SETip - Sistema Especialista para Tipificar Dados de uma Pesquisa: Variáveis Qualitativas ou Quantitativas

JOSÉ GONÇALO DOS SANTOS[1]
SILVIA MODESTO NASSAR[2]

[1]UNITINS – F undação Universidade do Tocantins
Departamento de Informática
Campus de Paraíso do Tocantins (TO)
goncalo@inf.ufsc.br

[2]UFSC – Universidade Federal de Santa Catarina
INE – Departamento de Informática e Estatística
Campus Trindade – Florianópolis (SC)
silvia@inf.ufsc.br

Resumo: Este artigo tem por finalidade, apresentar o protótipo de um sistema especialista que tem como função identificar dados categorizados, - variável estatística qualitativa – dentre as variáveis contidas numa base de dados. Evitando dessa forma, que usuários com pouca experiência em estatística e na utilização de *software* com rotinas estatísticas, trate os dados categorizados como dados quantitativos, realizando operações sem sentido, comprometendo dessa forma a análise estatística dos dados, e consequentemente levando-os a concluir a pesquisa de maneira incorreta. A linguagem usada para implementação, foi Borland DELPHI™, versão 4.0, devido ao fato de ser uma linguagem de programação muito difundida no Brasil, e também devido a sua facilidade em manipular base de dados.

Palavras-Chave: Inteligência Artificial, Sistemas Especialistas, Heurísticas, Análise Estatística .

1 Introdução

A análise estatística é fundamental para as áreas de ciência, tecnologia, comércio, entre outras. Ao lançar um novo produto no mercado, deseja-se saber se tal produto será aceito ou não pelos consumidores, isso é feito através de experimentos, experimentos estes, que geram dados que devem ser analisados de forma correta para que se chegue o mais próximo possível do resultado real sobre a opinião dos consumidores; ou quando deseja-se testar a confiabilidade de um novo *software*, isto também é feito através de experimentos; há uma infinidade de exemplos em que a análise estatística é aplicada. E hoje existem vários *softwares* estatísticos no mercado, que fazem todas as inferências estatísticas, deixando apenas a interpretação dos resultados a cargo dos estatísticos ou alguém que esteja trabalhando com estatística. Mas nenhum desses *softwares* evitam que os usuários tratem dados qualitativos como dados quantitativos. Existe atualmente no departamento de estatística e informática da UFSC, um *software* chamado SEstat (Sistema Especialista para auxílio ao ensino de estatística) [PRE99], que está sendo usado por alguns professores do departamento para ministrar aulas de estatística. O SEstat funciona muito bem, no sentido de tipificar variáveis, porém ele trabalha com uma base de dados fixa que foi previamente embutida na base de conhecimento do sistema, surge então a necessidade de criar um sistema especialista que consiga tipificar dados de bases fornecidas pelos usuários. A importância de tipificar tais dados, é devido ao fato de que dados categorizados tem um tratamento diferenciado de dados quantitativos em análise estatística. E se não for feita esta distinção, o resultado da pesquisa poderá estar seriamente comprometido.

Sistemas Especialistas (SE) têm experimentado tremendo crescimento e popularidade desde sua introdução comercial no início dos anos 80. Hoje, sistemas especialistas são usados em negócios, ciência, engenharia, fábrica e muitos outros campos [GIA98].

Todo sistema especialista é desenvolvido para desempenhar um papel específico, e o papel do sistema especialista ao qual este protótipo dará origem, é servir de intermediário entre um usuário e um *software*

estatístico, identificando os dados categorizados, para evitar inferências estatísticas inadequadas.

2 Análise Estatística de Dados

A análise estatística de dados, é fundamental para diversas áreas, tais como: ciências, tecnologia, comércio, política, etc., e para se ter uma boa análise estatística de dados, deve-se seguir as seguintes etapas:

- definir cuidadosamente o problema;
- formular um plano para a coleta adequada dos dados;
- coligir os dados;
- analisar e interpretar os dados;
- relatar as conclusões de maneira que sejam facilmente entendidas por quem as for usar para tomar decisões.

Todas essas etapas tem importância fundamental na análise estatística de dados e a conclusão final depende muito da fase de análise e interpretação dos dados. É nessa fase que os dados recebem vários tipos de tratamento, para que se possa calcular os erros que serão cometidos ao tirar algum tipo de conclusão. O tratamento dos dados é feito de acordo com a mensuração das variáveis: nominal, ordinal, intervalar e de razão. Os dois primeiros níveis levam às variáveis *qualitativas* e os outros às variáveis ditas *quantitativas*, um dado qualitativo tem um tratamento diferente de um dado quantitativo. Alguma variáveis como sexo, educação, estado civil, etc. apresentam como possíveis realizações de uma qualidade (ou atributo) do indivíduo pesquisado, ao passo que outras como número de filhos, salário, estatura, etc. apresentam como possíveis realizações números resultante de uma contagem ou mensuração. As variáveis do primeiro tipo são chamadas *qualitativas* e as do segundo tipo são chamadas *quantitativas* [BUS87]. Para saber mais, consulte: [MON97], [BAR98A], [BAR98B] e [KAR82].

3 Sistemas Especialistas

3.1 O que é um sistema especialista?

Há várias definições para sistemas especialistas, dentre elas pode-se citar: sistemas especialistas são programas que desempenham tarefas sofisticadas que antes pensavam ser possível apenas para especialistas humanos [BEN91]; sistema especialista é um programa inteligente de computador que usa conhecimento e procedimento de inferências para resolver problemas que requerem significativa especialidade humana para resolvê-los [GIA98].

Basicamente um sistema especialista é composto de: base de conhecimento, mecanismo de inferência, interface para aquisição de conhecimento e interface para usuário.

3.2 Arquitetura dos sistemas especialistas

A figura 1 mostra os quatro componentes de sistemas especialistas típicos:

- a base de conhecimento consiste de conhecimento específico sobre alguma área de domínio;
- o mecanismo de inferência usa regras gerais de inferência para analisar conhecimento explícito na base de conhecimento e para inferir conclusões adicionais, as quais podem ser explicitamente mantidas;
- a interface de aquisição de conhecimento auxilia os especialistas a expressar conhecimento de uma forma que possa ser armazenada na base de conhecimento;
- a interface para usuário auxilia os usuários na consulta ao sistema, providenciando para eles informações requeridas para resolver seus problemas.

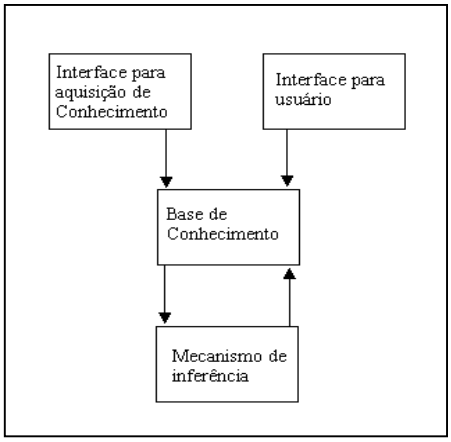

Figura 1: Componentes de sistemas especialistas.

3.3 Construção da base de conhecimento

Uma base de conhecimento para um sistema especialista, é construído através de um processo interativo de desenvolvimento, após projeto inicial e implementação de protótipo, o sistema cresce gradativamente. Para construir a base de conhecimento, basicamente necessita-se de: aquisição de conhecimento e validação das regras.

3.3.1 Aquisição de conhecimento

Aquisição de conhecimento é a transferência e transformação da habilidade de resolver problema, de alguma origem do conhecimento para o programa [BUC84].

A transferência e transformação necessária para apresentar habilidade para um programa computacional pode ser automatizada ou parcialmente automatizada em alguns casos especiais.

Na maioria das vezes, o profissional chamado de engenheiro de conhecimento, é requisitado para possibilitar a comunicação entre o especialista e o programa, este processo é mostrado na figura 2.

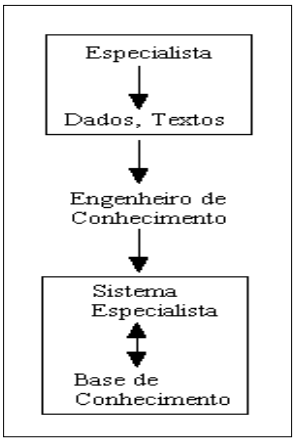

Figura 2: Comunicação entre o especialista e o programa.

3.3.2 Consolidação das regras

Os construtores de um sistema especialista baseado em regras, devem assegurar que o sistema irá dar aos seus usuários um resultado preciso ou uma solução correta de seus problemas [BUC84].

O processo de checagem de uma base de conhecimento para verificar se ela está correta e completa é o maior problema na construção da base de conhecimento. Este processo envolve teste e refinamento da base de conhecimento a fim de descobrir e corrigir uma variedade de erros que podem surgir durante o processo de transferência da habilidade de um especialista humano para um sistema de computador. Para mais detalhes, consultar [BUC84], [HAR92] e [KID87].

3.4 Validação do Sistema Especialista

O Sistema Especialista após a fase inicial de desenvolvimento, deve ser testado dentro do domínio para o qual fora proposto, aplicando-o para resolver problemas reais e verificando o seu comportamento, bem como os erros cometidos e, consequentemente corrigindo-os. Quando esses erros chegarem a um nível aceitável, então o Sistema Especialista estará validado para ser usado.

3.5 Vantagens dos Sistemas Especialistas

Sistemas Especialistas tem muitas vantagens sobre os especialistas humanos, dentre elas, tem-se:

- aumento de disponibilidade: o sistema especialista fica disponível em determinado computador para ser usado a qualquer momento;
- resposta rápida: resposta rápida ou em tempo real pode ser necessário para algumas aplicações. Dependendo do *software* ou *hardware* usado, um sistema especialista pode responder mais rápido e mais confiável que um especialista humano;
- permanência: o Sistema Especialista é permanente, diferente do especialista humano que pode afastar-se, sair ou morrer [GIA98].

4 Heurísticas

Heurísticas são critérios, métodos, para decidir qual entre muitas alternativas de ação promete ser mais efetivo no sentido de atingir algum objetivo. Tais critérios representam comprometimentos entre dois requisitos: a necessidade de tal simples critério e, ao mesmo tempo o desejo de vê-los discriminado corretamente entre boa e má escolha [PEA84].

Dentre tais critérios/métodos, tem-se representação de conhecimento, algoritmos de busca, árvore de decisão, dentre outros. O uso de tais critérios/métodos tem por objetivo, descobrir alguma solução ótima para determinado problema. Solução esta que serve apenas para o problema específico, porque não é fácil encontrar solução para os problemas em geral, levando em conta que cada problema tem sua característica e suas particularidades. Mais detalhes em [BAR97], [GIA98], [BIT98] e [PEA84] .

5 Descrição do SETip

O SETip carrega bases de dados do tipo: *.db, *.dbf e *.xls, bases estas contidas em algum arquivo do usuário. Depois que o usuário carrega a base de dados com a qual deseja trabalhar, lhe é dado a opção de escolher a variável com a qual deseja trabalhar, e ao fazer isso, o programa verifica o tipo da variável,

alertando ao usuário, com relação ao tipo de tratamento que pode ser

dado a tal variável. Isto é feito, seguindo o algoritmo a seguir.

6 Algoritmo do programa

O programa trabalha seguindo o algoritmo:

1 - Carregar a base de dados do usuário;

2 - Atribuir os valores da variável escolhida pelo usuário, a um vetor dinâmico;

3 – Verificar se os valores são do tipo numérico ou do tipo string. Se for do tipo string, devolver ao usuário a mensagem de que a variável é do tipo qualitativa, caso contrário, ir para o passo seguinte;

4 – Calcular $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, onde X_i é o vetor obtido no passo 2 e n é a quantidade de dados contidos na variável. Em seguida calcular: $D_2 = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$ e verificar se D_2 é menor que 0.9, se for, exibir a mensagem de que a variável pode ser qualitativa, caso contrário, ir para o passo seguinte;

5 – Calcular M = $\frac{Maior + Menor}{2}$, onde: Maior e Menor são o valor máximo e mínimo do vetor obtido no passo 2, respectivamente. Em seguida, calcular: P = Menor*1.5 e verificar se M é igual P, se for, exibir a mensagem de que a variável pode ser qualitativa, caso contrário, ir para o passo seguinte;

6 – Calcular $T_2 = \sqrt{\sum_{i=1}^{n} \frac{X_i - Min}{Max - Min}}$, onde Max e Min são os valores máximos e mínimos do vetor do passo 2, respectivamente e $\frac{X_i - Min}{Max - Min}$ é a aproximação dos valores do vetor do passo 2 da inversa de uma função do tipo $y = ax^2 + c$, com domínio $D \in [-1,1]$ e a imagem $IM \in [Min, Max]$. Em seguida, calcular : $D_t = \sqrt{\frac{\sum_{i=1}^{n} (X_i - T_2)^2}{n}}$ e verificar se D_t < 0.9, se for, exibir a mensagem de que a variável pode ser qualitativa, caso contrário, ir para o passo seguinte;

7 – Calcular: $E_n = \frac{L}{Max - Min}$, onde: $L = e^{-P}$, $P = \frac{\sum_{i=1}^{n} \frac{0.5}{X_i}}{n}$ (média ponderada dos valores do vetor, com peso de $\frac{1}{2X_i}$, i=1,2,...,n) e $e^{-P} = \frac{1}{\lim_{P \to \infty} (1 + \frac{1}{P})^P}$ e verificar se E_n > 0.11, se for, exibir a mensagem de que a variável pode ser qualitativa, caso contrário, exibir a mensagem de que a variável pode ser quantitativa.

7 Árvore de decisão

A figura 3 mostra a árvore de decisão do sistema, onde cada nodo da árvore indica um teste do tipo “se A, então B” , cada ramo indica um caminho a seguir e a palavra Quali e Quanti significam Variável do tipo qualitativa e quantitativa, respectivamente.

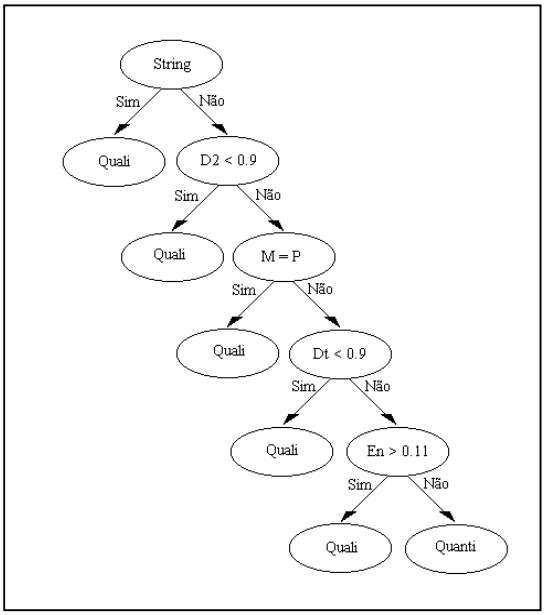

Figura 3: Árvore de decisão do sistema.

8 Modelagem do sistema

A modelagem do sistema fora feita segundo a metodologia UML (*Unified Modeling Language*), porém não serão apresentados todos os diagramas que o

modelo requer, que são: diagrama de *use-cases*, diagrama de classes, diagrama de objetos, diagrama de sequência, diagrama de colaboração, diagrama de atividade, diagrama de componentes e *deployment diagram* (este último foi mantido com o nome original, por não se encontrar uma tradução adequada). Os diagramas aqui mostrados são apenas os principais. A parte teórica da modelagem do sistema não será apresentada, por se tratar uma teoria muita ampla e está fora do contexto deste artigo. Para saber mais a respeito do assunto, consulte: [ERI98] e [RAT__].

8.1 Diagrama de *Use-Cases*

A figura 4 mostra o diagrama de *use-cases* do sistema, este diagrama serve para mostrar o número de atores – atores são usuários, outros sistemas, *hardware*, etc. – externos e suas interações com o sistema.

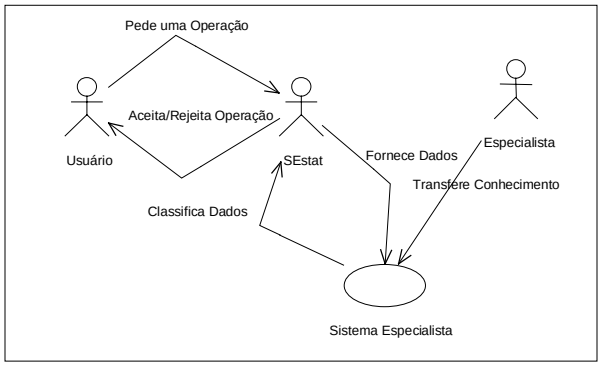

Figura 4: Diagrama de uses-cases do sistema

8.2 Diagrama de classes

A figura 5 mostra os três principais pacotes de classes do sistema, onde o pacote de interface guarda todas as classes destinadas a comunicação com o usuário, o pacote sistema especialista guarda as classes do sistema propriamente dito e o pacote persistente guarda as classes de manipulação de banco de dados.

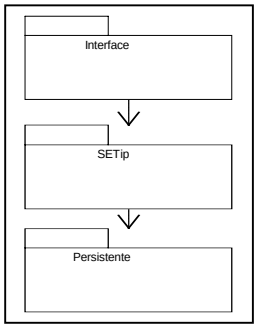

Figura 5: Pacotes de classes do sistema

8.2.1 Classes persistentes

A figura 6 mostra apenas uma classe componente da classe persistente. Essa classe executa apenas as rotinas de manipulação de banco de dados, como mostrado na figura: ler arquivo, selecionar coluna da tabela, gravar arquivo e fechar arquivo.

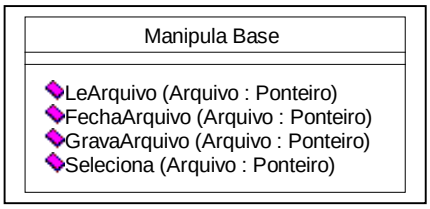

Figura 6: Classe persistente.

8.2.2 Classes do sistema especialista

A figura 7 mostra a principal classe do sistema especialista. Não serão mostrada todas as classes do sistema especialista, pois o objetivo aqui é apenas dar uma idéia do funcionamento do sistema.

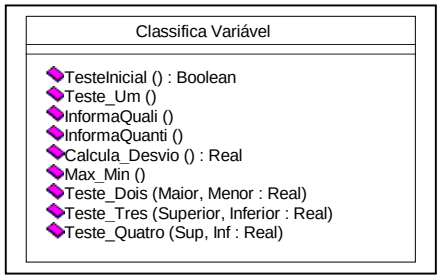

Figura 7: Classe principal do sistema

9 Testes realizados

Foram realizados vários testes com o protótipo, usando bases de dados de professores do departamento de estatística, dos alunos do curso de graduação em geral da UFSC e também dados conseguido através da internet, sendo que todas essas bases se diferenciavam ou no número de atributos (variáveis) ou na quantidade de casos (número de registros). Sendo que essas bases de dados abordavam diversos tipos de pesquisa, dentre elas: hábitos de uma determinada população, pesquisa de opinião pública, etc., o total das variáveis que foram testadas, chegou a 184.

10 Resultados obtidos

Das 184 variáveis testadas, obteve-se um bom índice de acerto por parte do protótipo, tais resultados são mostrados na tabela 1, observando que os percentuais foram arredondados.

Tabela 1: Resultados obtidos com os testes.

Classificação	Total	Percentual
Acertou	165	90%
Errou	8	4%
Não classificou	11	6%
Total	184	100%

11 Análise dos resultados

Como pode-se observar pela tabela 1, o erro cometido pelo protótipo é pequeno em relação ao percentual de acerto, mas vale lembrar que os testes foram feitos com um número não muito grande de variáveis, sendo que esse erro pode aumentar ou diminuir. Os teste iniciais, no começo do trabalho foram feitas com poucas variáveis - na ordem de 20 -, e a quantidade de variáveis foram aumentando à medida em que se conseguia mais bases de dados, porém o erro esteve oscilando em um intervalo de 3 a 7%, levando a crer que o protótipo está funcionando razoavelmente bem. O que observou-se durante os testes, é que quando as bases de dados eram originadas de uma pesquisa realizada por professores, o nível de acerto chegou até a 100%, ou seja, ou protótipo está muito bom para uma base de dados bem consistente. Mas é obvio que isso pode ser melhorado, e será feito à medida que mais testes forem realizados.

12 Conclusão

Como foi exposto no decorrer deste artigo, isto é apenas um protótipo de um sistema especialista que tem como objetivo, fazer uma classificação de variáveis estatísticas, para que o usuário trate corretamente cada variável, para que o resultado da pesquisa não seja comprometido. Portanto, até sua versão final, mais testes serão realizados para que esse objetivo seja alcançado.

13 Bibliografia

[BAR98A] BARBETTA, Pedro A. Estatística Aplicada às Ciências Sociais. Florianópolis: Editora da UFSC, 1998.

[BAR98B] BARBETTA, Pedro A. Tese de Doutorado. Florianópolis: PEPS 0753, 1998

[BAR98] BARRETO, Jorge M. Inteligência Artificial no Limiar do Século XXI. Florianópolis: $\rho\rho\rho$ Edições, 1997.

[BEN91] BENFER, Robert A. et. al. Experts Systems. Califórnia, USA: SAGE Publications,Inc.,1991.

[BIT98] BITTENCOURT, Guilherme. Inteligência Artificial: Ferramentas e Teorias. Florianópolis: Ed. Da UFSC, 1998.

[BUC84] BUCHANAN, Bruce G. Rule-Based Experts Systems. USA: Addison Wesley Publishing, Company, Inc., 1984.

[BUS87] BUSSAB, Wilton O. & MORETTIN, Pedro A. Estatística Básica. São Paulo: Atual, 1987.

[ERI98] ERIKSSON, Hans-Erik & PENKER, Magnus. UML toolkit. USA: John Wiley & Sons, Inc.,1998.

[GIA98] GIARRATANO, Joseph C. & RILEY, Gary. Experts Systems: Principles and Programing. USA: ITP, 2nd ed., 1998.

[HAR92] HART, Ana. Knowledge Aquisition for Experts Systems. N.Y. – USA: McGraw-Hill, 1992.

[KAR82] KARSMIER, Leonardo J. Estatística Aplicada à Economia e Administração. São Paulo: McGraw-Hill do Brasil, 1982.

[KID87] KIDD, Alison L. Knowledge Aquisition for Experts Systems. N.Y. – USA: Plenum Press, 1987.

[MON97] MONTGOMERY, Douglas C. Design and Analysis of Experiments. New York: John Wiley & Sons, Inc., 4th ed., 1997.

[PEA84] PEARL, Judea. Heuristics: intelligent Strategies for Computers Solving. Califórnia- USA: Addison-Wesley Publishing Company, Inc., 1984.

[PRE99] Anais do PRESTA – Conferência Internacional: Experiências e Perspectivas do ensino da Estatística, Florianópolis, 20,21 e 22 de setembro de 1999.

[RAT__] www.rational.com/uml.