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

Email: 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 Clave – Modelo de predição; Mineração de dados; Tráfego em rodovias; Árvores de decisão, CRISP-DM.

Abstract — The federal highways that cross metropolitan regions are constantly congested, not be by number of vehicles, but because they are target of manifestations of the most diverse, such as, strike, road traffic accident, holes, natural inclemencies, and other types of downtime. In extreme situations could even paralyze factories and in their environment. We propose a classification model for the behavior of federal highways that cross the state of Pernambuco in Northeast of Brazil

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 rodovias que cruzam o estado. A Polícia Rodoviária Federal e outros órgãos de controle público atendem e registram os acontecimentos nessas rodovias. A solução para absorver parte dessas informações requer o estabelecimento de várias etapas, para a definição dos dias, horários e rotas, escolhidos por critérios cientificamente estudados para acompanhar os veículos e conduzindo-os até seu destino de maneira segura e no menor tempo possível.

1. Fundamentação teórica

Utilizar técnicas de mineração de dados, além de extrair dados, extrai conhecimento, com isso é possível predizer resultados futuros. Essa 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.

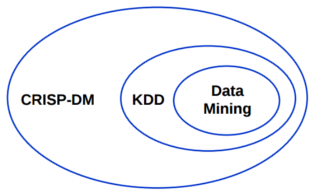
O contexto da aplicação do CRISP-DM[2] é guiado desde o nível mais genérico até o nível mais especializado, sendo normalmente explicado em quatro dimensões:

* 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 Maquina (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 dos 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. 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 ate que o modelo esteja adequado as necessidades especificas determinadas previamente.

Domínio das técnicas aplicadas à mineração de dados



Fonte: Neurotech-2012 [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). Esse modelo é usado para predizer quais rótulos de classes terão os objetos desconhecidos.

* 1. Árvores de Decisão

Uma árvore 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.

1. ideia proposta

A ideia metodologica 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 continham relatório de acidentes e a segunda a de interdições. A partir dos dados capturados nas 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” (segiu a mesma lógica do ajuste de horas). Data mining: O algoritmo escolhido para a pesquisa foi Árvore de decisão que possibilita um interpretação imediata e de fácil compreensão. Como ferramentas, foram escolhidas o Knime e o Weka e o R, com obejetivo 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 impelemntação Java do algoritmo C4.5, no R a biblioteca “rparty” implementa esse algoríttmo. 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 linar 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.

* 1. Dados encontrados após a mineração

Os gráficos gerados antes da mineração permitiram clarificar sobre o comportamento do condutor das rodovias. A causa dos acidentes a priori sugeriam ter sido causados pelo traçado

* 1. Los títulos de las Secciones

Hay dos tipos de títulos de Secciones: Títulos de Componentes e Títulos de Texto.

Los Títulos de Componentes identifican los diferentes capítulos tales como "Agradecimientos" ("Agradecimientos") y Referencias ("Referencias") y deben tener el formato con el estilo "Título 5". Utilice “figure caption” para figuras, y "table head” para los títulos de las tablas. Normalmente, el más sencillo es copiar y pegar ("copy-paste") de un título o subtítulo y cambiar el contenido existente convenientemente (donde la numeración se realiza automáticamente por la plantilla).

Títulos de Texto organizan los temas de artículo de moda relacional y jerárquica. Los estilos “Heading 1”, “Heading 2”, “Heading 3”, e “Heading 4” deben ser utilizados.

* 1. Figuras y Tablas
     1. Posicionamiento de Figuras y Tablas: Lugar pies de figura por debajo de las figuras y los títulos de las tablas sobre las tablas. Inserte figuras y tablas después de que se citan en el texto. Use la abreviación "Fig. 1 " para se refieren a las figuras, incluso al principio de una frase.

1. Ejemplo de un pie de figura (figure caption)

Leyendas de Figuras: Use 8 puntos Times New Roman. Utilice palabras en lugar de símbolos o abreviaturas. Buscar figuras legibles incluso si no se imprimen en color (por ejemplo, el uso de diferentes texturas y tipos de línea en lugar de los colores).

1. Título Ejemplo de una Tabla

| Tabla | Tabla | | |
| --- | --- | --- | --- |
| Columna 1 | Columna 2 | Columna 3 |
| Línea 1 | Contenido |  |  |
| Línea 2 |  |  |  |



1. Ejemplo de un pie de figura (figure caption)
2. Conclusiones

El modelo IEEE y otros modelos disponibles, facilitan el trabajo de los resúmenes de congresos de edición y la producción de material científico. Utilizar correctamente al escribir su artículo y deseos para un gran trabajo y contribuir al éxito de nuestra conferencia.

Agradecimientos (Heading 5)

Introduzca agradecimiento a las organizaciones y personas que contribuyeron a la obra. Si la presentación de la conferencia es Double-Blind, introduzca sólo texto DFBR (Deleted for Blind Review) en señal de agradecimiento.

Referências Bibliográfica

Los números de las citaciones son automáticos e entre corchetes [1] [2] [3]. A menos que haya seis autores o más se incluyen todos los autores y no utilizan "et al.". Los artículos que no hayan sido publicados deben ser citados como "unpublished" [4]. Los artículos aceptados para publicación deben ser citados como "in press" [5]. Para los artículos publicados en un idioma que no sea inglés, por favor poner la cita en inglés primero seguido de la cita en el idioma original [6].

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.