1
Introdução

1.1.
Motivação

Hoje em dia, com o mundo cada dia mais globalizado, vem-se acirrando a competitividade entre as empresas. Com isso, a informação vem ganhando mais destaque e importância, pois nela reside o conhecimento, diferencial para que as empresas atinjam seus objetivos, mantenham a sua competitividade e aumentem a eficiência de suas operações.

O avanço tecnológico vem proporcionando o armazenamento e recuperação de dados de forma cada vez mais rápida, confiável e segura. Isso permitiu um crescente investimento pelas empresas na migração de seus dados para o meio digital.

Essa enorme massa de dados gerada pelas diversas empresas diariamente pode conter informações importantes que não são fáceis de serem extraídas, pois geralmente o volume de dados é grande ou as informações que se buscam estão muito espalhadas pelos diversos bancos de dados. Com isso surge a necessidade de se analisar os dados automaticamente, buscando dessa forma extrair informação útil que possa agregar algum tipo de conhecimento (Evsukoff, 2005).

Existem várias formas de se analisarem os dados automaticamente, como a previsão, a análise de agrupamentos e a classificação de dados.

A previsão de dados (geralmente séries temporais) procura tratar de problemas onde é importante se ter uma idéia de como esses dados, ou variáveis, se comportarão no futuro, na maioria das vezes baseando-se em dados históricos.

A análise de agrupamentos, por sua vez, procura encontrar grupos de dados semelhantes entre si. Um agrupamento nada mais é do que um conjunto de dados com características similares. Esses agrupamentos determinam um modelo para a estrutura dos dados e, se analisados adequadamente, podem revelar informações importantes.

A classificação de dados pode ser considerada como o passo seguinte à análise de agrupamentos, onde se procura determinar a qual grupo (classe), entre os grupos pré-definidos, uma nova amostra pertence.

As técnicas de análise de agrupamentos estão ganhando cada vez mais mercado e se mostram bastante interessantes, pois revelam como os dados estão estruturados e possibilitam um melhor entendimento sobre o negócio. Por exemplo, em uma base de dados da área de telecomunicações, pode ser possível encontrar grupos que identifiquem diferentes tipos de clientes, permitindo, dessa forma, que a área de marketing trabalhe de forma diferenciada para cada grupo de cliente, de acordo com as características intrínsecas do grupo em questão.

O processo de agrupamento pode ser dividido basicamente em três etapas: seleção e tratamento de dados, agrupamento de dados e análise dos resultados.

Os principais itens que devem ser considerados nesse processo são (Berkhin, 2002):

- Tipo de atributos que o algoritmo opera
- Escalabilidade para grandes conjuntos de dados
- Habilidade de operar com uma dimensão grande de variáveis
- Habilidade de encontrar agrupamentos de forma irregular
- Tratar valores discrepantes (*outliers*)
- Tempo de execução
- Dependência de ordem dos dados
- Classificação
- Segurança no conhecimento a priori e parâmetros definidos pelo usuário (coleta de definições formais dos termos envolvidos)
- Interpretabilidade dos resultados

Nem todos os itens acima são considerados pela maioria dos métodos, e qualquer método que venha a ser escolhido irá gerar resultados que podem ser adequados (contêm informação relevante) ou não. Como a análise dos resultados gerados é muito subjetiva, o desconhecimento do assunto ou a escolha de um método não adequado ao problema podem levar a conclusões erradas sobre a estrutura dos dados. Por isso é importante conhecer e escolher um método que se ajuste ao problema e aos dados utilizados.

Além disso, devido ao grande número de parâmetros a serem escolhidos nos diversos métodos, em geral os autores recomendam que cada método seja executado várias vezes, sob condições iniciais diferentes.

Por todas essas razões é conveniente conhecer os principais métodos de agrupamento de dados e saber avaliar corretamente os resultados gerados. Existe ainda hoje uma escassez de ferramentas para esse fim. Em sua maioria, essas ferramentas são utilizadas no meio acadêmico ou são desenvolvidas em grandes empresas para uso próprio.

Academicamente, uma das ferramentas mais conhecida é o Matlab®, porém a ferramenta R® vem ganhando mercado. Essas ferramentas são bastante poderosas, mas as informações sobre a melhor forma de efetuar a modelagem do problema de agrupamento estão muito dispersas. Isto se deve ao fato de não serem ferramentas específicas para o problema e pela abrangência de diversos assuntos, o que as tornam de difícil utilização. A curva de aprendizado acaba sendo longa para que o usuário esteja apto a trabalhar em um determinado assunto de forma adequada.

As duas ferramentas citadas acima disponibilizam vários métodos para o processo de agrupamento de dados, porém existem diversos métodos que estão presentes em uma ferramenta e não estão presentes na outra. Em um problema real de agrupamento convém analisar os dados através da utilização de diferentes métodos, a fim de buscar aquele que melhor se adapte ao problema. Por não serem ferramentas integradas, o usuário fica limitado a uma ferramenta específica e acaba se conformando com o resultado obtido.

1.2. Objetivos

Deste modo, este trabalho tem como objetivos principais:

- *Estudo do processo de agrupamento de dados* ⇨
  As técnicas de agrupamentos de dados estão sendo cada vez mais pesquisadas e utilizadas, principalmente nas grandes empresas onde é importante a aquisição de informações estratégicas.

Esse estudo apresenta uma revisão detalhada do processo de análise de agrupamentos, bem como os principais métodos e os resultados gerados para análise de dados.

- *Desenvolvimento de aplicativo para análise de grupos* ⇨
  Desenvolvimento de um aplicativo que auxilie, de forma completa, todo o processo de agrupamento de dados. O aplicativo desenvolvido deve poder ser executado em qualquer plataforma (Windows, Unix, etc.), ser de fácil utilização e possibilitar o desenvolvimento de novos métodos internos, bem como a integração com outros aplicativos, permitindo ao usuário selecionar o(s) método(s) mais adequado(s) aos seus objetivos. O aplicativo deve disponibilizar também planilhas e gráficos que auxiliem na análise dos resultados obtidos.

- *Avaliação do Aplicativo em um caso real* ⇨
  De forma a demonstrar a facilidade de uso, assim como avaliar as vantagens do emprego de métodos de natureza fuzzy, é utilizada uma base de dados real. Os métodos de natureza fuzzy são técnicas onde um determinado dado pode pertencer a mais de um agrupamento. Este tipo de abordagem possibilita uma análise mais rica e menos rígida das distribuições dos dados.

1.3. Organização da Dissertação

Esta dissertação está organizada em 6 capítulos.

O Capítulo 2 apresenta alguns conceitos fundamentais em análise de agrupamentos, descrevendo todo o processo basicamente dividido em três etapas: seleção e tratamento de dados, agrupamento de dados e análise dos resultados.

O Capítulo 3 descreve os principais métodos e algoritmos de agrupamento de dados.

Já o Capítulo 4 apresenta a modelagem e o desenvolvimento de um aplicativo dedicado para a tarefa de agrupamento de um conjunto de dados. Este

aplicativo permite ao usuário a escolha do(s) método(s) de agrupamento de dados mais indicados(s) para o seu problema. O usuário pode escolher entre os métodos de agrupamento já implementados, permitindo, também que sejam a ele incorporados outros métodos já existentes ou implementados pelo próprio usuário.

O Capítulo 5 apresenta os estudos de caso e a análise dos resultados, os quais visam demonstrar a facilidade de uso do aplicativo, assim como avaliar as vantagens do uso de métodos de natureza fuzzy em uma base de dados real.

O Capítulo 6 apresenta a conclusão da dissertação e sugestões para pesquisas futuras.