Segmentação e classificação de Big Data

Gustavo Kendi Tsuji
Universidade de São Paulo

Introdução

- Hoje é possível encontrar mais de 60 trilhões de páginas indexadas pelo Google[6]. Só o Facebook possui *data warehouses* com mais de 300 *petabytes* e um tráfego de mais de 600 *terabytes* diários[7]. Mas o grande desafio já não é o armazenamento desses dados. Extrair conhecimento dessas informações se tornou o aspecto essencial para as empresas. Trata-se de um ativo intangível estratégico que viabiliza o desenvolvimento de novos produtos ou serviços bem como um melhor entendimento sobre comportamento de clientes e funcionamento de processos ou otimização da produção. Este trabalho apresentará uma abordagem para explorar o que é possível desenvolver num contexto de *Big Data* com algoritmos de *Machine Learning*.

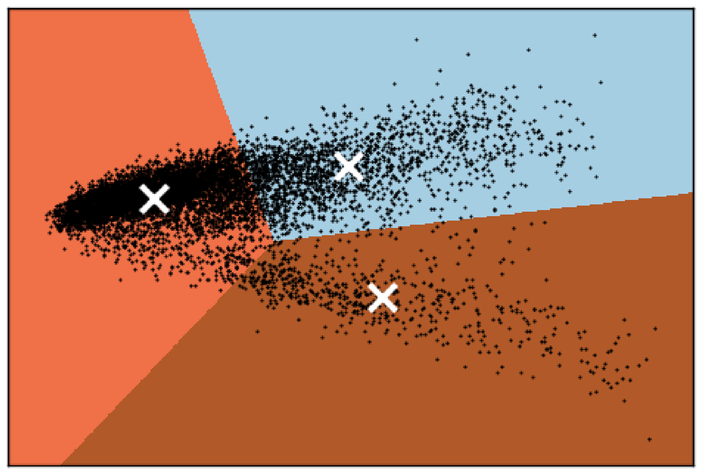
Metodologia

- Realização de experimento quantitativo de segmentação e classificação de dados. Este estudo utilizou uma base disponível no site kaggle[2] com mais de 800 mil registros. Ela possui informações de perfil de mutuários a detalhes das transações. Pertence a Loan Club, uma empresa de empréstimos com um sistema online.
- A análise consiste em:
  - Estudo dos algoritmos de segmentação K Médias (KM) e de classificação Regressão Logística (RL) e Random Forest (RF)
  - Análise descritiva dos clusters gerados pelo KM
  - Comparação da classificação da RL e RF

Fundamentação teórica

- Clusterização
  - O KM é um método de aprendizagem não supervisionado que reune os elementos em grupos baseados na similaridade entre eles. [5].

Figure: Ilustração da clusterização de uma amostra de 8000 registros

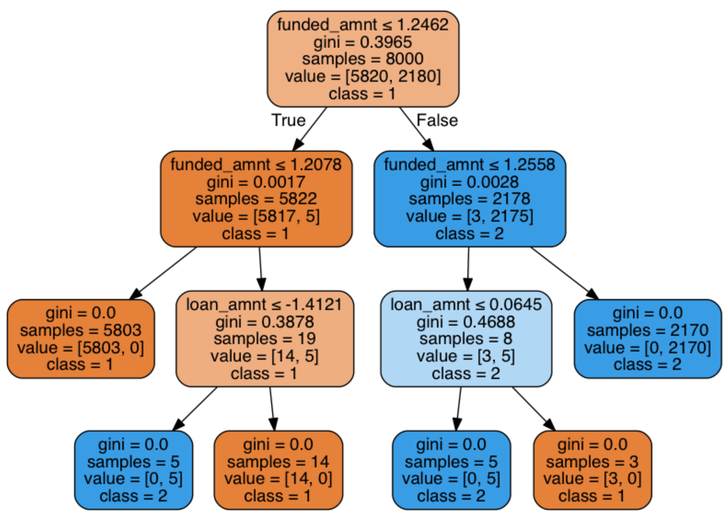


  - Essa similaridade é baseada na minimização da distância Euclidiana entre os vetores de atributos e os centróides dos clusters.
- Classificação
  - A RL é um modelo supervisionado que estuda a relação entre variáveis com o intuito de predizer a ocorrência de eventos. Ao ser treinado, a RL gera uma fórmula que pode classificar dados usando como base o *odd ratios*, isto é, as chances de que um determinado evento ocorra[4].

$$logit^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}} \quad (1)$$

  - A RF é um método que gera várias árvores para efetuar a classificação com o intuito de reduzir problemas de viés de usar somente uma.

Figure: Árvore de decisão para a base da Loan Club



  - A árvore de decisão é uma estrutura de modelo preditivo de fácil interpretação utilizado na aprendizagem supervisionada [4]. Ela define o fluxo que classifica os dados em cada uma de suas extremidades.

Objetivos

1. Segmentar e classificar dados usando Scikit Learn[3] e Apache Spark[1]

Resultados

- Clusterização
  - Usando o Scikit Learn foi possível visualizar diversas clusterizações. Foi escolhido o particionamento dos dados em 3 grupos.

Figure: Resultados da clusterização

(a) Ilustração do silhoutte score para 3 clusters

(b) Box plot por cluster do campo loan_amnt_by_cluster

- Classificação
  - Baseado na clusterização realizada pelo KM, utilizou-se uma *Cross Validation* treinando 80% da base com RL e RF e classificando os 20% restantes
  - Após o treinamento, gerou-se as Confusion Matrices de cada algoritmo

Figure: Confusion Matrices

(a) Confusion Matrix da RL

(b) Confusion matrix da RF

  - Pelos estudos, é possível compararmos as métricas de cada classificação e concluimos que a RL teve um melhor resultado.

Table: Métricas de assertividade e eficiência da RL e da RF

Classe	Precisão		Recall		Falso Positivo		F-measure	
	RL	RF	RL	RF	RL	RF	RL	RF
1	0,9349	0,8606	0,9474	0,9317	0,0087	0,0186	0,9411	0,8948
2	0,9167	0,8278	0,9264	0,8663	0,0264	0,0540	0,9523	0,8466
3	0,9690	0,9619	0,9628	0,9330	0,0555	0,0713	0,9524	0,9472

Conclusão

- Ambas ferramentas se mostraram que se complementam: o Scikit Learn possui muitos recursos que facilitam a visualização e a compreensão dos modelos, já o Apache Spark é robusto e eficiente, executando os algoritmos de forma rápida e escalável, ideal para o cenário de Big Data.
- A segmentação gerou 3 clusters e foi possível observar as diferenças entre si.
- Na classificação, mesmo com uma menor assertividade, a RF também é um algoritmo com alta eficiência. Possivelmente com outra configuração (redução de cluster, aumento de profundidade ou aumento de árvores) o resultado poderia ser diferente.

Referência Bibliográfica

[1] Apache spark - lightning-fast cluster computing.
http://spark.apache.org/.

[2] Kaggle: Your home for data science.
https://www.kaggle.com/.

[3] Scikit-learn: machine learning in python.
http://scikit-learn.org/stable/.

[4] Hastie, T., Tibshirani, R., and Friedman, J.
The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics).
Springer, 2011.

[5] MacQueen, J.
Some Methods for classification and Analysis of Multivariate Observations.
Berkeley, University of California Press, 1967.

[6] Smith, C.
By the numbers: a gigantic list of google stats and facts.
http://expandedramblings.com/index.php/by-the-numbers-a-gigantic-list-of-google-stats-and-facts/, 2016.

[7] Vagata, P., and Wilfong, K.
Scaling the facebook data warehouse to 300 pb.
https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/, 2014.