Aplicação de Métodos de Inteligência Artificial em Inteligência de Negócios

Rogério Garcia Dutra (POLI USP) rogerio.dutra@axiaconsulting.com.br

Resumo

*O objetivo deste artigo é combinar métodos de redes neurais artificiais e árvores de decisão para prospecção de padrões de classificação de clientes, como parte de um processo interativo de descobrimento e análise de regularidades, regras e associações em uma base de dados proveniente de processos empresariais de vendas .Como resultado dessa combinação de métodos de inteligência artificial, espera-se maximizar as vantagens de ambos métodos em um modelo denominado de Neural Tree Network (**NTN**), empregado como ferramenta de Inteligência de Negócios (BI) em atividades de exploração de dados (Data Mining) para utilização em sistemas de apoio à decisão.*

Palavras Chave: Redes Neurais Artificias, Árvores de Decisão, Sistemas de Apoio à Decisão.

1. Introdução

As corporações brasileiras sentiram nos últimos anos os sintomas de uma febre mundial que contaminou quase todas as empresas do globo, denominado de efeito ERP (*Enterprise Resource Planning*) ou simplesmente pacotes de gestão empresarial. Fenômeno típico da década de Noventa, que sucedeu a redução (*downsizing)* dos sistemas de grande porte em plataforma *mainframe*, essas soluções resolvem apenas o dia a dia operacional das companhias, isto é, os dados transacionais, gerando continuamente enormes quantidades de informação em estado "bruto". Com o propósito de garimpar e lapidar tais dados criou-se o conceito de **Inteligência de Negócios**.

Existem várias ferramentas para implementação do conceito de Inteligência de Negócios, variando desde planilhas eletrônicas até sofisticados sistemas de suporte à decisão baseados em *Data Warehouse* com ferramentas analíticas de *Data Mining*, cuja complexidade depende fundamentalmente da aplicação.
Este trabalho objetiva a exploração das ferramentas de *Data Mining,* visando a aplicação em gerenciamento de relações de empresas com seus clientes, através da combinação de métodos de Inteligência Artificial, tais como *Redes Neurais Artificiais* e *Árvores de Decisão* um modelo denominado de *Neural Tree Network* (**NTN**).

1.1 Inteligência de Negócios

Inteligência de Negócios ou ***Business Intelligence*** (BI) trata-se de um conjunto de conceitos e metodologias que, fazendo uso de dados transacionais e sistemas baseados nos mesmos, suporta a tomada de decisões em negócios com o objetivo de transformar informação em valor agregado ao negócio. Os sistemas de BI têm como principais características:

a) Extrair e integrar dados de múltiplas fontes;
b) Fazer uso da experiência e conhecimento adquirido por seus usuários;
c) Analisar dados dentro de uma cadeia de processos de negócios;
d) Trabalhar com múltiplas hipóteses e simulações;
e) Extrair padrões de comportamento e classificá-los em categorias.

Existem várias ferramentas para implementação do conceito de BI, cuja complexidade depende fundamentalmente da aplicação. Tais ferramentas abrangem desde planilhas

eletrônicas até complexos sistemas de suporte a decisão baseados em *Data Warehouse* com ferramentas analíticas de *Data Mining*[MILL98], foco deste artigo.

1.2 Descrição do problema alvo

O problema alvo desta dissertação de mestrado consiste em classificar clientes através dos dados reais provenientes da realização de processos de vendas e distribuição. Esta base de dados, pertencente a uma empresa multinacional do setor químico comercializa no Brasil produtos de beleza, será descrita em detalhes no capítulo 5 deste artigo.

A figura 1.1 ilustra os métodos de inteligência artificial, rede neural artificial e árvore de decisão, que compõem o modelo da *Neural Tree Network*, utilizado como ferramenta de inteligência de negócios para obtenção das classes que permitirão a estratificação de clientes necessária à aplicação de uma metodologia de CRM [PEPP00].

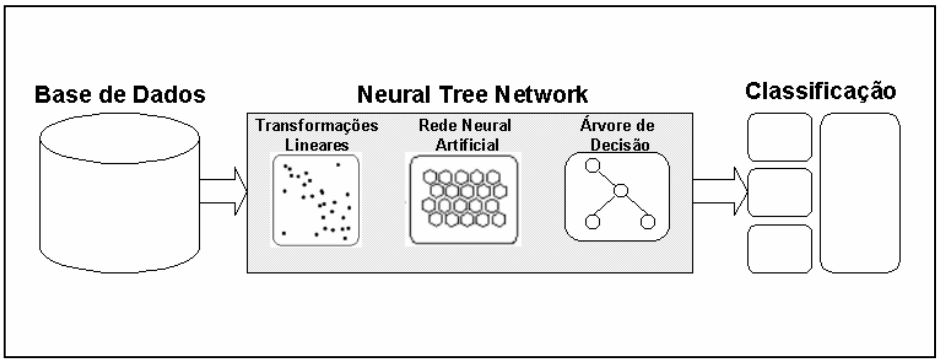

Fig. 1.1 – Modelo Neural Tree Network aplicado na solução do problema

Apesar o número de clientes que compõem a base de dados pertencerem a um conjunto aberto, neste artigo o mesmo será considerado como um conjunto finito e completamente conhecido dentro de um intervalo de tempo determinado. Porém o número de possíveis classes não é conhecido a priori, cabendo ao modelo **NTN** sua determinação.

2. Machine Learning

Bases de dados proveniente de processos empresariais podem atingir uma quantidade de registros elevadíssima, tornando impossível a realização de atividades de *data mining* por parte de um especialista humano. Face a essa necessidade, diferentes métodos foram criados para automatizar e sistematizar o processo de prospecção, análise de padrões e relacionamento dos mesmos com seus respectivos dados, denominados de *machine learning* (ML) [DHAR97].

Dentre os diversos métodos de ML, tais como: métodos baseados em regras, método do discriminante de Bayes, métodos heurísticos de lógica *fuzzy*, entre outros; escolheu-se nesta dissertação, trabalhar com métodos partitivos recursivos denominados de árvores de decisão ou *Decision Trees* (DT), devido às características citadas a seguir:

a) Permitem a redução do volume de dados através da transformação para um formato mais compacto, porém sem perder as principais características e relacionamento dos mesmos.
b) Permitem descobrir se os conjuntos de dados contém agrupamentos de objetos, que podem ser úteis para simplificações e particionamento dos mesmos.
c) Permitem mapear o relacionamento entre variáveis independentes e dependentes, objetivando a construção de um modelo classificatório preditivo.

O método C4.5 [QUIN92] tem sido largamente empregado para construir DT que implementam classificadores de elevada performance. Contudo, este algoritmo só permite

classes previamente definidas para classificação de atributos na fase de treinamento, não tendo a capacidade de interpolar ou deduzir novos padrões por inferência nesta fase.

3. Redes Neurais Artificiais

O modelo *Self-Organizing Map* (SOM) desenvolvido por Teuvo Kohonen [KOHO95] é um dos modelos mais populares de RNA. O algoritmo da SOM é baseado em um aprendizado competitivo e não supervisionado, o que implica em um treinamento direcionado exclusivamente pelos dados, sendo que os neurônios que constituem o mapa competem entre si para adquirir padrões dos dados, se aproximando deles. Algoritmos supervisionados, como o *Multi-Layered Perceptron* (MLP), requerem uma classificação pré-definida para cada vetor de treinamento, além de depender fundamentalmente do número de camadas internas (*hidden units*) para um aprendizado com baixo erro de classificação e performance, limitações as quais não ocorrem na arquitetura SOM.

Data Mining é uma área emergente de pesquisa em inteligência artificial e inteligência de negócios. O propósito de *data mining* é extrair conhecimento de base de dados cuja dimensão, complexidade e volume de dados seriam proibitivos a um observador humano. Algumas atividades típicas para realização de *data mining* são classificação, regressão, agrupamento, sumarização e modelagem de dependências. Devido às características de aproximação da função densidade de probabilidade e um alto poder de visualização propiciando uma melhor investigação dos dados, a SOM é o algoritmo de redes neurais artificiais que, combinado as vantagens das árvores de decisão, que possui um alto grau de aderência às atividades de *data mining*.

4. O modelo da Neural Tree Network

Evidentemente existem inúmeras formas de se combinar DT e RNA para formar o modelo da ***Neural Tree Network***, porém a arquitetura escolhida neste trabalho objetiva suportar as atividades de *Data Mining*, como é ilustrado na figura 4.1, a seguir:

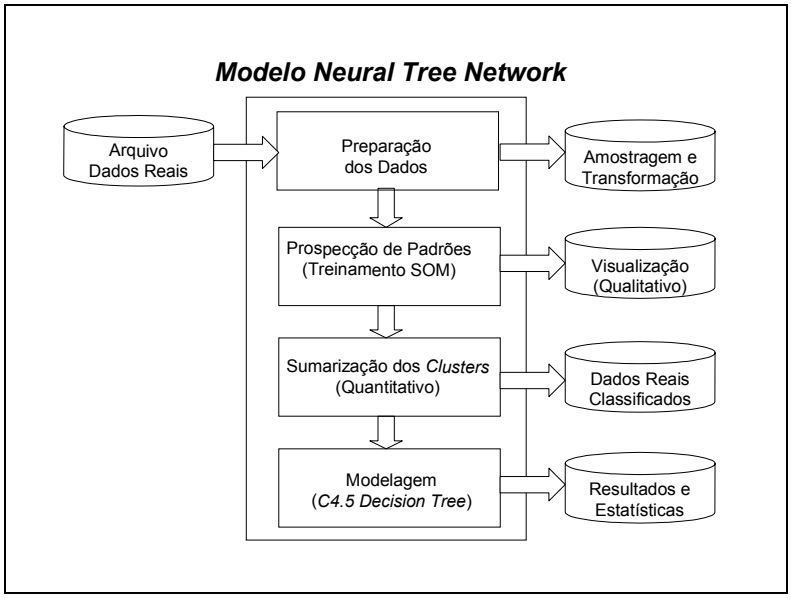

Fig. 4.1 – Diagrama de Blocos da Neural Tree Network (**NTN**)

A seguir serão descritos as variáveis entradas e saídas de cada bloco constituinte do modelo da **NTN** e as respectivas atividades realizadas para transformação dos dados.

5. Implementação e Aplicação da Neural Tree Network

Este capítulo terá como enfoque a obtenção e análise dos padrões de forma qualitativa e quantitativa, bem como a definição dos agrupamentos ou classes que permitam classificar amostras da base de dados.

A base de dados utilizada é composta de características inerentes ao processo de vendas e distribuição e indicadores (variáveis de conteúdo numérico) conforme ilustrado na tabela 5.1. A partir da tabela 5.1 construi-se dois conjuntos de dados para treinamento do algoritmo da SOM, denominados de *Treinamento_Parcial.data* e *Treinamento_Completo.data*. O conjunto *Treinamento_Parcial.data* não contem as variáveis **Num_Rem** e **Prc_Med** para verificar o efeito do acréscimo de variáveis sobre o número de classes obtidas a partir do treinamento da SOM.

Indicadores						**Classificação**
Volume de Vendas (Vol_Ven)	Quantidade de Vendas (Qtd_Ven)	Volume Devoluções (Vol_Dev)	Quantidade Devoluções (Qtd_Dev)	Preço Médio (Prc_Med)	Número de Remessas (Num_Rem)	A, B ou C (ABC)

Tabela 5.1 – Estrutura Completa da Base de Dados para treinamento da SOM

As próximas etapas resumem-se em preparar a base de dados para treinamento da rede neural artificial utilizando o algoritmo SOM e obter padrões qualitativos, através da visualização do formato e estrutura dos agrupamentos, mapa de componentes e projeções dos dados sobre o mapa, conforme ilustrado na figura 5.1.

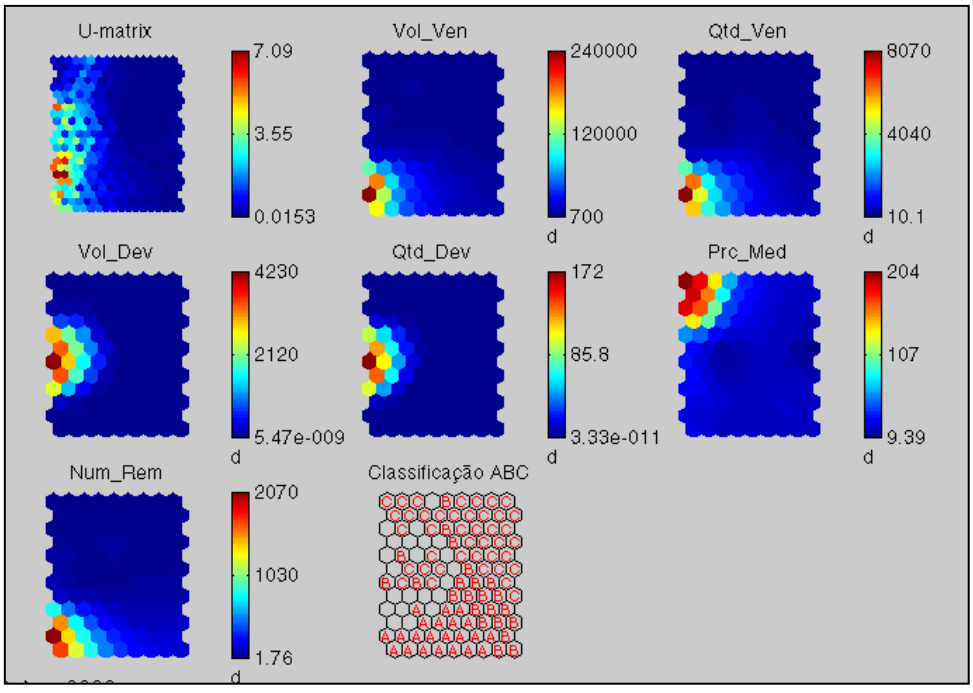

Fig. 5.1 – *U-Matrix* e mapas componentes após treinamento da SOM utilizando o arquivo *Treinamento_Completo.data*

Finalmente foi realizado uma quantificação do número de agrupamentos através da utilização do índice de *Davies-Boulding* [DAVI79] associado ao algoritmo *k-means*. Aplicando-se este algoritmo à base de dados *Treinamento_Completo.data,* o índice de *Davies-Boulding* atingiu o valor mínimo para 4 classes . A cada amostra da base de dados foi associada uma das 4

classes, formando um novo uma conjunto de dados classificados. A partir desse novo conjunto foram efetuadas diferentes tentativas para a determinação de uma árvores de decisão que minimizasse o erro de classificação.A técnica de validação cruzada dos dados de treinamento e testes foi utilizada para obter tal árvore.

Todas as funções utilizadas no treinamento do algoritmo da SOM e o resultados obtidos através do mesmo nos capítulos 5 e 6, basearam-se em funções previamente elaboradas em Matlab provenientes da SOMTOOLBOX 2.0 [VESA99].

6. Resultados Obtidos

A árvore de decisão resultante da 2ª. execução do algoritmo C4.5, utilizando a técnica de validação cruzada dos dados na etapa de modelagem, ilustrada através da figura 6.1, representa o principal resultado do modelo da ***Neural Tree Network***. Sua interpretação permitirá não somente a obtenção de resultados quantitativos, mas também qualitativos.

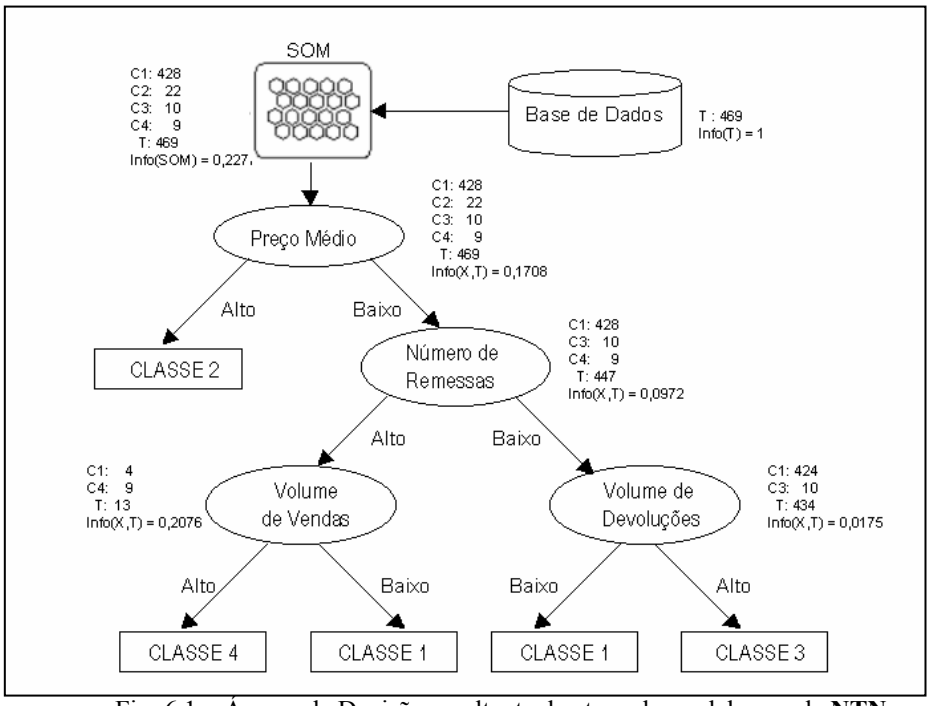

Fig. 6.1 – Árvore de Decisão resultante da etapa de modelagem da **NTN**

Na figura 6.1, C1 representa a Classe 1, *T* representa o total de amostras em determinado ramo da árvore e *Info(X,T)* representa a medida da informação devido a partição induzida pela variável *X* no conjunto *T*.

7. Conclusão

A conclusão deste trabalho avalia quantitativamente as atividades de transformação dos dados realizadas em cada etapa do modelo da **NTN** através da utilização do conceito de entropia de informação. Além da avaliação quantitativa, comparou-se qualitativamente aos requisitos do problema alvo e aos métodos que o compõem. Baseado nos resultados obtidos, definiu-se um significado para as classes definidas pela **NTN**, empregada como ferramenta de Inteligência de Negócios (BI) em atividades de extração de dados (*Data Mining*) para Gerenciamento de Relações com Clientes (CRM) no contexto do processo de vendas e distribuição analisado.

7.1 Avaliação do modelo da Neural Tree Network

A figura 7.1 resume os principais resultados em cada etapa do modelo da ***Neural Tree Network***, além de ilustrar a redução da entropia antes e após a **NTN**.

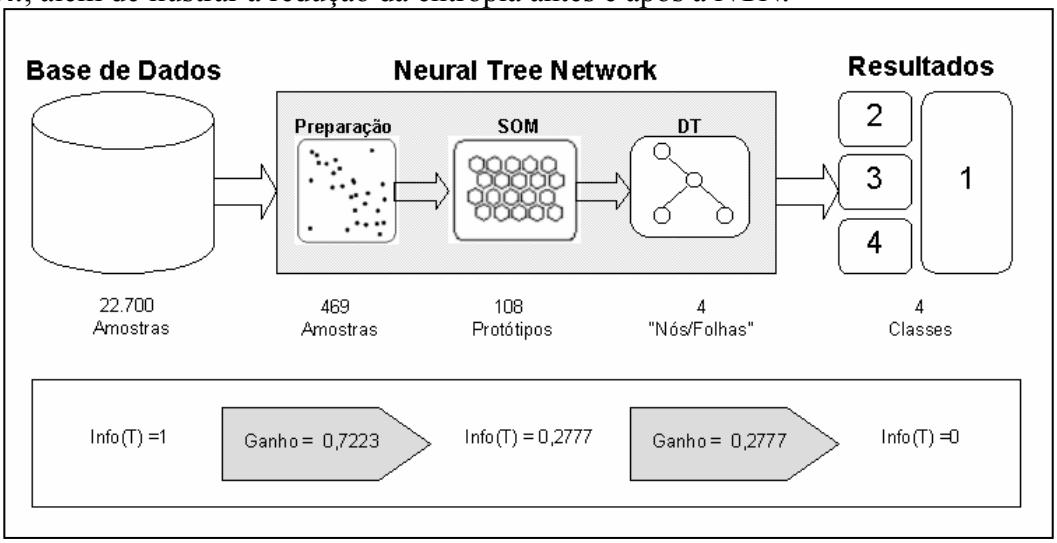

Fig. 7.1 – Principais resultados em cada etapa do modelo da **NTN**

A teoria da informação utiliza-se do termo entropia para quantificar o conteúdo de informação de um determinado agrupamento. Um valor de entropia elevado implica em conteúdo baixo de informação, ou seja, mais informação será necessário para identificar que uma determinada amostra de dados pertence a um determinado agrupamento ou classificação.

Baseado nessa afirmação, o modelo da **NTN** foi capaz de reduzir a entropia da base de dados inicial do valor máximo até seu valor mínimo com um grau de precisão médio, como será descrito a seguir. Observando-se cada etapa do modelo, a etapa de Preparação de dados não reduz a entropia do sistema, ao passo que a SOM foi responsável pela maior redução, isto é, um ganho de aproximadamente 72%, cabendo os 28% restantes à árvore de decisão.

A tabela a seguir sintetiza as principais características avaliadas nos modelos *Neural Tree Network* (**NTN**), *Self-Organizing Map* (SOM) e Árvore de Decisão (DT), através da qual conclui-se que a **NTN** maximiza as vantagens de ambas as técnicas.

CARACTERÍSTICA	SOM	DT	NTN
Precisão	Média	Alta	Média
Inteligibilidade	Baixa	Alta	Alta
Tempo de Resposta	Médio	Alto	Médio
Escalabilidade	Alta	Alta	Alta
Tolerância a ruído nos dados	Alta	Baixa	Alta
Tolerância a dados esparsos	Alta	Baixa	Alta
Curva de Aprendizado	Média	Baixa	Média
Independência de Especialistas	Baixa	Média	Média

Tabela 7.1 – Síntese de características do modelo **NTN** em comparação com as técnicas de inteligência artificial

Nesta tabela, a definição das características da SOM foram obtidas através de [VESA00] e da DT através de [DHAR96].

7.2 Definição do significado das classes resultantes da NTN

A partir da árvore de decisão ilustrada na figura 6.1 e de uma estatística efetuada sobre o número de clientes em cada classificação como mostrado na tabela 7.2, um especialista em

negócios pode inferir um significado para cada classe dentro do contexto do processo de negócios de vendas e distribuição:

Classificação	Clientes (Número)	Clientes (%)	Volume de Vendas (%)
Classe 1	428	91,3	58,3
Classe 2	22	4,7	1,1
Classe 3	10	2,1	1,8
Classe 4	9	1,9	38,8
Total	469	100	100

Tabela 7.2 – Estatística do número de clientes e percentual de faturamento por classe

Classe 1

Representa clientes regulares, isto é, os clientes que compram produtos de baixo preço médio, em pequenas quantidades e normalmente efetuam pequenos volume de devoluções. Estes clientes regulares representaram 91,3 % em número, porém respondem por 58,3% do faturamento em vendas.

Classe 2

Representa clientes que compram produtos de alto valor agregado, isto é, produtos de elevado preço médio em baixas quantidades. Estes clientes, apesar de representarem 4,7% em número, respondem por somente 1,1% do faturamento em vendas.

Classe 3

Representa clientes com elevada ocorrência de devoluções. Apesar do baixo percentual em número, ou seja 2,1%, representaram um faturamento em vendas de 1,8%.

Classe 4

Representam clientes ótimos, isto é, clientes que compram produtos de baixo preço médio, porém em quantidades elevadas. Respondem por um faturamento de vendas de 38,8 %, apesar de representarem somente 1,9% em número.

Baseado no significa de cada classe, um especialista de negócios utilizando o modelo da **NTN** como ferramenta para estratificação de clientes dentro do conceito de gerenciamento de relações com clientes, concluiria que:

a) Os clientes segmentados através da Classe 4 deverão receber tratamento personalizado. Todos os esforços de interação com os mesmos deverão buscar o melhor atendimento em função de suas necessidades e o aumento de satisfação;
b) Os clientes segmentados através da Classe 1 devem continuar fazendo parte do processo de prospecção e análise através da **NTN**, aguardando uma possível modificação para a Classe 4, porém sem aumento de esforço operacional para a empresa para que a mudança ocorra;
c) Os clientes segmentados através da Classe 2 representam clientes potenciais cujo aumento da quantidade de vendas poderá promovê-los à Classe 4 com menor esforço operacional que os clientes da Classe 1 e portanto justificam tratamento personalizado;
d) Os clientes segmentados através da Classe 3 deverão ter seu processo de vendas e distribuição revisto para que os possíveis problemas por parte da empresa sejam resolvidos resultando em uma mudança de classificação para a Classe 1. Caso contrário deverão ser eliminados do cadastro de clientes.

Evidentemente que as ações a serem tomadas baseadas nas conclusões inferidas através do *Neural Tree Network* como ferramenta de inteligência de negócios, deverão ser compatíveis

com o planejamento estratégico de médio e longo prazo da empresa para manter ou aumentar sua participação de mercado para os produtos rentáveis e diminuir sua participação nos produtos pouco rentáveis.

Além das conclusões baseadas na interpretação de um especialista em negócios, deve-se levar em consideração as limitações da ferramenta face os requisitos desejáveis para solucionar o problema de classificação de clientes baseado nos dados oriundos do processo de vendas e distribuição dos mesmos, como descrito a seguir.

7.3 Avaliação dos requisitos desejáveis para solução do problema

A tabela 7.3 permite realizar uma comparação qualitativa entre os requisitos desejáveis para a solução do problema de classificação de clientes descrito na seção 1.1 deste artigo.

CARACTERÍSTICA	DESEJÁVEL	NTN
Precisão	Alta	Média
Inteligibilidade	Alta	Alta
Tempo de Resposta	Alto	Médio
Escalabilidade	Alta	Alta
Tolerância a ruído nos dados	Alta	Alta
Tolerância a dados esparsos	Alta	Alta
Curva de Aprendizado	Baixa	Média
Independência de Especialistas	Alta	Média

Tabela 7.3 – Síntese de características do modelo **NTN** em comparação com os requisitos desejáveis

Através desta tabela, conclui-se que as limitações do modelo da **NTN** ocorreram em características pouco tangíveis, tais como Curva de Aprendizado e Independência de Especialistas, que dependem fundamentalmente da experiência e familiaridade do especialista em negócios com o problema e ferramenta em questão. As caraterísticas Precisão e Tempo de Resposta, relacionadas à qualidade e construção do modelo, foram avaliadas de forma conservadora devido a falta de padrões para uma avaliação mais quantitativa e menos qualitativa.

Referências

[DAVI79] DAVIES, D. L.; BOULDING, D.W. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. PAMI-1, no. 2, pp. 224-277. 1979.

[DHAR97] DHAR, V.; STEIN, R. Seven Methods for transforming corporate data into business intelligence. Prentice-Hall Press. 1997.

[KOHO95] KOHONEN, T. Self-Organizing Maps. Springer-Verlag. 1995.

[MILL98] MILLEY, A. H.; SEABOLT, J. D.; WILLIANS, J. S. Data Mining and the case for sampling. Solving business Problems using SAS® Enterprise Miner™ Software. Best Practices Paper. SAS Institute. 1998.

[PEPP00] PEPPERS & ROGERS GROUP DO BRASIL. Um Guia Executivo para Entender e Implantar Estratégias de Customer Relationship Management. CRM Series Marketing 1 to 1®. 1ª Edição. Janeiro. 2000.

[QUIN92] QUINLAN, J. R. C4.5 Programs for Machine Learning. Morgan Kaufmann. 1992.

[VESA00] VESANTO, J. Using SOM in Data Mining. Thesis for the degree of Licentiate of Science in Technology. Helsinki University of Technology. Finland.2000.

[VESA99] VESANTO, J.; ALHONIEMI, E.; HIMBERG, J.; PARHANKANGAS, J. Som Toolbox 2.0 BETA online documentation. Internet address http:// www.cis.hut.fi/projects/ somtoolbox. 1999.