Investigação da aplicação de algoritmos de agrupamento para o problema astrofísico de classificação de galáxias

Vanessa de Oliveira Gil [1]
Fabricio Ferrari [2]
Leonardo Emmendorfer [3]

Resumo: O surgimento de grandes bancos de dados dinamizou o armazenamento de registros e sua busca. Sendo assim, está à disposição um número exorbitante de informações que precisam ser analisadas. Para que seja possível explorá-las em tempo hábil, é utilizado o processo de mineração de dados, que os analisa sob diferentes perspectivas e permite a descoberta automática de informações e padrões, além de fornecer a capacidade de previsão de uma observação futura. Pretende-se explorar bases de dados astronômicos com parâmetros morfométricos de galáxias, por meio de algoritmos de agrupamento, a fim de identificar padrões naturais de agrupamentos como etapa anterior à classificação de galáxias. Para isso, serão aplicados os algoritmos Expectation Maximization (EM) e K-médias, que serão avaliados quando sujeitos a dados morfométricos reais do catálogo Extraction de Formes Idealisées de Galaxies en Imagerie (EFIGI) com galáxias de todos os tipos morfológicos, medidos pelo MORFOMETRYKA, e dados sintéticos. Após os dados serem agrupados pelos algoritmos, será utilizado o algoritmo Silhouette como método de validação para os resultados encontrados. Nesse espaço de parâmetros morfométricos serão detectadas classes de galáxias, permitindo o resultado estudar a continuidade morfométrica nas populações de galáxias espirais e elípticas.

***Palavras-chave*:** Análise de agrupamentos. Galáxia. Mineração de dados.

Abstract: *The emergence of large databases streamlined storage and search of records so there is a large quantity information that have to be analyzed. For the exploration of such data in a timely manner it's used the process of data mining to analyze data from different perspectives and allows automatic discovery of patterns and information. Besides it provides the ability to predict a future observation. It is intended to explore astronomical databases with morphometric parameters of galaxies. Clustering algorithms are used to identify natural clusters and patterns as a previous step of galaxies classification. The algorithms: Expectation Maximization (EM) and K-means will be applied to synthetic and real data from EFIGI survey (Extraction de Formes Idealisées de Galaxies en Imagerie) with galaxies of all morphologycal types as measured by MORFOMETRYKA. After the data is grouped by the algorithms the silhouette is used as a method of results validation. In this morphometric space of parameters the galaxies classes will be detected. With these results it is possible to study the morphometric continuity in populations of spiral and elliptical galaxies.*

***Keywords*:** *Clustering analysis. Data minning. Galaxy.*

1 Introdução

As galáxias são constituídas principalmente de quatro componentes – estrelas, gás, poeira e matéria escura –, e existem em grandes variedades de formas e tamanhos. Define-se a forma de uma galáxia em uma imagem de

[1]Programa de Pós-Graduação em Modelagem Computacional (PPGMC), FURG, *Campus* Carreiros - Av. Itália km 8, Rio Grande (RS) - Brasil. {vanessa.gil@furg.br}
[2]Instituto de Matemática, Estatística e Física, FURG, *Campus* Carreiros - Av. Itália km 8, Rio Grande (RS) - Brasil. {fabricio@ferrari.pro.br}
[3]Centro de Ciências Computacionais (C3), FURG, *Campus* Carreiros - Av. Itália km 8, Rio Grande (RS)- Brasil. {leonardo.emmendorfer@gmail.com}

http://dx.doi.org/10.5335/rbca.2015.4653

acordo com o comprimento de onda em que é observada. Seu formato original leva em consideração o intervalo dinâmico, o nível isofotal e a resolução da imagem.

Existem diversas abordagens para o problema de classificação de galáxias: morfológica, fotométrica, colorimétrica, espectroscópica. A mais usual é a abordagem morfológica, que se utiliza do diagrama de Hubble [1], baseada em critérios qualitativos e empíricos em relação à forma, à concentração e à estrutura das galáxias. Esse Sistema de Classificação de Hubble (SCH) consiste em três sequências principais de classificação: elípticas, espirais e espirais barradas.

As galáxias elípticas são caracterizadas basicamente por apresentar em uma única componente com morfologia esferoidal ou elipsoidal. Têm sua luminosidade distribuída suavemente e não apresentam aglomerações luminosas de estrelas azuis, nem manchas de poeira que as obscureçam. As galáxias espirais apresentam uma clara estrutura espiral, além de bojo, disco, halo e braços espirais. Algumas galáxias espirais compreendem uma estrutura em formato de barra, tendo, próximo ao seu centro, muitas estrelas. Existem diferenças quanto ao tamanho do bojo e ao grau de desenvolvimento e enrolamento dos braços espirais. Sendo assim, as galáxias são subdivididas de acordo com as características dos braços e do tamanho do bojo em relação ao disco. As galáxias irregulares são desprovidas de simetria circular ou rotacional, apresentando uma estrutura irregular e caótica que parece sofrer atividades de formação estelar intensa, sendo dominada por estrelas jovens e brilhantes e nuvens de gás ionizado.

Hubble propôs uma sequência baseada na complexidade das estruturas. As elípticas são visualmente mais simples que as lenticulares, já as galáxias espirais são mais complexas. Foram descobertas propriedades que se alteram de maneira sistemática ao longo do diagrama. Dentre essas características estão: a razão massa (luminosidade) do bojo/massa do disco, a razão massa do gás/massa das estrelas, a variação do índice de cor, a composição química do meio interestelar e a taxa de formação estelar.

2 Formação e evolução das galáxias

Não são todas as galáxias que se encaixam perfeitamente no esquema de Hubble, o que torna difícil sua classificação em termos morfológicos devido à assimetria, a núcleos descentralizados, dentre outros aspectos. Nos primeiros estudos sobre classificação de galáxias, acreditava-se que existia uma evolução das galáxias espirais e, com o tempo, se transformariam em elípticas. Essa hipótese surgiu devido ao fato de as galáxias elípticas geralmente terem estrelas mais velhas, contudo, as galáxias espirais também possuem estrelas tão velhas quanto as encontradas nas elípticas. Sendo assim, as galáxias elípticas formaram suas estrelas em um rápido surto consumindo grande quantidade de gás. Já nas espirais, o processo de formação foi lento, conservando seu gás e gerando estrelas por bilhões de anos.

Existem duas teorias sobre formação e evolução de galáxia: o modelo monolítico e o hierárquico. Contudo, ainda não existe uma única teoria que satisfaça todos os aspectos observacionais e que consiga prever como as galáxias se formaram e evoluíram. O cenário monolítico propõe que a formação e a evolução das galáxias aconteceram de maneira isolada pelo colapso de grandes nuvens de gás. Nessa situação, o formato das galáxias é determinado pela taxa de formação estelar dessa nuvem em contração e pelo momento angular da nuvem. Para nuvens com momento angular baixo, a taxa de formação estelar seria alta, então quase todo o gás seria consumido, originando uma galáxia elíptica ovalada. Por sua vez, sendo o momento angular da nuvem alto e a taxa de formação estelar baixa, o resultante seria uma galáxia espiral que teria gás para manter sua formação estelar por um longo tempo. Diferentemente do primeiro modelo, o cenário hierárquico tem como hipótese que pequenas nuvens de gás, contraindo-se, originariam sistemas puramente discoidais. Essas nuvens evoluiriam a galáxias espirais, se sofressem poucas interações, ou evoluiriam a galáxias elípticas, se houvesse fusões e encontros frequentes entre nuvens menores.

No passado, havia muitas galáxias pequenas com alta taxa de formação estelar. Isso não é observado atualmente, o que propõe que essas galáxias tenham se agregado, formando galáxias maiores. Outras contradições podem ser encontradas, como o fato de galáxias espirais serem raras em aglomerados densos de galáxias, onde as elípticas estão em maior número. Além disso, existe a indicação de que as estrelas de galáxias elípticas têm a mesma idade num determinado redshift. Essas contradições ora beneficiam um modelo, ora beneficiam outro, surgindo a possibilidade de que exista a formação monolítica para galáxias isoladas e a hierárquica para aglomerados

de galáxias.

A morfologia e a classificação de galáxias são o primeiro passo no estudo das galáxias como unidade fundamental da matéria no espaço. Qualquer teoria de formação ou evolução precisa explicar a distribuição das galáxias como uma função do tempo e do ambiente cósmico. A morfologia também está fortemente ligada ao histórico de formação de estrelas nas galáxias. O principal objetivo de estudos extragalácticos tem sido entender o que direciona a evolução morfológica das galáxias. É importante determinar os mecanismos dinâmicos e evolucionários que estão na base da enorme gama de formas que define vários esquemas de classificação de galáxias utilizados atualmente [2] [3], pois isso nos permitirá estabelecer relações, caso existam, entre diferentes tipos de galáxias.

3 Sistema CASGM

Medidas não paramétricas da morfologia das galáxias não assumem uma função analítica particular para a distribuição de luminosidade das galáxias. Foram introduzidos o índice de concentração (C) [4] e a assimetria rotacional (A) [5] como uma maneira de distinguir automaticamente as galáxias de tipo precoce e tardia definidas por Hubble. Autores subsequentes modificaram as definições originais para tornar C e A mais robustos para seleção da superfície de brilho e centralização de erros [6] [7]. A terceira quantidade do sistema de classificação morfológica "CAS"é uma medida de estruturas em pequena escala, a suavidade (S).

O CAS é um método não paramétrico simples, mas também tem seus pontos fracos. Como a concentração é medida dentro de diversas aberturas circulares sobre um centro pré-definido, fica implícito que assume uma assimetria circular, tornando-o um meio pobre para descrever galáxias irregulares. Outras duas medidas não paramétricas para quantificar a morfologia de uma galáxia foram acrescentadas ao sistemas CAS. A primeira é o coeficiente de Gini (G), adaptado à classificação morfológica por [8] para quantificar a distribuição relativa do fluxo dentro de um pixel associado com a imagem da galáxia; e a segunda, o momento de luz (M) da galáxia [9]. Além dessas medidas, foram utilizados o índice de Sérsic, que é uma função matemática que descreve como a intensidade I de uma galáxia varia de acordo com uma distância R de seu centro [10], e o índice σ_Ψ, que mede a quantidade de estruturas não radiais nas galáxias, em especial braços espirais e barras.

3.1 Concentração

A concentração (C) é a relação que mede a quantidade de luz dentro de uma abertura interna da galáxia, que pode ser circular ou elíptica, e também dentro de uma abertura mais externa. Foi adotada a definição de [2], de maneira que se calcula a taxa de luz presente na abertura mais interior, nesse caso 20%, e na mais exterior, que corresponde a 80% do fluxo total de luz,

$$C_1 = 5 \log \frac{R_{80}}{R_{20}} \tag{1}$$

em que R_{80} e R_{20} são aberturas circulares contendo, respectivamente, 80% e 20% do fluxo total de luz. Outro caso particular do índice de concentração é

$$C_2 = 5 \log \frac{R_{90}}{R_{50}} \tag{2}$$

que contempla as aberturas circulares R_{90} e R_{50} com, respectivamente, 90% e 50% do fluxo total de luz.

3.2 Assimetria

A assimetria consiste no parâmetro antigamente utilizado para quantificar a morfologia de galáxias que geralmente apresentavam alto *redshift*. Todavia, existem tentativas para calibrar e caracterizar a assimetria em galáxias mais próximas e modelos de formação de galáxias para entender sua importância na evolução dessas formações.

A assimetria é medida como o resíduo entre a imagem original e a sua versão rotada em 180° ,

$$A = \frac{\sum_{i,j} |I(i,j) - I_{180}(i,j)|}{\Sigma_{i,j} |I(i,j)|} - B_{180} \tag{3}$$

em que I é a imagem da galáxia e I_{180} é a mesma imagem rotacionada em 180° sobre o pixel central da galáxia. Devido ao ruído, a assimetria pode ser corrigida. No entanto, quando existe um baixo sinal médio de ruído por pixel ($< S/N >$), torna-se problemático defini-la.

Objetos elípticos com perfis de luz muito suaves têm um grande ângulo de simetria rotacional. Já galáxias com braços espirais são menos simétricas, e as irregulares e mergings são geralmente muito assimétricas.

3.3 Suavidade

Parâmetro utilizado para quantificar a fração de luz que está contida em estruturas de pequena escala como aglomerados de formação estelar. Galáxias que ainda apresentam formação estelar tendem a ter estruturas agrupadas e altos valores para a suavidade com $S = 0.1 - 1$, enquanto galáxias sem formação estelar apresentam uma distribuição de luz suave com $S < 0.1$. O método mais usual para calcular a suavidade foi desenvolvido por [11],

$$S = \frac{\sum_{i,j} |I(i,j) - I_s(i,j)|}{\sum_{i,j} |I(i,j)|} - \frac{\sum_{i,j} |B(i,j) - B_s(i,j)|}{\sum_{i,j} |B(i,j)|} \tag{4}$$

em que I e B são, respectivamente, as imagens da galáxia e do fundo, e I_s e B_s são as imagens suavizadas. A região até $0.25R_p$ do núcleo é descartada, já que sua inclusão contribuiria num acréscimo à suavidade que não está ligado a uma região de intensa formação de estrelas novas. A imagem suavizada é subtraída da original, produzindo um mapa residual que contém somente as estruturas de alta frequência da galáxia, então, é realizado o somatório desses resíduos pelo total de luz da imagem original, e disso é subtraído o ruído do céu depois de suavizado.

3.4 Coeficiente de Gini

O coeficiente de Gini é uma ferramenta estatística que infere a distribuição de riquezas em uma determinada população, portanto, é uma medida de desigualdade. Transpondo esse conhecimento para as imagens das galáxias, quando essa distribuição está em equilíbrio, significa que a distribuição de luz está dividida igualmente entre todos os pixels e o valor de G será igual a zero. Caso ocorra um desequilíbrio, toda a distribuição de luz estará presente em um pixel, e o valor de G será igual a um.

O valor de G é definido pela razão das áreas no diagrama da curva de Lorenz da distribuição de luz da galáxia, onde a posição espacial não é considerada [9]. Cada pixel é ordenado por seu brilho e percebido como parte de uma distribuição cumulativa. A maneira mais eficiente de computar o coeficiente de Gini é, num primeiro momento, ordenar crescentemente os valores de X_i

$$G = \frac{1}{\overline{X} n(n-1)} \sum_{i}^{n} (2i - n - 1) X_i \tag{5}$$

em que *n* representa a quantidade de pixels em uma galáxia e $\overline{X}$ é a média de todos os valores X_i do fluxo de pixels.

4 Materiais e métodos

A técnica de mineração de dados a ser utilizada é a de agrupamentos, também conhecida como clustering [12]. A análise de agrupamentos constitui uma metodologia numérica multivariada, com o objetivo de propor uma estrutura classificatória, ou de reconhecimento da existência de grupos, objetivando, mais especificamente, dividir o conjunto de observações em um número de grupos homogêneos, seguindo algum critério de homogeneidade [13].

A análise de agrupamentos explora semelhanças entre padrões, organizando os dados em grupos, de forma que as características dos objetos pertencentes ao mesmo grupo sejam mais parecidas entre si e distintas dos objetos presentes em outros grupos.

O processo de construção de agrupamentos compreende, essencialmente, cinco etapas principais, segundo [14] [15]. A primeira etapa é determinada pela preparação dos dados, pois, em alguns casos, é necessário transformá-los por

meio de normalizações e seleção de características. Após é preciso medir a similaridade ou dissimilaridade, que é computada mediante uma função definida entre pares de objetos. O próximo passo a ser realizado, simultaneamente com o anterior, consiste em aplicar as técnicas de agrupamento por meio dos algoritmos selecionados, nesse caso, EM e K-médias. De posse dos resultados, na quarta etapa é realizada a validação, na qual é medida, por meio de índices, a qualidade dos agrupamentos. Finalmente, após a execução de todas as fases, é necessária a interpretação dos resultados em relação aos grupos obtidos, examinando padrões para descrever sua natureza.

É utilizada a análise exploratória de dados por meio de técnicas de agrupamento, objetivando analisar os resultados para detectar classes de galáxias mediante parâmetros morfométricos. Fazem parte do conjunto de dados, os dados reais e sintéticos que contenham medidas morfométricas de galáxias. Após essas etapas serem realizadas, é aplicado um algoritmo para validar os resultados obtidos, o Silhouette.

Todo o procedimento metodológico é desenvolvido em Matlab R2012a. Os dados utilizados pertencem ao Extraction de Formes Idealisées de Galaxies en Imagerie (EFIGI) [6], que é um banco de dados projetado para uma amostragem densa de todos os tipos de galáxias e dezesseis atributos estimados visualmente. Nesse catálogo estão presentes dados de outros surveys e catálogos como o Sloan Digital Sky Survey (SDSS) [16], além de fornecer informações morfológicas detalhadas e imagens de 4.458 galáxias.

Por meio do MORFOMETRYKA [17], são calculados os índices estruturais do sistema CASGM, do índice de Sérsic e do raio Petrosiano. Esse algoritmo atua de maneira automatizada em grandes quantidades de dados, executando os processos de medições a partir do nome das galáxias.

4.1 Maximização de expectativa

A maximização de expectativa consiste na generalização da estimativa de máxima verossimilhança para bancos de dados com dados incompletos [18]. Esse é um algoritmo iterativo que considera uma mistura de modelos probabilísticos que descreve a distribuição dos grupos. O algoritmo assume que os componentes individuais são misturas de densidades.

No EM, a densidade de componentes é desconhecida, assim como o parâmetro de mistura, sendo necessário calculá-los com base em padrões. Existem dois passos principais para o funcionamento do EM. A primeira etapa consiste na expectativa que diz respeito às variáveis desconhecidas, utilizando a estimativa atual de parâmetros e condicionando as observações. Sendo assim, associa cada objeto x_i ao agrupamento C_k, com a seguinte probabilidade:

$$P(x_i \in C_k) = p\left(\frac{C_k}{x_i}\right) = \frac{p(C_k)p\left(\frac{x_i}{C_k}\right)}{p(x_i)} \tag{6}$$

em que

$$p(\frac{C_k}{x_i}) = N(x_i, E_k) \tag{7}$$

segue uma distribuição gaussiana de probabilidade com média m_k e valor esperado E_k.

A segunda etapa engloba a maximização, na qual é produzida uma nova estimativa de parâmetros até alcançar sua convergência, em que é utilizada uma função de verossimilhança das distribuições de probabilidade.

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i p(x_i \in C_k)}{\sum_j p(x_i \in C_j)} \tag{8}$$

Os passos de expectativa e maximização estão interligados, pois as novas probabilidades calculadas na fase de maximização serão utilizadas para realizar a inferência na fase da expectativa.

4.2 K-médias

Método nãosupervisionado de classificação que tem como objetivo minimizar a soma do erro quadrático sobre todos os grupos [19]. Para isso, requer três parâmetros específicos: o número de grupos, a inicialização do

grupo e a métrica da distância. O erro quadrático entre u_k e os pontos no grupo C_k são definidos como:

$$J(C_k) = \sum_{x_i \in C_k} \| x_i - \mu_k \|^2 \tag{9}$$

em que $X = x_i, i = 1, ..., n$ é o conjunto de n pontos d-dimensionais, $C = c_k, k = 1, ..., K$ é o conjunto de K clusters e μ_k é a média de clusters C_k. Como o objetivo é minimizar a soma do erro quadrado sobre todos os clusters, a equação é reescrita:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2 . \tag{10}$$

Um dos parâmetros mais complexos na análise de clusters é a definição do número de grupos (k) a serem encontrados no conjunto de dados. Em técnicas hierárquicas, é equivalente a decidir qual é o valor de dissimilaridade usado para cortar o dendrograma; já em métodos não hierárquicos, é o valor de k que deve ser fixado antes de começar o procedimento. O método para escolher o número de clusters no K-médias consiste em aplicar o algoritmo para k variáveis entre 2 e $m - 1$ e escolher o valor de k que resulte no melhor valor do silhouette para o conjunto de dados.

4.3 Método de validação

O processo de avaliação dos resultados dos clusters apresenta quatro componentes principais:

1. determinar se há uma estrutura não aleatória nos dados;

2. determinar o número de clusters;

3. avaliar como um resultado de clustering se ajusta a um conjunto de dados, sendo essa a única informação disponível;

4. avaliar o quão bem localizados estão os objetos dentro dos clusters de acordo com as partições obtidas baseadas em outras fontes de dados.

O Silhouette pode ser utilizado como índice de medição de qualidade do agrupamento final, independentemente da técnica escolhida [3]. A largura da silhueta avalia a qualidade de uma solução do agrupamento, considerando tanto a compacidade (distância entre os pontos de dados dentro do mesmo grupo) quanto a separação (distância entre os pontos de dados em dois grupos vizinhos). Esse método fornece a possibilidade de determinar o número apropriado de grupos, tal que o valor de k é escolhido de maneira que forneça o melhor valor médio do Silhouette, além de poder comparar grupos obtidos por diferentes algoritmos.

$$s(i) = \frac{b_i - w_i}{\max(b_i, w_i)} \tag{11}$$

com

$$b_i = \min_k (B_{i,k}) \tag{12}$$

em que w_i é a distância média do i-ésimo ponto até os outros pontos de um mesmo cluster e $B_{(i,k)}$ é a distância média do i-ésimo ponto até os pontos de outro cluster k. Quando essa medida apresenta um valor unitário positivo, indica que os pontos estão dispostos de maneira adequada; já ao ser caracterizado pelo valor zero, os pontos não estão bem separados em um cluster ou em outro. Para um valor unitário negativo, os pontos provavelmente estão nos clusters errados. Essa técnica também fornece uma representação gráfica do quão bem localizado está cada objeto nos grupos resultantes dos agrupamentos.

Medidas globais das silhuetas são dadas ou pela média por agrupamento ou por todo o conjunto de dados [20]. O valor médio de s_i, a largura média da silhueta de um grupo, para todo i em um dado grupo, é definido como a média de todas as silhuetas individuais. Então, $\overline{s}(k)$ é a média de todas as silhuetas individuais, sendo utilizada como um parâmetro para qualificar grupos com os mesmos dados, mas diferentes números de variáveis, procurando pelo melhor agrupamento, onde m é o número de objetos no conjunto de dados.

$$\overline{s}(k) = \frac{1}{m} \sum_{i=1}^{m} s(i). \tag{13}$$

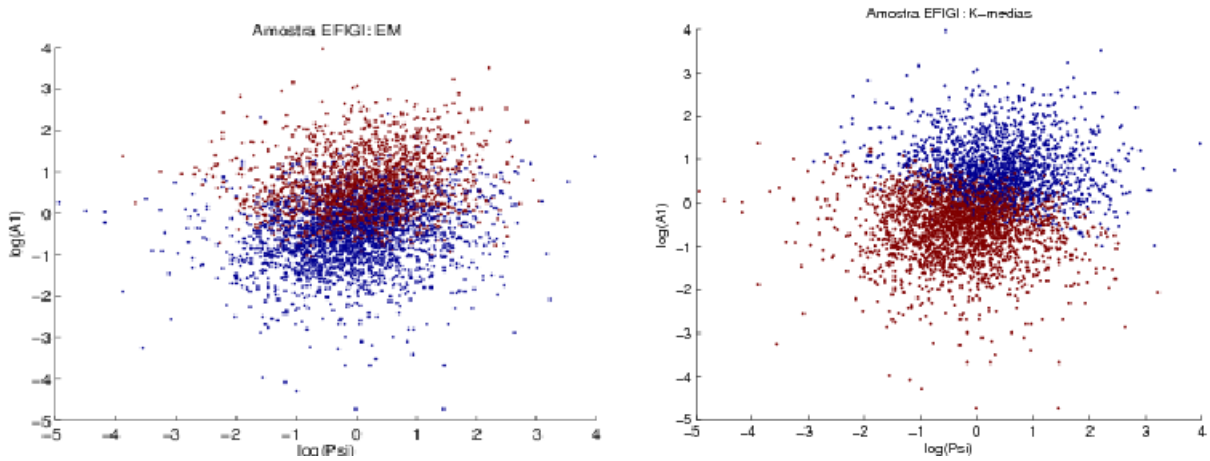
5 Resultados

Os primeiros testes foram realizados com dados sintéticos, privilegiando duas dimensões; após foram incrementados para cinco dimensões. Essa abordagem foi aplicada devido à necessidade de inferir a qualidade dos algoritmos de agrupamento e métodos de validação. Esses dados foram apresentados de duas maneiras. Quando as gaussianas estão bem separadas, percebe-se claramente a existência de dois grupos distintos. Contudo, quando essas gaussianas se aproximam até que seus centros coincidam, a separação entre os objetos é mais difícil. Foram aplicados os mesmos algoritmos nos dados do catálogo EFIGI, a fim de encontrar os grupos que separariam os diferentes tipos de galáxias presentes nessa amostra. Nas Figuras 1, 2 e 3, estão representados os grupos encontrados por meio dos algoritmos Expectation Maximization e K-médias.

Os parâmetros que se mostraram com maior potencial para separar os dados em dois grupos distintos de galáxias foram o coeficiente de Gini, a assimetria, a suavidade e o σ_Ψ. Os outros parâmetros promoveram pouca ou nenhuma separação dos dados.

O parâmetro σ_Ψ, quando associado a outros índices morfométricos, como a suavidade e a assimetria, promove uma distinção clara dos dados em dois grupos com características contrastivas. O mesmo acontece com o coeficiente de Gini, que apresenta um grande potencial para essa aplicação.

Figura 1: Comparação entre os resultados obtidos pelos algoritmos EM e K-médias, respectivamente, quando os atributos utilizados são $\log(A_1)$ x $\log(\sigma_\Psi)$ com dados provenientes do catálogo EFIGI. Ambos os algoritmos encontraram resultados equivalentes ao separarem galáxias elípticas e espirais.



Após a utilização das técnicas de agrupamento, o método de validação Silhouette foi aplicado para inferir a qualidade dos agrupamentos gerados pelos dois algoritmos selecionados. Os resultados obtidos são satisfatórios, pois, por meio do Silhouette, pode-se deduzir que todos os objetos estavam localizados em seus respectivos grupos. Nesse caso, o EM se mostrou mais adequado à aplicação, pois seu coeficiente de Silhouette é melhor que o apresentado pelo K-médias. Apesar desse fato, ambos os algoritmos deram origem a resultados semelhantes, o que os torna aptos para aplicação na classificação de galáxias. Os resultados obtidos são comparados com os dados rotulados por meio de uma matriz de confusão que mostra o número de classificações corretas em relação às preditas, com o objetivo de identificar os objetos presentes em cada grupo.

Figura 2: Comparação entre os resultados obtidos pelos algoritmos EM e K-médias, respectivamente, quando os atributos utilizados são $\log(G)$ x $\log(A_1)$ com dados provenientes do catálogo EFIGI. Os parâmetros utilizados evidenciaram uma separação efetiva dos dados e demonstram a continuidade morfométrica entre as galáxias elípticas e espirais.

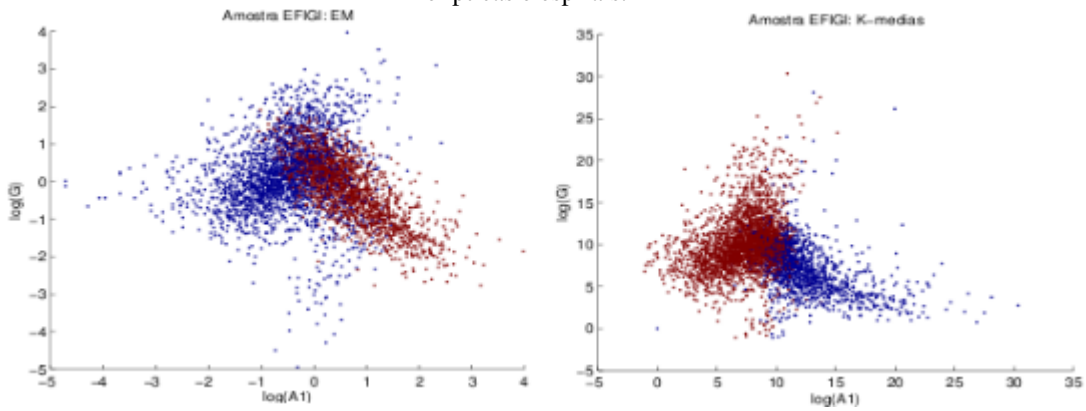
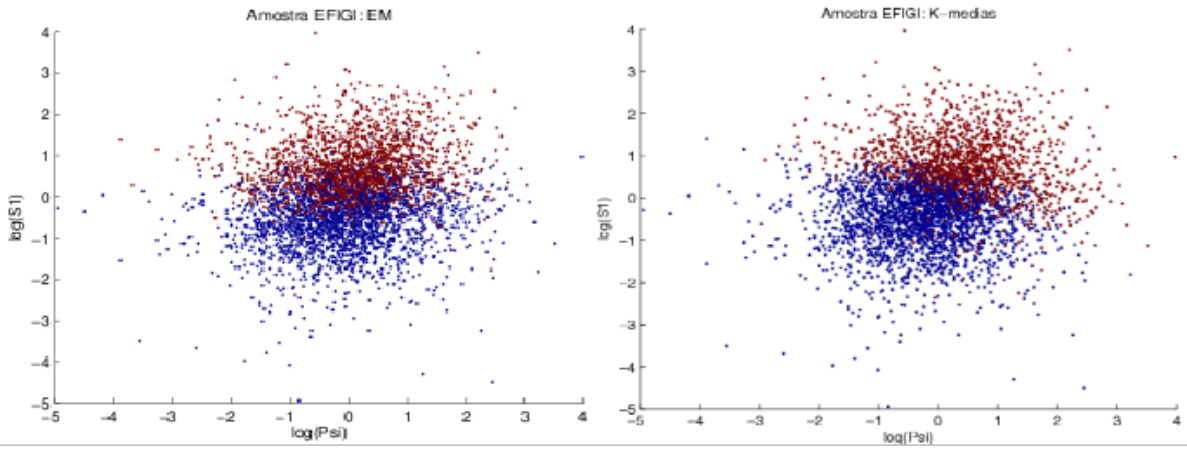


Figura 3: Comparação entre os resultados obtidos pelos algoritmos EM e K-médias, respectivamente, quando os atributos utilizados são $\log(S_1)$ x $\log(\sigma_\Psi)$ com dados provenientes do catálogo EFIGI. Ambos os algoritmos encontraram resultados equivalentes ao separar galáxias elípticas e espirais.



6 Conclusões e perspectivas futuras

A morfologia fornece informações importantes sobre as propriedades físicas das galáxias, como a taxa de formação estelar e a cinemática. Um dos objetivos principais dos estudos extragaláticos é entender o que direciona a morfologia das galáxias e como elas evoluem com o tempo e o ambiente cósmico.

Os algoritmos de agrupamento apresentam um grande potencial para serem utilizados na classificação entre galáxias elípticas e espirais. Espera-se encontrar um número correto de grupos tal que separe os diferentes tipos morfológicos de galáxias de acordo com suas propriedades morfométricas presentes nos surveys a serem utilizados.

Foram procurados grupos que separassem os diferentes tipos morfológicos de galáxias de acordo com suas propriedades morfométricas presentes no catálogo. Nesse espaço de parâmetros, surgem alguns grupos de objetos com diferenças nítidas, apesar de existir uma continuidade morfométrica entre os tipos de galáxias utilizados. Com esses resultados, pode-se perceber que as galáxias espirais e elípticas apresentam algumas características

morfométricas que as distinguem, contudo, devido a essa continuidade, não é possível caracterizá-las por uma visão bimodal como famílias claramente distintas. Outras análises de agrupamentos serão realizadas com catálogos de 14 mil objetos [21] e em grupos de Berlind [17] com cerca de 80 mil objetos, a fim de aprimorar a metodologia aplicada para a classificação de galáxias desses catálogos. Esses resultados também serão avaliados por uma matriz de confusão, e a localização dos dois tipos de galáxias presentes nos surveys poderá ser caracterizada por meio de uma função de probabilidade.

Agradecimentos

À Capes.

Referências

[1] HUBBLE, E. P. *Realm of the Nebulae*. [S.l.]: Yale University Press, 1936.

[2] BERSHADY, M.; CONSELICE, C. The asymmetry of galaxies: Physical morphology for nearby and high-redshift galaxies. *The Astrophysical Journal*, 2000.

[3] ROUSSEEUW, P. Silhouettes: a graphical aid to the interpretation and validation of clusters analysis. *Journal of Computational and Applied Mathematics*, v. 20, n. 53, 1987.

[4] ABRAHAM, R. G. et al. Galaxy evolution in abell 2390. *The Astrophysical Journal*, v. 471, p. 694–719, 1996.

[5] SANDAGE, A.; BEDKE, J. The carnegie atlas of galaxies. *Washington: Carnegie Institution*, 1994.

[6] BAILARD, A. et al. The efigi catalogue os 4458 nearby galaxies with detailed morphology. *A&A*, 2001.

[7] BUTA, J. et al. The de vaucouleurs atlas of galaxies. *Cambridge University Press*, 2007.

[8] ABRAHAM, R. G.; BERGH, S.; NAIR, P. A new approach to galaxy morphology. i. analysis of the sloan digital sky survey early data release. *The Astrophysical Journal*, v. 588, n. 218A, 2003.

[9] LOTZ, J. M.; PRIMACK, J.; MADAU, P. A new nonparametric aproach to galaxy morphological classification. *The Astronomical Journal*, v. 128, p. 163–182, 2004.

[10] SÉRSIC, J. L. Influence os the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la associacion Argentina de Astronomia La Plata Argentina*, v. 451, n. L1, 1963.

[11] CONSELICE, C.; BERSHADY, M.; JANGREN, A. Structural and photometric classification of galaxies. i. calibration based on a nearby galaxy sample. *The Astronomical Journal*, 2000.

[12] TAN, P.; STEINBACH M. ANDF KUMAR, V. *Introdução ao Data Mining.* [S.l.: s.n.], 2009.

[13] REGAZZI, A. Análise multivariada, notas de aula inf 766,. *Departamento de Informática da Universidade Federal de Viçosa*, v. 2, 2001.

[14] ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. [S.l.: s.n.], 1996. p. 226–231.

[15] FACELLI, K. *Um framework para análise de agrupamento baseado na combinação multi-objetivo de algoritmos de agrupamento*. Tese (183f. Tese de Doutorado em Ciências da Computação e Matemática Computacional) — Instituto de Ciências Matemática e de Computação. Universidade de São Paulo, São Carlos, 2006.

[16] YORK, D. G.; ADELMAN, J. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, v. 120, p. 1579–1587, 2000.

[17] FERRARI, F. Morfometryka: Galaxy morphometry in prep. 2015.

[18] DEMPSTER, A.; LAIRD, N.; RUBIN, D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, Series B, n. 39, p. 1–38, 1977.

[19] LLETÍ, R. et al. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, v. 515, p. 87–100, 2004.

[20] JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys (CSUR)*, p. 264–323, 1999.

[21] NAIR, P.; ABRAHAM, R. A catalogue of detailed visual morphological classifications for 14,034 galaxies in the sloan digital sky survey. *Astrophysical Journal*, v. 186, n. 427, 2010.