Trabalho Final: Reconhecimento de Padrões

Rúbia Reis Guerra

***Abstract*— Matrizes de proximidade obtidas por meio de clustering FCM podem ser utilizada como um kernel para a construcão de SVMs. Este trabalho avalia a utilizacão desta matriz como um kernel de SVM por meio de comparações empíricas.**

I. INTRODUÇÃO

Reconhecimento de Padrões (RP) é um ramo de aprendizado de máquina que se concentra em reconhecimento de padrões e regularidades em dados. Os sistemas de reconhecimento de padrões são, em muitos casos, treinados a partir de dados de "treinamento"rotulados (aprendizagem supervisionada), mas quando não há dados rotulados disponíveis, outros algoritmos podem ser usados para descobrir padrões desconhecidos anteriormente (aprendizado não supervisionado). Um tipo de problema possível de ser resolvido por técnicas de RP são problemas classificação, que se aplicam para muitos objetos e sistemas do mundo real. Neste trabalho será explorada solução de problemas de classificação utilizando máquinas de vetores de suporte (SVMs), cujo kernel será obtido a partir da criação de uma matriz de proximidade por meio de Fuzzy Clustering.

II. RESUMO TEÓRICO

A. Kernel

Em aprendizagem em máquina, os métodos de kernel são uma classe de algoritmos para a análise de padrões, cujo membro mais conhecido é a máquina de vetores de suporte (SVM). A tarefa geral de análise de padrões é encontrar e estudar tipos gerais de relações (por exemplo, agrupamentos, rankings, componentes principais, correlações, classificações) em conjuntos de dados. Para muitos algoritmos que resolvem essas tarefas, os dados brutos em representação devem ser explicitamente transformados em representações de vetores de características através de um mapa de características especificado pelo usuário. Em contraste, os métodos do kernel requerem apenas um kernel especificado pelo usuário, ou seja, uma função de similaridade em pares de pontos de dados brutos em representação.

Os métodos do Kernel devem seu nome ao uso de funções do kernel, o que lhes permite operar em um espaço de recursos implícitos de alta dimensão sem nunca calcular as coordenadas dos dados nesse espaço, mas simplesmente ao computar os produtos internos entre as imagens de todos os pares de dados no espaço de características. Esta operação é muitas vezes computacionalmente mais barata do que a computação explícita das coordenadas.

B. Máquinas de Vetores de Suporte (SVMs)

As Máquinas de Vetores de Suporte (SVMs) consistem em um algoritmo de aprendizagem de máquina supervisionado que pode ser usado para problemas de classificação ou de regressão. No entanto, é usado principalmente em problemas de classificação. Neste algoritmo, plota-se cada item de dados como um ponto no espaço n-dimensional (onde n é o número de dimensões do problema), sendo o valor de cada atributo o valor de uma determinada coordenada. Então, é realizada a classificação, encontrando o hiper-plano ótimo que diferencia as duas classes (figura 2).

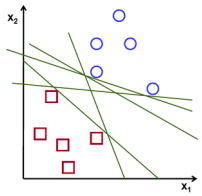

Fig. 1. Possíveis limiares de separação das classes

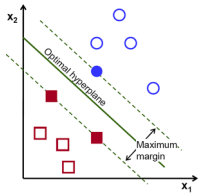

Fig. 2. Hiperplano resultante por SVM

C. Fuzzy Clustering (FCM)

Fuzzy c-means (FCM) é um método de agrupamento que permite que uma parte de dados pertença a dois ou mais clusters. Este método foi desenvolvido por Dunn em 1973 (e melhorado por Bezdek em 1981), e é freqüentemente usado em reconhecimento de padrões. Baseia-se na minimização da seguinte função objetivo:

$$J_m = \sum_{j=1}^{N}\sum_{i=1}^{N} u_{ij}^{m}\|x_i - c_j\|^2$$

onde m é qualquer número real maior do que 1, u_{ij} é o grau de associação de x_i no cluster j, x_i é o i-ésimo dado medido em dimensões d, c_j é o centro de dimensões d do cluster e $\|*\|$ é qualquer norma que expressa a semelhança entre qualquer dado medido e o centro.

III. COMPARAÇÃO EMPÍRICA

A solução proposta consistiu na utilização de uma matriz de proximidade obtida por meio de Fuzzy c-means como kernel de um classificador SVM. Para fins de comparação, o novo classificador foi testado nas seguintes bases de dados:

TABLE I
BASE DE DADOS UTILIZADAS

Nome	Origem	Atributos	Exemplos
BreastCancer	*mlbench*	11	699
Iris	Nativa (R)	5	150
Wine	UCI	13	178
Climate Model SC	UCI	18	540
Connectionist Bench	UCI	10	528

TABLE II
RESULTADOS

Base	Média de Acertos (%)
BreastCancer	96.12
Iris	80.00
Wine	87.50
Climate Model SC	91.41
Connectionist Bench	98.66

A partir dos resultados, observou-se uma alta taxa acertos para problemas em que as classes encontram-se espacialmente separadas. Nessas situações, observou-se que o clustering por FCM foi condizente com a classe a qual cada grupo de dados pertence. No caso do dataset Iris, como as classes 1 e 2 (vermelhor e azul no plot da figura 3) foram unidas e comparadas à classe 3, infere-se que a superposição das classes originais 2 e 3 influenciou na acurácia da classificação.

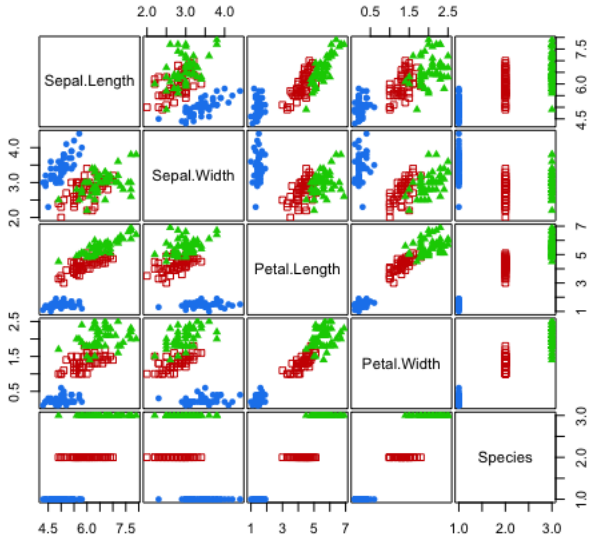

Fig. 3. Plot: Dataset Iris

REFERENCES

[1] A. Braga, A. Carvalho and T. Ludermir, Redes neurais artificiais. Rio de Janeiro: LTC Editora, 2007.
[2] R. Duda, D. Stork and P. Hart, Pattern classification and scene analysis. New York: Wiley, 2000.