A Novel Hybrid Ensemble Classification Approach to Candidate Well Selection for CBM Well Fracturing

Lei Xu¹,² and Fang Liu¹

¹ Neijiang Normal University, Neijiang, P.R.China
² Data Recovery Key Laboratory of Sichuan Province
E-mail: 1986_xulei@163.com

Abstract. For coalbed methane commercial development, hydraulic fracturing is the most common stimulation treatment. The selection of wells or layers is the first link. Fracturing with big yield potential and good effect can not only reduce fracturing risk but also offer more technological choices for reservoir stimulation. A method combined with qualitative and quantitative analysis is adopted to construct the index system of candidate wells selection according to coalbed reservoir properties and development characteristics. In order to study the complex uncertainty and unbalanced classification problem in the process of candidate well selection, a fusion of interval type 2 fuzzy logic system (IT2FLS) and extreme learning machine (ELM) based selective ensemble method is proposed in this paper. Illustrative examples are provided to show the validity of the proposed method.

1. Introduction
Hydraulic fracturing (HF) is one of the most important stimulation treatments for coalbed methane production. There are many factors that have influence on the effectiveness of hydraulic fracturing. In addition to the characteristics of the reservoir itself, the most important thing is the three aspects that human can change, namely to candidate well selection (CWS), fracturing design and field operation. There have been a large number of methods dealing with CWS [1-3]. Existing approaches to candidate well selection can be roughly divided into two categories: conventional and advanced. Conventional method is a kind of qualitative research methods based on geology factors including experience criterion method, screening method, reservoir dynamic analysis method and classification. For more details about conventional methods see [3]. Conventional methods have some weakness. For example, the result cannot be obtained when one or more of the key parameters were unavailable. In addition, the lack of data, data conflict, and data measurement error often exists in application process which causes the complexity and uncertainty of CWS.
Advanced method provides a technical framework to solve the problem with complex, uncertain, including artificial neural network (ANN), support vector machine (SVM), pattern recognition, grey correlation analysis, fuzzy logic (FL), etc. Those methods has been paid more attention by scholars and applied successfully including ANN and FL. However, an accurate ANN model needs a large number of labeled examples, and what's more does not account for uncertainty. FL is an extension of conditional logic which is introduced by Zadeh [4]. It also called type-1 fuzzy logic (T1FL). To some extent, it can overcome the conventional difficult. But the fuzzy membership degree is a precise value, so it can't completely handle the uncertainty.
Type-2 fuzzy logic (T2FL) is introduced by Zadeh as an extension of T1FL. Its membership function itself fuzzy so it has the ability to deal with uncertainties that caused by information loss. At present, it has become an efficient tool for dealing with uncertainty [5].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
Uncertainty refers to a kind of generalized error, including the concept of value and error, and also the error of measurable and unmeasurable. At present, it is well known that the petroleum reservoir data are inherently uncertain [6]. In terms of oil and gas, Zoveidavipoor et al [7] first introduced interval type 2 fuzzy set and system to study CWS for oil well. The author makes a contrastive study of the differences between type-1 fuzzy logic system (T1FLS) and Type-2 fuzzy logic system (T2FLS). However, directly use this method to CBM has some important weakness. On one hand, some important parameters such as permeability, skin factor, water cut, and porosity are hard to obtain. On the other hand, vertical well drilling account for the vast majority of well in China, daily average production per well is less than 1000 m$^3$. The proportion of high and low Wells is not equal to 1. As is well known, the greater the gap, the greater the imbalance. Therefore, the study of how to improve the classification performance of imbalanced data and the generalization ability of classifier has the important value and practical significance. However, the existing methods do not take into account the imbalance classification problem. Extreme learning machine (ELM), introduced by Huang [8], is a single-hidden layer feed-forward neural networks. Compared with BP Neural Network, ELM does not need to adjust the network weights and bias. It can get the only optimal solution by adjust the hidden nodes. It has advantages of fast learning speed and good generalization. However, ELM suffers from low performance when uncertainties are presumed to be rife [9].

The motivation of this paper is to extend the application of T2FLS and ELM to solver unbalanced classification of well and layer selection for CBM well fracturing. The article is organized as follow. In Section 2, we construct the production division and index system using fuzzy cluster analysis and grey correlation analysis. In Section 3, the frameworks of the hybrid integrated model are obtained. In Section 4, the accuracy of the proposed method is tested through several numerical examples.

2. Production division and index system construction

2.1. Production division
The reservoir of Hancheng area is unsaturated. Most of the production wells need to take a long water drainage stage. Therefore, to analyse the production expeditiously, classification standard is defined as average daily gas production after two years of production. In general, considering three classifications can be decomposed into three binary classifications, so in this paper we discuss binary classification only. High production gas well is greater than 1500 m$^3$/d and low production gas well is less than 1500 m$^3$/d.

2.2. Index system construction
In this paper, the main research is machine learning based candidate well selection which can be treated as a classification problem. For complex problems, a single model’s results cannot always be satisfied, so we call each model a weak predictor or weak classifier. When we have thousands of models, we can through fusion of the model to get a better effect, such as Bagging, Boosting and random forests.

The development of coalbed methane is a systematic project. Many factors would affect the production. geological analogy and statistical methods are the main methods to study the production characteristics and factors [10]. The existing research predominantly focused on the areas with a higher level of development, for example Qinshui Basin in china. Though the key factors in different areas vary greatly, yet gas content, coal thickness, permeability, fracture characteristics have the highest frequencies.

In recent years, only a few scholars have analyzed the factors affect the CBM well productivity of Hancheng area, and the key control factors are not clear [11]. However, the factors can be grouped into two main categories generally: those that cannot be controlled by human including gas content, permeability, buried depth, thickness, hydrodynamic system, etc; those can be controlled including drilling and well completion method, perforation program, fracturing design and field operation.

We all know that the quality of data collection and processing is the basis of machine learning. Compared with oil reservoir, the reason that it is difficult to get coal seam information or the obtained
information is not accurate. There are many reasons among which two aspects must be mainly mentioned here: on the one hand CBM reservoir has characteristics of heterogeneity and friability which affect the data from laboratory, on the other hand due to its low-cost development strategy many important parameters (permeability, porosity, compressibility, and relative permeability) are not available for development well[12]. Fortunately, scholars have done a lot of research to prediction these parameters based on the conventional well log response and got good effect [13].

To determine the predominant factors, we proposed a qualitative and quantitative evaluation method to determine key factors. We investigate the potential influencing factors extensively firstly, as shown in Table 1, and then exclude the factors which do not meet the principles of relative independentability and operability. For example, sand concentration is equal to the quantity total volume minus pad volume and divided by proppant volume, so sand concentration can be removed.

Fuzzy clustering can set up the description of the sample with the categories of uncertainty. Therefore, it can be used to data with poor separation. The advantage of grey correlation analysis (GCA) is that it requires only a little amount of data and workload, and can largely reduce the losses caused by the information asymmetry. In this paper, we combine fuzzy cluster analysis and grey correlation analysis methods to select appropriate inputs for CBM productivity forecasting. For more details see [12].

| Table 1. Primary input parameters and ranges |
|---------------------------------------------|
| Parameter       | Range     | Parameter       | Range     | Parameter       | Range     |
| thickness of coal, m | 2-12       | burial depth, m | 384-1005  | well diameter, cm | 18-41 |
| natural gamma, API | 10.5-47    | SP, mv         | -80-84    | acoustic, us/m   | 240-467  |
| CNL, v/v         | 22-60      | density, g/cm³ | 1-2.51    | LLD, Ωm         | 150-18343 |
| LLS, Ωm          | 56-17380   | RMSF, Ωm      | 36-13507  | proppant volume, m³ | 7-51 |
| pad volume, m³   | 70-500     | total volume, m³ | 207-1003 | injection rate, m³/min | 4.5-8.5 |
| delivery rate, m/day | -6.8-17.2 | distance from fault, m | 85-424   |

SP: Spontaneous potential; CNL: compensated neutron logging; LLD: deep lateral resistivity; LLS: shallow lateral resistivity; RMSF: microspheric focused resistivity.

![Figure 1. Secondary MFs at x = x’ for (a) T1 FS, (b) IT2 FS, and (c) GT2 FS.](image)

3. The constituent frameworks of the proposed hybrid integrated model
3.1. Type-2 fuzzy logic system
Zadeh first introduced type-2 fuzzy sets (T2FS) as an extension of ordinary fuzzy set which called type-1 fuzzy sets (T1FS). Unlike T1FS, the degrees of general type-2 fuzzy sets (GT2FS) memberships are themselves fuzzy. When the secondary MF equals to 1, then each GT2FS reduces to an IT2FS. IT2FS is the simplest kind of type-2 sets, and there are fast algorithms to compute the output. Therefore, it has practical value.

For a T1FS (Figure 1a), the primary membership at \( x = x' \) has only one value, so there is no uncertainty. For an IT2FS (Figure 1b), the primary membership at \( x = x' \) has values within the interval \([a, b]\), so it encompasses a large amount of uncertainty. For a GT2FS (Figure 1c), each point in the interval have a different secondary membership, so it has a large uncertainty than IT2FS.

As shown in Figure 2, a type-2 fuzzy logic system (T2FLS) contains five components: fuzzifier, rules, inference engine, type-reducer, and defuzzifier. For more details see [14].

![Figure 2. Type-2 fuzzy logic system.](image)

3.2. Extreme learning machine
In general, ELM can be expressed as:

\[
t_j = \sum_{i=1}^{N_h} \beta_i g(w_i x_j + b_i), \quad j = 1, \cdots, N_s
\]

(1)

If we have \( N_s \) samples \((x_i, t_i)\), \( N_h \) hidden node, where \( x_i = [x_{i1}, x_{i2}, \cdots, x_{in}] \in \mathbb{R}^n \), \( t_i = [t_{i1}, t_{i2}, \cdots, t_{im}] \in \mathbb{R}^m \), transfer function \( g(x) \). Or vector form, as: \( H\beta = T \), where \( H=\{h_i\} \) (i=1, \cdots, \( N_s \)) and \( j=1, \cdots, \( N_h \))

\[
h_j=g(w_j x_i+b_j).
\]

Then, the training procedures for ELM can be summarized as follows.

Step 1: Initialization. Assign random values to the input weight \( w_i \) and the bias \( b_i \).

Step 2: compute the hidden layer output matrix \( H \).

Step 3: compute the output weight \( b \) as follows: \( \beta = H' T \).

where \( H' \) is moore-penrose generalized inverse of a matrix of \( H \). See [8] for further details.

Compared with BP network, the learning speed of ELM is faster, generalization performance is better. However, this algorithm has some important weakness. One of the major is its inability to model uncertainties. Therefore, it is necessary to merge unique generalization ability of ELM and the ability of IT2FLS to better model uncertainty, in order to obtain better classification performance.

3.3. Imbalanced classification
Experiments show that: the base classifier with good predict performance cannot guarantee a good performance of the combined classifier. Diversity of the base classifiers is the key of multiple classifier system with high performance and generalization ability. The concept of "selective ensemble ", first proposed by Zhou et al [15], has caused great repercussions. The results reveals that it may be better to ensemble many instead of all of the classifiers at hand. However, there is no unified standard to measure the diversity of the base classifier yet.
At present, there are already many selective ensemble methods in literatures, but due to the different perspectives on which those methods are based, it is difficult to understand them clearly. In this paper, through literature research, we choose margin distance minimization (MDSQ), developed by Martínez-Muñoz G, Suárez A [16]. The method has the characteristics of good prediction performance, fast selection speed, but bigger target integrated classifier size.

Figure 3. Conceptual integration framework of the proposed IT2FLS-ELM hybrid system.

Figure 3 shows the conceptual integration framework of the proposed IT2FLS-ELM hybrid integrated system. In the first stage, the algorithm of SMOTE-ENN is adopted to get the training sets \{T_1, T_2, ..., T_n\}. The training set of \{T_i\} is passed to the IT2FLS block where all possible forms of uncertainties are adequately handled. Undergoing the five necessary processes, the final outputs from IT2FLS are then passed to the ELM block for training and finally get the base classifier \{C_i\}. In the second stage, the MDSQ method is adopted to select M optimal base classifiers based on the validation set. Finally we use the test set to validate the optimal base classifiers, and it is need to note that the weight of the optimal base classifiers is the same in this paper.

4. Experimental results and discussion

4.1. Criteria for performance evaluation

We consider a set of data sample \(U\). \(U = P \cup N\), where \(P\) represents the positive class and \(N\) the negative class. We define \(p=|P|\), \(n=|N|\), then the imbalance ratio (IR) is defined as \(IR=n/p\). The performance of the classifier can be expressed in a confusion matrix, as in Table 2.

For any given classifier, TP is the abbreviation of true positive means a correctly classified positive instance. Similarly, a true negative (TN) means a correctly classified negative instance. In the remaining cases, a false negative (FN) means a positive instance be misclassified as negative, a false positive (FP) means a negative instance be wrongly predicted as positive. G-mean take into account both the classification of positive and negative, the greater the value the better the performance of classifiers, defined as:

\[
G - \text{mean} = \sqrt{\text{TNR} \times \text{precision}}
\]  

(2)

where

\[
\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}
\]
Precision = TP / (TP + FP)

Table 2. A two-class dataset of confusion matrix

| Actual/Predicted | positive | negative |
|------------------|----------|----------|
| positive         | TP       | FN       |
| negative         | FP       | TN       |

4.2. Case 1

In order to compare the propose algorithm with existing methods, we consider 5 data sets from different area to evaluate our proposal. These data sets originate form the UCI machine learning repository [17] and we modified multiple class data sets into two-class. In the data set of Abalone, we take the attribute of no.17 as positive and the remaining as negative. In vehicle, attribute of van as positive and the remaining as negative. In vowel, attribute of hed as positive and the remaining as negative.

The characteristics of the data set can be found in Table 3, showing the number of attributes (Attr), the number of positive instances (PInst), the number of negative instances (NInst) and the imbalance ratio (IR) for each of them. Form Table 3 we can see that there were a broad range in the number of instances and IR. The value of IR in data set Sonar close to 1 which indicate that it not belong to imbalanced data. The choice of the data set is to verify the effectiveness of the proposed algorithm in the balance data set classification problem.

Table 3. Performances of different algorithms on UCI repository (G-mean)

| Data sets | Attr | PInst | NInst | IR | C4.5 | SMOTEBoost-C4.5 | EUSBoost-C4.5 | RotEasy-C4.5 | SMOTE-SVM | Our method |
|-----------|------|-------|-------|----|------|-----------------|---------------|--------------|-----------|------------|
| Abalone   | 8    | 58    | 4119  | 71.02 | 73.1 | 0.878           | 0.933         | 0.938        | 0.710     | 0.942      |
| Sonar     | 60   | 97    | 111   | 1.14  | 72.8 | 0.782           | 0.937         | **0.941**    | 0.905     | 0.936      |
| Credit-g  | 21   | 300   | 700   | 2.33  | 68.9 | 0.876           | 0.918         | 0.923        | 0.892     | **0.932**  |
| vowel     | 14   | 90    | 900   | 10.00 | 95.3 | 0.978           | 0.972         | 0.981        | 0.948     | **0.983**  |
| vehicle   | 19   | 199   | 647   | 3.25  | 89.4 | **0.952**       | 0.942         | 0.836        | 0.902     | 0.935      |

Table 4. Input variables for CWS

| categories | parameters |
|------------|------------|
| uncontrollable | thickness of coal, burial depth, GA, LLD, AC, SP, well diameter, distance form fault |
| controllable | total volume, proppant volume, injection rate, casing pressure, delivery rate |

Data in Table 3 is divided into training set, validation set and test set according to the proportion of 6:1:1. Considering the random of division, the experimental results were the mean of 100 times repeated experiments. In order to check the effectiveness of the proposed algorithm, a comparative study with four representative algorithms C4.5 [18], SMOTEBoost [19], EUSBoost [20] and RotEasy [21] were done in this paper. The experimental results are shown in Table 3. From the table it is evident that compared with other algorithms, the proposed algorithm can obtain relatively good results in view of the different data sets.

4.3. Case 2

In order to carry out an empirical study, real-industrial data from Hancheng area were acquired. Among them, positive number is 37, negative number is 135, the proportion of positive and negative classes is 0.215/0.785, IR is 3.65. According to input selection algorithm in Section 2.2, the factors are presented in Table 4. The entire data was divided into training set and test set according to the proportion of 6:1. Experimental results were the mean of 100 times repeated experiments. All data were normalized by equation (9) before training.
\[ z_{\text{normal}} = \frac{z - z_{\min}}{z_{\max} - z_{\min}} \]  

(3)

Table 5. Performances of different algorithms evaluated on Hancheng area

| Algorithm           | Number of base classifiers | TNR  | Precision | G-mean |
|---------------------|----------------------------|------|-----------|--------|
| Our method          | 20                         | 0.873| 0.914     | 0.893  |
| ensemble method of T1FLS-ELM | 20                  | 0.868| 0.891     | 0.879  |
| ensemble method of ELM         | 20                       | 0.815| 0.893     | 0.853  |
| ensemble method of IT2FLS      | 20                       | 0.841| 0.902     | 0.871  |
| IT2FLS-ELM          | 1                         | 0.756| 0.879     | 0.815  |
| ELM                 | 1                         | 0.681| 0.872     | 0.771  |
| IT2FLS              | 1                         | 0.733| 0.863     | 0.795  |
| T1FLS               | 1                         | 0.716| 0.854     | 0.782  |
| BP                  | 1                         | 0.512| 0.836     | 0.654  |
| SVM                 | 1                         | 0.645| 0.865     | 0.747  |

In the block of IT2FLS, consider a fuzzy system with uncertain mean Gaussian membership functions, singleton fuzzifier, max-product composition, product implication, centre of sums type-reducer, and average defuzzification.

In order to validate the proposed method, ensemble method of T1FLS+ELM, ELM and IT2FLS and advanced approach of BP network and SVM were also being used to predict the classification effect. BP network, T1FLS and SVM were performed by MATLAB toolbox, with the structure of BP network (13\times 18 \times 1), T1FLS 18 number of membership function with Gaussian membership function. A comparative analysis of our study with other methods is shown in Table 5. To our knowledge, no previous studies have been done using the proposed method for CWS.

From the table, it can be concluded that, the hybrid method can make up for the lack of a single method, complement each other’s advantages, and improve the effect of the classification. It can also be easily noted that the ensemble method has a better performance than the single method, and IT2FLS-ELM performed outstandingly better in all fronts.

The ELM method performed better than BP and SVM. The essence of the ELM and SVM are problems mapped to high-dimensional space, and then do regression or classification in high dimensional space. In principle, the classification or regression effect is directly affected by the mapping mode, and the more obvious differences in the high-dimensional space the better the classification effect. For ELM, the way map to high-dimensional space is infinite, and the training speed is very fast. However, once the SVM kernel function is settled, the mapping way is uniquely determined.

The overall results therefore indicate that the proposed model is able to consistently deal with the nature of coalbed CWS data due to its ability to cater for different forms of uncertainties, generalization and imbalance.

5. Conclusions
In this paper, we have presented a new hybrid ensemble pruning algorithm, a new algorithm level solution to two-class imbalanced classification problems of coalbed candidate-well selection. The solution is based on the combination of IT2FLS and ELM, and MDSQ. The comparison with four kinds of typical algorithm over a collection of imbalanced datasets from UCI repository verified the feasibility and effectiveness of the proposed algorithm. Experiment results in Hancheng area also show that the proposed method can outperform the classical algorithm such as IT2FLS, T1FLS, ELM, BP, SVM and its ensemble methods.

For future work, we will extend the method to handle multi-class problems and use the method to other reservoir engineering problems and look into its usefulness in other related fields.

6. References
[1] Barak, S., Mehrgini, B., Maghsoudlou, H., & Branch, Q 2014 Multi-Criteria Decision Making Approach To Candidate Well Selection For Hydraulic Fracturing Treatment CIE44 & IMSS. 14 2092-2106
[2] Xiang-jun, X., & Ting, Y. (2014) Application of TS Fuzzy Model in Candidate-well Selection for Hydraulic Fracturing In Fuzzy Information & Engineering and Operations Research & Management. Springer Berlin Heidelberg 563-570
[3] Zoveidavianpoor, M., Samsuri, A., & Shadizadeh, S. R 2012 Fuzzy logic in candidate-well selection for hydraulic fracturing in oil and gas wells: A critical review International Journal of Physical Sciences. 7 4049-4060
[4] Zadeh L A 1965 Fuzzy sets Information & Control. 8 338-353
[5] Zeng J, Liu Z Q 2006 Type-2 Fuzzy hidden Markov models and their application to speech recognition IEEE Transactions on Fuzzy Systems. 14 454-467.
[6] SUBBHEY S C M, SAMBRIDGE M 2004 Prediction under uncertainty in reservoir modeling Journal of Petroleum Science & Engineering. 44 143-53
[7] Zoveidavianpoor M, Gharibi A 2016 Applications of type-2 fuzzy logic system: handling the uncertainty associated with candidate-well selection for hydraulic fracturing Neural Computing & Applications. 27 1831-1851
[8] Huang, G. B., Zhu, Q. Y., & Siew, C. K 2006 Extreme learning machine: Theory and applications Neurocomputing. 70 489-501
[9] Olatunji, S. O., Selamat, A., & Abdulraheem, A 2014 A hybrid model through the fusion of type-2 fuzzy logic systems and extreme learning machines for modelling permeability prediction Information Fusion. 16 29-45
[10] Tao, S., Tang, D., Xu, H., Gao, L., & Fang, Y 2014 Factors controlling high-yield coalbed methane vertical wells in the Fanzhuang Block, Southern Qinshui Basin International Journal of Coal Geology. 134 38-45
[11] Zhao, J., Tang, D., Xu, H., Lv, Y., & Tao, S 2015 High production indexes and the key factors in coalbed methane production: a case in the hancheng block, southeastern ordos basin, china. Journal of Petroleum Science & Engineering 130 55-67
[12] Xu, L., Zhu, K., & Yang, X 2016 Production forecasting of coalbed methane wells based on type-2 Fuzzy logic system The Open Petroleum Engineering Journal. 9 268-278
[13] Fu, X., Qin, Y., Wang, G. G., & Rudolph, V 2009 Evaluation of coal structure and permeability with the aid of geophysical logging technology Fuel.88 2278-2285
[14] Mendel, J. M 2001 Uncertain rule-based fuzzy logic systems: introduction and new directions (pp. 131-184). Upper Saddle River: Prentice Hall PTR
[15] Zhou, Z. H., Wu, J., & Tang, W 2002 Ensembling neural networks: many could be better than all Artificial Intelligence. 137 239-263
[16] Martínez-Muñoz G, Suárez A 2006 Pruning in ordered bagging ensembles Proceedings of the 23rd international conference on Machine learning. ACM: 609-616
[17] Asuncion, A., & Newman, D 2007 UCI machine learning repository
[18] Quinlan, J. R 2014 C4. 5: programs for machine learning (Elsevier)
[19] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P 2002 SMOTE: synthetic minority over-sampling technique Journal of Artificial Intelligence Research. 16 321-357
[20] Wang, S., & Yao, X 2012 Multiclass imbalance problems: Analysis and potential solutions IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 42 1119-1130
[21] Parvin, H., & Minaei-Bidgoli, B 2013 A clustering ensemble framework based on elite selection of weighted clusters Advances in Data Analysis and Classification. 7 181-208

Acknowledgments
This study was supported by SCED of China grant 17ZB0220.