Prática - 12

Alunos:

- Pedro Henrique Faria Teixeira 11621BCC025
- João Daniel de Aquino Rufino 11621BCC033

ST-DBSCAN: Um algoritmo para agrupar dados espaciais-temporais

Sobre o trabalho:

Este trabalho apresenta um novo algoritmo de agrupamento baseado em densidade, ST-DBSCAN, que é baseado no DBSCAN. Propôs-se três extensões marginais ao DBSCAN relacionadas com a identificação de: ***(i) objetos centrais***, ***(ii) objetos de ruído*** e ***(iii) clusters*** adjacentes. Em contraste com os algoritmos de agrupamento baseados em densidade existentes, este algoritmo tem a capacidade de descobrir agrupamentos de acordo com os valores não espaciais, espaciais e temporais dos objetos. Neste trabalho, também foi apresentado um sistema de armazenamento de dados espaço-temporal projetado para armazenar e agrupar uma ampla gama de dados espaço-temporais. Mostra-se uma implementação do algoritmo usando uma data warehouse e mostra os resultados da mineração de dados.

Motivação:

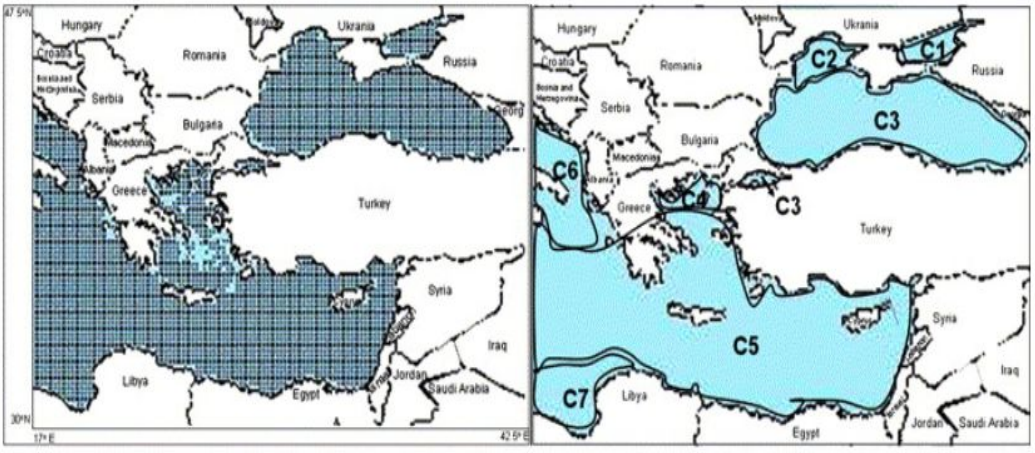
A motivação para o desenvolvimento do ST-DBSCAN foi fazer um melhoramento do DBSCAN dado três direções importantes que ele não consegue por si só resolver. Primeiro, ao contrário dos algoritmos de agrupamento baseados em densidade existentes, o algoritmo desenvolvido neste artigo pode agrupar dados espaciais-temporais de acordo com seus atributos não espaciais, espaciais e temporais. Em segundo lugar, o DBSCAN não pode detectar alguns pontos de ruído quando existem clusters de densidades diferentes. O algoritmo ST-DBSCAN resolve esse problema atribuindo a cada cluster um fator de densidade. Terceiro, os valores dos objetos de fronteira em um cluster podem ser muito diferentes dos valores de objetos de fronteira no lado oposto, se os valores não espaciais de objetos vizinhos tiverem pequenas diferenças e os clusters forem adjacentes uns aos outros. O algoritmo proposto consegue resolver esse problema comparando o valor médio de um cluster com o novo valor futuro.

Por que o DBSCAN foi usado neste trabalho:

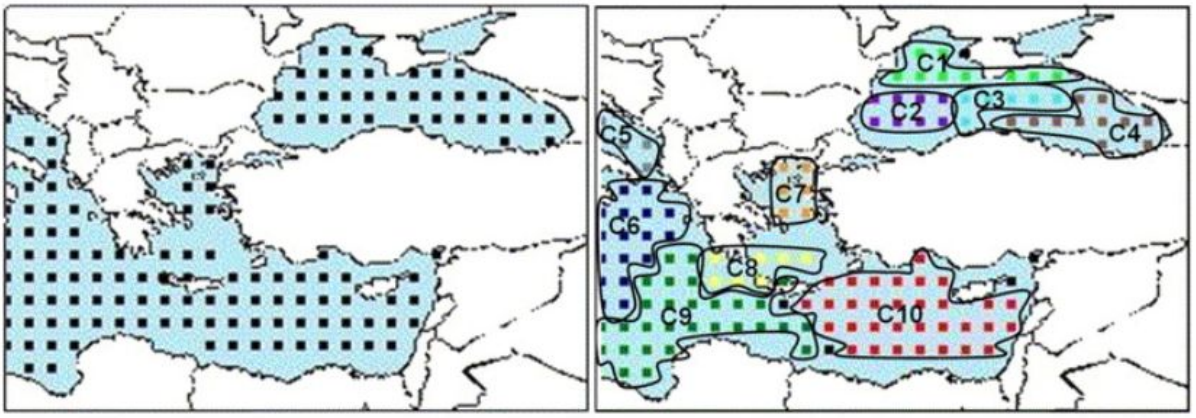
Neste estudo, escolheram o algoritmo DBSCAN, porque ele tem a capacidade de descobrir clusters de forma arbitrária, como linear, côncava, oval, entre outras. Além disso, ao contrário de alguns algoritmos de clustering, ele não requer a predeterminação do número de clusters. Além de ter a habilidade de processar bancos de dados muito grandes.

Resultados alcançados pelo trabalho:

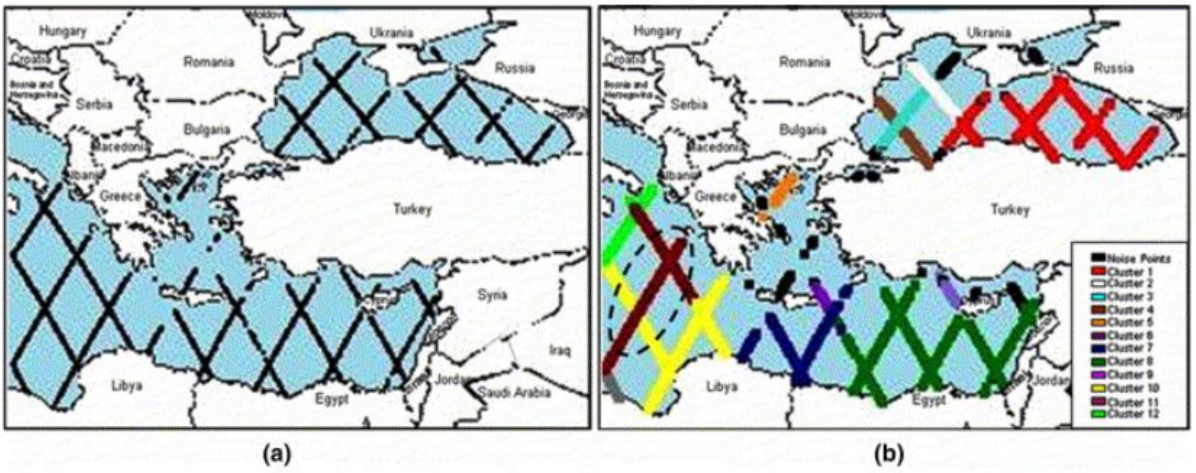
A avaliação dos resultados da mineração de dados foi gerada em cima de uma base de dados geográficos de regiões diversas no mundo. O banco de dados de exemplo contém registros semanais de temperatura diurna e noturna, que foram medidos em 5340 estações nos anos entre 2001 e 2004. Em outras palavras, os valores de temperatura da superfície do mar armazenados no banco de dados foram coletados em 5340 estações que são mostradas na abaixo em prontos pretos. A distribuição espacial da temperatura nas águas superficiais (30-47,5 ° N e 17-42,5 ° E) é mostrada na figura abaixo (b). Cada cluster possui pontos de dados com características semelhantes de temperatura da superfície do mar. O cluster número 1 faz fronteira com a Ucrânia e a Rússia. Esta região é a área mais fria. O cluster número 2 no norte da Ucrânia é a segunda área mais fria. As temperaturas da água do mar em outras partes do Mar Negro são semelhantes às do Mar de Mármara. O cluster número 4 cobre o norte do Mar Egeu. O cluster número 5 forma um grande cluster único. Os valores de temperatura das estações no Cluster 6 também têm características semelhantes. O cluster número 7 é a região mais quente, porque é a área mais próxima do equador. Nas estações de inverno, os clusters C5 e C7 podem ser marcados como um cluster, porque eles não podem ser distinguidos muito bem. No verão, o cluster C6 fica um pouco pequeno. Muitos fatores podem afetar essa distribuição da temperatura da água do mar. A temperatura varia tanto em termos de atitude quanto de profundidade em resposta às mudanças nas interações ar-mar. Fluxos de calor, evaporação, fluxo do rio, movimento da água e chuva influenciam a distribuição da temperatura da água do mar.



O satélite Topex / Poseidon fornece dados residuais da altura da superfície do mar como uma grade bidimensional separada por um grau em latitude e longitude. Portanto, os valores SSHR armazenados no banco de dados estão disponíveis em 134 estações, que são mostradas na figura abaixo como pontos pretos (a). Os clusters obtidos pelo uso da tabela Sea Surface Height são mostrados na figura abaixo do lado direito (b). Cada cluster tem pontos de dados com valores residuais de altura da superfície do mar semelhantes. Os clusters nomeados por C1, C2, C3 e C4 estão localizados no Mar Negro. O cluster nomeado por C7 está localizado no Mar Egeu. O restante dos clusters está localizado no Mar Mediterrâneo. Muitos fatores contribuem para mudanças na altura da superfície do mar, incluindo redemoinhos do mar, temperatura da parte superior da água do mar, marés, correntes marítimas e gravidade.



Valores significativos de altura de onda armazenados no banco de dados foram coletados em 1707 estações, que são mostradas na figura abaixo (a) como pontos pretos. A figura (b) mostra um exemplo de resultado de agrupamento obtido pelo uso do conjunto de dados medido em 28 de janeiro de 2001. Cada agrupamento tem pontos de dados com valores de altura de onda significativos semelhantes. Por exemplo, enquanto a região a leste da Ilha de Creta tem valores de altura de onda de aproximadamente 0,5 m, a região (cluster 11) que é circulada em linhas tracejadas tem valores de altura de onda de aproximadamente 3,6 m. A região que está circulada em linhas tracejadas tem os valores máximos de altura de onda.



Pontos positivos do uso do DBSCAN:

- O DBSCAN consegue fazer a clusterização da base de dados sem precisar de um número pré definido de clusters
- Tem a habilidade de descobrir clusters de forma arbitrária, como linear, côncava, oval, etc.
- Consegue processar base de dados muito grandes.

Pontos negativos do uso do DBSCAN:

- O DBSCAN puro não conseguiria resolver o problema para o qual buscavam solução, então foi preciso gastar tempo para desenvolver novas funcionalidades e incrementá-las.
- O DBSCAN não agrupa dados espaciais-temporais de acordo com seus atributos não espaciais, espaciais e temporais.
- O DBSCAN não pode detectar alguns pontos de ruído quando existem clusters de densidades diferentes, para isso foi preciso desenvolver novas funções.
- Os valores dos objetos de fronteira em um cluster do DBSCAN podem ser muito diferentes dos valores de objetos de fronteira no lado oposto, se os valores não espaciais de objetos vizinhos tiverem pequenas diferenças e os clusters forem adjacentes uns aos outros, com isso foi preciso desenvolver uma técnica que compara o valor médio de um cluster com o novo valor futuro para resolver.

Bibliografia:

ST-DBSCAN: An algorithm for clustering spatial–temporal data