Uma classificação de vinhos baseada em regras fuzzy utilizando o algoritmo FARC-HD

Vítor M. Magalhães[1], Jair O. G. Carmona[1], Giancarlo Lucca[1,2], Helida Santos[1], Eduardo N. Borges[1]

[1]Programa de Pós-Graduação em Computação
Centro de Ciências Computacionais, Universidade Federal do Rio Grande – FURG

[2]Programa de Pós-Graduação em Modelagem Computacional
Centro de Ciências Computacionais, Universidade Federal do Rio Grande – FURG

{vitor.magalhaes,jogonzalezc,giancarlo.lucca}@furg.br

{helida,eduardoborges}@furg.br

Abstract. *A widely used application of supervised machine learning is in classification problems. There are different approaches, with different techniques and algorithms, in order to deal with this problem. Fuzzy Rule-Based Classification Systems are a consolidated and well known approach. In this sense, this work aims to present an analysis of this technique, as well as one of its classifier algorithms, called FARC-HD. Precisely, this algorithm was applied to a classic wine classification problem, demonstrating its high interpretability and accuracy.*

Resumo. *Uma aplicação bastante utilizada do aprendizado de máquina supervisionado ocorre em problemas de classificação. Existem diferentes abordagens, com técnicas e algoritmos diversos, para lidar com este problema. Sistemas de classificação baseados em regras fuzzy são uma abordagem consolidada e amplamente utilizada. Nesse sentido, este trabalho tem por objetivo apresentar uma análise sobre esta abordagem, bem como um de seus algoritmos classificadores, chamado FARC-HD. Mais precisamente, este algoritmo foi aplicado em um problema clássico de classificação de vinhos, demonstrando sua alta interpretabilidade e precisão.*

1. Introdução

O Aprendizado de Máquina (AM) [Michalski et al. 2013] está presente em nosso cotidiano resolvendo diversos problemas que, por muitas vezes, sequer podemos notar. Pode-se citar suas aplicações em diversas áreas, como gestão de resultados de praticantes de atividade física [da Silva et al. 2020]; análise de reportagens em redes sociais buscando identificar *fake news* [Radicchi et al. 2019]; e até no sistema judiciário há inúmeras possibilidades de aplicação destes conceitos [Pedreira et al. 2021].

A aplicação dos conceitos de aprendizado de máquina naturalmente ocorre por um sistema de aprendizado, cuja definição conceitual é a de um programa de computador que toma decisões baseado em experiências acumuladas através das soluções bem sucedidas de problemas anteriores [Monard e Baranauskas 2003].

O AM é uma subárea da Inteligência Artificial [Russell e Norvig 2002], podendo ser considerado como um ramo em evolução de algoritmos computacionais

projetados para emular a inteligência humana aprendendo com o ambiente ao redor [El Naqa e Murphy 2015]. Diversas técnicas podem ser usadas para se aplicar o AM para resolução de problemas. Tratando-se de aprendizado de máquina supervisionado [Tan et al. 2005], mais especificamente para resolver problemas de classificação [Kotsiantis et al. 2007], uma das abordagens é a *Fuzzy Rule-Based Classification System* (FRBCS), ou sistema de classificação baseado em regras *fuzzy*. Sua principal diferenciação frente a outros classificadores é o fornecimento de modelos interpretáveis para o usuário final, justamente pela possibilidade de aplicação da lógica *fuzzy* [Zadeh 1996], discorrida no presente trabalho.

Sabendo-se que a compreensão detalhada dos algoritmos de processamento de informações para o AM pode levar a um melhor entendimento das habilidades de aprendizagem humana [Mitchell 1997], este trabalho tem como objetivo contribuir com o entendimento do funcionamento de um FRBCS. Nesse sentido, foi considerado o uso do algoritmo *Fuzzy Association Rule-based Classification model for High Dimensional problems* (FARC-HD) [Alcalá-Fdez et al. 2011], pelo fato de ser considerado um algoritmo estado da arte. Além disso, para verificar o método frente ao problema de classificação, o algoritmo foi aplicado em uma base de dados conhecida na literatura, com o objetivo de classificar o tipo de vinho - branco ou tinto - de acordo com resultados de análises.

O trabalho está organizado na seguinte forma: na Seção 2, a fundamentação teórica do presente trabalho é explicitada, partindo da lógica *fuzzy*, passando pelos *Fuzzy Rule-Based Classification Systems* e pelo algoritmo FARC-HD. Na Seção 3, os trabalhos relacionados são discutidos. Na Seção 4, a metodologia é apresentada, bem como as razões da escolha do algoritmo, os datasets e seu pré-processamento, e a configuração utilizada pelo algoritmo. Os resultados obtidos são apresentados e analisados na Seção 5. Por fim, na Seção 6, as conclusões são apresentadas.

2. Fundamentação teórica

Esta seção aborda os principais conceitos utilizados no desenvolvimento deste trabalho.

2.1. Lógica Fuzzy

Na teoria clássica (ou *booleana*) dos conjuntos, também conhecida como conjuntos *crisp*, um elemento pode pertencer (1) ou não pertencer (0) a um determinado conjunto. Entretanto, o mundo não funciona de maneira exata, e muito menos o raciocínio humano. A lógica clássica funciona, em sua plenitude, quando é aplicada a termos exatos e fronteiras bem definidas (não-vagas). Entretanto, quando ela é aplicada em ambientes incertos ou imprecisos, seu funcionamento deixa a desejar. Um exemplo que explica claramente sua inaplicabilidade a estas situações é o conhecido *Paradoxo de Sorites*: em que momento um monte de areia deixa de ser um monte, se formos retirando grão por grão?

Então, buscando expressar o funcionamento do mundo aplicado à teoria dos conjuntos, surge a teoria dos conjuntos *fuzzy*, na qual, um elemento pode pertencer parcialmente a um conjunto, abrindo uma infinidade de possibilidades entre o 0 e o 1 através dos graus de pertinência destes mesmos elementos aos conjuntos.

Assim sendo, a teoria dos conjuntos *fuzzy* é a predecessora da lógica *fuzzy*. Ela permite que esta última suporte modos de raciocínio que são aproximados ao invés de exatos. A modelagem linguística *fuzzy* permite lidar com sistemas através da construção de um

modelo linguístico que pode se tornar interpretável por seres humanos [Gacto et al. 2011]. Então, a lógica *fuzzy* pode ser vista como uma tentativa de formalização/mecanização de duas capacidades humanas notáveis: primeiro, a capacidade de conversar, raciocinar e tomar decisões racionais em um ambiente de informações imperfeitas. E, segundo, a capacidade de realizar uma ampla variedade de tarefas físicas e mentais sem quaisquer medições e cálculos [Zadeh 2008].

2.2. Sistemas de classificação baseados em regras *fuzzy*

Sistemas de classificação baseados em regras *fuzzy* são ferramentas úteis para lidar com problemas de classificação, enfatizando a significância das variáveis linguísticas e da lógica *fuzzy* [Zadeh 1975] e tendo por principal vantagem a alta interpretabilidade dos modelos de saída [Sanz et al. 2010].

Na Figura 1, verifica-se que um sistema *fuzzy* possui alguns componentes principais: dada uma entrada, inicialmente ela é *fuzzificada* através de um *fuzzificador*. O *fuzzificador* contém as funções de pertinência das variáveis linguísticas de entrada, recebendo um valor do universo de discurso e retornando o grau de pertinência do valor ao respectivo conjunto *fuzzy*. O próximo componente é a *máquina de inferência*, que é responsável por realizar todos os cálculos necessários, recebidos do componente *conhecimento*. Este é composto pelo banco de dados e pela base de regras do sistema. Por último, há ainda o *defuzzificador*, que além de possuir as funções de pertinência das variáveis linguísticas de saída, ele recebe os graus de pertinência para uma variável linguística - de saída - e retorna um valor para essa variável.

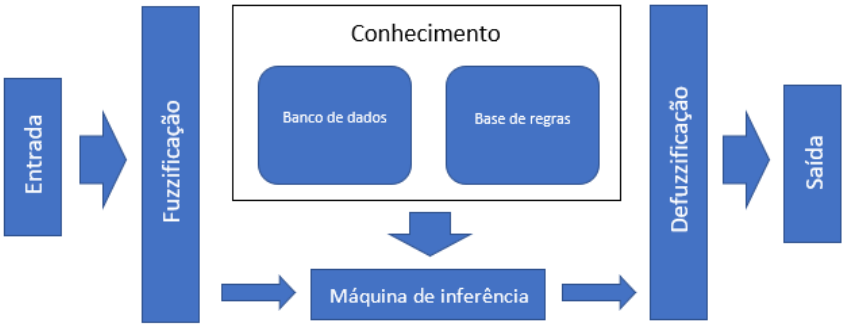

Figura 1. Componentes de um sistema *fuzzy*, adaptado de [Lucca 2018].

2.3. O algoritmo FARC-HD

A escolha por explorar o algoritmo FARC-HD em um problema de classificação deu-se pelo fato de o algoritmo basear-se em três grandes pilares de sustentação:

- obtenção das regras de associação *fuzzy* para classificação – uma árvore de busca é utilizada para listar todos os conjuntos possíveis de elementos *fuzzy* frequentes e para gerar regras de associação, limitando a profundidade dos ramos para encontrar um pequeno número de regras *fuzzy*;
- pré-seleção de regras candidatas – para diminuir o custo computacional do estágio de pós-processamento genético, o algoritmo usa a medida de acurácia relativa ponderada melhorada *wWRAcc* (sigla em inglês para *improved Weighted Relative Accuracy*) para pré-selecionar as regras mais importantes;

- seleção de regras genéticas e *tuning* lateral – um algoritmo genético é usado para selecionar e fazer a sintonia em um conjunto de regras de associação *fuzzy* com alta acurácia, para buscar a conhecida "sinergia positiva" apresentada por ambas as técnicas (seleção e *tuning*).

3. Trabalhos relacionados

Classificação *fuzzy* já foi empregada com sucesso em muitos contextos de problemas reais. No trabalho de [Assis Silva e Soares de Souza Lima 2009], foi utilizada para melhor visualizar as mudanças das classes de fertilidade do solo em cultivares de café, o que melhor definiu as zonas de transição gradual, ao invés de classificar as informações de forma exata. A lógica *fuzzy* foi aplicada para poder detectar e classificar falhas de curto-circuito em sistemas elétricos [Barros 2009]. Perturbações sutis diferenciavam o padrão em relação aos eventos caracterizados como de baixa e ou de média impedância. Um classificador baseado em regras *fuzzy* também já foi utilizado para combinar o conhecimento especializado do meteorologista com a velocidade e a objetividade de um computador, tendo suas regras formuladas através de conjuntos *fuzzy* para permitir a flexibilidade que o problema exigia [Bardossy et al. 1995]. Também pode-se citar a utilização de classificadores baseados em regras *fuzzy* para determinar o risco de um paciente sofrer de uma doença cardiovascular, fornecendo um modelo interpretável para explicar o resultado, servindo de base para a tomada de decisão da equipe médica [Sanz et al. 2014].

4. Metodologia

Esta seção apresenta a metodologia adotada para o desenvolvimento do trabalho, que baseou-se no processo de descoberta de conhecimento em bancos de dados (*Knowledge Discovery in Databases* - KDD) [Fayyad et al. 1996].

Após a definição do problema de classificação e consequente atributo-alvo, além das etapas de pré-processamento e transformação do *dataset*, é na etapa de *data mining* [Tan et al. 2005] que ocorre a aplicação do algoritmo FARC-HD. O método de funcionamento do algoritmo pode ser observado através da Figura 2.

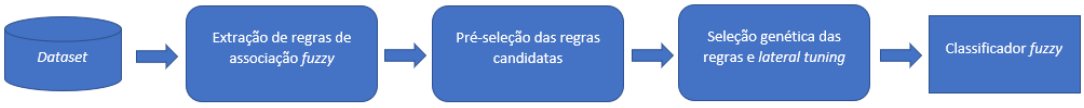

Figura 2. Método utilizado pelo algoritmo FARC-HD, adaptado de [Alcalá-Fdez et al. 2011].

Através da figura acima, verifica-se que, após receber o *dataset* como entrada, o primeiro passo realizado pelo algoritmo é a extração das regras de associação *fuzzy*. Após, é realizada a pré-seleção das regras candidatas e, por último, há a seleção genética das regras e o *lateral tuning*.

4.1. Datasets

Neste trabalho foram utilizados dois *datasets* distintos[1]. Um refere-se a resultados de análise para vinhos brancos, enquanto o outro analisa vinhos tintos. Os *datasets* possuem exatamente os mesmos 12 (doze) atributos, todos do mesmo tipo (ponto flutuante), diferenciando-se apenas pelo número de instâncias. O objetivo, então, é classificar o tipo de vinho - *branco* ou *tinto* - a partir dos resultados de análise.

4.2. Pré-processamento dos datasets

Ambos conjuntos de dados foram combinados em um mesmo *dataset*, tendo sido adicionado o atributo denominado *is red*, para identificar o tipo de vinho de acordo com o resultado de análise: 1 para vinho tinto e 0 para vinho branco, conforme a Tabela 1.

Tabela 1. Concatenação dos datasets considerados.

	acidez fixa	acidez volátil	ácido cítrico	açúcar residual	cloretos	dióx. de enx. livre	dióx. de enx. total	densidade	pH	sulfatos	álcool	qualidade	is red
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	1
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5	1
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6	1
4	7.4	0.66	0.00	1.8	0.075	13.0	40.0	0.99780	3.51	0.56	9.4	5	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5315	6.2	0.21	0.29	1.6	0.039	24.0	92.0	0.99114	3.27	0.50	11.2	6	0
5316	6.6	0.32	0.36	8.0	0.047	57.0	168.0	0.99490	3.15	0.46	9.6	5	0
5317	6.5	0.24	0.19	1.2	0.041	30.0	111.0	0.99254	2.99	0.46	9.4	6	0
5318	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34	0.38	12.8	7	0
5319	6.0	0.21	0.38	0.8	0.020	22.0	98.0	0.98941	3.26	0.32	11.8	6	0

Após, buscou-se eliminar instâncias que pudessem prejudicar o aprendizado em razão de sua baixa relevância. Não haviam instâncias com dados nulos, então foram apenas removidas as duplicadas. Após, foram analisadas as distribuições de alguns atributos e eliminadas instâncias pouco representativas. Por exemplo, a Figura 3 apresenta o histograma do atributo qualidade. Foram removidas as instâncias que possuíam valores inferiores a 5 e superiores a 7.

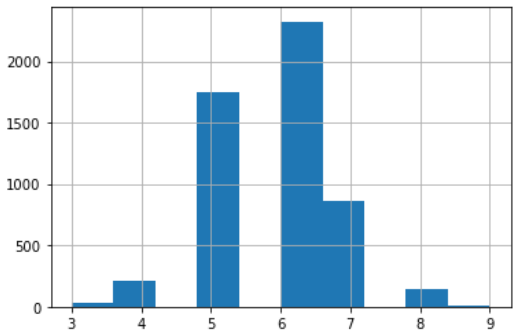

Figura 3. Histograma do atributo qualidade.

4.3. Configuração utilizada no FARC-HD

A técnica *K-Fold Cross Validation* (CV) é uma das abordagens mais usadas para validação de erro em classificadores [Tan et al. 2005]. Basicamente, consiste em dividir um conjunto de dados em *K* subconjuntos; então, iterativamente, *K-1* subconjuntos são usados

[1]Datasets do repositório de aprendizado de máquina - UCI, disponíveis em https://archive.ics.uci.edu/ml/datasets/wine+quality

para aprender o modelo, enquanto o restante é utilizado para avaliar o seu desempenho [Anguita et al. 2012]. Neste trabalho, foram utilizados 10 (dez) *folds* (subconjuntos).

Na configuração do algoritmo FARC-HD, para a obtenção dos resultados, foram utilizadas cinco variáveis linguísticas diferentes, com suporte mínimo de 0.05 (proporcionalidade de transações que contem o conjunto) e confiança máxima (estimativa de probabilidade) de 0.8. Destacamos a seguir os valores dos parâmetros usados no experimento, configurados no software KEEL[2]: *Seed = 1286082570; Number of linguistic values = 5; Minimun support = 0.05, Maximum confidence = 0.8, Depth of the trees (Depthmax) = 3, Parameter K of the prescreening = 2, Maximum number of evaluations = 15000, Population size = 50, Parameter alpha = 0.15, Bits per gen = 30, Type of inference = 1*.

5. Resultados

Após análise dos resultados obtidos pela saída do software KEEL, observou-se que justamente pela melhor definição dos limites de fronteiras entre as classes, característica de um sistema de classificação baseado em regras *fuzzy*, o algoritmo demonstrou um ótimo desempenho, obtendo uma acurácia média de 99.3% no subconjunto (partição) de treino e 98.9% no subconjunto (partição) de teste.

Por restrições de espaço, para demonstrar a assertividade do método, foi escolhido aleatoriamente um dos *folds*. Pode-se observar seus resultados abaixo:

- TotalNumberOfNodes: 4
- NumberOfLeafs: 5
- NumberOfAntecedentsByRule: 2.4

- NumberOfItemsetsTraining: 160
- NumberOfCorrectlyClassifiedTraining: 158
- PercentageOfCorrectlyClassifiedTraining: 98.75%
- NumberOfIncorrectlyClassifiedTraining: 2
- PercentageOfIncorrectlyClassifiedTraining: 1.25%

- NumberOfItemsetsTest: 18
- NumberOfCorrectlyClassifiedTest: 18
- PercentageOfCorrectlyClassifiedTest: 100.0%
- NumberOfIncorrectlyClassifiedTest: 0
- PercentageOfIncorrectlyClassifiedTest: 0.0%

Nos dados acima, verifica-se que, no *fold* supracitado, foram treinadas 160 instâncias, das quais 158 foram corretamente classificadas - uma acurácia do conjunto de treino de 98.75% - com apenas dois erros, um percentual de apenas 1.25%. Já no conjunto de teste, o resultado foi de 100% de acerto em todas as dezoito instâncias, demonstrando o desempenho excelente do classificador.

6. Conclusão

Problemas de classificação - aprendizado de máquina supervisionado - podem ser resolvidos com algoritmos de classificação baseados na lógica *booleana*. Nesse caso, os limites

[2]Ferramenta de *data mining* KEEL – https://www.keel.es

entre as diferentes classes de dados seriam mais facilmente identificados, tendo em vista que suas fronteiras são bem definidas. Todavia, em muitos problemas, a definição dessas fronteiras não é exatamente clara. Assim, objetivando predizer as classes às quais as instâncias pertencem através de um método de raciocínio *fuzzy*, considera-se infinitos valores no intervalo $[0, 1]$. Logo, as fronteiras entre as classes passam a ser definidas pelos graus de pertinência dos elementos, ou seja, o quanto um determinado elemento pertence a um determinado conjunto. Esta característica permite à lógica *fuzzy* trabalhar com termos linguísticos e, assim, a composição das regras se torna amigável ao entendimento humano, tornando os sistemas de classificação baseados em regras *fuzzy* muito úteis.

Neste trabalho, foi considerada a aplicação do algoritmo FARC-HD em um problema clássico de classificação de duas classes distintas de vinhos. Pode-se observar que tanto a delimitação dos triângulos das funções de pertinência, quanto o *lateral tuning* do algoritmo criaram fronteiras mais suaves e com limites mais bem definidos, resultando em uma acurácia de 98.75% para os dados de treino e 100% para os dados de teste, demonstrando de maneira objetiva a excelente performance neste problema.

Como sugestão para futuros trabalhos que complementem as avaliações aqui descritas, sugere-se um aprofundamento do estudo, diferenciando a presente abordagem da aplicação de técnicas de aprendizado de máquina tradicionais ou de agentes inteligentes e os impactos das possibilidades dos conjuntos *fuzzy* no mesmo problema, além da aplicação do mesmo algoritmo FARC-HD em diferentes problemas, explorando um pouco mais a alta interpretabilidade do modelo.

Agradecimentos

O presente trabalho foi realizado com apoio da CAPES (DS 88887.622570/2021-00, PNPD 464880/2019-00) e FAPERGS (19/2551-0001279-9).

Referências

Alcalá-Fdez, J., Alcala, R., e Herrera, F. (2011). A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. *IEEE Transactions on Fuzzy Systems*, 19(5):857–872.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., e Ridella, S. (2012). The 'k'in k-fold cross validation. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 441–446.

Assis Silva, S. d. e Soares de Souza Lima, J. (2009). Lógica fuzzy no mapeamento de variáveis indicadoras de fertilidade do solo. *Idesia (Arica)*, 27(3):41–46.

Bardossy, A., Duckstein, L., e Bogardi, I. (1995). Fuzzy rule-based classification of atmospheric circulation patterns. *Int. journal of climatology*, 15(10):1087–1097.

Barros, A. C. (2009). Detecção e classificação de faltas de alta impedância em sistemas elétricos de potência usando lógica fuzzy. Dissertação de Mestrado. Universidade Estadual Paulista (UNESP). Disponível em http://hdl.handle.net/11449/87092.

da Silva, H. d. B., Antoniazzi, R. L., Schuch, R. R., e Chicon, P. M. M. (2020). Gestão de atividades físicas com utilização de machine learning. *Anais do Seminário Interinstitucional de Ensino, Pesquisa e Extensão*.

El Naqa, I. e Murphy, M. J. (2015). What is machine learning? In *machine learning in radiation oncology*, pages 3–11. Springer.

Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

Gacto, M. J., Alcalá, R., e Herrera, F. (2011). Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences*, 181(20):4340–4360.

Kotsiantis, S. B., Zaharakis, I., Pintelas, P., et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24.

Lucca, G. (2018). Aggregation and pre-aggregation functions in fuzzy rule-based classification systems. Tese de Doutorado. Universidad Pública de Navarra. Disponível em https://academica-e.unavarra.es/handle/2454/34775.

Michalski, R. S., Carbonell, J. G., e Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, 18(3):11–11.

Monard, M. C. e Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, 1(1):32.

Pedreira, M. R. G. et al. (2021). Inteligência artificial e machine learning no judiciário: as mudanças processuais e os impactos da tecnologia no sistema brasileiro. Trabalho de conclusão de curso. Universidade Católica do Salvador. Disponível em http://ri.ucsal.br:8080/jspui/handle/prefix/4441.

Radicchi, L. B. M., Barion, M. C., e Ferreira, A. (2019). Arquitetura de machine learning para análise de reportagens textuais em redes sociais para a detecção de fake news.

Russell, S. e Norvig, P. (2002). *Artificial intelligence: a modern approach*. Prentice Hall.

Sanz, J. A., Fernández, A., Bustince, H., e Herrera, F. (2010). Improving the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets and genetic amplitude tuning. *Information Sciences*, 180(19):3674–3685.

Sanz, J. A., Galar, M., Jurio, A., Brugos, A., Pagola, M., e Bustince, H. (2014). Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. *Applied Soft Computing*, 20:103–111.

Tan, P.-N., Steinbach, M., e Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning – I. *Information Sciences*, 8(3):199–249.

Zadeh, L. A. (1996). Fuzzy sets. In *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*, pages 394–432. World Scientific.

Zadeh, L. A. (2008). Is there a need for fuzzy logic? *Information sciences*, 178(13):2751–2779.