# AggloCluster –Manual

Este manual apresenta uma descrição das funcionalidades do sistema computacional AggloCluster. O sistema disponibiliza a implementação do esquema *Matrix Updating Algorithmic Scheme* (MUAS), parametrizável para 4 estratégias de agrupamento (*Single Linkage*, *Complete Linkage*, UPGMA e WPGMA) bem como uma implementação do algoritmo AGNES, adaptada para induzir um número k de grupos. A versão clássica do AGNES [Kaufman & Rousseeuw 2005] implementa apenas a estratégia *Average Linkage* UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*) que induz um agrupamento hierárquico com o maior número possível de grupos. A versão do AGNES disponibilizada no AggloCluster, entretanto, contempla as quatro estratégias mencionadas anteriormente.

O AggloCluster é composto pelos módulos de pré-processamento (painel *Preprocess*), agrupamento (painel *Cluster*) e validação (painel *Cluster Validity*). O desenvolvimento do sistema foi fundamental para o melhor entendimento de como funcionam os algoritmos hierárquicos aglomerativos, além de oferecer um ambiente computacional para a investigação e realização de experimentos. O sistema foi projetado para ser executado como um aplicativo *Windows Form* na plataforma Microsoft Windows, e a linguagem C# em conjunto com o .NET Framework 4.0 foram utilizados para sua implementação. A seguir é apresentada sua arquitetura funcional e a descrição de cada um dos módulos.

## 5.1 Módulo de Pré-processamento (painel *Preprocess*)

O painel de pré-processamento (*Preprocess*) do AggloCluster permite a importação do conjunto de padrões por meio da leitura de arquivos texto em formato ARFF (*Attribute-Relation File Format*) e a visualização dos padrões em duas dimensões (se os padrões possuírem mais de duas dimensões, então são plotados os dois primeiros atributos de cada padrão) no plano cartesiano, como mostrado na Figura 5.1.

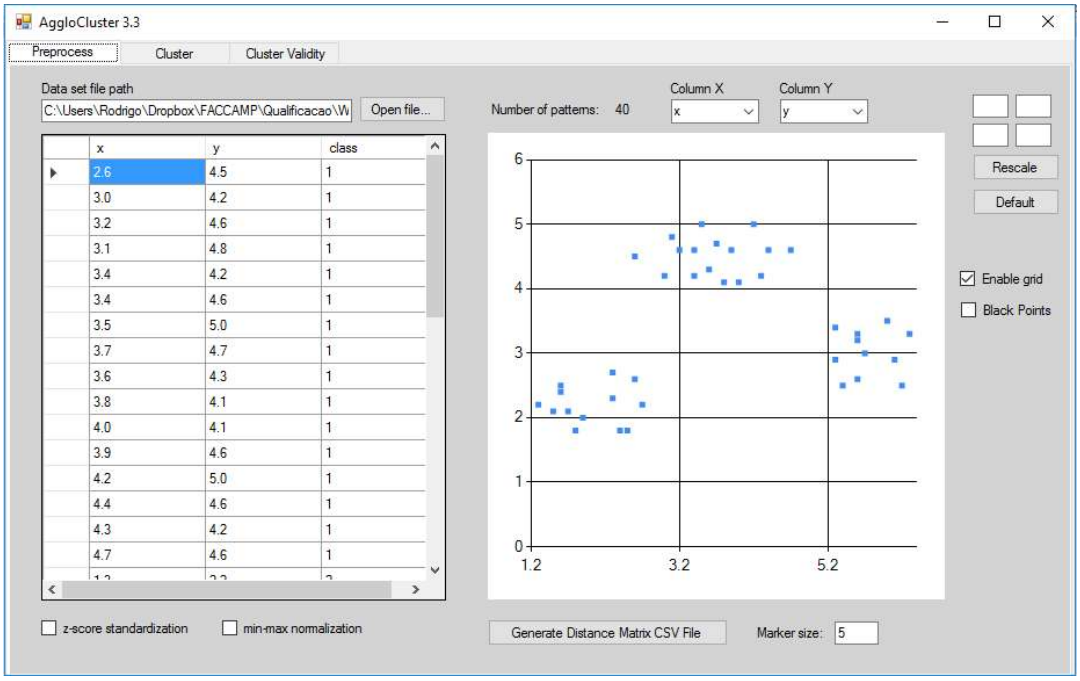


**Figura 5.1** Tela de Pré-processamento do AggloCluster.

Como pode ser observado no canto inferior esquerdo da Figura 5.1, existem duas opções para normalização de padrões. O *checkbox z-score standardization* permite realizar a transformação dos dados por meio de uma técnica conhecida como padronização. Outra técnica disponível para transformação dos dados que realiza a normalização dos valores de atributos trazendo-os para uma determinada faixa (*feature scaling*) é conhecida como Min-Max e pode ser acessada pelo *checkbox min-max normalization*. As técnicas de padronização e normalização aqui discutidas foram apresentadas no Capítulo 2.

## 5.2 Módulo de Agrupamento (painel *Cluster*)

Este módulo disponibiliza a implementação do MUAS e, também, as estratégias de agrupamento citadas anteriormente. Uma vez selecionada a estratégia de agrupamento e acionado o botão *Start Clustering*, o sistema executa o algoritmo e plota o seu resultado em uma árvore (*treeview*) com a hierarquia de grupos como mostra a Figura 5.2. A representação visual do agrupamento hierárquico na forma de um

dendrograma pode ser obtida acionando o botão *Show Dendrogram*, como mostra a Figura 5.3.

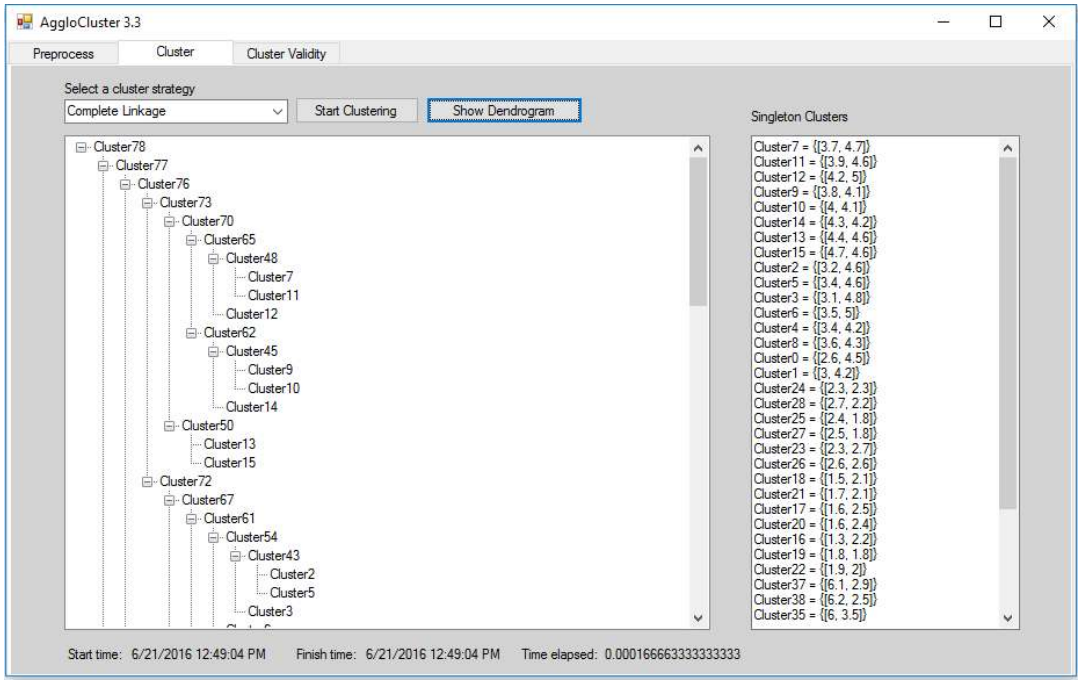


**Figura 5.2** Resultado da execução do MUAS utilizando como estratégia de agrupamento o *Complete Linkage*.

É importante observar que o AggloCluster somente habilita o painel *Cluster* se um conjunto de padrões foi previamente importado para o sistema por meio do painel *Preprocess*. O mesmo comportamento ocorre com o painel *Cluster Validity* discutido na Seção 5.3.

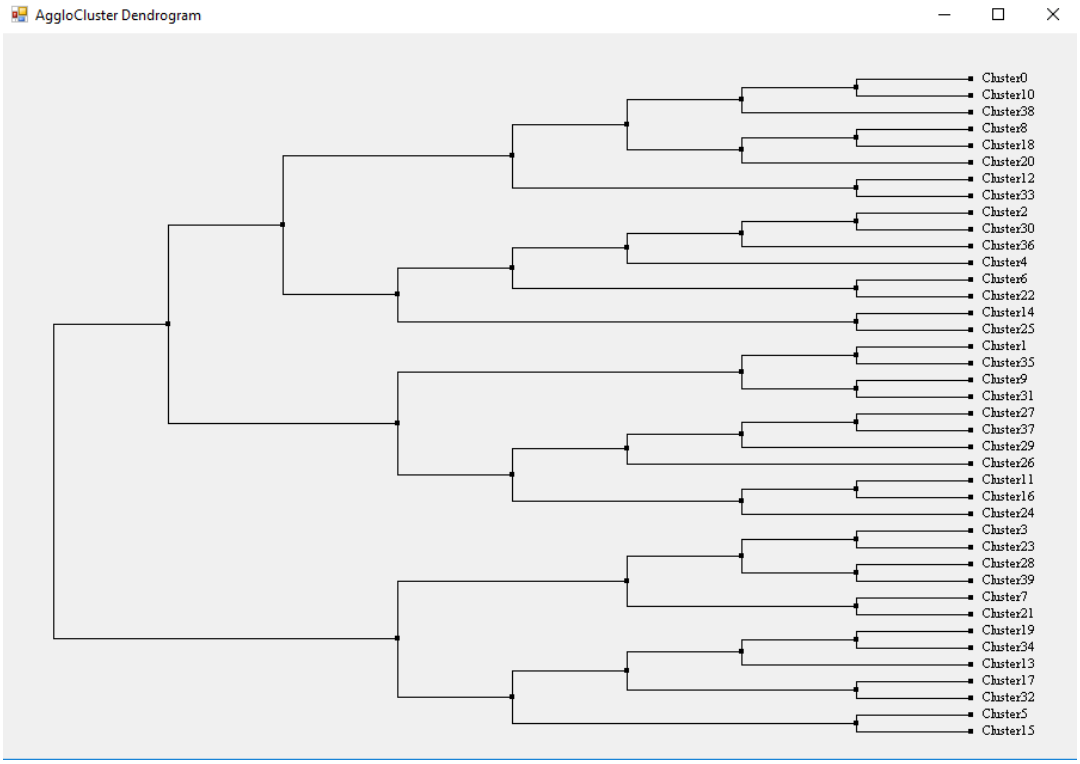


**Figura 5.3** Dendrograma resultante da execução do MUAS utilizando o método *Complete Linkage* como estratégia de agrupamento.

# 5.3 Módulo de Validação (Painel *Cluster Validity*)

O painel *Cluster Validity* disponibiliza recursos para realizar a validação interna de agrupamentos que já foram induzidos. Neste módulo estão disponíveis os índices de validação interna (índice de Dunn e índice Davies-Bouldin) e de validação externa (índice de Rand e índice de Jaccard) como mostra a Figura 5.4.

O módulo de validação disponibiliza, também, o algoritmo particionante K-Means, apresentado no Capítulo 4, cujos agrupamentos por ele produzido servirão de *baseline* na comparação dos resultados produzidos pelo algoritmos aglomerativos. A Figura 5.5 mostra o resultado do K-Means com a formação de 3 grupos (*Cluster0*, *Cluster1* e *Cluster2*) para o mesmo conjunto de padrões mostrado na Figura 5.1. Note que este módulo não se restringe apenas a validação, mas permite que os agrupamentos resultantes dos algoritmos sejam comparados.

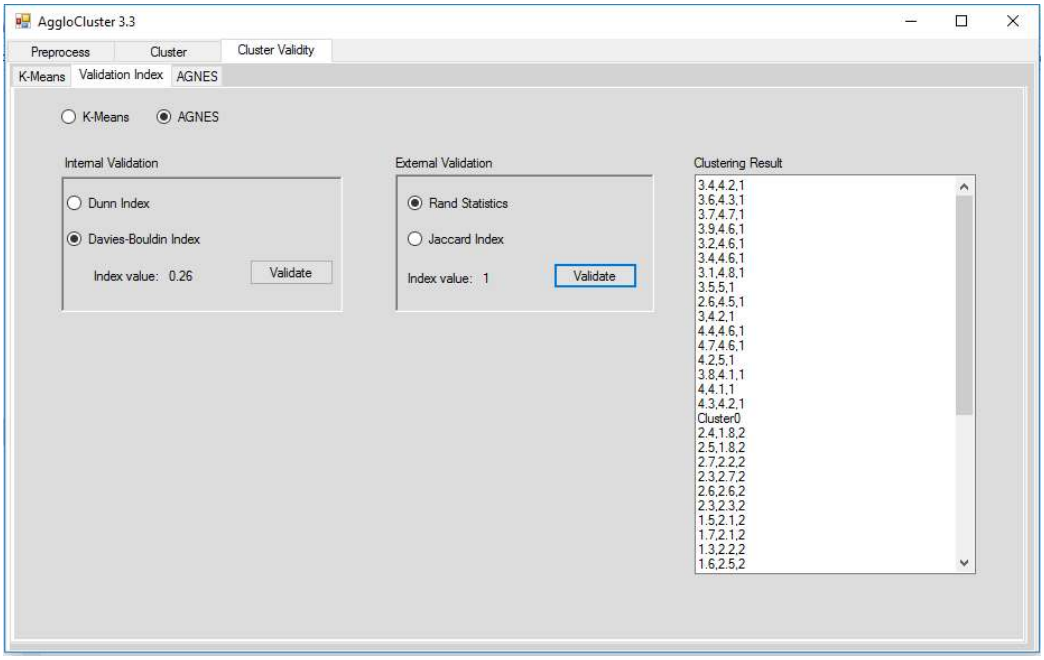


**Figura 5.4** Resultado do cálculo do índice de Davies-Bouldin e índice de Rand para o agrupamento resultante do AGNES.

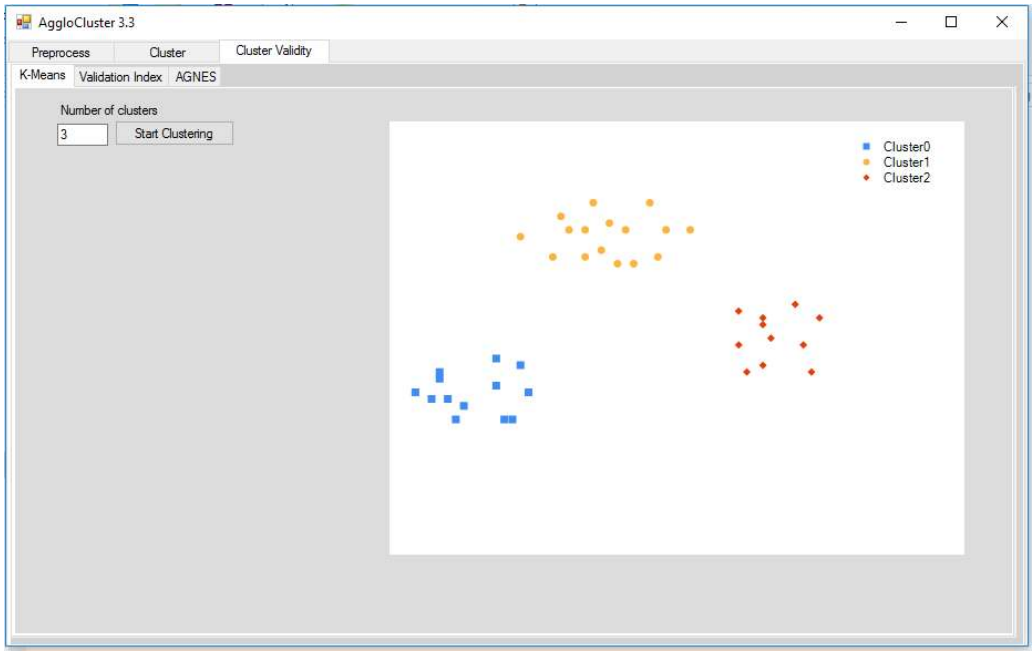


**Figura 5.5** Resultado do algoritmo K-Means com a formação de 3 grupos (*Cluster0*, *Cluster1* e *Cluster2*).

O algoritmo AGNES [Kaufman & Rousseeuw 2005], algoritmo aglomerativo baseado no esquema MUAS, foi implementado, mas não em sua forma clássica, ou seja, o ponto de parada do algoritmo é uma quantidade k de grupos informada e não um único grupo contendo todos os demais subgrupos. A Figura 5.6 mostra o resultado do agrupamento produzido pelo AGNES em que foi informada a quantidade 3 de grupos e a estratégia *Complete Linkage*.

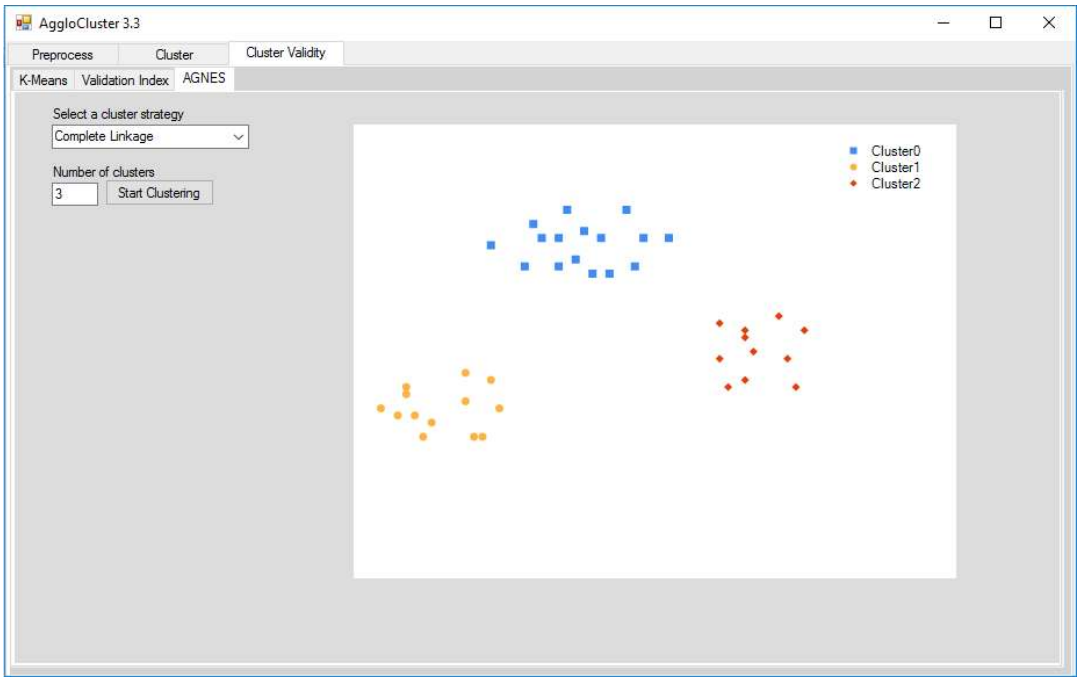


**Figura 5.6** Execução do algoritmo AGNES resultando na formação de 3 grupos (*Cluster0*, *Cluster1* e *Cluster2*).

É importante notar que tanto o esquema MUAS quanto o algoritmo AGNES realizam a tarefa de agrupamento criando uma estrutura hierárquica de grupos aninhados. Entretanto, para que os grupos produzidos pelo AGNES possam ser comparados com os grupos produzidos pelo K-Means, foi necessário adaptar a saída do algoritmo AGNES para transformar a hierarquia de grupos (e.g., como mostra a Figura 5.3) em grupos planos (*flat clusters*), ou seja, sem sobreposição de grupos (e.g., como mostra a Figura 5.6).