Data Mining with Spectral Clustering and Comparative Analysis: The Boston Housing Dataset

JR Waggoner
College of Computing and Technology
Lipscomb University
Nashville, TN, USA

Laura Gao
College of Computing and Technology
Lipscomb University
Nashville, TN, USA

ABSTRACT

Although the Boston Housing Dataset is of one of a handful of default data sets used to introduce basic regression modeling methods to users, clustering or classification of the data are rarely explored. Using spectral clustering and various other clustering and tree-based techniques, we examine the housing data in search of evidence that may support the existence of distinct housing classes that may go unnoticed in regression-based methods. We continue our analysis by generating class labels for the data and evaluating the performance of our clustering methods against the now labeled data. We conclude our analysis by examining evidence that suggests there may be three distinct clusters of housing observations within the dataset.

KEYWORDS

Clustering, Unsupervised Learning, Boston Housing, Spectral Clustering, Python

1. Introduction

The Boston Housing dataset exists alongside the Titanic and Iris as the easy-to-work-with data when introducing basic statistics and data mining methods. The Boston data in particular is often referenced used in lessons introducing regression-type methods[1], and as a result most analysis of the data has been done from this perspective. We question whether analyses fixed on the continuous price output variable of the data misses the potential to explore the data in other interesting and meaningful ways. Specifically, we seek to understand if there are distinct groups of

Therefore, we've chosen to explore the data using what may be considered unconventional methods for this type of data - spectral clustering, kmeans, and decision tree classifier. Our goal is to discover if the dataset can be clustered (and classified) around distinct groups of housing observations that reveal something new about the now forty-year-old data[2]. We have a particular interest in the ability of spectral clustering analysis to reveal patterns within the data but, though it is the focus of our exploration, we will not limit our analysis to the use of just this tool. Throughout our exploration, we will make use of a number of data mining methods to both explore the data and as a relative measure of the overall effectiveness of our modeling techniques

1.1 Hypothesis

H_0 = The data are homogenous and distinct clusters do not exist in the data.

H_a = The data are not homogenous and distinct clusters may exist in the data.

1.2 The Data

The Boston Housing dataset was originally constructed by Harrison and Rubinfeld in 1978. It contains a total of 506 observations of housing and related data for towns in the Boston area. Fourteen attributes were measured for each observation:

- CRIM – Per capita crime rate by town.
- ZN – proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS – proportion of non-retail business acres per town.
- CHAS – Charles River dummy variable.
- NOX – nitric oxides concentration (parts-per-10million.)
- RM – Average number of rooms per dwelling.
- AGE – Proportion of owner-occupied homes units built prior to 1940.
- DIS – weighted distances to five Boston employment centers.
- RAD – index of accessibility to radial highways.
- TAX – Full-value property-tax rate per $10,000.
- PTRATIO – pupil-teacher ratio by town.
- B – $1000(bk-0.63)^2$ where Bk is the proportion of African Americans by town.
- LSTAT – Percent lower status of the population.
- MEDV – Median value of owner-occupied homes in $1,000s.

It is noted that the measurements of median value appear to be censored at 50 given the housing prices are of a continuous nature up to 50, while there are sixteen observations of exactly 50 and zero observations higher than 50[3].

All measurements are numeric, and the dataset does not contain any missing values. Two of the fourteen attributes can be considered as output variables to be modeled with data mining techniques: NOX and MEDV, though the median price value is the most commonly modeled.

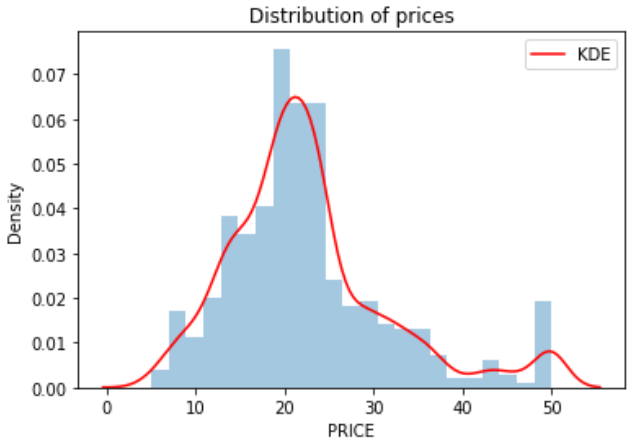

Figure 1: KDE and distribution of median home prices in the Boston Housing dataset.

Apart from generating new labels in the second half of our analysis, we used the data as-is without removing or transforming any of the variables.

2. Unsupervised Mining

To test our assumption that distinct clusters may exist in the data, we began our analysis by clustering and inspecting the results of a hierarchical clustering method.

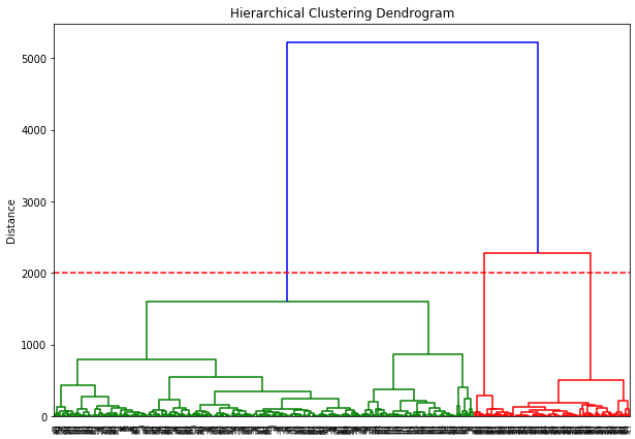

Figure 2: Hierarchical clustering dendrogram with

Inspection of the resulting dendrogram suggests the most well-defined clusters may occur when modeling for three or perhaps four clusters in the data. Further inspection reveals most of the observations fall into a single class, with the remaining observations split between two much smaller clusters.

We used the hierarchical clustering results to inform the remainder of our exploration of the data and the performance of a spectral clustering-based technique. To measure the relative performance of the algorithm, we compared its results to two other 'best-of-breed' data mining approaches: KMeans and Agglomerative Clustering. We used the Python library Scikit-learn's implementations of all three algorithms for our analysis[4]. For each of the three methods, we made three passes at the data and measured their ability to cleanly define two, three, and four discrete clusters in the housing data by computing and comparing the silhouette scores for each. The results of each pass are summarized in table 1.

	N = 2	N=3	N=4
KMeans	0.689184	0.720633	0.562631
Spectral	0.651076	0.274246	0.230872
Agglomerative	0.689184	0.718109	0.564658

Table 1: Result table of silhouette scores for each technique at N = 2, 3 and 4 clusters.

From the silhouette scores, we can draw several conclusions. First, there is little distinction between the three methods when the number of clusters is low. However, as the number of clusters increases, a very large gap forms between spectral and the other methods. This suggests spectral clustering may not be the most appropriate method for modeling distinct clusters in the data. Overall, the best silhouette results are obtained when modeling three clusters in the data, as our previous exploration with hierarchical clustering suggested. Kmeans appears to produce the most distinct clusters, though the heirerachical/agglomerative results were only a few thousands shy of kmeans. This small difference is most likely the result of our naive implementations of the algorithms, which all have room for further fine-tuning and improvement.

Finally, we applied the class labels generated by the kmeans algorithm to the data and visually inspected the nature and distribution of the labels.

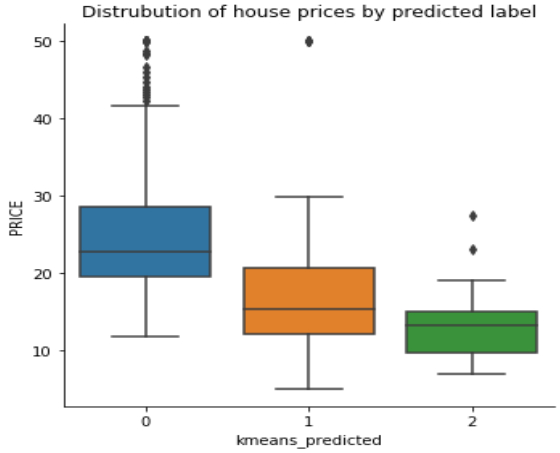

Figure 3: Distribution of house prices by predicted class.

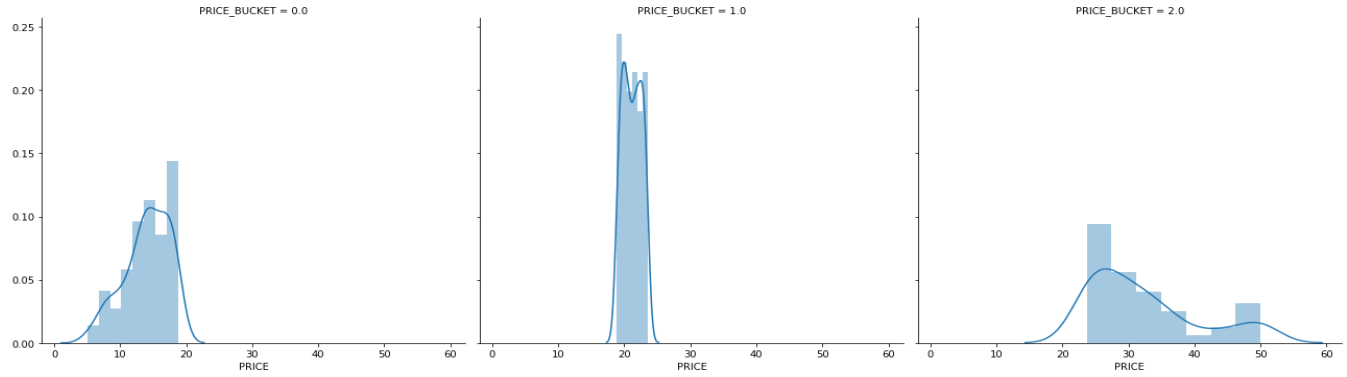

Figure 4. Distribution of prices by generated label

3. Semi-supervised Mining

With the knowledge gained from the unsupervised portion of our analysis, we turned our attention to modeling the data using generated labels. Though we considered and experimented with a number of labeling strategies, we decided to labelhe data based on the distribution of median housing prices. Observations were uniformly distributed between the three labels, see Figure 4.

Labeled-data in hand, we analyzed the ability of spectral clustering to accurately cluster and classify housing observations and compare its relative performance against two quintessential data mining methods – KMeans and a decision tree classifier. For each method, we rated the overall accuracy of the identified labels and generated confusion matrices to assess predicted labels. Finally, we tested each result set for significance with a chi^2 test, except for kmeans. The algorithm was only able to assign observations in the data to two of three generated classes and missed assigning any of the observations to the 'second' class, which the chi^2 test is unable to process.

3.1 Results

Both spectral clustering and kmeans methods struggled to accurately cluster and classify the data with respect to our generated labels, returning overall accuracies of 54% and 50%, respectively. The decision tree classifier, however, was able to much more accurately classify the labeled data and correctly classified 89% of the training data and 76% of unseen data withheld for testing. The confusion matrices for each run were as follows:

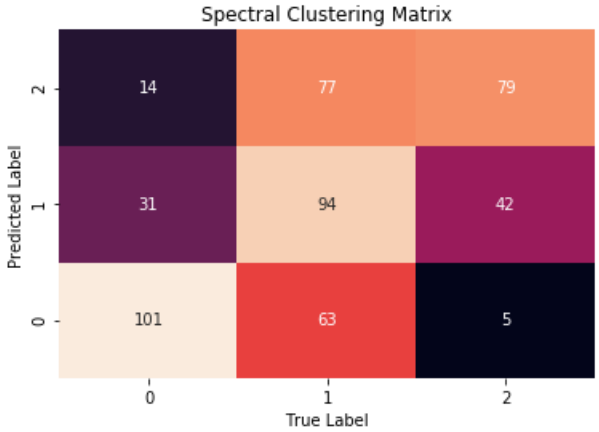

Figure 5. Spectral clustering confusion matrix.

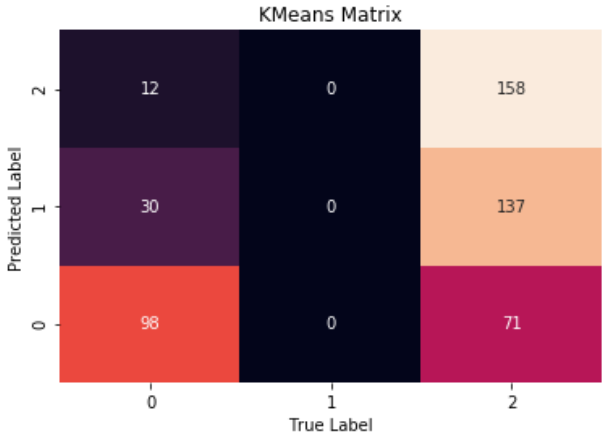

Figure 6. Kmeans confusion matrix.

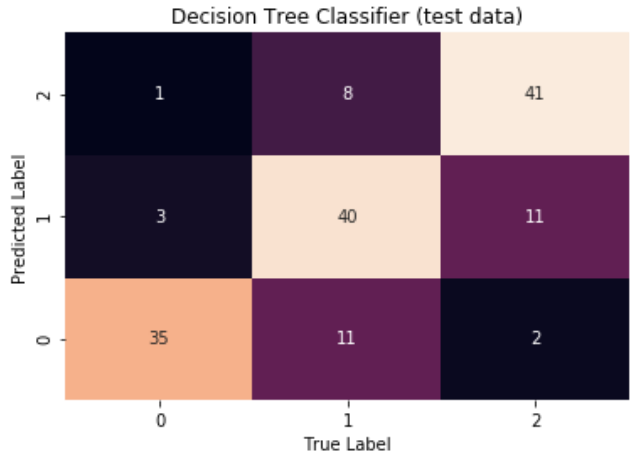

Figure 7. Decision tree classifier confusion matrix.

The results of our chi^2 tests for the spectral clustering and decision tree models are as follows:

```
Spectral Clustering Model p-value:
3.569659563600302e-33

Decision Tree (test) Model p-value:
2.599274332356367e-28
```

4. Conclusions

We set out to discover if distinct clusters existed in the Boston Housing dataset, and it appears that may be the case. Our unsupervised analysis suggests that the data may contain three distinct groups of housing observations, given the results of the kmeans and agglomerative algorithms. Spectral clustering fared much worse with respect to identifying the same three clusters. In the second half of our analysis, we demonstrated the imputation of class labels on the data that produced observations that diverge around PRICE = 21 and fall into low', 'medium', or 'high' price category. Of the three models tested on the labeled data, only the decision tree classifier produced results that could be considered accurate. Furthermore, because of the distribution of classes in the unsupervised portion of our analysis, we are confident that out naïve price bucketing approach to label generation was not ideal. Our technique generated equal numbers of observations across each class, but analysis of the hierarchical clustering results shows the data clusters are not distributed with this kind of uniformity. In future analysis, the use of generated labels will need to be more thoroughly worked through and should more closely match the distribution of class labels generated by unsupervised clustering of the data.

We believe the results of our analysis illustrate the benefits of approaching an old problem with a new perspective. Most simple analysis of the Boston Housing data involve regression-based techniques that preclude the exploration of distinct groupings within the data. Given the results of our analysis, we are compelled to reject our null hypothesis and conclude the data are not homogenous and distinct clusters may exist within the data.

REFERENCES

[1] SciPy-Lectures (2019). A Simple Regression Analysis on the Boston Housing Data. https://scipy-lectures.org/packages/scikitlearn/auto_examples /plot_boston_prediction.html.

[2] D. Harrison, D.L. Rubinfeld (1978). Hedonic prices and the demand for clean Air. Environmental Economics & Management, vol.5, 81-102.

[3] The Computer Science University of Toronto (1996). The Boston Housing Dataset. https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

[4] Scikit-learn (2019). 2.3 Clustering. https://scikit-learn.org/stable/modules/clustering.html

Appendix A. – Pairwise plot of input variables, shaded by KMeans predicted label.

Note – This view of the pairwise plot is included to help form a general impression of the distribution of predicted clusters among the data.