USING STACKED GENERALIZATION FOR ANOMALY DETECTION

Miguel Oliveira Sandim

Dissertação desenvolvida sob a orientação do *Prof. Carlos Soares*
e co-orientação do *Prof. Bernhard Pfahringer*

19 de setembro de 2017

---

1. Motivação

Análise de Dados tornou-se num tópico do mundo moderno, dado o grande número de possíveis aplicações em muitos diferentes domínios como marketing, investigação médica, visão por computador, análise de redes sociais, deteção de intrusões e deteção de fraude [1].

Deteção de Anomalias é uma área muito específica mas importante em Análise de Dados, dado o grande número de domínios em que pode ser aplicado [2]. De facto, o problema que motiva esta área é bastante comum e pode ser facilmente traduzido na seguinte questão: dada uma certa quantidade de dados, é possível detetar observações que se desviam do normal comportamento dos dados? Esta pergunta aparece, por exemplo, em áreas como deteção de fraude de cartões de crédito ou monitorização do estado de máquinas industriais.

A literatura relativa a técnicas de Deteção de Anomalias é muito extensa e diversa, com uma gama grande de técnicas que podem ter diferentes resultados (um *pontuação anómala* que indica o quão uma observação é uma anomalia, ou uma classificação - anómala ou normal), assim como diferentes pressupostos (técnicas baseadas em densidades têm pressupostos bastante diferentes das técnicas baseadas em agrupamentos). Esta heterogeneidade entre as várias técnicas de Deteção de Anomalias pode originar comportamentos diferentes para um mesmo conjunto de dados, o que torna a tarefa de escolher a(s) técnica(s) certa(s) para um domínio em específico bastante difícil e dependente dos dados.

2. Objetivos

Esta dissertação pretende aderessar este problema, usando várias técnicas de Deteção de Anomalias ao mesmo tempo e combinando os seus resultados num único. Esta é a ideia por detrás dos métodos de *Ensemble Learning*, que se baseiam na geração de um grupo de modelos (designado por *ensemble*) e combinar as suas previsões numa única. *Ensemble Learning* apresenta resultados empíricos de melhoria de performance em aplicações de aprendizagem computacional, como classificação, regressão, análise de séries temporais e sistemas de recomendação [3]. Mais especificamente, esta dissertação explora um método de *Sacked Generalization*, que consiste em adicionar um modelo extra que *aprende* a melhor maneira de combinar um grupo de modelos.

Assim, esta dissertação pretende responder à seguinte questão de investigação:

- Pode um método de *Stacked Generalization* melhorar a performance de técnicas de Deteção de Anomalias, mais especificamente a performance da melhor técnica para um determinado conjunto de dados?

3. Descrição da Dissertação

Este trabalho de investigação foi dividido em dois diferentes estudos de investigação. O primeiro estudo focou-se na análise da performance e diversidade das técnicas de Deteção de Anoamlias e teve os seguintes objetivos:

- Estudar a performance e diversidade dos diferentes tipos de técnicas de Deteção de Anomalias em vários conjuntos de dados bem conhecidos na literatura;
- Verificar se esta configuração experimental contêm modelos *exatos* e *diversos*.

O segundo focou-se na aplicação do método de *Stacked Generalization* às técnicas selecionadas e teve os seguintes objetivos:

- Determinar se combinando diferentes técnicas de Deteção de Anomalias com um modelo melhora a performance de cada uma das técnicas de Deteção de Anomalias usadas neste estudo;
- Caso o objetivo anterior se verifique, determinar em quanto a performance é melhorada.

Este trabalho de investigação incluiu várias técnicas do estado de arte de Deteção de Anomalias: *Classification and Regression Trees* (CART), *Support Vector Machine* (SVM), *Naive Bayes* (NB), *Random Forest* (RF), *Multilayer Perceptron* (MLP), *One-Class SVM*, *k-means*, *Density-based Spatial Clustering of Applications with Noise* (DBSCAN) e *Local Outlier Factor* (LOF).

Vários conjuntos de dados foram usados para avaliar a performance das técnicas de Deteção de Anomalias e dos métodos de *Ensemble Learning*. Estes

conjuntos de dados foram anterior usados na literatura de Deteção de Anomalias e recolhidos previamente por Campos et al. [4].

4. Conclusions

As principais conclusões desta dissertação podem ser brevemente sumariadas do seguinte modo:

- A maioria das técnicas de Deteção de Anomalias usadas neste estudo são *exatas* e *diversas* nos conjuntos de dados usados, permitindo assim as condições necessárias para que o método de *Stacking* melhore a performance da melhor técnica em cada conjunto de dados;
- A aplicação do método de *Stacking* garantiu valores mais altos na métrica F1 do que a melhor técnica de Deteção de Anomalias em mais de metade dos conjuntos de dados usados;
- Não existe uma indicação clara de que incluir técnicas de Deteção de Anomalias com diferentes modos de aprendizagem garanta valores mais altos de F1. Nos conjuntos de dados em que tal se verificou, a melhor combinação foi a de incluir técnicas de todos os modos de aprendizagem disponíveis.
- Não existe um meta-classificador que claramente ultrapasse os restantes na métrica F1 nos conjuntos de dados, por isso escolher um meta-classificador apropriado parece ser bastante dependente do conjunto de dados;
- Substituir o meta-classificador com um método simples de *Majority Voting* melhorou o valor da métrica F1 em ainda mais conjuntos de dados, com um aumento na melhoria média da métrica F1. Neste caso, *ensembles* com métodos de Deteção de Anomalias baseados em árvores (CART e Random Forest) foram os que obtiveram valores maiores da métrica F1 nos vários conjuntos de dados.

Referências

[1] Charu C Aggarwal. *Data Mining: The Textbook*. Springer Publishing Company, Incorporated, 2015. ISBN: 978-3-319-14142-8.

[2] Rupali Kandhari et al. "Anomaly detection". Em: *ACM Computing Surveys* 41.3 (2009), pp. 1–6. ISSN: 03600300. DOI: 10.1145/1541880.1541882.

[3] Charu C Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2017. ISBN: 1461463955, 9781461463955.

[4] Guilherme O. Campos et al. "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study". Em: *Data Mining and Knowledge Discovery* 30.4 (jul. de 2016), pp. 891–927. ISSN: 1573756X. DOI: 10.1007/s10618-015-0444-8. URL: http://link.springer.com/10.1007/s10618-015-0444-8.