

# Market Sector Selection Using Fuzzy Regression Trees

Richard E. Haskell

Computer Science and Engineering Department  
Oakland University  
Rochester, Michigan, USA 48309  
email: haskell@oakland.edu

## Abstract

Keywords: Trading Systems, Forecasting, Machine Learning, Decision Trees, Fuzzy Logic

Regression trees are binary decision trees that can predict continuous outputs based on historical training data. The trees are learned by assigning continuous output values to fuzzy classes. The nodes of the tree are made fuzzy such that test samples dropped through the tree will, in general, traverse all branches of the tree. Multiple trees are created for each of 31 Fidelity select funds with the goal of predicting the 13-week percent gain in the fund's net asset value.

## Introduction

There are many different strategies for investing in the stock market ranging from "buy and hold" to daily trading. The assumption made in this paper is that some portion of a portfolio will be invested in equities all the time. The only question then is which equities to own at any particular time. This paper will be concerned with relatively infrequent trades -- once per quarter rather than daily or weekly trades.

Market sectors as represented by the Fidelity select funds undergo continual rotation that provides an investment opportunity. These select funds each invest in a particular market sector such as computers, energy, autos, financial services, health, retail, and so forth. Currently there are 39 Fidelity select funds with more being added all the time. Thirty-one of these funds have been in existence since at least 1988.

The potential benefits of being able to choose the best Fidelity select funds at different periods of time are illustrated by Table 1 which shows the 13-week percent gains for the best and worst of the 31 Fidelity select funds for the four quarters of 1994. Note that the Select Construction and Housing fund goes from last place in the 1<sup>st</sup> quarter to first place in the 2<sup>nd</sup> quarter and the Select Software fund goes from last place in the 2<sup>nd</sup> quarter to first place in the 3<sup>rd</sup> quarter. However, the Select Software fund remains in first place in the 4<sup>th</sup> quarter. A more detailed look at the complete rankings in

the four quarters of 1994 will show that the top five funds in the second quarter are completely different from the top five funds in the first quarter. In fact, a total of 15 different funds appear in the top 5 in the four different quarters.

If one had invested equal amounts in the top five funds for each of the four quarters in 1994, then the total annual percent gain would have been  $5.69\% + 12.19\% + 17.73\% + 4.12\% = 39.73\%$ . This is to be compared with the actual annual percent change of the S&P 500 in 1994 of  $-1.67\%$ . The Dow increased only  $1.44\%$  in 1994. The best select fund in 1994 was Computer with a total gain of  $18.6\%$ . The worst select fund in 1994 was Air with a total gain of  $-31.6\%$ . Only 10 of the 31 select funds had positive gains in 1994.

It is clear that there would be a big payoff if one could train the computer to predict the top-performing Fidelity select funds for the next quarter. This paper describes a method of using fuzzy regression trees to predict the rankings of Fidelity select funds 13 weeks into the future.

Table 1

1994	Fund	13-week %gain
1 <sup>st</sup> Q: Best	Electronics	9.4
Worst	ConstrHousing	-33.4
2 <sup>nd</sup> Q: Best	ConstrHousing	20.8
Worst	Software	-22.2
3 <sup>rd</sup> Q: Best	Software	19.8
Worst	Broker	-5.3
4 <sup>th</sup> Q: Best	Software	6.7
Worst	HomeFinance	-22.5

## Fuzzy Regression Trees

In a binary tree classifier a decision is made at each non-terminal node of the tree based upon the value of one of many possible attributes or features [1,2]. If the feature value is less than some threshold then the left branch of the tree is taken, otherwise the right branch is taken. The leaves, or terminal nodes, of the tree represent the various classes to be recognized. If the classes to be recognized are distinct, we refer to the tree as a classification tree. If a continuous output variable is to be predicted, we refer to the tree as a regression tree.

In a fuzzy regression tree we assume that each sample can belong to all classes to varying degrees. For example, if the percent gains of a stock ranged from  $-12$  to  $+12$  then any value between  $-12$  and  $+12$  could be assigned degrees of membership in the seven fuzzy classes shown in Figure 1. For example, a percent gain

of +5 will belong to the fuzzy set 4 to degree 0.75 and to fuzzy set 8 to degree 0.25 as shown in Figure 1.

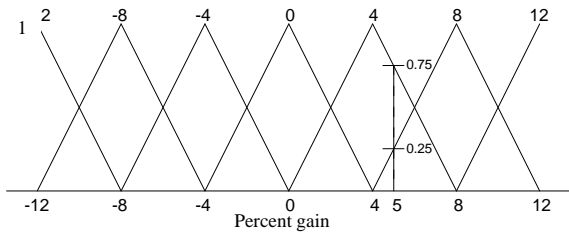


Figure 1 Converting continuous values to fuzzy classes

In a binary classification tree an error occurs when a test sample takes the wrong branch while being classified by the decision tree. This can occur due to noise in the test and/or training samples. Making the decision at each node a fuzzy set given by the membership functions shown in Figure 2 can minimize this problem.

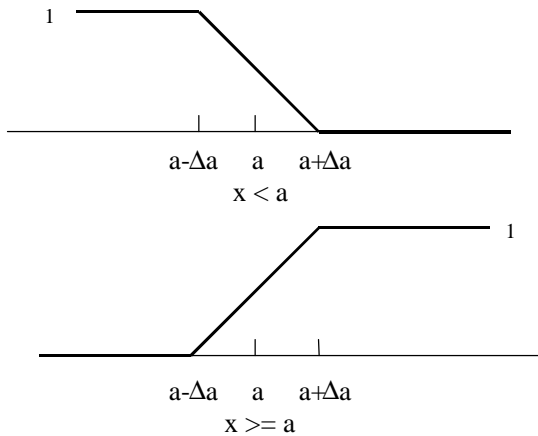


Figure 2 Fuzzy logic membership functions for  $x < a$  and  $x \geq a$

The left branch from a node has the weight,  $w_l$ , given by the membership function  $x < a$  associated with it, and the right branch from a node has the weight,  $w_r$ , given by the membership function  $x \geq a$  associated with it. When a test sample is being classified, both branches from the node are followed with the appropriate weight associated with each branch depending on the value of  $x$ . Note that if  $x$  is more than  $\Delta a$  from the threshold  $a$ , the weight value will be either 1 or 0.

The value  $\Delta a$  shown in Figure 2 is determined by a single user-defined parameter,  $p$ , called the *fuzzy percent*. The parameter  $p$  is the percent of the full range of values for a given feature within the subspace associated with a particular node. The fuzzy percent,  $p$ , is a measure of the fuzziness of the resulting tree. Increasing this fuzzy

percent often improves the performance of real-world problems as will be shown in the stock market example in the next section.

The method of producing such fuzzy regression trees from training data has been described in [3] and [4]. It uses a generalization of the Kolmogorov-Smirnoff distance for multiple classes to split the nodes of the tree [5, 6].

## Fidelity Select Funds

A database was established containing the weekly net asset value (NAV) for 31 select funds from 1988 to the present. One approach that was taken initially was to build a regression tree for each fund that would predict its 13-week percent gain. These percent gains could then be ranked in decreasing order from which the top funds would be selected.

An alternate approach that is used in the experiment described in this paper is to first create a new database containing ranking data. Each week the 31 select funds are ranked on their most recent 13-week percent gain on a scale of 1 - 31. The fund with the highest percent gain is ranked number 1 and the fund with the lowest percent gain is ranked number 31. This weekly ranking data was collected for all 31 funds from 1988 - 1998. An example of this ranking data for the two select funds Food and Computer for the year 1997 is shown in Figure 3. These plots, which are typical of all the select funds, show that a fund's ranking will vary from near the top to the list (1) to near the bottom of the list (31) over a period of several weeks.

This ranking data was used to build a separate regression tree for each select fund using eight years of weekly training data from 1989-1996. Ten features were used to build the tree for each fund. Four of the features were the fund's ranking at the current week as well as 13, 26, and 39 weeks in the past. The other six features were the current week's ranking for the six funds: Leisure, Transportation, Brokerage, Biotechnology, Computers, and Energy Services. These funds were designed to represent the six major sector categories: consumer, cyclicals, financial services, health care, technology, and utilities/natural resources.

The last day of training data was 12/27/96. However, we needed to use the data from the first quarter of 1997 because we needed to know the rankings 13-weeks in the future in order to build the regression trees. This means that the first test date that we could realistically use to test the trees would be 4/4/97. If we tested the trees using the date 12/27/96 we should expect good results because that date was part of the training data. In fact, the predicted ranking was almost identical

to the actual ranking as shown by the results in Figure 4. In this figure the funds are listed in their predicted ranking order. Their actual ranking number is also given and plotted by the stars at the top of the figure. Note that the actual rankings are numbered 0 - 30 rather than 1 - 31. A perfect prediction would yield a straight diagonal line of the type shown in Figure 4.

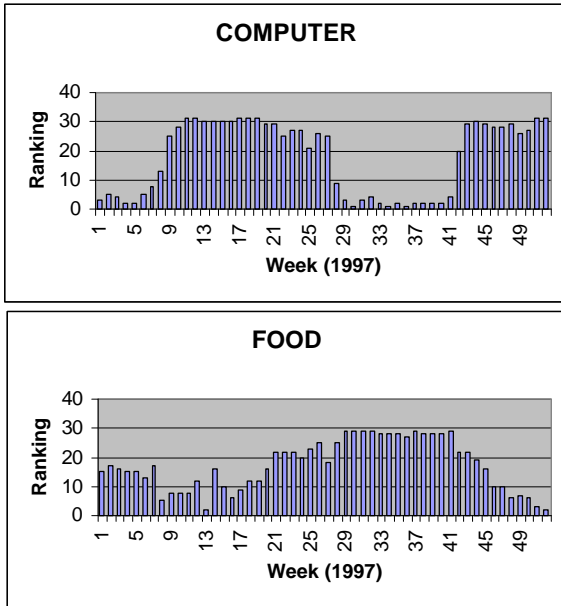


Figure 3 Ranking data for two of the select funds

The % Gain column in Figure 4 represents the actual percent gain over the next 13 weeks. The five averages at the bottom of Figure 4 are the percent gains that would be realized if one were to invest in the top 2, 3, 5, 7, and 9 funds in the predicted list. These can be compared with the “Average of all stocks” value at the very bottom of Figure 4. This would be the percent gain if one were to invest equally in all 31 select funds.

If we now test the tree on a date (7/4/97) not used for the training we get the discouraging results shown in Figure 5. The stars seem to be distributed pretty much randomly. We are looking for stars in the upper left corner with none in the upper right corner. While there are three stars in the upper left region there are also some in the upper right. In particular, note that MedDel, Health, and PrecMet are all low ranking funds in the upper right region. The actual average percent gains corresponding to Figure 5 are as follows:

Averages:  
 Top 2: 16.8757  
 Top 3: 12.6960  
 Top 5: 14.0855  
 Top 7: 15.8468  
 Top 9: 16.4937

Average of all stocks: 11.8503

The results in Figure 5 were obtained with a 0% fuzzy value at each tree node. If we change this to 20% fuzzy we obtain the results shown in Figure 6. Note that the stars are beginning to form a more diagonal line. In particular, note that MedDel, Health, and PrecMet have all moved to near the bottom of the list. The actual average percent gains corresponding to Figure 6 are as follows:

Averages:  
 Top 2: 16.8757  
 Top 3: 16.8313  
 Top 5: 17.5223  
 Top 7: 17.7297  
 Top 9: 17.5678  
 Average of all stocks: 11.8503

Our experience has been that if the test data does not closely follow the training data (which is the case in most real-world situations) then making the regression trees fuzzy almost always improves the overall performance.

Another way to improve performance is to use more than one tree for each fund. These different trees could use different training data, different features, or a different number of classes. For example, the results shown in Figures 5 and 6 converted the ranking data (with a range of 1 – 31) to nine fuzzy classes of the type shown in Figure 1. If we build another set of trees using seven classes then a different set of plots, similar to Figures 5 and 6, will be produced. The actual average percent gains corresponding to 20% fuzzy for this seven-class case are as follows:

Averages:  
 Top 2: 12.0742  
 Top 3: 12.4145  
 Top 5: 10.7493  
 Top 7: 13.9826  
 Top 9: 13.6792  
 Average of all stocks: 11.8503

Note that these values are significantly lower than the corresponding values for the 20% fuzzy nine-class case. However, if we average the predicted rankings from the two trees, we obtain the result shown in Figure 7. Note that the Top 2 and Top 3 averages are both higher than the Top 2 and Top 3 averages in each single tree. This is because funds that are highly ranked in both trees will tend to move to the top of the list. Note in Figure 7 that the top four predicted funds are all in the top 6 actual funds. Our experience has been that using multiple, fuzzy regression trees both improves performance and makes the predictions more robust.

As a final test, three sets of trees were made using 7, 9, and 11 classes. The rankings of all 31 funds were predicted for each of the 4 quarters starting on 4/4/97, 7/4/97, 10/3/97, and 1/2/98. For each quarter we calculated the percent gain realized if one were to invest

in the top 2, 3, 5, 7, and 9 funds in the predicted rankings. We then added the percent gains for each of the four quarters to get an annual percent gain. These results are summarized in Figure 8. Note that investing in the top 2 to top 7 predicted funds will more than double the annual return obtained by investing in all of the select funds.

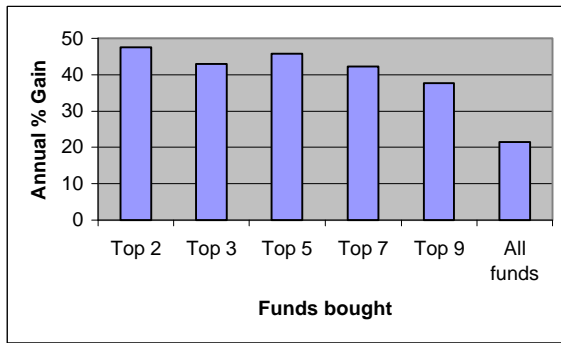


Figure 8 Result of averaging three trees for each fund (20% fuzzy)

## Conclusion

This paper has shown that fuzzy regression trees can be used to predict the rankings of Fidelity select funds in the next 13 weeks. Fuzzy regression trees learn a type of expert system rule by using historical training data.

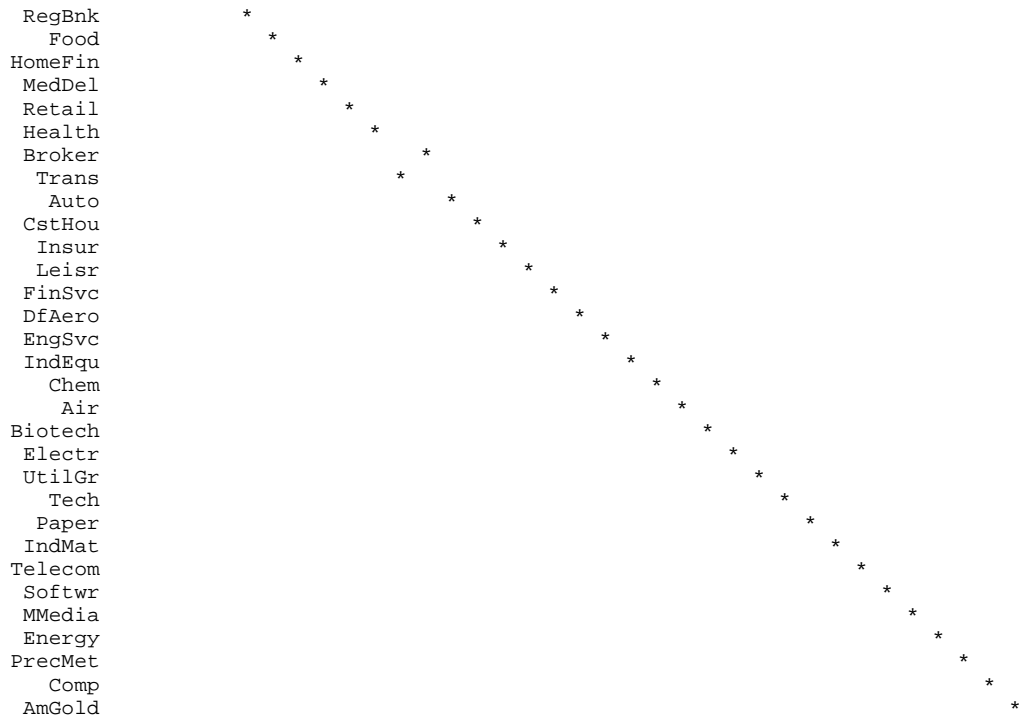
The performance of the regression trees is generally improved by making each node of the tree fuzzy. The performance can also be improved by using multiple trees that have been trained using different features, samples, or number of classes.

The regression trees are also a direct method of fuzzy inference and defuzzification providing an alternative to both neural networks and conventional fuzzy inference systems. They also provide automatic feature selection in the sense that unimportant features simply do not show up in the trees.

## References

1. G. R. Dattatreya and L. N. Kanal, "Decision Trees in Pattern Recognition," Progress in Pattern Recognition 2, L. N. Kanal and A. Rosenfeld (Editors), Elsevier Science Publishers B. V. (North-Holland), 1985.
2. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks/Cole, Monterey, CA, 1984.
3. R. E. Haskell, "Regression Tree Fuzzy Systems," Proc. ICSC Symposium on Soft Computing, Fuzzy Logic, Artificial Neural Networks and Genetic Algorithms, University of Reading, Whiteknights, Reading, England, pp. B.1-B.6, March 26 - 28, 1996.
4. R. E. Haskell, "Neuro-Fuzzy Classification and Regression Trees," Proc. Third International Conference on Applications of Fuzzy Systems and Soft Computing, Wiesbaden, Germany, October 5-7, 1998.
5. R. E. Haskell, G. Castelino and B. Mirshab, "Computer Learning Using Binary Tree Classifiers," Proc. 1988 Rochester Forth Conference, Programming Environments, Rochester, NY, pp. 77-78, June 14-18, 1988.
6. R. E. Haskell and A. Noui-Mehidi, "Design of Hierarchical Classifiers." In N. A. Sherwani, E. de Donker, and J. A. Kapenga, editors, *Computing in the 90's: The First Great Lakes Computer Science Conference Proceedings*, pp. 118-124, Berlin, 1991. Springer-Verlag. Conference held at Western Michigan Univ., Kalamazoo, MI, Oct. 18-20, 1989.

Current date: 122796 0 percent fuzzy



	Pred.	% Gain	Actual
RegBnk	1.0	5.8	0
Food	2.0	5.6	1
HomeFin	3.2	5.4	2
MedDel	4.0	5.4	3
Retail	5.0	4.8	4
Health	5.5	4.3	5
Broker	8.5	3.8	7
Trans	8.5	3.9	6
Auto	9.5	3.4	8
CstHou	10.0	3.2	9
Insur	11.0	2.5	10
Leisr	13.0	2.3	11
FinSvc	13.0	1.6	12
DfAero	14.0	1.5	13
EngSvc	15.0	0.9	14
IndEqu	15.2	0.5	15
Chem	17.0	0.5	16
Air	18.0	-0.5	17
Biotech	19.0	-1.2	18
Electr	20.0	-1.3	19
UtilGr	20.5	-1.5	20
Tech	22.0	-1.6	21
Paper	23.0	-2.3	22
IndMat	24.0	-2.6	23
Telecom	24.5	-2.8	24
Softwr	26.0	-4.8	25
MMedia	27.9	-5.2	26
Energy	28.3	-5.4	27
PrecMet	29.0	-5.8	28
Comp	30.0	-7.5	29
AmGold	31.0	-7.9	30

Averages:  
 Top 2: 5.69815  
 Top 3: 5.60783  
 Top 5: 5.38704  
 Top 7: 5.00652  
 Top 9: 4.69893

Average of all stocks: 0.157712

Figure 4 Results of testing 12/27/96 using 0% fuzzy

Broker  
 Softwr  
 MedDel  
 Comp  
 FinSvc  
 Telecom  
 EngSvc  
 Electr  
 Tech  
 Health  
 Leisr  
 Food  
 HomeFin  
 PrecMet  
 Air  
 Trans  
 Paper  
 IndMat  
 DfAero  
 CstHou  
 Insur  
 Auto  
 Retail  
 Chem  
 Biotech  
 AmGold  
 Energy  
 UtilGr  
 IndEqu  
 MMedia  
 RegBnk



Figure 5 Testing date 7/4/97 with 0% fuzzy

Broker  
 Softwr  
 Tech  
 EngSvc  
 FinSvc  
 Leisr  
 Comp  
 Electr  
 HomeFin  
 Food  
 CstHou  
 Air  
 Telecom  
 Trans  
 Paper  
 DfAero  
 IndMat  
 AmGold  
 Chem  
 Retail  
 MMedia  
 Insur  
 Energy  
 UtilGr  
 MedDel  
 IndEqu  
 PrecMet  
 Auto  
 Health  
 Biotech  
 RegBnk



Figure 6 Testing date 7/4/97 with 20% fuzzy



Broker	2.0	21.5	2
Electr	6.9	20.8	4
Comp	7.2	23.4	1
Tech	8.0	16.7	5
Food	9.5	4.5	27
Paper	9.9	10.2	17
FinSvc	11.2	9.0	20
DfAero	12.5	20.8	3
Chem	12.5	6.3	26
EngSvc	13.0	28.1	0
Leisr	13.2	13.1	9
Softwr	13.5	12.3	13
AmGold	13.7	8.6	21
Energy	14.0	13.1	10
HomeFin	15.5	13.2	8
Air	15.5	8.0	23
PrecMet	15.5	2.7	29
CstHou	15.6	12.7	11
Retail	16.5	11.0	15
Trans	16.5	13.7	7
MMedia	16.6	12.1	14
MedDel	18.5	4.3	28
IndEqu	18.5	8.2	22
Telecom	18.6	12.4	12
IndMat	20.5	9.9	18
UtilGr	22.5	7.4	25
Insur	22.6	7.8	24
Auto	24.1	10.6	16
Health	25.5	0.9	30
Biotech	27.0	14.5	6
RegBnk	27.5	9.8	19

Averages:  
 Top 2: 21.1274  
 Top 3: 21.871  
 Top 5: 17.3616  
 Top 7: 15.1391  
 Top 9: 14.7861

Average of all stocks: 11.8503

Figure 7 Averaging two trees (7 classes and 9 classes) with 20% fuzzy