ANÁLISE DE COMPORTAMENTO DO SISTEMA RODOVIÁRIO DO NORDESTE DO BRASIL

UMA ABORDAGEM DE MINERAÇÃO DE DADOS

*ANALYSIS OF THE BEHAVIOR OF THE ROAD SYSTEM IN THE NORTHEAST OF BRAZIL*

*A DATA MINING APPROACH*

Othon L. T. de Oliveira

Univerdidade de Pernambuco

Recife, Brasil

olto@ecomp.poli.br

Fernando B. de L. Neto

Universidade de Pernambuco

Recife, Brasil

fbln@ecomp.poli.br

Resumo — As rodovias federais que atravessam a Região Metropolitana de algumas cidades estão constantemente congestionadas, não apenas pela quantidade de veículos, mas por serem alvo de paralisações das mais diversas matizes, como protestos de trabalhadores, acidentes, buracos, intempéries naturais e outros tipos de paralisações. Em situações extremas poderiam paralisar até a produção das fábricas no seu entorno. Propomos modelo de classificação para o comportamento das rodovias federais que atravessam o estado de Pernambuco na região Nordeste do Brasil, de modo que seja possível antecipar eventos que poderão causar constrangimentos, como retenção, redução de fluxo de tráfego. A fonte de dados dessa pesquisa provém da base de dados da Polícia Rodoviária Federal de Pernambuco (PRF) a partir de 2007 tendo considerado veículos, traçado da via e trechos da rodovia relacionados a acidentes, dentre outros. Com base nas informações obtidas, foi realizada uma Mineração de Dados utilizando a metodologia CRISP-DM para encontrar padrões comportamentais nas rodovias e em seu entorno. Foram empregados algoritmos de aprendizagem de máquina para classificação e regressão, sendo priorizadas, Árvores de Decisão e Redes Neurais. Os valores da área sob a curva ROC (AUC) obtidos foram acima de 0.7 que reflete um alto grau de confiabilidade. O modelo de predição proposto significa um avanço em termos de mobilidade e gestão do transporte de cargas, uma vez que possibilita antecipar eventos e comportamentos, favorecendo a escolha de rotas alternativas e ampliando o espaço temporal de escolha para determinadas rotas.

Palavras Chave – Modelo de predição; Mineração de dados; Tráfego em rodovias; Árvores de decisão, CRISP-DM.

Abstract — Federal highways that cross the Metropolitan Region of some cities are constantly congested, not only by the number of vehicles, but also because they are subject to stoppages of the most diverse shades, such as workers' protests, accidents, holes, natural weather and other types of stoppages. In extreme situations they could paralyze even the production of factories in their surroundings. We propose a classification model of behaviorial patterns for the federal highways that cross Pernambuco, a state in the northeastern region of Brazil, so that it is possible to anticipate events that may cause constraints, such as retention, reduction of traffic flow. The data source of this research comes from the database of the Federal Highway Police of Pernambuco (PRF) since 2007. We having considered vehicles, track layout and road sections related to accidents, among others. Based on the information obtained, a Data Mining was performed using the CRISP-DM methodology to find behavioral patterns on highways and in their surroundings. Machine learning algorithms were used for classification and regression, being prioritized, Decision Trees and Neural Networks. The values ​​of the area under the ROC (AUC) curve obtained were above 0.7 which reflects a high degree of reliability. The proposed prediction model means an advance in terms of mobility and cargo transport management, since it allows anticipating events and behaviors, favoring the choice of alternative routes and increasing the time space of choice for certain routes

Keywords – Prediction model; Data mining; Road traffic; Decision tree; CRISP-DM.

1. Introdução

O transporte de cargas que atravessa as regiões metropolitanas das grandes cidades brasileiras é realizado principalmente pelas rodovias federais. Essas rodovias frequentemente se encontram congestionadas em determinados dias/horários. Recentemente, tem sido contabilizado aumento expressivo de veículos a cada ano. O estado de Pernambuco, localizado na região Nordeste do Brasil, possuía, em 2015, uma frota de 2.765.521 de veículos, sendo que boa parte dessa frota trafega pelas onze (11) rodovias federais que cruzam o estado. A Polícia Rodoviária Federal e outros órgãos de controle público atendem e registram os acontecimentos nessas rodovias. Nosso objetivo nessa pesquisa, foi o de propor um modelo de predição de comportamento das rodovias, de modo a tornar possível a escolha de dias, horários e locais para trafegar com menos interrupções devido ao fluxo do trânsito intenso. O mesmo interesse neste assunto foi despertado em outros pesquisadores. Costa, Bernardini, Lima e Viterbo em seu artigo “A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras” destacam algumas pepitas encontradas como sendo a **falta de atenção** do condutor a principal causadora dos acidentes. Eles utilizaram árvores de decisão e o algoritmo gerador de regras de associação Apriori proposto por Agrawal, Imielinki e Swami(1993).

1. Fundamentação teórica

*A* técnica de extração do conhecimento de grandes bases de dados é conhecida como “Knowledge Discovery Databases” (KDD). No processo de extração do conhecimento o KDD se caracteriza pela aplicação de algoritmos específicos para descoberta de padrões e/ou comportamentos em grandes bases de dados, também conhecidas como repositórios de dados. A mineração se distingue das técnicas estatísticas pelo fato de que não trabalha com dados hipotéticos, mas se apoia nos próprios dados para extrair os padrões[1]

* 1. O CRISP-DM

O “CRoss Industry Standard Process for Data Mining” CRISP-DM é um processo de mineração de dados que descreve como especialistas nesse campo aplicam as técnicas de mineração para obter os melhores resultados[2]. O CRISP-DM é um processo recursivo, em que cada etapa deve ser revista até quando o modelo apresentar os resultados satisfatórios, preliminarmente definidos. O Analista de Dados ou o Cientista de Dados é o profissional que acompanha e executa o processo.

O contexto da aplicação do CRISP-DM[2] é guiado desde o nível mais genérico até o nível mais especializado, sendo normalmente explicado em quatro dimensões:

* O domínio da aplicação - a área específica que o projeto de mineração de dados acontece;
* O tipo de problema - descreve as classes específicas do objetivo do projeto de mineração de dados;
* Os aspectos técnicos - cobrem as questões específicas como os desafios usualmente encontrados durante o processo de mineração de dados;
* As ferramentas e técnicas - dimensão específica que cada ferramenta/técnica de mineração de dados é aplicada durante o projeto.

A aplicação das técnicas de mineração de dados identifica padrões ocultos nos dados, inacessíveis pelas técnicas tradicionais, como por exemplo, consultas a banco de dados, técnicas estatísticas, dentre outras. Além disso, possibilita analisar um grande numero de variáveis simultaneamente, o que não acontece com o cérebro humano[3]. Fayyad[4] destaca a natureza interdisciplinar do KDD que contempla a intersecção de campos de pesquisa tais como Aprendizado de Máquina (Machine Learning), Reconhecimento de Padrões, Inteligência Artificial, estatística, computação de alto desempenho e outros. Propõe, ainda que o objetivo principal seja extrair um conhecimento de alto nível a partir de dados de baixo nível.

O modelo CRISP-DM contempla seis fases para um projeto de mineração de dados, sendo assim determina-se um ciclo de vida compreendido para cada uma dessas fases: A primeira fase, conhecida como Entendimento do negócio, ou “fase de entendimento dos objetivos é dos requerimentos sob a perspectiva do negocio” segundo CHAMPAN [5] e uma fase crucial da mineração. A fase Entendimento dos dados caracteriza-se pelo exame acurado dos dados, procurando identificar a sua qualidade. Dados ausentes “missing data” – são comuns em base de dados não estruturados, configurando-se como um problema a ser considerado. A terceira fase, Preparação dos dados diz respeito sobre a construção final do conjunto dos dados, cada conjunto de dados é “explicado” por um atributo, para selecionar quais dados serão mais relevantes calcula-se o coeficiente de correlação linear entre os atributos (variáveis). A formula a seguir exemplo do cálculo do coeficiente de correlação linear R.

* 1. Equação do coeficiente de correlação linear

R = (1)

Nesta equação x e y são as variáveis que se pretende avaliar quanto a relação de uma com a outra. Esse coeficiente pode variar entre -1 e +1, quando o coeficiente está próximo de +1 significa que ambas variam na mesma direção, ou seja caso uma aumente a outra aumenta, quando é negativo elas variam em direções opostas, se uma aumenta o coeficiente a outra diminui.

Na quarta fase, Modelagem de I.A. a tecnologia deve ser escolhida de forma criteriosa, baseada na experiência do analista de dados. Na fase cinco, Avaliação de desempenho um ou mais modelos devem ter sido construídos e testados de forma que o modelo esteja adequado aos objetivos do negócio. A sexta e última fase, Implementação caracteriza-se pela conclusão do processo, pode ser necessário retomar todo o ciclo até que o modelo esteja adequado as necessidades especificas determinadas previamente.

**Domínio das técnicas aplicadas à mineração de dados**

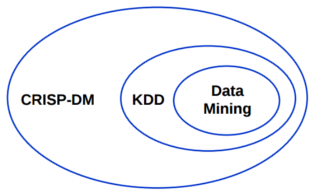


Figura 1: Domínio CRISP-DM [6]

* 1. Aprendizagem de Máquina

Aprendizagem de Máquina ou “Machine Learning” são técnicas para analisar dados de forma automatizada e interativa. Segundo Shalev-Shwartz & Bem-Davis[7] o termo Aprendizagem de maquina refere-se a detecção automática de padrões de dados.

Para Nilsson[8], o aprendizado ocorre quando uma maquina modifica sua estrutura interna, programa ou dados (baseados nos *inputs* ou em uma resposta para informação externa), de tal forma que melhora o desempenho futuro. Sistemas que executam tarefas de inteligência artificial, tais como Reconhecimento de padrões, Diagnósticos, Controle de Robôs, Predição e outros, precisam ser modificados para executarem “Machine Learning”.

* 1. Classificação e Regressão para análise preditiva.

Classificação é um processo para encontrar um modelo que descreve e distingue classes de dados. Esse modelo tem como base de análise um conjunto de treinamento (i.e. objetos de dados para os quais serão encontrados rótulos que os classifiquem) e é utilizado para predizer quais rótulos de classes terão os objetos desconhecidos.

* 1. Árvores de Decisão

Tem como entrada um conjunto de atributos (variáveis) para retornar como saída um decisão. O valor esperado da saída deve estar de acordo com o que foi dada a entrada.

Árvores de decisão são algoritmos rápidos que produzem regras de indução. A fase de extração dos dados é fortemente influenciada pelas variáveis escolhidas, isso pode representar um desafio maior para implementar essa técnica. Outro problema frequentemente encontrado em algoritmos de aprendizagem é o “overfitting” ou superadaptação aos modelos. Segundo RUSSEL E NORVIG[9] o “overfitting” ocorre quando o numero de atributos e grande e o algoritmo “deixa” de aprender.

A aplicação do CRISP-DM nesta pesquisa ajudou a guiar as escolhas nos momentos em que os resultados pareciam não fazer sentido algum, contudo por ser um processo recursivo, o retorno aos fundamentos dessa metodologia prevê que hajam os ajustes necessários para atingir os objetivos.

1. ideia proposta

A ideia metodológica proposta para esta pesquisa contemplou incialmente as fases do KDD conforme descrito a seguir.

* 1. Fases do KDD

*Target Data*: Nesta etapa foram coletadas as informações provenientes das bases de dados da Policia Rodoviária Federal (PRF) de 2007 a 2015, uma vez que nosso interesse era o de analisar os últimos dez anos, no entanto, como a base de dados so dispunha de informações a partir de 2007, foram considerados os nove anos disponíveis. A PRF dispõe em banco de dados relacionais alguns desses dados na Internet, contudo no artigo “Uma análise da qualidade dos dados relativos aos boletins de ocorrências das rodovias federais para o processo de Mineração de Dados” COSTA, BERNARDINI, LIMA[10] destacam a não padronização e não aceitação dos dados pela comunidade internacional. EAVES, D.[11] sugere que os dados sejam disponibilizados na maneira como foram coletados. A primeira base de dados coletada diretamente dos servidores da PRF continha relatório de acidentes e a segunda a de interdições. A partir dos dados capturados na base da PRF utilizamos como variáveis de entrada

* BR – Nomenclatura da rodovia (i.e. BR 101);
* Km – Quilômetro em que deu a ocorrência;
* Tipo de veículo – envolvido na ocorrência, ex.: carro, motocicleta, caminhão, etc.;
* Tipo de acidente – colisão lateral, frontal, traseira, etc.; atropelamento: com ou sem morte, envolvendo pessoas e/ou animais;
* Horário e data da ocorrência; dentre outros que serão apresentados mais adiante.

*Preprocessing*: Nesta fase foram retiradas as variáveis, sendo consideradas alguns atributos, pelo fato de conterem inconsistência e “missing data”, como, por exemplo, informações acerca de latitude e longitude. Cabe destacar que a base, como um todo, apresentava sérias inconsistências, uma vez que, por exemplo, um mesmo acidente, quando envolvia doi ou mais veículos, era lançado na base duas ou mais vezes, em função da quantidade de veículos envolvidos. Foram eliminadas variáveis em duplicidade (i.e. as variáveis Mês, Ano que apareciam separadamente, já aviam sido contempladas na variável Data.). *Transformation*: Foram criadas as variáveis “Tipo de paralisação”, contemplando acidentes sem mortos e com, no máximo, dois veículos envolvidos; “Dias da semana” (domingo, segunda-feira,….sábado); “Ajuste de horas” (i.e. 17h58, 17h59, 18h, 18h01, 18h02, arredondadas para 18h); “Ajuste de Km” (seguiu a mesma lógica do ajuste de horas). *Data mining*: O algoritmo escolhido para a pesquisa foi Árvore de decisão que possibilita uma interpretação imediata e de fácil compreensão. Como ferramentas, foram escolhidas o Knime e o Weka e o R, com objetivo de estabelecer uma comparação entre ambos, cuja intenção era produzir um classificador mais preciso. Nessa direção, a técnica Ensamble de classificadores [12] afirma que a combinação de um ou mais classificadores iguais, ou mais de um classificador diferente, aumenta a precisão. Tanto na ferramenta Knime com Weka o algoritmo é chamado de J48, uma vez que se trata da implementação Java do algoritmo C4.5, no R a biblioteca “rparty” implementa esse algoritmo. Para escolha das variáveis de *input* foi calculado a correlação linear entre todas as variáveis, entre as variáveis BR e Delegacia (variável que agrega municípios) obteve correlação linear de 0,653, já entre Tipo de Acidente e Traçado via a correlação foi baixa, apenas 0,14, variáveis com correlação linear abaixo disso foram descartadas. *Interpretation/Evaluation*: Produção de árvores de decisão a partir do estabelecimento de diferentes nós-raízes, definidos em virtude da correlação linear encontrada.

* 1. Dados encontrados antes da mineração

Acreditamos que os condutores dos automóveis têm um papel preponderante na causa dos acidentes. Os gráficos a seguir ajudam a elucidar o “modus operandi” do condutor das rodovias federais do nordeste brasileiro.

Tabela 1: Gráficos de acidentes BR X Hora

|  |  |  |
| --- | --- | --- |
| **BR 101** | **BR 232** | **BR 407** |
|  |  |  |
|  |  |  |
|  |  |  |

Fonte: Autor

A Tabela 1 contém gráficos de três BRs, as duas primeiras colunas são das rodovias mais utilizadas no Estado de Pernambuco (BR 101, BR 232), com tráfego intenso, a terceira coluna, BR 407, pouco utilizada em relação as duas primeiras, portanto com tráfego pouco intenso. Os gráficos da primeira linha da Tabela 1 representam os acidentes que ocorreram a cada hora (abcissa) em cada Km (ordenada) nos últimos 9 anos, na segunda linha são gráficos de frequência do local onde ocorreram esses acidentes, é possível perceber que há um determinado local, em cada rodovia onde se concentram os acidentes, na terceira linha os gráficos tipo boxplot exibem a concentração das ocorrências em torno da mediana dessa localidade (Km). Especulou-se a priori que a variável “traçado da rodovia” ou que as condições climatéricas poderiam ter influência na causa dos acidentes, contudo mais adiante descobrimos outros condicionantes influenciam mais fortemente essas ocorrências. É possível perceber na primeira linha da tabela 1 alguns padrões especialmente em determinados locais (Km), por exemplo na BR 101 entre os Km 40 e 100 ocorrem acidentes a partir da 05h da manhã até as 23h, já na BR 232 há um padrão nos acidentes nos Km 0, 90, 110, 260, 410 e 500, e na BR 407 os acidentes se concentram na altura do Km 130.

Dados encontrados após a mineração

A figura a seguir exibe a matriz de correlação linear que selecionaram as variáveis de *input* para os algoritmos de IA.

**Matriz de correlação linear**

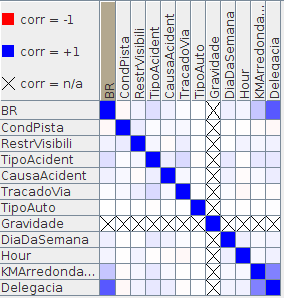


Figura 2: Correlação linear entre todas as variáveis

A tonalidade do azul significa que há uma correlação linear positiva entre as variáveis (vermelho seria negativa), com exceção da variável com ela própria. Variáveis como: ‘TipoAuto’ (tipo de automóvel envolvido no acidente) e ‘Hour’ (Hora da ocorrência) tiveram pouca influência no modelo, por outro lado variáveis como ‘RestrVisibili’ (restrição de visibilidade), ‘Delegacia’ (delegacia da PRF, agrega vários municípios), ‘TipoAcident’, KmArredondado, onde ocorreu o acidente, obtiveram índice de correlação linear alto, portanto foram relevantes para o modelo de predição.

Os resultados da classificação da árvore de decisão mais interessantes foram os seguintes:

* Colisão transversal ~ Região metropolitana (km<=55) ~ Automóvel ~ Acidente sem vítima => Falta de atenção
* Colisão transversal ~ Região metropolitana (Km>55) ~ Automóvel ~ Acidente com vítima => Desobediência à sinalização

A mesma configuração de acidentes foi encontrada para colisão lateral, colisão frontal.

A tabela a seguir resume a classificação para a variável “Causa do Acidente” encontrada no Weka

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Classe** | **Tx TP** | **Tx FP** | **Prec.** | **Recall** | **F-M** | **AUC** |
| Outras | 0.598 | 0.222 | 0.479 | 0.598 | 0.532 | 0.757 |
| Falta de atenção | 0.837 | 0.402 | 0.555 | 0.837 | 0.667 | 0.781 |
| Velocidade incompatível | 0.066 | 0.003 | 0.502 | 0.066 | 0.117 | 0.816 |
| Ultrapassagem indevida | 0.077 | 0.003 | 0.407 | 0.407 | 0.129 | 0.828 |
| Desobediência à sinaliz. | 0.064 | 0.002 | 0.512 | 0.064 | 0.114 | 0.849 |
| Defeito mecânico | 0.157 | 0.005 | 0.544 | 0.157 | 0.244 | 0.849 |
| Animais na Pista | 0.713 | 0.005 | 0.882 | 0.713 | 0.789 | 0.956 |
| Não guardar distância | 0.232 | 0.017 | 0.566 | 0.232 | 0.329 | 0.875 |
| Dormindo | 0.045 | 0.001 | 0.543 | 0.045 | 0.084 | 0.862 |
| Defeito na via | 0.168 | 0.004 | 0.568 | 0.168 | 0.26 | 0.87 |
| Ingestão de álcool | 0.065 | 0.003 | 0.467 | 0.065 | 0.114 | 0.812 |

TP: True Positive; FP: False Potive; Prec: Precison; F-M: f-measure ou f-score; AUC: Area Under Curva (Roc)

Prec = TP/(TP +FP)

Recall = TP/ (TP + FN)

F-Measure = 2 \* Precison \* Recall / (Precision + Recall)

A árvore construída pelo Knime para a mesma variável “Causa do Acidente é a seguinte:



1. Conclusões

A contribuição dessa pesquisa é de cunho metodológico-prático. Do ponto de vista metodológico pela contribuição pela aplicação do processo CRISP-DM, usado para construir o modelo preditivo; do ponto de vista prático pela proposição de um modelo que integre predição à API de mapas de posicionamento global, fornecendo informação suficiente a um gestor decidir quando enviar uma frota de caminhões por determinada rodovia que apresente retenções crescentes de logística de cargas.

As soluções disponíveis que existem, tais como: Google Mpas, Waze e outros dessa natureza somente exibem informações momentâneas, produzidas e compartilhadas pelos utilizadores dos aplicativos ou por informações provindas de GPS, contudo não analisam dados históricos dessas rodovias não fazem predições sobre seu comportamento.

Agradecimentos (Heading 5)

Introduza agradecimento a ...

Referências Bibliográfica

1. L. Castanheira, “Aplicação de técnicas de Mineração de Dados em Problemas de Classificação de Padrões”, 2008, pp553.
2. R. Wirth, “CRISP-DM 1.0 – Step-by-step data mining guide”, 2000, pp. 7– 10.
3. B. Possas, M. Carvalho, R. Rezende, and W. Meira jr., “Data mining: técnicas para exploração de dados”, Universidade Federal de Minas Gerais, 1998.
4. P. Fayyad, U. Piatetsky-Suapiro, and G. Smyth, “From data mining to knowledge discovery in databases”, Advances in Knowledge Discovery and Data Mining, 3rd ed, vol.17, 1996, pp.1– 36.
5. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Thomas, C. Shearer, and R. Wirth.
6. Neurotech SA – 2012, <http://www.neurotech.com.br/> acessado em: 20/01/2017
7. S. Ben-David, and S. Shalev-Shwartz, title: Understanding, “Machine Learning: From Theory to Algorithms, booktitle: Understanding Machine Learning: From Theory to Algorithms’, 2014, pp. 449.
8. J. N. Nilsson, “ Introduction to Machine Learning” vol. 56, pp. 387 – 389, 2005.
9. S. Ruseel and P. Norvig, “Inteligência Artificial.” .Elsevier, Rio de Janeiro, 3rd ed, pp. 716 – 721, 2004.
10. J. Costa, F. Bernardini, F. FILHO, “A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. AtoZ: novas práticas em informação e conhecimento”, v. 2, p. 1–26, 2015. Disponível em: <http://ojs.c3sl.ufpr.br/ojs/index.php/atoz/rt/printerFriendly/41346/25356>.
11. D. EAVES, “The three laws of open government data.”, v. 30, 2009.
12. F. Bernardini, “Combinação de classificadores simbólicos utlizando medidas de regras de conhecimento e algoritmos genéticos”, Instituto de Ciências e Matemática Computacional/USP, tese de doutorado, 2006.