DETECÇÃO DE FRAUDES BANCARIAS UTILIZANDO FERRAMENTAS DE CLUSTERIZAÇÃO DE DADOS

Rafael Duarte Beltran [1]

Alessandro Bof De Oliveira [2]

Resumo:

As instituições financeiras tem utilizado extensivamente sistemas informatizados para a prestação de seus serviços. Dessa forma é crescente o número de soluções e produtos oferecidos por meio digital. A conveniência da utilização dessas ferramentas traz consigo a importante questão da segurança dos dados e das operações financeiras realizadas pelas instituições e por seus usuários. Dentre essas questões de segurança, nosso trabalho se concentra na detecção de fraudes de operações financeiras. Ao utilizar o sistema financeiro, por exemplo com a utilização de cartões de crédito para efetuar compras, o usuário gera um histórico de dados que podem ser utilizados para construir um padrão de comportamento típico. Quando a segurança do sistema é comprometida, como por exemplo a utilização do cartão de crédito por outra pessoa que não o usuário, temos a ocorrência de uma fraude. O objetivo de nosso projeto é detectar comportamentos não usuais em bases de dados financeiras que indiquem um comportamento fraudulento. Para a realização de nosso trabalho serão utilizadas a linguagem de programação python 3.6 e a biblioteca de método de aprendizagem de máquina scikit-learning. Os dados utilizados são provenientes de uma base de dados real com histórico de transações de cartões de crédito e anônimas. A base de dados é rotulada com 284.807 transações normais e 492 fraudes. A abordagem de nosso trabalho é a realização de clusterização dos dados, utilizando diversos métodos (k-médias, mean shift, etc) e a posterior combinação dos resultados através de um método de votação. Os resultados preliminares, com a utilização do método de clusterização k-means, utilizando 20 clusters e o método de inicialização kmeans++ foi de 52.64% com a utilização de um cluster como fraude e o restante como transações normais. Ao considerar a união de dois clusters como fraude (e portanto 18 como transações normais), chegamos a precisão de detecção de fraudes de 74.39%. Os resultados obtidos são encorajadores e como próximos passos estão a combinação de diferentes métodos de clusterização para melhorar a precisão da detecção de fraudes.

Palavras-chave: Fraudes, Transações Bancarias, Clustering

Modalidade de Participação: Iniciação Científica

DETECÇÃO DE FRAUDES BANCARIAS UTILIZANDO FERRAMENTAS DE CLUSTERIZAÇÃO DE DADOS

[1] Aluno de graduação. rbduarte96@gmail.com. Autor principal

[2] Docente. alessandrooliveira@unipampa.edu.br. Orientador

DETECÇÃO DE FRAUDES BANCARIAS UTILIZANDO FERRAMENTAS DE CLUSTERIZAÇÃO DE DADOS

As instituições financeiras tem utilizado extensivamente sistemas informatizados para a prestação de seus serviços. Dessa forma é crescente o número de soluções e produtos oferecidos por meio digital. A conveniência da utilização dessas ferramentas traz consigo a importante questão da segurança dos dados e das operações financeiras realizadas pelas instituições e por seus usuários. Dentre essas questões de segurança, nosso trabalho se concentra na detecção de fraudes de operações financeiras. Ao utilizar o sistema financeiro, por exemplo com a utilização de cartões de crédito para efetuar compras, o usuário gera um histórico de dados que podem ser utilizados para construir um padrão de comportamento típico. Quando a segurança do sistema é comprometida, como por exemplo a utilização do cartão de crédito por outra pessoa que não o usuário, temos a ocorrência de uma fraude.

O objetivo de nosso projeto é detectar comportamentos não usuais em bases de dados financeiras que indiquem um comportamento fraudulento. Para a realização de nosso trabalho serão utilizadas a linguagem de programação python 3.6 e a biblioteca de método de aprendizagem de máquina scikit-learning. Os dados utilizados são provenientes de uma base de dados real com histórico de transações de cartões de crédito e anônimas. A base de dados é rotulada com 284.807 transações normais e 492 fraudes. A abordagem de nosso trabalho é a realização de clusterização dos dados, utilizando diversos métodos (k-médias, mean shift, etc) e a posterior combinação dos resultados através de um método de votação.

Os resultados preliminares, com a utilização do método de clusterização k-means, utilizando 20 clusters e o método de inicialização kmeans++ foi de 52.64% com a utilização de um cluster como fraude e o restante como transações normais. Ao considerar a união de dois clusters como fraude (e portanto 18 como transações normais), chegamos a precisão de detecção de fraudes de 74.39%. Os resultados obtidos são encorajadores e como próximos passos estão a combinação de diferentes métodos de clusterização para melhorar a precisão da detecção de fraudes.