Análise e visualização de dados utilizando redes neurais artificiais auto-organizáveis

Adriano Lima de Sá

Faculdade de Computação
Universidade Federal de Uberlândia

20 de junho de 2014

Objetivo

- Estudar a aplicabilidade da rede neural auto-organizável para extração de conhecimentos
- Aplicação em dados reais: Exame Nacional Ensino Médio (ENEM)
- Aprimorar e exercitar conhecimentos adquiridos na graduação

Contexto e Motivação de Self-Organizing Maps (SOM)

- Grandes volumes de dados não-lineares
- Alta complexidade para análise dos mesmos
- Aprendizado de Máquina e Inteligência Artificial
- Métodos para obtenção e distinção de tendências a partir das similaridades dos dados

Atualmente

- Análise em um grande volume de dados não-lineares com alta dimensionalidade é difícil e demorada
- Relações entre dados podem estar ocultas devido à complexidade e organização dos mesmos
- Desenvolvimento de métodos de análise computacional em Aprendizado de Máquina e Inteligência Artificial a fim de facilitar a compreensão de fenômenos complexos

Motivação e Uso

Self-Organizing Maps

- Treinamento competitivo e não supervisionado
- Mapeamento de um espaço L-dimensional em um conjunto de neurônios (vetores)
- Preservação da topologia dos dados
- Viabilização de análise de grande conjuntos de dados não-lineares e de alta dimensionalidade
- Usos
  - o SOM é utilizado em métodos de análise exploratória de fenômenos
  - Por exemplo, para fazer o estudo do Índice de Desenvolvimento Humano (IDH) calculado pela Organização das Nações Unidas é comum utilizar o SOM
  - Análises geo-demográficas
  - Mineração de Dados e Recuperação de Informação

Ideia do SOM

Ideia Geral

- Dados são mapeados de L para D dimensões
- Busca-se manter relações topológicas e métricas
- Uso de processos competitivos e não supervisionados
- Ocorre uma competição entre os neurônios para se tornarem o neurônio Best-Matching Unit (BMU)
- BMU é o neurônio com a menor distância euclidiana em relação a determinado padrão
- Não supervisionado devido à inexistência da entrada de dados externos providos por especialistas

Passo a passo

1. Atribuição de pequenos valores aleatórios a rede;
2. Apresentação de um vetor padrão;
3. Cálculo da similaridade entre cada neurônio e o padrão de entrada (usualmente euclidiana);
4. Seleção do neurônio com menor distância em relação ao vetor padrão (BMU);
5. Atualização da rede de acordo com neurônio vencedor, conforme a equação: $w_k(t+1) = w_k(t) + h_{ci}[x(t) - w_k(t)]$

A equação usa $w_k(t+1) \in R^n$ para representar os n valores do neurônio a ser atualizados no passo $t+1$ de execução do algoritmo; h_{ci} para representar uma função de distância no mapa de neurônios entre o neurônio vencedor e o neurônio sendo atualizado; e $x(t)$ o vetor de dados de entrada com n dimensões descrevendo suas características.

Mapeamento de Sammon (1969)

Abordagem

- Redução de dimensionalidade de conjunto de vetores para facilitar a visualização dos dados
- Objetivo: preservar a topologia (estrutura de distâncias entre pontos) dos dados quando possível
- Algoritmo que mapeia um espaço de alta dimensionalidade para 2 ou 3 dimensões
- Método não linear baseado em otimização com gradiente descendente do erro de mapeamento em menor dimensionalidade

Mapeamento de Sammon

Função de erro de mapeamento

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - dij)^2}{d^*_{ij}}$$

- O algoritmo de Sammon visa minimizar essa função de erro
- Função conhecida como estresse de Sammon ou erro de Sammon
- i, j são objetos no espaço original e $d*_{ij}$ representa sua distância
- d_{ij} representa a distância das projeções

Objetivo: mistura da SOM e do mapeamento de Sammon

Deseja-se usar o mapeamento de Sammon junto com o SOM

- Mapeamento de Sammon: redução de dimensionalidade preservando topologia
- Mapeamento de Sammon: exige obtenção de matriz de distâncias entre elementos: $O(n^2)$ de espaço
- SOM: redução do conjunto de dados usando neurônios para representar grupos de dados similares
- SOM: não reduz a dimensionalidade dos dados

Objetivo

- Obter método híbrido para reduzir dados a grupos similares em baixa dimensionalidade preservando topologia original
- Evitar $O(n^2)$ do mapeamento de Sammon e executar 2 algoritmos de uma só vez

Mistura da SOM e do mapeamento de Sammon

Metodologia (Como pretendemos fazer isso?)

- Definir um algoritmo de junção dos algoritmos propostos (Sammon e SOM) a fim de agilizar o processo de apredizagem de máquina (SOM).
- Executar, a cada etapa de apredizagem, a redução da dimensionalidade com o Sammon, para se obter o BMU da rede SOM, em vez da distância euclidiana usual.
- Na escolha do BMU deve-se avaliar qual neurônio juntamente com sua vizinhaça mais se aproxima de determinado padrão (vetor dados originais), ou seja, possuem a menor taxa de modificação das distâncias em menor dimensionalidade

Mistura da SOM e do mapeamento de Sammon

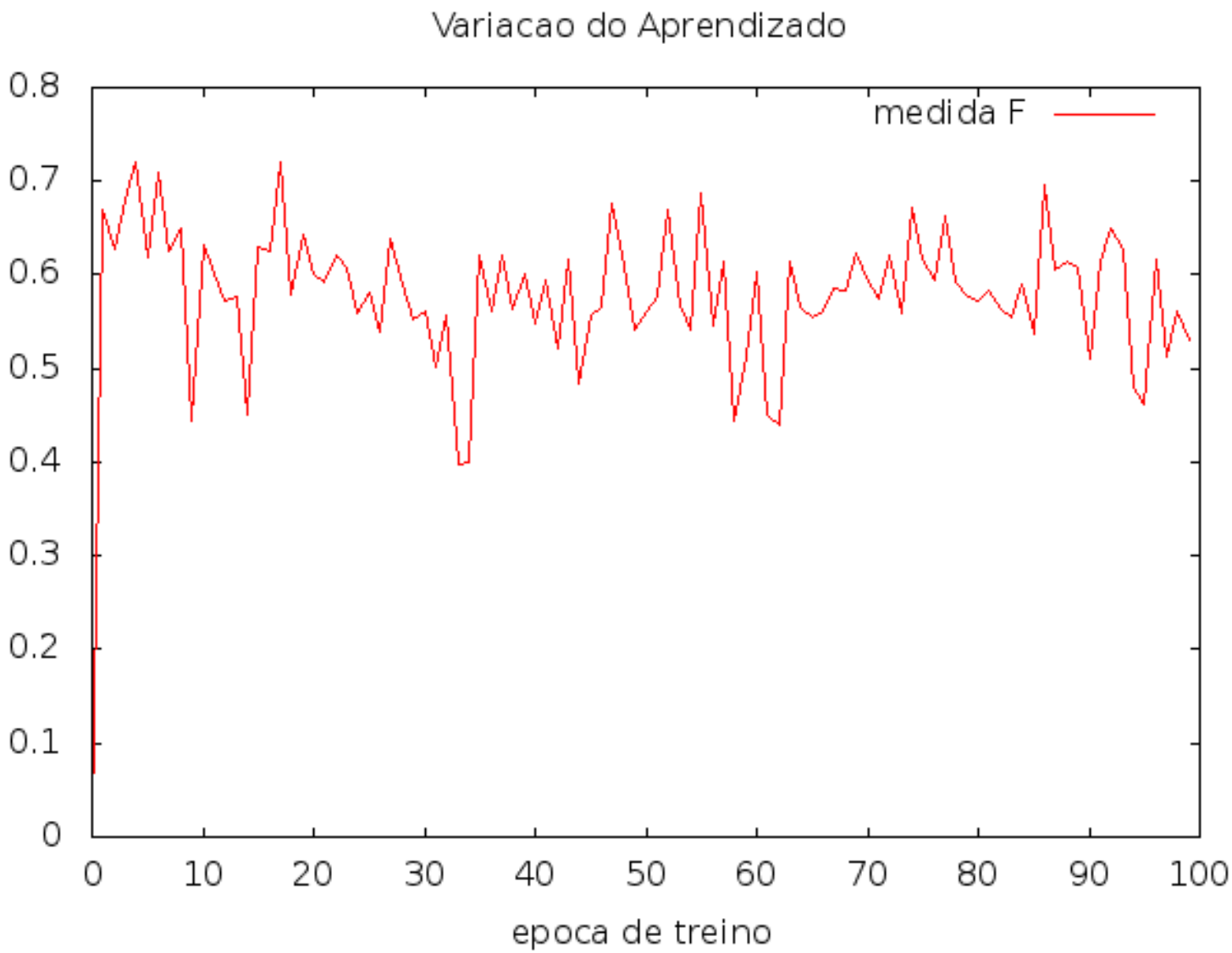
- Para avaliar a qualidade de aprendizagem ao longo do processo do SOM utilizamos uma função que calcula a medida F (medida externa de avaliação)

Medida F

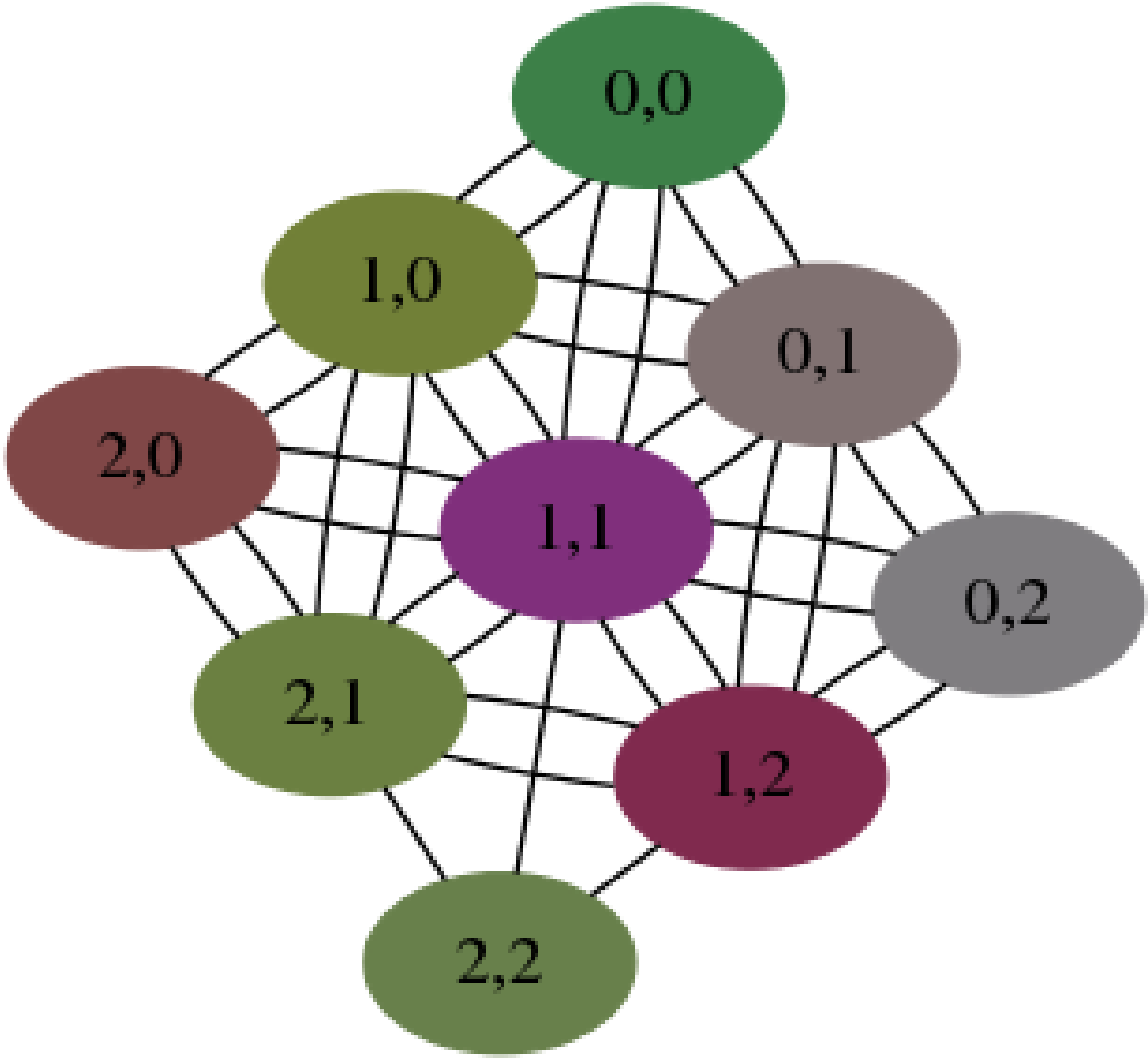
$$F = 2PR/(P + R)$$

  - Sendo que
  - P representa a taxa de precisão
  - R representa a taxa de recuperação

Resultado do SOM



Exemplo de resultado



Exemplo de resultado

- Para gerar o grafo utilizou-se sistema de cores HSV
- HSV abreviação de hue(matiz), saturation(saturação) e value(valor)
- Esse sistema de cores define a cor conforme os três parâmatros
- Matiz (tonalidade) define propriamente qual tom de cor a ser utilizada, abrangendo todas as cores do espectro, desde o vermelho até o violeta, mais o magenta
- Saturação define o quanto de "tinta" da cor definida será utilizado, quanto menor esse valor, mais cinza aparecerá a imagem, quanto maior mais "viva" é a imagem
- Valor (brilho) define o brilho da cor – quanto maior, mais brilhante estará a cor

Etapas do trabalho

- Estudos introdutórios;
- Mapa de Kohonen;
- Sammon;
- Junção SOM + Sammon;
- Implementação;
- Experimentos;
- Escrita de Artigo Científico.

Referências

- Marcelo Keese Albertini and Rodrigo Fernandes de Mello. A self-organizing neural network for detecting novelties. In Proceedings of the 2007 ACM symposium on Applied computing, SAC '07, pages 462-466, New York, NY,USA,2007.ACM.
- Sammon John W.Jr. A Nonlinear Mapping for Data Structure Analysis.IEEE Transactions on Computers, vol C-18, pp.401-409, 1969
- http://en.wikipedia.org/wiki/Sammon_mapping
- http://pt.wikipedia.org/wiki/HSV