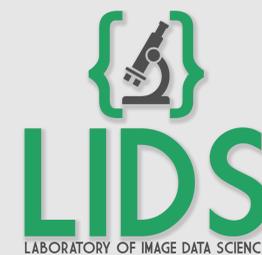


Anotação de Dados guiada por Projeção de Características

Autora: Bárbara C. Benato, barbarabenato@gmail.com
Orientador: Prof. Dr. Alexandre X. Falcão, afalcao@ic.unicamp.br



<Capes, Fapesp (Proc. #2016/25776-0)>

Palavras-chave: anotação de dados, analítica visual de dados, aprendizado semi-supervisionado.

Introdução

Muitos dados não-supervisionados? Poucos dados supervisionados? Baixa acurácia de classificação? Como explorar a capacidade do ser humano em *abstrair informações* e *identificar padrões* para reduzir o esforço do usuário e melhorar o projeto do classificador?

Objetivos

- Espaço de características com boa separabilidade das classes (muitos dados não-supervisionados);
- Propagar rótulos das amostras supervisionadas para não-supervisionadas com pouco esforço;
- Minimizar os erros de propagação de rótulos: alta acurácia na classificação de novas amostras.

Métodos

Propagação de rótulos semi-automática (SALP):
Autoencoder convolucional + t-SNE + OPF-Semi + scatterplot e usuário = *dados rotulados*

Experimentos e Resultados

- *Bases de dados:* 3% sup, 67% unsup, 30% teste MNIST (5.000) e Parasitos intestinais humanos: Larvas (3514), Ovos (5112), Protozoários (9568).
- *Validação:* Sem propagação, automática (nD, 2D), manual (ILP) e semi-automática (SALP).

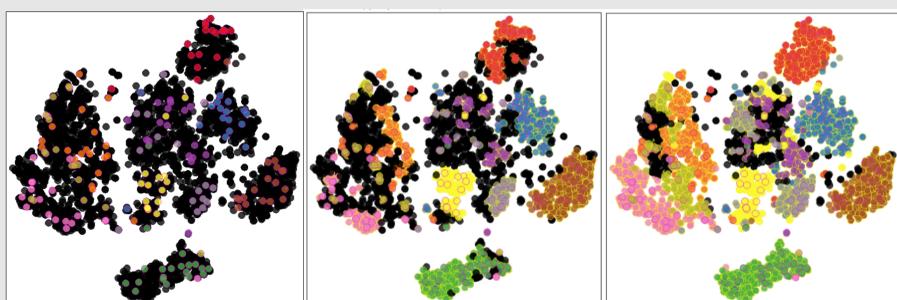


Fig. 1: MNIST: sup+unsup, OPF-Semi, usuário.



Fig.2: Parasitos e impurezas.

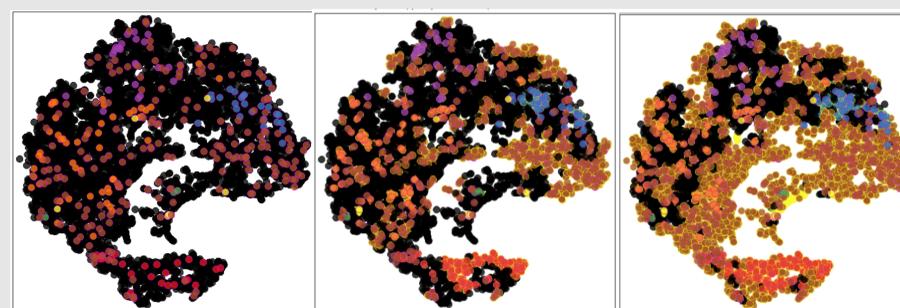


Fig.3: Larvas: sup+unsup, OPF-Semi, usuário.

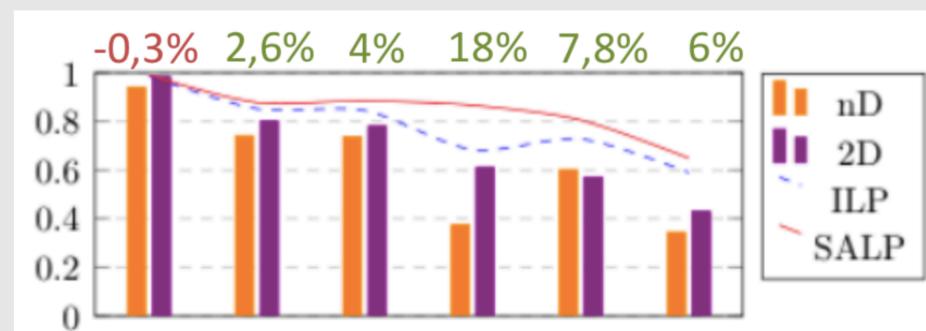


Fig.4: Resultados SVM (kappa) com OPF-Semi nD e 2D (barras) (O, P, M, OI, LI, PI).

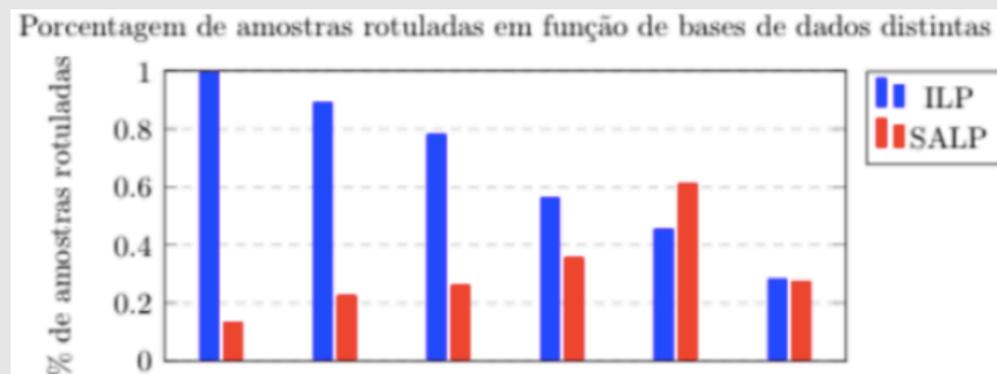


Fig.5: Esforço do usuário (O, LI, P, M, OI, PI).

Conclusões

SALP **combina** usuário e métodos automáticos. Propagação de rótulos automática no 2d obteve maiores acurácias do que nD. SALP obteve *melhores resultados* de classificação, comparado a propagação totalmente automática ou manual, além de **reduzir esforço** do usuário.