Predicting Car Model Classifications and City Gas Mileage

Ronald Surban Fatalla
Masters Student

Department of Computer and Electrical Engineering
University of Iceland
Spring 2005

Contents

1. Introduction
   - Abstract

2. Identifying the Normalized Losses
   - Selected Data Inputs
   - Considering 8 factors
   - Genfis2 function
   - Predicted Output vs. Fuzzy Design Output

3. Prediction of Car City Gas Mileage
   - Nonlinear Regression Problem
   - Exhsrch function
   - Different Plot for Input and Output
   - RMS error
   - Data Distribution

4. Conclusion
5. Database

Introduction

Abstract. Auto Imports Database is system information that demands different applications in understanding automobiles properties, design, and value. Applications such as Linear Regression Equation, Predicting the price of car using both numeric and Boolean attributes, Instant-base learning algorithm and more have been utilize for this database purposes. Exploiting this database is important for future research application, learning and understanding how it works is beneficial for an individual choosing the best fit car for themselves.

Two different applications were used in this project. One is the classification of cars in determining the normalized losses from different sets of car attributes, mostly using the numeric values. The other application is the prediction of city gas mileage and using as many attributes that could affect automobiles gas mileage performance. Architectures such as ANFIS and subtractive clustering were used in designing the project. The increasing number of automobiles and the demands of it in the society certainly call for improving and understanding databases such as this one.

Many applications such as the prediction of an automobiles car gas mileage definitely help an individual seeking for a car that can be trusted and valued. Another such aspects that are beneficial is knowing the average or normalized losses in cars worth yearly. Different size classification of cars is indeed one factors in insuring ones car and understanding it would definitely benefit the user. Knowing what we need to fix or what we need to have for a certain car to be optimal in terms of insurance policy is an advantage to anyone.

1. Identifying the Normalized Losses

Selected Data Inputs Used in Car Classifications

	Input								Output
Car Make	Wheel-Base	Length	Width	Height	Engine Size	Horsepower	City-mpg	Highway-mpg	Normalized Losses
alfa-romero	88.6	168.8	64.1	48.8	130	111	21	27	142
audi	99.8	176.6	66.2	54.3	109	102	24	30	164
bmw	101.2	176.8	64.8	54.3	108	101	23	29	192
bmw	103.5	193.8	67.9	53.7	209	182	16	22	164
chevrolet	94.5	158.8	63.6	52	90	70	38	43	81
dodge	93.7	157.3	63.8	50.8	90	68	37	41	118
honda	96.5	169.1	66	51	110	100	25	31	107
isuzu	94.3	170.7	61.8	53.5	111	78	24	29	152
jaguar	102	191.7	70.6	47.8	326	262	13	17	125
mazda	93.1	159.1	64.2	54.1	91	68	30	31	104
mercedes-benz	112	199.2	72	55.4	304	184	14	16	125
mercury	102.7	178.4	68	54.8	140	175	19	24	123
mitsubishi	93.7	157.3	64.4	50.8	92	68	37	41	161
nissan	99.2	178.5	67.9	49.7	181	160	19	25	231
peugot	107.9	186.7	68.4	56.7	120	97	19	24	161
plymouth	95.9	173.2	66.3	50.2	156	145	19	24	152
porsche	94.5	168.9	68.3	50.2	151	143	19	27	186
renault	96.1	176.8	66.6	50.5	132	155	23	31	125
saab	99.1	186.6	66.5	56.1	121	110	21	28	150
subaru	96.9	173.6	65.4	54.9	108	111	23	23	85
toyota	95.7	158.7	63.6	54.5	92	62	35	39	87
volkswagen	100.4	183.1	66.9	55.1	109	88	25	31	132
volvo	104.3	188.8	67.2	56.2	141	114	23	28	103

Figure 1

As we see from the data shown above we used mostly the continuous values and the physical aspects of a car that could change overtime and has a direct effect on the normalized losses in an automobiles yearly value. There are 206 instances where the training data is the upper 150 cases and the remainder is used as the checking data. The way the data was divided could have affected the plot of the output value predicted which will be shown later on. Eight input attributes were used to determine the output and the figure 2 shows the plot of these 8 input attributes.

8 Input At

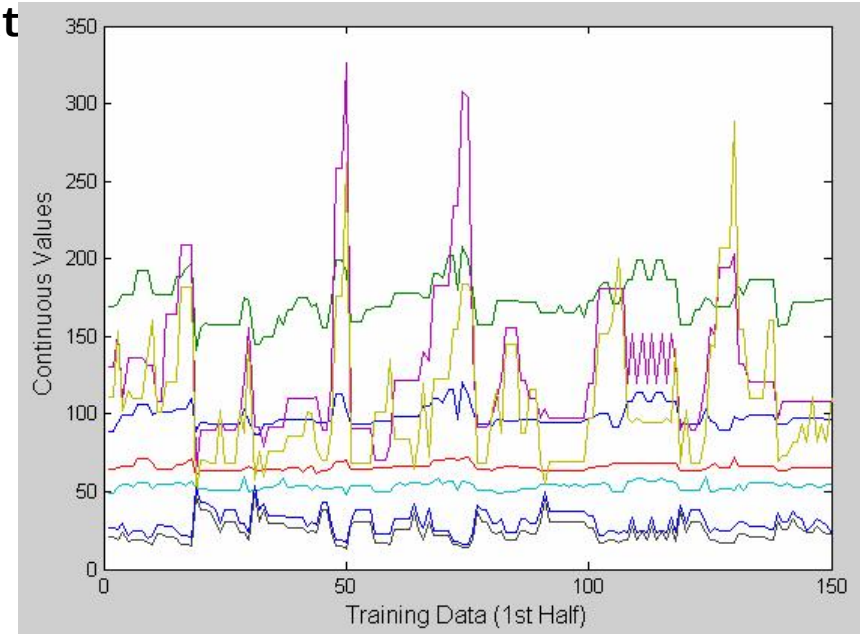

Figure 2

We can see from the graph that the input values are closely related and they are continuous values within the range of 0 to 350. The highest peaks are from the engine sizes, which vary raggedly from a small value to very large ones.

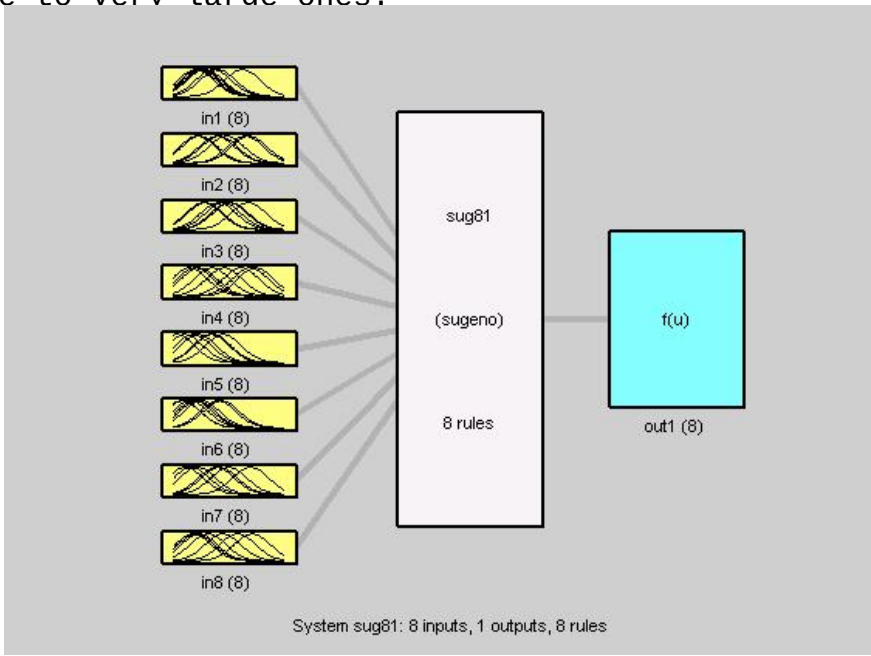

Figure 3

From the sets of inputs and the corresponding output “genfis2” is used to generate fuzzy inference system using subtractive clustering.

Genfis2 is used to initialize FIS or ANFIS training by first applying the subtractive clustering on the data. Using SUBCLUST to extract the set of rules that models the behavior of the data we can determine the number of rules and the antecedent membership function. The system generates 1 output from eight inputs and provided 8 rules. Subclust function simply finds the optimal data point to define the cluster centers, based on the density of the surrounding data points.

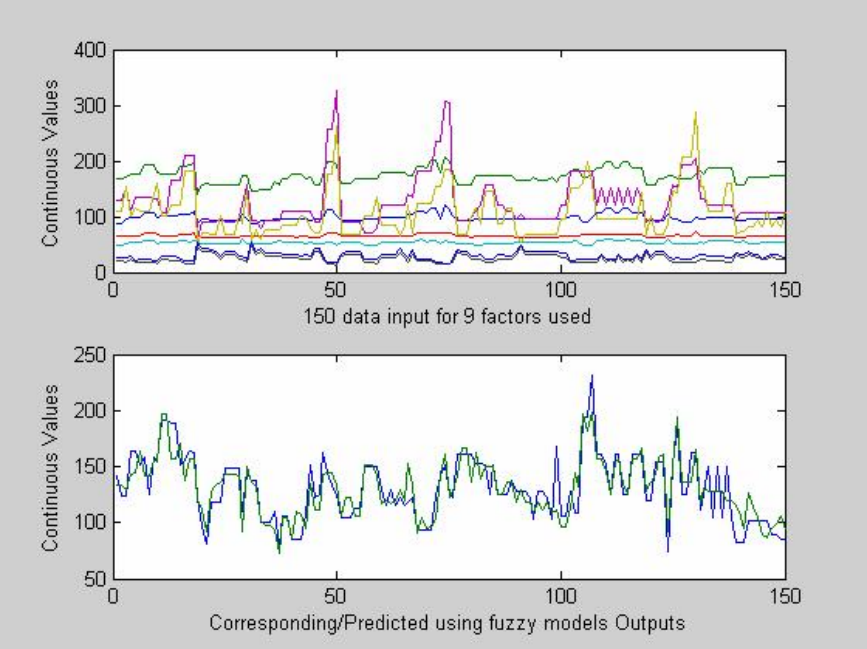

Figure 4

The graph above shows the first 150 instances with 8 inputs attributes while the graph below it displays the corresponding output and the output predicted by fuzzy model. The output graph have closely related curves but later on, we will determine how really close they are looking at the result values of the data. Looking at figure 4 could be deceiving in determining the output from different input attributes but the combination of “genfis2” and subtractive clustering makes the work load a little easier to do. We will compare the training data with the remaining data or the checking data and see how well it predicted the output from two different input sources.

The output data from figure 5 and figure 6 compares the training and checking data. A perfect prediction would show the values lying

along the diagonal line. Figure 6 is a useful measure of how good the model system is. The figures are not quite the same and the only reason is probably because of the division of the training and checking of data were most of the checking data corresponds to the sets of cars below the original database.

Output Gra

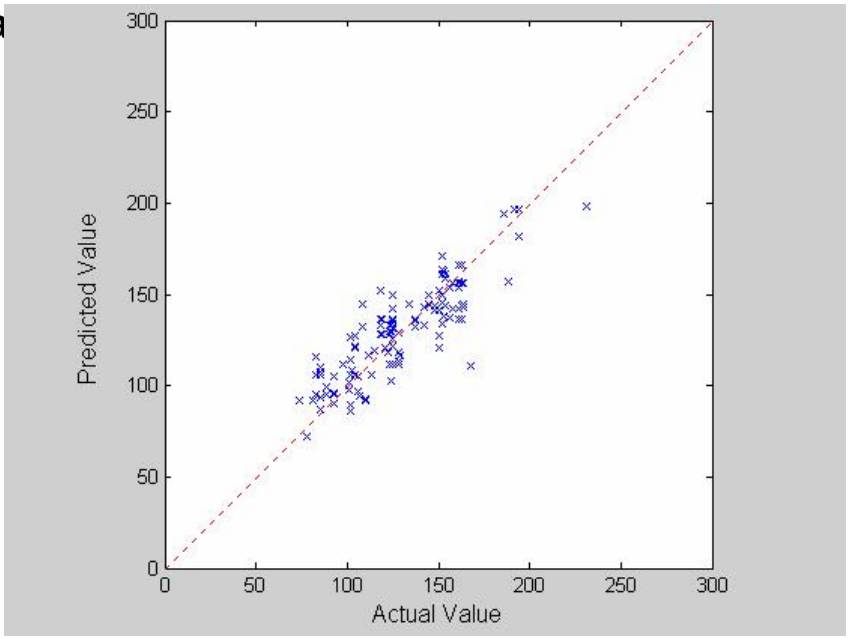

Figure 5

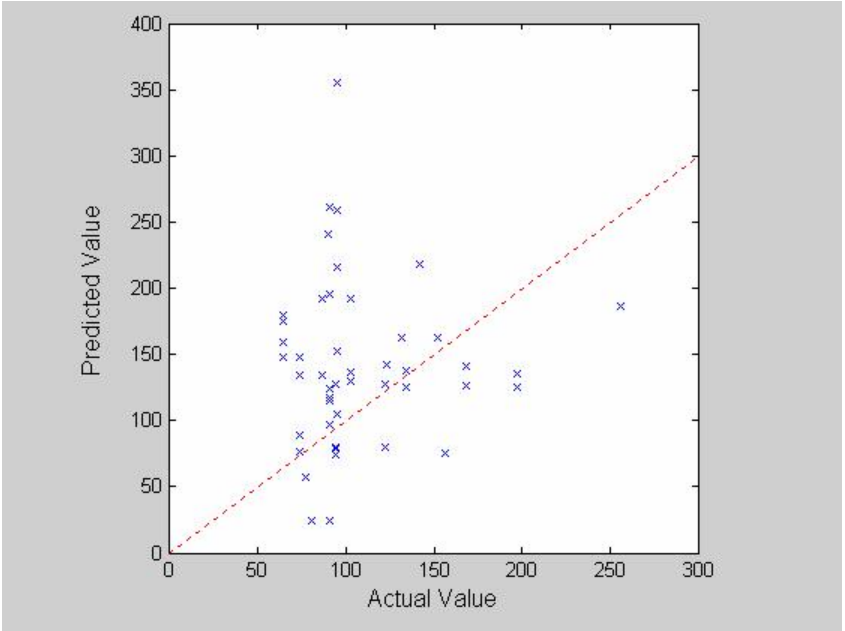

Figure 6

2. Predicting Cars City Gas Mileage

Selected Data Used

	Input Attributes											Output
Car Models	Fuel Type	Aspiration	Wheel-Base	Width	Height	Curb-Weight	Engine-Type	No. Cylinders	Engine Size	Fuel System	Horsepower	Peak rpm	City mpg
alfa-romero	gas	std	88.6	64	49	2548	dohc	four	130	mpfi	111	5000	21
audi	gas	std	99.8	66	54	2337	ohc	four	109	mpfi	102	5500	24
bmw	gas	std	101	65	54	2395	ohc	four	108	mpfi	101	5800	23
chevrolet	gas	std	88.4	60	53	1488	l	three	61	2bbl	48	5100	47
dodge	gas	turbo	93.7	64	51	2128	ohc	four	98	mpfi	102	5500	24
honda	gas	std	86.6	64	51	1713	ohc	four	92	1bbl	58	4800	49
isuzu	gas	std	94.3	62	54	2337	ohc	four	111	2bbl	78	4800	24
jaguar	gas	std	113	70	53	4066	dohc	six	258	mpfi	176	4750	15
mazda	gas	std	93.1	64	54	1890	ohc	four	91	2bbl	68	5000	30
mercedes-benz	diesel	turbo	110	70	57	3515	ohc	five	183	idi	123	4350	22
mercury	gas	turbo	103	68	55	2910	ohc	four	140	mpfi	175	5000	19
mitsubishi	gas	std	93.7	64	51	1918	ohc	four	92	2bbl	68	5500	37
nissan	gas	std	94.5	64	55	1889	ohc	four	97	2bbl	69	5200	31
peugot	gas	std	108	68	57	3020	l	four	120	mpfi	97	5000	19
plymouth	gas	std	93.7	64	51	1918	ohc	four	90	2bbl	68	5500	37
porsche	gas	std	94.5	68	50	2778	ohc	four	151	mpfi	143	5500	19
renault	gas	std	96.1	67	55	2579	ohc	four	132	mpfi	na	?	23
saab	gas	std	99.1	67	56	2658	ohc	four	121	mpfi	110	5250	21
subaru	gas	std	93.7	63	54	2050	ohcf	four	97	2bbl	69	4900	31
subaru	gas	turbo	96.9	65	55	2650	ohcf	four	108	mpfi	111	4800	23
toyota	gas	std	95.7	64	55	1985	ohc	four	92	2bbl	62	4800	35
volkswagen	diesel	std	97.3	66	56	2261	ohc	four	97	idi	52	4800	37
volvo	gas	std	104	67	56	2912	ohc	four	141	mpfi	114	5400	23

Figure 7

Predicting the cars city gas mileage depending on several attributes is not an easy task. From the original 26 different attributes that was provided by the database only few can be used in predicting the output. The hassle of missing attributes, Boolean attributes, and non-numeric data are several inconveniences that can be encountered in this design.

Other features that we need to consider are factors which directly affect the performance of the cars gas mileage. Mostly physical features but as well as type, model, and extra stuff such as qualities should be regarded in determining the car mileage. Another problem is the term “standard” or simply what is normal from this car company compare to the others. Similar sets of horsepower from one car type does not necessarily produced similar sets of weights, varying from one car model to another. Starting from huge sets of inputs the attributes were cut down into 12 candidates as inputs which are the most influential factors in determining the output of this design.

The Task

Finding 1 input from 12 candidates

Figure 8

Need to determine the best attributes from given candidates. Using the function “exhsrch” it constructs 12 ANFIS each with a single input attribute. We see from figure 8 that logically “horsepower” is one of the most influential attributes that affects mileage performance. The next most influential attribute is the curb-weight which again makes sense since driving a heavier load greatly effect the distance a particular car could travel. The training and the checking data were selected as the odd given data versus the even data. The training and the checking errors are comparable in size which simply implies that there is no overfitting and we can select more input variables.

Input-Out odel

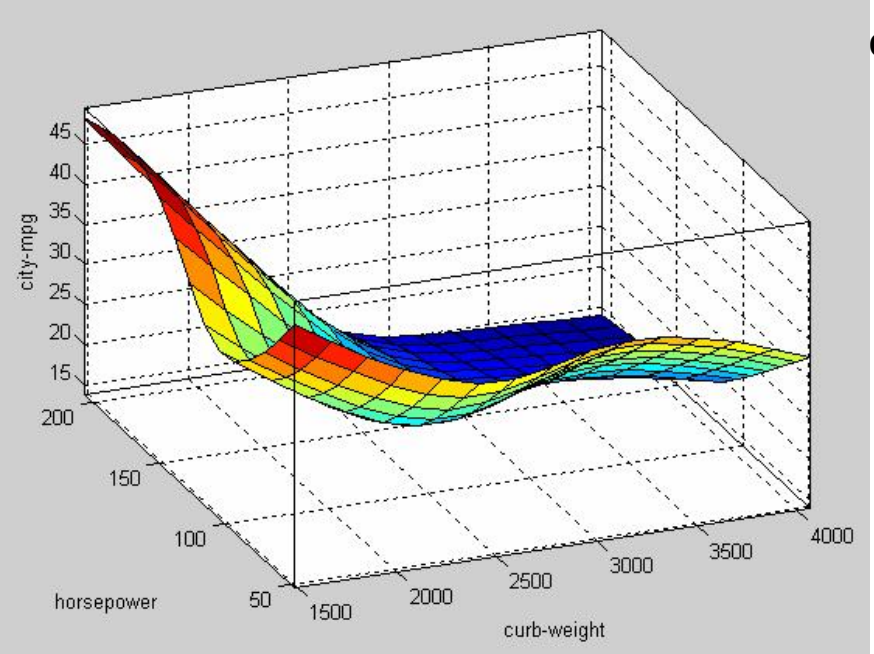

Figure 9

The resulting graph showed the horsepower and curb-weight as the best 2-input variables for ANFIS model. We can verify that the increase in city mpg is determined from a balance increase of curb-weight and an increase in the horsepower of the car.

Figure 10

Figure 10 is the output distribution of data. The lack of both the training and checking data is explicit, shown on the left upper

corner of the graph. These lacks of data is most likely the reasons why figure 9 behaves in a strange manner. The selected best two input attributes for ANFIS is reasonable but there are problems regarding the distribution of data from the two selected inputs. The training root mean square error is about 2.578 while the checking root mean square error is 3.008, in comparison a simple linear regression using all the input candidates training rmse is 3.007 while checking rmse is 3.177.

Conclusion

Clustering is a very effective way of dealing with large sets of data. Given a multi-dimensional data we can predict accurately an outputs, also a group of data using other sets of data.

The used of ANFIS architecture in the project showed a very good application for fuzzy inference system. Having large sets of data, we can use multiple inputs and select the best combination of inputs in arriving at the output.

Database:

1. Title: 1985 Auto Imports Database

2. Source Information:
   -- Creator/Donor: Jeffrey C. Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu)
   -- Date: 19 May 1987
   -- Sources:
     1) 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.
     2) Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038
     3) Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

3. Past Usage:
   -- Kibler,~D., Aha,~D.~W., \& Albert,~M. (1989). Instance-based prediction
      of real-valued attributes. {\it Computational Intelligence}, {\it 5}, 51--57.
        -- Predicted price of car using all numeric and Boolean attributes
        -- Method: an instance-based learning (IBL) algorithm derived from a localized k-nearest neighbor algorithm. Compared with a linear regression prediction...so all instances with missing attribute values were discarded. This resulted with a training set of 159 instances, which was also used as a test set (minus the actual instance during testing).
        -- Results: Percent Average Deviation Error of Prediction from Actual
           -- 11.84% for the IBL algorithm
           -- 14.12% for the resulting linear regression equation

4. Relevant Information:
   -- Description
      This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

      The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.

   -- Note: Several of the attributes in the database could be used as a "class" attribute.

5. Number of Instances: 205

6. Number of Attributes: 26 total

-- 15 continuous
-- 1 integer
-- 10 nominal

7. Attribute Information:

Attribute:	Attribute Range:
1. symboling:	-3, -2, -1, 0, 1, 2, 3.
2. normalized-losses:	continuous from 65 to 256.
3. make:	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type:	diesel, gas.
5. aspiration:	std, turbo.
6. num-of-doors:	four, two.
7. body-style:	hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels:	4wd, fwd, rwd.
9. engine-location:	front, rear.
10. wheel-base:	continuous from 86.6 120.9.
11. length:	continuous from 141.1 to 208.1.
12. width:	continuous from 60.3 to 72.3.
13. height:	continuous from 47.8 to 59.8.
14. curb-weight:	continuous from 1488 to 4066.
15. engine-type:	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders:	eight, five, four, six, three, twelve, two.
17. engine-size:	continuous from 61 to 326.
18. fuel-system:	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore:	continuous from 2.54 to 3.94.
20. stroke:	continuous from 2.07 to 4.17.
21. compression-ratio:	continuous from 7 to 23.
22. horsepower:	continuous from 48 to 288.
23. peak-rpm:	continuous from 4150 to 6600.
24. city-mpg:	continuous from 13 to 49.
25. highway-mpg:	continuous from 16 to 54.
26. price:	continuous from 5118 to 45400.