Controle de segurança para cartões de crédito usando K-Means

Fabrícia de Jesus Santos[1],
Thiago Alessandro dos Santos Pereira[1], **Weslley Henrique Campos Santos**[1]

[1]Departamento de Sistemas de Informação
Universidade Federal de Sergipe (UFS) – Itabaiana, SE – Brasil

fabriciacooper@gmail.com

thi.alessandro@hotmail.com, weslley_campos@outlook.com

1. Introdução

Devido a grande ocorrência de clone de cartões de crédito e débito, as empresas de cartão investem como mitigar esse tipo de crime. Recentemente, nos Estados Unidos passaram do uso da faixa magnética insegura em cartões de crédito e débito para cartões de chip e PIN mais bem protegidos, regulados pelo padrão EMV.

Aqui no Brasil o uso de cartões com chips já vem acontecendo alguns anos, porém pesquisadores descobriram que cibercriminosos brasileiros desenvolveram uma maneira de roubar dados e clonar cartões de chip e PIN. Na maioria dos casos, as vítimas não são informadas no seu devido tempo, ocasionando um transtorno, sendo que em alguns casos são solucionáveis e em outros não.

Geralmente, as empresas tentar informar aos seus clientes como manter-se informado sobre alguma anormalidade. Entre essas informações estão:

1. Verificar histórico de transações do cartão;
2. Utilizar aplicativo como AndroidPay ou o ApplePay para não revelar os dados do seu cartão;
3. Ter bastante cuidado ao realizar compras pela internet, pois é possível encontrar sites fraudados que podem roubar os dados;

O Brasil é um dos países campeões de fraude on-line, dentre outros motivos, por ter sido também um dos pioneiros na adoção do cartão chipado, que promove um alto grau de segurança para transações presenciais. Com o desuso da tarja magnética nos sistemas de pagamentos, os golpes migraram para a internet, tornando e-commerce um dos mais vulneráveis do planeta.

2. Objetivo

O objetivo do projeto é apresentar uma aplicação mobile que monitore as compras do consumidor ao usar seu cartão de crédito. Utilizando a aplicação, a pessoa poderá receber uma notificação, caso existe alguma compra realizada fora do padrão. Dessa forma, a própria pessoa estará informada sobre suas transições, e caso seja identificado alguma anormalidade, a vitima poderá entrar em contato com a empresa de cartão de crédito para tomar as medidas cabíveis.

A aplicação adota algoritmo K-means, que será explicado logo em seguida.

3. K-means

K-means é um tipo de aprendizado não supervisionado, que é usado quando você tem dados sem rótulo, ou seja, dados sem categorias ou grupos definidos. O objetivo deste algoritmo é encontrar grupos, sendo que o número de grupos é representado pela variável K. O algoritmo funciona iterativamente para atribuir cada ponto de dados a um dos grupos K com base nos recursos fornecidos. Os pontos de dados são agrupados com base na similaridade do recurso. Os resultados do algoritmo de agrupamento K-means são:

- Os centroídes dos clusters K, podem ser usados para rotular novos dados;
- Etiquetas para os dados de treinamento (cada ponto de dados é atribuído a um único cluster).

Em vez de definir grupos antes de examinar os dados, o armazenamento em cluster permite localizar e analisar os grupos que foram formados organicamente. Cada centroide de um cluster é uma coleção de valores de recursos que definem os grupos resultantes. Examinar os pesos dos recursos do centróide pode ser usado para interpretar qualitativamente o tipo de grupo que cada cluster representa.

3.1. Uso no ramo de negócios

O algoritmo K-means é usado para localizar grupos que não foram explicitamente rotulados nos dados. Isso pode ser usado para confirmar suposições de negócios sobre quais tipos de grupos existem ou para identificar grupos desconhecidos em conjuntos de dados complexos. Depois que o algoritmo é executado e os grupos são definidos, qualquer novo dado pode ser facilmente atribuído ao grupo correto.

Este é um algoritmo versátil que pode ser usado para qualquer tipo de agrupamento. Alguns exemplos de casos de uso são:

- Segmentação comportamental:
  - Segmentar por histórico de compras;
  - Segmento por atividades no aplicativo, site ou plataforma;
  - Definir personas com base em interesses;
  - Criar perfis com base no monitoramento de atividades.
- Categorização de inventário:
  - Inventário do grupo por atividade de vendas;
  - Inventário do grupo por métricas de produção;
- Classificando as medições do sensor:
  - Detectar tipos de atividade em sensores de movimento;
  - Agrupar imagens;
  - Áudio separado;
  - Identificar grupos no monitoramento de integridade;
- Detectando bots ou anomalias:
  - Separe os grupos de atividades válidos dos bots;
  - Agrupar atividade válida para limpar a detecção de outliers.

Além disso, o monitoramento se um ponto de dados rastreado alterna entre grupos ao longo do tempo pode ser usado para detectar alterações significativas nos dados.

3.2. Algoritmo

O K-means usa refinamento iterativo para produzir um resultado final. As entradas do algoritmo são o número de clusters K e o conjunto de dados. O conjunto de dados é uma coleção de recursos para cada ponto de dados. Os algoritmos começa com estimativas iniciais para os k centroides, que podem ser geradas aleatoriamente ou aleatoriamente selecionados a partir do conjunto de dados. O algoritmo então itera entre duas etapas.

3.2.1. 1ª Etapa de avaliação dos dados

Cada centróide define um dos clusters. Nesta etapa, cada ponto de dados é atribuído ao seu centróide mais próximo, com base na distância euclidiana ao quadrado. Mais formalmente, se c_i é a coleção de centróides no conjunto *C*, então cada ponto de dados *x* é atribuído a um cluster com base.

$$\underset{c_i \in C}{\arg\min}\ dist(c_i,\ x)^2$$

Onde *dist*( ) é a distância euclidiana padrão (L_2). Deixa o conjunto de atribuições de pontos de dados para cada i^{th} centróide aglomerado ser S_i.

3.2.2. 2ª Etapa de atualização do centroíde

Nesta etapa, os centroides são recalculados. Isso é feito tomando a média de todos os pontos de dados atribuídos ao cluster desse centróide.

$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i$$

O algoritmo itera entre as etapas um e dois até que um critério de parada seja atendido, ou seja, nenhum ponto de dados altera os clusters, a soma das distâncias é minimizada ou algum número máximo de iterações é atingido.

Este algoritmo é garantido para convergir para um resultado. O resultado pode ser um ótimo local, não necessariamente o melhor resultado possível, o que significa que a avaliação de mais de uma execução do algoritmo com centróides de partida randomizados pode dar um resultado melhor.

3.3. Escolhendo o K

O algoritmo descrito acima encontra os clusters e rótulos de conjunto de dados para um K pré-escolhido em particular. Para encontrar o número de clusters nos dados, precisa executar o algoritmo para um intervalo de valores K e comparar os resultados. Em geral, não há um método para determinar o valor exato de K, mas uma estimativa precisa pode ser obtida usando as seguintes técnicas.

Uma das métricas comumente usadas para comparar os resultados em diferentes valores de K é a distância média entre os pontos de dados e seu centróide de cluster. Desde

o aumento do número de clusters sempre reduzir a distância de pontos de dados, aumentando a K irá sempre diminuir essa métrica, ao extremo de se chegar a zero quando K é o mesmo que o número de pontos de dados. Assim, essa métrica não pode ser usada como o único destino. Em vez disso, a média de distância para o baricentro como uma função de K é plotado e o "ponto de cotovelo", onde a taxa de diminuição muda bruscamente, pode ser usado para determinar aproximadamente K.

Existem várias outras técnicas para validação de K, incluindo validação cruzada, critérios de informação, o método de salto teórico de informação, o método de silhueta e o algoritmo de G-médias. Além disso, o controlo da distribuição de pontos de dados entre os grupos fornece informações sobre a forma como o algoritmo é dividir os dados para cada K.

4. Aplicação

5. Conclusão

O presente projeto teve por objetivo aplicar o algoritmo de K-means sobre o problema citado anteriormente. O maior propósito é ganhar conhecimento sobre a área de Inteligência Artificial e ter suporte para aplicar os seus conceitos. Contudo, deve ser realizado um estudo mais aprofundado para solucionar de forma eficaz o problema.

O número de ocorrências sobre cartões clonados vem aumentando cada vez mais, dessa forma, muitas empresas investem para mitigar esse tipo de ameaça que traz grandes prejuízos ao consumidor. Adotando a ideia sugerida no projeto, a vítima ficará informada sobre alguma anormalidade.

Apesar da complexidade de utilizar K-means, este é o melhor algoritmo para solucionar o problema abordado.