Machine Learning competitions at Kaggle: Literature Review

Alan Lynch

registration: 100094667

1 Introduction

In this paper I will cover paper's that have contained useful information that will assist in my third year project. This will be done in two sections, section 3 will cover the technical definitions and formal algorithms of content from the papers collected. Section 4 will outline the content covered in each paper and conclude which papers are most useful.

2 Project Description

My project is Machine Learning Competitions at Kaggle, the competition that I will be doing is Zillow's Home Value Prediction(Zestimate), in which I will be looking at predicting house prices.

3 Technical Review

This section describes the core algorithms and techniques used in the papers covered.

3.1 Data Preprocessing

The process of cleaning, transforming, and discretizing data

3.1.1 Data Cleaning

can be defined as filling in missing data and removing outliers One way of addressing missing data is known as listwise deletion, whereby all records(rows) of data that contain missing data are removed from the sample. However there are problems that arise when using this method, namely reducing the sample size, learning algorithms often need large sample sizes to learn the underlying patterns in the data. Removing all records where missing data occurs can also cause bias. There are a number of different method for filling in missing data: Fill missing data with the mean of the column, use classification/regression to predict what missing data should be using other input variables.

3.1.2 Data transformation

The process of changing input values, using a function, so that the values fall within some bounds. Data transformation is useful as some machine learning algorithms require the values of the input to be within a certain range say, -1 and 1 or 0 and 1. Examples of data transformations are: log-transformations, min-max scaling and mean normalization.

3.1.3 Data discretization

Changing continuous variables into categorical variables. Used to improve accuracy and time taken to classify. Loses information about the output variable as it is now categorical which is less specific and more general than a continuous variable.

3.2 Feature Selection

The process of selecting a subset of features from the dataset that best describe the variable to be predicted. Used for several reasons: Simplifies models, quicker to train learning algorithms on, dimensionality reduction, more generalized model which avoids over fitting.

3.3 Feature Extraction

The process of creating new features based off the original set of features.

3.4 Learning Algorithms

Teaching a machine to classify / predict an output variable(y) based off input variables(x1,...,xn).

3.4.1 Classification Algorithms

Concerned with predicting the category to which the data belongs based off the input variables. (Used when output variable has finite number of values)

3.4.2 Regression Problems

Concerned with predicting the value of a continuous output variable based off the input variables. (Used when output variable is continuous) The Kaggle competition that I am looking at has a continuous output variable and so I will be looking at learning algorithms suitable for regression.

3.4.3 Linear Regression

Looks at predicting the output (Y) variable through the input (X) variable. Takes the form:

$$Y = B0 + B1 * X + e$$

Where Y is the outcome variable, B0 is the intercept, B1 is the regression coefficient and e is the error.

3.4.4 Multiple Regression

The same as linear regression but has multiple X variables. Takes the form:

$$Y = b0 + b1x1 + b2x2 + bnxn + e$$

Where n is the number of input variables

To find the regression coefficients:

$$B = (X'X)^{-1}X'Y$$

The equation for the MSE can be denoted as:

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Where N is the number of predictions, and y-hat is the estimate.

3.4.5 Ridge

Ridge regression is used when there is multicollinearity in the data, that is, there are strong relationships between input variables. It is also useful on complex models, as it is optimized for prediction, it does this by shrinking coefficients, setting them close to 0, which allows for better prediction on new unseen data. The estimation of regression coefficients can be given as:

$$B = (X'X + kI)^{-1}X'Y$$

Where k is a chosen value between 1 and 0 that will be applied to diagonal values. The only way of finding the optimum k value that will minimize MSE (Mean Squared Error), is by plugging values in and finding an optimal k value that produces the lowest MSE, this can be done graphically and cross-validation can be used.

3.4.6 Lasso

Similar to ridge in that it looks at minimizing coefficients and can be used on datasets with multicollinearity, but also performs feature selection as it can set coefficients to 0, essentially excluding them from the model.

3.4.7 Elastic-Net

A combination of Ridge and Lasso regressions penalties L1 and L2.

3.4.8 Support vector machine

A binary classifier that finds a line of separation between two classes that is as far away as possible from the nearest points of the two classes, unseen data will be classified based on which side of the line it falls under.

3.4.9 k-Nearest Neighbour Classification

A classifier that assigns a label based on the nearest neighbour points, where k is a chosen value.

3.4.10 k-Nearest Neighbour Regression

For regression, the output variable will be determined by the mean of the kth nearest neighbours output values.

4 Literature Review

In this section we discuss the papers researched, and the topics they cover.

4.1 Data Pre-processing

Schmitt P (2015) compares 6 data imputation techniques(Mean, K-near neighbours, fuzzy K-means, singular value decomposition, Bayesian principle component analysis and multiple imputations by chained equations) based on four evaluation criteria (Root mean square error, unsupervised classification error, supervised classification error and execution time). They concluded that bPCA and FKM performed the best of the 6 methods.

Gutierrez-Osuna and Nagle (1999) focuses on the evaluation of preprocessing transformations: baseline manipulation, compression and normalization. Baseline manipulation can be broken down into: differential, relative and fractional Compression is broken down into: Steady state, Transient integral and windowed time slicing Normalization is broken down into: vector normalization, vector auto-scaling and dimension auto-scaling. They propose two metrics for evaluating these techniques: information content and predictive accuracy. After defining both metrics they decide to use predictive accuracy. They then look to evaluate every possible combination of baseline manipulation, compression and normalization techniques across 4 datasets. They then show bar graphs of the different techniques on their predictive accuracy, and draw conclusions.

Leys et al. (2013) discusses the merits of using the median absolute deviation, and some of the pitfalls of using the standard deviation around the mean. They suggest that the standard deviation around the mean can be sensitive to outliers, something that the median absolute deviation overcomes. They also suggest that the median absolute deviation is âĂIJimmune to sample sizeâĂİ.

Ramaswamy et al. (2000) outlines a distance-based definition for detecting outliers suggested by Knorr and Ng. It continues by looking at distance-based algorithms: block-nested loop and Index based join, it also shows pseudo-code for both. They then suggest there own partition-based algorithm and compare performance with the previously discussed algorithms.

S. B. Kotsiantis and Pintelas (2007) This paper outlines: data cleaning, normalization, transformation, feature extraction and feature selection.

4.1.1 Feature Selection

Cawley and Talbot (2010) discusses the importance of low-variance in model selection criterion for avoiding over-fitting models and how over-fitting can cause selection bias in some performance evaluation metrics. The model selection criterion chosen is k-fold cross-validation. The paper looks at Kernel Ridge Regression as the main model and briefly describes it.

Kira and Rendell (1992) covers the strengths and weaknesses of current feature selection techniques and then introduces a new technique known as Relief. This appears to be quite an old paper but still useful. Compares Relief with Focus based on Accuracy and Learning time.

Hall (2000) introduces a new filter feature selection technique known as Correlation-based Feature Selection (CFS) and compares it to ReliefF.

4.2 Machine learning papers

Liaw and Wiener (2002) goes through the random forest algorithm for both classification and regression problems.

Schapire (2001) focuses on a method known as boosting to improve learning algorithms, In particular the paper looks at the AdaBoost algorithm.

Gunn (1998) looks at Support Vector Classification(SVC) and Support Vector Regression(SVR) to solve learning problems.

References

Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, 11:2079–2107.

Gunn, S. R. (1998). Support vector machines for classification and regression.

Gutierrez-Osuna, R. and Nagle, H. T. (1999). A method for evaluating data-preprocessing techniques for odour classification with an array of gas sensors. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 29 5:626–32.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, pages 129–134. AAAI Press.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY*, 49(4):764–766.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec.*, 29(2):427–438.

S. B. Kotsiantis, D. K. and Pintelas, P. E. (2007). Data preprocessing for supervised leaning. *International Journal of Computer and Information Engineering*.

Schapire, R. E. (2001). The boosting approach to machine learning an overview.

Schmitt P, Mandel J, G. M. (2015). A comparison of six methods for missing data imputation. *Journal of Bioinformatics and Biostatistics*, page 6.

Literature review

Introduction: brief description of project, areas of knowledge required, roadmap	First	2.1	2.2	3	Fail
Discovery of suitable quantity and quality of material	First	2.1	2.2	3	Fail
Description of key issues and themes relevant to the project	First	2.1	2.2	3	Fail
Evaluation, analysis and critical review	First	2.1	2.2	3	Fail

Quality of writing

Clarity, structure and correctness of writing	First	2.1	2.2	3	Fail
Presentation conforms to style (criteria similar to conference paper reviews)	First	2.1	2.2	3	Fail
References correctly presented, complete adequate (but no excessive) citations	First	2.1	2.2	3	Fail

Revised Workplan (if applicable)

Measurable objectives : appropriate, realistic, timely	First	2.1	2.2	3	Fail

Comments

Supervisor: Dr Gavin Cawley

Markers should circle the appropriate level of performance in each section. Report and evaluation sheet should be collected by the student from the supervisor.