FUZZY ASSOCIATION RULE–BASED CLASSIFIER FOR HIGH–DIMENSIONAL PROBLEMS(FARC–HD)

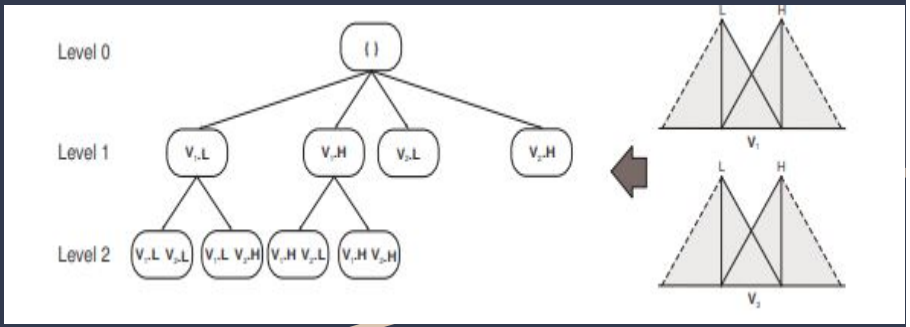
Conceitos Necessários

- Lógica Fuzzy: Uma extensão da lógica clássica, proposta por Zadeh em 1965. Visa medir incertezas, trabalhando com valores a mais do que apenas o 0 e 1, utilizando de seus graus de pertinência;
- Regras de Associação: São usadas para encontrar relações interessantes entre variáveis de qualquer base de dados. Conseguem se integrar com a lógica fuzzy por conseguir utilizar conjuntos fuzzy em sua composição. Tem o formato padrão de: *IF* X *THEN* Y;
- Fuzzy Rule-Based Classification Systems: Um modelo de classificação que utiliza das regras de associação em conjunto com os conjuntos fuzzy para predição de dados de um dataset.

Introdução e Objetivo do FARC-HD

- Esse algoritmo tem foco em lidar com a dificuldade dos FRBCS, grandes datasets. Quanto maior um dataset, maior o seu custo computacional, pois o crescimento do conjunto de regras fuzzy é exponencial.
- Utiliza um sistema de regras de associação fuzzy para a classificação de dados, que diferentemente do FRBCS(Fuzzy Rule-Based Classification System), gera as suas próprias regras de associação fuzzy a partir dos dados de entrada.
- Utiliza de Algoritmos Genéticos, técnicas de computação evolutiva e algoritmos como o Apriori em seu desenvolvimento;
- É dividido em 3 etapas principais:
    - Geração das Regras de Associação Fuzzy;
    - Pré-seleção de Regras Candidatas;
    - Avaliação e Seleção das Regras;

Funcionamento do FARC-HD



Primeira etapa - Geração das Regras de Associação Fuzzy

- Nessa etapa são criadas as regras de associação fuzzy que **poderão** ser utilizadas no modelo, ou seja, é desenvolvido uma base/conjunto de regras de associação fuzzy;
- Utiliza de uma árvore de busca para extrair as regras dos dados de entrada;
  - A árvore é limitada pelo parâmetro de $\mathbf{depth_{max}}$, que define a profundidade máxima que a árvore pode alcançar;
  - Ela lista todos os itemsets de cada classe, utilizando das métricas de suporte e confiança;
  - Cada nó inserido na árvore passa pelo cálculo do suporte e da confiança daquele nó.
  - Caso um nó tenha um valor de suporte inferior ao *suporte mínimo*, ele pode ser considerado uma folha (todos os nós na subárvore desse nó também terão valor inferior ao suporte mínimo).
  - o cálculo do suporte mínimo se dá pela equação: MinimumSupportCj = minSup · fCj, sendo minSup o valor determinado pelo expert e fCj a frequência da classe Cj
  - Caso um nó tenha um valor de confiança maior do que a *confiança máxima* estipulada ele também será considerado um nó folha, pois já atingiu a qualidade exigida, logo não precisa ser expandido (todos os nós na sua subárvore também terão a confiança acima da confiança máxima).

Funcionamento do FARC-HD

Segunda Etapa - Pré-seleção de Regras Candidatas

- Nesta etapa é realizada uma redução na quantidade de regras criadas na etapa anterior;
- Utiliza a técnica da descoberta de subgrupos para pré-selecionar as regras mais interessantes;
- Após isso usa de outra técnica chamada de esquema de ponderação de padrões, em que são atribuídos pesos para os padrões identificados de cada classe;
- Com esses pesos definidos o modelo analisa os pesos dos padrões positivos, ou seja, padrões que são cobertos pela regra a ser analisada, e atualiza os pesos de acordo com o resultado.
- Essa atualização ocorre reduzindo o valor do peso pela equação:

$$w(e_j, i) = \frac{1}{i+1}$$

- Para cada iteração as regras são classificadas de melhor a pior com base no critério de avaliação de regras dado por outra equação:

$$wWRAcc''(A \to C_j) = \frac{n''(A \cdot C_j)}{n'(C_j)} \cdot \left(\frac{n''(A \cdot C_j)}{n''(A)} - \frac{n(C_j)}{N}\right)$$

Funcionamento do FARC-HD

Segunda Etapa - Continuação

- Os padrões cobertos pela regra a ser analisada são eliminados caso sejam cobertos **kt** vezes;
- O processo acaba quando todos os padrões da classe tiverem sido cobertos **kt** ou quando não tiver mais nenhuma regra no conjunto de regras;
- Exemplo: R = If X1 is [0.0, 5.0[and X2 is [5.0, 10.0] → C

ID	X_1	X_2	$Class$	$Weight$
ID1	0.0	10.0	C_1	1.0
ID2	2.5	4.0	C_2	1.0
ID3	3.2	1.0	C_2	0.0
ID4	9.0	5.0	C_2	1.0
ID5	2.5	10	C_1	0.5

$$wWRAcc'(R) = \frac{1.0+0.5}{1.0+1.0+0.0+1.0+0.5} \cdot \left(\frac{1.0+0.5}{1.0+0.5} - \frac{2}{5}\right)$$
$$= 0.257$$

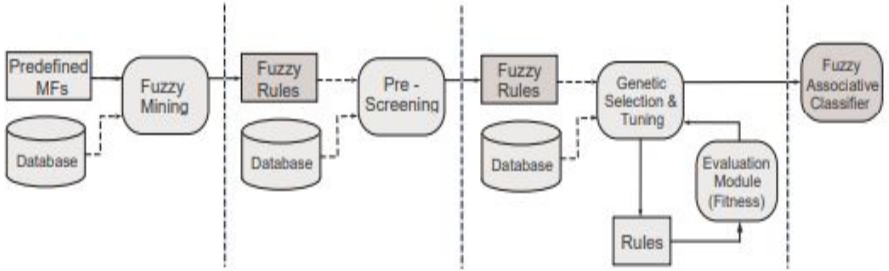
Quanto mais próximo de 1 no intervalo de [-1.0, 1.0], melhor é a regra para classificação seguindo esse métrica.

Funcionamento do FARC-HD

Terceira Etapa - Avaliação e Seleção das Regras

- Nesta etapa as regras de associação serão otimizadas, visando melhorar o desempenho do sistema o máximo possível;
- O conjunto compacto de regras de associação definido na etapa anterior agora irá passar por vários processos que utilizam da seleção de regras e ajuste lateral para definir o conjunto de regras de associação fuzzy final que será utilizado na predição de novos dados;
- As abordagens são:
  - Esquema de Codificação;
  - Avaliação de Cromossomos;
  - Pool de Genes Inicial;
  - Operador de Cruzamento;
  - Abordagem de Reinício;
  - Critério de Parada;

Visualização do Funcionamento do Modelo



O Algoritmo Seguido Pelo Modelo

- INPUT: A dataset with size T and m attributes, each with qj predefined linguistic terms.
- OUTPUT: A fuzzy associative classifier.

- Stage 1. Fuzzy Association Rule Extraction for Classification. For each class Cj :
  - Step 1: Calculate the minimum support of the class Cj according to eq. (8).
  - Step 2: Create the levels 0 and 1 of the tree.
  - Step 3: Create a new level in the tree.
  - Step 4: Prune nodes.
  - Step 5: If there are more than 2 nodes in the new level and the depth of the tree is less than Depthmax, go to Step 3.
  - Step 6: Generate the rules with the class Cj on the right-hand side.

O Algoritmo Seguido Pelo Modelo

- Stage 2. Candidate Rule Prescreening. For each class Cj :
  - Step 7: Set the weight of the patterns as 1.
  - Step 8: Calculate the wWRAcc” value for each rule.
  - Step 9: Select the best rule as a part of the initial RB for Stage 3 and remove it from the candidate rule set.
  - Step 10: Decrease the weight of the patterns covered by the selected rule.
  - Step 11: If any pattern has been covered less than kt times and there are more rules in the candidate rule set, go to Step 8.

- Stage 3. Rule Selection and Lateral Tuning.
  - Step 12: Generate the initial population with P chromosomes.
  - Step 13: Evaluate the population.
  - Step 14: Initialize the threshold value taking into account Gray codings, L = Linitial.
  - Step 15: Generate the next population:

Performance do Modelo em 26 datasets diferentes

	2SLAVE				FH-GBML				SGERD				FARC-HD			
Dataset	**#R**	**#C**	**Tra**	**Tst**	**#R**	**#C**	**Tra**	**Tst**	**#R**	**#C**	**Tra**	**Tst**	**#R**	**#C**	**Tra**	**Tst**
Iris	4.0	3.2	94.32	94.44	14.9	3.3	**98.89**	94.00	**3.4**	2.0	95.14	94.89	4.0	**1.1**	98.59	**96.00**
Phoneme	11.5	24.2	77.52	76.41	17.4	4.5	79.57	79.66	**3.6**	**1.9**	75.74	75.55	17.8	2.2	**83.52**	**82.14**
Monks	3.0	**1.3**	97.22	97.26	14.7	2.1	98.36	98.18	**2.2**	1.4	80.56	80.65	14.2	2.0	**99.92**	**99.77**
Appendicitis	4.4	7.5	91.20	82.91	13.8	7.0	93.19	**86.00**	**2.5**	2.0	87.88	84.48	6.8	**1.8**	**93.82**	84.18
Ecoli	12.6	9.6	89.51	**84.53**	10.3	4.2	75.83	69.38	**9.4**	**1.6**	76.53	74.05	33.8	2.4	**92.33**	82.19
Pima	7.8	8.8	76.35	73.71	10.6	6.0	77.18	75.26	**3.1**	**2.0**	74.01	73.37	22.7	2.4	**82.90**	**75.66**
Yeast	23.6	9.8	55.54	51.27	**7.5**	5.9	52.31	51.42	11.3	**1.5**	39.83	38.77	35.2	2.6	**63.81**	**58.50**
Glass	15.1	9.3	74.25	58.05	9.4	5.0	64.85	57.99	**6.9**	**2.0**	61.31	58.49	22.7	2.5	**81.10**	**70.24**
Page-blocks	7.5	10.3	91.39	91.39	7.4	8.1	94.37	94.21	**6.5**	**2.0**	90.83	90.72	19.1	2.3	**95.62**	**95.01**
Magic	4.1	10.5	73.97	73.96	9.9	8.2	81.23	81.30	**3.1**	**2.0**	72.17	72.06	43.3	2.5	**85.36**	**84.51**
Wine	5.5	10.3	92.52	89.47	9.2	4.7	95.51	92.61	**4.2**	2.0	93.67	91.88	8.7	**1.6**	**99.94**	**94.35**
Heart	4.3	10.7	75.35	71.36	12.7	3.2	84.65	75.93	**2.7**	**1.9**	74.83	73.21	27.0	2.6	**93.91**	**84.44**
Cleveland	11.9	12.8	54.24	48.82	6.9	4.5	58.29	53.51	**6.4**	**2.0**	56.55	51.59	61.3	2.9	**88.18**	**55.24**
Vowel	63.1	15.6	**82.06**	71.11	**9.2**	13.0	67.41	67.07	18.0	**1.9**	72.99	65.83	72.3	2.9	80.48	**71.82**
Crx	2.4	6.7	74.36	74.06	11.6	6.2	86.32	**86.60**	**2.1**	**1.9**	85.04	85.03	25.4	2.6	**91.17**	86.03
Pen-based	40.0	18.9	81.32	81.16	18.4	8.0	50.69	50.45	**15.9**	**2.0**	68.17	67.93	152.8	2.8	**97.04**	**96.04**
German	6.5	8.4	72.44	70.53	5.1	4.0	**87.11**	**87.01**	**3.4**	**2.0**	68.54	67.97	85.7	2.8	86.81	72.80
Twonorm	26.5	15.5	87.45	86.99	12.0	7.6	86.26	85.97	**3.1**	**2.0**	74.49	73.98	60.9	2.6	**96.64**	**95.28**
Ringnorm	**4.6**	23.7	80.12	79.63	6.9	11.3	87.34	86.92	6.8	2.0	73.21	72.63	24.0	**1.9**	**95.13**	**94.08**
Wdbc	5.2	8.1	92.43	92.33	7.2	4.9	95.12	92.26	**3.7**	2.0	91.79	90.68	10.4	**1.7**	**98.57**	**95.25**
SatImage	57.9	25.1	84.03	81.69	16.5	36.0	74.90	74.72	**12.2**	**2.0**	77.15	77.10	76.1	2.7	**88.68**	**87.32**
Texture	34.9	23.9	82.87	81.57	**14.6**	40.0	69.91	70.15	18.6	**2.0**	72.12	71.66	54.5	2.7	**93.71**	**92.89**
Spectfheart	6.1	21.7	80.71	79.17	10.8	44.0	79.28	72.36	**2.1**	1.9	79.05	78.16	12.9	**1.8**	**91.43**	**79.83**
Spambase	7.9	11.4	69.87	70.14	3.9	18.5	77.86	77.22	**3.7**	**2.0**	72.90	72.98	29.8	2.4	**92.37**	**91.93**
Sonar	9.6	17.5	77.92	71.42	10.3	4.7	80.56	68.24	**3.2**	**2.0**	74.22	71.90	18.0	2.3	**98.77**	**80.19**
Movementlibras	47.4	26.5	90.13	67.04	**12.1**	90.0	77.87	68.89	22.9	**2.0**	72.37	68.09	83.1	2.9	**95.52**	**76.67**
Mean	16.4	13.5	80.73	76.94	10.9	13.6	79.80	76.82	**6.9**	**1.9**	75.43	73.99	39.3	2.3	**90.97**	**83.94**

#R-> número médio de regras;

#C-> número médio de condições no antecedente da regra;

Tra -> classificação média obtida sobre a training data (em porcentagem)

Tst -> classificação média obtida no test data (em porcentagem)

Referências

ALCALA-FDEZ, J.; ALCALA, R.; HERRERA, F. A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems With Genetic Rule Selection and Lateral Tuning. IEEE Transactions on Fuzzy Systems, v. 19, n. 5, p. 857–872, out. 2011.

LUCCA, Giancarlo. **Aggregation and pre-aggregation functions in fuzzy rule-based classification systems**. 2018. 184 p. Dissertação de Doutorado — Universidad Pública de Navarra, Pamplona, 2018.