Comprehensive model optimization in pulp quality prediction: a machine learning approach

Chin-Wei Huang[a], Luc Baron[a], Marek Balazinski[a] and Sofiane Achiche[a]

[a]Department of Mechanical Engineering, École Polytechnique de Montréal C.P. 6079, succ. CV, Montréal, Québec, Canada H3C 3A7

ABSTRACT

Feature selection in machine learning is of great interest since it is reckoned as creating more efficient predictive models in several engineering domains. It is even of special importance in the pulp and paper transformation industry as the knowledge of this particular process is generally very limited. In this paper, we first compared the performance of rule-based genetic algorithm and that of adaptive neuro-fuzzy inference system; the latter is found to be more precise in predicting the pulp quality. We then combined several data mining algorithms such as genetic algorithm-partial least square regression, along with other statistical methods, to explore the relevancy of all the potential variables that could be used to predict the pulp ISO brightness, an important property that is usually linked to model performance and hence pulp quality prediction. A few highly relevant variables are thereby determined, and the full set of 79 variables obtained from a Chip Management System was trimmed down to an optimized combination of 3 inputs depending on their relevancy. Peroxide charge (P), average luminance (L) and hue (H) were chosen as the optimal subset to describe the ISO brightness of the pulp and the model was simplified without losing much of its accuracy. Finally, we derived the numbers of membership functions for each variable to further refine the fuzzy logic-based prediction model. The error then reached 2.18%. The loss on accuracy was compensated by adjusting to the fittest membership function numbers.

Keywords:
Pulp and paper quality prediction
Adaptive neuro-fuzzy inference system
Genetic algorithm-partial least square
Fuzzy logic-based modelling
Feature selection
Model optimization

1. Introduction

In the thermo-mechanical pulp (TMP) transformation process, the main goal is to optimize pulp quality while minimizing consumption of chemical materials in terms of either hydrosulfates or peroxides and possibly the energy consumption (not considered in this paper).

During the process, wood chips are first chopped and ground into pulps under mechanical forces. Bleaching agent such as peroxide charge is then added to whiten the pulp, and the whiteness, or brightness, of the paper is viewed as an important measure to evaluate its quality. Thus, to measure the whiteness, an index referred to as the ISO brightness is used. In this paper, this index will account for pulp quality, which describes the paper's ability to reflect light.

As we wanted to improve and control the quality of paper, the amount of bleaching agent needed to achieve certain degree of brightness need to be optimized. To do so, we sought to build a prediction model to estimate pulp ISO brightness based on chip properties and the amount of bleaching chemicals added in the process, in other words, to determine the consumption of peroxide charge. Besides, as the knowledge of the pulp transformation process is very sparse and imprecise, fuzzy logic, as a modeling approach that is used to handle the concept of partial truth, seemed to us a promising approach. However, one needs to note that to create fuzzy models one first needs to implement a learning tool to learn automatically from the experiment data. This will be explained in section 1.2.

Once the learning tool was selected and implemented, we needed to obtain an optimized selection of variables most related to the variability of the ISO brightness by means of feature selection process. We then further refine the model by looking into the membership function number for each variable to derive an optimal solution. The actions taken in this research are summarized in Fig. 1.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2749v1 | CC BY 4.0 Open Access | rec: 25 Jan 2017, publ: 25 Jan 2017

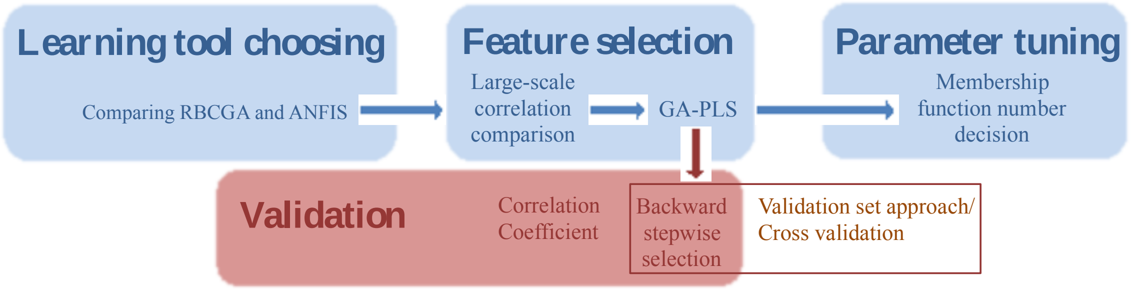
Learning tool choosing
Comparing RBCGA and ANFIS
Feature selection
Large-scale correlation comparison
GA-PLS
Parameter tuning
Membership function number decision
Validation
Correlation Coefficient
Backward stepwise selection
Validation set approach/ Cross validation

Fig. 1. Steps followed in this paper.

1.1 Variables Used to Predict Pulp Quality

To optimize pulp quality, the relationship between the end product's properties and the process inputs needs to be better understood. First off, pulp quality could be defined by a few variables, such as ISO brightness, fiber freeness or fiber length; several research works have already been carried out to predict them (Nilsson, 2005; Wimmer et al., 2002; Oluwadare and Sotannde 2007).

We could also take into account some material properties of the unprocessed wood chips, such as the concept proposed by Ding et al. (2005) pertaining to how chip quality could be defined. Afterward, in 2009 the same research group demonstrated that Chip Management System (CMS) measurements of chip properties could be used to predict pulp properties (Ding et al., 2009). The authors combined Projections to Latent Structures model and principal component analysis and finely interpreted some chip properties and their influence on pulp quality. They focused on explaining why these chip properties were regarded as critical parameters in determining pulp quality, including chip freshness, wood species, chip size distribution and bark percentage. The CMS measurement provided sufficient information to predict certain pulp properties, but when more than two species were present in the mix, the sensors failed to evaluate the pulp quality adequately. Hence, they could only predict chip brightness when wood species were precisely known, which however is rarely the case.

In the TMP process, wood chip quality and TMP refining process both are critical in affecting the final TMP properties (Li et al., 2011). On one side, physical and chemical properties of wood chips have a great effect on the refining process and the end products quality. The wood species and nature of fibers (Wood, 1996, 2000; Jackson, 1998) are directly linked to the primary characteristics of pulp quality. Other considerations also include the wood chip size (Lanouette, 2004), and the chip quality degradation (Koran and Nombi, 1994; Ding et al., 2003) has an influence on the bleaching process as well. On the other hand, during the refining process, energy consumption (Rusu et al., 2011) and refining intensity (Muhic et al., 2010) contribute to the structural strength of fibers and even the printing quality on the end product.

In brief, to previse the paper quality, we could rely on quite a few variables, while a comprehensive understanding of the relationship between these variables and the output is still lacking. To optimize both the prediction and then the process, in this paper, we will explore the use of machine learning algorithms that allow us to extract relevant information from the data for quality prediction purposes.

1.2 Methods Used to Predict Pulp Quality

Studies on pulp and paper quality prediction and optimization have focused on methods to understand the complex relationship between a wide selection of potential predicting variables and the desired pulp properties and on the convoluted reasoning that is required to elaborate the multivariate effect of incorporating sundry species of wood.

Wooten et al. (2011) developed an algorithm that was built on object identification, image capture, single value decomposition and logistic models to separate bark from wood chips. In this regard, wood chips could be individualized and the chips analyzed, which served to prevent the increase in ash content and optimize the pulp transformation and energy conversion processes. This way, utilization of wood chips could be improved by means of preceding classifications.

Another method was proposed by Thomas et al. (2011) to estimate wood chip brightness, and the TMP pulp ISO brightness could then be linearly correlated with the ISO brightness of the wood chips; the result came with an R2 value of 0.885. They took into consideration the monthly variations in chip and TMP brightness, which helped to solve the problem of higher dark percentages in chips and thus heavier consumption of bleaching agent during summertime. However, we should be aware that monthly change in climate has an influence on wood chips, whereas taking a direct record of chip properties provides more information on the mechanism that dominates the characterization of pulp quality. This encouraged us to look deeper into the pulp transformation process to obtain a clearer picture of the relationship between the wood chips and the pulp that will ultimately help us to more efficiently preprocess the raw materials and allocate resources.

Since very little human knowledge or experiences exist on the pulp and paper transformation dynamics, we aim at developing fuzzy logic based models in order to predict paper quality. We will take advantage of the capacity of fuzzy logic to model imprecise information as well as propose a fuzzy rule base that can be explicitly examined by human operators in order to increase the process knowledge for the users. More precisely, in this paper, we use Adaptive Neuro-Fuzzy Inference System (ANFIS) to generate the knowledge base. ANFIS modelling performance was already successfully evaluated and compared thoroughly (Jang, 1993; Schlechtingen et al., 2014).

However, one cannot claim an algorithm has an overall better performance over another while failing to specify the particular problematic setting where it is applied (Rao et al., 1995; Wolpert, 1994). For this reason, in this paper we would, in the first place, draw a comparison between the performance of a rule-based genetic algorithm (RBCGA) developed by the authors and that of the ANFIS with regard to their ability to minimize the error of the fuzzy models exploiting the tolerance of imprecision of the information provided by the selected pulp and paper variables. In 2005, a genetically generated fuzzy knowledge base was used to model the relationship between chips characteristics and pulp quality (Achiche et al., 2005). Specifically, average H, average S, average L (average HSL) and peroxide charge were chosen as input variables for the inference system and ISO brightness as output variable. The same variables are used in this paper so as to compare the performance of the above-mentioned learning tools.

1.3 Data Collection

In this paper, we aim at reducing a set of 79 potential input variables to the minimum required for pulp quality prediction represented by the ISO brightness as output. This set of data was obtained from an experiment carried out in Quebec in collaboration with Centre de Recherche Industrielle du Québec (CRIQ) and the University of Québec at Trois-Rivières (Achiche et al., 2006).

1.3.1 Wood Handling.

The experiment consisted of two different phases. First, a mixture of four representative species were used, including black spruce, balsam fir, jack pine and white birch. The trees were selected, cut, barked and chipped to obtain standard chips with known and controlled age. Then the samples were collected according to the arrangement described in Table 1. Second, experiments were carried out to investigate the effects of other variables on pulp quality such as species, density, initial dryness, and chip thickness. But this phase is not considered in this paper because the values could not be comprehensively measured and concluded by the chip management system (CMS) and they thereby represented only pure wood species (Achiche, 2006).

1.3.2 Pulping Transformation Processes.

Refining was conducted on the pilot unit Metso CD-300 at Centre Intégré en Pâtes et Papier of University of Québec at Trois-Rivières (CIPP of UQTR). Each sample was washed and refined first at 128 °C and then under atmospheric condition.

Bleaching using peroxide was applied to pulps with freeness ranging from 200 to 150 mL. Different concentrations of peroxide were used as bleaching chemicals: 0%, 1%, 2%, 3% and 5%.

1.3.3 Standard of Measurement.

Measure of ISO brightness and color coordinates was taken in conformity with the Pulp and Paper Technical Association of Canada (PAPTAC) standard.

1.3.4 The Chip Management System.

The physical properties of the wood chips were characterized through (1) classic laboratory measurements and (2) artificial vision measurements using chip management system (CMS).

The CMS allows online measurements of chip characteristics. Its main sensors consist of an artificial vision sensor, which combined RGB camera and a frame grabber, and a near-infrared sensor to measure properties such as chip brightness and moisture content. The geometry of RGB was then rearranged to provide information on wood chip's color representation Hue (H), Saturation (S) and Luminance (L).

As for chip luminance, the brightness of black is defined as zero whereas that of white is set to 150. The brightness measurements of wood chips were therefore in between.

Moreover, several auxiliary sensors were used on the CMS, providing information on the exterior and interior temperature, humidity, etc.

This helps us to understand the relationship between the wood chip properties and the pulp quality, and to improve the TMP transformation process in terms of cost of bleaching agent consumption and energy consumption in the long run.

1.4 Model Optimization

As discussed in the previous sections, there have been numerous sources of methods and models that draw on different combinations of variables and some of them have decent result of accuracy. In this paper, we take on a comprehensive generalization process to approximate the desired value (ISO brightness) at future observations of input variables, but this process often encounters a tradeoff between the bias and variance contributions to the estimation error, which depends on the level of flexibility of the model (Geman et al., 1992).

Table 1
Wood species handling (in percentage).

Test no.	Spruce	Balsam fir	Jack pine	Birch
1	0	20	40	40
2	100	0	0	0
3	0	100	0	0
4	60	0	0	40
5	0	60	40	0
6	60	0	40	0
7	0	60	0	40
8	20	0	40	40
The following are repetitions of tests 2 and 3 for experimental error determination.				
9	100	0	0	0
10	0	100	0	0
The following are additional tests.				
11	0	0	100	0
12	0	0	0	100

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2749v1 | CC BY 4.0 Open Access | rec: 25 Jan 2017, publ: 25 Jan 2017

To deal with the bias/variance dilemma, we consider the use of two different aspects (which will be explained further in section 1.4.1): Feature Selection and Membership Function Number Calibration.

1.4.1 Feature Selection

The selection of material and process variables used to determine the bleaching agent amount and other variables for the prediction of the pulp quality is a crucial step towards the optimization of the TMP process.

On one hand, more variables (if all relevant) provide more information to our fuzzy logic-based inference system, thus reducing the bias of the model as the additional fuzzy rules help to approach the complexity of our real-world problem. However, when more variables are selected as predictors, as they increase the dimensionality of the model and thus require more samples to lower the variance, the model created tends to produce larger error owing to the relatively greater variability, which is responsible for the overfitting of the model and thus worsened prediction performance. Our challenge lies in determining the fittest subset of variables so as to maximize the exploitation of our learning tool's potential while minimizing the generalization error of predictions based on observations beyond the learned dataset (Schaffer, 1994). The best subset of variables here accounts for the flexibility of the model corresponding to the best bias-variance trade-off.

In this paper, we use the genetic algorithm-partial least square (GAPLS) proposed by Leardi and Gonzalez (1998, 2000). It allows us to study each variable's univariate effect on the output variable ISO brightness.

As illustrated in the second block of Fig. 1, we started with a cross-comparison of correlation between any pair of the 79 potential variables, which yielded a number of 3081 combinations. Most of the variables were found to follow the same trends, whereof 76% were either highly positively correlated or practically symmetric, with a correlation coefficient higher than 0.9. As a second variable that is highly correlated with the first one barely provides more information, we deleted most of them and had the following five core variables left: average H, S and L, global moisture and peroxide percentage. These specific five variables were retained for further analysis in that they either can be explicitly measured or are manipulated variables of the experiment, like peroxide consumption.

After a set of core variables was obtained, we used the GAPLS toolbox to analyze their relevancy to ISO brightness. This step is then followed by two other methods to reconfirm GAPLS's suggestion, the Backward Stepwise Selection method and the correlation coefficients. Finally, we created a predictive model using ANFIS. The optimal dataset of selected variables was then categorized into groups so as to validate and cross-validate the re-productivity and accuracy of the automatically generated fuzzy model.

After the model was created and optimized, it can be used to predict the pulp quality based on ISO brightness. The use of a downscaled set of inputs that can satisfactorily describe our output variable can be easily manipulated and understood by a human operator.

1.4.2 Membership Function Numbers Calibration.

In this section we explain how to decide the number of membership functions (MF) in each premise trained by ANFIS. As stated by Bayarri, overly optimistic assessments of validity come along when the tuning process produces a good model for prediction in the range of the data, yet not for prediction outside of this range (Bayarri et al., 2007). In this case, the computer model does not adequately represent the reality.

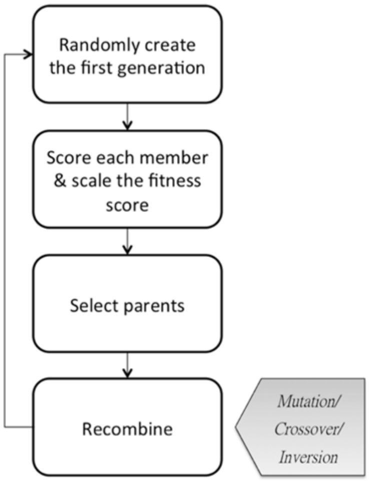
Randomly create the first generation
Score each member & scale the fitness score
Select parents
Recombine
Mutation/ Crossover/ Inversion

Fig. 2. Genetic algorithmic process.

As a last step to adequately optimize the model, we needed to calibrate the MF in terms of how many symmetric triangles (the basic MF that was used) were to be used, so that the phenomenon of over-tuning could be prevented via a further and final simplification of the model. We made a comparison among all the combinations of membership function numbers of selected variables within a designated range and obtained the best result. Shown on Table 2 is a fragment of the list of combinations of MF numbers in the premise. The designated range of MF numbers ranged from 2 to 5, which yielded 64 combinations. For each combination, we calculated the root-mean-square error and fitness level to compare how they performed.

2. Theories

2.1 Genetic Algorithm

Developed by John Holland in the 1960s, the genetic algorithm (GA) employed random exploitative search heuristic, often used to improve performance toward some optimal point or points (Goldberg, 1989). Basically, the GA process follows a simple four-step procedure shown in Fig. 2.

2.2 Partial Least Square Regression

When a partial least square regression (PLS) is used, we seek to first identify a set of linear combinations of the original predictors, the input variables that have potential relationship with the output to be discovered, and then to use this subset of inputs to fit a least square model (James et al., 2014; Hastie et al., 2009). This dimension reduction technique finds its way up toward higher variances and correlations with the predicted variable—the response—via weighing each input's strength and univariate effect.

2.3 Genetic Algorism Partial Least Square (GAPLS)

It is an efficient tool used for extracting important information from a set of variables (Leardi and Gonzalez, 1998; Leardi, 2000).

At the beginning of the optimization process, a pool of individuals, with respect to the core inputs variables we seek to analyze—average HSL, global moisture (M) and peroxide concentration (P)—are created as the first generation. Each individual, or chromosome, has as many genes as the variables we are analyzing. Each gene is set to be either 1 or 0. Think of it as an on/off switch that bridges or breaks an electrical circuit. When the switch is on, the input variable it represents is selected to check out its performance in estimating the output variable.

The toolbox can be modified to record the result of each run, yielding a probabilistic distribution that shows how each variable contributes to the process, or say, how much information we can draw out from it.

To extract information from the variables cited above, we have chosen the following parameters: auto-scaling for cross-validated explained variance; 5 deletion groups; 10 chromosomes; averagely 3 variables are present in each chromosome of the first generation; 30 variables as maximum; 1% mutation probability; 50% crossover probability. These parameters are decided in order to differentiate the variables as to how well they perform.

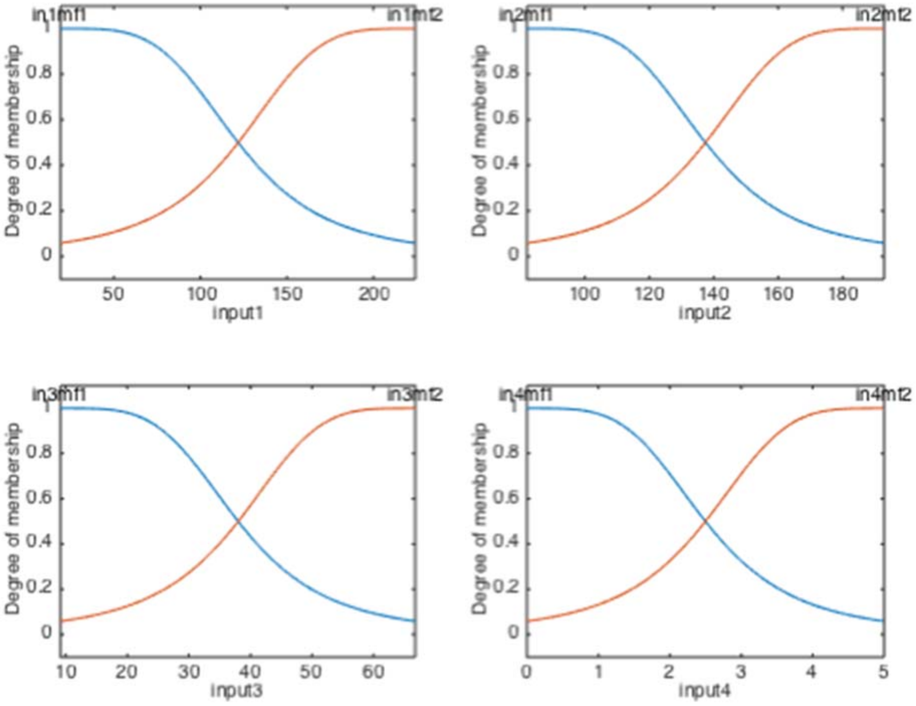

Fig. 3. Input membership functions.

Table 2
Partial list of MF analysis.

	MF number of		RMS	Fitness	
H	L	P	Error	Level	Runs
2	2	2	2.44	91.89	29
2	2	3	2.43	91.93	7
2	2	4	2.42	91.97	6
2	2	5	2.42	91.96	4
2	3	2	2.56	91.52	8
2	3	3	2.45	91.87	26

2.4 Correlation Coefficient

As will be explained in the following sections, a correlation coefficient is adopted so as to measure the linear association between the information provided by two variables. The Pearson product-moment correlation coefficient is used.

2.5 Adaptive Neuro-Fuzzy Inference System

It is designed to apply neural network algorithm to data mining methods for parameter tuning. It provides a method for the fuzzy system to learn from the training data, and helps manage a trade-off between precision and significance by mapping input space description to output space discourse for the fuzzy inference system. In this paper, the MF of the fuzzy inference system are adjusted, or decided, via a combination of back propagation algorithm and least squares method.

2.6 Performance Criterion

As we desire to evaluate the effect of the feature selection process, we need to have a statistical measure of error and a normalized performance criterion against which the testing result of the model can be benchmarked. In order to calculate the accuracy levels of the results of the dimension reduction process, we compare the outputs reproduced by the ANFIS model and that of the original experimental data. Hence the Root Mean Square error (RMS) of the prediction is defined:

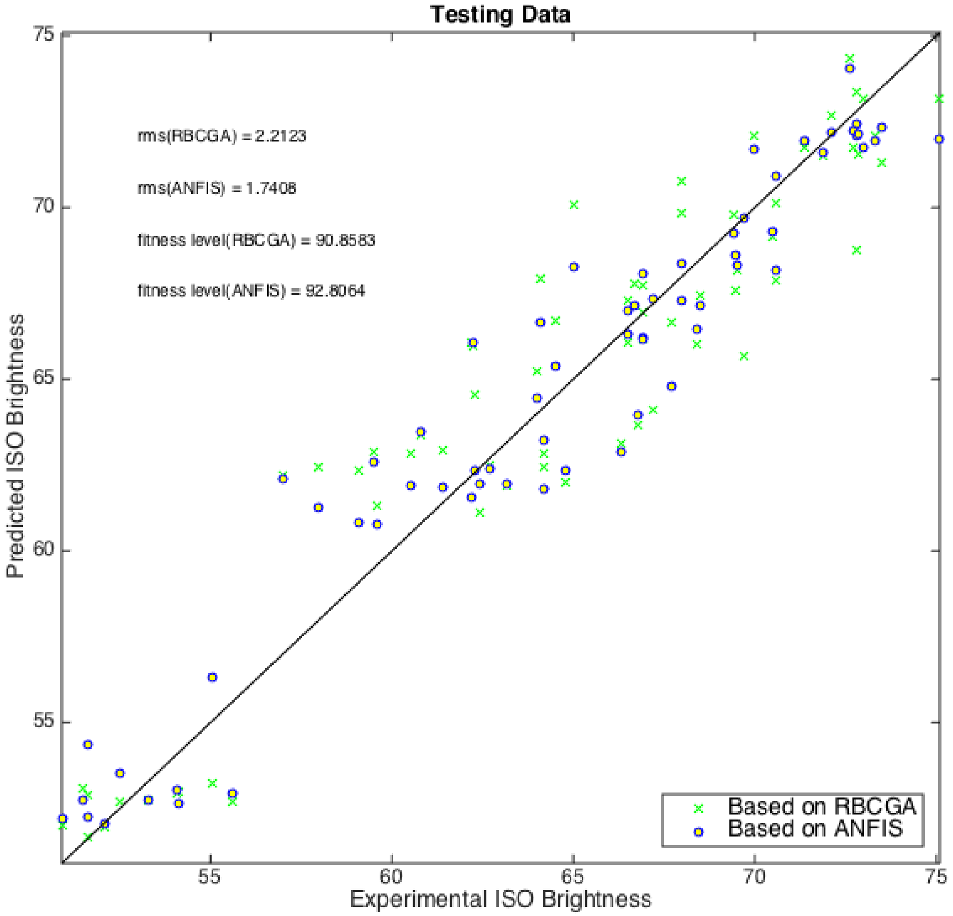

Fig. 4. Model comparison on testing file.

$$\text{RMS} = \left(\Sigma_{i=1}^{N} (\text{ANFIS}_{\text{output,i}} - \text{data}_{\text{output,i}})^2 / \text{N}\right)^{1/2}$$
(1)

where N represents the total number of training samples. To compare the result on the same scale, we have the fitness level (f) defined as a dimensionless function that is a percentage of the full range of the output:

$$f = (\text{L-RMS}) \times 100\% / L \qquad (2)$$

where L is the scale length of output, defined by the difference between the maximal and the minimal values of the conclusion.

3. Algorithm selection

We created ANFIS models to depict the relationships between the chip properties and pulp quality, to see if the new model can predict the pulp quality more precisely than the RBCGA models developed in (Achiche et al., 2005), which is used as a comparison benchmark.

3.1 Generation of Fuzzy Inference System

The input variables are chosen to be average HSL and peroxide charge, as in the case of RBCGA (Achiche et al., 2005). The predicted output variable is the ISO brightness of pulp. The original dataset was randomly divided into two parts: 90% were classified as learning file whereas 10% as testing file.

The basic setting to generate fuzzy inference system (FIS) includes numbers and types of MF. As shown in Fig. 3, we chose the same parameters for ANFIS as for RBCGA:

- Two triangles for HSL
- Three triangles for peroxide charge
- Linear type defuzzification for output (since it introduced less error)

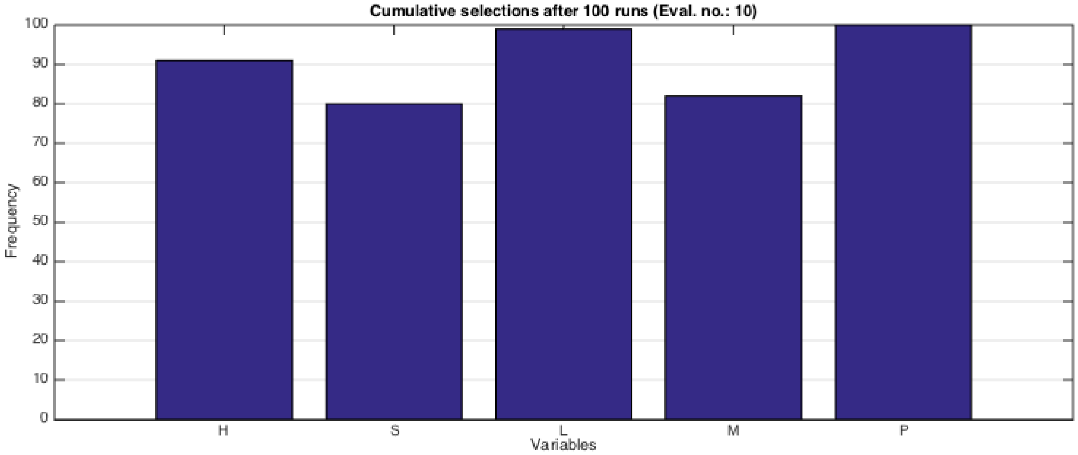

Fig. 5. cumulative selections after 100 runs (evaluate no: 10).

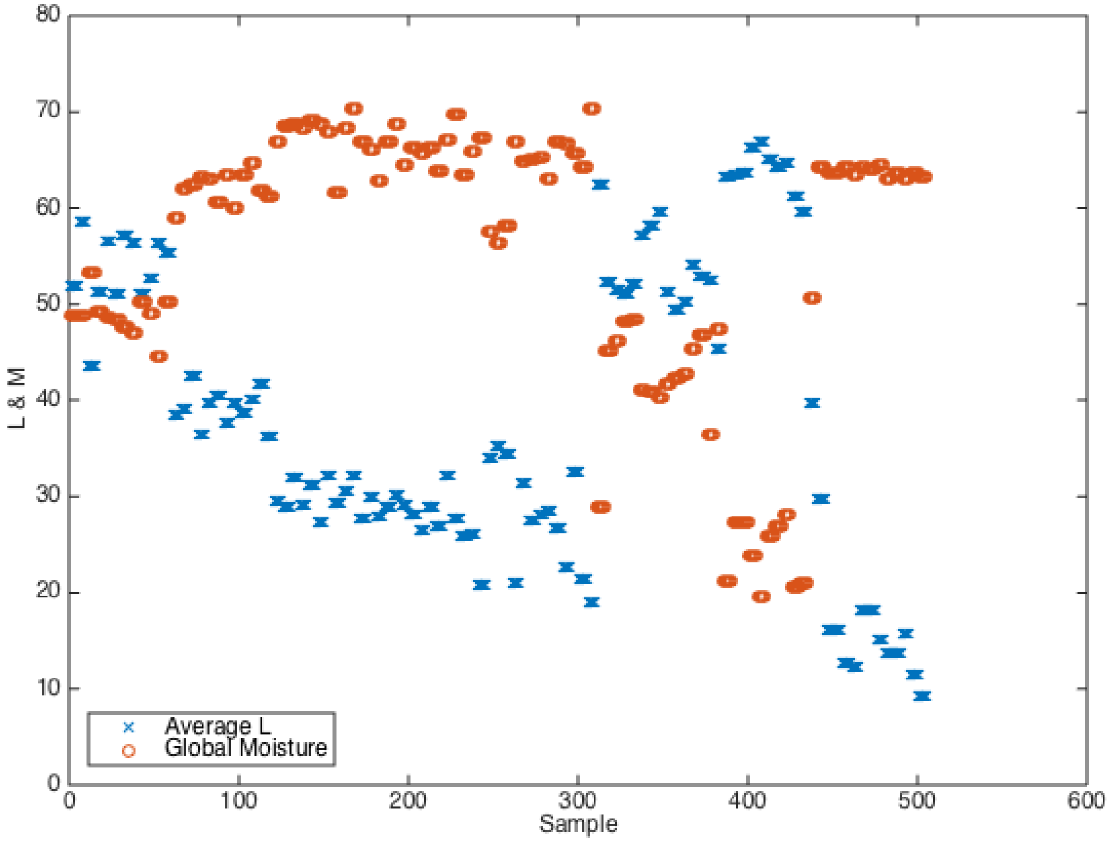

Fig. 6. Comparison between average luminance and global moisture.

Table 3
GA Selection Frequency for HSLMP.

Avg. H	Avg. S	Avg. L	Moisture	Peroxide Conc.
91	80	99	82	100

To train the FIS, we opted for the hybrid mode of optimization method. This adjustment allows the fuzzy systems to learn from the data they are modeling.

3.2 Model Performance Comparison

In this section, we already derived a model based on ANFIS and we sought to compare the result with that of RBCGA.

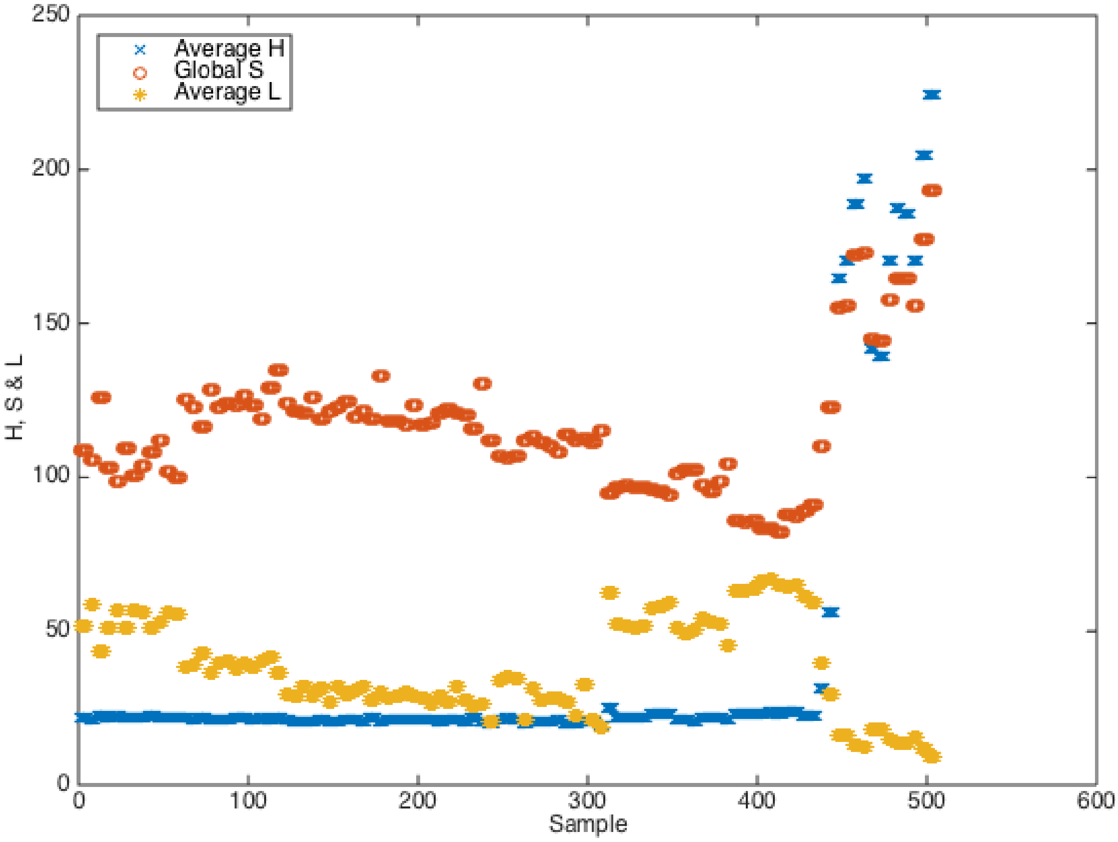

Fig. 7. Comparison between average H and S and L.

Table 4
Backward stepwise selection.

Var. selected	HSLMP	SLMP	HLMP	HSMP	HSLP	HSLM
Var. taken out	-	H	S	L	M	P
No of Var.	5	4	4	4	4	4
No of MF	2-2-2-2-3	2-2-2-3	2-2-2-3	2-2-2-3	2-2-2-3	2-2-2-2
RMS_learn	0.94165	1.5337	1.828	1.6076	1.5796	7.0004
f_learn	97.4138	95.7878	94.9793	95.5847	95.6616	80.7734
Loss on f	-	1.626	2.4345	1.8291	1.7522	16.6404
RMS_test	2.4566	2.4467	2.1771	2.4327	2.2102	8.6919
f_test	91.8521	91.8851	92.779	91.9313	92.6694	71.171
Loss on f	-	-0.033	-0.9269	-0.0792	-0.8173	20.6811

Table 5
Correlation coefficients between average H, S and L.

Variables	H & S	H & L	S & L
Correlation	0.81032067	-0.5601011	-0.8368392

The GA used in this paper to compare with ANFIS is a real/binary-like coded genetic algorithm developed by Achiche et al. (Baron et al., 2001).

For ANFIS, its RMS error of learning file is 1.64, which is lower than 2.11 as produced by the RBCGA model. This implies the ANFIS model fits the data better than the RBCGA.

In order to validate ANFIS's performance, we used the testing file and it yielded a RMS error of 2.21. Looking at Fig. 4, we found that the predicted values of ISO brightness based on ANFIS were generally closer to the experimental values, which were drawn as the diagonal line. The fitness level of ANFIS model, 92.81%, was higher than that of RBCGA model, 90.86% (Achiche et al., 2005).

As a result, ANFIS in this case has achieved a higher level of performance than RBCGA. So in the remaining part of this paper we will use ANFIS as the modeling tool.

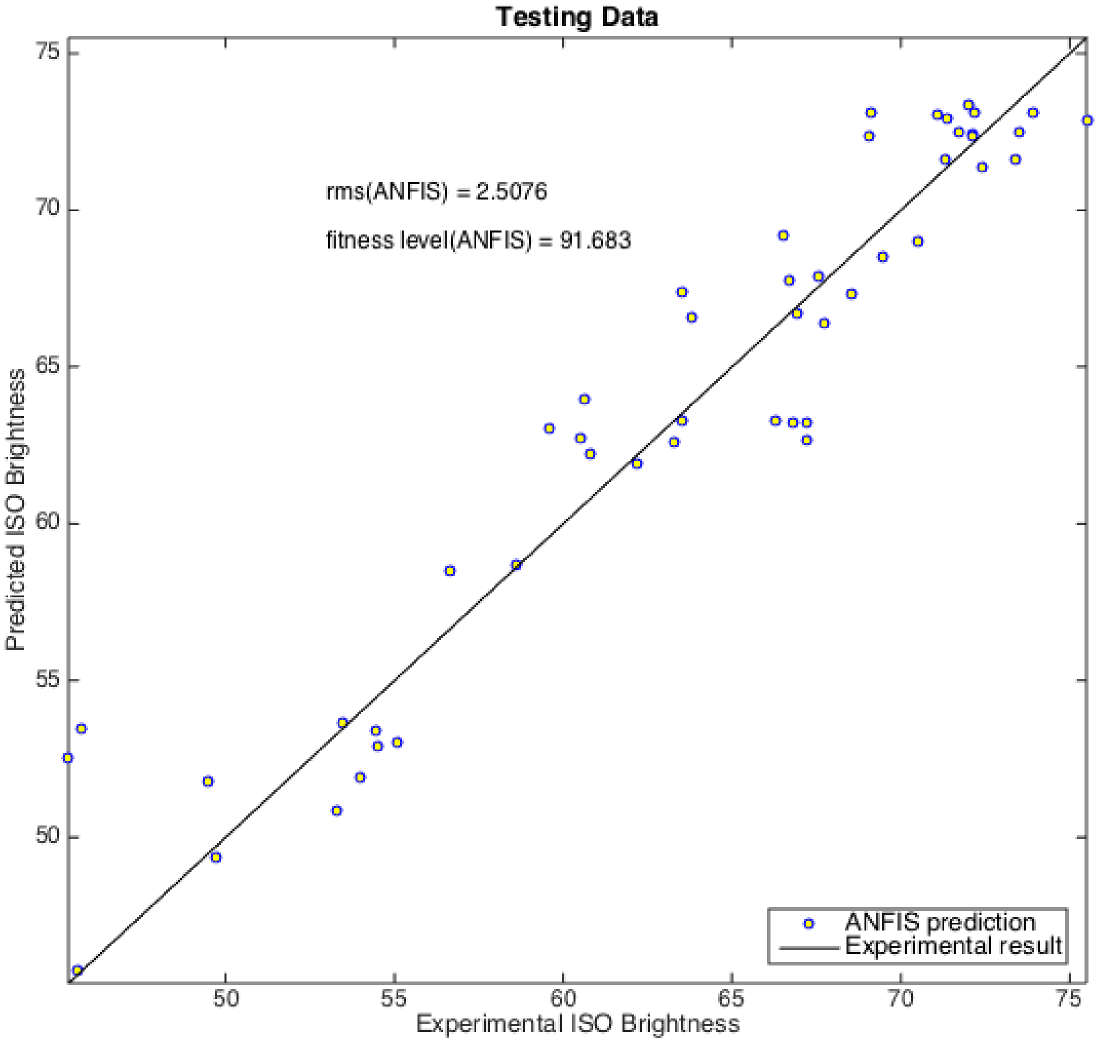

Fig. 8. Anfis prediction based on average luminance and peroxide charge.

Table 6
Best five mf numbers.

	MF number of		RMS	Fitness	
H	L	P	Error	Level	Runs
2	4	2	2.18	92.76	39
2	5	2	2.24	92.57	14
3	5	2	2.35	92.21	11
4	2	3	2.36	92.19	7
4	3	3	2.31	92.33	22

4. Feature selection of pulp and paper variables

Using GAPLS, we can find the combination of variables that best describes the desired output of the inference system.

Since most of the 79 variables were found highly correlated, we now apply GAPLS to the core variables that included average HSL, global moisture and peroxide concentration.

Figure 5 and Table 3 present the times each tested variable were selected; its frequency level got higher by one point every time the variable was chosen every one out of the 100 runs. By this it indicates a suggestion based on the relative importance of each variable; the cumulative count implies how well each variable did its job in describing the output as they were compared to each other. And we found the peroxide concentration was selected in each run and the average L was the second most frequently selected variable, followed by average H, global moisture, and average S.

The peroxide usage in the process was found to be most related to the ISO brightness of pulp. This came as no surprise given that peroxide has a major effect on pulp quality in terms of ISO brightness, as it is the bleaching agent.

This might also mean that the standard of using a 5% concentration might be too high as it dominates other properties of the wood chips, and that its usage amount can be optimally decided.

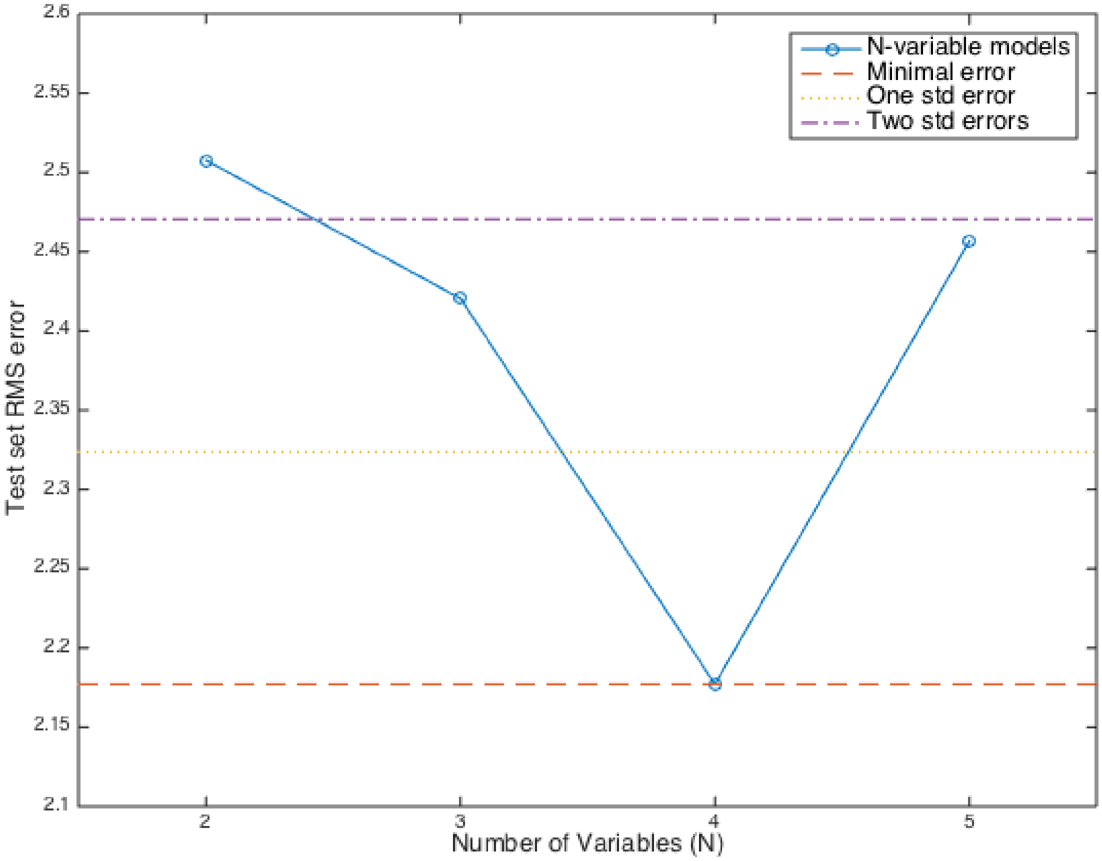

Fig. 9. Best models derived from backward stepwise selection algorithm.

On the other hand, the relatively low frequencies of average S and global moisture would be of high interest to be discussed later in the following part of this paper, pertaining to whether they are important factors to be included in a predictive model.

5. Verifying GAPLS suggestion

As illustrated in section 4, GAPLS helped us to sort through the core variables. Here we make use of three methods to validate the suggestion.

5.1 Backward Stepwise Selection

In this section we adopt a simple method–the Backward Stepwise Selection–as a crosscheck on our result of running GAPLS (James et al., 2014).

The Backward Stepwise Selection is an intuitive inspection on the importance of each variable that requires stripping the prediction model of one input at a time to compare its loss on accuracy. The more error is introduced, the more information is lost and of higher importance the variable being taken away might be. The first row of Table 4 shows the combinations of variables. Then the number of membership functions of each predictor is chosen for simplicity and in conformity to the author's experiences (Achiche et al., 2005).

Here we have 5 components in the main set, thus 5 combinations of subsets. For each combination, we carried out the RMS error and fitness level of both learning (training) and testing data, as in the model selection process discussed in section 3.1.

The first column of Table 4 indicates the over-fitting of the predictive model, which adopts all five variables, since the fitness value of the learning file, 97.4%, is a lot greater than that of the testing file, 91.9%.

The results of the five subsets show two things. Firstly, peroxide charge is an indispensable element of the prediction process; taking it away leads to a loss of 20.7% on the fitness level.

Secondly, it is shown that as long as peroxide is included, a 4-variable combination would have higher fitness level than the 5-variable one. While the error increases for the learning file and decreases for the testing file, the simplicity and accuracy of the model is confirmed.

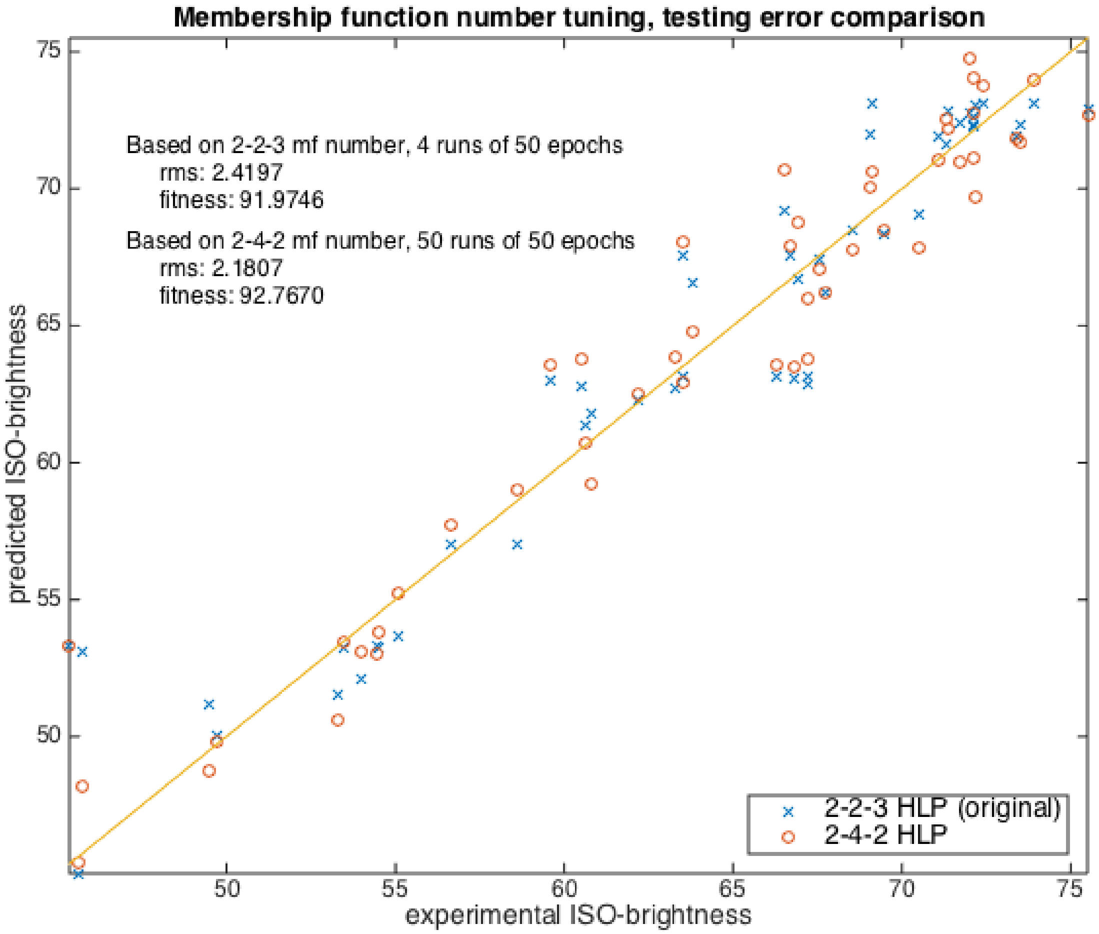

Fig. 10. Optimized mf numbers comparison on testing file.

As we look further into the numbers, it is found keeping the average H and L would ensure higher re-productivity, while the elimination of any of them results in a greater drop in the fitness level.

5.2 Correlation Comparison and Validations

In order to verify once more what is suggested by backward stepwise selection and GAPLS, we took a look at the correlation among the variables, and carried out the cross-validated root-mean-square (CV RMS) error of the HSL model.

For example, the correlation coefficient between average L and global moisture is -0.8645, implying negative linearity (symmetry) of the two variables, as shown in Fig. 6. Now turn to Fig. 7 and Table 5 to check up on average HSL: average H and L are poorly correlated, so we would lose a lot of information discarding any of them; average S is much more highly correlated with any of the other two, sharing a greater pool of information with them, so we would not sacrifice too much of the accuracy of the model if we had to trim it off the inference system.

We also carried out a simple validation and a 5-fold cross-validation on the HLP model, since it was suggested that average S and global moisture could be taken away to simplify the model.

As we took out one tenth of the data to be the testing file, the fitness level of this HLP model reached 91.97% and the RMS error was 2.42. Then the CV RMS error was calculated, yielding only 2.24. This confirmed that the result of deleting variables was satisfactory, since the prediction accuracy was maintained even when the model was further simplified, and this is believed to provide an unbiased estimate of the expected prediction error over training sets, as confirmed by the CV (Duchesne and Rémillar, 2005).

5.3 Fuzzy Model Based on L and P

Finally, we would like to examine the two most relevant variables, average L and peroxide charge (LP model), and their performance on predicting the ISO brightness. As shown in Fig. 8, most of the data points are found close to the experiment datum line. The fitness level is 91.68.

As we drew a comparison among all the best models suggested by backward stepwise selection process with decreasing numbers of variables, as in Fig. 9, we found the best 4-variable model (HLMP) makes the predictions that bear the least error. Then we calculated the standard error of the test RMS error for each of the suggested best models, finding that both the best 3-variable model’s (HLP) and the 5-variable model’s test errors fall within the 2-standard-error range. This approach led us to favoring the best 3-variable model over the adoption of all five variables. The LP model, however, dropped down to a lower level of accuracy because the variable average H was removed while it provided a great bulk of information.

The stepwise selection process helps us to elaborate on the relative relevancy of each chosen variable, which can be further exploited to study the trade-off between simplicity and prediction accuracy of the automatically generated models. Yet one should note that backward stepwise selection could not guarantee to find the best model out of all the combinations of variables, in that certain variables it rejects may provide more information when combined with a reduced set of variables. In this case, this algorithm fails to select the best model for a desired number of variables.

6. Selection of input Membership function (mf) numbers

Membership functions serve as the interface between the human and machine language; a curve that defines how input space is mapped to the degree of membership. To complete the optimization of ANFIS model, we aim at calibrating the MF in terms of numbers of MF used to describe the input variables.

Intuitively, the number of MF for a certain input should be as many as possible to have a refined covering of the universe of discourse. However, the more MF we have, the more complicated the model would be, in view of increased fuzzy rules. Besides, we need to take the variability of the model into consideration, which is inclined to increase as the model’s flexibility is improved as we allow more MF to be used. Thus the number of MF for each variable is an important parameter to be tuned.

We selected average H, L and peroxide charge P as the input variables to predict the ISO brightness. Then, we examined the predictive performance of each model whose input MF numbers for each variable ranged from 2 to 5 as maximum.

We then investigated the data using ANFIS. In each run, the epoch number was set to be 50; each run stopped when it reached the designated epoch number, which is defined as the number of training iterations for the ANFIS model before function evaluation and weight adjustments. Based on the author's experience, the stop criterion of the training process was as follows: if the difference between the maximal and minimal RMS errors falls below 0.003 and the difference between the first and last RMS errors is lower than 0.0002, or the standard deviation of RMS errors in any run is no more than 0.00001, the process stops. The only exception when the process stops while not meeting with the criteria above is when the process runs over 40 times.

As a result, we came by 64 combinations of MF numbers, out of which the best 5 combinations are listed in Table 6. According to the table, choosing 2, 4 and 2 MFs for variable H, L and P yielded the best result.

We then further explored its performance by extending the limitations on maximum runs to 100 runs, and found out that the process should stop after 50 runs to yield the least RMS error. The result was then compared to the original arrangement for MF numbers: 2 for H, 2 for L and 3 for P. As shown in Fig. 10, the fitness level of ANFIS model is raised by almost 0.8%. There were originally 12 fuzzy rules, which are now increased to 16. This means our model's predictive performance is improved while its simplicity is somewhat compromised.

7. Conclusion

In this paper, we first compared the performance of RBCGA and ANFIS, and then decided the latter to be the learning algorithm to be used for developing pulp quality prediction models.

We then combined methods including correlation comparison, GAPLS, backward stepwise selection to select the best variables to predict ISO brightness.

As implied by the results, peroxide charge is the most relevant variable to predict ISO brightness, followed by average L and H, which refine the model by raising the accuracy and re-productivity slightly at the cost of the simplicity of the fuzzy inference system. By reducing the dimension, we optimize the model and prevent complexity by picking up the more relevant variables to predict the pulp quality.

At the end of our paper, we also decided the best MF numbers via a few comparisons between the appointed combinations. By doing so, we sacrificed the simplicity to a slight extent, as the new model required more fuzzy rules than before.

This facilitates our inference's mechanism to describe the ISO brightness, and increases the efficiency in determining the bleaching agent amount. Through an examination on these variables' relevancy and by deciding the right numbers of MF, we can eventually achieve possible cost management and environmental impact reduction, as bleaching chemical could be accurately decided and cost of waste be reduced.

Besides, as we could decide which variables to include in the predictive model, other process variables and chip properties are deemed as less important and irrelevant, by which it indicates that a simpler measurement could be integrated into the pulping control operation that takes account of only average H and L. That is, a simple RGB camera should suffice for the ISO brightness prediction and the predictive performance is evidently satisfactory in terms of precision and simplicity.

9. References

Achiche, S., Baron, L., Balazinski, M., Benaoudia M., 2005. Chips Image Processing to Predict Pulp Brightness Using Fuzzy Logic Technique. PAPTAC, 173-176.

Achiche, S., Baron, L., Balazinski, M., Benaoudia, M., 2006. Online Prediction of Pulp Brightness Using Fuzzy Logic Models. Engineering Applications of Aritificial Intelligence 20, 25-36.

Baron, L., Achiche, S., Balazinski, M., 2001. Fuzzy Decision Support System Knowledge Base Generation Using a Genetic Algorithm. International Journal of Approximate Reasoning 28.2-3, 125-148.

Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.-H., Tu, J., 2007. A Framework For Validation Of Computer Models. Technometrics 49.2, 138-154.

Ding, F., Benaoudia, M., Bédard, P., Lanouette, R., Lejeune, C., Gagne, P., 2003. Wood chip physical quality definition and measurement, Internation Mechanical Pulping Conference, pp. 367-373.

Ding, F., Benaoudia, M., Bédard, P., Lanouette, R., Lejeune, C., Gagné, P., 2005. Wood Chip Physical Quality Definition and Measurement. Pulp and Paper Canada.

Ding, F., Benaoudia, M., Bédard P. Lanouette, R., 2009. Effects of Some Wood Chip Properties on Pulp Quality. Pulp and Paper Canada.

Duchesne, P.a.R., B., 2005. Statistical Modeling and Analysis for Complex Data Problems. Springer Science + Business Media, NY, USA.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural Networks And The Bias/Variance Dilemma. Neural Computation 4.1, 1-58.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley Longman.

Hastie, T., Tibshirani, R., Friedman, J, 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2 ed. Springer.

Hu, T., Zhao, M., Bicho, P., Losier, P., 2011. A method for estimating wood chip brightness and its applications. Canadian Journal of Forest Research 41(11), 2114-2119.

Jackson, M., 1998. The interaction of wood species and wood quality with the TMP process – A review. TAPPI Pulping Conference Proceedings, Montreal.

James, G., D. Witten, T. Hastie, R. Tibshirani, 2014. An introduction to Statistical Learning with Applications in R. Springer.

Jang, J., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Xplore: Systems, Man, and Cybernetics 23(3).

Koran, Z., Nombi, K., 1994. PTM pulping characteristics of budworm killed trees. WPDD Fiber Science 26(4), 489-495.

Lanouette, R., Bedard, P., Benaoudia, M., 2004. Effect of woodchips characteristics on the pulp and paper properties by the use of PLS analysis, 90th Annual Meeting - Pulp and Paper Technical Association of Canada (PAPTAC), Montreal, pp. 39-43.

Leardi, R., González, A, 1998. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemometrics and Intelligent Laboratory Systems 41(2), 195-207.

Leardi, R., 2000. Application of genetic algorithm–PLS for feature selection in spectral data sets. Journal of Chemometrics 14(5-6), 643-655.

Li, B., Li, H., Zha, Q., Bandekar, R., Alsaggaf, A., Ni, Y., 2011. Review: Effects of wood quality and refining process on TMP pulp and paper quality. BioResources 6(3), 3569-3584.

Muhić, D., Huhtanen, J.-P., Sundström, L., Sandberg, C., Ullmar, M., Petteri, V., Engstrand, P., 2010. Energy efficiency in double disc refining – Influence of intensity by segment design, 7th International Seminar on Fundamental Mechanical Pulp Research, Nanjing, China, pp. 109-117.

Nilsson, D., 2005. Prediction of Wood Species and Pulp Brightness from Roundwood Measurements, Department of Chemistry, Organic Chemistry. Umeå University

Oluwadare, A.O., Sotannde, A. , 2007. The Relationship Between Fibre Characteristics and Pulp-Sheet Properties of Leucaena leucocephala (Lam.) De Wit. Middle-East Journal of Scientific Research 2 (2), 63-68.

Rao, R.B., Gordon, D., Spears, W., 1995. For Every Generalization Action, Is There Really An Equal And Opposite Reaction? Analysis of the Conservation Law for Generalization Performance, Proceedings of the Twelfth International Conference on Machine Learning.

Rusu, M., Mörseberg, K., Gregersen, Ø., Yamakawa, A., Liukkonen, S., 2011. Relation between fibre flexibility and cross-sectional properties. BioResources 6(1), 641-655.

Schaffer, C., 1994. A Conservation Law For Generalization Performance. International Conference on Machine Learning – ICML, 259-265.

Schlechtingen, M., Achiche, S., Costa, T.-L., Raison, M., Santos, I., 2014. Using Data Mining Approaches for Force Prediction of a Dynamically Loaded Flexible Structure, ASME 2014 12th Biennial Conference on Engineering Systems Design and Analysis.

Wimmer, R., Downes, G., Evans, R., Rasmussen, G., French, J., 2002. Direct Effects of Wood Characteristics on Pulp and Handsheet Properties of Eucalyptus globulus. Holzforschung 56.3.

Wolpert, D.H., 1994. Off-Training Set Error And A Priori Distinctions Between Learning Algorithms. Santa Fe Institute, Santa Fe, NM.

Wood, J.R., 1996. Chip quality effects in mechanical pulping – A selected review, TAPPI Pulping Conference Proceedings.

Wood, J.R., 2000. Wood-induced variations in TMP quality – Their origins and control, TAPPI Pulping/Process and Product Quality Conference.

Wooten, J., S. Filip To, C. Igathinathane, L. Pordesimo, 2011. Discrimination of bark from wood chips through texture analysis by image processing. Computers and Electronics in Agriculture 79 (1).