DESENVOLVIMENTO DE UM MÉTODO HÍBRIDO INTEGRANDO OS MÉTODOS: HIERARCHICAL CLUSTERING E BISECTING K-MEANS

BENTO, Renan Delazari[1]
CHICON, Patricia Mariotto Mozzaquatro[2]

Resumo: O presente artigo apresenta o estudo sobre a técnica de clusterização integrante da Mineração de Dados aplicada na extração e agrupamento de informações similares. A pesquisa tem por objetivo construir uma método híbrido entre as técnicas de agrupamento Hierarchical Clustering e Bisecting K-Means. Como contribuição científica objetiva-se otimizar o desempenho no processo de descoberta de conhecimento em base de dados, ou seja, pretende-se comprovar cientificamente a performance de um método híbrido. Ainda, como justificativa social o método híbrido será disponibilizado para a comunidade científica. Este artigo irá apresentar o arcabouço da pesquisa, pois a mesma encontra-se em desenvolvimento, ou seja, é parte de um trabalho de conclusão de curso.

Palavras-chave: Técnica de Agrupamento. Hierarchical Clustering. Bisecting K-Means.

Abstract: *This article presents a study on the clustering technique member of Data Mining applied to the extraction and grouping similar information. The research aims to build a hybrid method between clustering techniques Hierarchical Clustering and bisecting K -Means . As a scientific contribution objective is to optimize performance in the process of knowledge discovery in databases, ie intended to scientifically prove the performance of a hybrid method. Yet, as social justification hybrid method will be made available to the scientific community. This article will present the framework of research, because it is in development, that is , it is part of a course conclusion work.*

Keywords: *Grouping technique. Hierarchical Clustering. Bisecting K-Means.*

1. INTRODUÇÃO

Atualmente com avanço de hardware e com uma grande quantidade de dados espalhados em bancos de dados no mundo inteiro, para que estes dados tenham uma importância relevante é preciso transforma-los em informação, para isso pode-se utilizar de recursos computacionais, como as técnicas de mineração de dados - MD (Data Mining), que foi proposta na década de 80.

A mineração de dados permite a exploração e analise destes bancos de dados, podendo ser de forma automática ou semiautomática, com objetivo de descobrir padrões e regras. Estes padrões e regras são analisados e transformados em informações pertinentes e

[1] Acadêmico do Curso de Ciência da Computação. E-mail: renanbento1@gmail.com
[2] Professora do Curso de Ciência da Computação E-mail: patriciamozzaquatro@gmail.com

valiosas que possibilitam montar estratégias e previsões melhorando assim a visão de quem as usa.

O processo de transformação de dados em conhecimento é chamado de Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Databases – KDD). Os autores FAYYAD et al. (1996) definem como sendo "o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente uteis, a partir dos dados armazenados em um banco de dados.

Segundo Prass (2004), o processo de KDD é definido em cinco etapas: Seleção: Classificação dos dados a serem processados. Pré-processamento: Melhorar os dados selecionados, com rotinas de limpezas, reduções, amostragens e fragmentações dos dados. Transformação dos dados: Transformação dos dados selecionados e pré-processados em formatos específicos de entrada de cada método da MD. Mineração de dados: Aplicação de um ou mais métodos da MD, com a finalidade de descobrir modelos de conhecimentos. Avaliação: Avaliação dos modelos obtidos na etapa anterior, com a finalidade de analisar os padrões e proporcionar novos conhecimentos.

Com a expansão da internet e a sua popularização a partir da década de 90, a mineração de dados se tornou de extrema importância para a descoberta de conhecimento.

A descoberta de conhecimento permite traçar perfis de clientes, hábitos de consumo, detecção de fraudes, prever oportunidades de novos negócios, melhorar a prestação de serviços de acordo com o cliente entre inúmeras possibilidades.

As empresas de médio e grande portes, instituições governamentais e ONGs são as que mais buscam estas informações, para que possam se destacar no cenário atual, com isso investem em sistemas de informações e pessoas capacitadas na descoberta destes conhecimentos tão valioso para elas.

Existem várias técnicas na área da mineração de dados. O objetivo desta pesquisa é criar um método híbrido a partir da técnica de agrupamento, integrando os métodos Hierarchical Clustering e a Bisecting K-Means.

A técnica de agrupamento (clustering) visa identificar e aproximar os registros similares. Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não supervisionado). Além disso, ela não tem pretensão de classificar, estimar ou predizer o valor

de uma variável, ela apenas identifica os grupos de dados similares (CAMILO et.al., 2009, p.9).

O método Hierarchical Clustering (HC) originalmente surgiu na década de 60, é do tipo hierárquico como seu próprio nome sugeri. Citado por Hahn apud de acordo com NETO (2012):

> HC é um método de aprendizagem não supervisionado e suas principais vantagens pode-se citar excelência de visualização dos agrupamentos e a sua generalidade, pois não se necessita de quaisquer informações de entrada de parâmetro ou número de grupos a serem formados. Também pode ser entendido como uma serie de grupos em que cada um desses grupos encontra-se alojado dentro de outro grupo mais próximo.

Método Bisecting K-Means (BKM) originalmente foi proposto por MacQueen em 1967, está dentro da MD na técnica de agrupamento, é do tipo partição. Ele trata do problema usando o critério da mínima soma de quadrados. Desde seu surgimento ele teve inúmeros aprimoramentos.

Segundo XAVIER (2012), BKM é um método iterativo simples para particionar um conjunto de dados em um número de grupos especificados pelo usuário.

2. REVISÃO DA LITERATURA

O presente capítulo irá tratar sobre o processo de Descoberta de Conhecimento em Base de Dados ou Knowledge Discovery Database, seus conceitos, etapas. Também irá abordar a técnica de agrupamento e seus métodos Hierarchical Clustering e Bisecting K-Means.

2.1 Descoberta de conhecimento em base de dados

.

O processo de transformar os dados armazenados em uma base de dados em conhecimento é conhecido como KDD, formalizado em 1989. De acordo com Fayyad (1996), o autor define sendo “o processo não trivial, de extração de dados, previamente desconhecidas e potencialmente úteis”. Fayyad (1996) enfatiza que o processo KDD possibilita descobertas de padrões compreensíveis que podem ser interpretados em forma de conhecimento úteis.

De acordo com Fayyad (1996), o processo KDD contém uma série de etapas: seleção, pré-processamento e limpeza, transformação, mineração de dados e interpretação/validação. Resumindo, pode-se dizer que o KDD é todo processo de transformar os dados de uma base de dados em informações potencialmente úteis.

Embora tenham uma ordem a ser executado, o processo é interativo e iterativo. Interativo porque o usuário, na hora que precisar intervir e controlar o curso das atividades. Iterativo por ser uma sequência finita e as operações posteriores sempre vão depender dos resultados das operações anteriores (FAYYAD, 1996),

A fase de seleção de dados é a primeira, nela são escolhidos um conjunto de dados, pertencentes a um determinado domínio, por um especialista neste domínio. Estes dados devem conter dados significativos descartando assim informações desnecessárias para a MD. Esta seleção deve ser de tal forma que pegue os mais variáveis dados (atributos) e registros (informações) que serão analisados, processados e transformados nas etapas seguintes.

A etapa de pré-processamento conforme Lopes (2003), Queiroga (2005) têm como principal objetivo contribuir para melhorar a qualidade dos dados selecionados. Para que possa melhorar tal qualidade, utiliza-se algumas rotinas como a limpeza, redução, amostragem e fragmentação, dentre outras. Este fase é umas das mais primordiais para alcançar uma solução satisfatória na MD.

A fase de transformação de dados é responsável pela adequação dos dados processados pelas etapas anteriores. Esta adequação é especifica para cada algoritmo da MD. Neste processo, ocorre a redução de discrepâncias, a generalização, a normalização e demais transformações, juntamente com a eliminação dimensional dos dados, assim evitando a ocorrência de dados com alta relação entre si.

A etapa de MD é onde são aplicados algoritmos voltados para o descobrimento de novos conhecimentos. Esta etapa são aplicadas técnicas de classificação, caracterização, associação e agrupamento. Exemplos de algoritmos: K-Nearest Neighbor, Hierarchical Clustering, K-Means e DBSCAN.

A última fase é a da interpretação e avaliação dos resultados. Este processo tem como principal objeto analisar os resultados obtidos pela a fase anterior MD. São gerados relatórios e representação formalizando os padrões descobertos durante as fases anteriores. É recomendado que um especialista do domínio veja os resultados obtidos e interprete os mesmos, com a finalidade de validar se o algoritmo é eficaz ou não. Casos os resultados obtidos não sejam satisfatórios, deve-se fazer uma revisão de todos os processos envolvidos no KDD, e até mesmo repetido quantas vezes for necessárias. A subseção a seguir irá abordar a MD.

2.2 Mineração de Dados

A MD é o processo de descoberta de informações acessíveis em grande base de dados. Ela utiliza equações matemáticas para extrair padrões e similaridade entre dados.

Uma vez estruturados, esses dados podem ser utilizados para o processo de descoberta de conhecimento.

A extração de informação é geralmente utilizada como um passo anterior a MD, e, portanto considerada uma etapa de pré-processamento (METZ, 2006).

Após o processo de extração de informações, ou seja, quando os dados semiestruturados ou desestruturados são transformados em dados estruturados pode-se aplicar as técnicas de MD.

MD é o processo de reconhecimento de padrões válidos ou não, existentes nos dados armazenados em grandes bancos de dados (FAYYAD et al., 1996).

Nos dias de hoje a MD está sendo aplicada nas mais diversas áreas onde se busca extrair conhecimento. As áreas mais frequentes são as seguintes: Marketings, detecções de padrões de um determinado cliente, investimentos (usada principalmente nos investimentos financeiros, mas podendo ser aplicada também em planejamentos orçamentários) e produção.

Segundo Metz (2006), a mineração de dados e dividida em três hierarquias de aprendizado.

A MD integra várias técnicas, sendo as mais citadas na literatura: Agrupamento, associação, classificação, predição/previsão e sumarização. A seguir serão descritas resumidamente cada uma delas, conforme (FAYYAD e STOLORZ, 1997).

Agrupamento: É um processo de partição, que visa dividir uma população em subgrupos mais heterogêneos entre si. É diferente da tarefa de classificação, pois não existem classes predefinidas, objetos são agrupados de acordo com a similaridade.

Associação: Segundo ONODA (2006), os algoritmos possuem como principal característica constatar padrões em associações e reciprocidade entre conjuntos de itens. Consiste em determinar fatos ou objetos que tendem a ocorrer juntos em um determinado eventos ou transação.

Classificação: Consiste em construir um modelo que possa ser aplicados a dados não classificados visando categorizar os objetos em classes. Seus algoritmos proporcionam o desenvolvimento de conjuntos de atributos similares pertencentes a uma classe predefinida.

Neto (2012) refere-se a técnica presente, responsável por classificar grandes volumes de dados em classes, tornando os dados mais compreensíveis ao entendimento

Predição/Previsão: Predição consiste em definir um provável valor para uma ou mais variáveis. Já a previsão consiste em prever valores futuros baseando-se em valores cronologicamente organizados anteriores.

Sumarização: Consiste na tarefa sumarização que envolve métodos para encontrar uma descrição compacta para um subconjunto de dados.

O autor Neto (2012) afirma que não existe uma técnica que resolva todos os problema em MD. Diferentes técnicas servem para diferentes problemas, e cada uma delas tem suas vantagens e limitações comparada a outra. A seção a seguir irá apresentar a técnica de agrupamento integrante do presente estudo.

2.3 Técnica de Agrupamento

A ideia principal é que a técnica da presente seção não necessita conhecer ou determinar o nº de *clusters* ou objetos a serem gerados.

Segundo Metz (2006), a técnica de Agrupamento (*Clustering*) na MD, é frequentemente utilizado em tarefas de exploração de dados e padrões, uma de suas principais utilizações é na área da bioinformática detectando características e segmentações em imagens. Os resultado obtidos são transformados em *cluster* de acordo com sua medida de similaridade. A técnica de agrupamento e dívida nas seguintes etapas:

Pré-processamento: Prepara e transforma os dados de acordo com sua similaridade. Os dados de uma base de dados podem assumir tipos de dados diferentes, de acordo com o método usado. Muitas vezes esta preparação requer algum tipo de normalização considerando medidas de similaridade. Além dos tipos dos atributos Metz (2006) menciona que o *clustering* é influenciado pela escala, a qual indica a significância relativa dos valores dos atributos. Estas escalas podem ser qualitativa (valores nominal ou ordinal) ou quantitativa (intervalos de valores ou proporção). Exemplos de dados qualitativa: nominal (CEP, cores, sexo) e ordinal (hierarquia militar, avaliações climatológicas). Exemplos de dados quantitativa: intervalo (diferença de valor quando significativa, como temperatura em graus célsius para *fahrenheit*) e proporção (seus valores tem um significado absoluto, como altura, salário e distancia).

Seleção da medida de similaridade: Esta etapa e considerada de extrema importância pois aqui devemos analisar o conjunto de dados, utilizados na construção do cluster, e

escolher qual a medida para o cálculo da similaridade. Existem diversas medidas de distância como euclidiana, *manhattan/city-block, minkowsky* e outras. Euclidiana: esta medida será usada na presente pesquisa, Metz (2006) cita como a mais usada na técnica de agrupamento, onde é defina pela equação apresentada no Quadro 1.Medida Euclidiana

Quadro 1 - Medida Euclidiana

$$[dist(Ei, Ej) = \sum_{l=1}^{M}(x_{il} - x_{jl})^2]$$

A medica de *Manhattan/city-block*: também conhecida como distancia, pode ser definida como a distância entre dois pontos no espaço euclidiano, com um sistema de coordenadas cartesianas fixo. É definida pela equação descrita no Quadro 2.

Quadro 2 - Manhattan/city-block

$$[dist(Ei, Ej) = \sum_{l=1}^{M}|x_{il} - x_{jl}|].$$

Avaliação de clusters: Nesta etapa os dados já preparados e tendo a medida usada também definida, o conjunto dos dados deve ser aplicado para execução no método desejado, tendo como resultado a construção de clusters gerando padrões. Este padrões são avaliados para determinar o grau de conhecimento obtido pela técnica de agrupamento. A validação da técnica de agrupamento é realizada com base em índices estatísticos que avaliam de maneira quantitativa.

2.3.1 Método Hierarchical Clustering

O algoritmo Hierarchical Clustering (HC) foi desenvolvido por King na década de 60. Entretanto, tornou-se conhecido após Johnson nessa mesma década. O método HC está contido na classe dos métodos não supervisionados. Sua principal vantagem é que não precisa de qualquer informação de entrada (parâmetro ou nº de grupo) a serem formados.

O autor Neto (2012), cita que o método HC por ser compreendido como uma serie de cluster (grupo de dados similares) e, que cada cluster encontra-se alojado dentro de outro mais próximo.

A identificação de grupos é geralmente realimentado recursivamente, utilizando tanto objetos quanto grupos já identificados previamente como entrada para o processamento. Assim construindo uma hierarquia de grupos de objetos no estilo de arvore (DINIZ e LOUZADA NETO, 2000).

Conforme Neto (2012), o método hierárquico cria uma decomposição da base de dados na forma de arvore, dividindo-a recursivamente em conjuntos de dados menores. Podem ser implementados baseando-se nas estratégias:

top-down (divisivos), o processo inicia todos os objetos do mesmo grupo, o qual vai sendo divido recursivamente até de cada grupo contenha um único elemento. Algoritmos divisivos crescem exponencialmente em relação ao conjunto de entrada.

Já o botton-up (aglomerativos), cada objeto é um grupo, e cada procedimento une os dois grupos mais próximo (similares) são unidos, até que, ao final seja um único grupo. Algoritmos aglomerativos são quadráticos em relação ao conjuntos de entrada.

2.3.2 Método Bisecting K-Means

O método K-Means(KM) ou K-Medida, foi proposto por J. MacQueen em 1967, sendo um dos métodos mais usados na área da MD.

Segundo Prass (2004) o método inicia com a escolha dos k elementos que formara as sementes iniciais. Esta escolha pode ser feita de muitas formas, sendo elas:

- selecionado as k primeiras observações;

- selecionado k observações aleatoriamente;

- escolhendo k observações de modo que seus valores sejam bastantes diferentes.

Escolhido a semente inicial, é calculada a medida euclidiana em relação aos outros elementos. O elementos mais similar e agrupado no cluster da semente, recalculando o centroide do mesmo. Este processo é repetido até que todos elementos façam parte de um cluster.

Após a criação dos cluster, procura-se uma partição melhor do que a gerada arbitrariamente. Para isto, calcula-se o graus de homogeneidade interna dos grupos através da Soma Quadrada Residual (SQRes), que é a medida usada para avaliar o quão boa é uma partição. A SQRes é definida na equação do Quadra 3:

Quadro 3 - SQRes

$$SQRes(j) = \sum_{i=1}^{n_j} d^2(o_i(j), \bar{o}(j))$$

Este método e considerado confiável, porem apresenta alguns problemas: Os autores HAN e KAMBER (2001), citam que os dados de entrada devem ser numéricos ou binários (exige uma maior preocupação em converter dados para que seja executado da melhor forma possível). É sensível a valores *outliers*, um único objeto com valor muito extremo pode modificar, a distribuição dos dados e na formação dos *clusters*.

2.4 Teste de Qui – Quadrado de Pearson

Quando se quer identificar relacionamento entre duas variáveis categóricas, pode-se usar o teste qui-quadrado (FIELD, 2009). Ele é uma estatística estritamente elegante baseada na ideia simples de comparar frequências observadas em certas categorias com as frequências em que se espera conseguir. A estatística resultante é o qui-quadrado de Pearson (χ^2) e é dada pela equação descrita no Quadro 4.

Quadro 4. Qui-quadrado de Pearson

$$\text{Desvio} = \Sigma\ (\text{modelo} - \text{observado})^2 \quad (1)$$

A estatística resultante é o qui-quadrado de Pearson (χ^2) e é dada pela equação apresentada no Quadro 5:

Quadro 5. Qui-quadrado de Pearson

$$\chi^2 = \Sigma \frac{(Oservado_{ij} - Modelo_{ij})^2}{Modelo_{ij}} \quad (2),$$

Conforme o Quadro 4, os dados observados são, obviamente, as frequências das variáveis estudadas. Uma vez que não se pode trabalhar com médias de variáveis categóricas, usa-se valores esperados. Para se calcular os valores esperados para cada uma das células da tabela, utiliza-se os totais das linhas e colunas para uma célula em particular a fim de calcular o valor esperado, assim apresentado no Quadro 6.

Quadro 6. Resultado esperado

$$Modelo_{ij} = E_{ij} = \Sigma \frac{Total\ da\ Linha_{\ i} \times Total\ da\ Coluna_{\ j}}{n} \quad (3),$$

Conforme o Quadro 6, n é simplesmente o número total de observações. A estatística dada pela equação (3) pode ser avaliada uma distribuição χ^2 com propriedades conhecidas. Tudo o que se precisa saber é o grau de liberdade e aplicar a fórmula (r-1)(c-1), onde r é p número de linhas e c é o número de colunas.

3. METODOLOGIA

A pesquisa em desenvolvimento classifica-se como qualitativa, pois visa obter um método híbrido único não podendo ser comparado.

A pesquisa qualitativa não se preocupa com representatividade numérica, mas, sim, com o aprofundamento da compreensão de um grupo social, de uma organização, etc. Os pesquisadores que adotam a abordagem qualitativa opõem-se ao pressuposto que defende um modelo único de pesquisa para todas as ciências, já que as ciências sociais têm sua especificidade, o que pressupõe uma metodologia própria. Assim, os pesquisadores qualitativos recusam o modelo positivista aplicado ao estudo da vida social, uma vez que o pesquisador não pode fazer julgamentos nem permitir que seus preconceitos e crenças contaminem a pesquisa (GOLDENBERG, 1997, p. 33).

O presente trabalho será desenvolvido nas seguintes etapas:

Etapa 1: Estudo teórico: Método híbrido: Conceito. Mineração de dados: Conceito, aplicação e técnicas. Descoberta de conhecimento em base de dados: Fases e tarefas. Técnica de Agrupamento: Conceito, aplicação, arquitetura e funcionamento. Métodos Hierarchical Clustering e Bisecting K-Means: Conceito, aplicação, arquitetura e funcionamento.

Etapa 2: Implementação: método Modelagem UML Diagrama de caso de uso Diagrama de sequência Criação da base de dados Implementação da método híbrido Implementação da aplicação

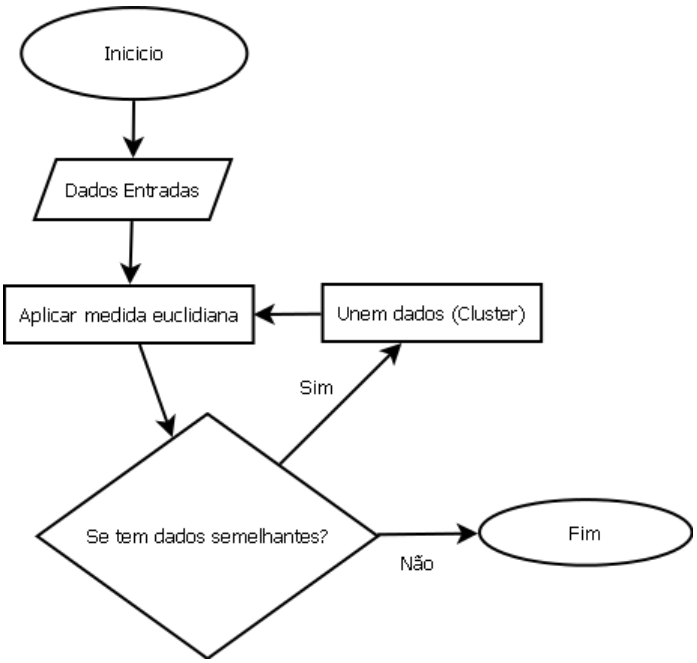
Etapa 3: Validação e resultado: Validação do método utilizando a métrica Alfa de Cronback Escrita dos Resultados

4. RESULTADOS PARCIAIS

O presente trabalho tem por objetivo otimizar o desempenho no processo de descoberta de conhecimento em base de dados, ou seja, pretende-se comprovar cientificamente a performance de um método híbrido. A modelagem da aplicação está sendo desenvolvida na linguagem AIML, construídos os diagramas de caso de uso e diagrama de sequencia.

O método hierarchical clustering é apresentado na Figura 1.

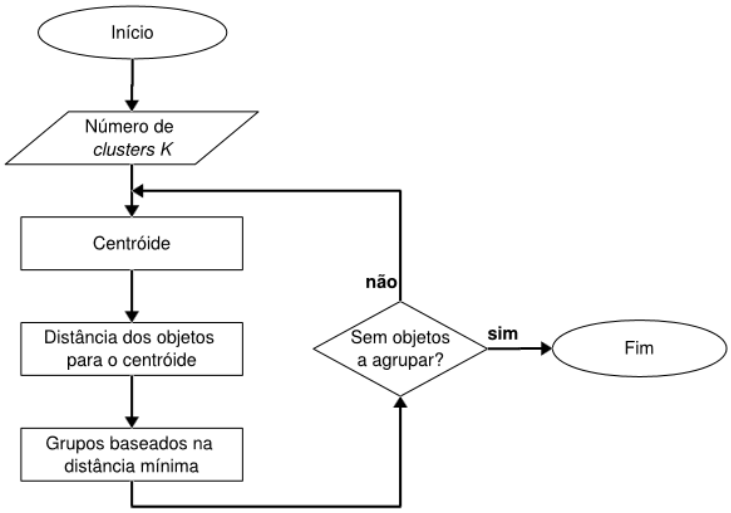
Figura 1. Fluxograma do Método Hierarchical Clustering



Conforme mostra a Figura 1, o fluxograma inicia-se com as entradas de dados, dados este que estão preparados para o processamento do método. Após esta entrada de dados o fluxo vai para aplicar a medida euclidiana dos dados mais próximos, sendo que esta etapa é recursiva ou seja até que encontre o dado mais próximo do dado selecionado efetua esta operação. No processo a seguir os dados são unidos formando um cluster. Então volta para faze anterior a fazer da medida euclidiana. Quando não haver mais dados soltos, o processo é finalizado.

A Figura 2 apresenta o fluxograma do método Bisecting K-Means.

Figura 2. Fluxograma do Bisecting K-Means



Conforme mostra a Figura2, o fluxograma inicia-se com a entrada de nº de *cluster* a serem formados. É escolhidos n dados, onde com estes dados são formados *clusters*. E comparado os dados soltos e verificada as distancias entre os dados e o centroide dos clusters formado. A cada adição dos dados no clusters é recalculada o centroide. Se houver dados a serem agrupados e feitos o processo novamente. Se não houver é finalizado o processo.

A pesquisa em desenvolvimento irá integrar os dois métodos citados a fim de gerar o agrupamento. Os testes serão aplicados utilizando a métrica Teste de Qui – Quadrado de Pearson descrita na seção 2.4.

5. CONSIDERAÇÕES PARCIAIS

Este artigo é parte integrante de um trabalho de conclusão de curso em andamento. Até o momento foi realizado todo o estudo teórico referente a implementação da técnica. A proposta está sendo modelada. Serão desenvolvidos os diagramas de caso de uso e sequencia na linguagem UML.

Como justificativa computacional está sendo desenvolvido um método híbrido integrando os métodos Hierarchical clustering e Bisecting K-Means, o mesmo objetiva otimizar o desempenho no processo de descoberta de conhecimento em base de dados, ou seja, pretende-se comprovar cientificamente a performance de um método híbrido.

Ainda, como justificativa social o método híbrido será disponibilizado para a comunidade científica.

REFERÊNCIAS BIBLIOGRAFICAS

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de Dados: Conceitos, tarefas, métodos e ferramentas. Goiás: Instituto de Informática, Universidade Federal de Goiás.

DINIZ, Carlos Alberto; LOUZADA NETO, Francisco. Data mining: uma introdução. São Paulo: ABE, 2000.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

FIELD, Andy. Descobrindo a estatística usando SPSS, Porto Alegre: Artmed, 2009.

GOLDENBERG, Mirian. A arte de pesquisar: Como fazer pesquisa qualitativa em Ciência Sociais. 8ª edição. Rio de Janeiro e São Paulo: RECORD, 2004. 107 p.

LOPES, Claudivan Cruz. Um sistema de apoio à tomada de decisão no acompanhamento do aprendizado em educação a distância. Dissertação de mestrado em informática, Universidade Federal de Campina Grande – UFCG, 2003.

METZ, Jean. Interpretação de clusters gerados por algoritmos de clustering hierárquicos. Dissertação de mestrado em ciência da computação e matemática computacional, Instituto de Ciências Matemáticas e de Computação – USP, 2006.

NETO, Gerson da Penha. Uso de algoritmos de mineração de dados para agrupamento e busca de erros em series temporais coletadas a partir de geossensores: Um estudo de caso na mata atlântica. Dissertação de mestrado em computação aplicada, Instituto Nacional de Pesquisas Espaciais – INPE, 2012.

ONODA, Mauricio. Metodologia de mineração de dados para análise do comportamento de navegar num web site. Tese de doutorado em engenharia civil, Universidade Federal do Rio de Janeiro – UFRJ, 2006.

PRASS, Fernando Sarturi. Estudo comparativo entre algoritmos de análise de agrupamento em *data mining* [trabalho obtenção do título de mestre]. Florianópolis: Universidade Federal de Santa Catarina; 2004.

QUEIROGA, Rodrigo Mendonça. Uso de técnicas de data mining para detecção de fraudes em energia elétrica. Dissertação de metrado em informática, Universidade Federal do Espirito Santo – UFES, 2005.

Relatório Técnico. RT-INF_001-09.

XAVIER, Vinicius Layter. Resolução do problema de agrupamento segundo o critério de minimização da soma de distancias [trabalho obtenção do título de mestre]. Rio de Janeiro: Universidade Federal do Rio de Janeiro; 2012.