Curso 2 – CD, AM e DM

Profa. **Roseli Ap. Francelin Romero**

MBA em Inteligência Artificial e BigData

Depto. de Ciências de Computação
ICMC - USP

PRE-PROCESSAMENTO DE DADOS (cont.)

DADOS DESBALANCEADOS

Profa. Roseli A. F. Romero

SCC – ICMC - USP

2º.ETAPA

PRE-PROCESSAMENTO

- Limpeza dos Dados
  - Ausentes
  - Inconsistentes
  - Redundantes
- Transformação dos Dados
  - Codificação
  - Normalização
  - Padronização
  - Redução de Dimensionalidade

Dados desbalanceados

- Número de objetos varia para as diferentes classes
  - Natural ao domínio
  - Problema com geração / coleta de dados
- Várias técnicas de AM não conseguem lidar com esse problema
  - Tendência a classificar na(s) classe(s) majoritária(s)

Motivação: Detecção de Fraudes

- Transações, em sua maioria pela internet, com dados de outras pessoas: os chamados fraudadores.

- Pesquisas indicam que em 2019, o prejuízo de lojistas e consumidores somam mais de 1.8 bilhão de reais.

- Portanto, é cada vez mais importante a existência de análises anti-fraude afim de prevenir que o evento fraude ocorra.

Motivação: Detecção de CRIMES

- FURTOS ocorrem diariamente:
  - FURTO DE CELULAR
  - FURTO DE VEÍCULOS
  - ROUBO DE CELULAR
  - ROUBO DE VEÍCULOS
- CRIMES
  - LATROCINIO
  - FEMINICIDIO
  - LESÃO CORPORAL, SEGUIDA DE MORTE

Desafio no Kaggle

- No site de desafios em Ciência de Dados, Kaggle, é fácil encontrar desafios relacionados a prevenção de fraudes.
- No caso do desafio ***Credit Card Fraud Detection*** vários modelos de AM tem sido testados.
- O objetivo deste desafio é encontrar o modelo que melhor discrimina fraudadores e não fraudadores.

Desafio no Kaggle

- Se considerarmos a base de comercio eletrônico no Brasil que **contem 1.041.356** registros de transações que ocorreram no período entre Out/2014 e Fev/2016.
- Para estimação dos parâmetros: o período entre Out/2014 e Mar/2015,
- Para avaliação do desempenho dos algoritmos o período entre Abr/2015 e Fev/2016
- A base possui 102 variáveis.

Desafio no Kaggle

Table 1. Distribuição da variável 'Frd'

Frd	Volume	Volume (%)
Não Fraude	1.015,043	97.5%
Fraude	26.313	2.5%

Dados estão desbalanceados

BASE DE DADOS de CRIMES

CATEGORIAS DE CRIMES	PERIODO	NUMERO DE REGISTROS
FEMINICIDIO	ABR/2015 a DEZ/2019	1.122
FURTO DE CELULARES	JAN/2017 a DEZ/2019	591.166
FURTO DE VEÍCULOS	JAN/2017 a DEZ/2019	408.294
LATROCINIO	JAN/2017 a DEZ/2019	32.867
LESÃO CORPORAL, SEGUIDA DE MORTE	JAN/2017 a DEZ/2019	487
HOMICIDIO DOLOSO (COM FURTO)	JAN/2017 a DEZ/2019	3.409
ROUBO DE CELULARES COM VIOLÊNCIA	JAN/2017 a DEZ/2019	878.069
ROUBO DE VEÍCULO COM VIOLÊNCIA	JAN/2017 a DEZ/2019	438.843
TOTAL		2.354.257

Dados desbalanceados

- Alternativas
  - Alteração do conjunto de dados
    - Balanceamento artificial
  - Utilizar diferentes custos de classificação para as diferentes classes
  - Induzir um modelo para uma das classes
  - Alteração do projeto de algoritmos para lidar com desbalanceamento

Como é feito o Balanceamento artificial ?

- Redefinir o tamanho do conjunto de dados:
  - Sobreamostragem (Oversampling). Acrescentar objetos
    - Replicar objetos da classe minoritária não adiciona informação
  - Subamostragem (Undersampling). Eliminar objetos
    - Ignorar objetos da classe majoritária
  - Abordagem híbrida

Oversampling

Os exemplos são replicados com base nos vários registros existentes até que a base fique balanceada, ou seja,

- 50% de fraudadores
- 50% não fraudadores

Undersampling

São retiradas várias amostras
da base até que a base fique balanceada, ou seja,
- 50% de fraudadores
- 50% não fraudadores

SMOTE - Synthetic Minority Oversampling Technique

Funcionamento: novas observações são adicionadas, porém com um ganho na informação, sem simplesmente duplicar registros.

SMOTE

- O SMOTE procura sintetizar novas instâncias minoritárias em instâncias reais, levando em conta o comportamento das instâncias mais próximas (chamados de vizinhos).

- O algoritmo seleciona os k vizinhos mais próximos, ou seja, com as menores distâncias euclidianas, de cada elemento da classe **minoritária** para criar novas amostras sintéticas.

SMOTE - Algoritmo

- Para cada registro da **classe minoritária**, encontra-se os k vizinhos mais próximos de tal modo que sejam todos desta mesma classe.
- Encontra-se a diferença entre o vetor de variáveis do registro considerado e os outros k vizinhos mais próximos, obtendo-se assim k vetores de diferenças.
- Cada um destes vetores é multiplicado por um valor aleatório entre 0 e 1.
- Adiciona-se estes vetores das diferenças, multiplicado por um valor aleatório, à cada instancia (classe minoritária original), para cada iteração, até encontrar uma nova base com dados balanceados.

No ex. Desafio do Kaggle

Foram considerados os 6 vizinhos mais proximos, ou seja, k = 6.

EXEMPLO 4

- SOBRE BALANCEAMENTO DE DADOS