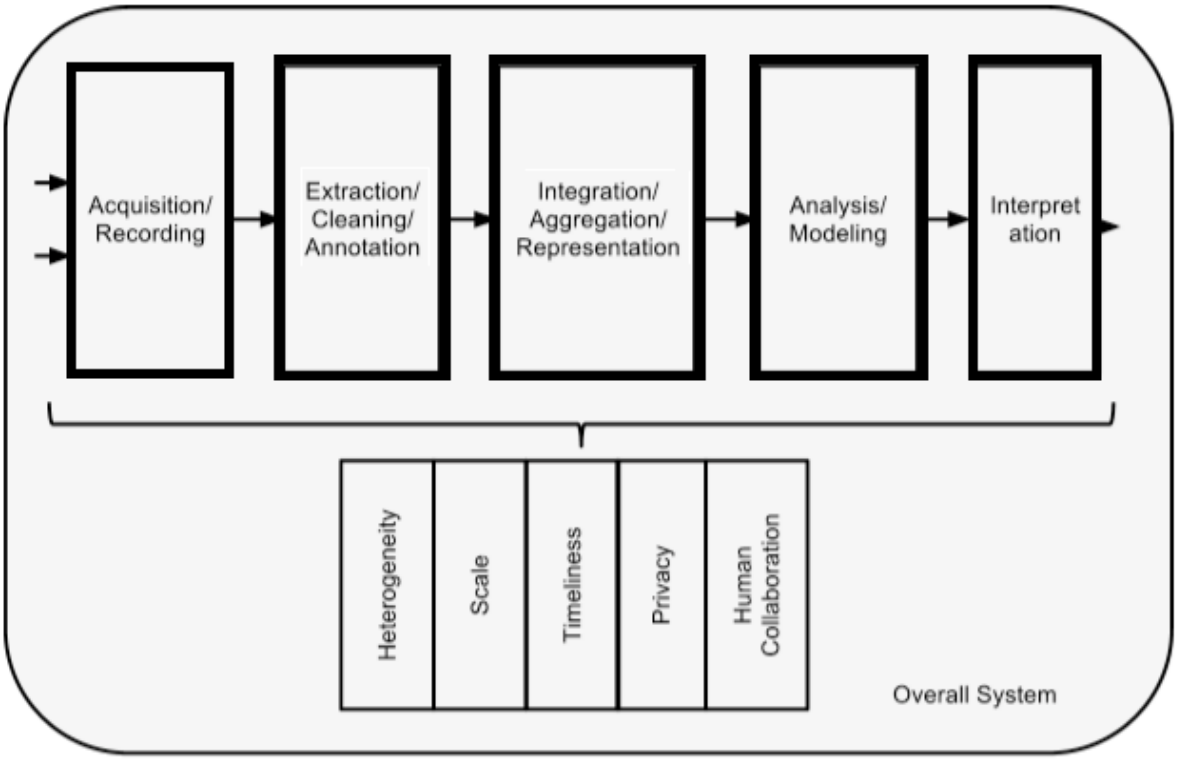

Figura 2 – Pipeline Análise Bigdata

- Seleção: Habilidade de gravar, ou recuperar os dados de uma fonte geradora de dados
- Pré-processamento: Fase responsável por extrair os dados, ainda não interpretáveis, para a análise propriamente dita.
- Transformação: Visa gerar uma representação lógica, preparando os dados para a modelagem
- Mineração de dados: Métodos de mineração dos dados necessários, para serem interpretados pela aplicação
- Interpretação/Avaliação: Interpreta os dados minerados para a aplicação.

Esses operadores são capazes de construir um sistema de análise de dados completo, primeiramente para agrupar dados e, em seguida, encontrar informações a partir desses dados e exibir o conhecimento para os interessados.

A coleta, seleção, pré-processamento e os operadores de transformação fazem parte da entrada no processo de análise. A seleção visa identificar qual o tipo de dados foi necessário para análise, selecionar as informações consideradas importantes a partir dos dados recolhidos. Assim, os dados coletados de diferentes fontes, serão integrados aos dados de destino. O operador de pré-processamento lida com os dados de entrada, sua função é detectar, limpar e filtrar os dados desnecessários para torna-los utilizáveis. Operadores de transformação convertem os dados em um formato que a mineração seja capaz de ser realizada (HAN; PEI; KAMBER, 2011).

2.2 SEPSE E CHOQUES

Sepse é uma Síndrome de Resposta Inflamatória (SIRS) (SALLES *et al.*, 1999), motivada por um agente agressor, associada à infecção sistêmica e tem alto índice de mortalidade (ENGEL *et al.*, 2007).

Quando o indivíduo é atacado por agentes infecciosos o corpo reage liberando mediadores químicos que provocam uma resposta inflamatória. Se grandes quantidades de bactérias chegam em massa à corrente sanguínea, espalhando-se pelo corpo, as células de defesa agem para combater a infecção, ocasionando um processo inflamatório difuso. Podendo ocasionar uma infecção generalizada, que é definido por Sepse.

A inclusão de dois ou mais critérios clínicos para diagnóstico, conforme a tabela 1, resulta em diagnóstico de sepse.

Quadro 1 – Critérios de Diagnósticos de Sepse (DELLINGER *et al.*, 2013).

Item	Sintomas
1	Temperatura Corporal > 38^oC ou < 36^oC.
2	Frequência respiratória > $20mpm$
3	Frequência Cardíaca > $90bpm$
4	Pressão arterial sistólica < $90mmHg$ ou pressão arterial média < $65mmHg$

Fonte – Elaborado pelo autor

2.2.1 Choques

O choque é uma síndrome caracterizada por insuficiência circulatória aguda com má distribuição generalizada do fluxo sanguíneo, que implica falência de oferta e/ou utilização do

oxigênio nos tecidos (FELICE *et al.*, 2011).

Os estados de choque podem ser classificados em:

- Hipovolêmico: É caracterizado pelo baixo volume intravascular, ou baixa capacitância.
- Obstrutivo: É ocasionado por um débito cardíaco deminuído por consequência de uma obstrução mecânica.
- Cardiogênico: É consequência da falência primária da bomba cardíaca.
- Distributivo: É caracterizado por inadequação entre a demanda tecidual e a oferta de oxigênio por uma alteração no fluxo sanguíneo.

2.2.1.1 Choque Séptico

Choque séptico é um choque distributivo, é definido pela presença de SIRS de origem infecciosa, podendo ser comprovada ou fortemente presumida. Como descrito anteriormente, o choque é caracterizado pela presença de dois ou mais critérios presentes na Tabela 1.

O choque séptico é definido como uma sepse grave associada à hipotensão refratária à reposição volêmica e com necessidade de uso de vasopressor para manter a pressão arterial (MORRELL; MICEK; KOLLEF, 2009), ou seja, uma insuficiência circulatória à reposição da quantidade de sangue circulando no corpo, causado pelo relaxamento das vias circulatórias devido a infecção generalizada, mesmo utilizando fármacos para manter a pressão arterial.

2.3 CLUSTERIZAÇÃO

Clusterização é o processo de agrupar um conjunto de objetos de dados em vários clusters de forma que objetos dentro de um cluster têm grande similaridade, mas são diferentes para objetos em outros clusters. Diferenças e semelhanças são avaliadas com base nos valores dos atributos que descrevem os objetos e a distância entre eles. Clusterização como uma ferramenta de mineração de dados tem suas raízes em muitas áreas de aplicação, tais como biologia, segurança, inteligência de negócios e busca na Web.

Clusterização em *Big Data* é um campo de pesquisa, com as seguintes restrições:

- Escalabilidade: Conseguir tratar uma grande quantidade de informação em um tempo computacionalmente escalável;
- Heterogeneidade: Tratar dados de diferentes tipos e atributos;
- Elasticidade: Tamanho dos clusters dever ser maleáveis para as diferentes quantidades de objetos;

- Tolerância a perturbação: Tratar elementos semelhantes, com um certo grau de tolerância.

As técnicas básicas de clusterização, são organizadas nas seguintes categorias: métodos de particionamento, hierárquicos, baseados em densidade, e grid.

2.3.1 Métodos de particionamento

Dado um conjunto de n objetos, um método de particionamento constrói k partições dos dados, onde cada partição representa um cluster e $k <= n$. Os métodos básicos de particionamento normalmente adotam a separação de cluster exclusivo. Este requisito pode ser relaxado, por exemplo, nas técnicas baseadas em particionamento fuzzy (YANG, 1993).

A maioria dos métodos de particionamento é baseado em distância. Dado K, o número de grupos para a construção, qualquer método cria uma partição inicial. Em seguida, através de busca locais iterativas tentam melhorar a partição, movendo objetos de um grupo para outro. O critério geral de uma boa partição é a distância entre elementos com alta similaridade "perto"ou relacionados uns aos outros. Existem vários tipos de outros critérios para julgar a qualidade de partições.

Métodos de particionamento tradicionais pode ser estendido para o agrupamento de um sub-espaço, em vez de procurar em todo espaço. Isso é útil quando existem muitos atributos e os dados são escassos, ou quando o tamanho do espaço é muito grande. Chegar em uma solução de clusterização otimizada baseada em particionamento é uma tarefa complexa, podendo exigir uma busca exaustiva de todas as partições possíveis. Muitas aplicações adotam métodos heurísticos a fim de contornar este problema, como os k-meios e algoritmos k-medoids, estas abordagens gulosas melhoram progressivamente a qualidade da clusterização e se aproximam de um ótimo local. Para grandes conjuntos de dados, esses métodos precisam ser estendidos.

2.3.2 Métodos hierárquicos

Os métodos hierárquico criam uma decomposição hierárquica de um dado conjunto de objetos de dados. Esses métodos podem ser classificados como sendo ou de aglomeração, ou de divisão, com base na formação da decomposição hierárquica. A abordagem de aglomeração, também chamada de $bottom-up$ (GUHA; RASTOGI; SHIM, 1998), cada objeto formando um grupo separado, mescla com objetos ou grupos próximos até fundir em um só (o nível mais alto da hierarquia). A abordagem de divisão, também chamada de $top-down$ (RASMUSSEN, 1992), todos os objetos no mesmo cluster, dividem-se em grupos menores a cada iteração até

que cada objeto está em um cluster.

Métodos hierárquicos podem ser baseados na ordem contínua dos elementos ou baseado na distância. Várias extensões de métodos hierárquicos consideram agrupamento em subespaços e sofrem com o fato de uma vez que a fusão ou divisão é feito, esta nunca pode voltar atrás. Esta tolerância é útil na medida em que conduz a custos computacionais menores, pois não tem que se preocupar com um número de combinações diferentes de opções. Em tais técnicas não é possível corrigir decisões erradas. No entanto, têm sido propostos métodos para melhorar a qualidade de agrupamento hierárquico.

2.3.3 Métodos baseados em densidade

A maioria dos métodos de particionamento agrupam objetos com base na distância entre eles. Tais métodos podem achar que somente estão aglomerados em forma esférica, e encontram dificuldade em descobrir grupos de outras formas. A ideia geral é a de continuar crescendo a partir de um dado grupo, desde que a densidade (número de objetos ou pontos de dados) na "zona"excede algum limiar. Por exemplo, para cada ponto de dados dentro em certo grupo, a vizinhança do raio dado tem que conter, pelo menos, um número mínimo de pontos (ESTER *et al.*, 1996).

Os métodos baseados em densidade, geralmente pode ser usado para filtrar o ruído e descobrir grupos de forma arbitrária. Métodos baseados em densidade pode dividir um conjunto de objetos em vários clusters exclusivos, ou uma hierarquia de clusters. Normalmente, os métodos baseados em densidade consideram somente conjuntos exclusivos. Além disso, os métodos baseados em densidade pode ser estendido a partir do espaço total ao cluster sub-espaço.

2.3.4 Métodos baseados em Grid

Métodos baseados em grid quantificam o espaço do objeto em um número finito de células que formam uma estrutura de grade. Todas as operações de clusterização são realizados sobre a estrutura da rede. A principal vantagem desta abordagem é o seu tempo de processamento rápido, que geralmente é independente do número de objetos dados, depende apenas da quantidade de células em cada uma das dimensões no espaço. O uso do grid é muitas vezes uma abordagem eficiente em muitos problemas de mineração de dados espaciais que incluem a clusterização. Portanto, métodos baseados em grade pode ser integrado com outros de agrupamento, tais como os baseados em densidade e hierárquicos (KU-MAHAMUD, 2013).

2.4 CORRELAÇÃO DE PEARSON

Semelhança de dados é a medida de quão parecidos dois objetos são. O contexto de similaridade na mineração de dados é geralmente descrito como um raio, com dimensões que representam características dos objetos. O grau de semelhança entre dois objetos é inversamente proporcional a sua distância no plano cartesiano. Similaridade é subjectivo e altamente dependente do domínio e aplicação. Por exemplo, dois frutos pode ser semelhante por causa de sua cor, tamanho ou gosto. Cuidados devem ser tomados quando se calcula a distância entre dimensões/recursos que não estão relacionados. Os valores relativos de cada recurso deve ser normalizada, ou uma característica pode acabar dominando o cálculo da distância.

Em estatística, uma medida de similaridade, é uma função real que a quantifica entre dois objetos (FREY; DUECK, 2007) sugerem que a medida de similaridade entre dois objetos seja definida pela distância euclidiana 2.1.

$$S(x,y) = -\|(x-y)\|_2^2 \tag{2.1}$$

Um método para determinar a correlação de similaridade é o de Pearson (FISHER, 1915; SAMPLE, 1921; GAYEN, 1951). O método estatístico mede a correlação linear entre duas variáveis X e Y, com valores entre -1 e +1. Quanto mais se aproximar da correlação positiva, mais forte é o nível de similaridade, ao contrário, quanto mais se aproximar da correlação negativa, mais fraco será o nível de similaridade. O coeficiente de correlação de Pearson é a covariância das duas variáveis, dividido pelo produto dos seus desvios padrão, como definido na equação 2.2:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2.2}$$

2.5 SÍNTESE DO CAPÍTULO

Nesse capítulo foi apresentado os conceitos gerais em *Big Data*, seus objetivos e desafios. Em seguida, os tipos e abordagens de clusterização, similaridade de dados e a correlação de Pearson. Na última seção foi definido o que é sepse, e choque séptico.

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados os trabalhos relacionados que serviram como contribuição para o trabalho proposto. Na Seção 3.1, será descrito o trabalho *"Selection of K in K-means clustering "*, que visa determinar o número de *k-clusters*, utilizando a clusterização K-meios. Na seção 3.2, será descrito o trabalho *"Smart Vital Signs Monitoring and Novel Falls Prediction System for older adults"*, que descreve um sistema que utiliza a arquitetura em *Big Data* para monitoramento de pessoas. Na seção 3.3, será descrito o trabalho *"Automated electronic medical record sepsis detection in the emergency department"*, que descreve um sistema automatizado para detecção de sepse baseado num valor de predição positiva.

3.1 SELECTION OF K IN K-MEANS CLUSTERING

K-meios é um algoritmo popular para clusterização de dados. Entretanto, uma das suas desvantagens é a exigência de que o número de *K* clusters, sejam especificados antes do algoritmo ser aplicado. Este artigo examina primeiro os métodos existentes para selecionar o número de clusters e propõe um algoritmo para determinar o número de *clusters* para o algoritmo K-meios para diferentes conjuntos de dados.

Há vários métodos para tentar determinar qual o melhor valor de *K*:

- Dentro de um intervalo: Ao invés de utilizar um único valor pré-definido, é determinado um conjunto relativamente grande de valores específicos, para refletir especificamente as características de cada conjunto de dados (AL-DAOUD; VENKATESWARLU; ROBERTS, 1995).
- Pelo usuário: Geralmente utilizado em *datamining*, requer o número de clusters definido pelo usuário, então a heurística deve executar várias rotinas com números diferentes de clusters visando a melhor performance entre eles(BOTTOU; BENGIO *et al.*, 1995).
- Processados depois: A escolha do número de clusters é feita após o pré-processamento dos dados (PHAM; DIMOV; NGUYEN, 2005).
- Gerados randomicamente: Gera números aleatórios, e a cada execução avalia o equilibrio entre os clusters gerados (HANSEN *et al.*, 1999).
- Medidas estatísticas: Utiliza várias técnicas para determinar a melhor configuração estatistica (BRADLEY; FAYYAD, 1998).
- Números de Classes: Determina o número de clusters, o número de classes de dados(PELLEG; MOORE *et al.*, 2000).

- Determinados Visualmente: Determina o número de clusters geometricamente (BILMES *et al.*, 1997).
- Medida de vizinhança: Adiciona fatores da vizinhança, para incluir os parâmetros na função K (KOTHARI; PITTS, 1999).

O método proposto baseia-se na perturbação que deve ocorrer a cada passo de inserção do elemento no conjunto. A clusterização é usada para encontrar irregularidade na distribuição dos dados e identificar aonde tem alta concentração de elementos. Para determinar a região de um cluster, não somente deve-se observar a distribuição interna dos elementos, mas também a interdependência com elementos de outros grupos do conjunto de dados.

A função de distorção de um cluster é uma função da população de dados e a distância entre objetos e centro do cluster 4.3:

$$D_i = \sum_{t=1}^{N_i} [d(k_i t, w_i)]^2 \tag{3.1}$$

onde:

- D_i é a distorção do cluster i;
- w_i é o centroide do cluster i;
- N_i é a quantidade de objetos pertencentes ao cluster i;
- $k_i t$ é o t-ésimo objeto pertencente ao cluster i;
- $d(k_i t, w_i)$ é a distância entre o objeto $x_i t$ e o centro do cluster w_j;

A partir do calculo da distorção, se houver o valor de K pode alterar a cada inserção de elementos no conjunto.

3.1.1 Clusterização K-meios

Algoritmos de clusterização (KAUFMAN; ROUSSEEUW, 2009): particionados constroem uma partição a partir de uma base de dados de n elementos, em conjuntos de k clusters, onde k é um parâmetro de entrada para estes algoritmos. O valor de k pode ser conhecido ou não. Um problema de clusterização é definido em duas classes, o problema K-clusterização, e problema de clusterização simples.

O problema de clusterização simples é encontrar uma maneira de clusterizar os dados de uma base X, de modo que os dados similares estejam agrupados, sem saber ao certo qual o numero máximo de clusters (BERKHIN, 2006; DOVAL; MANCORIDIS; MITCHELL, 1999).

No problema K-clusterização, o número de *K* clusters já são definidos previamente (FASULO, 1999).

Em uma k-clusterização, o número total de diferentes formas de agrupamento de n elementos de um conjunto em k clusters, equivale à função N(n, k) em 3.2:

$$N(n,k) = \frac{1}{k}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^n \tag{3.2}$$

Nesse processo de clusterização, é visível que a busca pela melhor solução no espaço de soluções viáveis é extremamente difícil, pertencendo a classe de problemas NP-Difícil.

A solução apontada na proposta, considera a clusterização por particionamento, onde o conjunto de elementos são divididos em K subconjuntos, com K conhecido (BERKHIN, 2006).

A técnica base para a clusterização é a K-meios, onde o elemento mais significativo no cluster é o centróide (HARTIGAN; WONG, 1979).

K-meios é uma boa escolha para conjuntos de dados que têm um pequeno número de clusters com tamanhos proporcionais e dados linearmente separáveis, utilizando a variação canopy (MCCALLUM; NIGAM; UNGAR, 2000), para usar em grandes conjuntos de dados.

A figura 3, mostra o funcionamento do algoritmo 1:

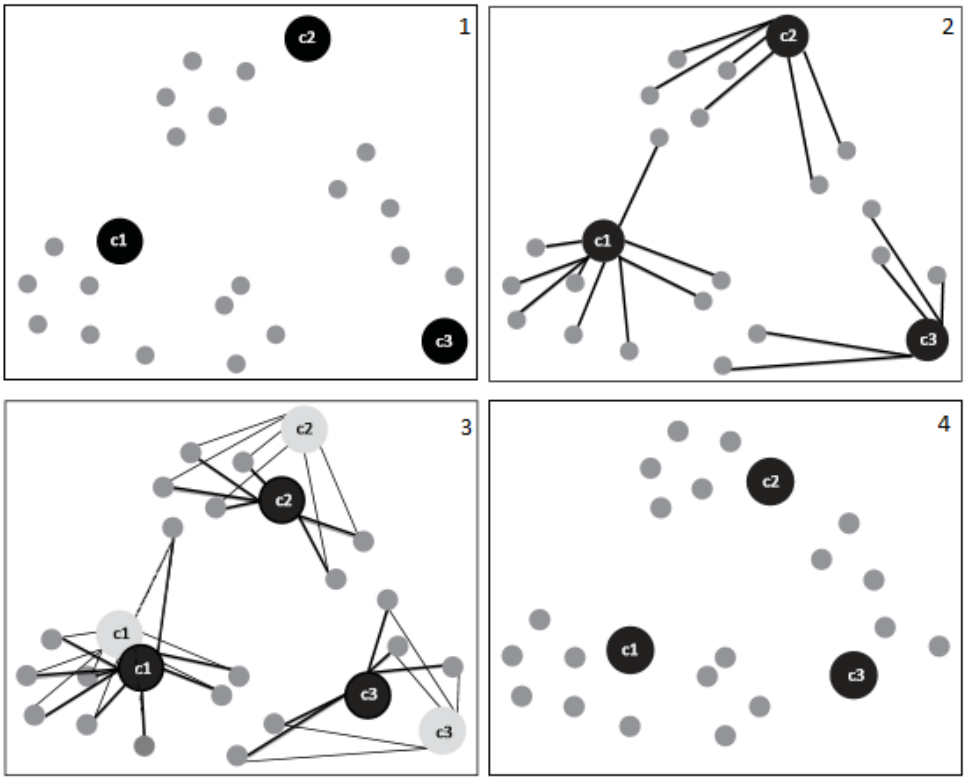

Figura 3 – Funcionamento K-means.

Algoritmo 1: Pseudo Algoritmo K-means.

```
Randomicamente é selecionado k clusters;
foreach C = Cluster Provisório do
    Atribui cada ponto x no raio R pra cada cluster C;
end
foreach C = Cluster provisório do
    foreach x = Ponto no cluster C do
        if x = centróide entre todos os pontos do cluster C then
            C = x;
        end
    end
end
return Retorna a melhor configuração da clusterização;
```

A técnica é barata, em custo de complexidade, e extremamente eficiente quando se trata de grande massas de dados.

Através dessa técnica conseguiremos montar todos os clusters, responsáveis por agrupar os dados similares possíveis.

3.2 SMART VITAL SIGNS MONITORING AND NOVEL FALLS PREDICTION SYSTEM FOR OLDER ADULTS

Esse trabalho propõe uma solução de monitoramento de idosos baseado em sinais vitais coletados por uma rede de sensores, e mostra como pode-se aplicar a arquitetura de análise em *Big Data* e clusterização para gerar soluções ótimas. É definido um framework que trabalha com os dados recebidos por sensores e vai desde o tratamento a nível de rede, até a análise do dado em si. O framework divide-se em 3 conceitos:

- Monitoramento Wireless: Sensores wireless são responsáveis por coletar os sinais vitais e transmitir na rede;
- Detecção de quedas e predição: Utilizando um fluxograma para alerta;
- Detecção precoce e interpretação de sinais vitais anormais: Utiliza uma lógica fuzzy para analisar os dados coletados para a tomada de decisão.

A figura 4, mostra o fluxograma desde a geração de dados pelo indivíduo, até a interface de saída para a tomada de decisão.

O framework coleta os sete tipos de dados mostrados na figura 4, via equipamentos médicos wireless, que alimentam o banco de dados. E a central de processamento é responsável por executar o fluxograma 5 de detecção de quedas.

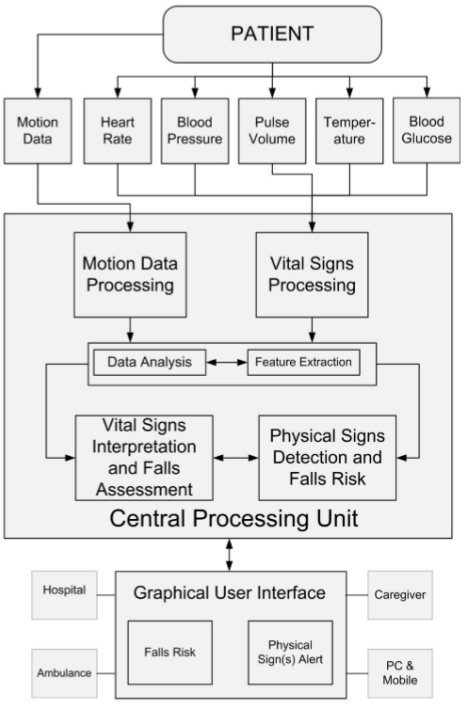

Figura 4 – Fluxograma de comunicação e geração de dados.

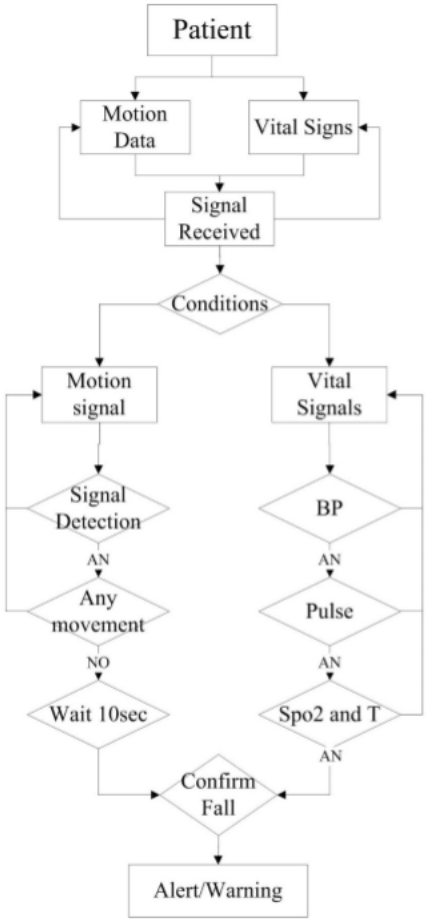

Figura 5 – Fluxograma de detecção de quedas.

Para detectar a anormalidade dos sinais vitais, o framework utiliza a arquitetura de análise em *Big Data* para facilitar a mineração dos dados, como mostra a figura 6:

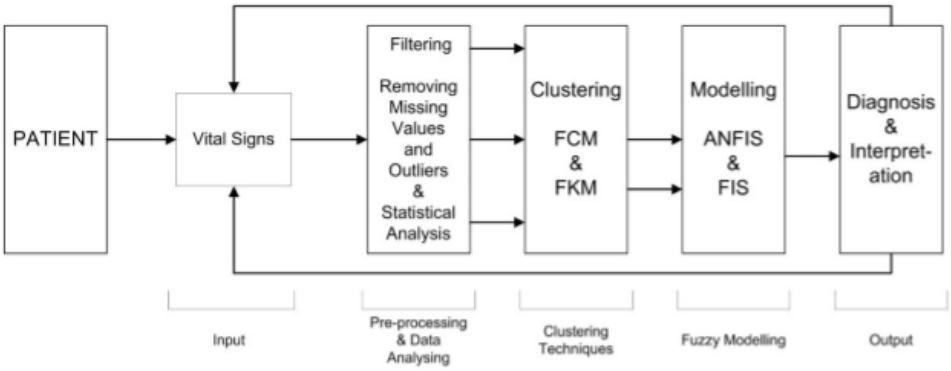

Figura 6 – Arquitetura de análise de dados.

Onde os sinais gerados passam por um pré-processamento como dito anteriormente, clusterização e um modelo fuzzy para gerar o diagnóstico como saída, a figura 7 mostra o modelo fuzzy para as sete variáveis.

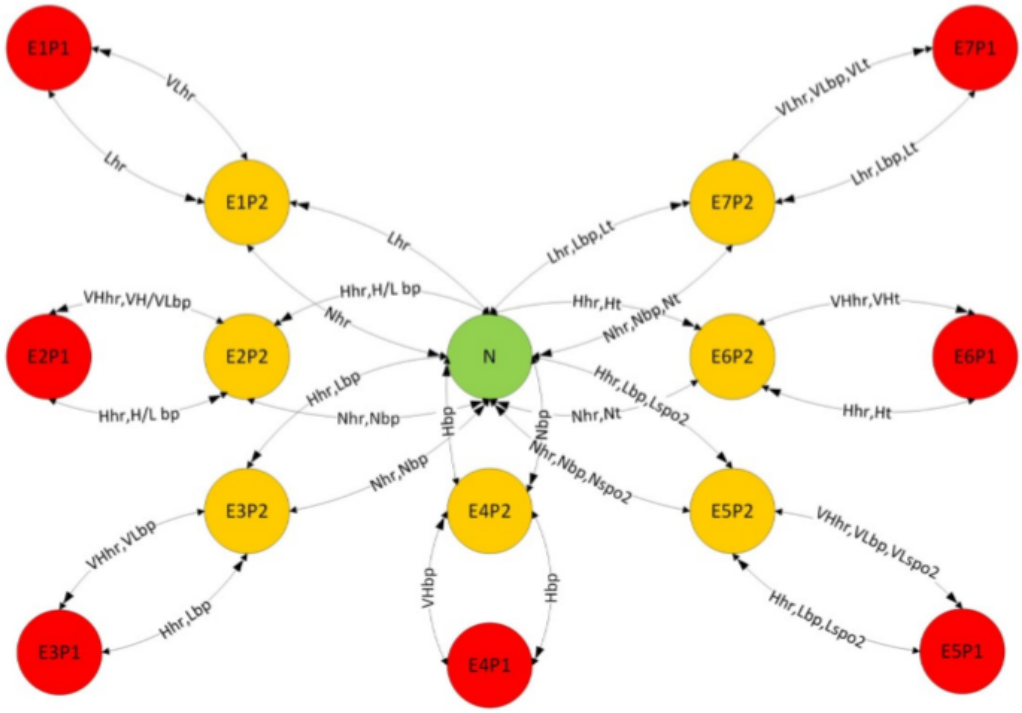

Figura 7 – Modelagem Fuzzy.

3.3 AUTOMATED ELECTRONIC MEDICAL RECORD SEPSIS DETECTION IN THE EMERGENCY DEPARTMENT

O trabalho propõe uma solução de triagem de pacientes numa emergência para determinar se o indivíduo está com suspeita, ou confirmação de sepse, através do índice de valor preditivo positivo *(PPV)* e de dois dos critérios especificados no quadro 1.

O sistema foi implementado utilizando uma ferramenta de identificação de sepsis baseada em *(EMR)*, *Cerner FirstNet* (Kansas City, Missouri), em um grande departamento de emergência acadêmico e urbano com 64 mil visitas anuais. O sistema *(EMR)* recolheu o sinal vital e a informação de teste de laboratório em todos os pacientes, desencadeando um "alerta de sepse"para aqueles com os critérios satisfeitos. Foi confirmado a presença de sepse por meio de revisão manual de médicos, enfermagem e registros laboratoriais. Também foi analisado uma seleção aleatória de casos de *(DE)* que não desencadeou um alerta de sepse. Foi avaliada a precisão diagnóstica da ferramenta de identificação de sepse.

O sistema demonstrou uma sensibilidade de 64%, PPV de 54% e valor preditivo negativo (NPV) de 99% para detectar sepsis grave com sinais de disfunção orgânica. Como esperado, esta estratégia aumentou o número de casos de sepse detectados, mas ao custo de falsos positivos positivos (diminuição de PPV). O período de testes incluem uma gama mais ampla de pacientes com ED em um período de tempo de 3 meses.

Uma diferença capital entre esse trabalho e a dissertação, é que os dados analisados pelo automated não são agrupados e analisado em conjunto. Diminuindo a precisão das análises.

4 UM SISTEMA E-HEALTH EM BIG-DATA PARA ANÁLISE E DETECÇÃO DE CHOQUE SÉPTICO EM PACIENTES ADULTOS.

Neste capítulo, é apresentada a proposta desta dissertação, Um sistema E-Health em Big-Data para análise e detecção de choque séptico em pacientes adultos, utilizando uma estrutura de análise em *Big Data* e clusterização para otimização de busca.

4.1 PRESSUPOSIÇÕES

- A geração de dados é feita através de uma base de dados relacional.
- O *Hadoop* é responsável pelo paralelismo no processamento dos dados.
- O analisador de dados irá alertar se há risco de sepse, com ciclos de 15 minutos.
- Cabe ao usuário determinar se o paciente selecionado evoluir para o estado de sepse, ou choque séptico.

4.2 VISÃO GERAL

O sistema proposto utiliza a arquitetura de processamento de dados como suporte para a análise dos dados gerados por sensores monitorando pacientes adultos. O dados coletados são frequência cardíaca, respiratória, temperatura corporal e pressão arterial sistólica. Esses dados coletados são injetados no sistema, armazenados no formato *Hadoop Distributed File System HDFS* e pré-processados para facilitar a clusterização. Após a clusterização, a cada 15 minutos o analisador irá ler os dados gerados e determinar se o paciente está com alto risco de sepse, satisfazendo dois ou mais critérios determinados na Tabela 1. Como mostra a figura 8:

Em uma visão geral, o processamento de dados em *BigData* é uma importante área de pesquisa. Em uma visão mais macro, podemos dividir o processamento de dados em 3 fases, como mostra a figura 9:

- Entrada: Armazena, coleta, transforma, pré-processa e prepara os dados para a ferramenta de mineração de dados.
- Data Mining: Realiza a mineração de dados dentre todos os dados pre-processados.
- Saída: Interpreta, e ou classifica os dados minerados

O trabalho atua no pré-processamento da massa bruta de dados da entrada, utilizando clusterização k-meios, agrupando elementos similares e utilizando o mapreduce prepara os dados pre-processados para ser analisados pelo minerador de dados, com objetivo de realizar o diagnóstico se o paciente está com sepse, ou com forte suspeita.

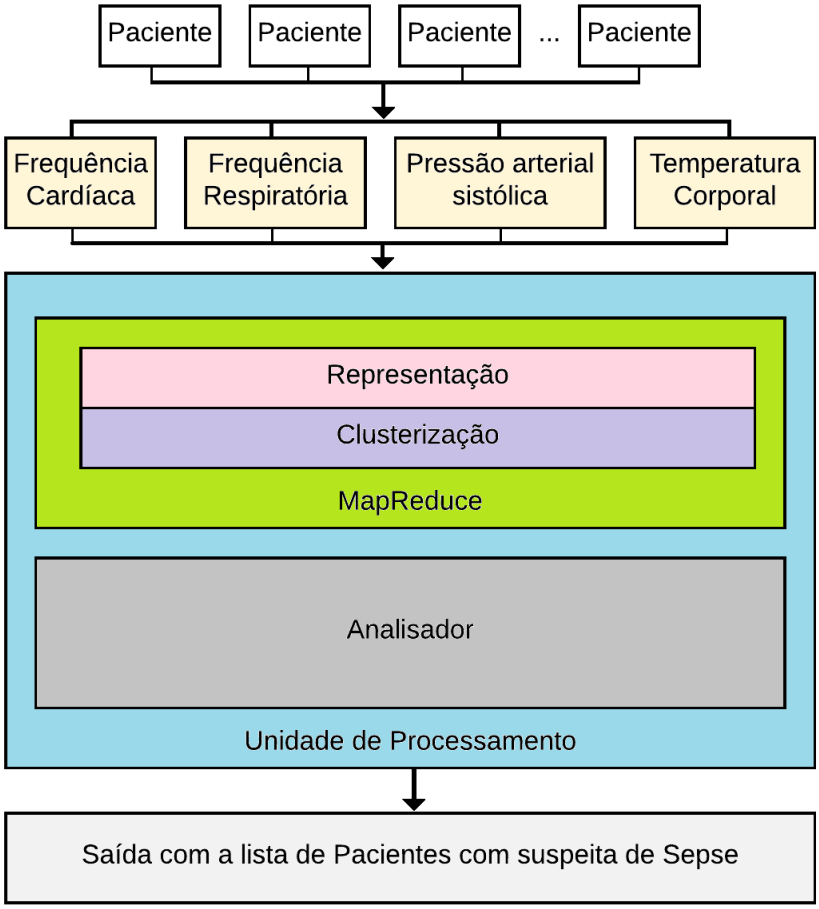

Figura 8 – Framework

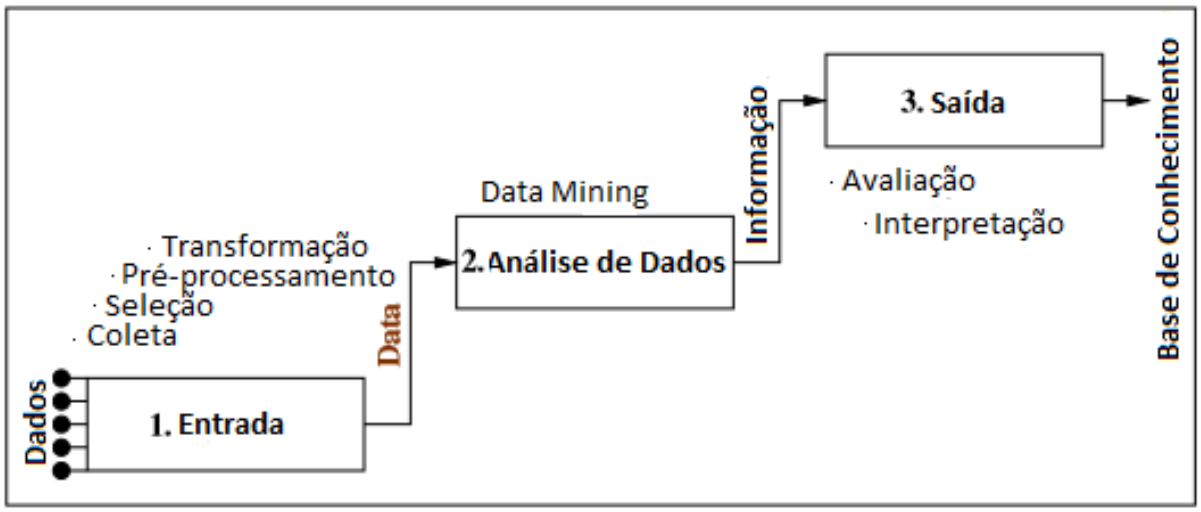

Figura 9 – Fases de processamento de dados em *Big Data*

Também será considerado a análise *batch*, ou seja, os dados necessitam estar armazenados numa base de dados relacional antes de serem introduzido na base de dados de analítica. As diferenças dessa abordagem para uma análise em tempo real (CHARDONNENS *et al.*, 2013), *stream de dados* é significativa de acordo com (HU *et al.*, 2014).

4.3 ARQUITETURA

A arquitetura da proposta está exposta na figura 10:

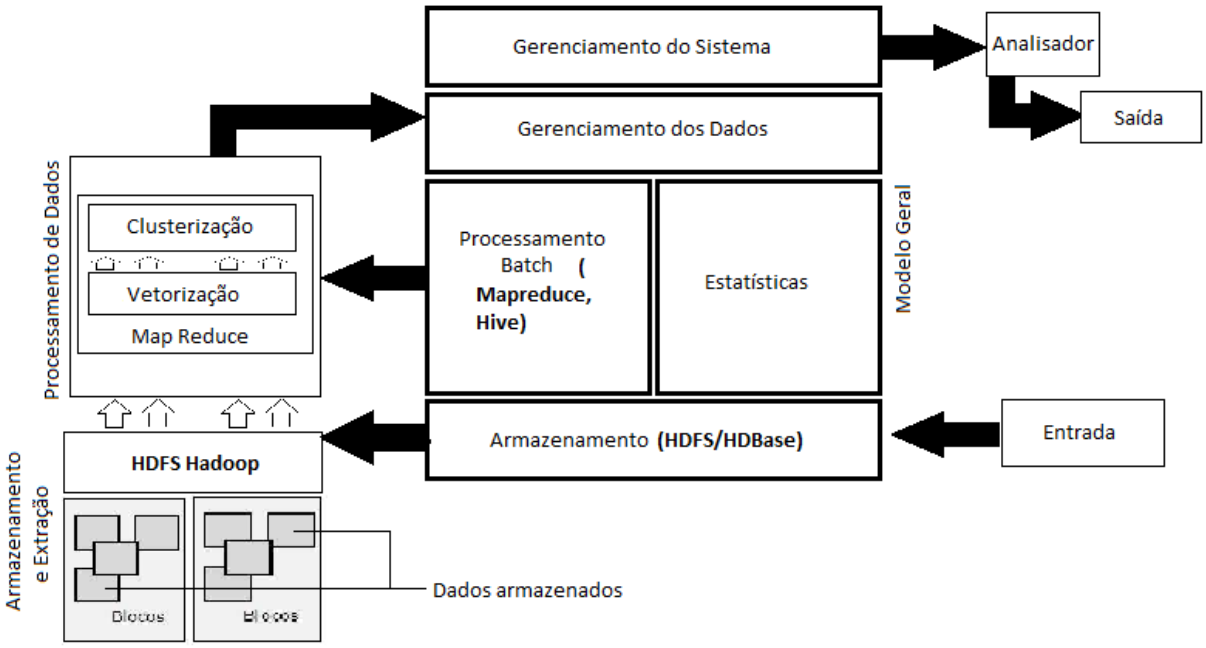

Figura 10 – Arquitetura

O tratamento da entrada de dados para mineração em *Big Data*, é feito em 3 passos: *extrair, transformar e carregar*, conhecido como arquitetura Extract, Transforming e Loading (ETL). Nesse passo, os dados de entradas são validados e descartados aqueles que forem incompletos. A entidade responsável pela extração da informação da base é gerenciada pelo mapreduce no Hadoop. Com os dados disponibilizados pelo Hadoop, começaremos a estratégia do trabalho, que segue em 3 passos:

- Representação dos Dados: Utiliza a técnica da vetorização e normalização de objetos;
- Clusterização: Agrupa elementos similares, de acordo com a correlação de pearson, em clusters, utilizando a clusterização particionada K-meios;
- Interpretação dos dados gerados: Utiliza um analisador minerador de dados simples para gerar a saída.

Com os dados vetorizados, será necessário determinar a similaridade entre os objetos para a clusterização

4.3.1 Representação dos Dados

Para obter uma boa clusterização, é necessário utilizar técnicas de vetorização, para representar os objetos como vetores. Isso permite que os algoritmos de clusterização compreendam os objetos e ajudam a calcular o valor de similaridade entre eles. Para o trabalho, utilizamos pontos num plano bidimensional agrupados com base nas distâncias entre elas. Na realidade, o agrupamento poderia ser aplicado a qualquer tipo de objeto, desde que você pode distinguir os itens semelhantes e diferentes, por exemplo, imagens podem ser agrupados com base em suas cores.

Alguns vetores com características incomuns distorcem os resultados de forma desproporcional, então será necessário normalizá-los. O processo de diminuição da magnitude de vetores grandes e aumentando a dos pequenos é chamada normalização, que é conhecida como *p-norma*. Por exemplo, a *p-norma* de um vector de 4-dimensões, [x, y, w, z], é dado 4.1:

$$p-norma = \frac{x}{(|x|^p+|y|^p+|w|^p+|z|^p)^{\frac{1}{p}}}, \frac{y}{(|x|^p+|y|^p+|w|^p+|z|^p)^{\frac{1}{p}}}, \frac{w}{(|x|^p+|y|^p+|w|^p+|z|^p)^{\frac{1}{p}}}, \frac{z}{(|x|^p+|y|^p+|w|^p+|z|^p)^{\frac{1}{p}}} \tag{4.1}$$

Onde, $(|x|^p+|y|^p+|w|^p+|z|^p)^{\frac{1}{p}}$, é a norma de um vector, o parâmetro p pode ser qualquer valor maior do que 0. No caso da norma-2 euclidiana, o p recebe valor 2. A escolha do p, depende da distância utilizada na métrica de similaridade. Pode se considerar a distância euclidiana d de dois elemento E_i e E_j no espaço dimensional p 4.2:

$$d(E_i,E_j) = \left[\sum_{k=1}^{p}(x_ik - x_jk)^2\right]\frac{1}{2} \tag{4.2}$$

4.3.2 Clusterização

Com os dados disponíveis, há a necessidade de agrupá-los. Há vários trabalhos direcionados nessa área, separar objetos em clusters é um problema complexo. Dentre todos os tipos de clusterização, a escolhida para o trabalho é a particionada sem sobreposição.

Quando está clusterizando um conjunto de dados, escolher o número k de clusters é um desafio. A atribuição do número de clusters é um problema NP-completo e (DEMPSTER; LAIRD; RUBIN, 1977; HAMERLY; ELKAN, 2004) apresentam um algoritmo melhorado para determinar o número de clusters k utilizando o método de expectativa máxima gaussiana . O

algoritmo descobre um k apropriado, usando um teste estatístico para decidir se vai dividir um centro k-meios em dois centros. Ao invés de utilizarmos a expectativa máxima gaussiana, analisaremos a clusterização, de acordo com a função de distorção.

Clusterização é usado para encontrar irregularidades na distribuição de dados e identificar as regiões em que os objetos estão mais concentrados. No entanto, nem todas as regiões com uma elevada concentração de objetos é considerado um cluster. Para uma região ser identificada como um cluster, é importante analisar não só a sua distribuição interna, mas também sua interdependência com outros agrupamentos de objetos no conjunto de dados.

Em K-meios, a distorção de um cluster é uma função da população de dados e a distância entre objetos e centro do cluster, adaptando essa função a correlação de Pearson, gera função de distorção 4.3:

$$D_i = \sum_{t=1}^{N_i} ([d(k_i t, w_i)]^2 * (1 - \frac{\sum_{j=1}^{n}(x_j - \bar{x}) * (y_j - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2} * \sqrt{\sum_{j=1}^{n}(y_j - \bar{y})^2}})) \tag{4.3}$$

onde:

- D_i é a distorção do cluster i;
- w_i é o centroide do cluster i;
- N_i é a quantidade de objetos pertencentes ao cluster i;
- $k_i t$ é o t-ésimo objeto pertencente ao cluster i;
- $d(k_i t, w_i)$ é a distância entre o objeto $x_i t$ e o centro do cluster w_j;
- $(x_1, x_2, ..., x_n)$ e $(y_1, y_2, ..., y_n)$ são as séries de valores;
- $\bar{x}$ e $\bar{y}$, são as médias das medidas do conjunto X e Y;
- $\sum_{j=1}^{n}(x_j - \bar{x}) * (y_i - \bar{y})$ é a cov(X, Y);
- $\sum_{j=1}^{n}(x_j - \bar{x})^2$ é a var(X);
- $\sum_{j=1}^{n}(y_j - \bar{y})^2$ é a var(Y);

Cada cluster é representado por sua distorção e o seu impacto sobre a totalidade do conjunto de dados, e é avaliada pela sua contribuição para a soma de todas as distorções 4.4:

$$S_k = \sum_{i=1}^{K} D_i; \tag{4.4}$$

onde K é o número de clusters. Essas funções estão sujeitas a uma função de avaliação $F(K)$ (TIBSHIRANI; WALTHER; HASTIE, 2001):

$$F_K = \begin{cases} 1, & K = 1 \\ \frac{S_k}{\alpha_k * S_k - 1}, & S_k - 1 \neq 0, \forall K > 1 \\ 1, & S_k - 1 = 0, \forall K > 1 \end{cases} \tag{4.5}$$

$$\alpha_K = \begin{cases} 1 - \frac{3}{4N_d}, & K = 2eN_d > 1 \\ \alpha_K - 1 + \frac{1-\alpha_K - 1}{6}, & K > 2eN_d > 1 \end{cases} \tag{4.6}$$

onde, α_K é um fator de peso aplicado para reduzir os efeitos das dimensões. Os resultados da função $F(K)$ são esperados valore iguais a 1, é considerado uma perturbação quando $F(K) < 0.85$ (PHAM; DIMOV; NGUYEN, 2005).

4.3.2.1 Funcionamento

Inicialmente toda a massa de dados não está vinculada a nenhum cluster, logo é selecionado um ponto aleatoriamente para ser o cluster-head (PHAM; DIMOV; NGUYEN, 2005). A cada novo cluster-head inserido no conjunto de dados, é calculado o valor da distorção em cada cluster-head e o valor da função de avaliação resultante. Se o valor da função de avaliação for menor que um certo limiar, então há necessidade de inserir um novo cluster-head. Se a inserção de um elemento não gera perturbação no cluster corrente, ou seja o elemento causa pouca distorção ao conjunto, conclui-se que esse elemento pertence ao cluster.

O funcionamento do algoritmo 2 é exemplificado na figura 11.

Sendo determinado a expectativa de número de clusters, agregado ao algoritmo original k-means, Podemos determinar o algoritmo 2.

No sistema utilizado, a clusterização é feita em conjunto com o *mapreduce*. Determinado os clusters, vamos descrever como os dados estarão dispostos para o agente analisador de mineração de dados.

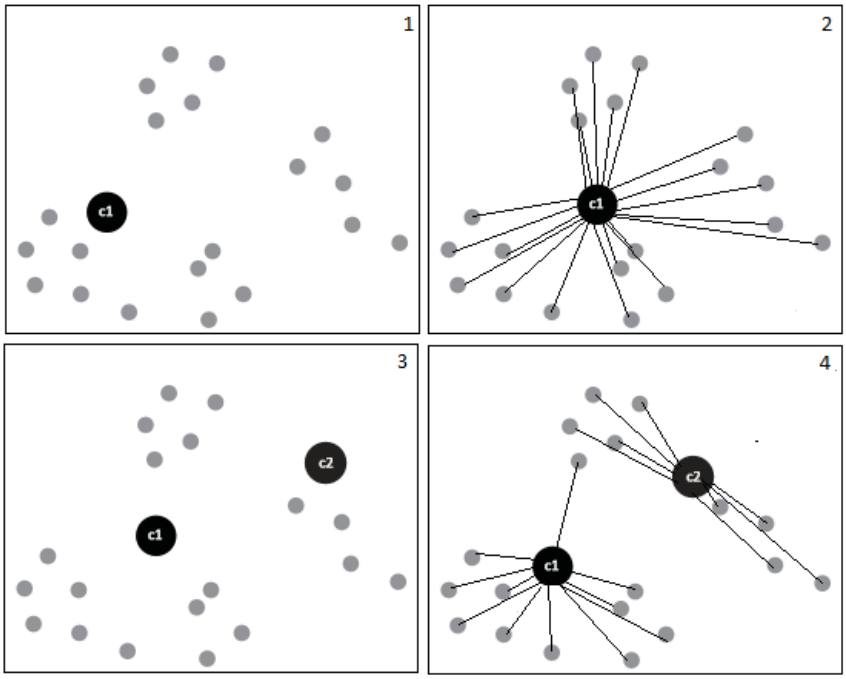

Figura 11 – Funcionamento K-means Proposto.

Algoritmo 2: Algoritmo Kmeans with clustering modificado. K-means-C(X, α)

```
Faça C o conjunto de clusters iniciais(C = x̄);
while Não Convergiu do
    C = K-means(C, X);
    Faça [x_i|classe(x_i) = j] o conjunto dos pontos atribuidos ao cluster c_j;
    Usa a função F(K) para detectar se há perturbações;
    if F(K) < 0.85 then
        Particiona o cluster c_j em dois outros clusters;
        InsereNovoCluster(Cluster com a maior distorção);
    else
        Mantém c_j como cluster;
        Avaliar se a solução converge;
    end
end
```

4.3.3 Analisador de Dados

O analisador de dados trabalha com uma heurística simples de lógica de primeira ordem. Ele é responsável por a cada 15 minutos fazer uma verificação na saída gerada pela unidade de processamento. Para cada cluster, o analisador lê um elemento randomicamente do cluster e determina a qual classe pertence, de acordo com os critérios de identificação de sepse exposto no Quadro 1. No Quadro 2 todas as classes possíveis.

As classes que não satisfazem os critérios de buscas, são aquelas que tem o somatório da linha menor que dois, ou seja, classes que não satisfazem o critério de sepse descrito

Quadro 2 – Classes correspondentes.

Classe	Temperatura Corporal > 38^oC ou < 36^oC	Frequência respiratória > $20mpm$	Frequência Cardíaca > $90bpm$	Pressão arterial sistólica < $90mmHg$ ou média < $65mmHg$
1	0	0	0	0
2	0	0	0	1
3	0	0	1	0
4	0	0	1	1
5	0	1	0	0
6	0	1	0	1
7	0	1	1	0
8	0	1	1	1
9	1	0	0	0
10	1	0	0	1
11	1	0	1	0
12	1	0	1	1
13	1	1	0	0
14	1	1	0	1
15	1	1	1	0
16	1	1	1	1

Fonte – Elaborado pelo autor

anteriormente.

O algoritmo 3 representa o pseudo código para execução do analisador, os valores que ele retorna são os ID's dos pacientes.

Algoritmo 3: Algoritmo do Analisador

```
Faça N o número total de clusters e (k = 0);
while (N > 0) do
    ElementoTeste = CentróideDoCluster(k, número de elementos(Cluster_k));
    Sum = VerificarParâmetros(ElementoTeste);
    if Sum >= 2 then
        for Cada elemento no Cluster_k do
            Retorna todos os ID's não repetidos
        end
    else
        Ignora cluster;
        N = N − 1;
        k = k + 1;
    end
end
```

4.4 SÍNTESE DO CAPÍTULO

Nesse capítulo foi apresentado a técnica utilizada para o trabalho de dissertação. A técnica é feita nos passos aqui explicados, primeiramente foi exposto como representar os dados na matriz de elementos, então como clusterizar esses dados para preparar os dados para o agente minerador de análise de dados.

5 SIMULAÇÕES E RESULTADOS

Para os testes da proposta, foi desenvolvido um sistema para simular a inserção dos dados, utilizando a base de dados Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) (SAEED *et al.*, 2002) e o *Sqoop* (TING; CECHO, 2013) para a inserção desses dados para o HDFS.

5.1 AMBIENTE

Foi utilizado o Hadoop (SHVACHKO *et al.*, 2010) e a implementação do *mapreduce* para os testes. Para a base de dados inicial, foi utilizada a base de dados MIMIC II.

5.2 CENÁRIO

A figura 12 mostra a arquitetura utilizada nos testes.

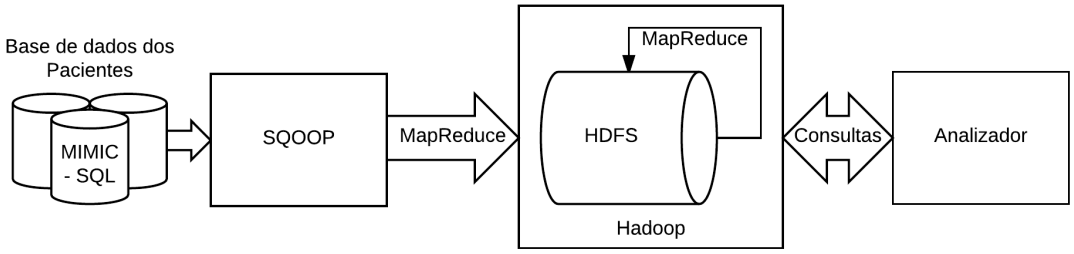

Figura 12 – Arquitetura do Cenário de teste.

A base de dados MIMIC II, foi convertida para Structured Query Language (SQL) externo ao sistema proposto. A comunicação e injeção dos dados do banco de dados relacional SQL externo para dentro do sistema foi feita utilizando o Sqoop. Os dados inseridos no HDFS estão ordenados por tempo, assim facilitando a limpeza de dados incompletos na base.

Com os dados inseridos no HDFS, o sistema realiza o algoritmo do trabalho proposto através da implementação do *mapreduce*, um analisador de dados externo extrai o arquivo de saída e através da análise determina o resultado obtido.

Foram feitos cinquenta testes para cada métrica, totalizando cento e cinquenta testes no geral.

5.3 PARÂMETROS DOS TESTES

Todos os testes foram realizados utilizando o Hadoop (SHVACHKO *et al.*, 2010). Os parâmetros dos testes estão sumarizados no quadro 3. O número de pontos é obtido após a limpeza de todos os registros, eliminando os pontos com dados incompletos. A quantidade de clusters esperados para o número de pontos obtidos, variam de 50 a 5000, de acordo com (PELLEG; MOORE *et al.*, 2000).

Quadro 3 – Sumário dos parâmetros de simulação.

	Parâmetros
Número de Pacientes	4.000
Número Total de Registros	3.537.582
Número de Pontos	805.609
Número máximo de clusters esperados	5000
Número minimo de clusters	50
Falso Negativo	91%
Falso Positivo	81% - 98%
Intervalo de confiança	95%

Fonte – Elaborado pelo autor

5.4 MÉTRICAS

Para a avaliação dos resultados são consideradas as seguintes métricas:

- Quantidade de Falso Positivo (Especificidade);
- Quantidade de Falso Negativo (Sensibilidade);
- Precisão na análise dos dados;

Os parâmetros e métricas utilizados nos testes são descritas nas seções anteriores. O cenário padrão utilizado é a arquitetura proposta, que são os mesmos utilizados para expor os resultados obtidos. O critério de avaliação do resultado é inspirado na analise Kappa (SIM; WRIGHT, 2005).

A análise Kappa é uma área da estatística que lida com julgamentos categóricos. São colocados objetos em categorias que não precisam ter nenhuma ordem inerente entre eles, e os mesmos são convertidos em um número que representa a concordância entre dois ou mais avaliadores. Essa medida fornece a ideia do quanto as observações obtidas se afastam daquelas esperadas, indicando-nos assim o quão legítimas as interpretações são.

Kappa destina-se a dar ao leitor uma medida quantitativa da magnitude de concordância entre dois ou mais observadores, iremos comparar com o trabalho *Automated electronic medical record sepsis detection in the emergency department* (NGUYEN *et al.*, 2014). O tra-

balho propõe uma ferramenta de identificação de sepse no Departamento de Emergência (DE), utilizando Electronic Medical Record (EMR). O sistema do trabalho (NGUYEN *et al.*, 2014), aciona alerta de sepse quando os dados coletados satisfazem a dois ou mais dos critérios de SIRS e pelo menos um sinal de choque. Os critérios de SIRS são os mesmos propostos no quadro 1. Para análise Kappa, os índices de concordância positiva (P_{pos}) e a negativa (P_{neg}) são definidos por:

$$P_{pos} = \frac{VP+VP}{(VP+FP)+(VP+FN)} \tag{5.1}$$

$$P_{neg} = \frac{VN+VN}{(VN+FP)+(VN+FN)} \tag{5.2}$$

Onde:

- VP = Verdadeiro Positivo;
- FP = Falso Positivo;
- VN = Verdadeiro Negativo;
- FN = Falso Negativo.

O índice Kappa (K) é calculado como:

$$K = \frac{P_0 - P_e}{1 - P_e} \tag{5.3}$$

onde:

P_0 é a taxa de aceitação relativa, dada por:

$$P_0 = \frac{VP+VN}{\sum(VP+VN+FP+FN)} \tag{5.4}$$

P_e é a taxa hipotética de aceitação, dada por:

$$P_e = \frac{(VP+FP)*(VP+FN)+(FN+VN)*(FP+VN)}{\sum(VP+VN+FP+FN)} \tag{5.5}$$

O erro padrão de K é dado por:

$$EP = \sqrt{\frac{P_0(1-P_0)}{n(1-P_e)^2}} \tag{5.6}$$

Como descrito em (BAIG, 2014), o intervalo de confiança de 95%(IC) para K pode ser obtido por:

$$IC = K \pm 1.96 * EP \tag{5.7}$$

As diretrizes de interpretação de K são dadas no quadro 4 seguinte:

Quadro 4 – Medidas de concordância observada para dados categóricos (LANDIS; KOCH, 1977)

Valor de K	Interpretação
< 0	Sem concordância
0 - 0.20	Concordância Mínima
0.21 - 0.40	Concordância Razoável
0.41 - 0.60	Concordância Moderada
0.61 - 0.80	Concordância Substancial
0.81 - 1.00	Concordância Perfeita

Fonte – Elaborado pelo autor

As categorias quantitativas, como precisão e sensibilidade podem ser obtidos por:

$$Precisao = \sum (\frac{VP + VN}{VP + VN + FP + FN}) \tag{5.8}$$

$$Sensibilidade = \frac{VP(Alarms)}{VP(Alarms) + FN(Alarms)} \tag{5.9}$$

$$Especificidade = \frac{VN(Alarms)}{VN(Alarms) + FP(Alarms)} \tag{5.10}$$

5.5 RESULTADOS

Nessa seção, serão analisados os resultados obtidos pelos testes e compará-los com os protocolados e registrados pela base de dados de teste.

O sistema proposto foi projetado como um dispositivo de monitoramento auxiliar, que visa registrar, monitorar, detectar, interpretar e diagnosticar pacientes que podem ter alta probabilidade de evoluir para o estado de sepse durante a internação hospitalar através da análise dos sinais vitais.

A avaliação do sistema proposto foi direcionada para três aspectos principais:

- Armazenamento e monitoramento de sinais vitais gerados por sensores refletido na sensibilidade e especificidade do sistema proposto, ou seja, os problemas devem ser identificados e classificados corretamente, quando for gerado um alarme, ou aviso em resposta a uma condição detectada.
- Interpretação dos sinais vitais através de filtros, de acordo com as normas padrões para identificação de sepse em pacientes.
- Otimizar as análises dos dados assim diminuindo o tempo de classificação dos dados para diminuir a taxa de mortalidade.

A Figura 12 mostra a arquitetura do sistema proposto e o fluxo dos dados. O objetivo do projeto é utilizar a infraestrutura *Big Data*, juntamente com um algoritmo de clusterização, coletar os parâmetros fisiológicos (pressão arterial, frequência cardíaca, frequência respiratória e temperatura) para realizar o diagnóstico precoce de sepse. A formação dos clusters é realizado por uma variação do k-means incremental, conforme mostrado na Figura 13.

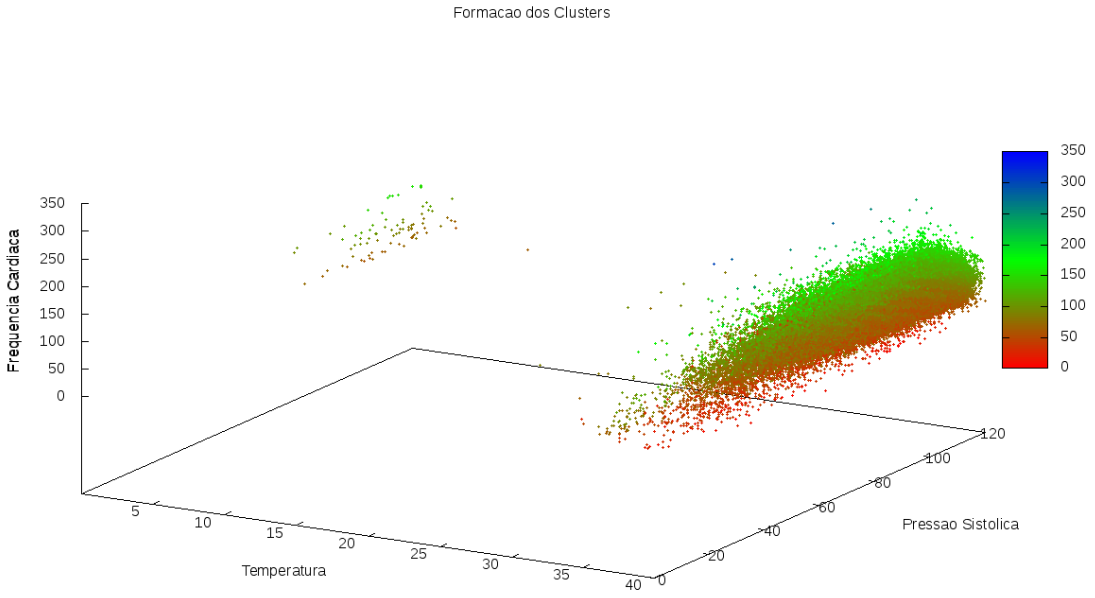

Figura 13 – Formação dos Clusters

Para detecção de sepse é necessário outros dois parâmetros que estão intrínsecos aos parâmetros base para auxílio da interpretação e determinação do diagnóstico. São eles, a pressão arterial média e a frequência respiratória. Devido a difícil visualização de gráficos multidimensionais, a figura 14 mostra a formação dos clusters em todas as dimensões utilizadas, ou seja todas as combinações dois a dois das variáveis do sistema.

O banco de dados utilizado contêm sinais fisiológicos e sinais vitais de séries

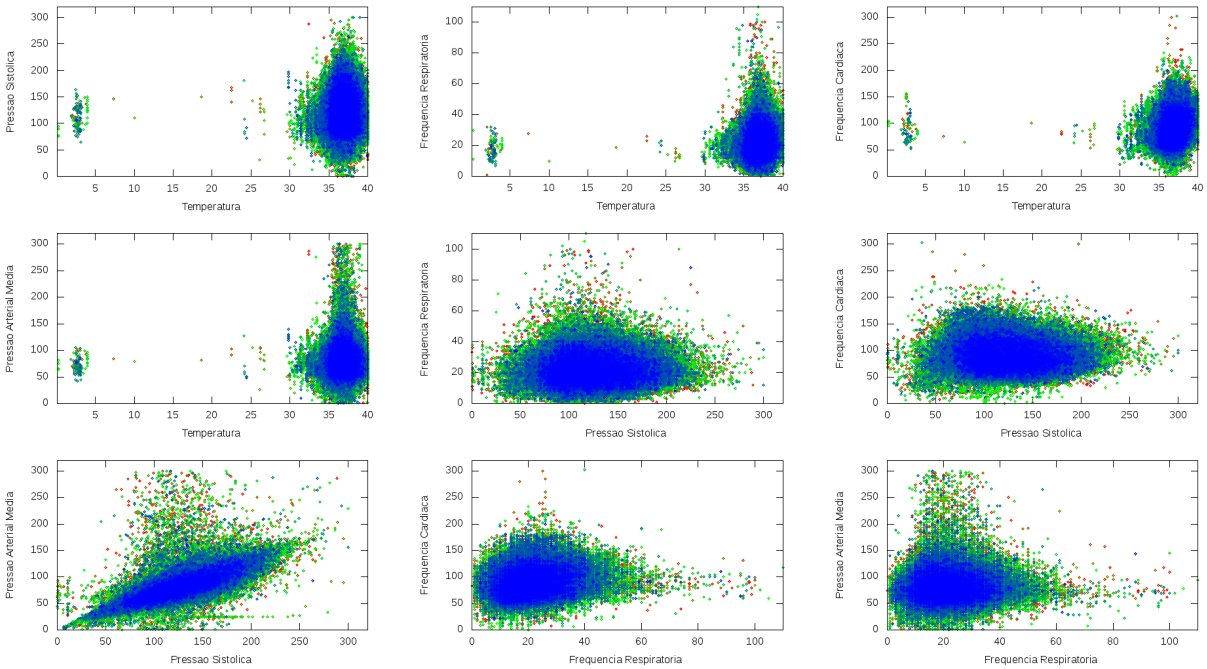

Figura 14 – Formação dos clusters em todas as dimensões utilizadas

temporais capturados por monitores de pacientes e dados clínicos abrangentes obtidos de sistemas de informações médicas hospitalares de pacientes na UTI. Os dados foram coletados entre 2001 e 2008 de uma variedade de UTI (atendimento médico, cirúrgico, coronário e neonatal) em um único hospital de ensino superior.

A conversão de dados, leitura de arquivos e módulos de pré-processamento convertem os dados recebidos para um formato legível. Para os testes são 3.537.582 registros de 4000 pacientes. Os mesmos após a fase de limpeza da base, são condensados em 805.609 pontos.

Após a clusterização, o sistema vai analisar os clusters no qual os centros satisfazem as condições, conforme a Tabela 1 no capítulo 2, para determinar se um paciente está evoluindo para sepse. Os dados analisados são reduzidos a 7,76% do número total de pontos.

Os resultados obtidos pelo sistema proposto utilizando a base de dados MIMIC II para detecção de sepse são mostrados na Tabela 2. Os resultados são comparados com um algoritmo similar de monitoramento de pacientes (NGUYEN *et al.*, 2014) para avaliação de precisão e sensibilidade na análise dos dados.

A precisão atingida foi de 91,99% dos dados analisados com sensibilidade de 76,88% e especificidade de 93,85%, isso significa que os teste feitos pelo sistema proposto são de alta consistência nos resultados obtidos com a capacidade de diagnosticar indivíduos verdadeiramente positivos com taxa de 76,88%, ou seja indivíduos doentes, e verdadeiramente negativos com taxa de 93,85%, ou seja indivíduos sadios.

Como dito anteriormente, o índice kappa é uma maneira muito utilizada para expres-

Tabela 2 – Resultados do sistema proposto comparando com o proposto em (NGUYEN *et al.*, 2014)

Alarms	Sistema Proposto	Automated
VP	8.061	1.058
VN	49.517	492.522
FP	2.586	288.918
FN	2.423	23.111
Total	62.587	805.609
Precisão(%)	91,996%	61,267%
Sensibilidade(%)	76,888%	4,377%
Especificidade(%)	93,855%	63,027%

Fonte – Elaborado pelo autor

sar a confiabilidade de um teste. Os índices de concordância e proporção serão obtidos com os resultados positivos e negativos reais. A taxa geral de concordância do teste foi de 91,85%, obtendo a taxa Kappa K_a = 0.704, que significa, de acordo com o Quadro 4, um resultado substancial. Assim podemos gerar os índices de concordância que estão no Quadro 5. A taxa de aceitação relativa é a proporção de pacientes doentes entre os testes positivos encontrados, no caso dos teste a taxa de aceitação relativa obtida é de 75,3%, ou seja, a cada 100 registros 75 são realmente positivos. No caso da taxa hipotética de aceitação, é a proporção de pacientes sadios que estão entre os negativos obtidos, no caso dos testes é de 95,1%, ou seja, a cada 100 testes negativos 95 seriam pacientes sadios. O erro padrão é de 0.015 fazendo assim o intervalo de confiança de 95% para K variar de 67,3% a 73,5%.

Quadro 5 – Resultados da análise Kappa e dos índices de concordância

Taxa de aceitação relativa	Taxa hipotética de aceitação	Índice de concordância Positiva	Índice de concordância Negativa	Erro Padrão	Intervalo de Confiança 95% para K
P_o	P_e	P_{pos}	P_{neg}	EP	$IC_{95\%}$
0.918	0.724	0.753	0.951	0.015	0.673 - 0.735

Fonte – Elaborado pelo autor

5.6 SÍNTESE DO CAPÍTULO

Nesse capítulo foi apresentado os resultados obtidos pelo trabalho de dissertação, cenário utilizado, parâmetros e um método de avaliação de diagnósticos visando classificar o sistema proposto.

6 CONCLUSÕES

Este trabalho teve como objetivo principal elaborar uma infraestrutura de análise de dados em Big Data para E-health, visando elevar a sensibilidade e especificidade no diagnóstico de sepse em pacientes monitorados por sensores eletrônicos, assim aumentando a precisão, sensibilidade e especificidade dos dados analisados. Para desenvolvê-lo foi proposto uma técnica de clusterização incremental que tem como objetivo diminuir o espaço de busca das soluções.

Problemas de clusterização são NP-completo por natureza, a solução aqui apresentada escolhe por meio otimizado, de acordo com a distribuição dos dados na base, o menor número de clusters necessários para maximizar o desempenho do agrupamento de maneira eficiente. A grande contribuição do trabalho é prover uma infraestrutura Big Data clusterizando a grande quantidade de dados, de forma eficiente e otimizada.

Foi estudado e aplicado um algoritmo incremental de seleção de clusters (PHAM; DIMOV; NGUYEN, 2005) para determinar a formação dos agrupamentos, o algoritmo alivia o problema com mínimos locais, permitindo que os centros de cluster se movam de forma tão radical quanto para reduzir a distorção geral do cluster. No entanto, o método é muito sensível aos erros na estimativa de distorção.

Os experimentos mostram que o trabalho de dissertação proposto consegue ter um desempenho satisfatório tanto na precisão dos diagnósticos, quando na sensibilidade e especificidade, segundo (IMHOFF; KUHLS, 2006).

6.1 TRABALHOS FUTUROS

Para trabalhos futuros, são feitas duas sugestões que podem trazer melhoras significativas. A primeira se trata de aplicar um algoritmo mais inteligente para o agente analisador, assim ele pode ser bem mais preciso que o atual. A segunda sugestão é expandir o sistema para análise e diagnóstico para outras doenças, assim o sistema ficaria mais robusto e completo, e poderia ser utilizado como um grande framework na área da saúde. Tentar aplicar uma técnica para aumentar a sensibilidade do sistema, diminuindo a taxa de falsos negativos.

REFERÊNCIAS

AL-DAOUD, M.; VENKATESWARLU, N.; ROBERTS, S. **Fast K-means clustering algorithms**. [S.l.]: University of Leeds, School of Computer Studies, 1995.

ALBERTI, C.; BRUN-BUISSON, C.; GOODMAN, S. V.; GUIDICI, D.; GRANTON, J.; MORENO, R.; SMITHIES, M.; THOMAS, O.; ARTIGAS, A.; GALL, J. R. L. Influence of systemic inflammatory response syndrome and sepsis on outcome of critically ill infected patients. **American journal of respiratory and critical care medicine**, Am Thoracic Soc, v. 168, n. 1, p. 77–84, 2003.

ANSHARI, M.; ALAS, Y. Big data era: Big challenges for asean economic community. In: **International Conference on Asean Studies**. [S.l.: s.n.], 2015. v. 2, p. 3–5.

BAIG, M. M. **Smart Vital Signs Monitoring and Novel Falls Prediction System for Older Adults**. Tese (Doutorado) — Auckland University of Technology, 2014.

BERKHIN, P. A survey of clustering data mining techniques. In: **Grouping multidimensional data**. [S.l.]: Springer, 2006. p. 25–71.

BILMES, J.; VAHDAT, A.; HSU, W.; IM, E.-J. Empirical observations of probabilistic heuristics for the clustering problem. **Technical Report TR-97-018, International Computer Science Institute**, Citeseer, 1997.

BOTTOU, L.; BENGIO, Y. *et al.* Convergence properties of the k-means algorithms. **Advances in neural information processing systems**, MORGAN KAUFMANN PUBLISHERS, p. 585–592, 1995.

BRADLEY, P. S.; FAYYAD, U. M. Refining initial points for k-means clustering. In: CITESEER. **ICML**. [S.l.], 1998. v. 98, p. 91–99.

CHARDONNENS, T.; CUDRE-MAUROUX, P.; GRUND, M.; PERROUD, B. Big data analytics on high velocity streams: A case study. In: IEEE. **Big Data, 2013 IEEE International Conference on**. [S.l.], 2013. p. 784–787.

DELLINGER, R. P.; LEVY, M. M.; RHODES, A.; ANNANE, D.; GERLACH, H.; OPAL, S. M.; SEVRANSKY, J. E.; SPRUNG, C. L.; DOUGLAS, I. S.; JAESCHKE, R. *et al.* Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock, 2012. **Intensive care medicine**, Springer, v. 39, n. 2, p. 165–228, 2013.

DEMCHENKO, Y. The big data architecture framework (bdaf). **Outcome of the Brainstorming Session at the University of Amsterdam**, v. 17, p. 1494–1499, 2013.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the royal statistical society. Series B (methodological)**, JSTOR, p. 1–38, 1977.

DOVAL, D.; MANCORIDIS, S.; MITCHELL, B. S. Automatic clustering of software systems using a genetic algorithm. In: IEEE. **Software Technology and Engineering Practice, 1999. STEP'99. Proceedings**. [S.l.], 1999. p. 73–81.

DUMBILL, E. **Planning for big data**. [S.l.]: "O'Reilly Media, Inc.", 2012.

ENGEL, C.; BRUNKHORST, F. M.; BONE, H.-G.; BRUNKHORST, R.; GERLACH, H.; GROND, S.; GRUENDLING, M.; HUHLE, G.; JASCHINSKI, U.; JOHN, S. *et al.* Epidemiology of sepsis in germany: results from a national prospective multicenter study. **Intensive care medicine**, Springer, v. 33, n. 4, p. 606–618, 2007.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Kdd**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

EYSENBACH, G. What is e-health? **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 3, n. 2, p. e20, 2001.

FASULO, D. **An analysis of recent work on clustering algorithms**. [S.l.], 1999.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

FELICE, C. D.; SUSIN, C. F.; COSTABEBER, A. M.; RODRIGUES, A. T.; BECK, M.; HERTZ, E. Choque: diagnóstico e tratamento na emergência. **Revista da AMRIGS**, v. 55, n. 2, p. 179–196, 2011.

FISHER, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. **Biometrika**, JSTOR, v. 10, n. 4, p. 507–521, 1915.

FREY, B. J.; DUECK, D. Clustering by passing messages between data points. **science**, American Association for the Advancement of Science, v. 315, n. 5814, p. 972–976, 2007.

GAYEN, A. The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. **Biometrika**, JSTOR, v. 38, n. 1/2, p. 219–247, 1951.

GUHA, S.; RASTOGI, R.; SHIM, K. Cure: an efficient clustering algorithm for large databases. In: ACM. **ACM SIGMOD Record**. [S.l.], 1998. v. 27, n. 2, p. 73–84.

HAMERLY, G.; ELKAN, C. Learning the k in a> means. **Advances in neural information processing systems**, The MIT Press, v. 16, p. 281, 2004.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011.

HANSEN, L. K.; LARSEN, J.; NIELSEN, F. Å.; STROTHER, S. C.; ROSTRUP, E.; SAVOY, R.; LANGE, N.; SIDTIS, J.; SVARER, C.; PAULSON, O. B. Generalizable patterns in neuroimaging: How many principal components? **NeuroImage**, Elsevier, v. 9, n. 5, p. 534–544, 1999.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, JSTOR, v. 28, n. 1, p. 100–108, 1979.

HU, H.; WEN, Y.; CHUA, T.-S.; LI, X. Toward scalable systems for big data analytics: A technology tutorial. **IEEE Access**, IEEE, v. 2, p. 652–687, 2014.

IMHOFF, M.; KUHLS, S. Alarm algorithms in critical care monitoring. **Anesthesia & Analgesia**, LWW, v. 102, n. 5, p. 1525–1537, 2006.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis**. [S.l.]: John Wiley & Sons, 2009. v. 344.

KOTHARI, R.; PITTS, D. On finding the number of clusters. **Pattern Recognition Letters**, Elsevier, v. 20, n. 4, p. 405–416, 1999.

KU-MAHAMUD, K. R. Big data clustering using grid computing and ant-based algorithm. In: **Proceedings of the International Conference on Computing and Informatics**. [S.l.: s.n.], 2013. p. 6–14.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, JSTOR, p. 159–174, 1977.

LATHAM, H. E.; BENGTSON, C. D.; SATTERWHITE, L.; STITES, M.; SUBRAMANIAM, D. P.; CHEN, G. J.; SIMPSON, S. Q. Stroke volume guided resuscitation in severe sepsis and septic shock improves outcomes. **Journal of critical care**, Elsevier, v. 42, p. 42–46, 2017.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C.; BYERS, A. H. Big data: The next frontier for innovation, competition, and productivity. 2011.

MARSHALL, J. C.; VINCENT, J.-L.; GUYATT, G.; ANGUS, D. C.; ABRAHAM, E.; BERNARD, G.; BOMBARDIER, C.; CALANDRA, T.; JØRGENSEN, H. S.; SYLVESTER, R. *et al.* Outcome measures for clinical research in sepsis: a report of the 2nd cambridge colloquium of the international sepsis forum. **Critical care medicine**, LWW, v. 33, n. 8, p. 1708–1716, 2005.

MCCALLUM, A.; NIGAM, K.; UNGAR, L. H. Efficient clustering of high-dimensional data sets with application to reference matching. In: ACM. **Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2000. p. 169–178.

MORRELL, M. R.; MICEK, S. T.; KOLLEF, M. H. The management of severe sepsis and septic shock. **Infectious disease clinics of North America**, Elsevier, v. 23, n. 3, p. 485–501, 2009.

NGUYEN, S. Q.; MWAKALINDILE, E.; BOOTH, J. S.; HOGAN, V.; MORGAN, J.; PRICKETT, C. T.; DONNELLY, J. P.; WANG, H. E. Automated electronic medical record sepsis detection in the emergency department. **PeerJ**, PeerJ Inc., v. 2, p. e343, 2014.

PELLEG, D.; MOORE, A. W. *et al.* X-means: Extending k-means with efficient estimation of the number of clusters. In: **ICML**. [S.l.: s.n.], 2000. v. 1, p. 727–734.

PHAM, D. T.; DIMOV, S. S.; NGUYEN, C. Selection of k in k-means clustering. **Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science**, SAGE Publications, v. 219, n. 1, p. 103–119, 2005.

RASMUSSEN, E. M. Clustering algorithms. **Information retrieval: data structures & algorithms**, v. 419, p. 442, 1992.

SABIA; ARORA, L. Technologies to handle big data: A survey. **ICCCS - International Conference on Communication, Computing & Systems**, 2014.

SAEED, M.; LIEU, C.; RABER, G.; MARK, R. G. Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In: IEEE. **Computers in Cardiology, 2002**. [S.l.], 2002. p. 641–644.

SALLES, M.; SPROVIERI, S.; BEDRIKOW, R.; PEREIRA, A.; CARDENUTO, S.; AZEVEDO, P.; SILVA, T.; GOLIN, V. Síndrome da resposta inflamatória sistêmica/sepse 3/4 revisão e estudo da terminologia e fisiopatologia. **Revista da Associação Médica Brasileira**, SciELO Brasil, v. 45, n. 1, p. 86–92, 1999.

SAMPLE, C. D. F. A. S. correlation coefficients covering the cases (i)“the frequency dis-tribution of the values of the correlation coefficient in samples from an indefinitely large population,” biometrika, vol. 10, pp. 507v521, 1915. here the method of defining the sample by the coordinates of. 1921.

SHVACHKO, K.; KUANG, H.; RADIA, S.; CHANSLER, R. The hadoop distributed file system. In: IEEE. **Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on**. [S.l.], 2010. p. 1–10.

SILVA, E.; PEDRO, M. de A.; SOGAYAR, A. C. B.; MOHOVIC, T.; SILVA, C. L. O.; JANIS-ZEWSKI, M.; CAL, R. G. R.; SOUSA, É. F. de; ABE, T. P.; ANDRADE, J. de *et al.* Brazilian sepsis epidemiological study (bases study). **Critical Care**, BioMed Central, v. 8, n. 4, p. R251, 2004.

SIM, J.; WRIGHT, C. C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. **Physical therapy**, Oxford University Press, v. 85, n. 3, p. 257–268, 2005.

TANKARD, C. Big data security. **Network security**, Elsevier, v. 2012, n. 7, p. 5–8, 2012.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 63, n. 2, p. 411–423, 2001.

TING, K.; CECHO, J. J. **Apache Sqoop Cookbook**. [S.l.]: "O’Reilly Media, Inc.", 2013.

TORIO, C. M.; ANDREWS, R. M. **National inpatient hospital costs: the most expensive conditions by payer, 2011: statistical brief# 160**. [S.l.]: Agency for Health Care Policy and Research (US), Rockville (MD), 2006.

WEBER, K.; OTTO, B.; ÖSTERLE, H. One size does not fit all—a contingency approach to data governance. **Journal of Data and Information Quality (JDIQ)**, ACM, v. 1, n. 1, p. 4, 2009.

WESTPHAL, G. A.; LINO, A. S. Systematic screening is essential for early diagnosis of severe sepsis and septic shock. **Revista Brasileira de terapia intensiva**, SciELO Brasil, v. 27, n. 2, p. 96–101, 2015.

WU, X.; ZHU, X.; WU, G.-Q.; DING, W. Data mining with big data. **Knowledge and Data Engineering, IEEE Transactions on**, IEEE, v. 26, n. 1, p. 97–107, 2014.

YANG, M.-S. A survey of fuzzy clustering. **Mathematical and Computer modelling**, Elsevier, v. 18, n. 11, p. 1–16, 1993.