2015-09-30

An Evaluation of Selection Strategies for Active Learning with Regression

Jack O'Neill

Technological University Dublin

Recommended Citation

O'Neill J. (2015) *An Evaluation of Selection Strategies for Active Learning with Regression.* A dissertation submitted in partial fulfilment of the requirements of Technological University Dublin for the degree of M.Sc. in Computing (Data Analytics), September 2015.

An Evaluation of Selection Strategies for Active Learning with Regression

Jack O’Neill

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

September 2015

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Jack O Neill

Jack O'Neill

Date: ***04 September 2015***

ABSTRACT

While active learning for classification problems has received considerable attention in recent years, studies on problems of regression are rare. This paper provides a systematic review of the most commonly used selection strategies for active learning within the context of linear regression. The recently developed Exploration Guided Active Learning (EGAL) algorithm, previously deployed within a classification context, is explored as a selection strategy for regression problems. Active learning is demonstrated to significantly improve the learning rate of linear regression models. Experimental results show that a purely diversity-based approach to active learning outperforms more traditional algorithms such as Query-By-Committee.

Key words: *Active Learning, Regression, EGAL, Hypothesis Testing, Diversity*

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to Dr. Sarah Jane Delany and Dr. Brian Mac Namee for their invaluable guidance and expertise, right from the very beginning of this project. This thesis would not have been possible without their insights, advice and encouragement.

I would like to thank all of the staff at DIT, particularly the school of computing, whose dedication and passion for the subject have been an ongoing inspiration over the course of my Masters' programme. Also, a big thank you to my classmates who have made this past year unforgettable.

I would like to say a special thank you to my family; in particular my mother, Dervilla, and father, Domhnall, for all the support they have given me, particularly over the last number of months.

Finally, I would like to thank Lily Michailidis, my future family, for her encouragement, throughout the past year, and for all the sacrifices she has made to make this dissertation possible.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1 INTRODUCTION

1.1 Background

Machine learning is a branch of artificial intelligence which aims to allow computers to "optimize a performance criterion using example data or past experience" (Alpaydin, 2014). Within the domain of data analytics, this performance criterion is usually the ability to predict an outcome given a set of input data. Machine learning has been applied to a wide range of problems, such as understanding natural language (Zhu et al., 2008), document classification (Lewis and Gale, 1994), and sentiment analysis (Blitzer et al., 2007), among others. Machine learning algorithms are typically trained using data which has been labelled by an oracle, usually a human expert.

Early approaches to developing and training these algorithms were rooted in the framework of *supervised* learning. Supervised learning is a two-step process, whereby data is first gathered and labelled by an oracle (usually a human expert); after this has been done the labelled data is then used to train a model to make predictions about similar, but as-yet unseen data. Labelled data is not always easy to collect, and there are many cases in which it is "very difficult, time-consuming, or expensive to obtain" (Settles, 2010).

In recent years, there has been more focus on the cost involved in labelling data, (Settles et al., 2008a, Margineantu, 2005), which has coincided with the increasing popularity of active learning. Active Learning rests on the assumption that a machine learning algorithm "can perform better with less training if it is allowed to *choose* the data from which it learns." (Settles, 2012). Active learning helps avoid unnecessary time, money and effort being spent on labelling data which will not improve the resulting model.

Machine learning problems can be categorised as either classification or regression problems, depending on the output type of the data on which they operate. Classification problems assign a class label to input data, separating it into one of a number of distinct groups. Many classification problems are binary, for example "spam" or "not spam" in the case of emails, "customer likely to leave" or "customer not likely to leave" in the domain of churn prediction, etc. Classification problems need not be binary, however, and may have multiple potential labels, or "class values", as with problems of categorizing text documents based on content, or word-sense disambiguation in natural language processing.

Regression problems, on the other hand, have continuous-valued outputs. Unlike classification problems, for which there are a set finite number of answers from which to choose, regression problems, being continuous, have an infinite number of potential answers. Regression problems occur when the output is numeric, as in the case of house prices, for example, or scale-based, as when predicting the level of suspicion of a credit-card transaction.

The output of a classification model is always either right or wrong. If the predicted class value matches the actual class value the prediction is correct, otherwise it is incorrect. Where there are multiple potential class values, all incorrect class values are generally considered to be "equally incorrect". This is not the case with regression problems. The "error" of a regression model is measured as the difference between the predicted class value and the actual class values; thus a prediction which is "close" to the actual value is "more correct" than a prediction which significantly misses the mark. This difference is fundamental for active learning algorithms; its ramifications will be explored more fully in Section 2.4.

Most of the existing research in Active Learning has been on problems of classification, a 2013 study in active learning for regression points out that while active learning has been "extensively studied for classification problems … there is still very limited work on active learning for regression" (Cai et al., 2013). As active learning for regression has yet to gain wider popularity, the choice of suitable active learning algorithms lacks a solid statistical underpinning (see Section 2.5).

1.2 Research Project

While a number of active learning algorithms have been proposed for use in regression problems, and demonstrated to outperform a baseline measurement, this project aims to provide a comparison between the current state-of-the-art approaches; and identify the most effective general-purpose selection strategies.

In addition to comparing the current state-of-the-art, this project investigates the applicability of the newly proposed Exploration Guided Active Learning (EGAL) selection strategy – previously studied in the context of classification – to active learning with linear regression.

The research question this project aims to answer is

“Can active learning selection strategies based on integral dataset properties combined with an analysis of prediction model output be used successfully for linear regression models?”

1.3 Research Objectives

The objective of this project is to establish the effectiveness of the current state-of-the-art active learning algorithms in the context of regression problems. In order to achieve this objective, the following goals have been defined

1. To explore the current state-of-the-art in active learning for regression
2. To establish optimal parameters for applying EGAL to regression problems
3. To identify suitable performance measures for evaluating active learning algorithms for regression.
4. To compare all algorithms on multiple datasets, and establish the statistical significance of differences in performance

1.4 Research Methodologies

In order to accomplish the research goals defined in Section 1.3, this project uses both secondary research, in the form of a literature review, and empirical research, consisting of the implementation and evaluation of multiple active learning selection strategies across a range of real-world datasets. The approach this dissertation takes can be broken down into the following intermediary goals.

1 A literature review is conducted to explore the current state-of-the-art in active learning for regression (objective 1)

2. A review of the most commonly used statistical tests and performance measures for verifying algorithm effectiveness is carried out identifying those most appropriate for the problem at hand (objective 3)
3. The EGAL selection strategy is implemented in Java and tested on an artificial dataset to establish the optimal parameters for use in regression problems (objective 2)
4. The selection strategies identified in objective 1 and 2 are implemented in Java, and evaluated across ten real-world datasets; these evaluations are statistically validated using the findings from objective 3 (objective 4)

1.5 Scope and Limitations

- This project discusses the application of active learning selection strategies to linear regression models. While a number of alternatives to linear regression have been developed, a full treatment of these approaches, though merited, is beyond the scope of the current work.
- Due to the computational intensity of training a linear regression on extremely large datasets, data reduction was performed on a number of datasets in the form of feature selection and observation filtering.
- Although a potential improvement to the EGAL algorithm when used in the context of regression is suggested; the time required to fully develop and test a new derived algorithm makes a concrete implementation infeasible under the scope of the current project.

1.6 Document Outline

This dissertation is organized into the following sections

- Chapter 2 provides an overview and background of the current state-of-the-art approaches to active learning; both in the context of classification and regression. The most prominent selection strategies applicable to continuous-valued outputs are discussed; and EGAL, a recent innovation, is introduced. This chapter also examines the recommended approaches to hypothesis verification, focussing particularly on the work of Demšar (2006).

- Chapter 3 discusses the experiment design and research methods employed, outlining the datasets used, the steps taken in preparing the data and the choice of global parameters used in the experiment. This chapter also gives an overview of the selection strategies used, and the perfomance measures chosen on which the algorithms are evaluated.
- Chapter 4 reports the findings of the experiments conducted, both in determining optimal parameters for the EGAL selection strategy and in comparing the active learning selection strategies under review.
- Chapter 5 builds on the preceeding chapter, discussing the findings in detail and suggesting an optimal algorithm for active learning for regression.
- Chapter 6 provides a summary of the dissertation and outlining its contributions to the current body of knowledge. Future research is recommended.

2 STATE-OF-THE-ART

2.1 Introduction

A major application of machine learning is the development of algorithms which use known examples to make predictions about previously unseen data. Very often, these algorithms are interested in a single output, or response, variable; also known as a label. Machine learning models are usually trained on samples of previously labelled data which allow it to make inferences about the population as a whole. These inferences can then be used to deduce likely labels for unseen data.

The quality and quantity of the data used to train a model has a significant impact on the resulting algorithm's accuracy in labelling unseen data. Provided samples are drawn from the population without bias, increasing the number of samples, will increase the approximation to the overall population. Figure 1 below demonstrates this by plotting x against sin(x). The sine wave pattern becomes more apparent as the number of observations plotted increases.

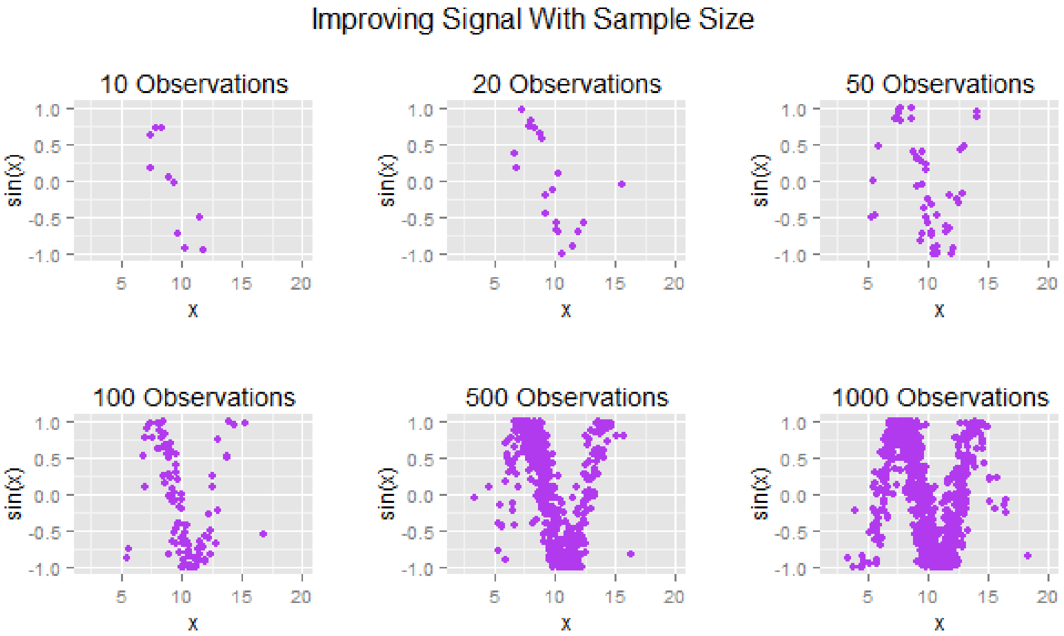

Figure 1 In the example above, values of x were drawn i.i.d from N(10,2) and corrupted using noise levels of N(0, 0.5). The sine wave becomes more apparent as the number of observations increases

There are many situations in which labelling data for use in training a model is expensive; for example, recognition of parts of speech, or extracting entity-related features from documents (Settles, 2010). In these cases; it is not always possible to simply increase the number of labelled examples to improve an algorithm's efficiency. However, judicious selection of which observations to label and add to the training set can vastly improve the predictive power of even a small number of observations; as shown in Figure 2.

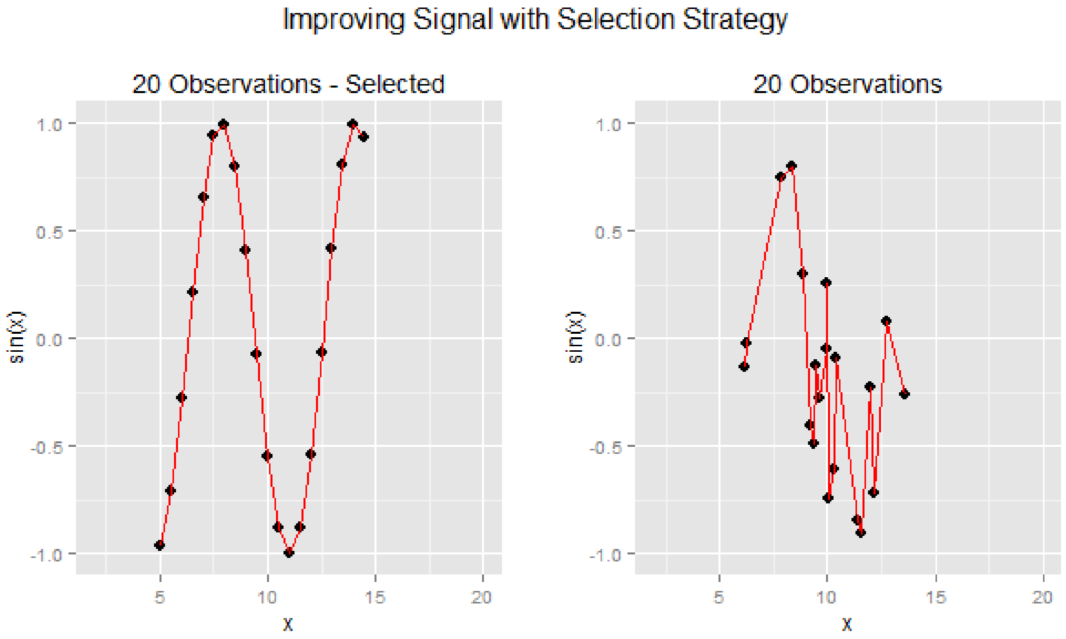

Figure 2 The sine wave function is apparent even with a small number of observations when these observations are carefully chosen. The figure on the left selects only observations with no noise, evenly distributed across the x axis. The figure on the right relies on random sampling from a normal distribution,

Exploring approaches to determining which observations will be of maximum utility in training a model, otherwise known as active learning, constitutes an entire sub-field of machine learning. This chapter will examine the state-of-the-art practices in active learning research, particularly as it applies to problems of regression. Section 2.2 gives an overview of the main subdivisions within the field of active learning. Section 2.3 reviews the most commonly employed selection strategies in active learning problems. Section 2.4 discusses the challenges in developing selection strategies particular to problems of regression. Finally, Section 2.5 explores statistical methods for comparing the performance of Active Learning algorithms, both in problems of classification and regression.

2.2 Active Learning

Active learning is a subfield of *semi-supervised* learning. Unlike *supervised* learning, which requires data to be labelled (or annotated) with its outcome, or target value; semi-supervised learning aims to harness the power of both labelled and unlabelled data in generating models. Although not as prominent as *supervised* learning, interest in semi-supervised learning has grown in recent years, stemming from the growing recognition of data labelling as an "additional, error-prone preparation process" (Schwenker and Trentin, 2014). Active learning aims to minimize labelling costs by ensuring that only the most useful observations are labelled. There are three major approaches to active learning, defined by the manner in which the algorithm gains access to the data to be labelled; *membership query synthesis, pool-based active learning, and stream-based active learning*.

2.2.1 Membership Query Synthesis

One of the earliest approaches to active learning, membership query synthesis, was introduced by Angluin (1988). Within a framework of Membership Query Synthesis, the model is allowed to 'invent' the data, often referred to as generating a query *de novo*. While this method allows the model to learn intelligently by optimizing the input data for rapid improvement; studies have shown (Baum and Lang, 1992) that humans tend to have difficulty accurately labelling such artificial data. Recent research, however, by King *et al.* (2009) and King *et al.* (2004) has demonstrated that this approach may be effective when in the context of automated scientific experiments, where the label does not depend on human interpretation.

2.2.2 Stream-based Active Learning

Stream-based active learning is employed when models select existing queries for labelling, rather than generating them *de novo*. In a stream-based active learning context, the algorithm must consider each query in isolation from all others. In some cases (Loy et al., 2012), the incoming query is checked against the existing classifier to measure uncertainty; those about which a classifier is uncertain are more likely to be sampled, as they are expected to add to the predictive power – or confidence – of the current classifier. In other cases, the incoming queries are labelled if they appear to shed light on underlying unobservable patterns, as in the Hidden Markov Model (Anderson and Moore, 2005). The main limitation of stream-based learning on the selection strategy is that the pool of unlabelled data is not available to the model when the next suitable query is being selected, meaning the usefulness of each observation must be considered in isolation. In situations where this is not the case, a pool-based active learning approach is popular.

2.2.3 Pool-based Active Learning

Pool-based active learning is perhaps the most prominent subfield of active learning, and has received increased attention in recent years as unlabelled data has become easier to collect (Settles, 2010). Pool-based active learning assumes that the model has access to the entire set of unlabelled data at selection time. As with transductive learning (Vapnik, 2013), a related branch of semi-supervised learning, pool-based active learning leverages information gleaned from examining the distribution and features of the as-yet unlabelled data to rank all instances in the unlabelled pool (or some subset thereof), according to some chosen informativeness measure.

Some of the most common informativeness measures for pool-based active learning are uncertainty sampling (Lewis and Gale, 1994), query-by-committee (Seung et al., 1992) and expected model change (Settles and Craven, 2008). These strategies, along with some less popular but equally important alternatives will be discussed in greater detail in Section 2.3.

2.3 Active Learning Selection Strategies

This section describes the most prominent selection strategies used in active learning. Section 2.3.4, however, describes EGAL, a recently introduced approach (Hu, 2011), which has yet to gain widespread adoption.

2.3.1 Uncertainty Sampling

Uncertainty sampling seeks to label the observations for which the current classifier is least confident. This approach is visualized in Figure 3. The reasoning behind this is that labelling observations close to the decision boundary will enable the model to fine tune its knowledge of the spatial limits of each group. Labelling a query which is not close to the decision boundary will add little extra knowledge to the current model.

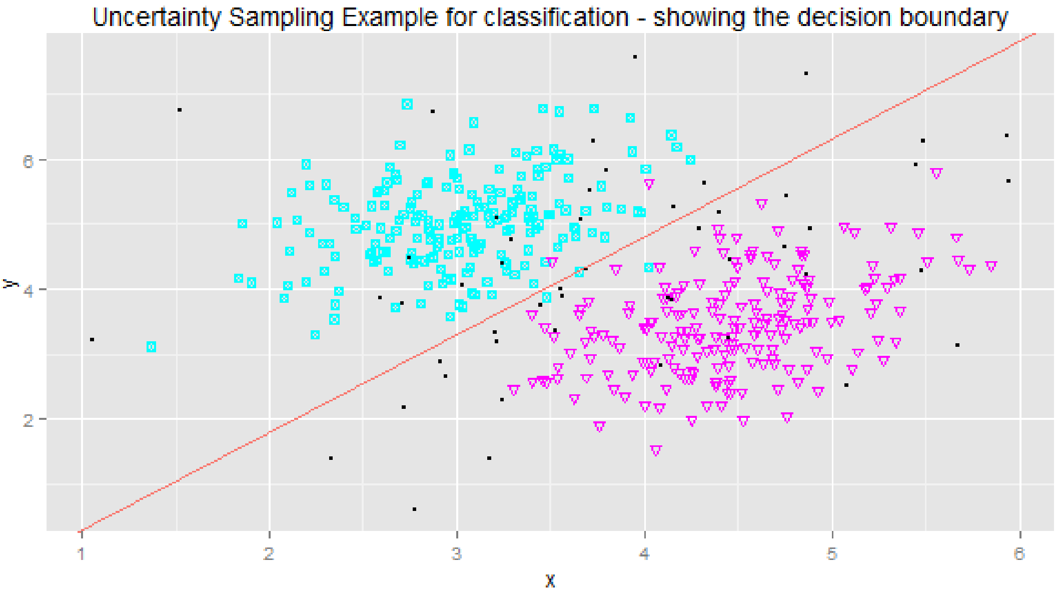

Figure 3 Uncertainty Sampling in a clustering problem. The model is more likely to select unlabelled queries (shown in black) which are close to the decision boundary, depicted as a red line

In the context of regression problems, uncertainty sampling is a common approach for neural networks (MacKay, 1992) or support vector machines (Tong and Koller, 2002). Uncertainty sampling has been criticised on the basis that it sometimes queries outliers which add little or no value to the model (Roy and McCallum, 2001). Its utility when used in conjunction with linear regressions is limited; as these models do not readily offer localized measures of confidence; and any changes to the model's parameters are global. This means that the accuracy of all observations is affected; as opposed to classification problems, where a small change in the decision boundary will affect only observations in the immediate proximity.

These problems can be overcome by using regression models with localized variances, such as kernel ridge regression, which through the use of localized weights, can adapt the predictive function depending on the 'position' of the observation in the general feature space. Recent work (Douak et al., 2013) has begun to explore applications of uncertainty sampling to kernel ridge regression. Douak's study, predicting wind speeds in Algeria showed that uncertainty sampling consistently outperforms a random baseline. This direction of research could prove to be an exciting and innovative field in the coming years; and increased attention appears to be merited.

2.3.2 Query-by-committee

Query-by-committee (QBC) is an ensemble-based selection strategy. The use of ensembles – or combining predictions from multiple sources – has a long history, predating the field of machine learning. Clemen, in his review of the practice of combining forecasts, quotes Laplace, a 19th century mathematician as observing that "In combining the results of these two methods, one can obtain a result whose probability law of error will be more rapidly decreasing" (Clemen, 1989). In the field of machine learning, ensemble models combine the output of various individual classifiers, which, ideally, are all accurate and tend to "make their errors on different parts of the input space" (Opitz and Maclin, 1999).

In the context of active learning, query by committee selection strategies train *K* different classifiers in such a way that each classifier has a slightly different *view* of the data. Technically speaking, each classifier represents a competing hypothesis consistent with labelled dataset (Settles and Craven, 2008). By definition, the most informative instance in the case of QBC, is the one on which the individual members of the committee disagree most.

One of the most common methods of generating a committee of classifiers, and that used by Burbidge *et al.* in their study of active learning for regression (Burbidge et al., 2007) is to train multiple classifiers (or regressors) on different subsets of the labelled data. This can be done either through bootstrap aggregating (bagging) the labelled dataset; i.e. generating subsamples of the data by sampling uniformly and with replacement, as described by Breiman (1996), or using a leave-one-out method, where the data is divided into k equal subsamples, where k is the number of classifiers in the committee, and each classifier is trained on (k – 1) subsamples of the data. The latter is the approach used by Burbidge.

QBC rests on a number of assumptions which are unlikely to be met in many real-world scenarios. These assumptions include that the data is noise-free, a perfect deterministic classifier exists, and that it is possible to draw classifiers randomly from the version space (Lewis and Gale, 1994). Although from this perspective, QBC is usually employed for 'unsuitable' problems, empirical findings from Burbidge *et al.* (2007) employing QBC for linear regression models, on both artificial and real-world data, numerous studies by Cohn *et al.* (1994), and Cohn (1996) using QBC with neural networks, and research by McCallum and Nigam (1998) in text classification using Bayesian classifiers have all shown that QBC is a powerful Active Learning selection strategy; which, for the purposes of this paper is considered the established state-of-the-art.

2.3.3 Expected Model Change

The Expected Model Change algorithm is derived from the earlier Expected Gradient Length, (EGL) introduced by Settles *et al.* (2008b). The idea behind Expected Gradient Length is to favour instances which "would impart the greatest change to the current model *if we knew its label*". (Settles, 2010). The intuition behind this approach is that "it prefers instances that are likely to influence the model" (Settles, 2010). It does not explicitly seek to guarantee that this influence results in increased accuracy, instead relying on the fact that after repeated applications of the process, the maximal model change, and hence accuracy over a given training set will quickly converge as more labels are added. Cai *et al.* summarize this, saying that "if the model is changed due to an outlier, this sampling strategy will certainly choose a good example that can maximize change again in the next data round, so that the negative effect of the outlier will be relieved" (Cai et al., 2013).

The Expected Gradient Length algorithm introduced in *Multiple Instance Active Learning* (Settles et al., 2008b) uses the learner's "current belief" to approximate an instance's actual class label. In the context of a classification problem, this current belief can be inferred from the posterior probabilities for each potential class label. The algorithm can be generalized however, to adopt any feasible approach to determining a probable label for a given instance. Cai *et al.* (2013), in their implementation of EGL for regression, referred to in their paper as Expected Model Change (EMC), use an approach similar to QBC described in 2.3.2. A committee of regression learners is generated using bootstrap sampling (Efron, 1979); *i.e.* repeatedly sampling from the training set with replacement, to generate a number of datasets which, while representative of, differ slightly from the initial training data. The "current belief" of the current learner is approximated as the average value obtained from the committee of learners. Whereas QBC scores an instance based on the level of disagreement within the committee, EMC scores an instance based on the level of disagreement between the committee and the actual learner.

2.3.4 Exploration Guided Active Learning

Exploration Guided Active Learning (EGAL) was introduced by Hu, (2011) who explored its application to problems of text classification. Whereas QBC and EMC approaches to active learning are diversity based strategies, which aim to quickly 'explore' the data; basing their prediction of the usefulness of an unlabelled instance on the perceived difference between that instance and the labelled dataset; EGAL is an example of a density-weighted approach, which also takes into account the representativeness of each instance of the dataset as a whole.

Zhu *et al.* (2008) have demonstrated that combining density-related information can help learners avoid querying outliers, which could otherwise end up reducing the accuracy of the model. By calculating the similarity between an instance and its K-nearest neighbours; outliers *i.e.* those instances which have very little similarity to their nearest neighbours, can be avoided as being unrepresentative of the data. The advantage of this approach is illustrated in Figure 4.

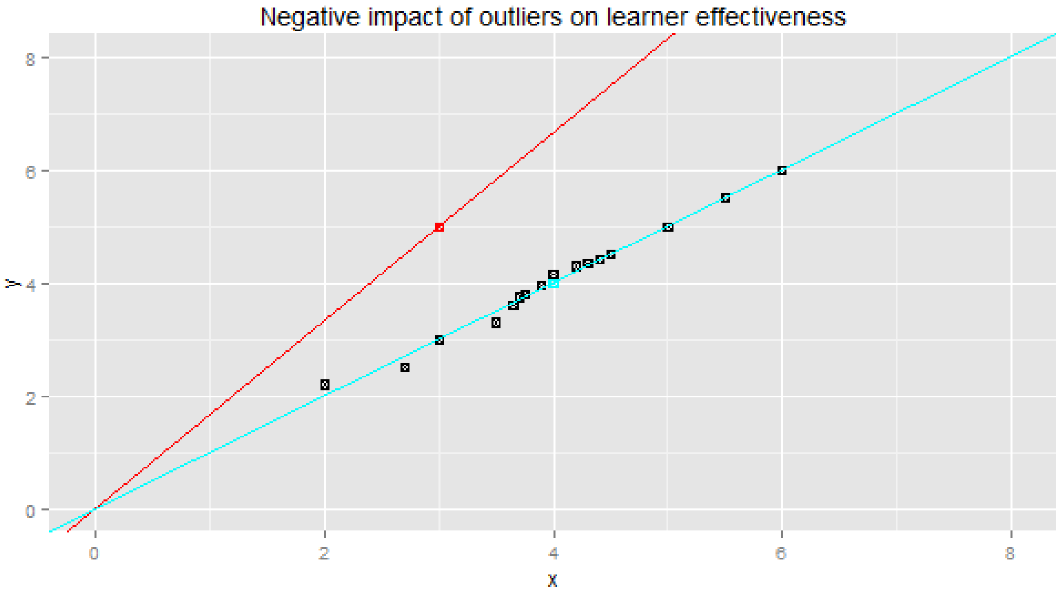

Figure 4 Negative Impact of outliers on learner effectiveness - Querying an outlier (red) can sometimes reduce the accuracy of the overall model. In this case, the most representative example, (cyan) provides a greater overall accuracy to the model

It is important to note that diversity of some form is still important in density-weighted approaches. Zhu *et al.* (2008), combine density with an uncertainty sampling metric to select the most uncertain samples while disregarding outliers. The discussion of EMC (see Section 2.3.3), demonstrates that some level of diversity is required in order to change the output of the leaner. Density weighted approaches seek to balance this requirement with a level of robustness against unhelpful, or even harmful outliers, as demonstrated above.

2.3.4.1 Density in EGAL

The EGAL approach to active learning calculates the density of each observation within the entire dataset, both labelled and unlabelled using any suitable similarity measure. The density of a given example, x is calculated as the sum of similarities between that example and all examples falling within a certain predefined neighbourhood. Formally speaking, density is defined by the equation

$$\sum_{x_r \in N_i} sim(x_i, x_r)$$

where

$$N_i = (x_r \,\epsilon\, \mathcal{D} | sim(x_i, x_r) \geq \alpha)$$

Equation 1 Density function in EGAL

In the equations above, α is a parameter which controls the minimum similarity required between two examples for them to be considered neighbours. Thus, the similarity between examples which are less similar than the α parameter are ignored when calculating density.

2.3.4.2 Diversity in EGAL

While density is calculated between all examples, regardless of whether they are labelled or unlabelled, the diversity measure used in EGAL is the inverse similarity between an unlabelled example, x, and its nearest labelled example. Candidates for labelling are drawn from a subset of the unlabelled data, where the diversity measure exceeds a given threshold, β. Once all candidates have been labelled, the β threshold is dynamically decreased, to allow previously excluded examples to be considered.

The β threshold is determined is determined by a parameter ω, which controls the proportion of the unlabelled dataset which are added to the candidate set when β is updated.

2.3.4.3 Density and Diversity Combined

When selecting an example for labelling, the EGAL algorithm first produces a candidate set of examples sufficiently diverse from the currently labelled examples to be considered. These examples are then labelled in order of density, with the densest examples considered to be the most useful, and therefore labelled first.

The choice of ω parameter was shown to play an important role in the effectiveness of the EGAL selection strategy. The ω parameter controls the size of the candidate set, and therefore the level of bias of the EGAL algorithm towards density. An ω parameter of

0.5 would result in 50% of the remaining unlabelled data being added to the candidate set after each update. A parameter value of 1 results in a density-only approach, as all unlabelled examples are added to the candidate set, regardless of their diversity score. A parameter value of 0 results in purely diversity-based sampling, as only the most diverse example is added to the candidate set; and so will be selected regardless of density. When an ω parameter of 0.25 was used, the EGAL algorithm was shown to consistently outperform random-sampling, diversity-only and density-only approaches to active learning for text-classification. The results of Hu's study (Hu, 2011), show that EGAL is a promising approach to active learning; but further research is needed to establish EGAL as an effective general-purpose algorithm, across a broad range of problem contexts.

2.4 Active Learning for Regression

The approaches outlined above have been applied primarily to problems of classification. Classification models map each input, or observation, to a real-valued vector space, χ, and to each of these vector spaces assign a single class label ϒ, drawn from a finite set of potential labels, "representing the ground truth of the classification problem at hand" (Schwenker and Trentin, 2014). Regression models, on the other hand map each input, *x* to a real-valued output. Accuracy cannot be measured in binary terms of correct answers vs. incorrect answers as with classification, leading to differences in the way models are evaluated.

A common evaluation metric for Linear Regression models is the Root Mean Squared Error (RMSE); which approximates the average numeric difference between a model's predictions and actual class values. As error is measured continuously, any change in the model will affect the error of all labelled examples. Uncertainty sampling, discussed in Section 2.3.1 aims to adjust the model's decision boundaries; a small change to the decision boundary between classes can increase the accuracy of the model in the immediate locality, and leave the accuracy of the rest of the input space unchanged. This approach, however, is not possible for regression problems.

Much of the recent literature in Active Learning for regression has applied approaches which have been tried and tested in a classification setting to regression problems.

Burbidge *et al.* (2007) apply QBC to linear regression problems, while Cai *et al. (2013*) carry out a similar experiment using EMC.

2.5 Statistical Tests for Algorithm Performance

The use of statistical tests for comparing machine learning algorithms has increased in recent years, which is attributed by Demšar (2006) both to the maturity of machine learning as an academic discipline, and to the publication of a study on the use of statistical tests for comparing classification algorithms by Dietterich (1998). In recent years, systematic hypothesis testing has come to be seen as not just a desirable, but a necessary step in confirming whether a new proposed method offers a significant improvement over the existing alternatives (Derrac et al., 2011). Growing awareness of the importance of choosing an appropriate statistical test to the problem at hand has led to a number of studies evaluating the strengths and weaknesses of different methods of hypothesis verification. (Derrac et al., 2011, Luengo et al., 2009, Trawiński et al., 2012)

2.5.1 Evaluating Hypothesis Tests

When a machine learning paper introduces a new algorithm or enhancement to an existing algorithm, "an implicit hypothesis is made that such an enhancement yields an improved performance over the existing algorithm(s)" (Demšar, 2006). In order to verify this hypothesis, the common statistical approach of *rejecting the null hypothesis* is taken. The researcher hypothesises that there is no statistically significant improvement offered by the new approach. Statistical tests are then used to assess the likelihood of the null hypothesis being the case. If the probability of the null hypothesis being true is sufficiently small, it can be rejected, implying that the new approach does, in fact, offer a statistically significant improvement over the alternatives. This probability is usually represented in statistical tests as a *p value*, which is the probability of the null hypothesis being true. If a statistical test yields a *p* value lower than a pre-selected threshold – usually 0.05 (5%) and 0.01 (1%) (James et al., 2014) - the researcher can then reject the null hypothesis.

Tests for statistical significance can yield two types of errors. Firstly, the null hypothesis may be rejected in error. This is known as a Type I error, and leads to a false positive *i.e.* an algorithm is found to be significantly better than the alternatives when in fact it is not. Alternatively, a statistical test may fail to reject the null hypothesis when it should, in fact, be rejected. This is known as a Type II error, and leads to false negatives *i.e.* an algorithm is not found to be significantly better than the alternatives when in fact it is. Using hypothesis tests which are not suited to the data can increase the probability of a Type II, or more worryingly, a Type I error, as Dietterich has shown (1998).

2.5.2 Testing Multiple Classifiers over Multiple Datasets

When comparing two classifiers on a single dataset, the McNemar test (Salzberg, 1997), or more recently, T or F tests after cross validation (Alpaydin, 1999, Dietterich, 1998) are commonly used. Demšar (2006), however, has warned that these tests are prone to Type I errors when applied repeatedly to multiple classifiers. An example will help to illustrate why this is the case.

A researcher is comparing 5 algorithms on a single dataset. The accuracy of each classifier is then computed, and the results of each algorithm are compared to all others, and tested for statistical significance. For each test, the null hypothesis is formulated as *there is no significant difference between the performance of Algorithm A and Algorithm B on the given dataset.* The researcher decides to reject the null hypothesis if tests indicate that there is at most a 5% chance of a Type I error. If the null hypothesis is rejected the researcher can then say with 95% confidence that there is a significant difference between the performances of the algorithms under scrutiny. The researcher repeats this test on each pair of algorithms, leading to a total of $\binom{5}{2}$ or 10 tests. The probability of making a Type I error on any single test is 0.05. However, the probability of making a Type I error on at least one test is now $(1 - 0.05)^{10}$ or roughly 60%.

Although pairwise testing when comparing multiple algorithms across many datasets can significantly increase the chance of Type I errors, as shown above, it is still used in the literature. (Cai et al., 2013, Zhou et al., 2002). Tests across multiple domains, such as the Analysis of Variance (ANOVA) test, and the related Friedman test for statistical significance, are potential alternatives to pairwise T-testing. Instead of testing each algorithm separately against all others, the null hypothesis of tests across multiple domains is that all classifiers perform equally well and that any observed difference are due to randomness (Demšar, 2006). If the null hypothesis is rejected, *post-hoc* tests can be used to ascertain which algorithms stand out.

2.5.3 Parametric and Non-Parametric Tests

Parametric statistical tests make assumptions about the underlying population from which the data has been drawn. When these assumptions are met, parametric tests can be more accurate than their assumption-free, non-parametric equivalents. However, Demšar (2006) stresses that using parametric tests inappropriately can lead to "elevated Type I errors". It is therefore important to understand the assumptions underpinning parametric tests before utilising them in research. The ANOVA test is a parametric test which has been recommended for multiple comparisons across datasets (Vázquez et al., 2001), provided its assumptions are met.

The ANOVA test seeks to divide the variance found in the results between "variability between the classifiers, variability between the data sets and the residual (error) variability" (Demšar, 2006). The researcher is generally interested in the variability between the classifiers, as when this is high enough, he or she is then in a position to conclude that there is a significant difference in classifier performance. The ANOVA test assumes independence between performances on each dataset, making it unsuitable where datasets have been resampled. It also assumes that scores across each dataset are normally distributed. Demšar, however, points out, that in practice this is rarely a problem and most statisticians "would not object to using ANOVA unless the distributions were […] clearly bi-modal" (Demšar, 2006). More importantly, ANOVA assumes *sphericity*; *i.e.* that the variances in scores between all groups are equal. This is a particular problem for regression analysis, where root mean squared error (RMSE) is used as a performance measure. The scale of the output variable has a large impact on absolute RMSE values. The RMSE when calculating house prices in dollars will naturally be higher than the RMSE when calculating subjective ratings given on a scale of 1 – 10. Because of this, there is no commensurability between results across different datasets, and the ANOVA test is not usually suitable for determining statistical significance.

The Friedman test (Sheskin, 2003), is a popular alternative to the ANOVA test when the assumptions of ANOVA are not met. The Friedman test replaces absolute score values with ranks. Assuming an experiment is conducted over *K* datasets, a rank of 1 will be assigned to the best performing algorithm on each dataset, 2 to the second-best up to a rank of *K* for the worst-performing algorithm. This approach eliminates the problem of incommensurability of results between datasets; as absolute measures are discarded and only the relative performance of each algorithm to all others is retained.

One potential shortcoming of the Friedman test is that ranks are applied only within datasets, so it does not differentiate between decisive and marginal ‘wins’. Trawiński *et al.* (2012) have demonstrated that the Friedman aligned rank tests is a more powerful test, which assigns ranks across all datasets. In order to compare across datasets, the scores within each dataset are first ‘aligned’; using distance from the average score for each classifier on that dataset. This ensures commensurability of results between datasets; which in turn allows us to assign ranks across, rather than within datasets.

3 DESIGN / RESEARCH METHODS

3.1 Experimental Methods

The following section outlines the datasets used when comparing active learning algorithms, outlining the features of each dataset, the preparatory work carried out on the data prior to conducting the experiment, as well as the global experimental parameters used.

3.1.1 Datasets Used

Each of the selection strategies under consideration were tested on ten separate datasets, taken, mainly from the UCI machine learning repository.

Dataset	# Attributes	Size	Provenance
House Prices	14	506	UCI Machine Learning Repository
Bike Sharing Demand	9	2000	UCI Machine Learning Repository
Scale Dataset (Dennis Schwartz)	500	1027	Association for Computational Learning
Online News Popularity	60	2000	UCI Machine Learning Repository
Auto MPG	8	392	UCI Machine Learning Repository
Concrete	9	1030	UCI Machine Learning Repository
Red Wine	12	1599	UCI Machine Learning Repository
White Wine	12	2000	UCI Machine Learning Repository
Treasury	16	1049	UCI Machine Learning Repository
Yacht	7	309	UCI Machine Learning Repository

Table 1 Overview of datasets used

3.1.2 Data Preparation

Three nominal attributes were removed from the bike sharing demand dataset as these are not useful in a regression model. Furthermore, two derived attributes which could be used to directly calculate the output value (number of registered bike users and number of casual users) were omitted. The number of observations in each dataset was limited to the first 2000 encountered in the original dataset to facilitate computation. The table above outlines the features of the datasets after these treatments. Datasets were chosen to have real-valued, rather than class-valued outputs; to facilitate the use of a linear, rather than logistic, regression model. All attributes were normalized using feature rescaling as shown in Equation 2, for use both in the linear regression model and when computing distances between observations.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Equation 2 Feature Scaling Function

The subjectivity dataset was converted to a unigram bag of words; each word was included as a feature, with the number of occurrences in a particular review as the feature-value. This resulted in an extremely large dataset, presenting onerous computational requirements. In order to facilitate computational performance, and following the approach of Blitzer *et al.* (2007); extensive feature selection was performed, selecting only the most 'informative' features. In the case of the subjectivity dataset, the top 500 features (from a total set of roughly 20,000) were selected, with occurrence count used as a rough heuristic to determining the informativeness of a particular feature. Before selecting the most commonly occurring words, commonly occurring *stop words* were removed from the feature-set. The list of stop words used was a combination of those collated by the Natural Language Toolkit Python project; and a small number of domain-specific stop words (e.g. film) identified manually. While feature selection on this scale runs the risk of generating underspecified models; we are interested only in the relevant performance of multiple regressors on the same dataset, and not their absolute values. No normalization was performed during data preparation, as normalization occurs when generating the linear regression models.

3.1.3 Resampling

When establishing reliable estimates of an algorithm's performance on a given dataset, it is usual to employ cross validation or resampling techniques (Demšar, 2006). Cross validation and resampling make it possible to run multiple tests on a single dataset; the mean square error and variance between tests can then be used to determine significant differences between model performances. However, care must be taken when evaluating the output of experiments conducted using the above techniques, as functions for computing statistical significance make the assumption of independence between tests.

Cross validation aims to approximately simulate multiple independent test sets; and provide a guideline to the expected generalization error of a given model. Dietterich's (1998) improved 5x2 cross validation, and Alpaydin's 5x2 cv F test (Alpaydin, 1999); go some way towards compensating for the inherent dependence and overlap found between test sets generated randomly from a given sample. These approaches are not suited to the current experiment, however, as the datasets are not split into training and test sets as is customary in model evaluation. The motivation behind this decision is discussed in Section 3.2.

Demšar demonstrates that when comparing model performances across multiple domains, the sources of the variance are the *differences in performance over (independent) data sets and not on (usually dependent) samples,* (Demšar, 2006). The datasets used for this experiment have thus been resampled using disjoint sampling into five separate subsamples; each consisting of 20% of the total data. Five new samples of the data are created by combining four of these five subsamples, so that each consists of 80% of the data. It is expected that computing the mean error over each of the five subsets will improve the reliability of model performance. The dependency between the subsamples is not an issue as we are not interested in the variance over samples. This is in line with Demšar's suggestions (2006).

3.1.4 Experimental Framework

The role of an Active Learning selection strategy is to choose observations for labelling from a pool of unlabelled data. As is standard in active learning contexts, selection is performed in batches. A batch size *b* is chosen which determines the number of observations labelled in each iteration of the Active Learning process. During each iteration, the selection strategy is allowed to choose *b* observations from the unlabelled dataset *U* which are labelled and added to the labelled dataset, *L*. The labelled dataset is then used to train a Linear Regression, which is assessed on its ability to accurately predict the outputs of the remaining unlabelled dataset. A batch size of 2% of the total number of observations was chosen for each dataset. This was done to ensure that the selection strategies have an equal number of iterations across all datasets. Where a constant batch size is used across all datasets, datasets with fewer observations will have fewer iterations, and consequently the effect of the selection strategy will be less pronounced. The first batch of labelled data (seed data) was chosen randomly from the unlabelled dataset, and consisted of 2% of the total observations. Each selection strategy was seeded with the same initial data.

3.2 Selection Strategies Used

This section gives a brief overview of the parameters used in implementing the selection strategies under examination.

3.2.1 Random Baseline

10 models are induced using random selection strategies initialized with differing random seeds. The final value is the mean of these results. The graphs in Section 4.3 show error bars of 1 standard deviation above and below the mean.

3.2.2 QBC Implementation

Settles reports that there is “no general agreement in the literature on the appropriate committee size to use”, but that “even small committee sizes (two or three) have been shown to work well in practice” (Settles, 2010). Committees of 2, 3, 4, and 5 linear regression models were considered during preliminary testing, and a committee size of 5 was determined to be optimal. Committee sizes greater than 5 were not considered due to the additional computation involved in maintaining large committees.

3.2.3 Expected Model Change

As with QBC, committee sizes of 2, 3, 4 and 5 were considered for the Expected Model Change algorithm. A committee size of 5 was determined to be optimal, while remaining computationally feasible.

3.2.4 EGAL Implementation

A number of values for the α and ω parameters of the EGAL selection strategy were considered, as outlined in Section 4.2. Parameter values of 0.75 and 0.25, respectively, were determined to be optimal and used in the EGAL implementation on real-world datasets.

3.2.5 Diversity (EGAL)

By creating an EGAL selection strategy with an ω parameter of 0, we ensure that only the most "diverse" observation in the current unlabelled dataset is selected for labelling. This is a purely diversity-based approach; as the algorithm will not have to choose between observations in the candidate set, so the density measure will not be used.

3.2.6 Density (EGAL)

By creating an EGAL selection strategy with an ω parameter of 1, we ensure that no observation is excluded from the candidate set. This eliminates diversity from the selection strategy, as instances will be queried in order of density alone.

3.3 Performance Measures

3.3.1 Raw Model Performance Measures

A performance measure for that iteration will be calculated using the following formula, closely modelled on the standard Root Mean Squared Error formula.

$$Peformance = \sqrt{\frac{1}{(|U| + |L|)} \sum_{i=1}^{|U|} \left(\hat{y}_{U_i} - y_{U_i}\right)^2}$$

Equation 3 Performance measure used to calculate regressor accuracy

Where U is the unlabelled dataset, |U| its cardinality, |L| is the cardinality of the labelled dataset, $\hat{y}_{U_i}$ represents the predicted output value for the ith observation of U, and y_{U_i} represents its actual value. The use of this equation as opposed to the standard root mean square error formula effectively assigns a score of 0 to all labelled examples; and ensures the performance of each algorithm tends towards 0 as more examples are labelled. The justification for this is that, as with RayChaudhuri and Hamey (1995), we are interested only in labelling the current dataset; rather than training a model robust to generalization, and so would not attempt to predict an output value which has already been supplied to us.

For each algorithm over each subsample of each dataset, the performance measure of each iteration is reduced to a measurement of the area under the curve, where the x axis represents the number of labelled observations, and the y axis our performance measure defined in Equation 3. An algorithm's performance on the dataset is defined as the average performance measure across all subsamples of the data. In this context, a more accurate model will result in a lower performance score.

4 EXPERIMENTAL ANALYSIS

4.1 Introduction

The following chapter consists of two sections. There is as yet no literature on the application of EGAL to regression problems; Section 4.2, therefore, explores configurations of the α and ω parameters for the EGAL selection strategy, using datasets with varying degrees of noise. A combination of a high α, and low ω is determined optimal and used as the parameter for the EGAL selection strategy explored in 4.3. Section 4.3 compares the performance of five active-learning algorithms (EGAL, Query-by-Committee, Expected Model Change, Density and Diversity) across ten real-world datasets; finding that all but density outperform a random baseline. The difference is shown to be statistically significant in the case of the Query-By-Committee, Expected Model Change and Diversity-based approaches. Finally, the diversity-based approach is recommended as the most effective selection strategy for linear regression.

4.2 EGAL Exploration

In order to explore the impact of the α parameter on the learning rate of a linear regression model a simple artificial dataset was used. The dataset consists of two features *x* and *y*. Noise was generated by drawing randomly from a normal distribution N(0, δ). The δ parameter, representing the standard deviation of the noise distribution is adjusted to increase or decrease the level of noise in the data. The *x* variable is drawn from uniform distribution between 0 and 100. Three datasets were generated using δ 2, 10 and 27 for low, moderate and high noise. The resulting distribution of the data is visualized in Figure 5. Nine EGAL selection strategies were created, using pairwise combinations of ω and α parameters of 0.25, 0.5 and 0.75. Five samples of 400 observations each were created for each dataset, and the performance of each algorithm was averaged over each sample. The performance was measured according to the framework laid out in 3.2.

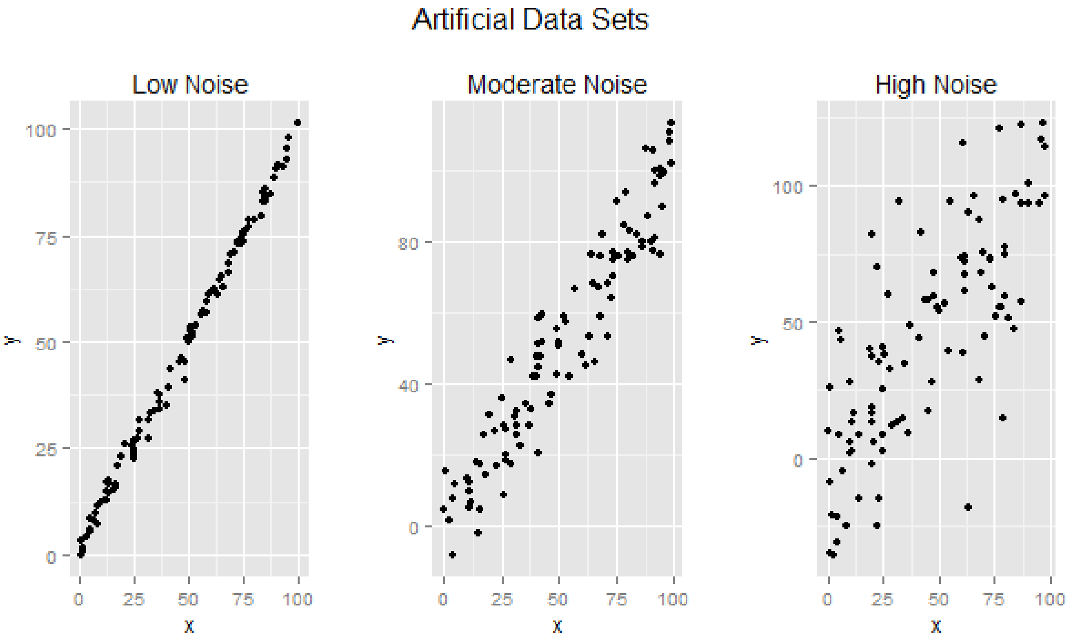

Figure 5 Distribution of x and y in uniform artificial datasets showing 100 observations from each

4.2.1 Impact of the Alpha Parameter

The impact of the α parameter, which controls the size of the neighbourhood for the purposes of measuring density had very little effect on the resulting model performance. This is immediately apparent from the performance graphs in Figure 6. However, for completeness, the results have been tabulated in Table 2.

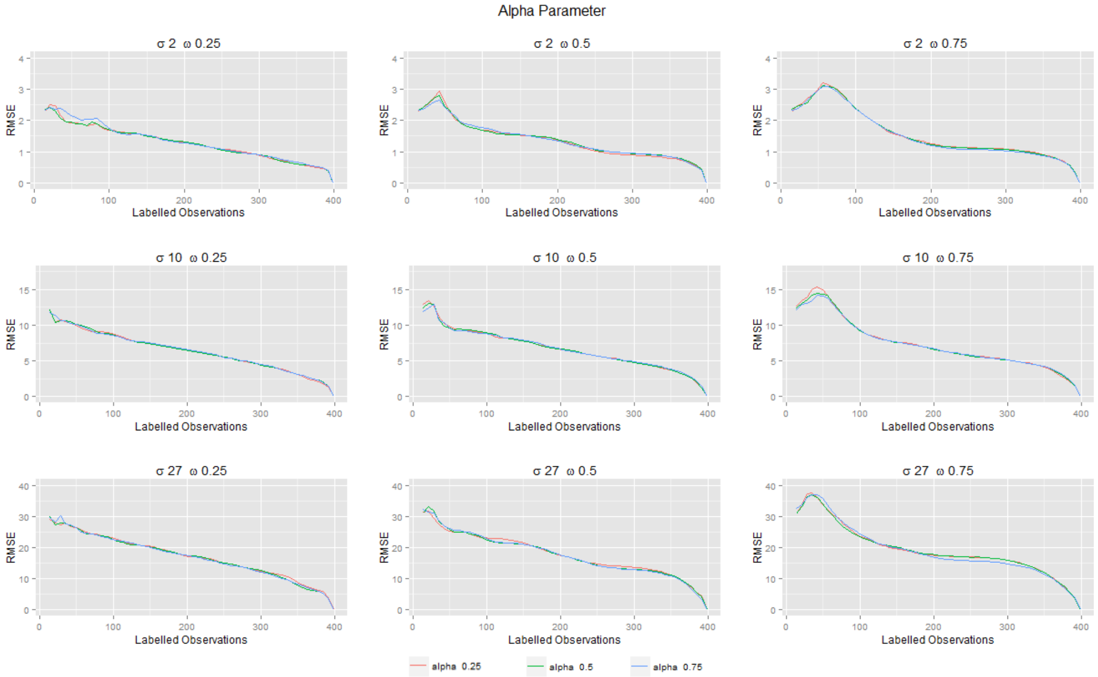

Figure 6 Linear Regression error rates for alpha parameters at varying levels of noise and omega values. This graph is reproduced and enlarged in Appendix A.

Low Noise	*α = 0.25*	*α = 0.5*	*α = 0.75*
ω = 0.25	493.8541	**491.2902**	504.4977
ω = 0.5	**525.412**	531.2243	529.3337
ω = 0.75	603.8843	600.1709	**591.0677**

Moderate Noise	*α = 0.25*	*α = 0.5*	*α = 0.75*
ω = 0.25	**2439.089**	2439.864	2452.423
ω = 0.5	2565.946	**2559.582**	2565.52
ω = 0.75	2827.304	2800.801	**2796.037**

High Noise	*α = 0.25*	*α = 0.5*	*α = 0.75*
ω = 0.25	6600.094	6530.479	**6525.273**
ω = 0.5	6863.346	6804.748	**6786.643**
ω = 0.75	7462.606	7440.471	**7359.806**

Table 2 Comparison of Area under the curve for EGAL selection strategies by Alpha Parameter, the best performing algorithm on each evaluation is highlighted in bold.

While results are mixed across all evaluations, an α parameter of 0.75 appears to outperform consistently on highly noisy data. The Friedman Test for statistical significance fails to reliably establish a difference between the three parameter settings, so no particular parameter value can be claimed to outperform all others. The results, however, tentatively suggest that a value of 0.75 for alpha may be desirable. Section 5 will discuss potential reasons for the apparent lack of significance of the α parameter in this experiment.

4.2.2 Impact of the Omega Parameter

The impact of the ω parameter, which controls the selection strategies' bias towards either density or diversity was shown to have a more pronounced effect on the accuracy of the resulting linear regression model.

ω

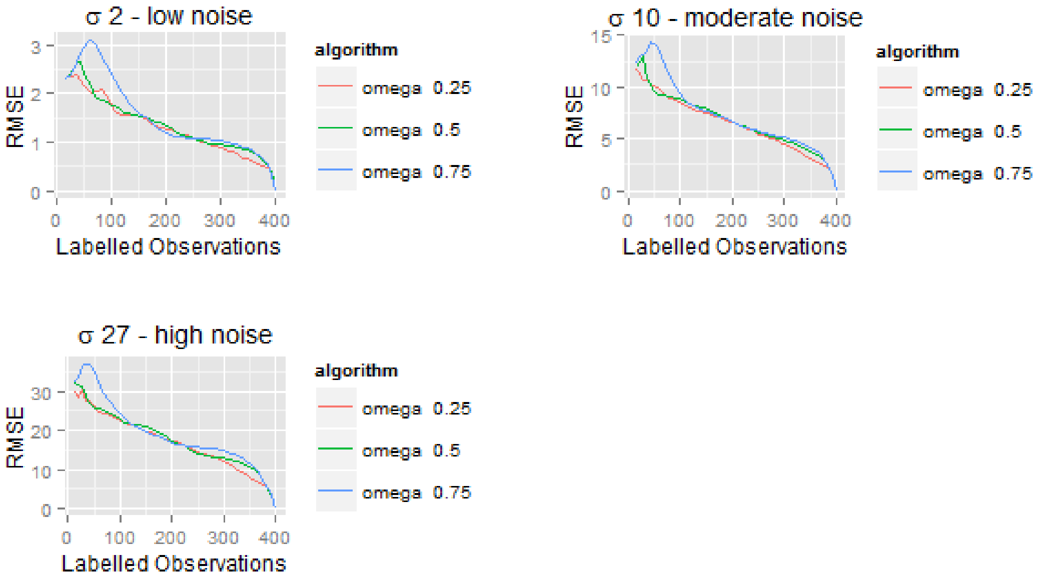

Figure 7 Impact of the omega parameter on model accuracy. This graph is reproduced and enlarged in Appendix A

A higher value for the ω parameter results in a bias towards density. In each of the cases above, an ω value of 0.75 results in an initial spike in root mean square error. This strategy is outperformed by the others across all datasets. The difference between an ω value of 0.5 and a value of 0.25 is less apparent. However, in each of the three datasets, the lower parameter is less prone to spikes in RMSE, and the resulting curve has a smoother gradual descent than the alternatives. The areas under the curve (AUCs) for each of these algorithms are tabulated below, showing that performance tends to increase as the value of ω decreases.

	δ = 2	*δ = 10*	*δ = 27*
ω = 0.25	**504.4977**	**2452.423**	**6525.273**
ω = 0.5	529.3337	2565.52	6786.643
ω = 0.75	591.0677	2796.037	7359.806

Table 3 Comparison of Area under the curve for EGAL selection strategies by omega Parameter, the best performing algorithm on each evaluation is highlighted in bold.

The results above suggest that a lower value for omega is generally desirable. While there are insufficient independent datasets for a reliable test for statistical significance, the above results suggest that a low value of omega should be used regardless of the expected noise of the dataset.

4.2.3 Suggested Default Parameters

The primary aim of the current study is to evaluate the applicability of a number of active learning algorithms proven in the context of classification problems to active learning for regression. Statistically verifying the optimal parameters for EGAL is an area deserving of further study, but outside the scope of this work. The findings above suggest that the impact of the α parameter may not be statistically detectable, however, there is possibly an inverse linear relationship between the ω parameter and model performance. The results suggest that a low ω and high α value are desirable parameters regardless of the expected noise in the dataset under consideration. On the strength of these findings, the α and ω parameters used in the following section were fixed at 0.75 and 0.25 respectively.

4.3 Comparison of Active Learning Selection Strategies for Linear Regression

Having selected the optimal parameters for EGAL, each of the selection strategies outlined in Section 3.2 were evaluated on ten real-world datasets, described in Section 3.1.1. The parameters for the experimental framework follow those in Section 3.1.4. This section presents the empirical findings across all datasets. Section 4.4 reports on the tests for statistical significance carried out on the results.

Most of the datasets yielded relatively accurate models using the random baseline. However, the impact of good selection strategies was apparent. Figure 8 shows the results of testing on the treasury and ratings dataset. As indicated in 3.2.1, the random baseline shown is the mean over ten runs, with error bars showing one standard deviation above and below the mean. The RMSE of effective selection strategies on these datasets drops sharply as the most informative observations are labelled early on,

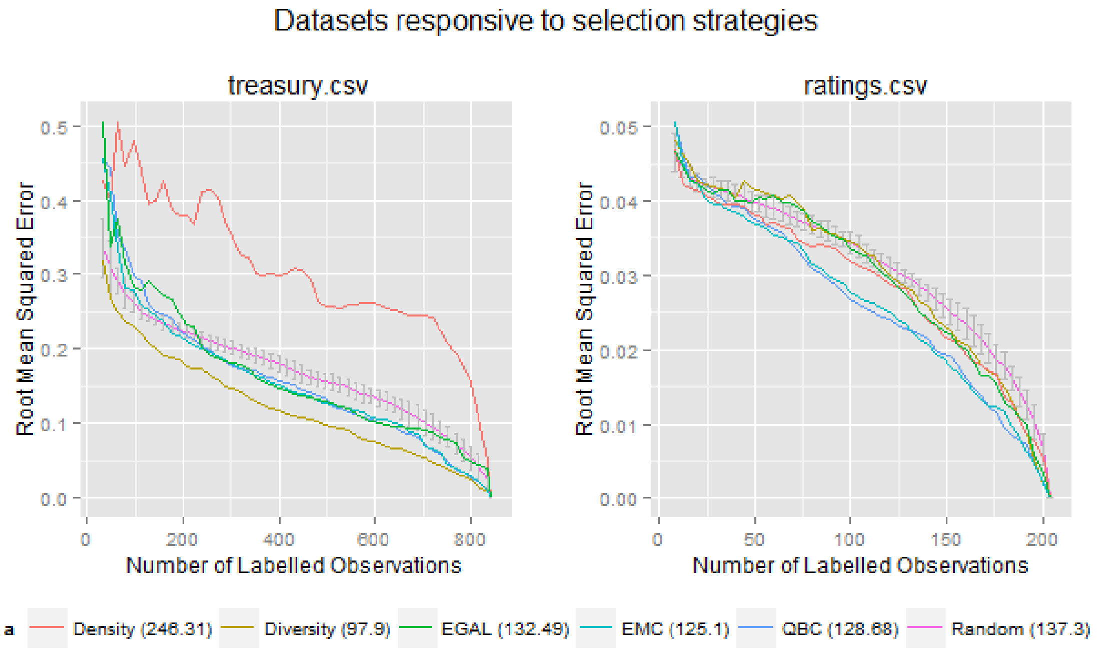

Figure 8 Algorithm performances on the treasury and ratings datasets. The AUC of each algorithm is shown between brackets in the legend.

The impact of using a selection strategy varied across datasets. For example, both of the wine quality datasets showed little reduction in RMSE over iterations. The error curves for these datasets decrease steadily and gradually, reflecting the fact that as each observation is labelled it reduces the overall error, but adds little predictive power to the regression models. While the active learning strategies outperformed the baseline in these datasets, there is less improvement than in the datasets above.

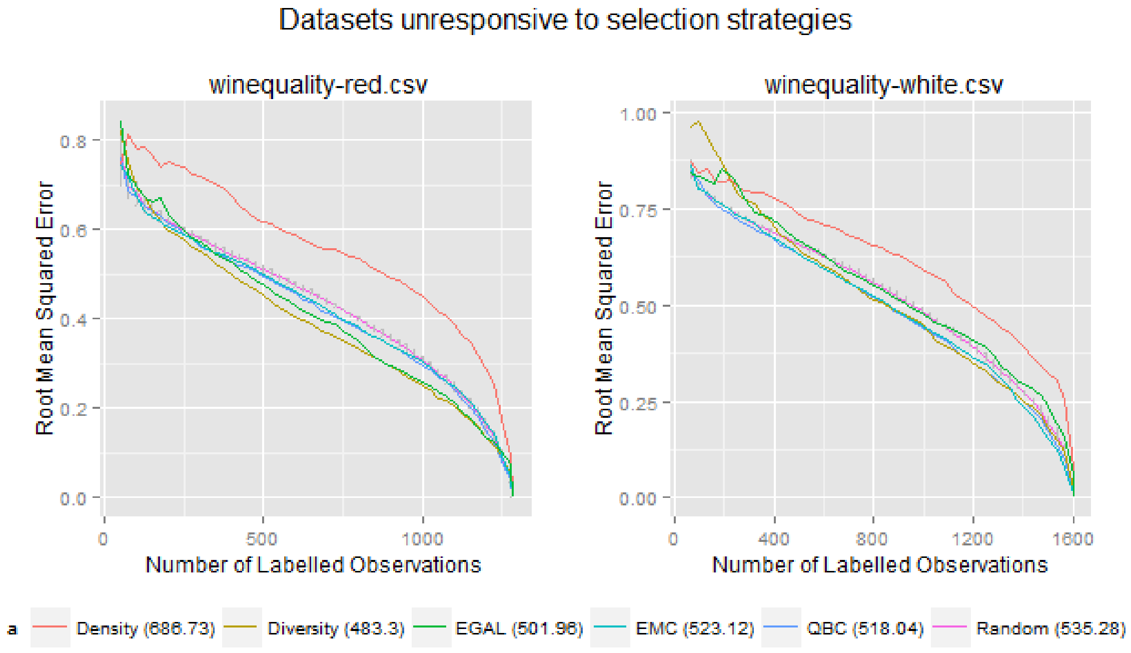

Figure 9 Datasets showing little response to active learning selection strategies. The AUC of each algorithm is shown between brackets in the legend.

The online news dataset showed a pronounced improvement with effective selection strategies. QBC, Expected Model Change and Diversity had a significantly reduced error compared to Density and Random. The EGAL selection strategy performed poorly on this dataset.

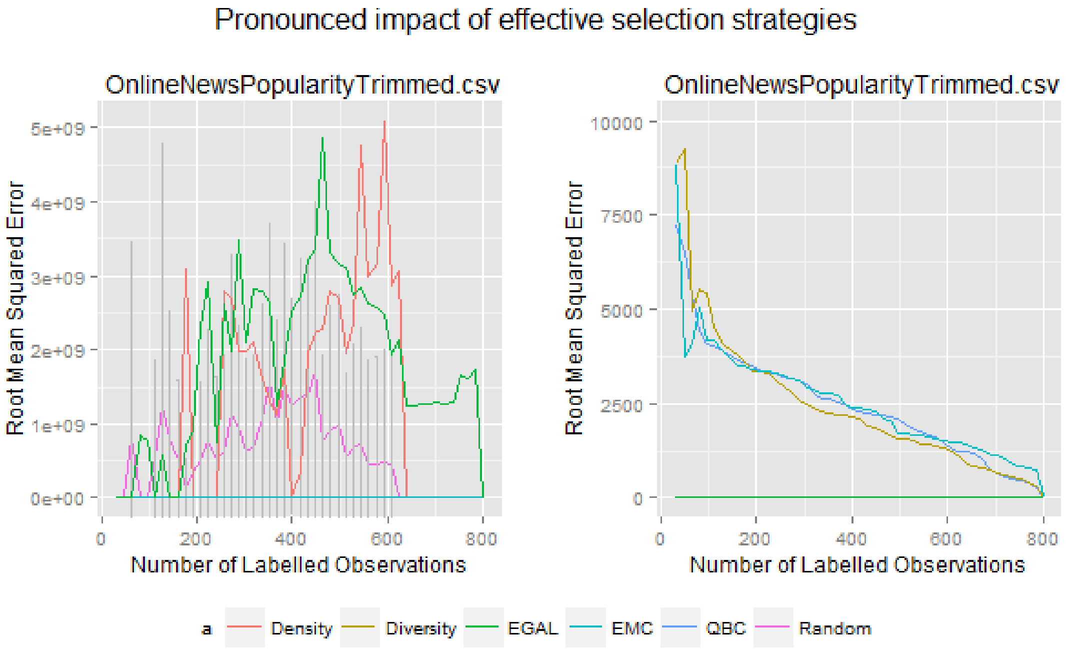

Figure 10 Selection strategy performance on the Online News dataset. The graph on the right is rescaled to show the relative performance of the effective selection strategies

There appears to be a strong correlation between the number of outliers in the dataset and the effectiveness of active learning selection strategies. The class value output distributions for four of the datasets used are depicted in Figure 11. This indicates that the datasets on which the active learning selection strategies had the greatest impact were those in which the highest concentration of outliers were found.

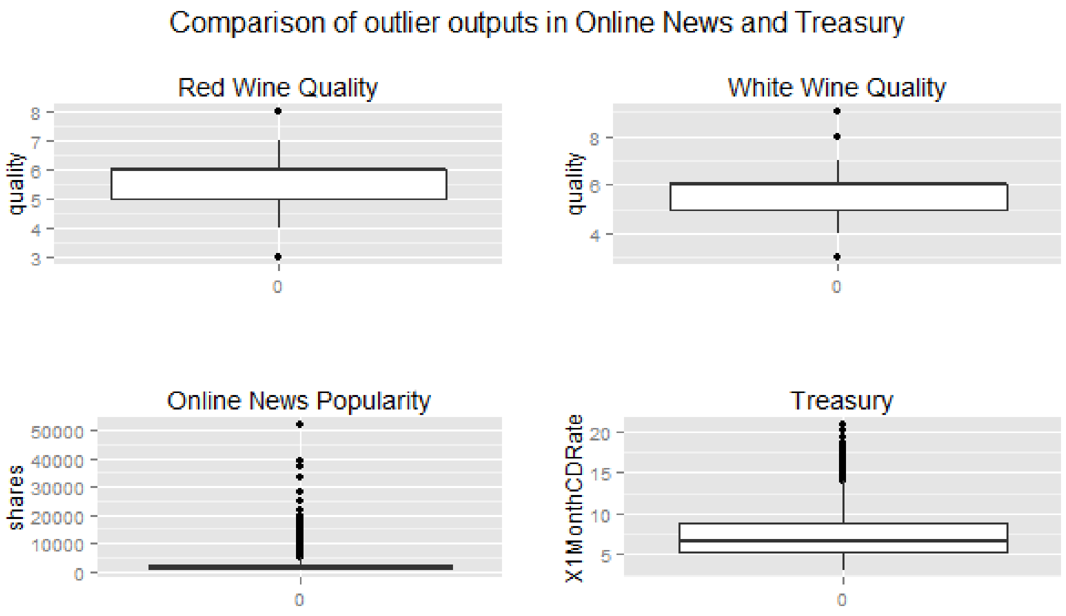

Figure 11 Comparison of class output value distributions between datasets

4.4 Experimental Results and Statistical Testing

The experimental results in Section 4.3 suggested that active learning algorithms perform consistently better than a random baseline, across a range of datasets. This section aims to verify this claim by testing the results for statistical significance. An aligned ranks Friedman test, as described in Section 2.5.3 is used to account for relative performance differences across datasets, while adjusting for the large differences in AUCs found between different datasets. The aligned rank test is adapted from that described by Wobbrock *et al.* (2011), and applied to active learning for regression by Trawiński *et al.* (2012). Whereas Trawiński *et al*. calculate aligned values by subtracting the mean performance of all algorithms on a given dataset from the performance of each individual algorithm; this implementation expresses the performance of each algoirthm as a proportion of the mean. Without this adjustment, the increased AUC in larger datasets could mask an algorithm's true improvement over the mean.

	QBC5	Density	EGAL	Diversity	EMC	Random
House	**1645.138**	6869.11	2658.679	1675.837	1699.345	2487.919
Treasury	128.6814	246.3101	132.4929	**97.90293**	125.1006	137.2967
Bikes	80845.4	82868.55	80577.95	79912.81	**78550.19**	81723.57
Concrete	5295.104	7678.782	5632.93	**4925.694**	5363.381	5558.064
News	1879243	9.43E+11	3.19E+11	**1837987**	1893041	4.41E+11
Red Wine	518.0445	686.7273	501.9587	**483.296**	523.1188	535.2783
White Wine	**755.7992**	943.2882	825.2064	791.7877	759.2629	797.0211
Cars	658.0499	1009.814	698.0459	**617.5846**	660.4059	705.7087
Yacht	1599.275	2346.35	1486.415	**1462.768**	1522.813	1570.506
Ratings	5.122294	5.602077	5.806533	5.932432	**5.112155**	6.076917

Table 4 Raw AUC scores for all algorithms across all datasets. The most effective algorithm on each dataset is highlighted in bold

Table 4 summarizes the raw AUC scores over each of the datasets. Due to the large difference in average scores between datasets, the absolute values depicted above have been aligned and ranked, as shown in Table 5.

	QBC5	Density	EGAL	Diversity	EMC	Random
House Prices	6	59	54	5	7	38
Treasury	20	58	13	8	10	15
Bikes	45	50	36	22	46	44
Concrete	26	56	29	11	30	35
News	3	4	9	2	1	60
Red wine	23	55	21	12	27	41
White Wine	40	52	34	19	33	42
Cars	24	57	31	14	16	37
Yacht	32	53	17	25	39	49
Ratings	18	43	48	47	28	51

Table 5 Aligned Rank scores for each algorithm over each dataset

The Friedman aligned-ranks test for statistical significance returned a p value < 0.00001, indicating an extremely strong likelihood that there is a statistically significant difference in the performance of at least one of the algorithms in the group. The results of the post-hoc Friedman-Nemenyi test are summarised in Table 6 and Table 7.

	QBC	Density	EGAL	Diversity	EMC
Density	**0.0018***	-	-	-	-
EGAL	0.6292	0.206	-	-	-
Diversity	0.9607	**0.000003***	0.154	-	-
EMC	1	**0.0018***	0.6292	0.9607	-
Random	0.0902	0.8394	0.8912	**0.007***	0.0902

Table 6 Post-hoc Friedman Nemenyi test for statistical differences between algorithms. Algorithms with significant differences in performance are highlighted in bold and marked with an asterix

Algorithm	Groups		
QBC	**-**	**B**	**C**
EMC	**-**	**B**	**C**
EGAL	**A**	**B**	**C**
Diversity	-	**B**	**C**
Density	**A**	**-**	**-**
Random	**A**	**B**	**-**

Table 7 Summary of post-hoc findings. No statistical significance was detected between algorithms sharing membership of any group.

The results of the post hoc test indicate that there is a significant difference between the performance of QBC, EMC and Diversity on the one hand, and Density on the other. The only algorithm which could be statistically verified to outperform a random baseline was Diversity; though EMC and QBC came close. It seems likely that this could be proven in a test using more datasets. While no statistically significant difference was detected between Diversity and any of the other effective selection strategies, empirical evidence from Section 4.3 suggests that this algorithm is most likely to perform best on an unseen dataset.

4.5 Conclusion

Section 4.2 tentatively established generally optimal parameters for the EGAL algorithm in the context of regression problems. While the α parameter was shown to have little impact on the overall accuracy of the resulting linear regression model, the effect of the ω parameter was demonstrated to be more pronounced. As a rule of thumb, high α values and low ω values are recommended for optimal results.

Section 4.3 empirically established that most active learning algorithms tend to outperform a random baseline. The Density selection strategy, however, tended to perform consistently worse. The findings suggest a correlation between the presence of outliers in the dataset and the effectiveness of active learning algorithms. The results of the random selection strategy on a dataset containing a large number of outliers suggest that active learning algorithms perform best where passive learning algorithms fail.

Section 4.4 statistically verified the hypothesis that active learning algorithms can increase the accuracy of linear regression models. While no statistical difference was detected between any of the effective algorithms, the evidence suggested that a diversity-based approach to query selection usually outperforms the alternatives.

5 FINDINGS

5.1 Importance of Density and Diversity in Active Learning for Regression

The experimental analysis in Section 4 has shown that a diversity-based approach to active learning regularly outperforms the state-of-the-art alternatives. The consistently poor performance of the density-based approach reinforces the notion that choosing observations for labelling based on their density has a significant impact on the performance of the learning model. Diversity-only approaches to active learning are rarely mentioned in the literature; so it is highly likely that this phenomenon is particular to active learning for regression problems.

A linear regression algorithm attempts to find the line of "best fit" which minimizes the error across the dataset. This can be represented on a 2 dimensional plane, where the input features are mapped to an *x* value, and the class value, or output is represented on the *y* axis. As new observations are labelled and added to the model, the regression function, mapping observations to output values is updated to accommodate the new data. Figure 12 illustrates how diversity-based approaches to active learning exploit this property.

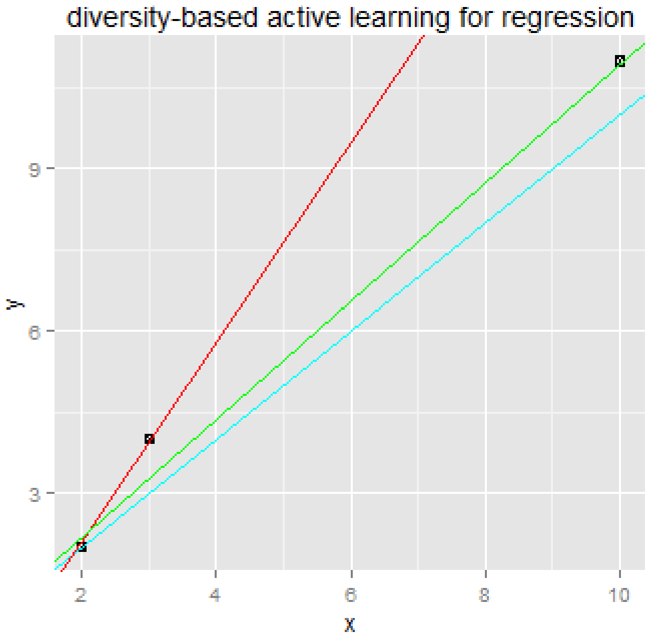

Figure 12 Illustrating the benefit of diversity in active learning for regression

Consider the simple regression function $y = x$ illustrated above. Having labelled the point (2, 2), the algorithm now has to choose between two observations, equally corrupted with noise of 1. The shorter distance between (2, 2) and (3, 4) will result in a greater change to the slope of the regression function, meaning the impact of the noise will be more pronounced. On the other hand, labelling the observation at (10, 11) will cause the slope to shift only slightly, reducing the impact of the noise on the accuracy of the overall model. This effect may be exacerbated by the evaluation framework which removes labelled data from the model scoring. Labelling outliers early means the model will not be asked to re-evaluate them in future iterations, and can thereby avoid repeated large errors.

The same effect may be occurring in both the QBC and EMC frameworks. Linear regression models aim to minimize the total error across the entire dataset, so the regression function will be adjusted to fit areas with a higher density of labelled data, at the expense of lower density areas if doing so reduces the overall error. Because of this, models trained on differing subsets of the data are more likely to disagree on areas which are sparsely labelled, as these are the areas in which the function is most likely to deviate from the data. Both of these frameworks will therefore, to a certain extent approximate the diversity-based approach, which may explain their superior performance to EGAL and the baseline.

5.2 Low Performance of EGAL in Active Learning for Regression

EGAL was the only effective active learning algorithm which failed to outperform the random baseline. This appeared to be a surprising result given its strong performance in classification tasks. Section 5.1 has shown the drawbacks of density-weighted algorithms, however; and suggested that this effect may be more pronounced in regression tasks than in classification. The EGAL selection strategy favours the densest observations for labelling once they are part of the candidate set. The candidate set, however, is recalculated only at the beginning of each batch. As the algorithm is density-weighted, it is likely to label entire clusters in a single batch, which may explain EGAL's oscillating error curves. If the algorithm were to be adjusted, such that the candidate set was pruned after every label was selected, this could be avoided and the accuracy of EGAL could be expected to improve as a consequence.

6 CONCLUSION

This section summarizes the key objectives and goals of this project. The key findings are summarised, and the project's contributions to the body of knowledge is outlined. Finally, further work and research is recommended.

6.1 Problem Definition & Research Overview

This project aims to establish whether or not active learning strategies can be successfully applied to training linear regression models. A lack of research in the area of active learning for regression problems, and more specifically, for a solid statistical underpinning ground to this claim was the catalyst for the research. As the field of active learning for regression has received little attention, the opportunity was taken to apply EGAL, a strategy proven in the field of classification, to a regression context.

6.2 Experimentation, Evaluation & Limitations

This dissertation established that active learning algorithms can significantly improve the accuracy of linear regression models. However, the only algorithm which could be conclusively shown to outperform a random baseline was Diversity. Empirical evidence suggests that most of the algorithms under scrutiny provide consistent benefits, but further evaluation on a larger number of datasets would be required to verify this claim.

The EGAL algorithm did not perform as well as was hoped and may need further adjustments in order to make it suitable for regression problems.

The data used in this research was taken mostly from the UCI Machine Learning repository; and the size of the datasets were limited by the computational effort required to train models. The increasing public use of the internet has made ever larger datasets available in recent years, and work is being done on sentiment analysis on websites such as Twitter (Go et al., 2009) and Amazon (Blitzer et al., 2007). Only one comparable dataset was included in this study. In order to provide more of a "real-world" context, ideally more such datasets would have been utilized.

It is important to remember that the evaluation criteria used are not typical of most machine learning problems. Whereas prediction models are required to label previously unseen data; our evaluation task makes so such assumption. Models are trained purely to aid in evaluating the *current dataset*; and the extent to which a model has been over-fitted to the data is not measured. This is not necessarily a limitation in itself, but it becomes one when comparing this research to existing experiments. Allowing the model to ignore the generalisation error makes this an easier problem than training a model which must aim to generalize easily.

6.3 Contributions to Body of Knowledge

This dissertation has provided a statistically verified comparison of the effectiveness of multiple active learning selection strategies across a broad selection of real-world datasets. While traditional approaches which have worked well in a classification context appear to consistently outperform a random baseline; a diversity-only approach, which has received less attention in the literature was determined to be the only algorithm for which this claim could be statistically verified.

The applicability of EGAL to regression problems has been explored; with optimal parameters suggested for use across datasets. The performance of EGAL on datasets with different class value distributions has been explored; laying the groundwork for future improvements to the algorithm in the context of regression-based learning.

This study suggests the use of the Friedman aligned ranks test when comparing multiple regression classifiers, to cope with the incommensurability of the raw performance measures for regression, while taking into account the intuitively apparent relative "magnitude" of performance differences across datasets.

6.4 Future Work & Research

While the EGAL selection strategy has been applied in the context of regression problems, the evidence suggests that the algorithm may need modification before it is suitable for use in this new environment. There is scope for further research on how the algorithm can be improved, and whether density can be utilised, under certain circumstances to improve the performance of an active learning algorithm.

Only a single active learning algorithm was shown to significantly outperform a random baseline. However, the results would suggest that EMC and QBC may also be statistically significant. Further testing across a broader range of datasets may well to establish this hypothesis statistically.

6.5 Conclusion

This project aimed to answer the question of whether "active learning selection strategies based on integral dataset properties combined with an analysis of prediction model output be used successfully for linear regression models". The research has conclusively shown that the answer is an emphatic "yes". While there is little doubt that active learning selection strategies can be used successfully for linear regression models, only a single strategy could be verified to significantly outperform the baseline. Fortunately, the empirical results are encouraging and it seems likely that all mainstream classification algorithms tested are also useful in the domain of regression problems. The success of the Diversity based approach, however, suggests that regression problems and classification problems cannot be treated identically; and that perhaps algorithms successful in a classification context may benefit from some altering before being applied to problems of regression.

7 REFERENCES

Alpaydin, E. 1999. Combined 5× 2 cv F test for comparing supervised classification learning algorithms. *Neural computation,* 11**,** 1885-1892.

Alpaydin, E. 2014. *Introduction to machine learning,* MIT press.

Anderson, B. & Moore, A. Active learning for hidden markov models: Objective functions and algorithms. Proceedings of the 22nd international conference on Machine learning, 2005. ACM, 9-16.

Angluin, D. 1988. Queries and concept learning. *Machine learning,* 2**,** 319-342.

Baum, E. B. & Lang, K. Query learning can work poorly when a human oracle is used. International Joint Conference on Neural Networks, 1992.

Blitzer, J., Dredze, M. & Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. ACL, 2007. 440-447.

Breiman, L. 1996. Bagging predictors. *Machine learning,* 24**,** 123-140.

Burbidge, R., Rowland, J. J. & King, R. D. 2007. Active learning for regression based on query by committee. *Intelligent Data Engineering and Automated Learning-IDEAL 2007.* Springer.

Cai, W., Zhang, Y. & Zhou, J. Maximizing expected model change for active learning in regression. Data Mining (ICDM), 2013 IEEE 13th International Conference on, 2013. IEEE, 51-60.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting,* 5**,** 559-583.

Cohn, D., Atlas, L. & Ladner, R. 1994. Improving generalization with active learning. *Machine learning,* 15**,** 201-221.

Cohn, D. A. 1996. Neural network exploration using optimal experiment design. *Neural networks,* 9**,** 1071-1083.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research,* 7**,** 1-30.

Derrac, J., García, S., Molina, D. & Herrera, F. 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation,* 1**,** 3-18.

Dietterich, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation,* 10**,** 1895-1923.

Douak, F., Melgani, F. & Benoudjit, N. 2013. Kernel ridge regression with active learning for wind speed prediction. *Applied Energy,* 103**,** 328-340.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *The annals of Statistics***,** 1-26.

Go, A., Huang, L. & Bhayani, R. 2009. Twitter sentiment analysis. *Entropy,* 17.

Hu, R. 2011. *Active Learning For Text Classification.* Doctorate, Dublin Institute of Technology.

James, G., Witten, D. & Hastie, T. 2014. An Introduction to Statistical Learning: With Applications in R. Taylor & Francis.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P. & Soldatova, L. N. 2009. The automation of science. *Science,* 324**,** 85-89.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G., Bryant, C. H., Muggleton, S. H., Kell, D. B. & Oliver, S. G. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature,* 427**,** 247-252.

Lewis, D. D. & Gale, W. A. A sequential algorithm for training text classifiers. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 1994. Springer-Verlag New York, Inc., 3-12.

Loy, C. C., Hospedales, T. M., Xiang, T. & Gong, S. Stream-based joint exploration-exploitation active learning. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012. IEEE, 1560-1567.

Luengo, J., García, S. & Herrera, F. 2009. A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests. *Expert Systems with Applications,* 36**,** 7798-7808.

MacKay, D. 1992. Information-based objective functions for active data selection. *Neural computation,* 4**,** 590-604.

Margineantu, D. D. Active cost-sensitive learning. IJCAI, 2005. 1622-1623.

McCallum, A. K. & Nigam, K. Employing EM and pool-based active learning for text classification. International Conference on Machine Learning (ICML), 1998. Citeseer.

Opitz, D. & Maclin, R. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research***,** 169-198.

RayChaudhuri, T. & Hamey, L. G. Minimisation of data collection by active learning. Neural Networks, 1995. Proceedings., IEEE International Conference on, 1995. IEEE, 1338-1341.

Roy, N. & McCallum, A. 2001. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*.

Salzberg, S. L. 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery,* 1**,** 317-328.

Schwenker, F. & Trentin, E. 2014. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters,* 37**,** 4-14.

Settles, B. 2010. Active learning literature survey. *University of Wisconsin, Madison,* 52**,** 11.

Settles, B. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning,* 6**,** 1-114.

Settles, B. & Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Honolulu, Hawaii: Association for Computational Linguistics.

Settles, B., Craven, M. & Friedland, L. Active learning with real annotation costs. Proceedings of the NIPS workshop on cost-sensitive learning, 2008a. 1-10.

Settles, B., Craven, M. & Ray, S. Multiple-instance active learning. Advances in neural information processing systems, 2008b. 1289-1296.

Seung, H. S., Opper, M. & Sompolinsky, H. Query by committee. Proceedings of the fifth annual workshop on Computational learning theory, 1992. ACM, 287-294.

Sheskin, D. J. 2003. *Handbook of parametric and nonparametric statistical procedures,* crc Press.

Tong, S. & Koller, D. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.,* 2**,** 45-66.

Trawiński, B., Smętek, M., Telec, Z. & Lasota, T. 2012. Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *International Journal of Applied Mathematics and Computer Science,* 22**,** 867-881.

Vapnik, V. 2013. *The nature of statistical learning theory,* Springer Science & Business Media.

Vázquez, E. G., Escolano, A. Y., Riaño, P. G. & Junquera, J. P. 2001. Repeated measures multiple comparison procedures applied to model selection in neural networks. *Bio-Inspired Applications of Connectionism.* Springer.

Wobbrock, J. O., Findlater, L., Gergle, D. & Higgins, J. J. The aligned rank transform for nonparametric factorial analyses using only anova procedures. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011. ACM, 143-146.

Zhou, Z.-H., Wu, J. & Tang, W. 2002. Ensembling neural networks: many could be better than all. *Artificial intelligence,* 137**,** 239-263.

Zhu, J., Wang, H., Yao, T. & Tsou, B. K. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 2008. Association for Computational Linguistics, 1137-1144.

APPENDIX A

The following graphs, from Section 4.2, have been reproduced, enlarged and rotated for readability.

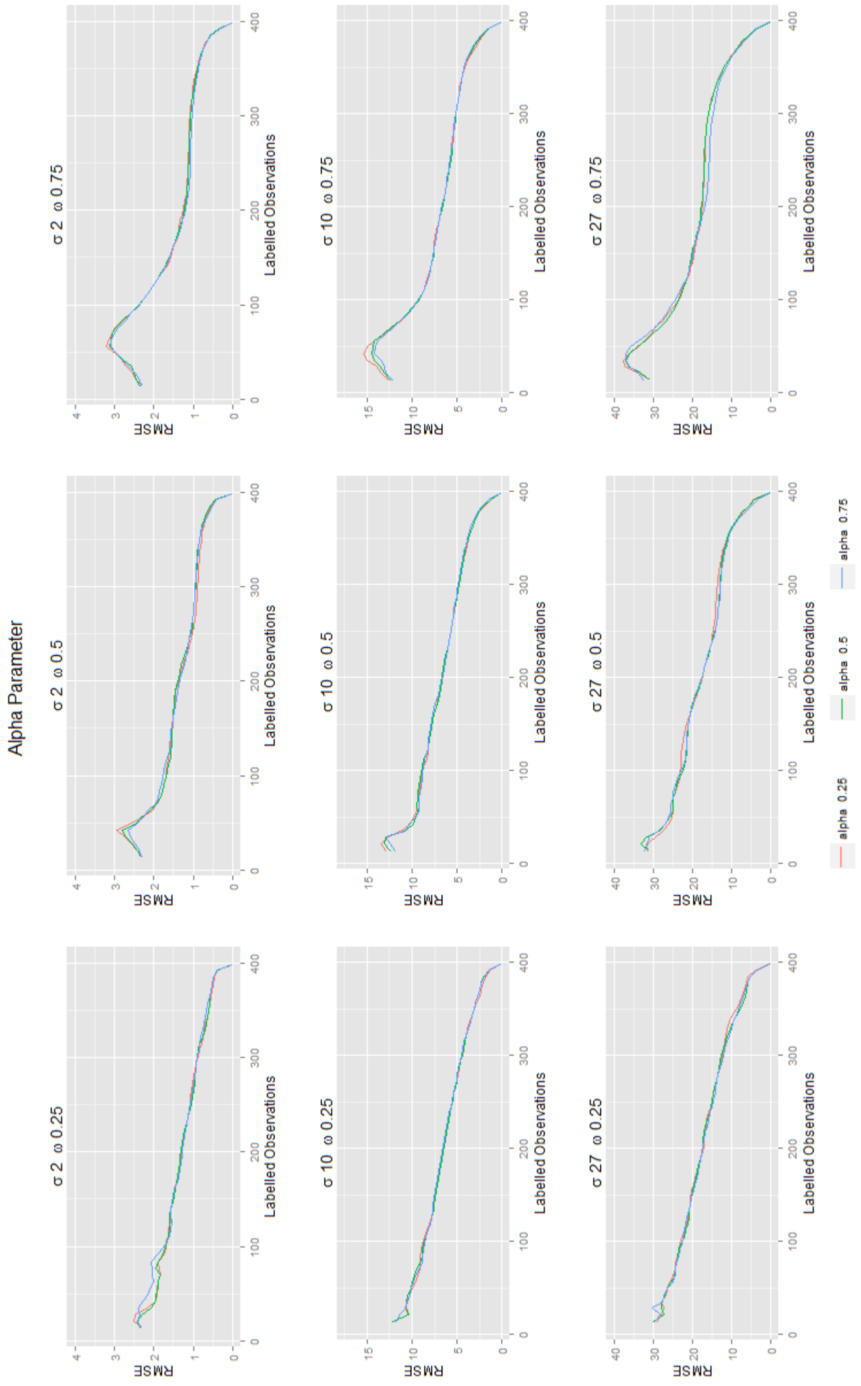

Figure 13 An enlarged reproduction of Figure 6

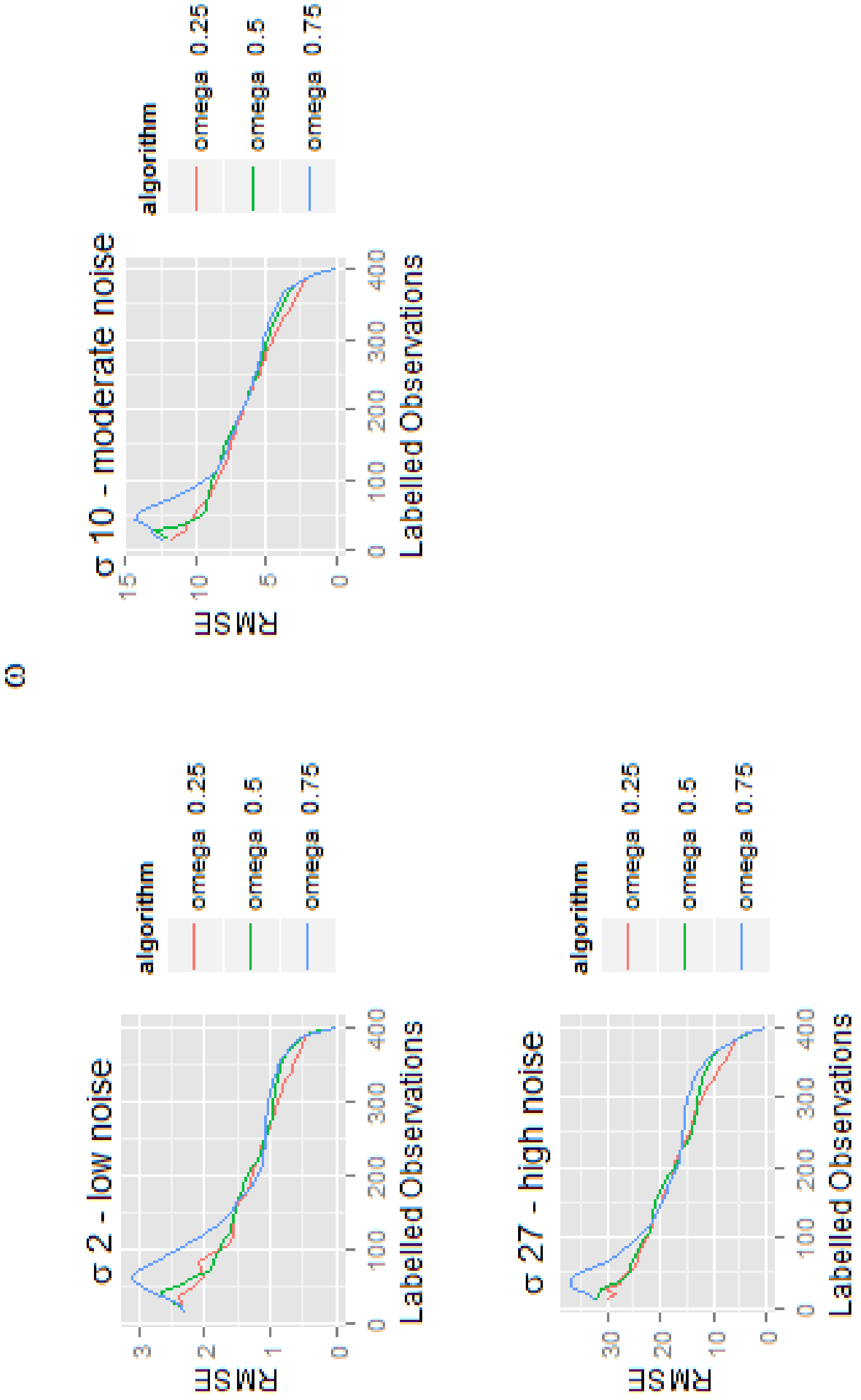

Figure 14 An enlarged reproductin of Figure 7