Aprendizado de Máquina no Gerenciamento de Recursos da Computação em Nuvem: Uma Revisão Sistemática

Lucas Casagrande, Guilherme Koslovski, Marcelo Hounsell, Maurício Pillon

Programa de Pós-Graduação em Computação Aplicada (PPGCA)
Universidade do Estado de Santa Catarina – UDESC – Joinville – SC – Brasil

lucas.casagrande@edu.udesc.br, {guilherme.koslovski, marcelo.hounsell, mauricio.pillon}@udesc.br

Abstract. *Resource management in cloud computing is one of the most crucial tasks to guarantee the efficiency of its computational resources. Nevertheless, machine learning has become very popular due to its great capacity in pattern's recognition and decision-making tasks. The objective of this work is to search in the literature the current methods of machine learning applied to the resource management problem in cloud computing. Following a defined protocol, 80 papers from IEEE were found in which 25 were selected for further analysis. The results indicate a predominance of Artificial Neural Networks to workload's predictions and resource allocation.*

Resumo. *O gerenciamento de recursos em um ambiente em nuvem é um mecanismo vital para a eficiência de seus recursos computacionais. Não obstante, a área de aprendizagem de máquina tem se tornado cada vez mais popular devido a sua capacidade no reconhecimento de padrões e tomadas de decisões. O objetivo do presente trabalho é levantar na literatura o panorama atual da aprendizagem de máquina aplicada a problemática de gerenciamento de recursos em nuvem. Com base em uma estratégia de busca, 80 artigos da IEEE foram encontrados sendo 25 artigos selecionados para análise. Os resultados apontam uma predominância na utilização de Redes Neurais Artificiais na estimação da demanda e na alocação de recursos.*

1. Introdução

A Computação em Nuvem (CN) é um paradigma da computação definida como um modelo que permite o acesso sob demanda à um conjunto de recursos computacionais compartilhados entre múltiplos consumidores ao mesmo tempo [Mell e Grance 2011].

Em outras palavras, a CN pode ser considerada um tipo especial do paradigma da computação distribuída, que, através da virtualização, provê uma camada de abstração sob os seus recursos computacionais, que passam a ser vistos como um conjunto infinito de recursos, e possibilita o encapsulamento das aplicações de modo que elas possam ser dinamicamente configuradas de modo isolado e seguro [Foster et al. 2008].

Dentre os aspectos fundamentais da tecnologia de virtualização em ambientes em nuvem, encontra-se o mecanismo de gerenciamento de recursos, responsável pela coordenação e otimização dos recursos disponíveis em uma infraestrutura para CN. A partir deste mecanismo é possível desligar máquinas físicas que estejam em estado de espera, alocar e desalocar dinamicamente máquinas virtuais, migrando estas aplicações

para máquinas físicas sendo pouco utilizadas, e provisionar recursos de acordo com a demanda. Logo, a eficiência deste mecanismo impacta nos custos provenientes da operação de uma infraestrutura em nuvem sendo a sua otimização um objeto de estudo, e.g., alcançar maior economia de energia [Manvi e Shyam 2014; Younge et al. 2010].

A busca incessante pela maximização da eficiência de ambientes em nuvem, tanto por questões financeiras quanto ambientais, relacionado ao desperdício de energia proveniente da má utilização dos recursos, incentiva o desenvolvimento de métodos eficientes para a problemática do gerenciamento de recursos [Jennings e Stadler 2014]. Dada a relevância do tema, é possível encontrar na literatura trabalhos secundários que realizam um levantamento acerca do estado da arte, técnicas e objetivos de estudos nesta problemática.

Lopez-Pires e Baran (2015) analisaram 84 artigos quanto as suas abordagens de otimização, objetivos, técnicas e configuração dos ambientes em nuvem utilizados. Os autores concluíram que as técnicas mais utilizadas são baseadas em heurísticas sendo a minimização do consumo de energia o objetivo mais abordado. Demirci (2015) realizou uma revisão sobre métodos baseados em Aprendizado de Máquina (AM) focados na redução dos custos com energia em um ambiente em nuvem. Neste estudo foi possível perceber que redes neurais artificias e regressão linear são as técnicas mais utilizadas no contexto analisado. Boutaba et al. (2018) realizaram um extenso estudo sobre as aplicações de AM na área de redes apontando uma série de questões a serem consideradas em pesquisas futuras e no trabalho de Tyagi e Gupta (2018) foi realizado um levantamento sobre técnicas de escalonamento em sistemas distribuídos e paralelos sendo descrito um conjunto de técnicas e algoritmos com potencial para a solução do problema.

O presente artigo difere dos demais apresentando uma revisão acerca das produções científicas que utilizam AM na problemática do gerenciamento de recursos em CN. Dada a dinamicidade de um ambiente em nuvem, AM se torna interessante nesta problemática devido a sua capacidade no reconhecimento de padrões e na tomada de decisões. É uma área que contém um conjunto de métodos para treinamento de modelos que se divide conforme o seu tipo de aprendizagem, podendo ser de modo: supervisionado, onde os dados de treinamento contém a resposta esperada, não supervisionado, onde o modelo não tem conhecimento da resposta esperada, ou por reforço, quando o modelo aprende através da experimentação com um ambiente [Hashem et al. 2015; Murphy 2012].

O presente trabalho está organizado em quatro seções. Na segunda seção é descrito o protocolo que descreve os passos realizados na condução deste trabalho. Na terceira seção são apresentados os resultados obtidos e a devida classificação de cada trabalho analisado. Na última e quarta seção são discutidos os resultados e as conclusões.

2. Protocolo de Revisão

Para guiar o desenvolvimento do presente trabalho, um protocolo de revisão é definido com base nas recomendações de Budgen e Brereton (2006). Portanto, no decorrer desta seção são definidas as perguntas de pesquisa, os critérios de seleção, estratégias de busca, método para avaliação da qualidade dos artigos, e etapas de coleta e análise dos dados.

2.1. Perguntas de Pesquisa

Com a finalidade de identificar sob quais aspectos a problemática de gerenciamento de recursos tem sido abordado com técnicas de AM, a seguinte pergunta de pesquisa é realizada: *como tem sido aplicada as técnicas de aprendizado de máquina na problemática do gerenciamento de recursos da computação em nuvem?*

Com base na pergunta principal, duas outras perguntas de pesquisa são definidas para delimitar o escopo do trabalho:

- $\mathbf{QP}_1$: Em que parte do gerenciamento de recursos as técnicas de AM estão sendo utilizadas?
- $\mathbf{QP}_2$: Quais são as cargas de trabalho utilizadas na avaliação e treinamento dos modelos?

2.2. Critérios de Seleção

Para definir e limitar o escopo de artigos selecionados, critérios de inclusão e exclusão são definidos. Como critério de inclusão, foi levado em consideração artigos que: contenham definido claramente em seu título uma proposta para o gerenciamento de recursos em nuvem; são pesquisas primárias de periódicos ou conferências; e foram publicados a partir de 2013.

Como critério de exclusão, artigos que contenham as seguintes características são excluídos deste estudo: não estejam publicados em Inglês; não estejam diretamente relacionados a computação em nuvem; é um artigo duplicado de outro que já se encontra no conjunto; e não aplicam aprendizado de máquina (supervisionado, não supervisionado ou por reforço).

Para avaliar os critérios de seleção em cada artigo, um pesquisador foi responsável pela avaliação onde, caso não esteja claro a sua adequação, um segundo pesquisador foi consultado para avaliar o artigo. Caso haja conflito em ambas as avaliações, um terceiro pesquisador foi responsável pelo voto de desempate.

2.3. Estratégia de Busca

Para realizar a busca pela literatura, uma busca manual é realizada em Mecanismos de Busca Acadêmicos (MBA). Com base no trabalho de Lopez-Pires e Baran (2015), foi possível perceber uma predominância de artigos sendo publicados na IEEE. Portanto, o presente trabalho limita-se aos artigos encontrados na biblioteca digital IEEEXplore.

Com base no MBA mencionado, a seguinte frase de busca é realizada: *"("Document Title":"Resource*" AND "Cloud Computing" AND ("Machine Learning" OR "Deep Learning" OR "Neural Network" OR "Reinforcement Learning" OR "Supervised Learning" OR "Unsupervised Learning"))*".

Esta busca resultou em artigos que continham em seus títulos alguma indicação referente a problemática de gerenciamento de recursos e, em seu corpo, palavras-chaves que os limitam para o contexto da CN e AM, onde variações comuns encontradas no AM, como as redes neurais artificiais e a aprendizagem profunda, são utilizadas com a finalidade de ampliar os resultados.

2.4. Avaliação da Qualidade

Assegurar que somente trabalhos de qualidade sejam levados em consideração nas próximas etapas do estudo é importante para a análise da pergunta de pesquisa. Com esta finalidade, o mesmo processo de avaliação descrito na seção 2.2 é utilizado com base em uma tabela classificatória que leva em consideração se: foi realizado experimentos passíveis de replicação; o ambiente de teste está devidamente definido; os dados utilizados no treinamento estão definidos; e a arquitetura do modelo está devidamente descrita. Trabalhos que não contenham as características mencionadas são descartadas da análise.

2.5. Coleta de Dados

O conteúdo extraído de cada artigo consiste em: ano de publicação; o título do trabalho; técnicas de AM utilizadas; tipos de aprendizagem utilizados; a carga de trabalho utilizada; e o tipo de problema abordado. Caso o artigo possua mais de uma carga de trabalho, será considerado somente a de maior importância seguindo, da maior para a menor, a seguinte ordem: dados coletados de aplicações reais; gerados por aplicações de *benchmarking*; e dados simulados.

Em relação ao tipo de problema abordado nos trabalhos, esta característica é extraída de acordo com os objetivos do modelo sob um aspecto geral. Alguns trabalhos podem abordar mais de um tipo de problema, logo somente os problemas que o modelo de AM busca resolver, ou aqueles em que o modelo está embutido, são extraídos. Para a avaliação desta etapa, o mesmo processo descrito na seção 2.2 é utilizado.

2.6. Análise dos Dados

Os dados coletados são tabulados seguindo as seguintes características: técnicas utilizadas; tipo de aprendizado do modelo; banco de dados ou gerador de carga utilizado; e o tipo de problema que o modelo resolve.

3. Resultados

3.1. Estudos Levantados

Com base na estratégia de busca definida, foram encontrados um total de 80 trabalhos onde 75 são selecionados após a aplicação dos critérios de inclusão baseado na leitura dos resumos e na aplicação de filtros no MBA utilizado. Destes 75 trabalhos, após aplicação dos critérios de exclusão, com base na leitura da introdução, arquitetura do modelo e experimentos realizados, 31 trabalhos são selecionados para a etapa seguinte de avaliação da qualidade.

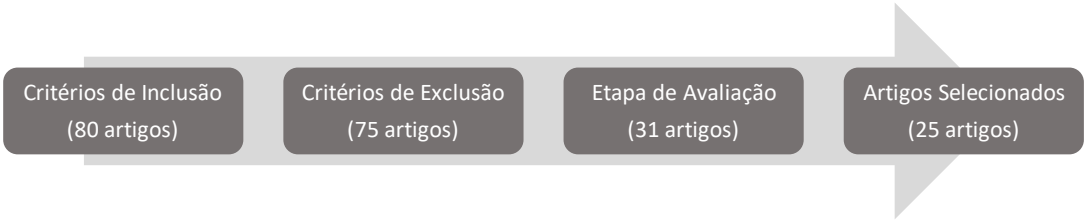

Figura 1. Processo de seleção de artigos

Na avaliação da qualidade, uma leitura detalhada foi conduzida nos experimentos realizados em cada trabalho, onde 6 artigos são excluídos por, principalmente, não apresentarem experimentos passíveis de replicação ou parcialmente descritos. Como resultado deste processo, 25 trabalhos são selecionados para a etapa de análise. Um resumo deste processo pode ser visualizado na Figura 1 enquanto na Figura 2 é possível visualizar a quantidade de artigos selecionados por ano de publicação.

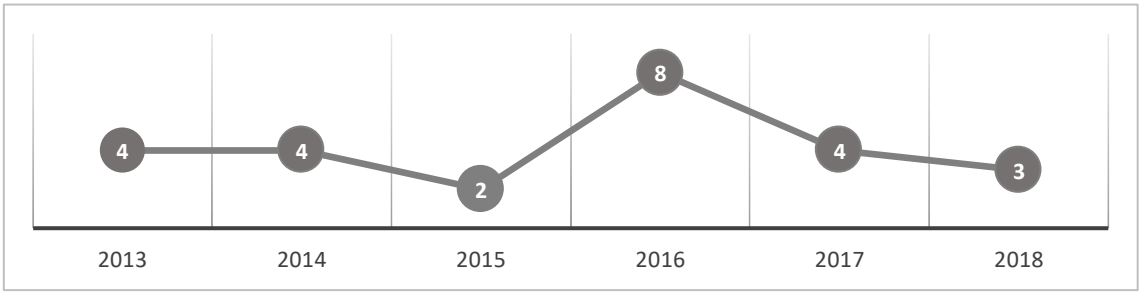

Figura 2. Artigos classificados por ano de publicação

3.2. Tipos de Problema

A problemática de gerenciamento de recursos em CN pode ser abordada sob diferentes aspectos, como: provisionamento de recursos e alocação, migração, escalonamento e posicionamento de máquinas virtuais. Analisando os trabalhos selecionados, é possível perceber uma predominância na utilização de AM na abordagem do problema de provisionamento de recursos, este compondo mais da metade dos trabalhos analisados. Não obstante, o problema de alocação de recursos tem sido o segundo problema onde AM tem sido mais utilizada como possível solução.

Tabela 1. Tipos de problemas abordados com aprendizado de máquina

Problema	Qtd.	Percentual
Provisionamento	15	60 %
Alocação	7	28 %
Migração	2	8 %
Escalonamento	2	8 %
Posicionamento	1	4 %

Demais problemas como migração, escalonamento e posicionamento de máquinas virtuais com técnicas de AM são abordados em uma quantidade pequena de trabalhos. Também é possível notar que alguns trabalhos abordam mais de um aspecto da problemática em questão. Na Tabela 1 encontram-se os tipos de problemas objetivo das técnicas de AM empregadas nos trabalhos analisados e a frequência com que são abordadas.

3.3. Técnicas Utilizadas

A área de AM dispõe de um conjunto considerado de técnicas de aprendizado supervisionado, não supervisionado e por reforço. No contexto de gerenciamento de recursos em CN, é possível perceber uma predominância na utilização de *Artificial Neural Networks* (ANN), sendo vista em 44% de todos os trabalhos analisados, seguido pelo Q-

Learning, utilizado em 24% dos trabalhos, e *Support Vector Machines* (SVM), visto em 21% dos trabalhos. Na Figura 3 é possível visualizar a frequência de utilização das técnicas de AM nos trabalhos analisados enquanto na Figura 4 é demonstrado a frequência de utilização destas técnicas em função dos tipos de problemas que abordam.

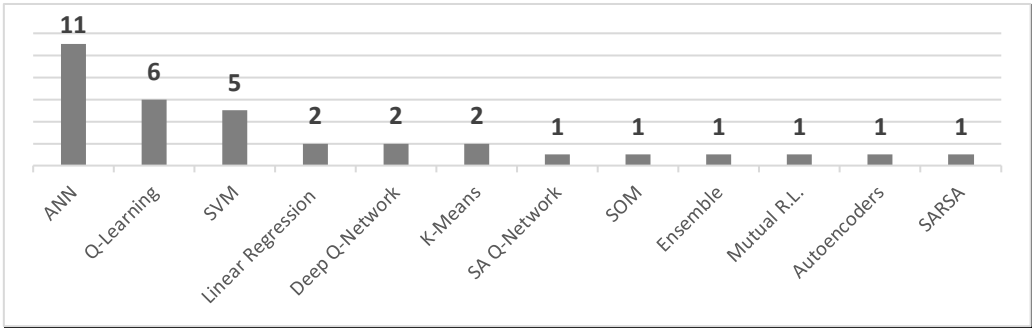

Figura 3. Técnicas de aprendizado de máquina utilizadas na problemática do gerenciamento de recursos em nuvem

Analisando os trabalhos selecionados em função do tipo de problema que abordam e a técnica de AM que utilizam, é possível perceber uma predominância na utilização de ANN no problema de provisionamento de recursos, principalmente na estimação da demanda por recursos, seguido pelo Q-Learning e SVM. No problema de alocação, o cenário se mantém com ANN e Q-Learning sendo as duas técnicas mais utilizadas.

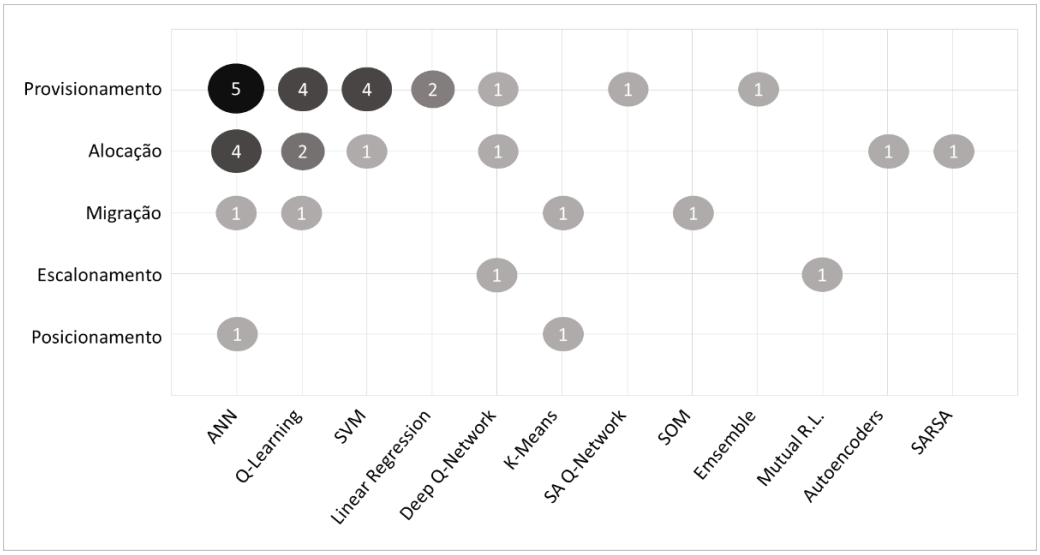

Figura 4. Mapeamento das técnicas de aprendizado de máquina em função dos tipos de problemas abordados

Em relação ao problema de migração de máquinas virtuais, ANN, Q-Learning, K-Means e *Self Organizing Map* (SOM) são utilizadas na mesma proporção. Já nos problemas de escalonamento e posicionamento de máquinas virtuais, uma parcela pequena de produções aborda estes problemas com a utilização de AM. Também é possível notar em alguns trabalhos a composição de múltiplas técnicas de AM na solução de um único problema.

3.4. Cargas de Trabalho

Para o treinamento e avaliação dos modelos treinados com base nas técnicas de AM, um banco de dados ou gerador de carga de trabalho, no caso do aprendizado por reforço, se faz necessário. Analisando este critério foi possível perceber que 48% dos trabalhos selecionados utilizam *traces* coletados de aplicações reais, sendo que a maioria destes utilizam o *trace* do *cluster* da Google, disponibilizado gratuitamente na internet.

Tabela 2. Tipos de cargas de trabalho utilizadas

Dados	Qtd	Percentual
Traces Reais	12	48 %
Gerados	7	28 %
Simulados	6	24 %

Em relação aos demais trabalhos, estes se dividem entre a utilização de dados simulados e dados gerados por geradores de carga específicos, tipicamente vistos em aplicações de *benchmarking*. Na Tabela 2 é possível visualizar a frequência com que cada tipo de carga de trabalho é utilizada.

4. Discussão e Conclusão

A busca pela maior eficiência na utilização dos recursos em um ambiente em nuvem motiva a adoção de novas técnicas e abordagens para a problemática de gerenciamento de recursos. Dada a dinamicidade de um ambiente em nuvem e a quantidade de informações coletadas com base no *tracing* de sua infraestrutura, a aplicação de AM se torna interessante tanto no reconhecimento de padrões como na tomada de decisões. Neste contexto, o presente artigo trouxe uma visão geral sobre o panorama do AM aplicada à problemática de gerenciamento de recursos em nuvem.

Seguindo um protocolo de revisão, 80 artigos foram coletados onde, após avaliação dos critérios de seleção e qualidade dos trabalhos, 25 artigos foram selecionados para análise. Com base nos resultados, foi possível perceber a predominância da utilização de ANN's, na aprendizagem supervisionada, e do Q-Learning, na aprendizagem por reforço. Técnicas de aprendizado supervisionado têm sido mais utilizadas na estimação da demanda por recursos, para auxiliar no problema de provisionamento, enquanto técnicas de aprendizado por reforço têm sido vistas tanto no provisionamento como no problema de alocação.

Analisando os tipos de cargas de trabalhos utilizados, observou-se uma predominância na utilização dos dados coletados do *cluster* da Google e na utilização de aplicações de *benchmarking*, principalmente como geradores de carga de trabalho em modelos que utilizam aprendizagem por reforço.

Em relação ao tipo de aprendizado mais utilizado, foi possível notar a predominância do aprendizado supervisionado em 60% dos trabalhos analisados em relação ao aprendizado por reforço, visto em 40% dos trabalhos, e do aprendizado não supervisionado, visto em apenas 12% dos trabalhos. A utilização de múltiplas técnicas de AM tem sido um cenário comum em alguns trabalhos, onde é utilizado aprendizado não supervisionado no pré-processamento para posterior aplicação do aprendizado supervisionado para fins de classificação e regressão.

Durante a condução deste trabalho, foi possível notar a falta de ferramentas padronizadas para avaliação dos modelos. A não padronização dos experimentos e das avaliações entre os diferentes modelos pode acarretar em um aumento no viés, dificuldade na interpretação e confiança nos resultados. Este aspecto fica em maior evidência nos trabalhos que utilizam aprendizado por reforço, onde é necessário modelar um ambiente para que o modelo aprenda através da experimentação. Portanto, verificou-se uma necessidade maior por métodos de avaliações e modelos de ambientes padronizados para a aplicação do AM no gerenciamento de recursos em CN. Todas as referências e classificações dos artigos levantados podem ser encontradas em http://goo.gl/W4B9Th.

Referências

Boutaba, R., Salahuddin, M. A., Limam, N., et al. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, v. 9, n. 1, p. 16.

Budgen, D. and Brereton, P. (2006). Performing systematic literature reviews in software engineering. In *Proceeding of the 28th international conference on Software engineering - ICSE '06*. . ACM Press. http://portal.acm.org/citation.cfm?doid=1134285.1134500.

Demirci, M. (2015). A Survey of Machine Learning Applications for Energy-Efficient Resource Management in Cloud Computing Environments. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. . IEEE.

Foster, I., Zhao, Y., Raicu, I. and Lu, S. (2008). Cloud Computing and Grid Computing 360-degree compared. In *Grid Computing Environments Workshop, GCE 2008*.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., et al. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, v. 47, p. 98–115.

Jennings, B. and Stadler, R. (2014). Resource Management in Clouds: Survey and Research Challenges. *Journal of Network and Systems Management*, v. 23, n. 3, p. 567–619.

Lopez-Pires, F. and Baran, B. (2015). Virtual Machine Placement Literature Review. *arXiv preprint arXiv:1506.01509*,

Manvi, S. S. and Krishna Shyam, G. (2014). Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. *Journal of Network and Computer Applications*, v. 41, n. 1, p. 424–440.

Mell, P. and Grance, T. (2011). The NIST Definition of Cloud Computing. *National institute of standards and technology (NIST) Special Publication 800-145*

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge,: The MIT Press.

Tyagi, R. and Gupta, S. K. (2018). A Survey on Scheduling Algorithms for Parallel and Distributed Systems. *Silicon Photonics & High Performance Computing*. Singapore: . v. 718p. 51–64.

Younge, A. J., Von Laszewski, G., Wang, L., Lopez-Alarcon, S. and Carithers, W. (2010). Efficient resource management for Cloud computing environments. In *International Conference on Green Computing*. . IEEE.