5CONTECSI June 04 - 06, 2008 São Paulo, Brazil

5th International Conference on Information Systems and Technology Management
5º Congresso Internacional de Gestão da Tecnologia e Sistema de Informação
De 04 a 06 de Junho de 2008 - São Paulo - Brasil

PS-1056

DATA MINING APPLIED TO THE PROCESSES CELERITY OF PERNAMBUCO'S STATE COURT OF ACCOUNTS

Maria Uilma R. S. de Sousa (Universidade Federal de Pernambuco, Pernambuco, Brasil) - murss@cin.ufpe.br
Kelly Patrícia da Silva (Universidade Federal de Pernambuco, Pernambuco, Brasil) - kps@cin.ufpe.br
Paulo Jorge L. Adeodato (Universidade Federal de Pernambuco, Pernambuco, Brasil) - pjla@cin.ufpe.br e (NeuroTech Ltda.) - paulo@neurotech.com.br
Adrian L. Arnaud (Universidade Federal de Pernambuco, Pernambuco, Brasil) - ala2@cin.ufpe.br e (NeuroTech Ltda.) - adrian@neurotech.com.br

Colaborador
Francisco de Assis Tenório de Carvalho (Universidade Federal de Pernambuco, Pernambuco, Brasil) - fatc@cin.ufpe.br

The judgment of the public managers' actions, which is responsibility of the court of accounts, is performed through the use of processes. The celerity in the elaboration of processes leads to an efficient social control tool. In despite of the increasing technology investments and the growth of the number of technicians; the stock of processes has been growing annually. That is, the amount of litigated processes (input) is greater than the amount of judged processes (output). This paper presents a data mining solution, with knowledge acquisition from the court of accounts data base, for a decision support system focusing on the process celerity. In this work, we used the CRoss Industry Standard Process for Data Mining (CRISP-DM) methodology. The proposed solution, based on neural networks, was able to correctly classify 85% of the delayed processes.

Keywords: Decision support systems, Public administration, Data mining, Neural networks, Knowledge discovery in databases.

MINERAÇÃO DE DADOS APLICADA À CELERIDADE PROCESSUAL DO TRIBUNAL DE CONTAS DO ESTADO DE PERNAMBUCO

O julgamento dos atos dos gestores públicos, atribuição dos tribunais de contas, se materializa através de processos. A celeridade na elaboração dos processos resulta em instrumento eficaz de controle social. Apesar dos crescentes investimentos tecnológicos e aumento do quadro de pessoal técnico, o estoque de processos do Tribunal de Contas do Estado de Pernambuco (TCE-PE) vem crescendo anualmente, isto é, a quantidade de processos autuados é maior que a de processos transitados em julgado. Este trabalho apresenta uma solução de mineração de dados, com a extração do conhecimento contido na base de dados do referido tribunal, para o desenvolvimento de um sistema de apoio à decisão com foco na celeridade processual. Para a realização do trabalho, foi seguida a metodologia *CRoss Industry Standard Process for Data Mining* (CRISP-DM). A solução apresentada, baseada em redes neurais artificiais, conseguiu classificar corretamente 85% dos processos com permanência prolongada no TCE-PE.

Palavras-chave: Sistemas de apoio à decisão, Controle externo, Mineração de dados, Redes neurais artificiais, Descoberta de conhecimento em bases de dados.

1. Introdução

O controle sobre a totalidade da administração pública, exercido pelos que representam, por delegação, a sociedade politicamente organizada, é denominado Controle Externo e constitui-se em um dos pilares das democracias modernas. No Brasil, o controle externo é exercido pelos Tribunais de Contas, órgãos integrantes dos poderes legislativos estaduais e federal [Art. 71, CF, 1988], que visam a garantir o estrito respeito aos princípios fundamentais da administração pública - legalidade, impessoalidade, moralidade, publicidade e eficiência [Art. 37, CF, 1988].

O Tribunal de Contas do Estado de Pernambuco (TCE-PE) é responsável pelo julgamento dos atos exercidos pelos gestores públicos, tanto na esfera estadual quanto na municipal, constituindo-se num instrumento de controle social à disposição do cidadão pernambucano. Por controle social "entende-se a participação da sociedade no acompanhamento e verificação das ações da gestão pública na execução das políticas públicas, avaliando objetivos, processos e resultados" [TVE Brasil].

O julgamento dos atos dos gestores é, portanto, o serviço público prestado pelo TCE-PE, o qual se materializa através do processo, formalmente autuado. A celeridade na elaboração dos processos, que resulta em resposta tempestiva à sociedade, consiste em um dos principais indicadores de excelência daquele tribunal.

Apesar dos crescentes investimentos tecnológicos, infra-estrutura, logística, etc. e considerável aumento do quadro de pessoal técnico, o estoque de processo do TCE-PE vem crescendo anualmente, isto é, a quantidade de processos autuados (entradas) é maior que aquela de processos transitados em julgado (saídas).

Por força de lei [inciso II, Art. 71, CF, 1988], os Tribunais de Contas são obrigados a julgar todo e qualquer gasto de recursos públicos. Por essa razão, a sua atuação no gerenciamento da **entrada** de processos (aumento do estoque) está restrita a ações de natureza organizacional. Resta-lhe, portanto, atuar diretamente na **saída** de processos (redução do estoque).

Desde a entrada, no TCE-PE, até a saída, os processos passam por cinco fases: formalização, instrução, julgamento, publicação e encerramento. Essas fases são realizadas de maneira seqüencial. É a celeridade na instrução e julgamento dos processos, chamada celeridade processual, a ação apontada como alternativa para resolver o problema da evolução negativa do estoque de processos.

Diante do exposto, observa-se a necessidade de busca de novas alternativas para a solução da evolução negativa do estoque de processo, que se enquadra num problema da teoria das filas [Prado, 2004].

Este trabalho apresenta uma solução de mineração de dados, com a extração do conhecimento contido na própria base de dados que compõe o estoque de processo do TCE-PE, para a proposição de um sistema de apoio à decisão como instrumento de atuação efetiva na celeridade de execução dos processos daquela Corte de Contas.

Tendo como objetivo principal propor um sistema de apoio à decisão para fase de instrução, foram tratados apenas os dados *a priori* a essa fase, ou seja, dados gerados na fase de formalização. A fase de instrução foi escolhida por tratar-se da fase que consome mais tempo. Nessa fase, são coletadas e organizadas as informações-chave para o julgamento do processo.

O artigo está organizado em 7 seções. A seção 2 descreve a metodologia utilizada para o desenvolvimento das tarefas de mineração. As seções 3 e 4 apresentam o entendimento e a preparação dos dados. A seção 5 traz a modelagem da solução e em seguida, na seção 6, são avaliados os resultados. Finalmente, na seção 7, são apresentadas as conclusões e propostas de trabalhos futuros.

2. Metodologia

Para o desenvolvimento do presente trabalho, dentre as metodologias disponíveis para a execução de projetos de mineração de dados, foi seguido o *CRISP-DM (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)* [Colin Shearer 2000]. Trata-se de uma metodologia padrão, não proprietária, que está estruturada em torno das tarefas e objetivos para cada uma das fases do projeto de mineração de dados. A Figura 1 apresenta o fluxograma das fases de trabalho do *CRISP-DM*.

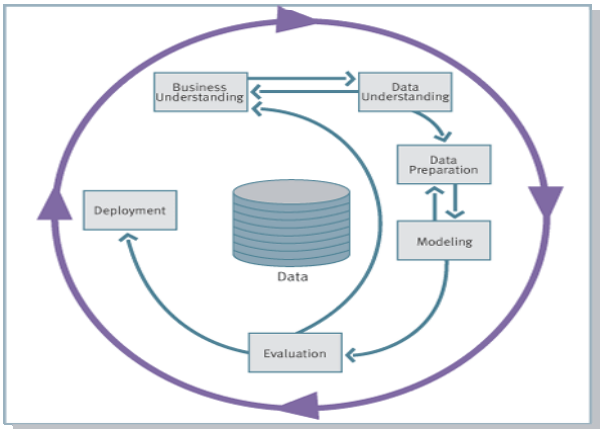

Figura 1- Fases do CRISP-DM

Segundo essa metodologia, a execução de um projeto de mineração de dados está estruturada em seis fases interdependentes. A saber:

Fase 1 – Entendimento do negócio (*Business Understanding)* – tem por objetivo identificar as metas e requisitos a partir de uma perspectiva de negócio, e então convertê-los para uma aplicação de mineração de dados e um plano inicial de ataque ao problema.

Fase 2 – Entendimento dos dados (*Data Understanding)* –essa fase tem por finalidade determinar quais os dados disponíveis e onde se encontram, e, como atividade principal, extrair uma amostra dos dados a serem usados e avaliar o ambiente em que eles se encontram.

Fase 3 – Preparação dos dados (*Data Preparation)* –essa fase tem por objetivo adaptar e transformar os dados no formato apropriado às respostas que se procuram para a criação de programas de extração, limpeza e transformação dos dados para uso pelos algoritmos de *data mining*. Alguns algoritmos necessitam dos dados em formatos específicos, o que acaba causando vários retornos à fase de preparação dos dados.

Fase 4 – Modelagem (*Modeling)* – nessa fase, são criados modelos explicativos das necessidades a satisfazer, seleção do(s) algoritmo(s) a ser(em) utililizado(s) e efetivo processamento do modelo.

Fase 5 – Avaliação (*Evaluation)* - tem por finalidade verificar se os resultados obtidos satisfazem os objetivos do projeto. Ao final da fase de modelagem, vários modelos devem ter sido avaliados sob a perspectiva do analista responsável. Agora, o objetivo passa a ser avaliar os modelos com a visão do negócio, se certificando de que não existem falhas ou contradições com relação às regras do negócio.

Fase 6 – Implantação (*Deployment)* – tem por objetivo disponibilizar os resultados do projeto aos tomadores de decisão. A criação e validação do modelo permitem avançar mais um passo, no sentido de tornar o conhecimento gerado acessível. Isto pode ser feito de várias maneiras, desde a criação de um software específico para tal até a publicação de um relatório para uso interno. Neste trabalho, não foram executadas as tarefas referentes à fase 6. O trabalho apresentado encerra-se com a conclusão das atividades previstas na fase 5.

3. Desenvolvimento

3.1 Entendimento dos dados

Esta etapa do trabalho corresponde à segunda fase do *CRISP-DM*, *Data Understanding*, em que é feita uma análise inicial dos dados disponíveis. A primeira fase, *Business Understanding*, foi abordada na introdução.

Os dados utilizados neste estudo foram obtidos do sistema de Acompanhamento de Processo (AP) do Tribunal de Contas do Estado de Pernambuco (TCE-PE) e referem-se ao estoque total de 98.624 processos julgados até 30 de novembro de 2006.

Os dados foram extraídos em arquivo único do tipo Access® (extensão mdb) diretamente do banco de dados corporativo do TCE-PE, cujos atributos foram definidos com o auxílio de alguns especialistas do domínio. Posteriormente, durante o pré-processamento, fase em que os dados são colocados no formato apropriado para utilização das técnicas, os dados foram estratificados por tipo de processo e gravados em arquivo Excel® (extensão xls), onde foram realizadas consultas *OLAP* (*On line Analytical Processing*) [Chaudhuri & Dayal, 2003] e algumas tarefas de pré-processamento, as quais foram complementadas com a ferramenta WEKA[1] [Witten & Frank, 2005].

A estratificação por tipo de processo foi necessária tendo em vista a diversidade entre o tempo ideal de permanência de cada tipo. Na Tabela 1 são relacionados os tipos de processo e o resultado da análise do volume dos dados. Partindo de um montante de 98.624 registros, excluíram-se 53.079 o que resultou num total de 45.545 processos. Do total de registros excluídos, 49.257 são de um único tipo de processo que, no momento da execução deste trabalho, já estava extinto, o que justifica sua exclusão da base. Os demais apresentavam dados faltosos que não foi possível recuperar.

Como instrumento para entendimento dos dados, é apresentado um histograma (Figura 2) com a distribuição do tempo total de permanência dos processos no TCE-PE. Considerando a particularidade dos dados, em que o tempo de permanência está diretamente relacionado com o tipo de processo e existência de grande variação entre os tempos de cada tipo, o histograma foi construído a partir do tempo de permanência normalizado por tipo de processo.

O especialista do domínio definiu o terceiro quartil como referência de tempo de permanência satisfatório dos processos no TCE-PE; sendo assim, 75% do total de processos apresenta tempo de permanência satisfatório.

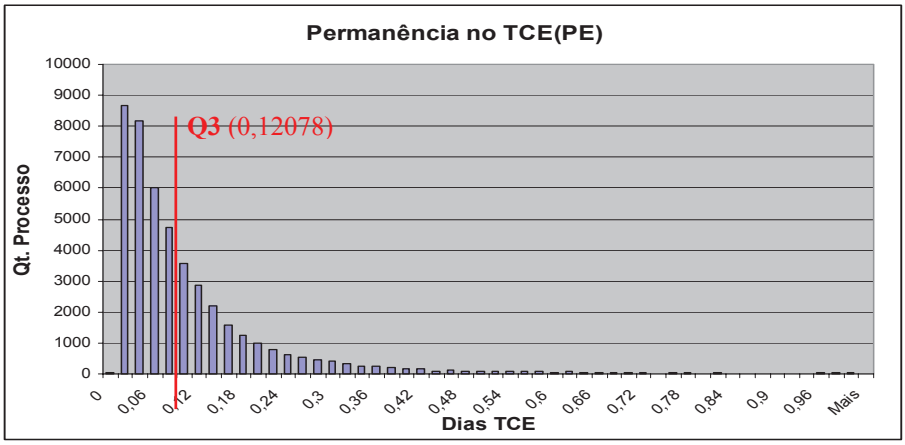

Figura 2 - Tempo de Permanência no TCE-PE.

3.1.1 Análise do Volume de Dados

A base de dados original apresentava um estoque de 98.624 registros (processos julgados) distribuídos em 38 diferentes categorias (tipos de processos).

Após a análise dos dados, foram selecionados 45.545 registros distribuídos em 26 categorias. Na Tabela 1, é apresentada a análise do volume de dados, estratificada por tipo de processo onde, partindo do total de processos extraídos da base (original), são apresentados os registros válidos (válidos), que representam o resultado da exclusão dos registros com campos não preenchidos. Apresenta ainda o volume de dados usado no projeto (mantido) e o resultado da análise que, rotulou as categorias em:

Mantido, para as categorias que tiveram seus registros mantidos na base;

Agrupado, para as categorias que foram agrupadas em categoria nova ou preexistente;

Excluído, para as categorias excluídas da base e

Acrescido, para nova categoria atribuída.

Por fim, é apresentada a justificativa para o resultado da análise, ressaltando-se que:

- De um total de 38 (trinta e oito) categorias observadas nos dados originais, apenas 26 (vinte e seis) foram usadas no projeto, uma vez que 06 (seis) foram excluídas, 06 (seis) agrupadas em uma nova denominada OUTROS e 01 (uma) agrupada em uma categoria preexistente;
- Os tipos de processo que apresentavam quantidades inferiores a 40 registros foram agrupados na nova categoria denominada OUTROS;
- Os processos do tipo SINDICÂNCIA E INQUÉRITO foram excluídos da base por terem natureza estranha ao problema;
- O tipo de processo APOSENTADORIA OU REFORMA, apesar de apresentar a maior quantidade de registros de toda a base, foi excluído porque, em 2005, foi extinto e substituído por dois tipos, APOSENTADORIA e RESERVA E REFORMA, que apresentam boa representatividade na base;
- O tipo de processo RECURSO, apesar de extinto, foi mantido na base porque engloba tipos de processo atuais de semelhante natureza jurídica, que não estão representados na base;
- O tipo PROCESSO LICITATÓRIO (extinto) foi agrupado no tipo AUDITORIA ESPECIAL porque os processos dessa natureza jurídica são formalizados atualmente nesse tipo.

ANÁLISE DO VOLUME DOS DADOS						
Cód	**Tipo do Processo**	**Original**	**Válidos**	**Mantido**	**Resultado da análise**	**Justificativa**
1	CÂMARA	2638	2524	2524	Mantido	-
2	PREFEITURA	2372	2290	2290	Mantido	-
3	AUTARQUIA	339	329	329	Mantido	-
4	ECONOMIA MISTA	143	141	141	Mantido	-
5	FUNDAÇÃO	205	198	198	Mantido	-
6	EMPRESA PÚBLICA	169	165	165	Mantido	-
7	GOVERNO DO ESTADO	17	15	0	Agrupado em OUTROS	Ocorrência de registros menor que 40.
8	UNIDADE GESTORA ESTADUAL	451	433	433	Mantido	-
9	FUNDO	741	735	735	Mantido	-
10	SECRETARIA DO PODER EXECUTIVO	12	11	0	Agrupado em OUTROS	Ocorrência de registros menor que 40.
11	MINISTÉRIO PÚBLICO	1	1	0	Agrupado em OUTROS	Ocorrência de registros menor que 40.
12	TRIBUNAL DE CONTAS	0	0	0	Excluído	Ausência de registros
13	TRIBUNAL DE JUSTIÇA	0	0	0	Excluído	Ausência de registros
14	ASSEMBLÉIA LEGISLATIVA	0	0	0	Excluído	Ausência de registros
15	REPASSE A TERCEIROS	15137	14454	14454	Mantido	-
16	RELATÓRIO DE GESTÃO FISCAL	295	279	279	Mantido	-
17	PROCESSO PRINCIPAL	612	574	574	Mantido	-
18	AUDITORIA ESPECIAL	948	914	1335	Mantido	Agrupou o tipo PROCESSO LICITATÓRIO (extinto).
19	DESTAQUE	83	77	77	Mantido	-
20	CONTRATAÇÃO TEMPORÁRIA	779	693	693	Mantido	-
21	CONCURSO	4099	3850	3850	Mantido	-
22	PROVIMENTO DERIVADO	15	13	0	Agrupado em OUTROS	Ocorrência de registros menor que 40.
23	PENSÃO	2686	2596	2596	Mantido	-
24	NOVAÇÃO DE PORTARIA	2061	1984	1984	Mantido	-
25	APOSENTADORIA OU REFORMA (extinto)	50391	49257	0	Excluído	Tipo de processo extinto e desmembrado em dois novos distintos já representados na base: APOSENTADORIA e RESERVA E REFORMA.
26	RESERVA E REFORMA	277	256	256	Mantido	-
27	APOSENTADORIA	4083	3843	3843	Mantido	-
28	RECURSO (extinto)	4677	4603	4603	Mantido	Apesar de atualmente extinto, este tipo de processo agrupa tipos atuais não representados na base
29	AGRAVO	5	5	0	Agrupado em OUTROS	Ocorrência de registro menor que 40.
30	EMBARGOS DE DECLARAÇÃO	10	8	0	Agrupado em OUTROS	Ocorrência de registros menor que 40.
31	RECURSO ORDINÁRIO	231	210	210	Mantido	-
32	PEDIDO DE RESCISÃO	178	171	171	Mantido	-
33	DENÚNCIA	1336	1261	1261	Mantido	-
34	CONSULTA	3074	2454	2454	Mantido	-
35	AUTO DE INFRAÇÃO	42	37	37	Mantido	Apesar de a representatividade ser inferior a 40 registros, este tipo de processo é extremamente relevante para o estudo.

36	SINDICÂNCIA	3	3	0	Excluído	Processo de natureza estranha ao problema.
37	PROCESSO LICITATÓRIO (extinto)	453	421	0	Agrupado em AUDITORIA ESPECIAL	Os atos desta natureza atualmente são auditados sob o tipo AUDITORIA ESPECIAL.
38	INQUÉRITO	1	1	0	Excluído	Processo de natureza estranha ao problema.
39	OUTROS	60	53	53	Acrescido	Agrupa os tipos GOVERNO DO ESTADO, SECRETARIA DO PODER EXECUTIVO, MINISTÉRIO PÚBLICO, PROVIMENTO DERIVADO, EMBARGOS DE DECLARAÇÃO e AGRAVO.
TOTAL		98624	94859	45545		

Tabela 1 - Análise do Volume de Dados

3.2 Visão Original dos Dados

A Tabela 2 apresenta um resumo da visão original dos dados. Uma análise detalhada dessa visão é apresentada abaixo.

VISÃO ORIGINAL DOS DADOS					
ID	**LISTA DE VARIÁVEIS**	**DESCRIÇÃO**	**TIPO DE VARIÁVEL**	**NÍVEL (%) DE PREENCHIMENTO**	**NÚMERO DE DISTINTOS**
1	LocalFormalizacao	Os segmentos administrativos do TCE que são responsáveis pela formalização.	Categórica	100%	11
2	DiasTCE	Quantidade de dias que o processo passa no TCE.	Numérica	100%	-
3	DiasFormalizacao	Quantidade de dias que o processo passa na fase de formalização.	Numérica	100%	-
4	DiasInstrucaoJulgamento	Quantidade de dias que o processo passa na fase de instrução e julgamento.	Numérica	100%	-
5	DiasPublicacao	Quantidade de dias que o processo passa na fase de publicação.	Numérica	100%	-
6	TipoDeliberacao	Legalmente existem 3 categorias possíveis: Acórdão, Decisão e Parecer. Foi acrescentada a categoria NEDI (Informação não disponível) na migração do sistema anterior para o atual.	Categórica	100%	4
7	Situacao da Deliberação	Apresenta a situação da deliberação do processo	Categórica	100%	31
8	Exercício	Indica o ano. Varia entre 1969 e 2006.	Numérica	82%	-
9	Relator	Responsável pela proposta de julgamento	Categórica	100%	20
10	Modalidade	Classificação dos Processos quanto à natureza jurídica.	Categórica	100%	13
11	Tipo	Classificação pormenorizada da Modalidade. Rotula os tipos de processo.	Categórica	100%	38
12	Natureza	Classificação quanto à natureza da pessoa jurídica	Categórica	100%	17
13	OrgaoJulgador	Câmaras e Pleno	Categórica	100%	4
14	UnidadeGestora	Denominação da Unidade Gestora.	Categórica	100%	825
15	Esfera	Indica se o processo pertence à esfera estadual ou municipal	Categórica	100%	2

Tabela 2. Variáveis selecionadas para o desenvolvimento da solução

Os campos TipoDeliberacao e SituacaoDeliberação apresentaram campos preenchidos com o texto "Não informado".

O administrador do banco de dados informou que, no momento da migração dos dados de um sistema anterior para o atual, todos os campos inexistentes no sistema antigo foram preenchidos no sistema atual com o texto "Não informado".

O campo Exercício, que significa exercício financeiro não estava preenchido em 17.050 registros, ou seja, 37,5% do total porque, no sistema anterior, o preenchimento deste campo não era obrigatório. No entanto, este dado é conhecido a partir do número do processo, em que os dois primeiros dígitos representam o exercício financeiro. No pré-processamento, estes campos foram preenchidos a partir do número do processo possibilitando a recuperação total dessa informação.

Verificou-se que algumas variáveis tiveram seus domínios modificados ao longo dos 37 anos do estoque de processo analisado. No entanto, em reunião com o administrador do banco de dados, obteve-se a equivalência entre as categorias antigas e suas correspondentes atuais, e os registros que apresentavam categorias em desuso foram atualizados.

Os campos numéricos DiasTCE, DiasFormalizacao, DiasInstrucaoJulgamento e DiasPublicacao apresentaram valores negativos em uma pequena quantidade de registros. Tais registros foram excluídos.

4. Preparação dos Dados

A preparação dos dados corresponde à terceira fase do CRISP-DM, *Data Preparation*. Nela são realizadas todas as atividades de construção da base de dados para apresentação à ferramenta de modelagem.

Usualmente, até a conclusão desta etapa, de um trabalho de descoberta de conhecimento a partir de dados (*Knowledge Discovery in Databases* – KDD) [Fayyad 1996] é consumido cerca de 80% do tempo gasto no projeto. Em nosso caso, este período consumiu cerca de 90% do tempo dedicado ao projeto. Por tratar-se de um problema do mundo real, todas as decisões de preparação dos dados foram validadas com os respectivos geradores da informação em reuniões formais junto à instituição cliente, perfazendo um total de seis reuniões realizadas.

Nas subseções que se seguem, são detalhadas as tarefas realizadas.

4.1 Seleção de atributos

Para a modelagem do problema, foram mantidos somente os atributos que são conhecidos até o fim da fase de Formalização, pois representam os dados *a priori* da fase seguinte, Instrução, que é a fase alvo do sistema de apoio à decisão apresentado. Para tanto, foram excluídos os atributos: DiasTCE, DiasInstrução, DiasJulgamento e DiasPublicação. Além desses, o atributo UnidadeGestora foi excluído por tratar-se de um atributo categórico com 825 categorias distintas que apresentavam empecilhos de agrupamento por representarem pessoas jurídicas diversas.

4.2 Redução do Número de Categorias

Alguns campos categóricos precisaram ter uma redução na quantidade de categorias para facilitar a conversão para atributos binários ou para refletir a atual organização do TCE. Abaixo, apresentamos os critérios utilizados para o agrupamento das categorias:

- O atributo LocalFormalização teve a substituição da ocorrência DICO pelo valor DIPR para refletir uma mudança de nomenclatura de um segmento administrativo do TCE;
- O atributo Exercício apresentava 37 categorias diferentes. Foi adotado como critério, para reduzir a quantidade de categorias, selecionar os 19 exercícios mais recentes e agrupar os demais numa categoria "outros", resultando em 20 categorias;
- O campo Tipo teve os valores "Secretaria do Poder Executivo", "Governo do Estado", "Ministério Público", "Tribunal de Contas", "Tribunal de Justiça", "Assembléia Legislativa", "Provimento Derivado", "Embargos de Declaração" e "Agravo" agrupados na categoria “outros”.

4.3 Campos Calculados

Foi adicionada à base um atributo chamado OperaçãoEleição. A operação eleição é uma atividade que acontece em anos eleitorais e possui duração de 90 dias. Inicia-se 45 dias antes do dia de votação e prolonga-se por 45 dias após o dia da votação. Durante o período de OperaçãoEleição, os processos da esfera municipal que estão em fase de instrução têm seus trabalhos suspensos, o que acaba impactando o tempo de permanência destes no TCE. O atributo binário OperaçãoEleição indica se existe possibilidade de o processo ter seu tempo de permanência no TCE prolongado em virtude da OperaçãoEleição. Esse dado é obtido a partir da data de finalização da fase de Formalização do processo juntamente com o calendário dos períodos eleitorais.

A classe atribuída a cada processo foi calculada da seguinte maneira: se a quantidade de dias de permanência no TCE estiver acima do terceiro quartil dos processos de seu tipo, o processo será considerado com tempo de permanência Ruim; caso contrário, a permanência será considerada Boa. O ponto de corte de terceiro quartil foi estabelecido juntamente com o especialista do domínio. Assim, 75% das instâncias foram classificadas como “Boa” e 25% como “Ruim”.

4.4 Normalização dos Dados

O campo DiasFormalização, único campo numérico entre os atributos selecionados, foi normalizado pela média e desvio-padrão, levando-se em consideração os diferentes tipos de processo. Ou seja, se um registro pertence ao tipo Aposentadoria, seu atributo DiasFormalização é normalizado pela média e desvio-padrão do rol de registros desse tipo de processo.

4.5 Conversão dos Atributos Categóricos em Numéricos

Os campos categóricos foram convertidos em números binários, sendo cada categoria convertida em um atributo que pode assumir valor 0, se o registro não possui a categoria correspondente, ou 1, caso contrário. A tabela 3 apresenta um exemplo da conversão para o atributo TipoDeliberacao que possui 04 (quatro) categorias distintas: Decisão, Acórdão, Parecer e NEDI.

Categoria	Número Binário			
Parecer	1	0	0	0
Decisão	0	1	0	0
Acórdão	0	0	1	0
NEDI	0	0	0	1

Tabela 3 - Exemplo de conversão de atributo categórico em número binário.

4.6 Seleção das Instâncias

A base de dados foi dividida em três conjuntos: treinamento, validação e teste. A divisão foi feita por amostragem aleatória estratificada por classe de processo da seguinte maneira: 50% das instâncias (processos) foram colocadas no conjunto de treinamento, 25% no conjunto de validação e os 25% restantes no conjunto de teste. Todos os conjuntos também foram estratificados pelo tipo de processo. Com o intuito de balancear o conjunto de treinamento, replicamos as instâncias da classe "Ruim", fazendo também a replicação estratificada por tipo de processo.

5. Modelagem

Na fase de modelagem, *Modeling* do CRISP-DM, selecionam-se as técnicas de modelagem, aplicam-se as técnicas escolhidas e ajustam-se seus parâmetros para os valores ótimos. Várias técnicas de aprendizado de máquina são apropriadas ao problema em estudo [Rezende 2003]: redes neurais artificiais, árvores de decisão, máquinas de vetores suporte, regras de classificação, etc. Nesse projeto decidimos utilizar redes neurais do tipo MLP (*Multilayer Perceptron*) [Haykin 1999]. A rede neural treinada foi gerada no MatLab[2].

5.1 Redes neurais MLP

O primeiro passo da fase de modelagem é a escolha da técnica que será utilizada na solução do problema. Nós escolhemos as redes neurais *multilayer perceptron* (MLP) para a modelagem de nossa solução. Nossa escolha pode ser justificada pelos bons resultados obtidos por essas redes em diversos problemas do mundo real [PAKDD 2007].

Uma Rede Neural Artificial (RN) é formada por certo número de unidades de processamento interconectadas. Tais unidades são inspiradas nos neurônios biológicos do cérebro. Os neurônios de uma RN são conectados por ligações que permitem a transmissão de sinais de um neurônio para outro. Cada neurônio recebe certo número de sinais de entrada e produz um único sinal de saída. O sinal de saída é então transmitido através das conexões de saída do neurônio. Essas redes são, geralmente, formadas por uma ou mais camadas de neurônios interligadas por um grande número de conexões. A cada conexão é associado um peso, que representa o conhecimento representado e serve para ponderar as entradas recebidas na rede. A Figura 3 apresenta a estrutura de uma rede neural multicamadas com uma única camada intermediária (camada escondida) e um neurônio na camada de saída.

[2] MATLAB (MATrix LABoratory) é um pacote de *software* proprietário, de alta performance, destinado à computação científica e engenharia.

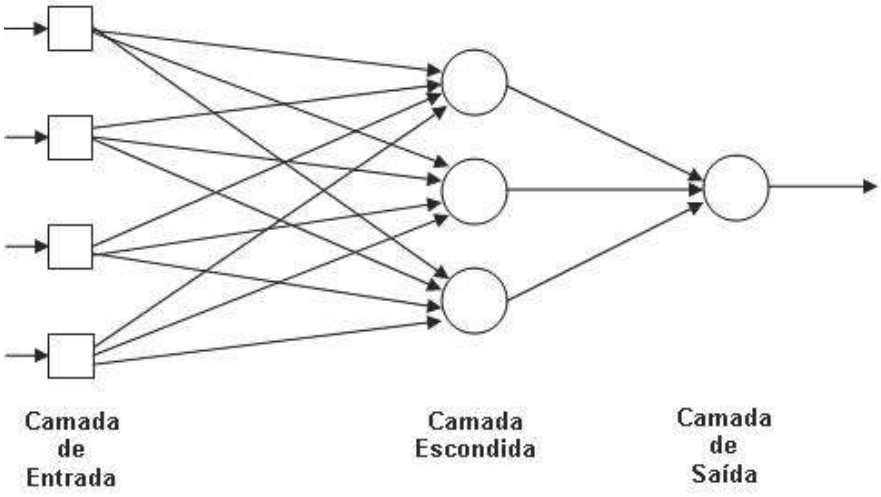

Figura 3 - Rede Neural Perceptron Multi-Camadas

O primeiro passo na construção de uma rede neural artificial é a definição de sua arquitetura. Para tanto, são definidos o número de neurônios utilizados e a maneira como estes serão conectados para formar a rede. Definida a estrutura, escolhe-se o algoritmo de aprendizado que será utilizado. Finalmente, treina-se a rede neural; isto é, iniciam-se os pesos da rede e os atualizamos iterativamente, de acordo com o conjunto de instâncias de treinamento [Haykin 1999].

Além do número de neurônios das camadas de entrada da rede e da camada escondida, é preciso também definir alguns parâmetros relativos à execução do treinamento. Nesse trabalho utilizamos os seguintes parâmetros:

- Taxa de Aprendizado: 0,001
- Momento: 0,01
- Número de neurônios na camada escondida: 05
- Número máximo de iterações: 100.000

Após a definição dos parâmetros da rede, passamos para a execução de seu treinamento no MatLab, utilizando o algoritmo *backpropagation* [Haykin 1999].

A rede neural treinada apresenta como saída um *score* para cada um dos processos. Esse *score* é um valor entre 0 e 1, que pode ser entendido como a probabilidade de o processo ser da classe "boa" ou da "ruim". A determinação do limiar que separa as duas classes é fundamental para a determinação do desempenho do sistema. Tal limiar será definido na seção seguinte.

O gráfico abaixo (Figura 4) apresenta a importância de cada variável para o sistema de decisão. A importância de cada variável foi calculada somando-se todos os pesos das conexões entre a camada de entrada (considerando apenas os neurônios de entrada ligados à variável computada) e a única camada intermediária da rede MLP. Todos os valores de importância apresentados na figura 4 foram normalizados para valores entre 0 e 1.

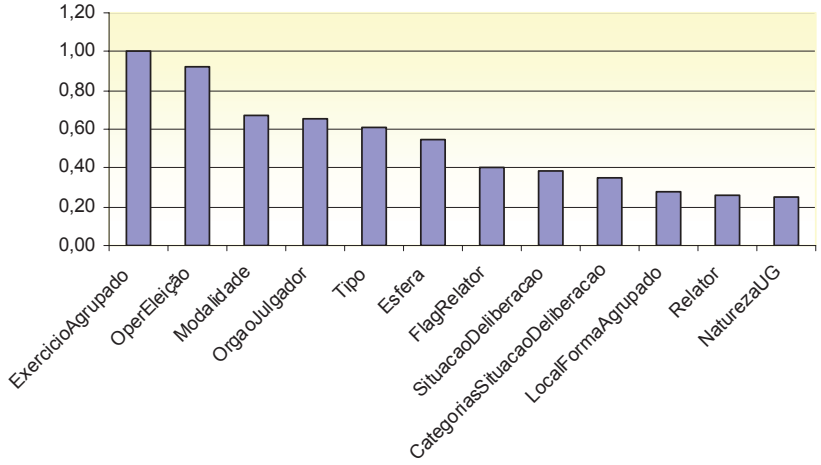

Figura 4. Importância das variáveis na determinação da classe

5.2. Indução de Regras

A indução de regras é uma das técnicas que podem ser empregadas para identificar relações ou padrões que permitem uma melhor compreensão sobre as dependências existentes entre as variáveis da massa de dados e o nosso alvo: os processos com permanência boa ou ruim.

Neste trabalho, as regras foram induzidas utilizando-se o algoritmo *A Priori* [Han, J., & Kamber 2006]. Cada regra possui uma condição ou premissa, que determina o universo de casos da massa de dados para os quais a regra é aplicável. Adicionalmente, cada regra possui três medidas: o suporte, o percentual de processos ruins existentes no universo de casos associados à regra e o *lift*. O suporte representa a porcentagem de casos da base de dados para os quais a condição da regra se aplica. E o *lift* mede a relação entre o número de processos ruins associados à regra e a média de processos ruins existentes na base de dados. Mais precisamente, neste trabalho o *lift* foi calculado utilizando-se a seguinte formulação: $(\Phi/\mu - 1) \times 100$. Onde Φ é o percentual de processos ruins associados à regra e μ é a média de processos ruins, considerando-se todos os casos (ou registros) da base de dados.

A Tabela 4 apresenta as 8 regras que melhor caracterizam a classe dos processos com tempo de permanência ruim. As regras foram selecionadas a partir do conhecimento do especialista e a partir das medidas de suporte, percentual de processos ruins e *lift*.

A Regra de número 1 confirma o conhecimento do especialista: processos instruídos durante a deflagração de uma operação eleição, na sua maioria (cerca de 71,29% dos casos), tiveram tempo ruim de permanência.

A Regra de número 2, cujo suporte é de 49,57%, revela que caso um processo da esfera municipal seja formalizado na sede (DIPR), a permanência foi ruim em cerca de 30,80% dos casos. Isso se explica porque após a formalização, decorrerá algum tempo até que o processo chegue à inspetoria da área municipal onde será instruído e ainda, a formalização fora do local apropriado resulta, possivelmente, em documentação precária.

Todos os processos são classificados em duas esferas: estado ou município. A regra de número 3, com suporte de 63,15% e *lift* de 114,75%, revela que os processos da esfera municipal apresentam permanência ruim. Esta regra é reforçada pelas regras seguintes, de número 4 e 5, que demonstram os processos de prefeitura com permanência ruim com suporte de 48,39% e, mais especificamente, os processos do tipo pensão de prefeitura que

apresentam permanência ruim com *lift* de 204,30%. Para o especialista, este conhecimento é uma novidade, considerando a priorização de recursos técnicos, financeiros e de pessoal para a área municipal. Um fator que poderá explicar, em parte, é a distância física entre os municípios auditados.

A regra 6 demonstra que, caso um processo sofra deliberação desfavorável, seu tempo de permanência é ruim em 31,17% das vezes. Isso se justifica pela prudência do julgador ao proferir uma decisão que resultará em punição ao interessado no processo.

A regra 7 reforça o conhecimento explicitado pelas de número 3, 4 e 5. Nela se observa que, apesar da categoria da situação da deliberação ser favorável, caso o processo seja da área municipal, apresenta permanência ruim.

A regra de número 8 revela conhecimento novo para o especialista no domínio. Isso se justifica pelos procedimentos atuais adotados pelo TCE-PE, quando já na fase de julgamento, o processo é distribuído para a relatoria de um auditor na ausência funcional do conselheiro relator, tendo o tempo de permanência, sob a responsabilidade desse conselheiro relator, computado como fase de julgamento.

Regra	Condições ou Premissas	Suporte (%)	Ruim (%)	*Lift* (%)
1	Se OperEleição = SIM	4,59	71,29	279,59
2	Se LocalFormaAgrupado = DIPR e Esfera = M (Município)	49,57	30,80	121,69
3	Se Esfera = M	63,15	29,02	114,75
4	Se NaturezaUG = P (Prefeitura)	48,39	30,44	120,14
5	Se Tipo = Pensão e NaturezaUG = P (Prefeitura)	2,61	53,27	204,30
6	Se CategoriasSituacaoDeliberacao = Desfavorável	17,13	31,17	121,73
7	Se CategoriaSituacaoDeliberação = Favorável e Esfera = M (Município)	30,72	30,11	118,32
8	Se FlagRelator = Auditor	33,71	35,62	140,47

Tabela 4. Regras Induzidas pelo algoritmo A Priori.

6. Avaliação dos Resultados

Toda a avaliação de desempenho foi realizada sobre um conjunto de teste, estatisticamente independente dos dados de modelagem. A subseção seguinte apresenta os tipos de erro de classificação associados ao modelo. Em seguida, o modelo será avaliado sob quatro diferentes medidas de desempenho: as curvas ROC (*Receiver Operating Characteristic Curve)*, o coeficiente de GINI, o KS2 (Teste Kolmogorov-Smirnov) e a avaliação de custo.

6.1 Erros de Classificação

A avaliação do modelo foi realizada em termos dos erros de classificação tipo I e tipo II, ao invés de considerar somente a taxa de erro geral. O Erro tipo I é o erro de classificar os processos com bom tempo de instrução (classe boa), como processos de instrução ruim – Falsos Negativos. Enquanto que o Erro tipo II é o erro de classificar os processos com tempo de instrução ruim como se fossem bons – Falsos Positivos.

Assim temos:

- Erro tipo I = (RUIM dado que é BOA)/BOA
- Erro tipo II = (BOA dado que é RUIM)/RUIM
- Erro geral = ((RUIM dado que é BOA) + (BOA dado que é RUIM)) / (RUIM+BOA)

6.2 Curvas ROC

A curva Roc (*Receive Operator Characteristic Curve*) [Fawcett 2003] mostra a relação dos verdadeiros positivos (sensibilidade) com os falsos positivos (especificidade), na qual o percentual varia de acordo com o limiar ou ponto de corte aplicado [Spackman 1989]. Esta análise é feita por meio de um método gráfico simples e robusto, que permite estudar a variação da sensibilidade e especificidade, para diferentes valores de ponto de corte.

A escolha do limiar ou ponto de corte de um sistema de apoio à decisão recai sobre a decisão entre aumentar a sensibilidade à custa de redução da especificidade ou vice-versa. Deve-se avaliar cuidadosamente a importância relativa da sensibilidade e especificidade do teste para estabelecer o ponto de corte mais adequado. A estratégia, em geral, é a seguinte:

a) Se a principal preocupação é evitar resultado falso-positivo, então o ponto de corte deve objetivar o máximo de especificidade.

b) Se a preocupação maior é evitar resultado falso-negativo, então o ponto de corte deve objetivar o máximo de sensibilidade.

- A sensibilidade de um teste é dada por:

$$sensibilidade = \frac{Vp}{Vp + Fn}$$

- A especificidade de um teste é dada por:

$$especificidade = \frac{Vn}{Vn + Fp}$$

A área abaixo da curva ROC está associada ao poder discriminante de um classificador e pode ser determinada através de métodos de resolução numérica. A curva ROC é um gráfico de pares x e y que correspondem, à especificidade e à sensibilidade, respectivamente.

Uma das vantagens deste método é que as curvas de diferentes modelos podem ser comparadas; quanto melhor o classificador, mais perto estará sua curva do canto superior esquerdo do gráfico.

Essa relação prevê o desempenho do modelo independentemente da distribuição da classe e custo associado com o tipo de erro.

Como neste estudo a principal preocupação é evitar resultado falso-positivo, então o ponto de corte deve objetivar o máximo de especificidade. A partir da análise do gráfico apresentado na Figura 5, definimos o ponto de corte para nosso sistema em 0,2, obtendo assim uma sensibilidade de 78%.

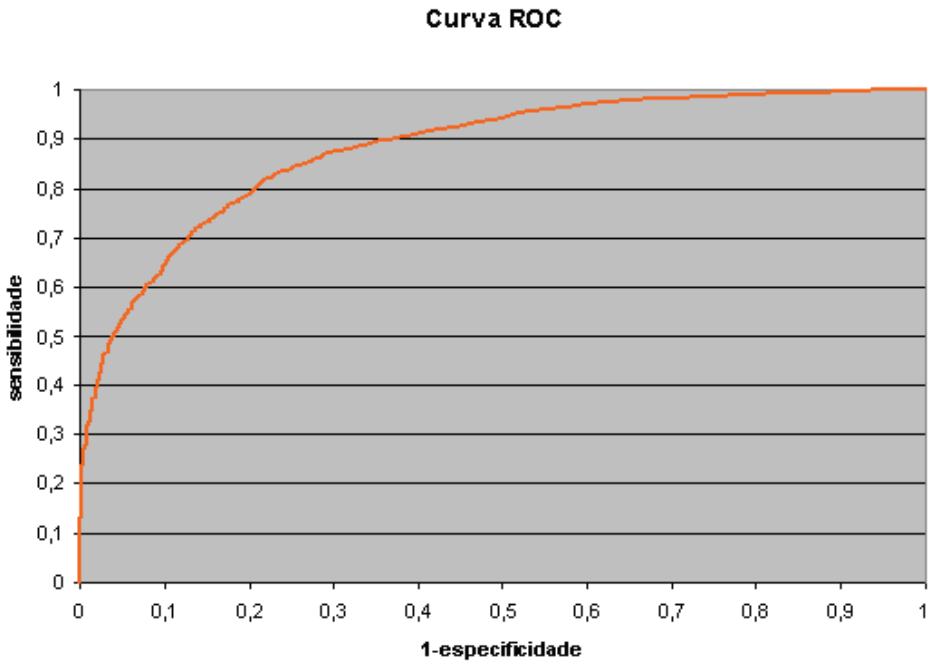

Figura 5 - Curva ROC

Esse ponto de corte em 0,2 resulta em:

- Taxa de Erro I: 29,3%
- Taxa de Erro II: 15,5%
- Taxa de Erro global: 26%

6.3 Teste Kolmogorov-Smirnov (KS-2)

O KS-2 é um teste não-paramétrico utilizado para medir a aderência de dados a uma distribuição. Em sistemas decisórios em geral, ele serve para medir a separabilidade entre duas distribuições a partir da função de distribuição acumulada de cada uma [Adeodato et. al 2005]. O teste é baseado na maior diferença absoluta entre a freqüência acumulada das duas classes.

Submetendo os processos à rede neural treinada, pudemos medir o KS-2 entre as distribuições dos processos que apresentaram permanência no TCE-PE "boa" daqueles com permanência "ruim". Neste caso, quanto maior o KS máximo, mais distintos são os perfis das duas classes.

Vemos, na Figura 6, que a pontuação dos processos com permanência "boa" é bastante superior à dos processos com permanência "ruim", o que mostra que o sistema apresentado é capaz de discriminar as duas classes apresentadas. O valor do KS máximo obtido pelo modelo desenvolvido foi 0,6.

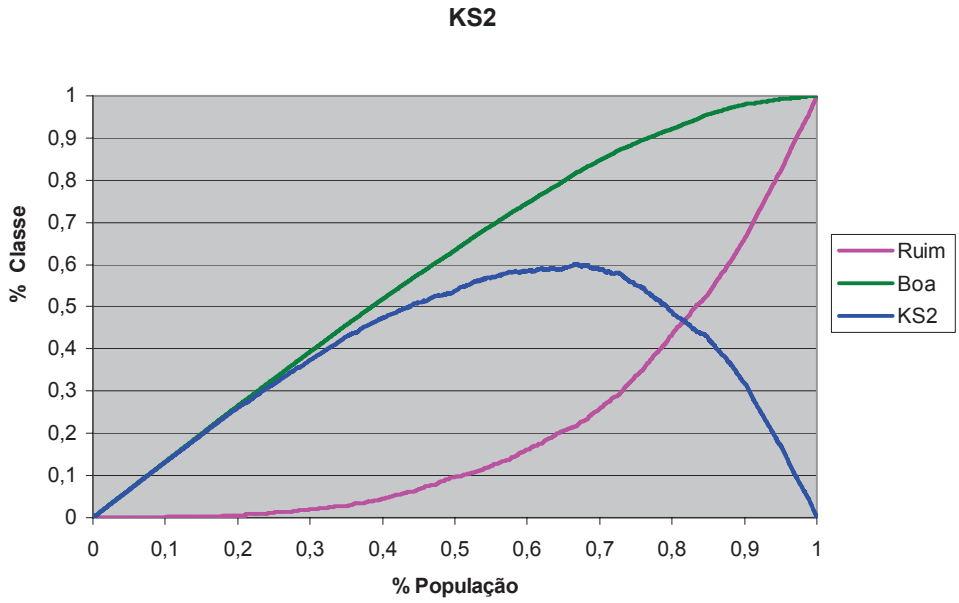

Figura 6 - Gráfico do KS-2

6.4 Coeficiente de GINI

O coeficiente de GINI é uma medida de desigualdade utilizada normalmente para calcular a desigualdade de distribuição de renda, mas pode ser usado para qualquer distribuição [Hoffman 1998]. Numericamente, varia de zero a um, onde o valor zero representa a situação de igualdade, ou seja, todos têm a mesma renda e o valor um está no extremo oposto, isto é, uma única pessoa detém toda a riqueza.

O índice de GINI pode ser utilizado na avaliação de um classificador, medindo o grau da concentração dos seus acertos. O valor zero indica uma perfeita igualdade da distribuição, enquanto o valor unitário indica a concentração máxima.

Se o modelo não for capaz de distinguir as duas classes, o gráfico é representado por uma reta de 45 graus. Assim, quanto mais distante a curva dessa reta, melhor a qualidade do classificador.

O gráfico apresentado na Figura 7 foi construído a partir das duas distribuições. Podemos observar que as duas curvas estão afastadas de uma hipotética reta de 45 graus.

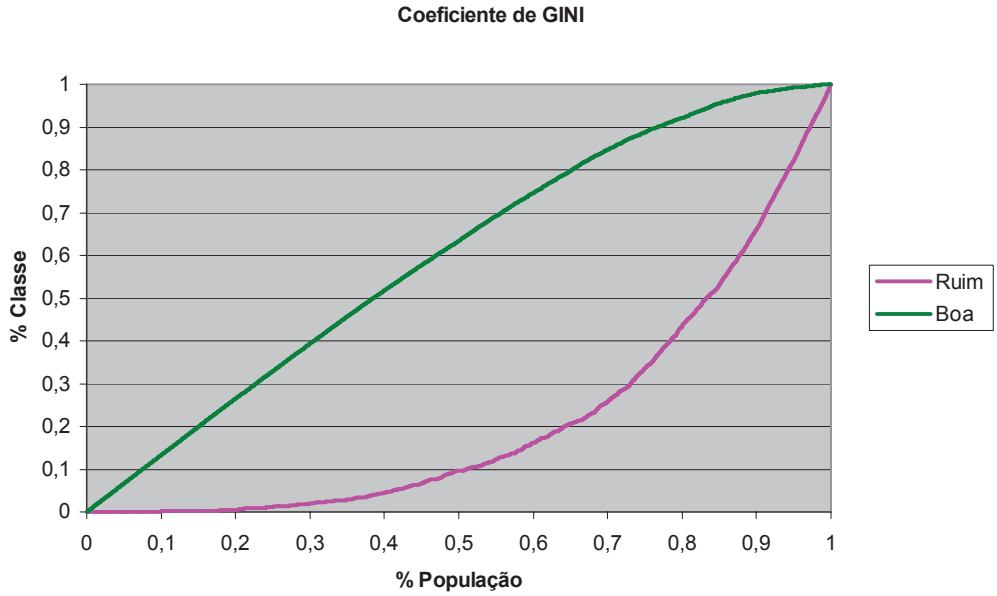

Figura 7 - Coeficiente de GINI

6.5 Avaliação do Custo

Para o problema em estudo, observa-se que os custos associados aos erros tipo I e II são significativamente diferentes. As conseqüências de classificar um processo com permanência ruim como se fosse boa (erro tipo II) são potencialmente maiores que classificar um processo com permanência boa como se fosse ruim (erro tipo I), uma vez que o erro tipo II agrava, consideravelmente, o tempo de produção do processo, pois esse deixará de ser tratado como ruim, passando a ocupar posição na fila de processo com celeridade boa, ou seja, atrás de todos os demais classificados como ruim.

Foi atribuído peso 01 (um) para o custo dos erros tipo I e peso 09 (nove) para o custo do erro tipo II, considerando as seguintes premissas:

1. A celeridade processual é um indicador de excelência do TCE-PE;
2. As regras de distribuição na fase de Instrução são uniformes para todos os tipos de processo;

Dessa forma, o erro ponderado pelos custos associados é dado por:

$$Ep = \frac{(custoI * erroI) + (custoII * erroII)}{(custoI + custoII)}$$

As Tabelas 5-7 mostram as matrizes de confusão e erros ponderados obtidos para os pontos de decisão com os limiares: 0,3 - 0,2 e 0,1.

A partir dos erros ponderados, observa-se que o melhor desempenho é obtido pelo modelo de limiar 20%, em que o erro ponderado atinge seu valor mínimo em 0,64175 e sobe quando o limiar assume valor 0,1 ou 0,3.

MATRIZ DE CONFUSÃO – Ponto de corte 0,3						
CLASSE VERDADEIRA		Classificado como		TOTAL	Erro	
		BOM	RUIM			
	BOM	6656	1970	8626	Erro I	0,228
	RUIM	520	2236	2756	Erro II	0,189
	TOTAL	7176	4206	11382	Erro geral	0,219
Erro Ponderado						0,65325

Tabela 5- Matriz de Confusão com ponto de corte 0,3

MATRIZ DE CONFUSÃO – Ponto de Corte 0,2						
CLASSE VERDADEIRA		Classificado como		TOTAL	Erro	
		BOM	RUIM			
	BOM	6097	2529	8626	Erro I	0,293
	RUIM	428	2328	2756	Erro II	0,155
	TOTAL	6525	4857	11382	Erro geral	0,26
					Erro Ponderado	**0,64175**

Tabela 6- Matriz de Confusão com ponto de corte 0,2

MATRIZ DE CONFUSÃO – Ponto de corte 0,1						
CLASSE VERDADEIRA		Classificado como		TOTAL	Erro	
		BOM	RUIM			
	BOM	5275	3351	8626	Erro I	0,388
	RUIM	338	2418	2756	Erro II	0,123
	TOTAL	5613	5769	11382	Erro geral	0,324
					Erro Ponderado	0,66475

Tabela 7- Matriz de Confusão com ponto de corte 0,1

Vê-se que o erro ponderado atinge o mínimo no ponto de corte = 0,2 e sobe para os valores 0,1 e 0,3.

O gráfico a seguir (Figura 8) apresenta a curva do custo. Considerando a relação entre custos 9/1, o ponto de corte ideal do modelo é 0,18.

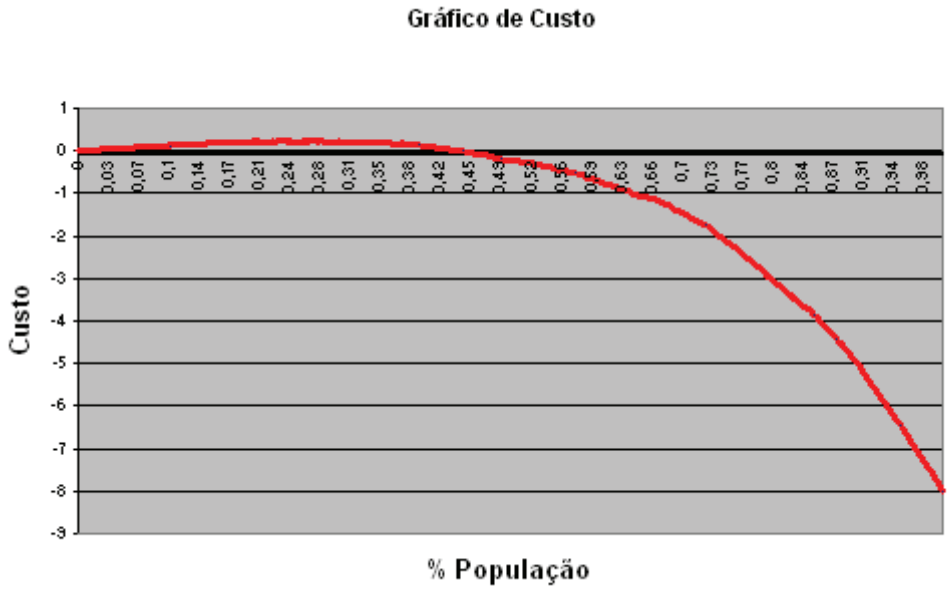

Figura 8 - Gráfico de Custo

7. Conclusões

Este trabalho propõe um sistema de apoio à decisão para o Tribunal de Contas do Estado de Pernambuco. Tal sistema foi desenvolvido a partir de um processo de mineração de dados na base de dados do TCE-PE. Os dados utilizados nesse processo de descoberta de conhecimento se referem ao estoque de processos julgados pelo referido tribunal até 30 de novembro de 2006.

O sistema de apoio à decisão apresentado atua ao final da fase de formalização, apontando aos gestores a possibilidade de um processo ter tempo de permanência "ruim" e, a partir desse momento, poderem ser tomadas várias medidas preventivas no intuito de evitar a confirmação dessa predição. O grande impacto do modelo desenvolvido é a redução do tempo gasto em processos formalizados no TCE-PE. Tal melhoria permite ao cidadão pernambucano uma resposta célere sobre o julgamento dos atos exercidos pelos gestores públicos.

A solução desenvolvida é baseada em uma rede neural artificial do tipo MLP. O limiar de decisão foi escolhido de maneira a minimizar o número de falsos positivos (processos que terão tempo de permanência "ruim" e são apontados pela rede como permanência "boa"). Dessa forma, a rede neural comete erro em apenas 15% dos processos com tempo de permanência "ruim".

Este projeto focou a atuação na fase de Instrução. Em trabalhos futuros, é previsto estender o objeto de estudo para apoio à tomada de decisão nas fases de Julgamento e Publicação.

Referência Bibliográfica

[Adeodato et. al 2005] ADEODATO, P. J. L et. al. Sistema de Apoio à Decisão para Estimação do Sucesso do Aluno no Programa de Mestrado em Ciência da Computação da UFPE. CBRN, 2005.

[CF 1988] BRASIL. Constituição (1988).

[Chaudhuri & Dayal 2003] CHAUDHURI, S.; DAYAL, U.. An overview of data warehousing and OLAP technology. SIGMOD Rec. 26 (1) , p. 65-74, 1997.

[Colin Shearer 2000] SHEARER, C. The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing, v. 5, n. 4. Washington,fall 2000.

[Duda 2000] DUDA, R. O.; HART, P. E.; STORK, D. G. Patter Classification : USA. 2nd. Wiley, 2000.

[Fayyad 1996] FAYYAD, U; PIATETSKY, G.; SMYTH, P..The KDD process for extracting useful knowledge from volumes of data, *Commun. ACM*, v.39, n. 11, 1996, p. 27-34.

[Fawcett 2003] FAWCETT, T. "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers".2003 (http://citeseer.comp.nus.edu.sg/fawcett03roc.html, retrieved July, 2007.)

[Han, J., & Kamber 2006], HAN, J., & KAMBER M. Data Mining: Concepts and techniques. Morgan Kaufmann, San Francisco, CA. 2006

[Haykin 1999] SIMON HAYKIN. Neural Networks: A comprehensive foundation, 2ª edição. Prentice Hall, 1999.

[Hoffman 1998] HOFFMAN, R.. Estatística para Economistas, 3. São Paulo: Editora Atlas, 1998.

[PAKDD 2007], The 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Nanjing, China, 22-25 May 2007.

[Prado 2004] PRADO, D. Introdução à teoria das filas e da simulação. Rio de Janeiro: INDG, 2004.

[Rezende 2003] REZENDE, S. O. (organizadora). Sistemas Inteligentes: Fundamentos e Aplicações. Editora Manole Ltda, 2003.

[Spackman 1989] SPACKMAN, K. A.. Signal detection theory: valuable tools for evaluating inductive learning. In Proceedings of the Sixth international Workshop on Machine Learning (Ithaca, New York, United States). A. M. Segre, Ed. Morgan Kaufmann Publishers, San Francisco, CA, p. 160-163, 1989.

[TVE Brasil] Disponível em: <http://www.tvebrasil.com.br/salto/boletins2001/cont/cont0.htm>. Acesso em: maio 2007.

[Witten & Frank 2005] WITTEN, I. H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Technique with Java Implementation. San Francisco, CA: Morgan Kaufman, 2005.