UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DA COMPUTAÇÃO

Disciplina: Computação Difusa e Intervalar

Aplicação de um FRBCS

Maurício Balboni

1 Introdução

O presente trabalho é resultado de uma atividade proposta pela disciplina de "Computação Difusa e Intervalar", ofertada pelo curso de pós-graduação em Engenharia da Computação, pela Universidade Federal do Rio Grande (FURG), a disciplina é ministrada pela professora Hélida Salles em conjunto com o professor Giancarlo Lucca, e tem como objetivo realizar uma apresentação de como se aplicar um classificador difuso em ferramentas que aprendizado de maquina, como Keel, Weka, Scikit-Learn ou outras ferramentas a escolha do aluno. Para a realização desse trabalho, foi escolhido a ferramenta Keel [**?**], a escolha dessa ferramenta se deu pelo fato de que entre as citadas, essa era a única que eu não conhecia, e por ter o seu código aberto e de fácil edição dos algoritmos fornecidos, se tornou um fato relevante na escolha da mesma.

O presente trabalho busca realizar além de uma apresentação de como se aplicar um classificador difuso, um comparativo aos resultados referentes a aplicação de um treinamento do modelo em R, com a ferramenta Keel sem a utilização de um sistema de classificação difuso utilizando uma validação cruzada e por fim os resultados referentes ao mesmo dataset utilizando o sistema de classificação difuso FURIA [**?**].

2 Dataset Fifa

O dataset Fifa, foi retirado de uma base de dados da plataforma "sofifa", esse dataset se refere as informações referentes aos jogadores de futebol do jogo "FIFA 2018", o mesmo contem 88 atributos referentes a cada um dos 18207 jogadores. O objetivo proposto é que a partir do dataset em questão, se possa estimar o atributo "Value".

3 Pré-processamento

Inicialmente o dataset é muito grande, ele tem dimensão inicial de 18208 x 84, e com dados fora de padrão, e também com muito valores faltantes, fazendo necessária a redução de atributos e dimensionalidade do modelo.

Para isto, foi excluído todas as instancias de teste em que tinham valores faltantes, além de exclusão de atributos em que não eram relevantes para o modelo, tais como, "ID, PHOTO, Nationality, Flag, logo, ..." dentre outros. Além disto, também foi padronizados valores do dataset como no lugar de K foi substituído por 1000, e no lugar de M foi substituído por 1000000, além da retirada de caracteres não benéficos para meu modelo. Com isto foi reduzido um total de 40 atributos, tornando meu dataset 17605 x 44, ainda continua muito grande. Para tratar meu atributo "alvo", primeiramente foi verificado que ele possuem valores numéricos, então foi realizado a discretização dos valores, para que ele tenha classes na qual se possa classificar, as classes foram definidas como "Entre 5700 e 8500, Entre 8500 e 11500, Maior que 11500 e Menor que 5700". Tem muitas instancias de jogadores com baixo valor e poucos com valores altos, então meus jogadores com valor alto seriam os dados discrepantes, mas eu não quero retirar eles, quero que ele consiga classificar, e com regressão, a curva de aprendizado dele seria linear, e quando chegasse em valores altos a curva de aprendizado não iria conseguir acompanhar o salto do gráfico pra cima, então se eu categorizar meus *"Value"* eu consigo fazer um aprendizado mais abrangente (por mais que eu diminua minhas possibilidades).

Ainda assim, o meu dataset continua muito grande, então foi realizada uma analise mais detalhada, e verifiquei como se comportavam os valores baixos, percebi que todos as instancias que tinha o valor menor que 3000 eram muito parecidas, praticamente iguais, mudavam pouca coisa, então eu decidi que iria excluir essas instancias, me restaram 1583 instancias, então eu reduzi 91% do meu dataset. Então meu dataset ficou 1582x44.

4 Resultados

Foi realizado o treinamento do meu modelo em R, e na Tabela 1 podemos visualizar a matriz de confusão. Na Figura 1 podemos ver as medidas de qualidade do modelo e na Figura 2 se observa as medidas de qualidade entre as classes.

Table 1: **Matriz de Confusão do modelo em R**

Predição	Menor que 5700	Entre 5700 e 8500	Entre 8500 e 11500	Maior que 11500
Menor que 5700	133	6	0	0
Entre 5700 e 8500	1	100	9	0
Entre 8500 e 11500	1	10	22	3
Maior que 11500	8	8	13	82

Figure 1: Medidas de Qualidade

```
                Accuracy : 0.851
                  95% CI : (0.8121, 0.8846)
    No Information Rate : 0.3611
    P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.7919

 Mcnemar's Test P-Value : 0.0001529
```

Figure 2: Medidas de Qualidade das Classes

```
Statistics by Class:

                     Class: Entre 5700 e 8500 Class: Entre 8500 e 11500
Precision                            0.9091                     0.61111
Recall                               0.8065                     0.50000
F1                                   0.8547                     0.55000
Prevalence                           0.3131                     0.11111
Detection Rate                       0.2525                     0.05556
Detection Prevalence                 0.2778                     0.09091
Balanced Accuracy                    0.8848                     0.73011
                     Class: Maior que 11500 Class: Menor que 5700
Precision                            0.7387                0.9568
Recall                               0.9647                0.9301
F1                                   0.8367                0.9433
Prevalence                           0.2146                0.3611
Detection Rate                       0.2071                0.3359
Detection Prevalence                 0.2803                0.3510
Balanced Accuracy                    0.9357                0.9532
```

Na tabela 2 encontra-se a matriz de confusão das predições realizadas pelo "Keel" com o algoritmo "Public-C". Além disto, temos como acurácia do modelo 0.8571371706.

Table 2: **Matriz de Confusão do modelo Public-C**

Predição	Menor que 5700	Entre 5700 e 8500	Entre 8500 e 11500	Maior que 11500
Menor que 5700	575	35	0	24
Entre 5700 e 8500	40	356	19	20
Entre 8500 e 11500	1	18	146	32
Maior que 11500	0	6	31	279

E por fim, com base na tabela 3 encontra-se a matriz de confusão das predições realizadas pelo "Keel" com o algoritmo de classificação fuzzy "FURIA". Além disto, temos como acurácia do modelo 0.8659581244.

Table 3: **Matriz de Confusão do modelo FURIA**

Predição	Menor que 5700	Entre 5700 e 8500	Entre 8500 e 11500	Maior que 11500
Menor que 5700	596	15	1	22
Entre 5700 e 8500	35	357	28	15
Entre 8500 e 11500	3	37	124	33
Maior que 11500	6	3	14	293