Selecting Key Predictor Parameters for Regression Analysis using Modified Neighbourhood Component Analysis (NCA) Algorithm

[1]B. Amankwaa-Kyeremeh, [2]C. Greet, [1,3]M. Zanin, [1]W. Skinner and [1]R. K. Asamoah
[1] University of South Australia, Future Industries Institute, Mawson Lakes, Adelaide, SA 5095, Australia
[2]Magotteaux Pty Ltd, Rear of 31 Cormack Rd, Wingfield SA 5013, Australia
[3]University of Adelaide, Adelaide SA 5005, Australia

Amankwaa-Kyeremeh, B., Greet, C., Zanin, M., Skinner, W. and Asamoah, R. K., (2020), "Selecting key predictor parameters for regression analysis using modified Neighbourhood Component Analysis (NCA) Algorithm", *Proceedings of 6th UMaT Biennial International Mining and Mineral Conference*, Tarkwa, Ghana, pp. 320-325.

Abstract

Selecting the most useful features for the purpose of regression analysis is very critical in ensuring good prediction. In this research, modified Neighbourhood Component Analysis (NCA) algorithm has been used as a feature selection criterion for selecting the most relevant parameters from 25 rougher flotation parameters. Predictor parameters selected to be relevant for regression analysis included throughput, feed particle size, frother dosage, xanthate dosage and froth depth as confirmed in literature. This result is a clear indication that modified NCA Algorithm can select relevant features for the purpose of regression analysis.

Keywords: Neighbourhood Component Analysis, Regularisation Term, Feature Selection, Machine Learning

1 Introduction

The application of supervised machine learning algorithms in froth flotation modelling is becoming popular because of the advantages they exhibit over traditional first order kinetic models (Aldrich *et al.*, 1997; McCoy and Auret, 2019). Unlike traditional empirical and first order kinetic models which have phenomenological interpretation, machine learning models are nonparametric and therefore the number of model parameters are not dependent on domain knowledge (Von Stosch *et al.*, 2014). Supervised learning algorithms are often used for the purpose of regression and in recent years, researchers like Jahedsaravani *et al.* (2014), Jahedsaravani *et al.* (2016), Massinaei and Doostmohammadi (2010) have been successful in applying this knowledge for the prediction of key flotation performance indicators (recovery and grade) as a function of process measurements. However, one main challenge posed to the development of regression models is the handling of data set made up of so many features.

Froth flotation is known to be affected by so many parameters. Arbiter and Harris (1962) estimated that, up to about a hundred parameters can affect froth flotation process. In such a situation, feature selection which is the process of selecting only relevant parameters for the purpose of modelling must be carried out. Feature selection methods helps in identifying irrelevant and redundant parameters that do not contribute significantly to the accuracy of predictive models. Feature selection methods can significantly improve accuracy, reduce learning times and simplify learning results (Liu and Motoda, 2012; Zhao *et al.*, 2010). Feature selection can broadly be grouped into wrapper, embedded and filter methods (Bolón-Canedo *et al.*, 2014). Common examples of the various categories that have been used over the years include Principal Component Analysis (PCA), Independent Component Analysis (ICA), Pearson correlation criteria, Mutual Information (MI), Sequential feature selection algorithm and Genetic algorithm (Battiti, 1994; Chuang *et al.*, 2008; Goldberg, 2006 ; Guyon and Elisseeff, 2003; Reunanen, 2003). However, we intend to use modified Neighbourhood Component Analysis (NCA) as a feature selection criterion for this work. Modified NCA as a feature selection technique is known to provide good performance, reduce the number features and comes with high accuracy as compared to the other feature

selection techniques (Yang *et al.*, 2012a). It can be observed from literature that, most of the applications of NCA are for classification purposes, however, modified NCA Algorithm in MATLAB Simulink R2020a has been applied for regression purpose in this work.

The main purpose of this work is to ascertain whether the modified NCA can select meaningful features for regression purpose. This work is grouped into four sections. Following this introduction, we defined the methodologies adopted for this paper in Section 2. Section 3 emphasises on results and discussions of the main findings of this work and finally concluding in Section 4.

2 Methodology

2.1 Data collection and pre-processing

A total of 25 rougher flotation (1.5 million observations each) parameters and its corresponding time stamped rougher recovery data have been extracted from the historical data of a typical copper flotation plant. For the purpose of confidentiality, the name of the company cannot be disclosed. The data set extracted was made up of elemental assays of rougher feed as well as data on other rougher flotation parameters. Typical industrial data especially those from complex process like froth flotation are known to have inherent issues that make them inappropriate for machine learning algorithms (Ge *et al.*, 2013 ; Kadlec *et al.*, 2009). Issues associated with data set included missing and faulty values, repeated values, outliers and varying sample sizes as a result of different sampling rate of each parameter. An observation was considered an outlier when it fails to fall in the range (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR) where Q1= lower quartile, Q3 = upper quartile and IQR = interquartile range. The large size of the data set (> 1 million observations each) made it convenient to delete observations that had issues. Now, using copper recovery data as reference parameter, any of its observations that had issue were deleted along with its corresponding observations in the various parameters. Copper recovery was used as the reference parameter because that was the main key flotation performance indicator. This approach was adopted not only to filter the data but also to ensure that the cleaned data sets had equal sizes in terms of their observations.

2.2 Feature selection by Neighbourhood Component Analysis (NCA)

Selection of appropriate features for flotation modelling is very critical in ensuring the success of good predictive model. Modified NCA Algorithm has been used for feature selection in this research. NCA is a non-parametric method that selects relevant features with the goal of maximizing prediction accuracy of regression models. Considering a training set

$$\mathrm{S} = \{(\mathrm{x_i}, \mathrm{y_i}), \mathrm{i} = 1,2,3,4, \ldots . \mathrm{n}\}$$

Where n = total number of observations, x = input variable and y = output variable. A randomised regression model picks a point from the input variables in S say point x_j with its corresponding output value y_j. The probability $P(x_j|S)$ that x_j is picked from S as the reference point for x is higher if x_j is closer to x as measured by the distance function d_w, where $d_w(x_i, x_j) = \sum_{r=1}^{p} w_r^2 |x_{ir} - x_{jr}|$ and w_r are the feature weights assuming that $P(x_j|S) \propto k\left(d_w(x, x_j)\right)$. k is some kernel or a similarity function that assumes large values when $d_w(x_i, x_j)$ is small. $P(x_j|S)$ for all j must be equal to 1 and therefore it is possible to write $P(x_j|S) = \frac{k(d_w(x,x_j))}{\sum_{j=1}^{n} k(d_w(x,x_j))}$ (Yang *et al.*, 2012b).

In the same vein, considering the leave-one-out application of this randomised regression model using data in S^{-i} (i.e.) the training set without the point (x_i, y_i), the probability that point x_j is selected as the reference point for x_i is given by:

$$P_{ij=} P(x_j|S^{-i}) = \frac{k(d_w(x_i, x_j))}{\sum_{j=1, j\neq i}^{n} k(d_w(x_i, x_j))}$$

Now assigning y_b to be the output value predicted by a randomised regression model, y_a be the true value for x_i and l be the loss function that measures the difference between y_a and y_b, the average value of $l(y_b, y_a)$ becomes $l_i = E(l(y_b, y_a)|S^{-i}) = \sum_{j=1, j\neq i}^{n} p_{ij}(y_i, y_j)$. The loss function $l(y_b, y_a)$ used for this work is mean squared error $(y_i - y_j)^2$ (Yang *et al.*, 2012b). A regularisation term λ is added to the final objective function to avoid overfitting of the randomised regression model. The objective function $f(w)$ for minimisation after adding a regularisation term λ becomes:

$f(w) = \frac{1}{n}\sum_{i=1}^{n} l_i + \lambda \sum_{r=1}^{p} w_r^2$

Modified NCA function fsrnca was called to carry out the feature selection using 25 parameters as predictors and copper recovery as a corresponding response variable. Both default and tuned λ values have been used for feature selection.

3 Results and Discussion

In this section, the results for applying modified NCA feature selection algorithm on 25 predictors have been presented. The initial regularisation term λ (lambda) was first set to a default value of $\frac{1}{n}$ where n is the total number of observations. Feature weights of the 25 predictor parameters have been computed as shown in Fig.1. A parameter was considered relevant if its weight is greater than zero. As seen from Fig.1, more than half of the predictors had weights greater than zero.

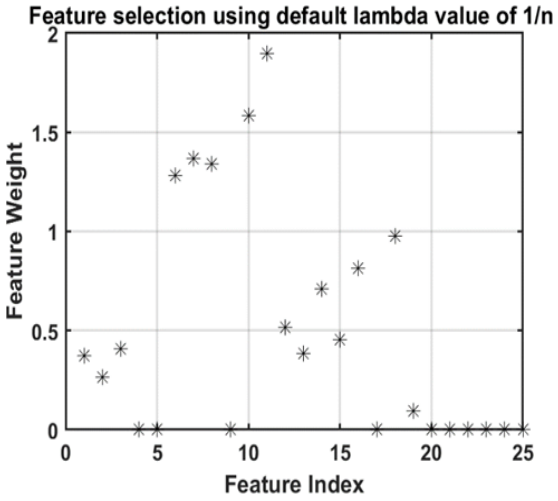

Fig. 1 Predictors feature weights using default lambda values

Now to improve performance, λ was tuned using five-folds cross validation. A five-fold cross validation was chosen in order to trade off between variance and bias error as well as reduce the computational cost and memory usage considering the large data size. NCA randomised regression

model was trained for each λ using training set in each fold. The regression loss for the corresponding test or validation test in each fold was computed. Average losses for the folds were computed and the λ value that gives the minimum loss was used as the new default regularisation term in calling the modified NCA function. As seen from Fig.2, the best λ was determined to be 0.0269 with a corresponding loss of 0.5881.

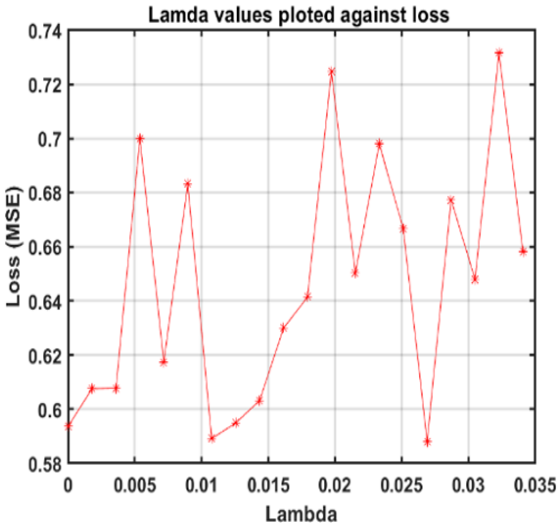

Fig. 2 Plot of average λ values against loss

Figure 3 shows the new feature weights of the 25 parameters using the best λ value.

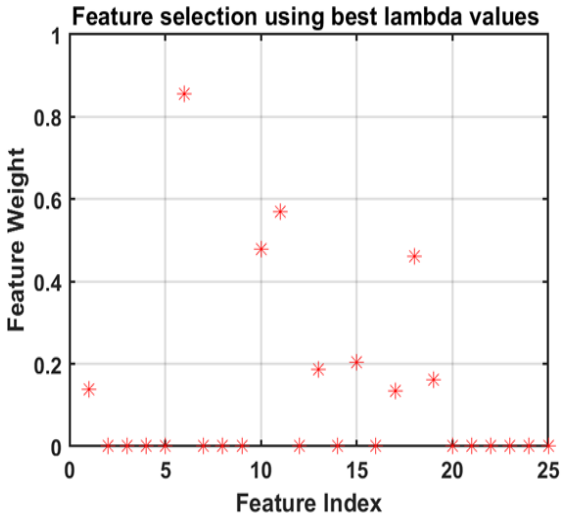

Fig. 3 Predictors feature weights using best lambda value.

It can be seen from Fig. 3 that the number of parameters with feature weights greater than zero has reduced from 14 to 9 which implies that only the most relevant features get selected. Table 1 gives the summary of selected parameters with their corresponding feature weights.

Table 1 Feature weights of relevant predictor parameters

Parameter	Feature weight
Throughput	0.1375
Feed particle size	0.8564
Total xanthate to tank 1	0.4787
Total xanthate to tank 4	0.5703
Froth depth of tanks 4 and 5	0.1862
Froth depth of tanks 2 and 3	0.2041
Froth depth of tank 1	0.1344
Frother to tank 4	0.4611
Frother to tank 1	0.1615

As shown in Table 1, relevant predictor parameters for regression included throughput, feed particle size, xanthate dosage, frother dosage and froth depth.

Due to decline in copper ore grades over the years, most industrial plants now focus on keeping throughput at maximum without compromising residence time. Throughput has a general trend of increasing recovery to a maximum before it begins to decline upon further increment. High throughput implies high solid mass recovery from flotation process explaining its significance in predicting copper recovery.

Feed particle size is an important parameter which has gained attention when it comes to the processing of minerals by froth flotation (Trahar, 1976). Fine, intermediate and coarse particles are known to respond differently to froth flotation. Studies have shown that best flotation recoveries are obtained from particles in the range of 10-150 μm (Shergold, 1984). In a research conducted by Chelgani *et al.* (2010), they predicted froth flotation recovery and collision probability based on flotation operational parameters using artificial neural network and regression procedures. In their research, feed particle size was one of the predictor parameters that helped in predicting flotation recovery with correlation coefficient of 0.98 respectively.

Furthermore, Nakhaei *et al.* (2012) predicted copper grade and recovery of a pilot flotation plant using non-linear regression model approach. Chemical reagent dosage and froth depth were the among the predictor parameters which help in predicting copper recovery with a correlation coefficient of 0.86.

These literature references confirm the robust efficiency of the modified NCA algorithm for selecting relevant features as explanatory variables in regression analysis which contribute significantly to the accurate prediction of the response variable.

4 Conclusion

Modified Neighbourhood Component Analysis (NCA) algorithm has been applied as a feature selection criterion for selecting relevant parameters for regression analysis. A total of 25 flotation parameters were used as predictors with copper recovery as the main response variable. Both default and tuned regularisation term (λ) were used to select the most relevant predictor parameters. Default λ value selected 14 relevant predictor parameters. However, when λ was fine-tuned to average λ value that gives the minimum loss (mean squared error), the number of relevant predictor parameters reduced from 14 to 9 and included throughput, feed particle size, frother addition, xanthate addition and froth depth as confirmed in literature. These results imply that the modified NCA Algorithm can be used as additional feature selection tool to improve the prediction accuracy of a model in regression analysis.

Acknowledgements

This research has been supported by the South Australian Government through the PRIF RCP Industry Consortium. The authors will also like to appreciate the support of the Future Industries Institute of the University of South Australia.

References

Aldrich, C, Moolman, D, Gouws, F and Schmitz, G (1997), "Machine learning strategies for control of flotation plants", *Journal of Control Engineering Practice*, vol. 5, no. 2, pp. 263-269.

Arbiter, N and Harris, CC (1962), "Flotation kinetics", *Journal of Froth flotation*, vol. 50, no. 1.

Battiti, R (1994), "Using mutual information for selecting features in supervised neural net

learning", *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537-550.

Bolón-Canedo, V, Sánchez-Marono, N, Alonso-Betanzos, A, Benítez, JM and Herrera, F (2014), "A review of microarray datasets and applied feature selection methods", *Journal of Information Sciences*, vol. 282, pp. 111-135.

Chelgani, SC, Shahbazi, B and Rezai, B (2010), "Estimation of froth flotation recovery and collision probability based on operational parameters using an artificial neural network", *International Journal of Minerals, Metallurgy and Materials*, vol. 17, no. 5, pp. 526-534.

Chuang, L-Y, Chang, H-W, Tu, C-J and Yang, C-H (2008), "Improved binary PSO for feature selection using gene expression data", *Journal of Computational Biology Chemistry*, vol. 32, no. 1, pp. 29-38.

Ge, Z, Song, Z and Gao, F (2013), "Review of recent research on data-based process monitoring", *Industrial Engineering Chemistry Research*, vol. 52, no. 10, pp. 3543-3562.

Goldberg, DE (2006), *Genetic algorithms*, Pearson Education India,

Guyon, I and Elisseeff, A (2003), "An introduction to variable and feature selection", *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157-1182.

Jahedsaravani, A, Marhaban, MH and Massinaei, M (2014), "Prediction of the metallurgical performances of a batch flotation system by image analysis and neural networks", *Minerals Engineering*, vol. 69, 12/01, pp. 137–145.

Jahedsaravani, A, Marhaban, M, Massinaei, M, Saripan, M and Noor, S (2016), "Froth-based modeling and control of a batch flotation process", *International Journal of Mineral Processing*, vol. 146, pp. 90-96.

Kadlec, P, Gabrys, B and Strandt, S (2009), "Data-driven soft sensors in the process industry", *Computers & Chemical Engineering*, vol. 33, no. 4, pp. 795-814.

Liu, H and Motoda, H (2012), *Feature selection for knowledge discovery and data mining*, Springer Science & Business Media,

Massinaei, M and Doostmohammadi, R (2010), "Modeling of bubble surface area flux in an industrial rougher column using artificial neural network and statistical techniques", *Journal of Minerals Engineering*, vol. 23, no. 2, pp. 83-90.

McCoy, J and Auret, L (2019), "Machine learning applications in minerals processing: A review", *Journal of Minerals Engineering*, vol. 132, pp. 95-109.

Nakhaei, F, Mosavi, MR, Sam, A and Vaghei, Y (2012), "Recovery and grade accurate prediction of pilot plant flotation column concentrate: Neural network and statistical techniques", *International Journal of Mineral Processing*, vol. 110-111, 2012/07/18/, pp. 140-154.

Reunanen, J (2003), "Overfitting in making comparisons between variable selection methods", *Journal of machine learning research*, vol. 3, no. Mar, pp. 1371-1382.

Shergold, H 1984, 'Flotation in mineral processing', *The Scientific Basis of Flotation*, Springer, pp. 229-287.

Trahar, W (1976), "The selective flotation of galena from sphalerite with special reference to the effects of particle size", *International Journal of Mineral Processing*, vol. 3, no. 2, pp. 151-166.

Von Stosch, M, Oliveira, R, Peres, J and de Azevedo, SF (2014), "Hybrid semi-parametric modeling in process systems engineering: Past, present and future", *Journal of Computers Chemical Engineering*, vol. 60, pp. 86-101.

Yang, W, Wang, K and Zuo, W (2012a), "Neighborhood Component Feature Selection for High-Dimensional Data", *Journal of Computers in Biology*, vol. 7, no. 1, pp. 161-168.

Yang, W, Wang, K and Zuo, W (2012b), "Neighborhood Component Feature Selection for High-Dimensional Data", *Journal of Computers*, vol. 7, no. 1, pp. 161-168.

Zhao, Z, Morstatter, F, Sharma, S, Alelyani, S, Anand, A and Liu, HJAfsr (2010), "Advancing feature selection research", pp. 1-28.

Authors

Bismark Amankwaa-Kyeremeh is a PhD candidate at the Future Industries Institute (University of South Australia). Bismark has a B.Sc. Hons. in Minerals Engineering from the University of Mines and Technology, Tarkwa, Ghana. His research interest includes froth flotation and application of machine learning in froth flotation operation.

Christopher Greet holds a PhD in Mineral and Resources Engineering, University of South Australia. He has worked with several mine projects within Australia. His research interest includes flotation, mineral surface chemistry etc. He is currently working as a manager metallurgy, Minerals Processing Research at Magotteaux, Adelaide Australia.

Max Zanin is Associate Research Professor in Mineral Processing. He holds a BEng (Hons) in Mineral Processing Engineering (University of Trieste) and a PhD in Geo-engineering (University of Cagliari). His research interest specialises in Mineral processing: froth flotation, physical separation, Urban Mining and solid waste treatment, Sustainable use of resources and optimization of processes. He is member of the Australasian Institute of Mining & Metallurgy (AusIMM), and of the Australian Colloid and Interface Society (ACIS). Max is currently part of the CSIRO Sustainable E-waste Processing initiative.

William Skinner is a Research Professor and Strand Leader - Minerals and Resource Engineering, Future Industries Institute (FII), University of South Australia, Australia. Areas of his research interest include Flotation (pulp and surface chemistry); Leaching (including heap); Physical Separation and other nit Operations (including surface effects); Surface Chemical Control in Grinding/Milling; Mineral formation-Processing Relationships; Bulk Property-Surface Reaction Relationships in Processing Contexts (oxidation, activation, dissolution and molecular adsorption); Agglomeration Chemistry; Impact of Water Chemistry on Processing, Mineral Sands; Synthetic Rutile and Pigment Processing, etc.

Richmond K. Asamoah is a research fellow of the minerals and resources engineering at the Future Industries Institute (University of South Australia), having over eight years in-depth knowledge of and hands-on experience in mineral characterisation and extractive metallurgy in tandem with molecular chemistry and interfacial science. He has specialised diagnostic and prognostic skills in laboratory operations for industry-specific methodical investigations, and simulation and modelling of process flowsheets.