Determinação de Classes e Critérios de Classificaçaõ de Solo - Profundidade B

Joel Maurício Corrêa da Rosa

1 de fevereiro de 2007

Neste relatório é apresentada uma seqüência de operações no software R feita com o objetivo de primeiramente criar uma classificação para diferentes observações de solo e apresentar estas regras de classificação através de uma árvore de decisão binária.

1 Métodos de Clusterização

Algoritmos de clusterização, ou agrupamento, são métodos utilizados para dividir um conjunto de n objetos em k grupos que sejam internamente homogêneos e heterogêneos entre si em relação à um conjunto de variáveis. Estes algoritmos são divididos em duas classes: hierárquicos e não-hierárquicos.

Neste trabalho são aplicadas duas técnicas de clusterização; uma do tipo *crisp* na qual o conjunto de dados é particionado em k subconjuntos e cada objeto deve pertencer somente a um dos subconjuntos e outra técnica, classificada como *fuzzy*,em que cada objeto pertence a todos os k grupos, mas com diferentes graus de pertinência.

Nos algoritmos do tipo *fuzzy*, os objetos pertencem a todos os grupos, mas com diferentes graus de pertinência. Desta forma, ao objeto são associados o conjunto e o seu grau de pertinência àquele conjunto.

Os agrupamentos são formados mediante a avaliação de uma medida de distância ou dissimilaridade.

2 Agrupamentos no Solo Segundo Atributos Químicos

A análise a ser apresentada nesta seção refere-se à formação de agrupamentos no solo com base em vetores observados do solo para 9 variáveis 'químicas'.

Os comandos abaixo, procedem a leitura dos dados e suas respectivas coordenadas para o processamento da análise de agrupamentos.

```
> rm(list = ls())
> dados.ori <- read.table("http://leg.ufpr.br/~joel/pesquisa/dados/Quimicos_C2.txt",
+     h = T, sep = "", dec = ",")
> names(dados.ori)
```

```
 [1] "X"      "Y"      "Al"     "Ca"     "Mg"     "PhH2O"  "PhKCL"  "K"
 [9] "P"      "C"      "AB"     "H"      "Prof"   "Regiao"
```

```
> head(dados.ori)
```

```
       X       Y  Al   Ca   Mg PhH2O PhKCL    K   P    C AB   H Prof Regiao
1 688713 7190023 2.0 1.55 1.51  5.15  4.20 0.06 0.4 12.4  2 3.9   22      1
2 688694 7190000 3.5 1.05 1.06  5.10  4.30 0.10 0.2 16.0  2 6.2   35      1
3 688457 7189578 1.4 0.85 0.41  5.30  4.25 0.04 0.2  8.8  2 2.7   38      3
4 688790 7190115 6.0 0.23 0.03  4.85  4.05 0.08 0.4  6.9  2 2.0   40      1
5 688592 7189756 3.0 0.61 0.62  5.10  4.25 0.04 0.6 12.4  2 4.7   42      3
6 688601 7189843 2.3 1.13 0.98  5.25  4.20 0.03 0.2 10.6  2 4.6   45      2
```

Para evitar problemas em função das diferentes escalas de medição das variáveis químicas, abaixo é feita a padronização das variáveis, subtraindo-as da média observada e dividindo pelos respectivos desvios-padrão.

Ressalta-se aqui que alguns pacotes do R para clusterização contém em suas funções, parâmetros para a padronização das variáveis.

```
> DadosClust <- dados.ori[, c(3:10, 12)]
> medias <- apply(DadosClust, 2, mean)
> desvpad <- apply(DadosClust, 2, sd)
> DadosClust.p <- DadosClust
> for (i in 1:9) {
+     DadosClust.p[i] <- DadosClust.p[i] - medias[i]
+ }
> for (i in 1:9) {
+     DadosClust.p[i] <- DadosClust.p[i]/desvpad[i]
+ }
```

Conforme pode ser verificado, através da matriz de correlações, há indícios de fortes associações entre variáveis químicas tal como ocorre entre Alumínio (Al) e Potássio (K) cujo coeficiente de correlação linear é igual a -0,7464

```
> cor(DadosClust)
```

	Al	Ca	Mg	PhH2O	PhKCL	K
Al	1.00000000	-0.145865155	-0.16177153	-0.41532358	-0.79025474	0.160913458
Ca	-0.14586515	1.000000000	0.50256315	0.12357576	0.14640758	-0.004615275
Mg	-0.16177153	0.502563149	1.00000000	0.12857666	0.21424913	0.018640192
PhH2O	-0.41532358	0.123575762	0.12857666	1.00000000	0.41179120	0.049801649
PhKCL	-0.79025474	0.146407580	0.21424913	0.41179120	1.00000000	0.022025355
K	0.16091346	-0.004615275	0.01864019	0.04980165	0.02202536	1.000000000
P	0.11018646	0.086904836	0.12811349	-0.09723479	-0.14291278	0.129478407
C	-0.02592455	0.045412441	0.04489215	0.13487250	0.02944880	-0.001993542
H	-0.23320377	0.146490918	0.12154734	0.28470817	0.22827132	0.011006393

	P	C	H
Al	0.110186462	-0.025924553	-0.23320377
Ca	0.086904836	0.045412441	0.14649092
Mg	0.128113493	0.044892149	0.12154734
PhH2O	-0.097234791	0.134872497	0.28470817
PhKCL	-0.142912785	0.029448802	0.22827132
K	0.129478407	-0.001993542	0.01100639
P	1.000000000	0.004206076	0.01568585
C	0.004206076	1.000000000	0.45237489
H	0.015685850	0.452374893	1.00000000

Abaixo, são carregados 3 pacotes no R para utilização de algoritmos de clusterização, definido o número de clusters e também as cores que representarão os clusters na visualização espacial.

```
> require(flexclust)

[1] TRUE

> require(cclust)

[1] TRUE

> require(cluster)

[1] TRUE

> numclust = 3
> ind.c <- colors()[c(562, 45, 499, 10, 614)]
> ind.c

[1] "royalblue"    "cadetblue3"   "orange1"      "aquamarine2"  "springgreen4"
```

2.1 Clusterização Fuzzy Através da Função fanny

Utiliza-se abaixo a função fanny presente no pacote cluster para a determinação dos clusters nebulosos.

Em função da expectativa, anterior a coleta de dados, em relação à quantidade de diferentes classes de solo na região da coleta de dados, foi feita a opção inicial por K=3 agrupamentos.

Arbitrariamente foi definido o parâmetro memb.exp que determina o grau de nebulosidade da clusterização.

```
> clust.info.fuz <- fanny(DadosClust.p, numclust, memb.exp = 1.3,
+     maxit = 3000, tol = 1e-20)
> clust.fuz <- clust.info.fuz$clustering
> max.memb <- apply(clust.info.fuz$membership, 1, max)
```

A partir dos graus de pertinência, é possível formar agrupamentos, alocando as observações ao grupo ao qual está associado o seu maior grau de pertinência. Veja, por exemplo, os grau de pertinência da observação 48 aos K=3 diferentes clusters.

```
> memb.48 <- clust.info.fuz$membership[48, ]
> memb.48

[1] 0.19956142 0.71595313 0.08448546

> clust.48 <- which(memb.48 == max(memb.48))
> clust.48

[1] 2
```

Conforme a seqüência de pertinências mostrada acima, a observação 48 deve ser alocada ao cluster 2 que detém a maior pertinência (0.716). De acordo com este procedimento de alocação das observações aos grupos, abaixo mostra-se a distribuição de freqüências dos agrupamentos.

```
> table(clust.info.fuz$clustering)

 1  2  3
44 42 47

> DadosClust.p$fuzclust <- factor(clust.info.fuz$clustering)
> head(DadosClust.p)
          Al         Ca          Mg      PhH2O       PhKCL          K
1 -0.3163010  1.1382793  2.86548522 -0.1963890 -0.59685159  0.1230950
2  0.6933319  0.2265041  1.72610647 -0.3649034 -0.03782862  0.8349054
3 -0.7201542 -0.1382059  0.08033715  0.3091543 -0.31734010 -0.2328102
4  2.3760533 -1.2688071 -0.88180491 -1.2074754 -1.43538603  0.4790002
5  0.3567876 -0.5758580  0.61204724 -0.3649034 -0.31734010 -0.2328102
6 -0.1143745  0.3723882  1.52355024  0.1406398 -0.59685159 -0.4107627
          P           C           H fuzclust
1  0.1899219  0.24588662 -0.02551318        1
2 -0.4187435  0.70220643  1.21329361        1
3 -0.4187435 -0.21043320 -0.67184716        2
4  0.1899219 -0.45126865 -1.04887531        3
5  0.7985873  0.24588662  0.40537614        3
6 -0.4187435  0.01772671  0.35151497        1

> means <- by(DadosClust.p[, -10], DadosClust.p$fuzclust, mean)
> centroid <- NULL
> for (i in 1:numclust) {
+     centroid <- rbind(centroid, means[[i]])
+ }
```

2.2 Clusterização Crisp através do Método das K-Médias

Utiliza-se abaixo o comando `kcca` para a determinação dos clusters crisps. O algoritimo das k-médias é um procedimento iterativo que necessita a definição de valores iniciais para os centróides de cada cluster. A escolha por estes valores iniciais foi feita com base nos resultados obtidos através da fuzzy-clusterização. Deste modo, os centróides obtidos ao final do processo anterior serão utilizados como valores iniciais neste algoritmo.

```
> require(cclust)

[1] TRUE

> clust.info.crisp <- kcca(DadosClust.p[, -10], centroid)
> clust.info.crisp

kcca object of family 'kmeans'

call:
kcca(x = DadosClust.p[, -10], k = centroid)

cluster sizes:

 1  2  3
36 50 47

> clust.crisp <- predict(clust.info.crisp)
```

Para entender melhor a característica de cada cluster, na sequência é apresentado, através de um gráfico de barras, os centróides de cada cluster.

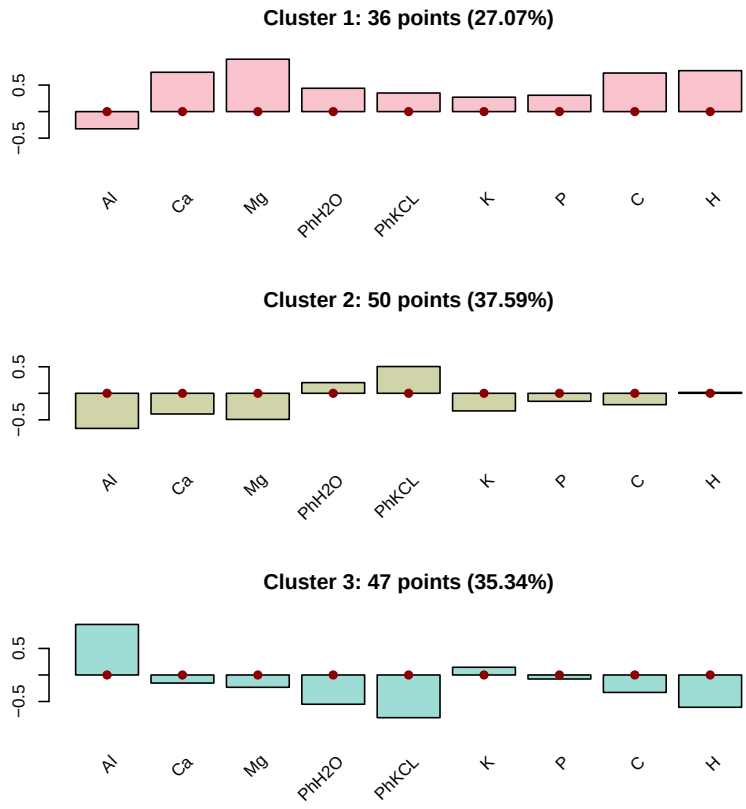
Para avaliar a robustez da média como medida de informação do cluster, repete-se abaixo o procedimento com a mediana.

```
> clust.info.crisp.md <- kcca(DadosClust.p[, -10], centroid, family = kccaFamily("kmedians"))
> clust.info.crisp.md

kcca object of family 'kmedians'

call:
kcca(x = DadosClust.p[, -10], k = centroid, family = kccaFamily("kmedians"))

cluster sizes:

 1  2  3
41 45 47

> clust.crisp.md <- predict(clust.info.crisp.md)
```

Para entender melhor a característica de cada cluster, na sequência é apresentado, através de um gráfico de barras, os centróides de cada cluster.

```
> table(clust.crisp, clust.crisp.md)

           clust.crisp.md
clust.crisp  1  2  3
          1 30  1  5
          2  8 42  0
          3  3  2 42
```

```
> barplot(clust.info.crisp)
```



3 Distribuição Espacial dos Agrupamentos

3.1 Distribuição Espacial dos Agrupamentos Nebulosos

O mesmo processo anterior é repetido para os clusters nebulosos.

```
> coord <- dados.ori[, 1:2]
> DadosClustGeo.fuz <- data.frame(coord, clust.fuz)
> head(DadosClustGeo.fuz)
```

```
      X       Y clust.fuz
1 688713 7190023         1
2 688694 7190000         1
3 688457 7189578         2
4 688790 7190115         3
5 688592 7189756         3
6 688601 7189843         1
```

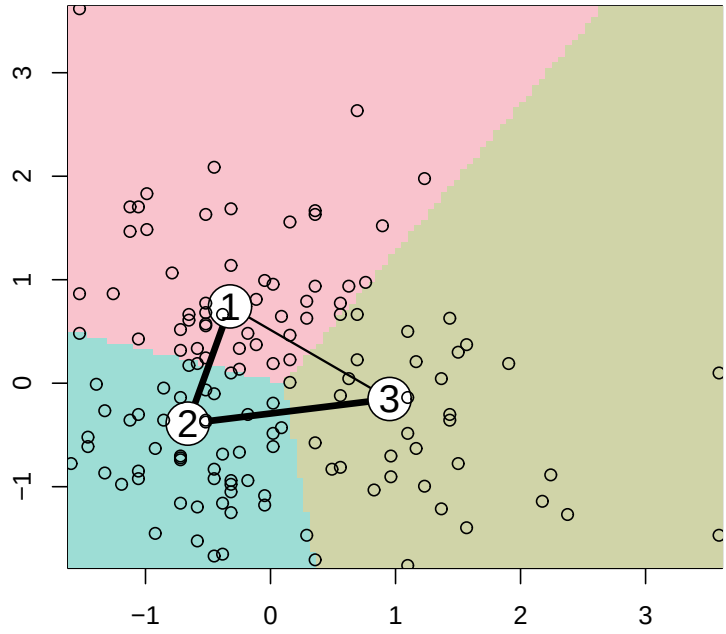
Carregando o pacote geoR

```
> require(geoR)
```

```
[1] TRUE
```

```
> ViewClust.fuz <- as.geodata(DadosClustGeo.fuz, coords.col = 1:2,
+     dados.col = 3)
```

```
> image(clust.info.crisp)
> points(DadosClust.p[, -10])
```



3.2 Distribuição Espacial dos Agrupamentos Crisp

Abaixo é criado um *data frame* com as coordenadas e os respectivos clusters obtidos pelo método das k-médias.

```
> coord <- dados.ori[, 1:2]
> DadosClustGeo <- data.frame(coord, clust.crisp)
> head(DadosClustGeo)
```

```
       X       Y clust.crisp
1 688713 7190023           1
2 688694 7190000           1
3 688457 7189578           2
4 688790 7190115           3
5 688592 7189756           3
6 688601 7189843           1
```

```
> require(geoR)
```

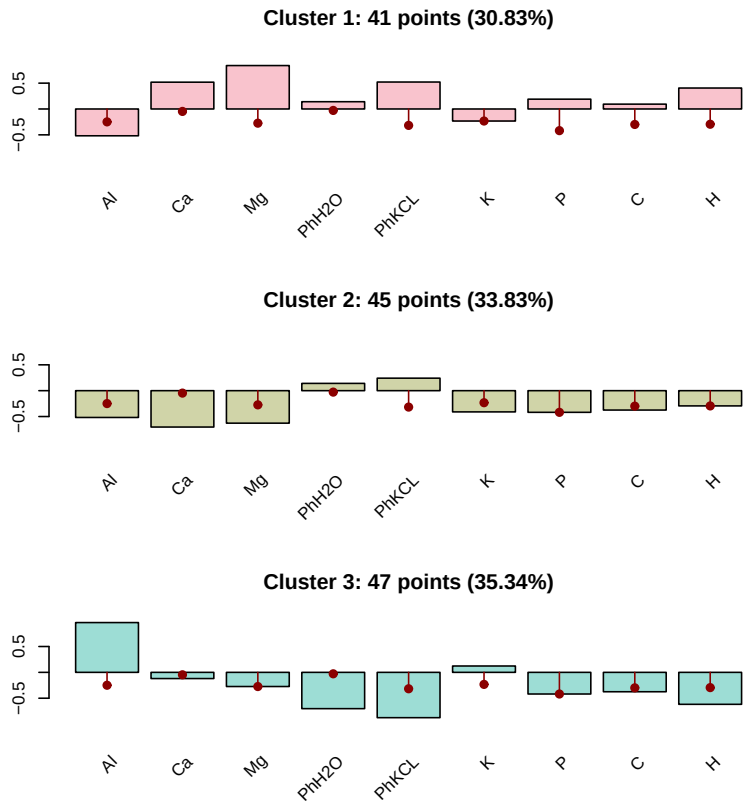
```
[1] TRUE
```

```
> ViewClust <- as.geodata(DadosClustGeo, coords.col = 1:2, dados.col = 3)
```

4 Extração das Regras de Clusterização

Nesta seção, as classes determinadas pela análise de cluster serão assumidas como verdadeiras e utilizadas como níveis de uma variável dependente em uma árvore de classificação.

```
> barplot(clust.info.crisp.md)
```



Com este procedimento, o objetivo é extrair as regras lógicas que determinaram a formação dos agrupamentos.

Carregando os pacotes para a construção de árvores de classificação.

```
> require(tree)
```

```
[1] TRUE
```

```
> require(maptree)
```

```
[1] TRUE
```

4.1 Extração das Regras para Agrupamentos Nebulosos

```
> d.fuz <- data.frame(DadosClust, Cluster = clust.fuz)
> tree.fuz <- rpart(factor(Cluster) ~ ., data = d.fuz)
> tree.fuz
```

```
n= 133

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 133 86 3 (0.33082707 0.31578947 0.35338346)
   2) Al< 2.15 68 30 2 (0.42647059 0.55882353 0.01470588)
     4) Mg>=0.36 27  4 1 (0.85185185 0.11111111 0.03703704) *
```

```
> plot(ViewClust.fuz$coord)
> for (i in 1:numclust) {
+     i.c = ind.c[i]
+     ViewClust.fuz$data <- clust.fuz == i
+     points(ViewClust.fuz, pt.div = "data.proportional", col = i.c,
+         add = T)
+ }
```

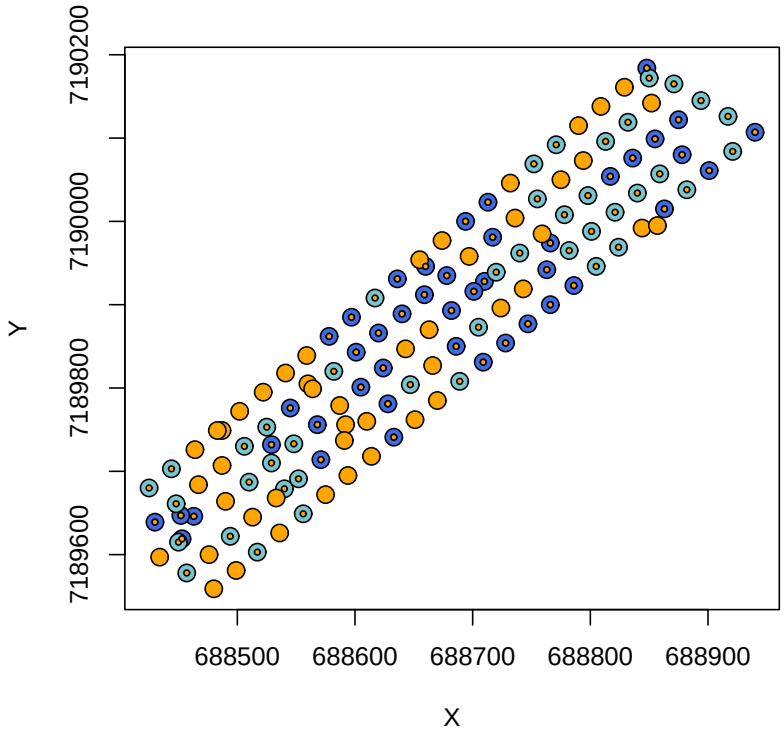
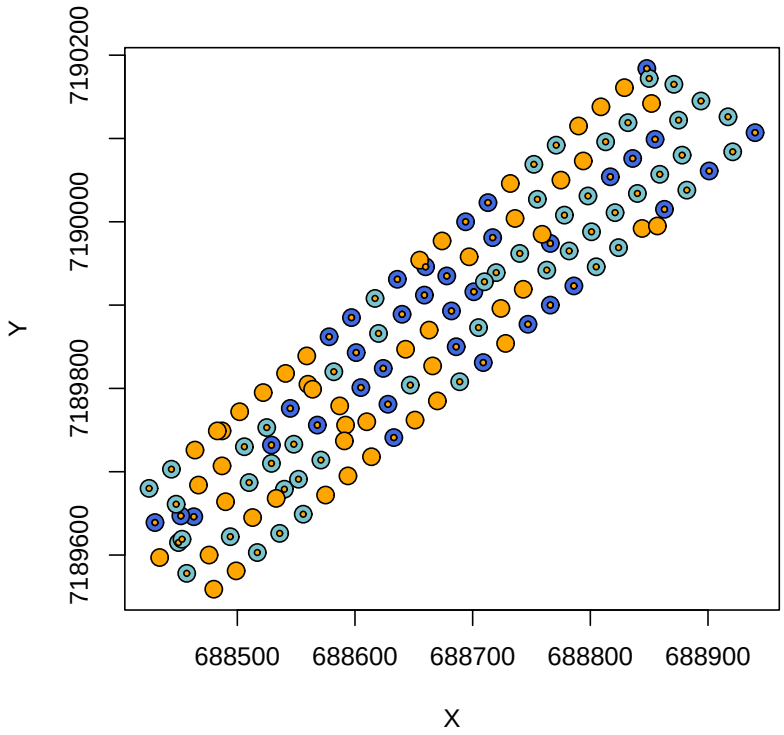


Figura 1: Distribuição Espacial dos Agrupamentos Nebulosos

```
> plot(ViewClust$coord)
> for (i in 1:numclust) {
+     i.c = ind.c[i]
+     ViewClust$data <- clust.crisp == i
+     points(ViewClust, pt.div = "data.proportional", col = i.c,
+         add = T)
+ }
```




```
    5) Mg< 0.36 41  6 2 (0.14634146 0.85365854 0.00000000) *
  3) Al>=2.15 65 19 3 (0.23076923 0.06153846 0.70769231)
    6) Mg>=0.685 11  1 1 (0.90909091 0.00000000 0.09090909) *
    7) Mg< 0.685 54  9 3 (0.09259259 0.07407407 0.83333333)
     14) C>=13.6 7  2 1 (0.71428571 0.14285714 0.14285714) *
     15) C< 13.6 47  3 3 (0.00000000 0.06382979 0.93617021) *
```

4.2 Extração das Regras para Agrupamentos Crisp

```
> DadosTreeCrisp <- DadosClust
> DadosTreeCrisp$regiao <- factor(dados.ori$Regiao)
> DadosTreeCrisp$prof <- dados.ori$Prof
> d.crisp <- data.frame(DadosTreeCrisp, Cluster = clust.crisp)
> tree.cl <- rpart(factor(Cluster) ~ ., data = d.crisp)
> tree.cl

n= 133
```

```
> draw.tree(tree.fuz, nodeinfo = T)
```

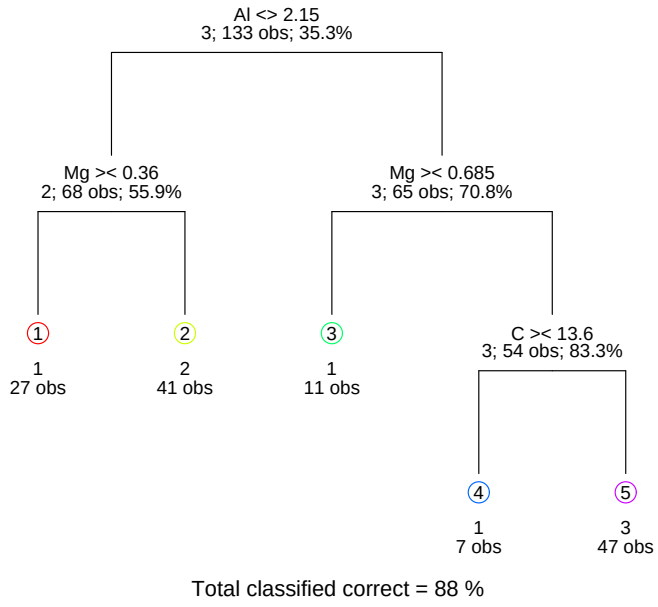


Figura 2: Árvore de Classificação para os Grupos Nebulosos

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 133 83 2 (0.27067669 0.37593985 0.35338346)
   2) Al< 2.15 68 23 2 (0.32352941 0.66176471 0.01470588)
     4) Mg>=0.385 26  8 1 (0.69230769 0.26923077 0.03846154)
       8) H>=3.45 18  1 1 (0.94444444 0.05555556 0.00000000) *
       9) H< 3.45 8  2 2 (0.12500000 0.75000000 0.12500000) *
     5) Mg< 0.385 42  4 2 (0.09523810 0.90476190 0.00000000) *
   3) Al>=2.15 65 19 3 (0.21538462 0.07692308 0.70769231)
     6) H>=4.8 11  3 1 (0.72727273 0.27272727 0.00000000) *
     7) H< 4.8 54  8 3 (0.11111111 0.03703704 0.85185185)
      14) Mg>=0.685 7  2 1 (0.71428571 0.00000000 0.28571429) *
      15) Mg< 0.685 47  3 3 (0.02127660 0.04255319 0.93617021) *
```

```
> draw.tree(tree.cl, nodeinfo = T)
```

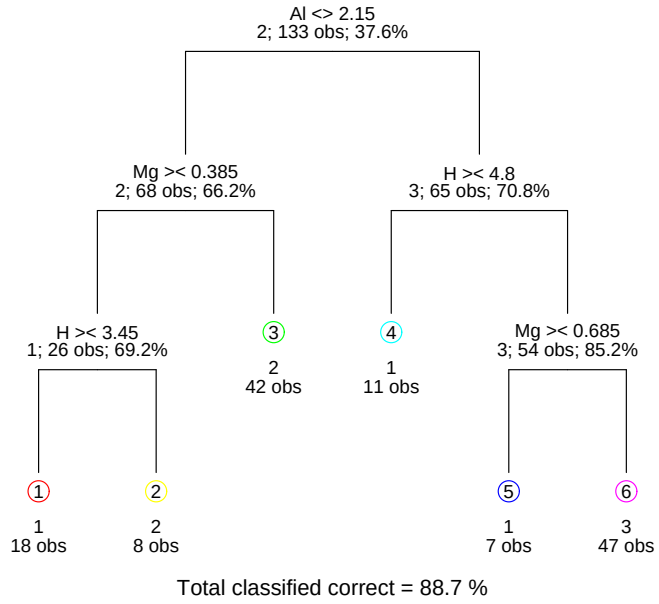


Figura 3: Árvore de Classificação para os Grupos Crisp

5 Estatísticas Descritivas dos Clusters

Cluster Crisp x Cluster Fuzz

```
> table(clust.fuz, clust.crisp)
```

```
         clust.crisp
clust.fuz  1  2  3
        1 36  7  1
        2  0 42  0
        3  0  1 46
```

Cluster x Regiao

```
> table(dados.ori$Regiao, clust.crisp)
```

```
   clust.crisp
     1  2  3
  1 11 25 13
  2 15  5  8
  3 10 20 26
```