APLICAÇÃO DO DATA *MINING* NA DESCOBERTA DE PERFIS DE RISCO DE CÂNCER

MATHEUS, Vinicius Quevedo[1];
CHICON, Patricia Mariotto Mozzaquatro[2]
ANTONIAZZI, Rodrigo Luiz[3]

Resumo*:* Atualmente obter informação como apoio a tomadas de decisões eficientes, tem sido de grande importância no ramo da saúde, dentre os recursos de tecnologia disponíveis para essa melhor tomada de decisão, destaca-se a mineração de dados aplicada em base de dados específicas para se obter o resultado desejado. O presente trabalho tem por objetivo apresentar o método de mineração de dados (*Dbscan),* seus comportamentos e resultados, com objetivo de extrair o conhecimento para uma tomada de decisão mais precisa. Para melhor apoio nas decisões, busca-se neste trabalho, desenvolver uma aplicação na base de dados real de valores da doença de câncer, para auxílio aos profissionais da saúde obtendo informações de melhor compreensão, possibilitando uma melhor tomada de decisão eficaz em ambientes reais.

Palavras-chave: Mineração de dados, *Dbscan*. Oncologia

Abstract: Nowadays obtaining information as support to decision-making efficiency, has been of great importance in the field of health among the technology resources available for better decision-making, there is a data mining applied on the basis of specific data to obtain desired result. This work aims to present two methods of data mining (DBSCAN), their behaviors and outcomes, in order to extract the knowledge for decision making more accurate. To approach the best support the decisions, we seek in this task , to develop an application on a real database with values from cancer, to help health professionals better understanding of obtaining information, allowed for a better effective decision making in real environments.

Keywords: *Data Mining, Dbscan, Oncology*

1. INTRODUÇÃO

Devido a grande disponibilidade de dados em formas eletrônicas, observou-se a necessidade de extrair informações úteis para uma melhor tomada de decisão. Na visão de alguns pesquisadores a aplicação de técnicas computacionais é considerada um passo primordial na descoberta de conhecimento. Neste contexto, uma área bastante utilizada é a Mineração de dados.

[1] Acadêmico do Curso de Ciência da Computação. E-mail: viniciusqm@gmail.com.
[2] Professora do Curso de Ciência da Computação. E-mail: patriciamozzaquatro@gmail.com.
[3] Professor do Curso de Ciência da Computação. E-mail: rantoniazzi@unicruz.edu.br.

Mineração de dados é uma das principais etapas que estão unificados no processo de Descoberta de Conhecimento em Bases de dados, ("*Knowledge Discovery in Databases – KDD*"). O objetivo do KDD é aplicar técnicas e ferramentas no intuito de descobrir informações úteis dentro de bases de dados, o KDD é uma tecnologia computacional com finalidade de descoberta de padrões, ou seja, obtenção de conhecimento a partir de um conjunto de dados transformados (FAYYAD, 1996; COELHO, 2005), podendo ser aplicada em diversas áreas como banco de dados corporativos, bancários, hospitalares, clínicas médicas dentre outros.

MD designa-se em aplicar algoritmos para a extração de padrões consistentes dentro de uma grande quantidade de dados (TARAPANOFF, 2001). Em geral o processo de conhecimento é separado nas seguintes etapas:

- Preparação: é a etapa onde os dados são preparados para que possam ser aplicáveis as técnicas de *data mining*. Os dados mais importantes são selecionados e purificados (retirada de inconsistências).
- *Data Mining*: é a etapa onde os dados preparados são processados, é a mineração propriamente dita, com o objetivo de transformar os dados em informações.
- Análise de dados: é a etapa onde se analisa o resultado gerado na *data mining*, que tem por objetivo verificar se o processo permitiu a descoberta de alguma informação antes desconhecida, definindo também a importância desses fatos gerados. Normalmente os resultados são expressos em gráficos.

O artigo proposto tem como objetivo a descoberta de perfis de risco de câncer com base em dados históricos da doença, para isso, será utilizada a técnica *clusterização* de dados implementando o método *Dbscan*.

O *Dbscan* descobre agrupamentos em bases de dados com ruídos, ou seja, encontra regiões densas que são separadas por regiões de baixa densidade agrupando os objetos na mesma região densa. O método encontra agrupamentos verificando a vizinhança de *r* de cada ponto da base de dados HAN e KAMBER (2001). O desempenho do algoritmo depende da utilização ou não de um índice indexador na base de dados, independente ou não do uso do indexador, seu custo computacional é comparado a métodos hierárquicos, pois precisa necessariamente comparar todos os elementos presentes na base de dados.

O número de mortes por câncer em nível mundial deverá aumentar 45% entre 2007 e 2030 (de 7,9 milhões para 11,5 milhões de mortes), segundo a Organização Mundial da Saúde (HEALTHY,2012). Desse modo o trabalho proposto irá contribuir socialmente auxiliando na identificação de possíveis perfis de risco com base em análise de um banco de dados composto por dados de pacientes que obtiveram a doença.

2. REVISÃO DA LITERATURA

As subseções a seguir irão apresentar a mineração de dados, a técnica de agrupamento e o método dbscan.

2.1 Mineração de Dados

Atualmente grande parte das organizações utiliza um sistema de gerenciamento de banco de dados (SGBD), onde são extremamente eficientes em organizar e armazenar dados obtidos em suas operações cotidianas, porém, a maioria ainda não consegue usar adequadamente todos esses dados para que sejam transformados em informações ou conhecimentos, que possam ser usados em favor de suas organizações auxiliando em tomadas de decisões.

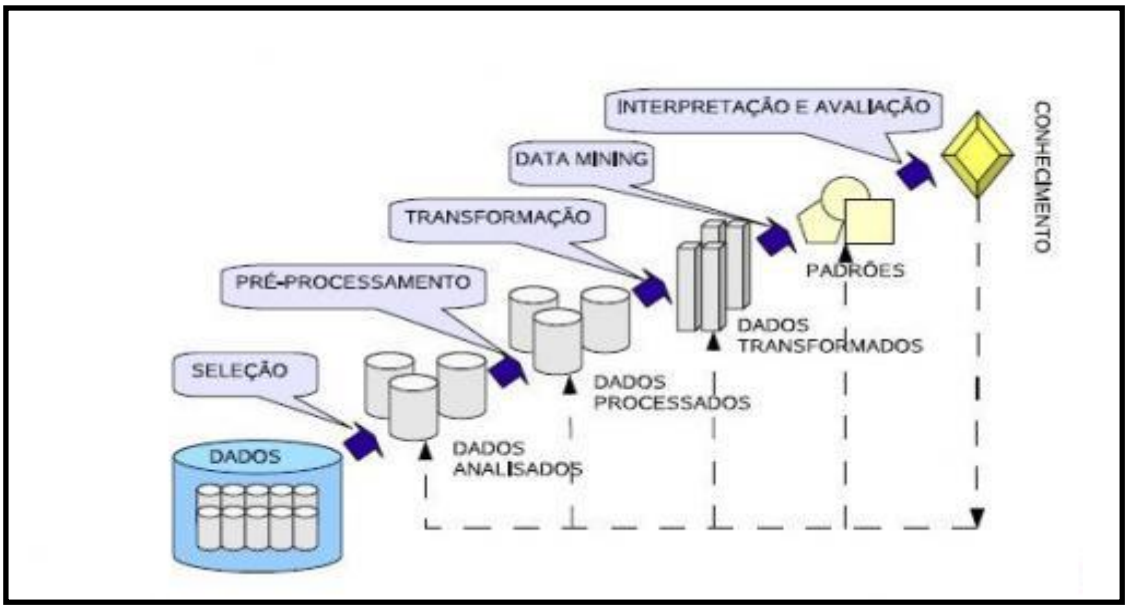
O conceito de mineração é cada vez mais utilizado como uma tarefa de descoberta de informações, ela possibilita extrair de grandes quantidades de dados um resultado que melhore e facilite as escolhas baseados nos dados minerados (HOSKING , 1997).

Segundo Han e Kamber (2011), aplicar mineração de dados tem relação nas descrições de classes e conceitos da discriminação dos dados pesquisados, por exemplo, a análise de uma base de dados de um hospital ou clínica que classifica pessoas com maior risco para um determinado tipo de doença (criança até 12 anos têm maior probabilidade de ter problemas respiratórios, idosos entre 60 a 80 anos têm maior probabilidade de ter alguma espécie de câncer). Frequentemente a mineração tem sido considerada uma mistura de pesquisas e estatísticas, inteligência artificial e bancos de dados. Mineração é a parte de um processo de pesquisa denominado Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database - KDD*), no qual possui uma metodologia própria para a preparação e exploração dos dados, interpretação dos seus resultados e assimilação dos conhecimentos minerados (Han; Kamber, 2011).

O processo de transformação da informação em conhecimento teve um crescimento exponencial devido à evolução dos computadores, a utilização do método de KDD possibilitou um grande avanço e uma maior facilidade e agilidade na extração de informação, possibilitando as pessoas analisarem bases de dados antes inacessíveis e obtendo conhecimento antes impossível. (PASSOS, 2006).

A técnica de KDD pode ser dividida em processos e etapas, segundo o autor Fayyad (1996) um dos primeiros autores a escrever sobre o assunto, apresentou um processo baseado em cinco fases conforme apresentado na Figura 1.

Figura 1. Etapas de Mineração de dados



Fonte: Adaptado de Han e Kamber (2001)

Conforme ilustra a Figura 1, a fase inicial é onde se escolhe a base de dados a ser avaliada, possui impacto significativo sobre a qualidade do resultado final, nesta fase é escolhido o conjunto de dados contendo todas as possíveis variáveis (também chamadas de características ou atributos) e registros (também chamados de casos ou observações) que farão parte da análise (PRASS, 2004). Normalmente essa escolha dos dados fica a critério de um especialista do domínio. Logo após, a escolha é realizada a limpeza dos dados, que é uma parte fundamental no processo de KDD, pois a qualidade dos dados vai determinar a eficiência dos algoritmos de mineração. Nesta etapa deverão ser feitos as tarefas que retirem dados redundantes e inconsistentes, recuperem dados que não estão completos e avaliem possíveis dados diferentes do conjunto. Transformação dos dados é a fase do KDD que

antecede a fase de mineração, após serem selecionados, limpos e pré-processados, os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos possam ser aplicados. *Data Mining* é a principal etapa no processo, onde é feito a exploração e análise, de forma automática ou semiautomática com o objetivo de descobrir padrões e regras em grandes bases de dados. A avaliação é a fase final, onde se identifica e extrai o conhecimento da base de dados observando os resultados obtidos nas fases anteriores (FAYYAD ,1996).

A mineração de dados tem sido aplicada em diferentes domínios, mas para sua aplicação é necessário uma estrutura com algumas tarefas para uma execução de forma organizada. As tarefas de mineração possuem divisões como a predição e descritiva. A predição envolve a extração dos dados que prevê um valor futuro de uma forma conclusiva, ou seja, a atividade tem o objetivo de auxiliar na tomada de decisão de forma não supervisionada. Na atividade de descrição busca-se formas e padrões de dados que apresentam uma interpretação humana, que demonstra um apoio à decisão, as duas tarefas são implementadas por meio de técnicas (SILVEIRA, 2003). A subseção a seguir irá abordar de forma geral as técnicas de mineração de dados classificação, *clusterização*, associação e padrões sequenciais.

2.1.1 Técnica de agrupamento

Clustering ou agrupamento é descrito como um conjunto de dados, de objetos, onde é possível agrupá-los de forma que os elementos que compõem cada grupo sejam mais parecidos entre si do que parecidos com os elementos dos outros grupos. É colocar os dados iguais (ou quase) juntos em um mesmo grupo e os desiguais em grupos distintos Han e Kamber (2001). A *clusterização* permite determinar qual o número de grupos e os grupos existentes em um conjunto de dados.

Segundo (COLE, 1998), o problema de *clustering* é a busca por aquelas partições que refletem a estrutura de um conjunto de objetos, é um procedimento exploratório que busca por uma estrutura natural com relação ao conjunto específico. Este processo envolve ordenar os casos de dados, ou objetos, em grupos, ou *clusters*, tal que os objetos no mesmo *cluster* são mais parecidos entre si do que em relação aos objetos em outro *cluster*.

Hruschka e ebecken (2001) definem *clustering* como sendo uma tarefa onde se busca identificar um conjunto finito de categorias ou *clusters* para descrever os dados.

Os autores Han e kamber (2001) falam que *clustering* é o processo de agrupar os dados em *clusters* tal que os objetos dentro de um *cluster* são muito parecidos em comparação

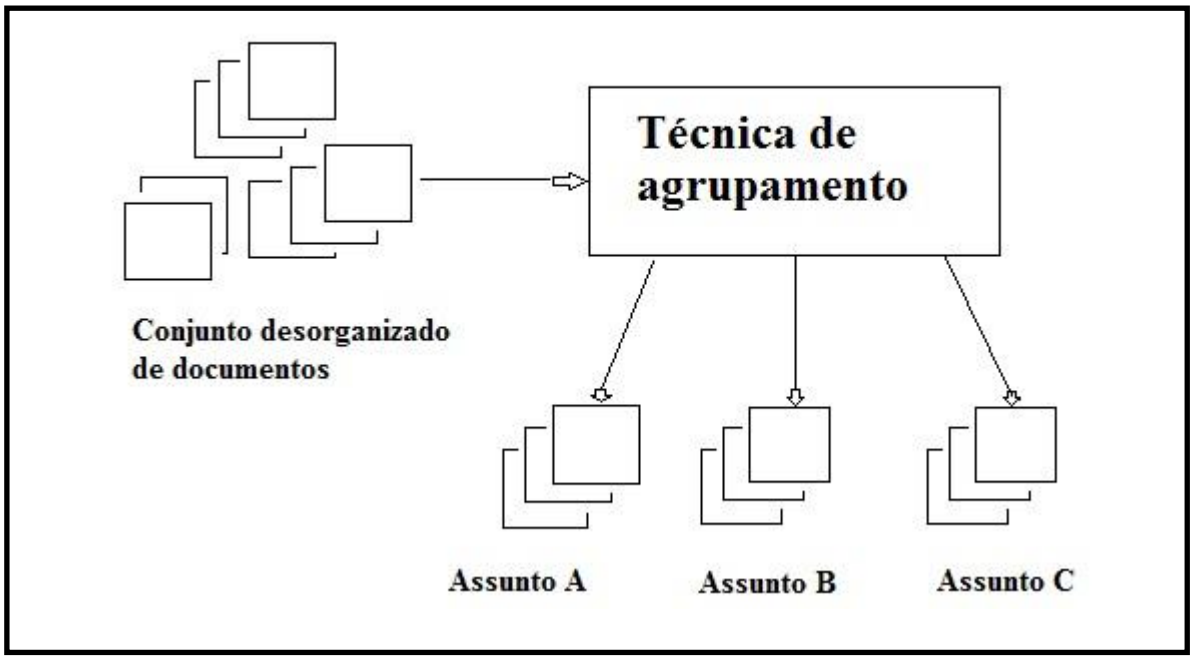
uns com os outros, mas são muito diferentes para objetos em outros *clusters*. Também comparam classificação com *clusterização* escrevendo que ao contrário da classificação, *clustering* não conta com classes predefinidas e exemplos de treinamento de classes rotuladas. Por esta razão, *clustering* é uma forma de aprendizado por observação, em lugar de aprendizado por exemplo.

Entretanto Ankerst (1999) destaca que o propósito de descobrir *clusters* é encontrar a partição de um banco de dados de registros tal que os registros que tem características similares são agrupados juntos. Isso, então, permite que as características de cada grupo possam ser descritas.

A *clusterização* pode ser aplicada em várias áreas em que se deseje agrupar dados, sejam eles dados de supermercados, por exemplo, onde é possível agrupar dados de compras dos clientes, os sintomas de doenças, os documentos existentes da internet, transações bancárias dos clientes de um determinado banco Han e Kamber (2001).

Conforme ilustra a Figura 2, tem-se um conjunto de documentos desorganizados que representam a base de dados, após é aplicada à técnica de *clusterização* nesses documentos, conseguiu-se agrupar os documentos por assunto, onde na base de dados seriam os *clusters* semelhantes.

Figura 2. Organização de documentos com agrupamento



Fonte: Adaptado de Wives (1999)

2.1.1.1 Método Dbscan

Os métodos de *clusterização* por densidade utilizam critérios de *clusterização* local, por considerarem a densidade de ligações entre os dados. A possibilidade de encontrar *clusters* de formas arbitrárias e o fato de não precisar da definição do número de *clusters* como parâmetro inicial são as principais vantagens dos métodos baseados em densidade (Yip, 2006). Entretanto, alguns algoritmos podem exigir a definição de outros parâmetros, como é caso do algoritmo *DBSCAN*.

Um *cluster* baseado em densidade é um conjunto de objetos conectados por densidade, todo objeto não contido em qualquer *cluster* é considerado um ruído (HAN; KAMBER, 2001). O método *Dbscan* é baseado em densidade onde o mesmo cresce em regiões com densidade alta o suficiente nos *clusters* e descobre *clusters* em base de dados espaciais com ruídos (HAN e KAMBER, 2001).

O algoritmo encontra regiões densas que são separadas por regiões de baixa densidade, regiões de ruídos, e agrupa os objetos na mesma região densa (AGRAWAL, 1998). A ideia desse método é que cada um dos *clusters* deve manter uma vizinhança com um número mínimo de vizinhos dentro de uma esfera de raio R. Os *clusters* que possuem uma vizinhança com densidade mínima, e estão a uma distância menor que R, pertencem ao mesmo *cluster*. O método é aplicável a qualquer base de dados contendo dados de um espaço métrico, isto é, bases de dados com uma função de distância para pares de objetos.

O algoritmo recebe como entrada o tamanho da vizinhança *Eps*, o número mínimo de pontos *MinPts*, e também um conjunto de dados X. Se a distância entre dois pontos centrais é menor que *Eps*, eles são colocados no mesmo *cluster*. Os pontos periféricos são colocados no mesmo *cluster* que os pontos centrais e pontos ruidosos são descartados da classificação, por não pertencerem a *cluster* algum. Ainda ESTER (1996), esclarece que a ideia principal do método é que para cada objeto de um *cluster*, sua vizinhança, para algum dado raio (*Eps*), tem que conter ao menos um número mínimo de objetos (*MinPts*), isto é, a densidade da vizinhança tem que exceder algum limite.

O objeto de borda, como por exemplo, o M, está na vizinhança de dois objetos centrais, T e R, que pertencem aos *clusters* C1 e C2, sendo assim, pode ser atribuído a qualquer um dos dois, pois está em uma região de fronteira dos dois agrupamentos, quando isso ocorre, uma convenção diz que o objeto M será atribuído ao primeiro *cluster* encontrado.

O objeto S não é atribuído a nenhum dos *clusters*, então ele será definido como um ruído (ESTER,1996; HAN; KAMBER,2001).

Foi observado que o algoritmo é muito sensível aos parâmetros definidos pelo usuário, e esses parâmetros, são difíceis de determinar (HAN e KAMBER, 2001). ESTER (1996) afirma que o método *DBSCAN* somente necessita de um parâmetro de entrada (*Eps*), pois o outro seria fixo para todas as bases de dados, contanto que sejam de duas dimensões, e o algoritmo ajuda o usuário a determinar um valor apropriado para este parâmetro.

Com os parâmetros *Eps* e *Minpts* já informados pelo usuário conhecedor da BD, o algoritmo inicialmente escolhe um ponto arbitrário Xp, então N*Eps* (Xp) é recuperada, e se Xp for um objeto de borda então não irão existir pontos diretamente alcançáveis por densidade a partir de Xp, pois a N*Eps* (Xp) <= *Minpts*. O ponto Xp é marcado como ruído e o *Dbscan* prossegue no próximo ponto. Se um ponto for marcado como ruído pelo algoritmo, posteriormente ele pode estar na *Eps*-vizinhança de outro ponto que ainda não foi visitado, assim essa classificação pode ser retirada caso o objeto seja diretamente alcançável por densidade a partir de um ponto central ainda não visitado. Se N*Eps (*Xp*)* contenha ao menos *Minpts*, um novo *cluster* é criado contendo o ponto Xp e todos os pontos na *Eps*-vizinhança de Xp. Com o novo *cluster* formado a *Eps*-vizinhança de cada ponto ainda não visitado é recuperado iterativamente e a densidade de cada ponto nessa vizinhança é calculada, permitindo que novos pontos possam ser adicionados ao *cluster* (ANKERST,1999;ESTER,1996).

GUHA (1998) afirma que o algoritmo *Dbscan* também sofre do problema de falta de robustez que atinge os métodos hierárquicos de *clusterização* que utilizam todos os objetos, e como o método não desempenha qualquer etapa de *pré-clusterização* e trabalha diretamente sobre a base de dados inteira, ele pode ter alto custo computacional no caso de bases de dados grandes. O autor ainda diz que métodos baseados em densidade usando amostragem aleatória para reduzir o tamanho da entrada podem não ser possíveis, a razão para isto é que a menos que os tamanhos das amostras sejam grandes, podem existir variações substanciais na densidade dos objetos dentro de cada *cluster* na amostra aleatória.

3. METODOLOGIA

A pesquisa desenvolvida classifica-se como quantitativa, pois conforme o autor AZEVEDO (1998), uma pesquisa ou método científico é definido quantitativo quando normalmente existe uma medida numérica que possibilita uma análise estatística dos dados.

Para desenvolver o trabalho serão realizadas algumas fases metodológicas, conforme descritas seguir:

A etapa um mostra um estudo teórico sobre: estudar processos cancerígenos; extração de informações em uma base de dados com valores de diversos tipos de câncer; aplicação das técnicas de mineração de dados na tomada de decisão; estudar a descoberta de conhecimento em base de dados; analisar técnicas de mineração de dados, como também a implementação e funcionamento do método *Dbscan.*

A etapa dois mostra o desenvolvimento prático: necessário modelar o sistema para aplicação dos métodos; desenvolver o sistema e integrar um banco existente; implementar o algoritmo *Dbscan* no sistema citado.

A etapa três refere-se à validação: comparar os métodos utilizados, buscando identificar de igualdade e divergência; efetuar análises comparativas dos resultados obtidos conforme os parâmetros de entrada. Observar os resultados realizando uma análise quantitativa demostrando em forma de tabelas e gráficos.

4. RESULTADOS E DISCUSSÕES

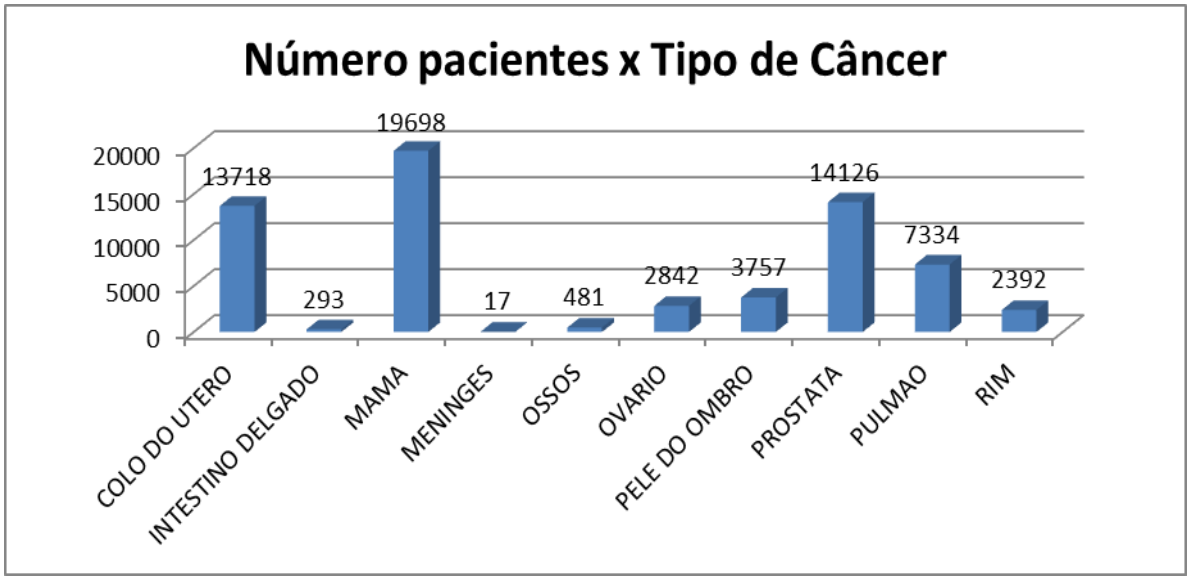
A pesquisa desenvolvida apresenta os resultados da aplicação das técnicas de mineração de dados no diagnóstico de câncer.

Alterações em determinados genes – mutações – causam o câncer, mas falhas na interação que cada célula do organismo mantém com as demais células e com moléculas presentes na chamada matriz extracelular também estão envolvidas na origem e evolução dos tumores. Os cientistas vêm desvendando detalhes desses mecanismos (os genéticos e os interativos), o que poderá levar a novas drogas que previnam ou combatam a doença e talvez a terapias capazes de reverter o processo que resulta no câncer BELIZÁRIO (2002).

Foi desenvolvido um sistema computacional no qual foi integrada a técnica de mineração de dados clusterização onde foi implementado o método dbscan. O sistema identifica o tipo de câncer, agrupa pacientes com sintomas silimares, idade e sexo.

A Figura 3 mostra os resultados obtidos na mineração efetuada em uma base de dados, com dados retirados do Institudo de Oncologia de São Paulo, os números mostram uma confirmação de maior incidência de câncer de mama entre as mulheres, e de próstata entre os homens.

Figura 3 – Número de pacientes versus câncer

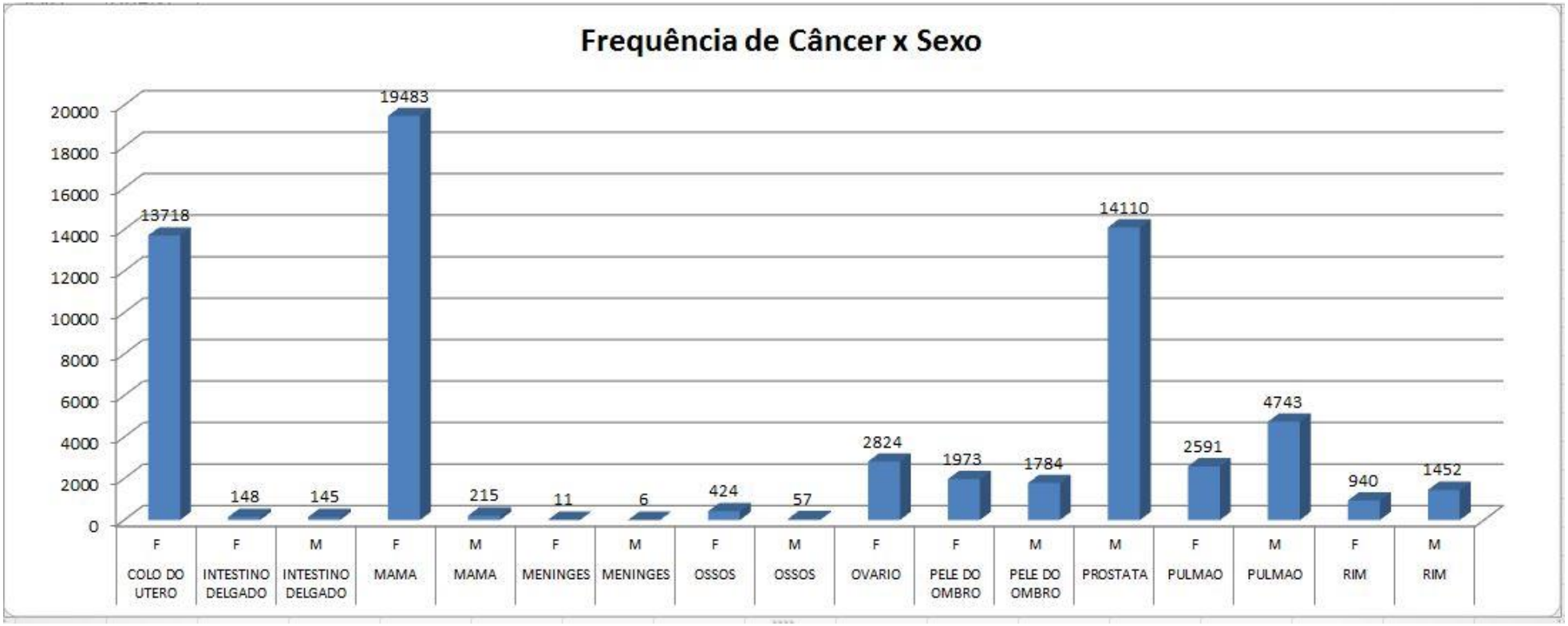


Fonte: Do autor

Para realizar essa análise, for retirada uma amostra da base de dados, pois a mesma resultou em um número muito grande de resultados. O gráfico leva em consideração a frequência dos tipos de câncer estudados nesse artigo, não levando em consideração o sexo.

Alguns tipos de câncer podem ocorrer em ambos os sexos, já outros não, como é o caso de colo de útero, que ocorre somente em mulheres. Já o câncer de pulmão pode ocorrer em ambos os sexos, então foi feito um melhor detalhamento dos resultados obtidos, como mostra a Figura 4.

Figura 4 – Frequência de câncer versus sexo



Fonte: Do autor

A Figura 4 mostra um melhor detalhamento dos resultados obtidos, sendo agora detalhado por tipo de sexo. Analisando o gráfico, verifica-se uma maior incidência de câncer de pulmão entre homens do que em mulheres, o mesmo ocorre no câncer de rim.

No câncer de pele há uma maior incidência de câncer entre as mulheres, mas observa-se que a diferença é pequena. Também há uma maior incidência de câncer de ossos entre as mulheres do que entre os homens, e no câncer de intestino há quase um empate.

Observa-se também que o câncer de mama entre os homens é relativamente raro, pois é alta a discrepância entre os números apresentados no gráfico.

5. CONSIDERAÇÕES FINAIS

O presente artigo tem por objetivo aplicar o método de mineração de dados (*Dbscan*), em uma base de dados, para descobrir as maiores tendências e perfis de determinados tipos de câncer, para que profissionais da área médica possam analisar os dados e tomar algumas decisões ou até mesmo tentar descobrir o porquê de determinado tipo de câncer é tem mais incidência em um determinado sexo do que em outro.

O trabalho proposto irá contribuir socialmente auxiliando na identificação de possíveis perfis de risco com base em análise de um banco de dados composto por dados de pacientes que obtiveram a doença.

Essa contribuição científica viabiliza o estudo da aplicação do método (*Dbscan*) de *data mining* discriminando o funcionamento e resultados, com o objetivo de extrair informações úteis como à busca de padrões de perfis de risco.

Como contribuição social se traz com a tentativa de descoberta de perfis de riscos dentro de uma base com dados reais, assim a sociedade poderá se beneficiar com a informação e verificar se poderá enquadrar-se em algum perfil de risco ou não.

Dificuldade encontrada foi em relação ao volume de dados, a base de dados do Instituto de Oncologia era relativamente grande, com vários tipos de câncer, então se optou em definir dez tipos de câncer mais comuns para realizar o agrupamento, assim não gerando muita lentidão na importação e posteriormente no processamento desses dados.

Pode-se obter em projetos futuros, processar as informações em plataformas móveis, assim profissionais da área poderão fazer uma primeira análise do paciente sem necessariamente estarem em um consultório.

REFERÊNCIAS BIBLIOGRAFICAS

ANKERST, M., BREUNIG, M., M., KRIEGEL, H.-P., et al., 1999, **"OPTICS: Ordering Points to Identify the Clustering Structure"**, In: Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 49-60, Philadelphia, PA, USA, June.

AGRAWAL, R., GEHRKE, J., GUNOPULOS, D., et al., 1998, **"*Automatic Subspace Clustering on High Dimensional Data for Data Mining Applications*"**, *In: Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 94- 105, Seattle, Washington, USA, June

AZEVEDO; Israel Belo de. **O prazer da produção científica: diretrizes para a elaboração de trabalhos acadêmicos**. Piracicaba: Ed. da UNIMEP, 1998.

BELIZÁRIO, José Ernesto. Departamento de Farmacologia, Instituto de Ciências Biomédicas, Universidade de São Paulo. Ciência Hoje • v. 3 1 • n º 184 , 2002.

COELHO, Paulo Sérgio de S. **Um Sistema para Indução de Modelos de Predição Baseados em Árvores.** Tese de Doutorado (Doutorado em Ciência em Engenharia Civil) – Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2005.

COLE, R. M., 1998, ***Clustering with Genetic Algorithms***, *M. Sc., Department of Computer Science, University of Western Australia*, Australia.

ESTER, M., KRIEGEL, H.-P., SANDER, J., et al., 1996, **"*Incremental Clustering for Mining in a Data Warehousing Environment"*,** *In: Proceedings of the 24th International Conference on Very Large Data Bases* (VLDB), pp. 323-333, *New York City, New York, USA, August*

FAYYAD, Usama; SHAPIRO, Gergory P.; SMYTH, Padhraic. ***The KDD Process for Extracting Useful Knowledge fomVolumes of Data***. *ACM Press, New York*, v. 39 , n. 11, p. 24-26, nov. 1996.

HAN, Jiawei; KAMBER, Micheline. ***Data Mining***: *Concepts and Techniques. San Diego: Academic Press*, 2001.

HOSKING , J.R.M., PEDNAULT , SUDAN,M.**A *statistical perspective on data mining***. *New York* , NY, EUA,1997.

HRUSCHKA, E. R., e EBECKEN, N. F. F., 2001, "**A Genetic algorithm for cluster analysis", Submitted to: IEEE Transactions on Evolutionary Computation** , January 2001.

PASSOS,M. **Modelos de dispositivos de microondas e Ópticos através de redes neurais artificiais**. Natal,2006.

PRASS, Fernando Sarturi. **Estudo comparativo entre algoritmos de análise de agrupamentos em *data mining.*** Dissertação de Mestrado em Ciências da Computação, Universidade Federal de Santa Catarina – UFSC,2004.

SILVEIRA, R de FREITAS. **Mineração de dados aplicada à Definição de Índices em Sistemas de Raciocínio Baseados em casos** . UFRGS, 2003.

TARAPANOFF,Kira. (Org). **Inteligência organizacional e competitiva**. Brasília : Editora Universidade de Brasília, 2001.

WITTEN, I. H.; FRANK, E. ***Data Mining: Pratical Machine Learning Tools and Techniques with JavaImplementations***. Morgan Kaufmann Publishers. San Francisco, Califórnia, 2000.

WIVES, Leandro Krug. **Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de "*Clustering*",** Universidade Federal do Rio Grande do Sul, 1999.

YIP, A. M.; DING, C.; CHAN, T. F. ***Dynamic cluster formation using level set methods.*** *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 28, 2006.