House Price Prediction Using Clustering and Genetic Programming along with Conducting a Comparative Study

1st Fateme Azimlu
Electrical, Computer, and Software Engineering Department
Ontario Tech University
Oshawa , Canada
fateme.azimlu@uoit.net

2nd Shahryar Rahnamayan
Electrical, Computer, and Software Engineering Department
Ontario Tech University
Oshawa , Canada
shahryar.rahnamayan@uoit.ca

3rdMasoud Makrehchi
Electrical, Computer, and Software Engineering Department
Ontario Tech University
Oshawa , Canada
masoud.makrehchi@uoit.ca

ABSTRACT

One of the most important tasks in machine learning is prediction. Data scientists use different regression methods to find the most appropriate and accurate model for each type of datasets. This study proposes a method to improve accuracy in regression and prediction. In common methods, different models are applied to the whole data to find the best model with higher accuracy. In our proposed approach, first, we cluster data using different methods such as K-means, DBSCAN, and agglomerative hierarchical clustering algorithms. Then, for each clustering method and for each generated cluster we apply various regression models including linear and polynomial regressions, SVR, neural network, and symbolic regression in order to find the most accurate model and study the genetic programming potential in improving the prediction accuracy. This model is a combination of clustering and regression. After clustering, the number of samples in each created cluster, compared to the number of samples in the whole dataset is reduced, and consequently by decreasing the number of samples in each group, we lose accuracy. On the other hand, specifying data and setting similar samples in one group enhances the accuracy and decreases the computational cost. As a case study, we used real estate data with 20 features to improve house price estimation; however, this approach is applicable to other large datasets.

KEYWORDS

Genetic Programming, symbolic Regression, Regression, Machine Learning, Clustering, Multi-level-model, House Price Prediction

ACM Reference Format:
1st Fateme Azimlu, 2nd Shahryar Rahnamayan, and 3rdMasoud Makrehchi. 2021. House Price Prediction Using Clustering and Genetic Programming along with Conducting a Comparative Study. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3449726.3463141

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07…$15.00
https://doi.org/10.1145/3449726.3463141

1 INTRODUCTION

Prediction is a key task in science and engineering. Scientists utilize various methods to discover patterns in a given dataset to predict future-related parameters. The invention of computers in the 1940s, revolutionized scientific prediction techniques because it empowered scientists to fulfill complicated calculations. Another revolution gained by the development of the internet after 1990, which generated a large amount of data that was not possible to be analyzed by using traditional methods. Therefore, data mining techniques were created as important tools to empower scientists in analyzing large amounts of exponentially growing data. Data mining is a collection of methods for recognizing hidden patterns in large data sets [8]. The first time a computer gaming and artificial intelligence scientist, Arthur Samuel, suggested the term "machine learning" in 1959 [9]. Machine learning (ML) consists of several algorithms, which are able to learn from data and predict desired variables [12]. ML includes various tasks such as regression, clustering, and classifying. Machine learning and data mining influenced many fields, including physics, health and medical science, social science, and economics. In a report published by the European Public Real Estate Association [1], it was shown that real-estate covers nearly 20% of economic activities. Price estimation including house price prediction is a very important task in economics, marketing, and even politics. Machine learning techniques have recently had an important role in price prediction [3], [15].

In machine learning, usually the predictive model is applied to the whole dataset. But when there is a high dispersion in some featured values, it is not easy to fit a proper model into a whole dataset. For example, in social or medical predictions there is a high dispersion in the age of the people or in one of the most famous machine learning problems, house price prediction, there is a high diversity in features such as size, age, and price of houses and dispersion in their values. Therefore, it is hard to fit a model into the whole data. In this paper, we have proposed a multi-level model which first, categorizes data with machine learning clustering methods before prediction, and assigns similar samples in different smaller sub-datasets. Then, applies different regression and prediction methods to each sub-dataset. Because we have customized the model for each cluster, we can enhance prediction accuracy. Moreover, we can examine various predictive models for each created sub-dataset and select the best model for each cluster. At the end, we can calculate the overall error for the whole dataset by averaging. In this study, we not only employed conventional regression and prediction techniques, we also utilised Genetic Programming (GP) as a symbolic

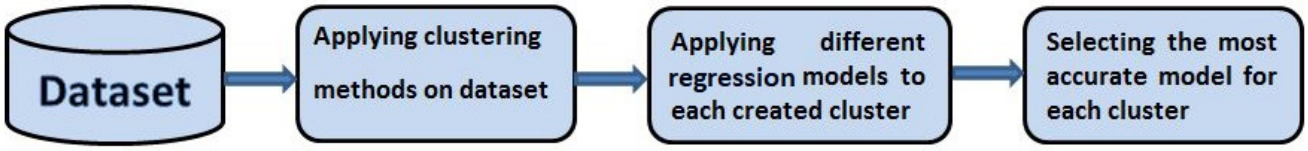

Figure 1: Summary of the proposed method.

regression tool and compare it with other prediction methods. GP is powerful in finding Hidden patterns in datasets with complicated structure and performs well in noisy datasets. In addition, Gp has lower limitation comparing to other regression methods which need high volume data to perform well. As a case study, we used house sales data for King County, located in the USA [2].

2 PROPOSED METHOD

When there is a high dispersion in differnt variable values in a dataset, fitting a single model to the whole dataset may not be easy. Especially when the dataset has a complicated structure or each part of the dataset has various structures, the created model for the whole dataset may not be reliable. Sometimes, in real world applications, especially when we have high volume of samples, a single model for the whole data may not be accurate. For example, when we model price estimation of all houses in range of prices in $100K to $10M. In addition, we compare a small house with a larger one, located in a big lot with many bedrooms, bathrooms, a swimming pool, and other features. In these cases, we can split data to sub-datasets (clusters) and apply customised models to each sub-dataset. To test our proposed method, we used different clustering and prediction techniques sequentially. The main contribution of this study is improving prediction accuracy by customising prediction models for sub-datasets or groups which are created based on similarity among the members of each cluster. The process is depicted in Figure 1. Therefore, if we group the instances to different categories based on the similarities and then find the best model for each group or cluster, we may be able to create accurate models. We can group the data points based on the most important features. For this task, we can rely on expert knowledge to know which variables are the most effective ones. For instance, for predicting house price, we can use a real-estate agent's information to know which feature is the main variable in order to estimate the house price. For example, location, size, number of bedrooms, etc. On the other hand, when we have various features, it may not be easy or possible for human to find the most similar instances for grouping task. In such conditions, data scientists utilize algorithms to fulfil this difficult task using ML clustering methods. In clustering techniques we create sub-sets of samples in such a way that members in the same cluster are more similar to each other than the objects in other sub-sets. There are many different clustering methods, which result in different types of clusters.

In the first stage of experiments, we apply different prediction methods, such as linear and polynomial regressions, Support Vector Regression (SVR), Artificial Neural Network, Genetic Programming (GP) and other methods on the entire data in order to find the best model with the lowest error in predicting the target value. For the second stage, we apply our proposed method. Our meta-model consists of two steps which are illustrated in Figure 2.

In the first step, we cluster or group data using **Algorithm-based clustering** or **Expert-based grouping**.

In **Algorithm-based clustering**, we use different machine learning clustering methods such as k-means, DBSCAN and agglomeration hierarchical clustering algorithms.

In **Expert-based grouping**, human groups data based on her/his knowledge.

In the second step of our proposed method, for each cluster or group, we apply different prediction techniques that we have used in the first stage for the whole data. Now we can compare the prediction accuracy of different models on each group and select the best one. Moreover, we can compare the best model after clustering, with the most accurate model in the first stage that is applied to the whole data before clustering. Many parameters affect each model's accuracy such as structure of the data, the number of features and size of the dataset (the number of data points in the dataset). Therefore, we examined the proposed method's performance on house sales data for King County in the USA. When a new sample, such as a new house, comes to the market and we need to predict the target value, first, based on the object's distance from the centroid of the clusters, we assign it to one of the generated clusters. Then, using the best model that previously found for the closest cluster to the new sample, we are able to predict the target value such as the house price.

To avoid over fitting, we used cross validation method. We splited data 70% for training and 30% for testing data. If the model has problem of over fitting in training data, we can not get a good accuracy in the test data.

2.1 Regression Models and Prediction Techniques

Regression is a supervised machine learning technique that creates a model or function "f" from the input variables "X" to predict output values of a desired target "Y" when the target values are continuous. Whenever there is a new input data (X), the output variable Y = f(X) is predicted value. There are numerous regression

and prediction techniques. Each method has its own importance, advantages, disadvantages, and limitations. We employed the most common approaches for the first stage of our research and selected the most accurate ones for the rest of our study.

Linear Regression: Linear regression is a supervised machine learning linear approach for modeling the relationship between dependent variable (Y) which is the target feature and one or more independent variables (X). If the best fit to this data is straight line. Lasso [18] and Ridge models are also linear models that are trained with L1 and L2 regularization approaches.

Polynomial Models: In some cases, the straight line which presents linear model, cannot fit the given data points. In this regressions we obtain low accuracy in RMSE and R^2 score for linear models. In some datasets, increasing the order of predicting model and as a result, improving the complexity of the model may solve the fitting problem. We only need to create a model similar to the linear regression and only add higher order of dependent variables to the equation.

Support Vector Regression (SVR): Support vector regression, SVR, considers the data as a series of points inside a space between the specific margin boarders. The model is a hyperplane with maximum margin such that maximum number of data points are located within that margin [5].

Artificial Neural Network (ANN): The ANN idea is inspired by human brain and its neural systems and simulates the brain learning. These systems learn to accomplish tasks by using examples without any specific programs or special rules similar to the brain learning procedure that learns from experience. For the first time, Warren Mc Culloch and Walter Pitts [13] suggested a mathematical model for neural networks using threshold logic algorithms. After that, many scientists developed regression and classification methods based on ANN. In 1992, White gathered many articles about ANN and advanced statistics [19]. In the early stage of using ANN, the computers were not powerful enough to accomplish ANN tasks beneficially but still many research works employed it for accurate regressions comparing to other conventional methods [14]. Today, ANN is one of the most important machine learning techniques.

Symbolic Regression In symbolic regression, we try to create the model which best fits the measured data [20]. In 1985, Cramer proposed one of the first tree-structured evolutionary algorithms that could be used in basic symbolic regression. John Koza [11] was the first person, who developed Genetic Programming in LISP, one of the earliest programming languages, and also it was shown that GP is a powerful tool in problem solving, including symbolic regression. In addition, John Koza showed GP can be applied in automatic functions by discovering an approximate value for the impulse response function in time invariant systems. It was a great improvement in machine learning regression methods [10].

2.2 Clustering Algorithms

Clustering is a fundamental technique extensively used for discovering the internal data structure in machine learning and recognizing hidden patterns. Clustering is an unsupervised machine learning task that partitions the data that has not been labelled, classified or grouped. Most of the conventional methods focus on modeling the similarity among data points. Based on the nature and structure of dataset, various clustering methods such as; K-means, DBSCAN, Mean-shift, Spectral clustering, and Hierarchical agglomerative clustering, should be examined to recognize the hidden patterns in data and its structure successfully [4].

K-means Clustering: K-means was the first clustering technique that we utilized. It is an extensively used clustering technique that minimizes the average squared distance between instances in the same cluster. This technique groups instances into a predefined number (k) of clusters based on the nearest mean distance of each data point to the cluster members [7].

DBSCAN Clustering: Density-based spatial clustering of applications with noise (DBSCAN), is a clustering algorithm which detects core samples of high density areas and expands clusters from them. It is practically beneficial for type of data that include groups of samples with similar density. First time, Martin Ester, Hans-Peter Kriegel et. al suggested this method in 1996 [6].

Hierarchical Agglomerative Clustering (HAC): Hierarchical clustering [16] is one of the most common clustering techniques. It builds clusters by combining clusters or splitting them. This hierarchy creating clusters can be illustrated by a tree shape or dendrogram.

2.3 Performance Assessment

In this study, we applied different prediction and regression techniques on house price dataset to find out if the proposed approach is practically able to improve predictions. For calculating accuracy, we used the relative absolute error which is calculated by the Eq. 1:

$$Er = \mid y_e - y_r \mid / y_r, \tag{1}$$

(where in Eq. 1, Er, y_r and y_e are the error, real value, and estimated value, respectively.)

If we consider the average error of n instances, we call it normalized mean absolute(NMA) error or NMAE which is one of the most common metrics for evaluating accuracy of continuous variables, Eq. 2:

$$Er = 1/n \sum_{j=1}^{n} \mid y_e - y_r \mid / y_r \tag{2}$$

NMAE measures the average value of the errors in predicted values, without considering their direction. It measures the absolute disparity between prediction and real value in the test sample of the absolute differences between prediction and real value. The advantage of using NMAE is that it put error in prediction into observable insight. In addition, it provides the error for the whole process.

3 EXPERIMENTS AND RESULTS

Our case study is house price prediction. As many variables affect the final price, high dimensionality of the house price datasets challenges the prediction of the price of each house. For our study, we used house sales data for King County in the USA [2]. It contains 21,614 instances and 20 features such as price, number of bedrooms, number of bathrooms, house size, floors (number of floors), condition, grade, building and renovation date and location. This is a multivariate dataset with both real and integer values with no sparsity. As illustrated in Figure 3, the mean price, μ, is 540,088.14

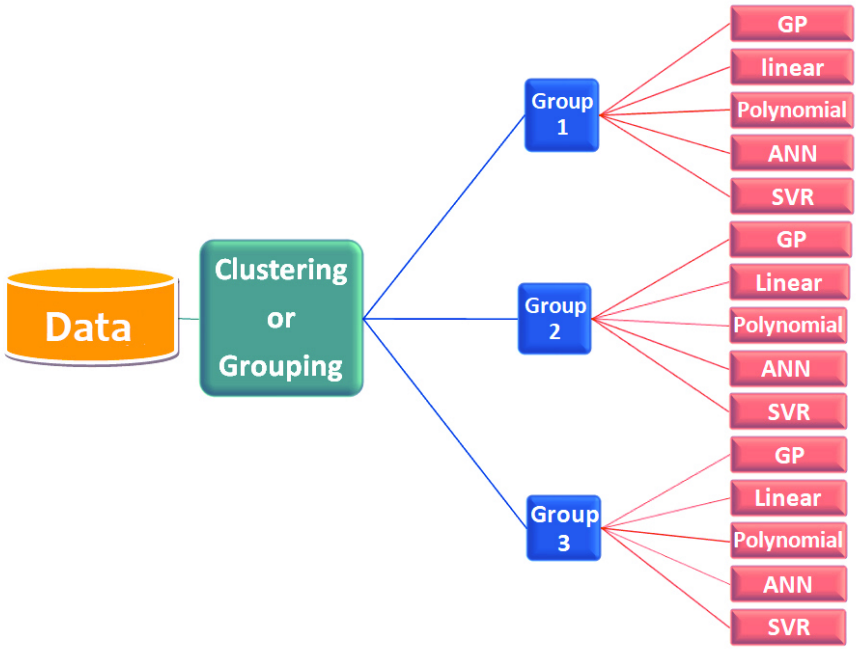

Figure 2: Configuration of the proposed method: after clustering, we apply different prediction techniques to each group.

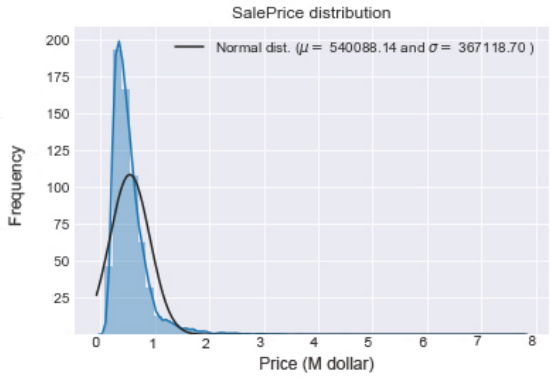

Figure 3: Comparison of price distribution in the house price dataset and the normal distribution. The prices are based on US dollar, μ is the mean price and σ is the data standard deviation. Both the plot shape and the σ value, confirm that this dataset has considerable dispersion and our data does not follow a normal distribution.

Dollar and the data standard deviation, σ, is 367,118.70 Dollar. Both the plot shape and the σ value, confirm that this dataset has considerable dispersion. Usually, a large dispersion, makes it hard to fit a simple model on entire data. In our case, the house price varies within four order of magnitudes. Therefore, our proposed method may work effectively for this dataset. In the first stage of the estimation task, various prediction models such as linear and polynomial regressions, SVR (shrinking heuristic and RBF kernel), ANN (multilayer perceptron), GP, and some other methods, are applied on the whole dataset. We were interested to find out how the results may change if we create smaller sub-sets of our data without any condition (no correlation between sub-sets members). Therefore, we randomly divided the dataset to smaller groups with 2000 and 1200 members to investigate the effect of the data size on prediction accuracy for each model. As expected, the accuracy declines by reducing data size, but the decreasing accuracy rate and improving the error, varies widely for different models. Table 1 and Figure 4 display the NMA errors for each prediction model. For calculating error, we consider the NMA error which is calculated by the Eq. 1 For symbolic regression, we create an equation using Eureqa [17] to create the models which have the best fit to our data. Eureqa not only has the ability to discover the functions, it also has the power to find the relevant coefficients of that function. Eq 3 is an example of one of the models that GP generated for the house price dataset.

$$price = a + (b + c * waterfront + d * grade + f * condition + g * sqftliving + h * sin(i + j * lat) - sin(k + l * long) * sin(m + n * lat))^p \quad (3)$$

(Where a=71287.47, b=8.655, c=4.1929, d=0.8507, f=0.3901, g=0.00108, h=1.59789, i=5.17567, j=11.532, k=6.277, l=7.101, m=5.1757, n=11.532, p=4.3536)

The next step is clustering data using K-means, DBSCAN, Hierarchical Agglomerative Clustering, and at last, Human Knowledge

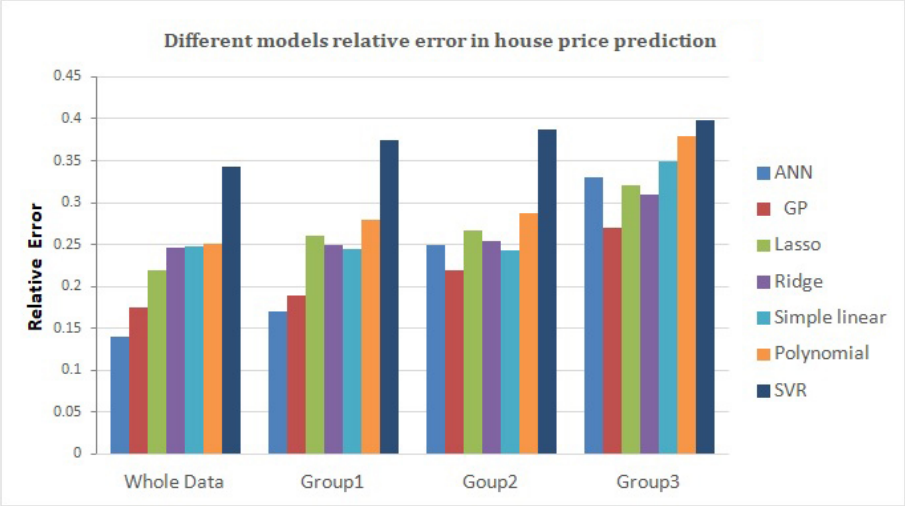

Figure 4: Different methods NMA error in house price prediction dataset and randomly created smaller sub-sets. Group 1 has 2000-2500 instances, Group 2 has 1200-2000 instances and Group 3 has 1200 instances.

Models:	ANN	GP	Lasso	Ridge	Linear	Polynomial	SVR
Whole Data	**0.14**	0.175	0.22	0.246	0.248	0.251	0.343
G1: 2000-2500	**0.17**	0.19	0.261	0.25	0.244	0.28	0.375
G2: 1200-2000	0.25	**0.22**	0.267	0.254	0.243	0.287	0.388
G3: 1200	0.33	**0.27**	0.32	0.31	0.35	0.38	0.399

Table 1: Different methods NMA error in house price prediction dataset and randomly created smaller sub-sets. Group 1 has 2000-2500 instances, Group 2 has 1200-2000 instances and Group 3 has 1200 instances.

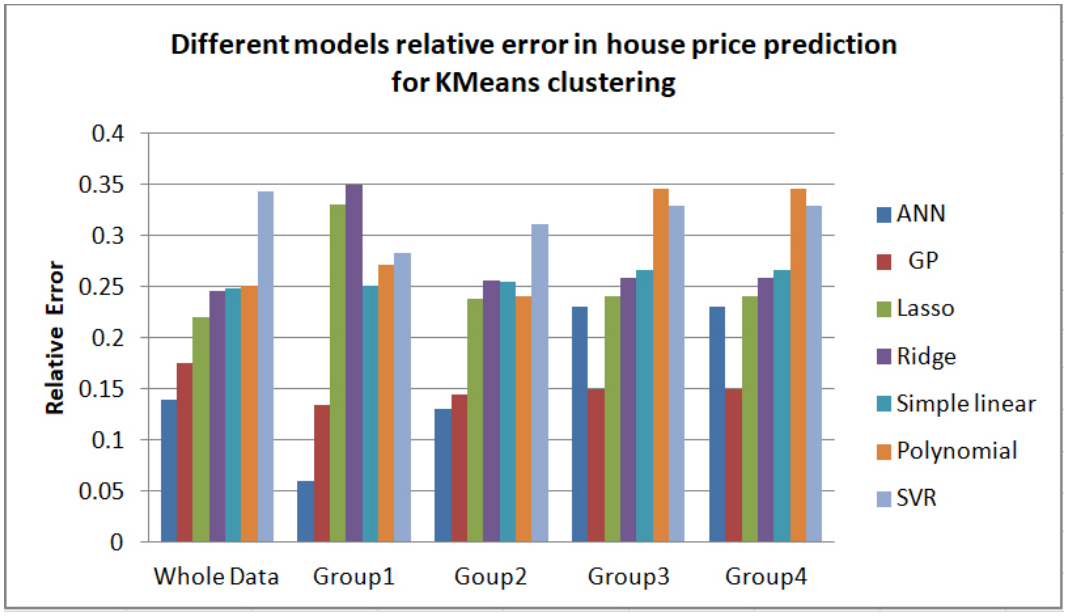

Figure 5: Different methods NMA error in house price prediction dataset and small sub-sets created by K-means . Group 1 has more than 2500 members, Group 2 has 2000-2500 instances, Group 3 has 1200-2000 members and Group 4 has 1200 samples.

Models:	ANN	GP	Lasso	Ridge	Linear	Polynomial	SVR
Whole dataset	**0.14**	0.175	0.220	0.246	0.248	0.251	0.343
G1: >2500 mem	**0.06**	0.14	0.2389	0.243	0.241	0.228	0.28
G2: 2000-2500	**0.13**	0.143	0.2389	0.2565	0.2544	0.2410	0.3109
G3: 1200-2000	**0.1445**	0.145	0.2365	0.2527	0.2517	0.271	0.3132
G4: <1200	0.23	**0.15**	0.2409	0.2588	0.2660	0.346	0.3282
Average	**0.09**	0.144	0.239	0.253	0.252	0.23	0.308
Overall	0.087						

Table 2: Different methods NMA error in house price prediction dataset and smaller sub-sets created by K-means . Group 1 has more than 2500 members, Group 2 has 2000-2500 instances, Group 3 has 1200-2000 instances and Group 4 has 1200 samples. If we select the best model for each cluster, the Overall error for prediction after clustering, is 0.087.

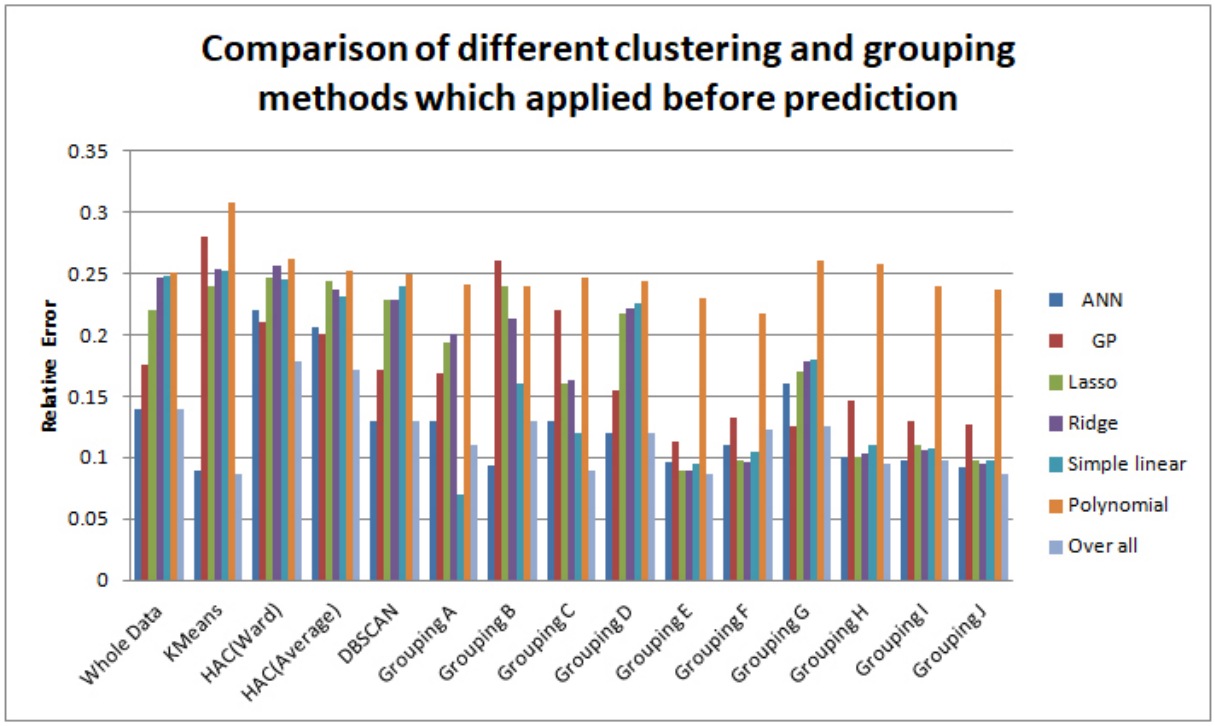

Figure 6: Comparison between different clustering and grouping methods effects on house price prediction. In this table, HAC has two linkage method: ward and average. Grouping A is grouping based on price. Grouping B is grouping Based on grade. Grouping C presents grouping based on predicted price. Grouping D is grouping based on location. Grouping E: 6 groups based on location and price. Grouping F: 9 groups based on location and price. Grouping G: 18 groups based on location, size and number of bedrooms. Grouping H: 12 groups based on location, size and number of bedrooms. Grouping I: 9 groups based on location and size. Grouping J: 9 groups based on location and size adding boarder members to the groups.

Base Grouping technique, to know how the prediction may change if we have smaller datasets which its members in a cluster have a higher similarity. We only illustrated K-means clustering results as a sample of clustering data before prediction. Table 2 and Figure 5 compare the NMA error for the employed models and the last row in Table 2, shows the average error for each model if we use clustering. In addition, if we select the best model for each group and then calculate the average, the Overall error for prediction through clustering, is **0.087**. K-means results, illustrated in Table 2, indicate an important point that in all sub-datasets which include large number of instances, ANN has better performance especially in the most voluminous cluster, its accuracy is surprising. All prediction models except Lasso, could decrease error for this cluster. None of linear models have better average error comparing to the entire dataset predictions. The most surprising results belongs to **GP**. For all K-means clusters it has very good performance and

Models:	ANN	GP	Lasso	Ridge	Linear	Polynomial	overall
Whole dataset	**0.14**	0.175	0.220	0.246	0.248	0.251	0.14
K-means	**0.09**	0.28	0.239	0.253	0.252	0.308	0.087
HAC,Ward	0.22	**0.21**	0.246	0.256	0.245	0.262	0.178
HAC,Average	0.206	**0.20**	0.243	0.236	0.231	0.252	0.172
DBSCAN	**0.13**	0.171	0.228	0.228	0.24	0.249	0.129
Grouping A	**0.129**	0.168	0.193	0.2	0.07	0.241	0.11
Grouping B	**0.093**	0.26	0.24	0.213	0.16	0.24	0.093
Grouping C	**0.13**	0.22	0.16	0.163	0.12	0.246	0.09
Grouping D	**0.12**	0.154	0.217	0.221	0.225	0.243	0.12
Grouping E	**0.096**0	0.113	0.089	0.09	0.095	0.23	0.087
Grouping F	0.11	0.133	**0.097**	0.096	0.104	0.217	0.123
Grouping G	0.16	**0.125**	0.17	0.178	0.18	0.26	0.125
Grouping H	**0.101**	0.146	0.10	0.103	0.11	0.257	0.095
Grouping I	**0.098**	0.13	0.11	0.106	0.108	0.24	0.098
Grouping J	**0.092**	0.127	0.098	0.095	0.097	0.237	0.086

Table 3: Comparison between different clustering and grouping methods effects on house price prediction. In this table, HAC has two linkage method: ward and average. Grouping A is grouping based on price. Grouping B is grouping Based on grade. Grouping C presents grouping based on predicted price. Grouping D is grouping based on location. Grouping E: 6 groups based on location and price. Grouping F: 9 groups based on location and price. Grouping G: 18 groups based on location, size and number of bedrooms. Grouping H: 12 groups based on location, size and number of bedrooms. Grouping I: 9 groups based on location and size. Grouping J: 9 groups based on location and size adding boarder members to the groups.

even its average accuracy is not lower than the best model before clustering, ANN. K-means clustering significantly has improved GP's performance in prediction. In the smallest sub-dataset GP is the most accurate method. As we observed a little improvement in accuracy after clustering data with K-means, we repeated the experiment with other clustering and grouping methods. For human Knowledge Base Grouping, we grouped data based on the selected features such as location (Zip code), size and other features.

Summary of all experiments results for house price prediction are illustrated in the Table 3 and Figure 6. Before clustering, neural network is the most accurate model with the error equal to 0.14 in price prediction. Therefore, if our method can predict with lower error, this confirms our hypothesis. Hierarchical Agglomerative clustering was not successful in improving accuracy compared to applying prediction models to the whole data. But compared to randomly selected sub-datasets it has a better performance and demonstrates the effect of similarity between the samples in each sub-dataset. DBSCAN somewhat enhanced prediction for ANN that decreased the error from 0.14 to 0.13 and overall to 0.129, which was not a remarkable improvement. The most successful clustering technique in our method is K-means, which decreases the overall error to 0.087. It reveals that K-means can detect similarity between data points in house price dataset. The only problem with K-means is its computational time that makes it an expensive technique especially when the dataset includes a large number of instances and has many features. Therefore, if we have time limitation, using K-means and GP in our method, may not be effective. Based on the results which show lower error after applying K-means, we can conclude that K-means which select similar data points based on their distances to the other data points, has acceptable performance for this dataset even if it has too many features. For house price data, DBSCAN could not improve the average accuracy, but it confirms that in small sub-dataset, we can trust GP to provide more accurate prediction comparing to other models. Moreover, weak performance of HAC clustering assert that all clustering techniques may not work effectively for our proposed method. Figures 4 represents the result of applying different models on the sub-datasets of house price prediction. As shown, when we apply different models to the whole data, neural network technique has better performance and offers a lower error in prediction. It reveals that the ANN model can be trained effectively when we have a sufficient number of samples. But when the number of instances is small, creating an appropriate model fails, while, even with a small number of samples, GP can still generate a model which both fits training data and predicts the test data with acceptable accuracy. Furthermore, it is surprising that simple linear model is less sensitive to the number of training data and linear predictions in grouping based on human knowledge, which we consider one or two features for grouping and adding boarder members to the sub-groups, have better performance compared to algorithmic clustering methods. This grouping scheme is faster than clustering techniques, but we need to examine all combinations of important features with a different number of members in created groups and this task is time consuming and computationally expensive. However, GP has larger overall error than some other models, it is surprising that GP presents a somewhat better performance in prediction when samples have a higher similarity. When we randomly create smaller datasets by decreasing the number of samples, this increases the error in all models. Polynomial and neural network regressions are apparently very sensitive to the sample size. Consequently, as GP can perform better in small

size datasets compared to other methods, in sub-datasets with a low number of instances, we can rely on GP predictions.

4 CONCLUSION REMARKS AND FUTURE WORKS

Generally speaking, we can conclude that K-means, which select similar data points based on their distances to the other data points, has an acceptable performance for house price data. For different datasets based on the nature and structure of the data, we need to examine different clustering methods. Moreover, using expert knowledge, grouping based on one or two features can be as accurate as the machine learning clustering. As in our case study we have a large number of features, linear and polynomial regressions do not fit well with the data. But in grouping schemes, in smaller groups, linear and polynomial models perform excellent. It means that even if the whole dataset fits with the complicated models, small sub-datasets can be considered as a linear or polynomial models. In these cases, fitting a model and predictions can be easier and faster with acceptable accuracy. In addition, in large sub-datasets, which we have enough training data, neural network is very successful to create accurate model, but when we cannot provide a large size training data, GP is powerful in creating accurate models. In our meta-model, after clustering in the first step, it is very probable to gain small-size sub-datasets, and conventional prediction methods are not efficient for these sub-datasets. For a large number of iterations in GP, if we apply a proper mutation and crossover rate, we have the chance to discover an appropriate model even for small-size datasets. In low size datasets, GP has the problem of over-fitting but if we select a large proportion of the data, (30 to 40 percent) as the test data, we can avoid over-fitting. Even if GP is computationally expensive, it can create symbolic models for regression with reasonable accuracy, especially in cases where we are unable to gather a large volume of instances, it outperforms other models. If we compare the GP's accuracy in the datasets, which include approximately the same number of instances, but differ in number of variables, as it is expected, regression accuracy decreases by increasing the number of features. Although, its accuracy is lower for high dimensional data, it can still perform more accurately in comparison with other models when we have low size dataset. Artificial Neural network is very powerful in detecting patterns and trends in complicated, or imprecise data. Generally, intricate patterns can not be discovered by human analysis or other computer algorithms. As it works like the human brain and needs to be trained, similar to human experts, its success is very dependent on the volume of relevant training data and as we demonstrated in Table1 by decreasing the data size its accuracy drops dramatically. Our experiments verified our hypothesis for a large dataset that if we split data in sub-datasets based on similarities between each group, or cluster data points, most models especially neural network, provide more accurate predictions. In addition, if we take the advantage of differnt models in each sub-dataset and select the best model for each sub-dataset, overall we can reduce the prediction error. The other advantage of proposed method is that customizing models for smaller sub-datasets with more similar samples comparing to the whole data, not only improves accuracy, but it also may decrease the computational complexity because a model, which is fitted to the entire data is more complicated than customized models. Moreover, after clustering, prediction procedure for sub-datasets can be done in parallel and for a smaller size sub-dataset, prediction process is faster.

Future Works

While our prediction method demonstrated higher accuracy in prediction compared to applying the models to whole data, there are still several cases and schemes that we would like to consider for future research.

There are several clustering methods and none of them can work efficiently for all types of datasets. Therefore, if we examine more clustering methods, we may be able to improve the prediction accuracy.

It is not possible to find a model which performs well for all type of datasets, but if we apply our model on other different datasets, we can figure out that proposed method can work accurately for which type of data with which characteristics.

REFERENCES

[1] [n.d.]. European Public Real Estate Association. http://alturl.com/7snxx.

[2] [n.d.]. House Sales Data set in King County, USA. https://www.kaggle.com/harlfoxem/housesalesprediction/version/1.

[3] Patrick Bajari, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. 2015. Machine learning methods for demand estimation. *American Economic Review* 105, 5 (2015), 481–85.

[4] Pavel Berkhin. 2006. A survey of clustering data mining techniques. In *Grouping multidimensional data*. Springer, 25–71.

[5] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[6] Martin Ester and Kriegel. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

[7] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108.

[8] M Kantardzic. 2003. Data Mining Concepts, Models, Methods, and Algorithms. A John Wiley & Sons. *Inc., Chichester* (2003).

[9] R Kohavi and F Provost. 1998. Glossary of terms: Machine learning. *30: 271* 274 (1998).

[10] J Koza, Martin A Keane, and James P Rice. 1993. Performance improvement of machine learning via automatic discovery of facilitating functions as applied to a problem of symbolic system identification. In *IEEE International Conference on Neural Networks*. IEEE, 191–198.

[11] John R Koza. 1994. Genetic programming as a means for programming computers by natural selection. *Statistics and computing* 4, 2 (1994), 87–112.

[12] John McCarthy and Edward A Feigenbaum. 1990. In memoriam: Arthur samuel: Pioneer in machine learning. *AI Magazine* 11, 3 (1990), 10–10.

[13] Warren S McCulloch and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5, 4 (1943), 115–133.

[14] Janardan Misra and Indranil Saha. 2010. Artificial neural networks in hardware: A survey of two decades of progress. *Neurocomputing* 74, 1-3 (2010), 239–255.

[15] Vahid Moosavi. 2017. Urban data streams and machine learning: a case of swiss real estate market. *arXiv preprint arXiv:1704.04979* (2017).

[16] Lior Rokach and Oded Maimon. 2005. Clustering methods. In *Data mining and knowledge discovery handbook*. Springer, 321–352.

[17] Michael Schmidt and Hod Lipson. 2013. Eureqa (version 0.98 beta)[software]. *Nutonian, Somerville, Mass, USA* (2013).

[18] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[19] Halbert White. 1992. *Artificial neural networks: approximation and learning theory*. Blackwell Publishers, Inc.

[20] Chi Zhang. [n.d.]. Genetic programming for symbolic regression. *University of Tennesse, Knoxville, TN* 37996 ([n. d.]).