FICE
3ª FEIRA DE INICIAÇÃO CIENTÍFICA E EXTENSÃO
17 e 18 de setembro/2014
ENSINO DE QUALIDADE
CURSOS GRATUITOS
PROFESSORES MESTRES E DOUTORES

SEGMENTAÇÃO DE GALÁXIAS BASEADAS EM CARACTERÍSTICAS ÓPTICAS ATRAVÉS DO WATERSHED

Thiago Crestani[1] *; Fábio José Rodrigues Pinheiro*[2]

INTRODUÇÃO

O desenvolvimento de novas tecnologias tem resultado na produção de grandes volumes de dados em diversas áreas de conhecimento. Entretanto, esta produção de dados não possui um relacionamento direto com a descoberta do conhecimento. Para tal, a utilização das tecnologias de Banco de Dados e *Data Mining* (Mineração de Dados) tem se mostrado de fundamental importância para o processamento e a classificação desses dados. A área da astronomia é uma das ciências que atualmente envolve volumes de dados gigantescos. Numa dessas linhas está a análise espectroscópica e fotométrica de galáxias, mantendo em alguns casos informações de milhões de objetos *(surveys)*. Para a manutenção e recuperação destes dados, são criados bancos de dados específicos. Neste sentido, o projeto aplica técnicas de mineração de dados em *surveys* de galáxias. Ele é uma continuação de um projeto anterior com o mesmo tema, porém neste projeto procurou-se aperfeiçoar os algoritmos utilizados e automatizar o processo de segmentação.

PROCEDIMENTOS METODOLÓGICOS

No projeto anterior foram obtidos dados de dois observatórios astronômicos internacionais o *WISE* [WISE,2010] e o *SDSS* [SDSS,2014] que vieram inicialmente em formato *FITS* e para armazená-los foi escolhida ferramenta *PostgresSQL*. Considerando que ela não suporta a entrada de dados nesse formato, foi desenvolvida uma ferramenta para fazer a tradução dos dados de formato *FITS* para

[1] Aluno do Instituto Federal Catarinense - Campus Videira - Curso de Ciência da Computação - E-mail: thiagocrestani@gmail.com
[1]

[2] Professor Orientador. Instituto Federal Catarinense - Campus Videira - Curso de Ciência da Computação E-mail: fabio@ifc-videira.edu.br
[2]

SQL, utilizando a linguagem de programação *Java*. Os dados já em formato *SQL* foram inseridos no *PostgreSQL*.

Alguns dos dados armazenados no banco de dados possuíam valores de capturas incorretas, para eliminar estes valores realizou-se uma filtragem e a eliminação de registros com *REDSHIFT* menor que 0.01 e maior que 1.5, e campos com os registros *DIST_X, W1PRO, W2PRO, W3PRO e W4PRO* igual a *NULL*.

Com os dados armazenados foram realizados testes para verificar a coesão deles. Alguns gráficos do tipo *BPT* [BALDWIN, PHILLIPS, TERLEVICH, 2014] e do tipo *WHAN* [FERNANDES et al., 2014] foram gerados para mostrar visualmente como estavam dispostos os dados.

Para a integração com o banco e a criação dos gráficos de pontos foi utilizado a linguagem de programação *Python*. Após isso foram realizadas divisões nos gráficos gerados, feitas de forma estática, ou seja, os valores para traçar as linhas foram previamente calculados e aplicados em todos os desenhos. Não foi necessário nenhum tipo de processamento, já que para todos os gráficos as linhas passavam pelo mesmos locais.

Uma forma de classificar as galáxias é através da separação dos grupos de acordo com cada característica, existem padrões que fazem a divisão através de valores pré-definidos, porém foi necessário criar uma forma que permita classificar essas galáxias de forma automática para cada conjunto de características e com isso melhorar a classificação das mesmas. Então foi realizada uma ampla pesquisa para verificar qual o melhor método para realizar a separação automática do grupo de galáxias, pois não bastava que a divisão fosse apenas realizada visualmente (nos gráficos) era preciso que cada registro fosse classificado e estimado o desvio da reta de corte.

Após pesquisa constatou-se que os métodos *Watershed* e *Simple-k-Means* seriam os mais efetivos para aplicação das técnicas de divisão dos dados. Entretanto o *Watershed* foi aplicado e *Simple-k-Means* foi utilizado apenas para testes de comparação.

Antes de iniciar a aplicação do *Watershed* aconteceram algumas etapas de preparação dos dados. A primeira delas foi a transformação em matrizes tridimensionais, que através de um programa escrito em *Python* carrega valores do banco e insere em uma posição de uma determinada matriz, essa matriz definida com as dimensões fixas de 800x800 contém o número total de registros presentes em cada parte dos gráficos. Esse processo foi realizado com todas as combinações de subtração entre os parâmetros presentes no banco de dados, que são *w1, w2,*

w3 e *w4* para o eixo x e todas a combinações de subtração de u, r, i, g e z para o eixo y, o que gera 60 gráficos, porém 8 foram descartados devido inconsistência na captura dos dados.

Em seguida foram retiradas as áreas sem valores dos gráficos, que devido ao fato que terem sido gerados com tamanho fixo, em certas áreas não existe nenhum ponto, o que prejudica processo de *limiarização* que inicialmente foi executada com um valor fixo para todos os gráficos, porém como os picos mudavam de formato e também de distância entre eles, uma função foi utilizada para obter o valor ideal de *limiarização* para cada gráfico e aplicá-la.

Para separar os dois picos e aplicar o *Watershed,* a parte mais baixa do gráfico foi corroída deixando apenas dois picos, só que da mesma maneira que na *limiarização* os picos eram desuniformes e as vezes os valores fixos da corrosão eram muito baixos para alguns gráficos e altos para outros. Então fazendo testes percebeu-se que para que o *Watershed* tivesse melhor resultado era necessário fazer a menor corrosão possível, ou seja, separar os dois picos com a menos distancia possível entre eles. Então foi implementado um *script* que encontra a menor corrosão e aplicava sobre o gráfico. Entretanto alguns gráficos tinham picos totalmente diferentes, então só a corrosão não era suficiente, por isso foi desenvolvido um *script* que testa se a divisão foi bem sucedida, caso contrário ele corta os valores mais baixos gradativamente, até que se tenha um bom resultado.

Após isso foi aplicado a dilatação dos limites, para que os picos fiquem mais próximo e então executado o *Watershed*. A linha de tendência da divisão dos picos foi traçada nos gráficos através de uma função que usa polinômios de grau 2.

E no final pode-se obter a função da reta para cada gráfico, então comparando com os valores do banco utilizados para gerar o gráfico foi possível marcá-los se estavam acima ou abaixo da linha de divisão por idade. Para realizar esse procedimento de forma automática foi utilizado um *script* desenvolvido na linguagem de programação *Python*.

RESULTADOS E DISCUSSÕES

O projeto foi construído com base em uma projeto anterior do qual herdou alguns recursos. No projeto inicial os dados foram armazenados em um banco de dados específico para o projeto e também foram realizados alguns testes para verificar a consistência desses dados.

Após isso foram realizados os primeiros testes com o método de divisão *watershed.* Esses foram desenvolvidos com gráficos experimentais, feitos com vários recursos manuais e pouco refinamento.

No projeto atual foi construído um *script* em *python* capaz de criar as combinações de todos os campos e gerar os gráficos tridimensionais. São 60 gráficos gerados respeitando os parâmetros presentes no banco de dados.

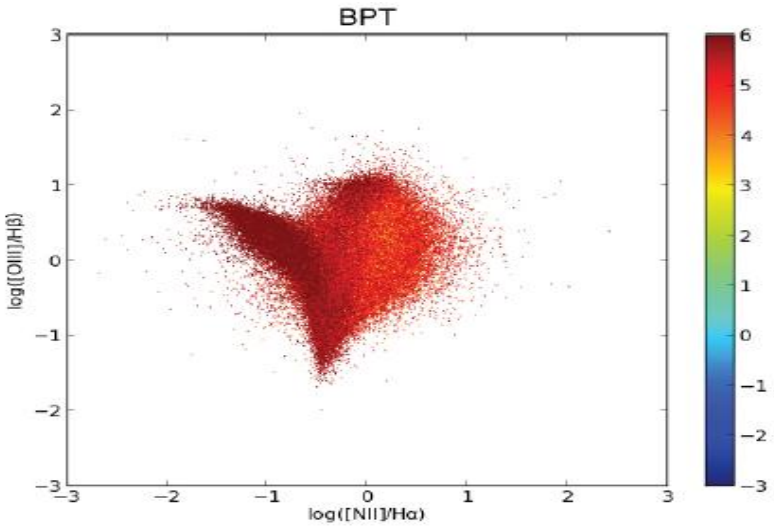

Figura 1 - Gráfico *BPT*

Ainda no primeiro projeto, foram realizados classificações simples nos dados. Estas consistiam em traçar linhas em pontos específicos do gráfico e conforme os dados ficam distribuídos é possível perceber a quais agrupamentos estão presentes na imagem.

No segundo projeto essas divisões puderam ser implementadas em todos os gráficos uma vez que eles foram gerados de forma automática.

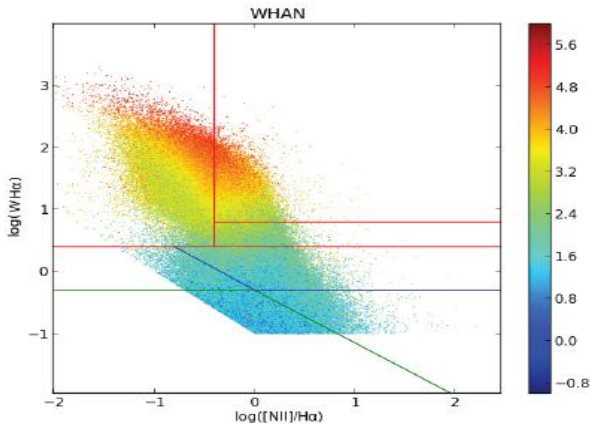

Figura 2 - Gráfico *WHAN* com as divisões fixas traçadas

Para a classificação das galáxias por idade é necessário dividir os gráficos do tipo *WHAN* conforme mostra a Figura 2, em duas categorias, no entanto essa divisão tem uma metodologia que deve ser seguida. Os gráficos são compostos por dois

agrupamento de dados (picos), esses agrupamentos devem ser separados ao meio exatamente na menor concentração de pontos.

Para realizar essa tarefa foi escolhido o método *Watershed*, mas antes de iniciar a aplicação da técnica, acontecerem algumas etapas de preparação dos dados. A primeira delas foi a criação de matrizes tridimensionais, que através de um programa escrito em *Python* carrega valores do banco e insere em uma posição de uma determinada matriz. Essa matriz tem as dimensões fixas de 800 X 800 e contém o número total de registros presentes em cada parte do gráfico.

Então foram retirados os limites excessivos do gráfico para evitar erros de processamento. Em seguida foram aplicadas técnicas de *limiarização*, dilatação de limites e corrosão da imagem para posteriormente dividir os dois picos utilizando o método *Watershed*. Processados os dados é possível observar a linha que corta os dois conjuntos de galáxias.

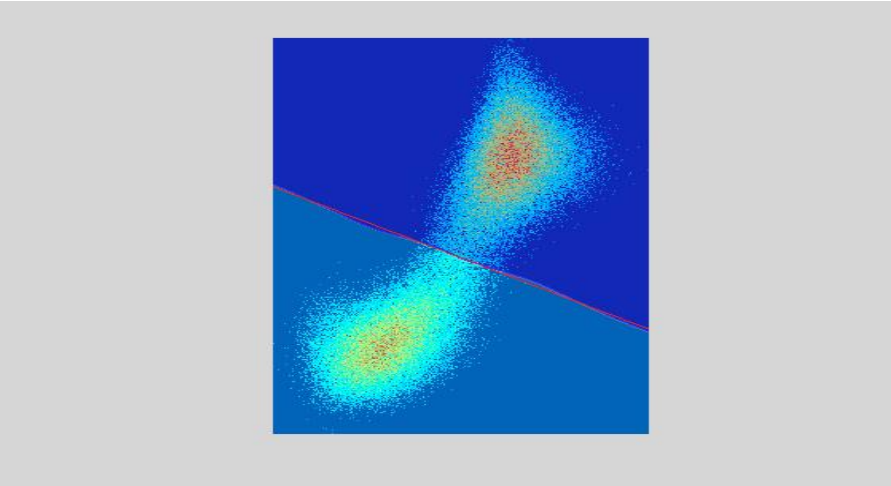

Figura 3 - Gráfico com a Linha de *Watershed* e a Linha de Tendência Traçada

Tendo em vista que é possível recuperar os pontos por onde passa a linha do watershed, utilizando um polinomio de grau 2 foi possível função da reta de cada divisão, que foi utilizado para comparação dos registros no banco de dados.

```
%armazena as coordenadas da linha nas variaveis x e y
[ x y ] = find(m > 1);
%junta as coodenadas
xy = [ x y ];
%polinomio de grau 2
eq = polyfit(x,y,1);
xf = eq(1);
yf = eq(2);
% gera equção da reta em formato ideal para o script python
disp([nome,'="',num2str(xf), '*(valor1-',num2str(contx),')+',num2str(yf),'"']);
disp([nome,'dify','=',num2str(conty)]);
ybest=xf*x+yf;
```

Figura 4 - Trecho de código que gera a equação da reta dos gráficos.

Essas equações foram armazenadas em um arquivo que contem a equação e o valor de linhas que foram cortadas na vertical e horizontal de cada gráfico. E por fim, usando a lista de funções de reta dos gráficos, os dados foram comparados e implantados no banco mapeando os registros e marcando se eles estavam acima ou abaixo da linha de tendência desenhada no gráfico. No total, o banco possui cerca de 900 mil registros e foram gerados 5 gráficos, resultando em uma quantidade de cerca de 50.400.000 registros que foram classificadas no banco de dados .

CONSIDERAÇÕES FINAIS

No projeto anterior foi realizada a divisão visual dos dados, enquanto nesse obteve-se a divisão real, marcando os registros um a um dentro do banco de dados. Desta forma é possível obter exatamente quais galáxias estão acima ou abaixo de cada gráfico e também fazer a comparações entre elas. Alem disso foram armazenados no banco a distancia que cada registro está da linha divisória, isto é muito importante visto que quanto mais perto ou longe da linha a galáxia pode ser mais jovem ou nova.

Ao final, alcançou-se o objetivo de realizar a marcação dos dados presentes no banco de dados de acordo com o agrupamento que pertencem, ainda foram registrados as distâncias que cada galáxia tinha da linha divisória entre esses agrupamentos. Ambas as técnicas testadas mostraram bom desempenho. Considera-se importante a realização de outros testes e ainda com outras técnicas para fazer uma comparação mais precisa entre elas.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. N. Mining association rules between sets of items in large databases. Proceedings of the ACM International Conference on Management of Data, p. 207–216, 1993.

AGRAWAL, R.; SRIKANT, R. Mining sequential patterns. In: YU, P. S.; CHEN, A. S. P. (Ed.). Eleventh International Conference on Data Engineering. Taipei, Taiwan: IEEE Computer Society Press, 1995. p.3–14. Disponível em: <citeseer.ist.psu.edu/agrawal95mining.html>

Baldwin, J. A.; Phillips, M. M.; Terlevich, R. - SAO/NASA ADS Astronomy Abstract Service - Disponível em: http://adsabs.harvard.edu/abs/1981PASP...93....5B - Acesso em: 01 de Agosto de 2014

Fernandes, Cid et al - SAO/NASA ADS Astronomy Abstract Service - Disponível em: http://adsabs.harvard.edu/abs/2010MNRAS.403.1036C - Acesso em: 01 de Agosto de 2014

DUNHAM, M.H. – Data Mining Introductory and advanced topics [S.I.]:Prentice Hall, 2003

SDSS - Disponível em: <http://www.sdss.org/> - Visto em 1 de Agosto de 2014

Wright, E. L. et al. The Wide-field Infrared *Survey* Explorer (WISE): Mission Description and Initial On-orbit Performance. v. 140, p. 1868–1881, dez. 2010.