Universidade do Minho

"Improve Learning from Crowds via Generative Augmentation"

Mineração de Dados

- Filipa Pereira - PG46978
- Luísa Carneiro - PG46983
- Luís Pinto - PG47428
- Rita Peixoto - PG46988

---

1. Contextualização

O artigo "Improve Learning from Crowds via Generative Augmentation" pretende implementar *Generative Augmentation* de forma a resolver o problema de *sparsity* relacionado com o *crowdsourcing*.

O *crowdsourcing* é uma metodologia de obtenção de anotações para uma coleção de dados através de um conjunto de entidades, que tem como objetivo reduzir os custos de obtenção dessas anotações. Este objetivo implica que o número de entidades (anotadores) envolvidas nas anotações não pode ser muito grande, trazendo por isso problemas de *sparsity* que estão relacionados com a falta de anotações para cada instância da coleção de dados.

Para resolver este problema, o artigo tenta implementar uma *framework* **CrowdInG** que tem o objetivo de gerar mais anotações através da modelação da distribuição das anotações nas instâncias ou nos anotadores. Para isso o artigo propõem dois critérios: as anotações geradas devem seguir a distribuição das anotações autenticas e as anotações geradas devem ter informação comum com as *ground-truth labels*, isto é, as *labels* reais de cada instância.

2. Questões da Pesquisa

Este artigo científico pretende responder a diversas questões das quais se destacam:

1. De que modo se poderá realizar *crowdsourcing* sem que este implique altos custos?
2. Sendo os dados etiquetados (*label data*) escassos, como é que se pode efetuar a sua obtenção, de forma eficiente e pouco custosa, para serem utilizados em *Machine Learning* supervisionado?
3. Como é que se pode lidar com o problema de *sparsity* recorrendo a métodos de *Data Augmentation* de forma a gerar as anotações em falta?
4. De forma a obter um modelo classificador preciso, quais são as anotações que mais impactam o processo de treino desse modelo?
5. Quanto trabalho humano poderá ser economizado através da solução de aumento de dados proposta?

3. Enquadramento do Caso de Estudo

O presente documento insere-se no contexto da UC de Mineração de Dados, e trata de efetuar um resumo e apreciação do artigo científico lido e analisado detalhadamente, de forma a perceber como

podem ser gerados dados através de *crowdsourcing* e tem como objetivo fornecer uma perspetiva mais ampla sobre conceitos da área de *data mining*.

Neste artigo é tratado o problema da aprendizagem por *crowds*, que teve por base outras investigações pertinentes, como por exemplo: estimar a eficiência de aplicar um classificador de regressão logística a dados obtidos por *crowdsourcing* recorrendo à modelação de anotações. Esta solução foi mais tarde refinada com classificadores de redes neuronais para substituir a regressão logística e a forma de modelar as anotações. No entanto, estas metodologias apenas são aplicadas às anotações observadas, continuando a haver escassez nos dados.

Para este caso em concreto, as pesquisas relacionadas com redes generativas (GANs) foram as que causaram mais impacto, tendo-se focado no estudo de *semi supervised GANs* para gerar novas *labels* para os dados, e também na aplicação de GANs para lidar com escassez de dados através do aumento dos mesmos. Além disso, também é relevante destacar o contributo com uma GAN *framework* que permite visualizar graficamente a aprendizagem do modelo.

Todas estas pesquisas oferecem conceitos chaves que foram da *framework* em análise.

4. Metodologia

O caso de estudo atual foca-se num relatório **experimental**, em que são criadas e testadas ferramentas para tornar o processo de *crowdsourcing* através de anotações mais eficiente. Trata-se de um estudo **quantitativo**, recorrendo à linguagem matemática para descrever as experiências e os resultados obtidos.

A metodologia adotada para a recolha de dados baseou-se, primeiramente, em usar **crowdsourcing** de modo a obter os dados e anotações iniciais. De seguida, pretende-se aumentar a quantidade de anotações através de **GANs** (Generative Adversarial Networks), ou mais concretamente **InfoGANs** (Information Maximizing Generative Adversarial Networks). Estes conceitos serviram de base para a criação da *framework* dos autores - o **CrowdInG** (Crowdsourced data through Informative Generative augmentation).

Posteriormente, esta *framework* foi testada tendo em conta 3 *datasets* do mundo real (LabelMe, Music e CIFAR-10H) para colocar à prova a sua eficácia, quer através de testes de performance com diferentes *baselines*, quer por estudos de ablação.

5. Sobre o Caso de Estudo

Tal como se referiu anteriormente, o problema descrito neste artigo científico consiste no facto de existir uma clara escassez de dados etiquetados, o que compromete a qualidade dos modelos de *Machine Learning*, além de que a obtenção deste tipo de dados, através de *crowdsourcing* revela-se custosa.

Deste modo, a solução implementada, CrowdIng, começa por prever uma anotação de um *input* através do *classifier* do *Generative Module*. Esta anotação vai ser utilizada pelo o *Generator* para assim gerar a distribuição das anotações do input recebido. De seguida esta distribuição vai ser utilizada pelo *Descriminative Module* que vai determinar se os valores da distribuição são próximos dos reais. Estes dados são depois utilizados na próxima iteração do *CrowdIng* para gerar anotações cada vez mais próximas das *ground-truth labels*.

Durante o processo de desenvolvimento da solução, foram enfrentados alguns desafios. Primeiramente, foi necessário ter em conta que o processo de treino da *framework CrowdInG* não é trivial. Visto que o número de instâncias não observadas (sem anotadores associados) é bastante superior ao número de instâncias observadas, se forem usadas apenas anotações sintéticas no treino do modelo discriminativo da GAN, esta irá classificar todas as anotações criadas como geradas e não reais. Uma solução que o artigo propõe para resolver este problema consiste em utilizar uma estratégia de seleção de anotações baseada na entropia. Além disto, outro desafio encontrado é o *model collapse*.

Este fenómeno ocorre uma vez que tanto o modelo classificador como o gerador (na GAN) poderão sofrer mudanças significativas de modo conseguir adaptar dados de treino complexos. Uma possível solução para este fenómeno seria utilizar uma estratégia de treino com dois passos principais: 1) utilizar instâncias com baixa entropia para atualizar o *generator*, 2) usar o *generator* atualizado nas restantes instâncias para atualizar o modelo classificador.

Por fim,neste artigo foram utilizados três *datasets* diferentes de forma a testar o *CrowdInG* implementado. A cada um deles foi aplicado um conjunto de modelos de *Data Augmentation* para além do *CrowdIng* de forma a comparar os resultados e foi-se iterativamente reduzindo o número de anotações de forma a analisar e comprar os resultados. O resultado desta análise sugere que o custo de geração de anotações reduz significativamente, mantendo a qualidade do classificador. Desta forma um grande quantidade de mão-de-obra pode ser reduzido.

6. Conclusão

A criação da *framework CrowdInG* , apresentada ao longo do artigo, permite resolver de forma eficiente os problemas derivados do *sparsity* sem aumentar o custo do *crowdsourcing*. Além disso, também é possível com a *CrowdInG* informar os anotadores de potencias confusões nas anotações realizadas, sendo que a *framework* só gera anotações cujo o modulo generativo tenha baixa confiança.