SEAR
Scalable, Efficient, Accurate, Robust kNN-based Regression

Aditya Desai, Himanshu Singh
International Institute of Information Technology, Hyderabad, India

Vikram Pudi
International Institute of Information Technology, Hyderabad, India

Keywords: Regression, Prediction, *k*-nearest neighbours, Generic, Accurate.

Abstract: Regression algorithms are used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships. Statistical approaches although popular, are not generic in that they require the user to make an intelligent guess about the form of the regression equation. In this paper we present a new regression algorithm SEAR – Scalable, Efficient, Accurate *k*NN-based Regression. In addition to this, SEAR is simple and outlier-resilient. These desirable features make SEAR a very attractive alternative to existing approaches. Our experimental study compares SEAR with fourteen other algorithms on five standard real datasets, and shows that SEAR is more accurate than all its competitors.

1 INTRODUCTION

Regression analysis has been studied extensively in statistics (Humberto Barreto; Y. Wang and I. H. Witten, 2002), there have been only a few studies from the data mining perspective. The algorithms studied from a data mining perspective mainly fall under the following broad categories - Decision Trees (L. Breiman, J. Friedman, R. Olshen, and C. Stone, 1984), Support Vector Machines (W. Chu and S.S. Keerthi, 2005), Neural Networks, Nearest Neighbour Algorithms (E. Fix and J. L. H. Jr., 1951)(P. J. Rousseeuw and A. M. Leroy, 1987), Ensemble Algorithms(L. Breiman, 1996)(R.E. Schapire, 1999) among others. It may be noted that most of these studies were originally for classification, but have later been modified for regression (I. H. Witten and E. Frank).

In this paper we present a new regression algorithm SEAR (Scalable, Efficient, Accurate and Robust kNN-based Regression). SEAR is generic, efficient, accurate, robust, parameterless, simple and outlier resilient. The remainder of the paper is organized as follows: In Section 2 we discuss related work and in Section 3 we describe the SEAR algorithm. In Section 4 we we experimentally evaluate our algorithm and show the results. Finally, in Section 5, we conclude our study and identify future work.

2 RELATED WORK

In this section, we describe work related to the new regression algorithm proposed in this paper.

Traditional Statistical Approaches. Most existing approaches (Y. Wang and I. H. Witten, 2002) (Humberto Barreto) follow a "curve fitting" methodology that requires the form of the curve to be specified in advance. This requires that regression problems in each special application domain are separately studied and solved optimally for that domain. Another problem with these approaches is outlier (extreme cases) sensitivity.

Global Fitting Methods. Approaches like Generalized Projection Pursuit regression (Lingjaerde, Ole C. and Liestøl, Knut, 1999) is a new statistical method which constructs a regression surface by estimating the form of the function. Lagrange by Dzeroski et. al. (Dzeroski S., Todorovski L., 1995) generates the best fit equation over the observational data but is computationally expensive due to huge search space. Lagramge (Todorovski L., 1998) is a modification of Lagrange in which grammars of equations are generated from domain knowledge. However, the algorithm requires prior domain knowledge. Support vector machine (W. Chu and S.S. Keerthi, 2005)

Desai A., Singh H. and Pudi V..
SEAR - Scalable, Efficient, Accurate, Robust kNN-based Regression.
DOI: 10.5220/0003068703920395
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 392-395
ISBN: 978-989-8425-28-7

(SVM) is a new data mining paradigm applied for regression. However these techniques involve complex abstract mathematics thus resulting in techniques that are more difficult to implement, maintain, embed and modify as the situation demands. Neural networks are another class of data mining approaches that have been used for regression and dimensionality reduction as in Self Organizing Maps (Haykin, Simon). However, neural networks are complex and hence an in depth analysis of the results obtained is not possible. Ensemble based learning (L. Breiman, 1996) (R.E. Schapire, 1999) is a new approach to regression. A major problem associated with ensemble based learning is to determine the relative importance of each individual learner.

Decision Trees. One of the first data mining approaches to regression were *regression trees* (L. Breiman, J. Friedman, R. Olshen, and C. Stone, 1984), a variation of decision trees where the predicted output values are stored at the leaf node. These nodes are finite and hence the predicted output is limited to a finite set of values which is in contrast with the problem of predicting a continuous variable as required in regression.

***k*-Nearest Neighbour.** Another class of data mining approaches that have been used for regression are *nearest neighbour techniques* (E. Fix and J. L. H. Jr., 1951)(P. J. Rousseeuw and A. M. Leroy, 1987). These methods estimate the response value as a weighted sum of the responses of all neighbours, where the weight is inversely proportional to the distance from the input tuple. These algorithms are known to be simple and reasonably outlier resistant butt they have relatively low accuracy because of the problem of determining the correct number of neighbours and the fact that they assume that all dimensions contribute equally.

From the data mining perspective, we have recently developed a nearest neighbour based algorithm, PAGER (Desai A., Singh H. 2010) which enhances the power of nearest neighbour predictors. In addtition to this PAGER also eliminates the problems associated with nearest neighbour methods like choice of number of neighbours and difference in importance of dimensions. However, PAGER suffers from the problem of high time complexity $O(nlogn)$, bias due to the use of only the two closest neighbours and decrease in performance due to noisy neighbours. In addition PAGER assigns equal weight to all neighbours which is not typical of the true setting as closer neighbours tend to be more important than far away ones. In this paper we use the framework of PAGER and present a new algorithm **SEAR** which not only retains the desirable properties of PAGER, but also eliminates its major drawbacks mentioned above.

3 SEAR ALGORITHM

In this section we present the SEAR (**S**calable, **E**fficient, **A**ccurate and **R**obust) regression algorithm. SEAR is uses the nearest neighbour paradigm which makes it simple and outlier-resilient. In addition to these, SEAR is also scalable and efficient. These desirable features make SEAR a very attractive alternative to existing approaaches.

In the remainder of this section, we use the notation shown in Table 1. First in Section 3.1 we present an algorithm that has a low space and time complexity for computing k nearest neighbours followed by Section 3.2 which removes over-dependence on the two closest neighbours for constructing a line. Finally in Section 3.3, we describe our technique for noisy neighbour elimination.

Table 1: Notation.

k	The number of closest neighbours used for prediction.
D	The Training Data
d	Number of dimensions in D
n	Number of training tuples in D
A_i	Denotes a feature
X	The feature vectors space. $X = (A_1,\ldots,A_d)$.
T	A tuple in X-space. $T = (t_1,\ldots,t_d)$
T_{RID}	Tuple in D with record id RID
$v_{i,RID}$	Value of attribute A_i of T_{RID}
$N_{L,i}$	Value of attribute A_i for the L^{th} closest neighbour of of T
y	The response variable.

3.1 Approximate *k*NN

SEAR rectifies the problem of high-dimensionality and high response time by providing a variation that reduces the space complexity by a factor of d and the time complexity by a factor of n. This is achieved by a heuristic which approximately computes k nearest neighbours in about $log(n)$ time for $k << n$. For this we generate a list L_i for all attributes A_i where each row is a two tuple consisting of tuple id and attribute value. The list is sorted on the basis of attribute values. For any tuple T and corresponding to attribute A_i, we search the list for T_i using binary search (Knuth

D., 1997). Let *ind* be the index of the closest value. We then load tuples at indexes from $(ind-k, ind+k)$. This is done for all attributes and the union of all is maintained in a list. We then compute the nearest k-neighbours from among these points. Upon simple analysis, it is clear from the mentioned algorithm that the space complexity is $O(n+kd)$ and the time complexity is $O(log(n)+(kd)log(kd))$, where n is the number of tuples in D and d is the number of feature variables. This is because the union can contain a maximum of of $2*kd$ points (k-neighbours are found for d dimensions) and as $k \ll n$, kd can be ignored with respect to n. Thus, $O(n+kd) \approx O(n)$. Even though this algorithm does not compute the closest set of neighbors, experimental results show that the algorithm is still good enough to return a relatively good set of closest neighbors.

3.2 Line Construction using k Neighbours

An approach using the first two neighbours for line construction(used in PAGER) suffers from the problem of over−dependence on the closest two neighbors which may result in biased or incorrect results, if one of the two neighbors has noisy dependant variable values. The argument used in PAGER *i.e.* lower the error in prediction of the neighbourhood, the better the line fits the neighbourhood helps us formalize the concept of the best fitting line which is done as follows - *The best fitting line in this plane is the line which minimizes the weighted mean square error in prediction of the neighbors and hence will be a good representative of the neighborhood.* The weights assigned to neighbours are inversely proportional to the distance between the neighbour and T, thus ensuring that there is lower mean error for closer points.
It can be proved that, this line is equivalent to the best fitting line for k weighted points. However, due to lack of space, the proof is not included in this paper.

3.3 Noisy Neighbour Elimination

It can be clearly seen from (Desai A., Singh H. 2010) that the stability of prediction is inversely proportional to the error in prediction in each dimension. Due to this relation, noisy neighbours may have an adverse effect on the relations between feature variables. Hence it becomes absolutely necessary to eliminate these neighbors (noisy). The algorithm to do the same proceeds as follows :

1: $\mu \leftarrow Mean(y1,\ldots,yk)$
2: $std_1 \leftarrow \frac{\sum_1^k |y_i-\mu|}{k}$
3: $med \leftarrow Median(y1,\ldots,yk)$
4: $std_2 \leftarrow \frac{\sum_1^k |y_i-med|}{k}$
5: **if** $std1 < std2$ **then**
6: $dev \leftarrow std1$, $center \leftarrow \mu$
7: **else**
8: $dev \leftarrow std2$, $center \leftarrow med$
9: **end if**
10: **if** $N_y \notin [center-\eta*dev, center+\eta*dev]$ **then**
11: Eliminate N
12: **end if**

The reason for using both mean (μ) and median (med) is that the noisy neighbours may seriously affect the value of the mean and thus may give a large std_1 value which may not be of any use to eliminate the noisy point. It was observed that the algorithm performed significantly well for values of $\eta \sim 3$.

4 EXPERIMENTAL RESULTS

In this section we compare our regression algorithm against fourteen other algorithms on 5 datasets including a mix of synthetic and real life datasets. The algorithms used are available in the Weka toolkit (I. H. Witten and E. Frank). We experimented with different parameter values in these algorithms and selected those values that worked best for most datasets. We used five datasets in our experimental study. All the results and comparison have been done using the *leave one out* comparison technique which is a specific case of n-folds cross validation, where the n is set to the number of instances of the dataset. The metrics used for comparison are Root Mean Square Error (RM) and Absolute Mean Error (AB). Only a maximum of top three performing algorithms are indicated (in boldface) where the ordering is in accordance with the principle of "Pareto Dominance".

5 CONCLUSIONS

In this paper we have presented and evaluated SEAR, an algorithm for regression based on nearest neighbor methods. Evaluation was done against 14 competing algorithms on 5 standard synthetic and real-life datasets. Although simple, it outperformed all competing algorithms on most datasets.

Future work in this field includes application of SPACER to industrial and real world applications like stock market prediction among others. In addition SEAR can be applied on data-streams and can be used for active learning. It is also of great interest to experiment with different curve fitting measures instead of

Table 2: Experimental Results on Multi, Friedman-1, Friedman-2, Slump and Concrete Dataset.

	Multi		Friedman-1		Friedman-2		Slump		Concrete	
Algorithm	**AB**	**RM**	**AB**	**RM**	**AB**	**RM**	**AB**	**RM**	**AB**	**RM**
SEAR	***0.53***	***0.67***	***1.69***	***2.19***	**29.09**	**42.43**	**5.43**	**7.65**	***5.17***	***7.34***
Gaussian	**0.48**	**0.61**	**1.74**	**2.21**	**35.02**	**53.21**	**5.81**	**7.26**	**6.58**	**8.38**
Isotonic	0.90	1.13	3.59	4.32	219.69	285.11	**5.89**	**7.52**	10.84	13.52
LMS	**0.56**	**0.71**	2.28	2.90	114.37	161.02	6.71	10.58	9.28	16.55
LinearReg	**0.56**	**0.71**	2.28	2.88	107.94	144.30	6.55	7.79	8.29	10.46
Pace	**0.56**	**0.71**	2.28	2.89	107.75	144.47	6.73	7.9	8.32	10.52
RBF	1.01	1.26	3.77	4.71	244.93	315.83	6.84	8.56	13.43	16.67
SL	0.90	1.10	3.56	4.27	220.86	284.07	6.53	7.85	11.87	14.52
MLP	0.63	0.77	2.09	2.61	**30.08**	**38.62**	6.03	8.85	6.58	8.61
SMOReg	0.56	0.72	2.27	2.89	107.27	144.31	5.79	7.95	8.23	10.97
SVMReg	0.56	0.72	2.27	2.89	107.41	144.45	5.79	7.95	8.23	10.97
KStar	0.62	0.78	1.96	2.55	56.06	79.14	6.45	9.75	**6.08**	**8.36**
LWL	0.83	1.05	3.47	4.16	168.19	228.41	5.83	7.74	10.55	13.25
Additive	0.68	0.85	2.30	2.83	142.93	173.57	6.05	8.36	6.67	8.45
IBK	0.57	0.71	**1.74**	**2.21**	50.57	68.70	5.83	8.03	**6.10**	**8.55**

the linear curve fitting measure currently used.

REFERENCES

I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2 edition, 2005.

Y. Wang and I. H. Witten, Modeling for optimal probability prediction, 2002.

Humberto Barreto, An Introduction to Least Median of Squares, Chapter contribution to Barreto and Howland, Econometrics via Monte Carlo Simulation

Lingjaerde, Ole C. and Liestøl, Knut, Generalized projection pursuit regression, SIAM Journal on Scientific Computing, 1999.

L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and regression trees, Wadsworth Inc., 1984.

Haykin, Simon, Self-organizing maps, Neural networks - A comprehensive foundation (2nd edition ed.). Prentice-Hall.

W. Chu and S.S. Keerthi, New approaches to support vector ordinal regression, Proc. of International Conference on Machine Learning(ICML-05), pages 142-152, 2005.

L. Breiman, Bagging Predictors, Machine Learning, vol. 24, no. 2, pp. 123-140, 1996.

R.E. Schapire, A Brief Introduction to Boosting, Proc. 16^{th} International Joint Conf. Artificial Intelligence pp. 1401-1406, 1999.

E. Fix and J. L. H. Jr. Discriminatory analysis, nonparameteric discrimination: Consistency properties, Technical Report 21-49-004(4), USAF school of aviation medicine, Randolf field, Texas, 1951.

P. J. Rousseeuw and A. M. Leroy. Robust Regression and Outlier Detection, Wiley, 1987.

Donald Knuth. The Art of Computer Programming, Volume 3: Sorting and Searching, Third Edition. Addison-Wesley, 1997. ISBN 0-201-89685-0. Section 6.2.1: Searching an Ordered Table, pp. 409–426.

A. Asuncion and D. Newman. UCI Machine learning repository, 2007.

Yeh, I-Cheng, Modeling slump flow of concrete using second-order regressions and artificial neural networks, Cement and Concrete Composites, Vol.29, No. 6, 474-480, 2007.

I.-C. Yeh. Modeling of strength of high performance concrete using artificial neural networks, Cement and Concrete Research, 28, No. 12:1797–1808, 1998.

Todorovski, L. (1998) Declarative bias in equation discovery. M.Sc. Thesis. Faculty of Computer and Information Science, Ljubljana, Slovenia.

Dzeroski, S. and Todorovski, L. (1995) Discovering dynamics: from inductive logic programming to machine discovery. Journal of Intelligent Information Systems, 4: 89-108.

Desai A., Singh H., Vikram Pudi, October 2009. Technical Report PAGER: Parameterless, Accurate, Generic, Efficient kNN-based Regression. http://web2py.iiit.ac.in/publications/default/view_publication/ techreport/59