ORIGINAL ARTICLE

Modeling of breakout prediction approach integrating feature dimension reduction with K-means clustering for slab continuous casting

Haiyang Duan[1,2] · Xudong Wang[1,2] · Yu Bai[1,2] · Man Yao[1,2] · Yu Liu[3] · Qingtao Guo[4]

Abstract
Breakout is a catastrophic accident in continuous casting. The existing breakout prediction methods based on logical judgment and neural networks need to constantly adjust the prediction parameters or prepare high-quality samples as the input, resulting in poor robustness and unstable precision stability. Therefore, it is particularly important to develop a breakout prediction method that not only can predict breakout accurately but also avoid human intervention significantly. This work proposes a novel approach for breakout prediction combining K-means clustering and feature dimension reduction. The method uses feature dimension reduction to obtain the typical feature vector (TFV) that can characterize the original temperature change trend, and then a K-means clustering model is established to realize online detection of breakout prediction. The results show that the model has a 100% alarm rate for the true breakout, and meanwhile, reduces the number of false alarms from 555 to 217 compared with the on-line breakout prediction system (BPS). The proposed method does not need to adjust the prediction parameters frequently or prepare the input samples carefully, which not only avoids the human intervention but also meets the requirements of online monitoring for the practicality and applicability of the breakout prediction method.

Keywords Feature selection · Dimension reduction · K-means clustering · Breakout prediction · Continuous casting

1 Introduction

Breakout is a serious accident in continuous casting, which is known as the rupture of weak shell and exudation of molten steel below the mold [1]. Breakout not only damages the caster equipment but also reduces the product quality and efficiency [2]. Therefore, breakout prediction is always the top priority of the whole process control of continuous casting.

✉ Xudong Wang
  hler@dlut.edu.cn

1 School of Materials Science and Engineering, Dalian University of Technology, Dalian 116024, China

2 Key Laboratory of Solidification Control and Digital Preparation Technology (Liaoning Province), Dalian University of Technology, Dalian 116024, China

3 School of Mechanical Engineering, Northeast Electric Power University, Jilin 132012, China

4 State Key Laboratory of Metal Material for Marine Equipment and Application, Anshan 114009, Liaoning, China

The mold level and the friction between the mold and shell fluctuate significantly when breakout occurs [3]. Through discussing the characteristics of mold friction, Ma et al. [4] considered that the friction can be used to predict abnormal situations before breakout. Salah et al. [5] developed a model integrating adaptive principal component analysis and mold level to detect and evaluate breakout. Zhang et al. [6] proposed a model for simulating the breakout caused by friction and illustrated the mechanism of the breakout in detail. However, the friction or mold level is difficult to reflect the local features of breakout. Therefore, the above works are insufficient in the accuracy and sensitivity of the breakout prediction, so they are mostly used for offline analysis.

The temperature change trend of thermocouple (TC) embedded in mold copper plates can directly show the process of generation, propagation, and expansion of breakout. Hence, the breakout prediction methods based on temperature of TC are widely used. Bhattacharya et al. [7] established a breakout prediction system to transform temperature into a "breakoutability." By using the Levenberg-Marquardt (LM) algorithm and genetic algorithm (GA) to optimize weights and thresholds, Liu et al. [8] constructed a back-propagation (BP)

neural network model for breakout prediction. He et al. [9] developed the breakout detection model using logic judgment. The breakout prediction models based on logical judgment usually have to adjust the prediction parameters frequently under different steel grades, casting speeds, and casting conditions. Neural networks can determine the weights and thresholds by self-learning, but have strong dependence on the quantity and quality of training samples.

In the previous work [10], we established a prediction model of breakout based on temperature timing characteristics. The model combines dynamic time warping (DTW) with K-means clustering to measure the similarity of temperature under different casting conditions to detect breakout. The detection results confirm the feasibility of machine learning algorithms for continuous casting anomaly prediction. However, taking the full-time temperature series of each thermocouple as the input at any time will dramatically increase the subsequent calculation of similarity and clustering evaluation, which brings a huge burden to the processing speed and real-time performance for on-line monitoring.

In order to ensure the real-time performance of detection accuracy and avoid as much as possible the impact of human intervention on the model, this work uses feature dimension reduction to obtain the TFV that can represent the change trend of original temperature. And then, a K-means clustering breakout prediction model based on TFV is established to detect breakout online.

The rest of this paper is organized as follows: Section 2 introduces the breakout mechanism, as well as the temperature features under different casting conditions; section 3 describes the process of obtaining TFV with feature selection and dimension reduction; section 4 elaborates the establishment of K-means clustering breakout prediction model based on TFV; and section 5 shows the process and results of online detection. In the end, the conclusion of this paper is given.

2 Mechanism and temperature features of the breakout

The main parameters of the caster are summarized in Table 1. Figure 1 illustrates the arrangement of TCs in the mold. The mold consists of four copper plates with 120 TCs in 40 columns and 3 rows, of which 57 TCs are installed in 19 columns and 3 rows on each wide face (loosed and fixed wide faces) and 3 TCs are installed in a single column and 3 rows on each narrow face (left and right narrow faces).

Under the normal casting conditions, the shell thickness and the heat resistance increase along casting direction, and the temperature of lower row TCs

continuously decreases, as shown in Fig. 1(a). Breakout mainly occurs near the meniscus inside the mold. During the downward movement along the casting direction, the breakout will expand longitudinally and horizontally. When breakout gradually expands to ① and ②, the temperature of TC in the first row will rise first and then fall, as shown in Figs. 1(b)–(c). When breakout expands to ③ after a period of propagation and expansion, the temperature of TC in the second row has the similar change trend with that in the first row, as shown in Fig. 1(d). It can be seen from Fig. 1 that the change trend of the temperature of TCs in the same column and the adjacent columns, rising first and then falling, is the typical characteristic of the temperature when breakout occurs.

Figure 2 displays the change trend of temperature under different modes, such as normal casting condition, breakout, and false alarm. The temperature changes steadily with time under normal casting conditions, as shown in Figs. 2(a)–(c). Figures 2(d)–(f) are the temperature of TC in the first and second rows during breakout. It can be seen that the temperature of TC in the first row rises and then falls, and subsequently the temperature of TC in the second row starts to rise, which is completely consistent with the change trend of the temperature described Fig. 1.

False alarm refers to an incorrect alarm issued by BPS which misjudges the temperature under normal casting conditions as that of breakout, such as the temperature shown in Figs. 2(g)–(i). In Fig. 2(g), although the temperature of TC in the second row has a rising trend after that in the first row rises for a period of time, there is no falling trend in the temperature of the first row. In Fig. 2(h), the temperature of TC in the second row has an obvious rising and falling trend,

Table 1 Main parameters of the cast

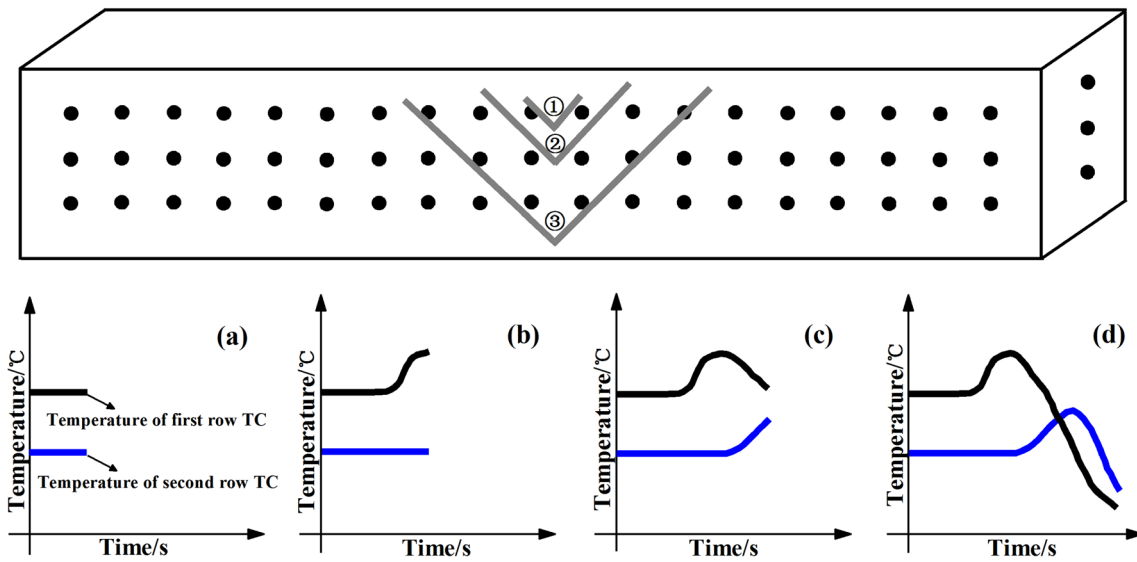| Item	Value
Caster type	Curved-type
Caster radius (m)	10.75
Number of strands	1
Casting speed (m/min)	0.6–1.4
Mold length (mm)	900
Effective mold length (mm)	800
Metallurgical length (m)	28.8
Lubrication	Powder
Mold flux feeding	Manual

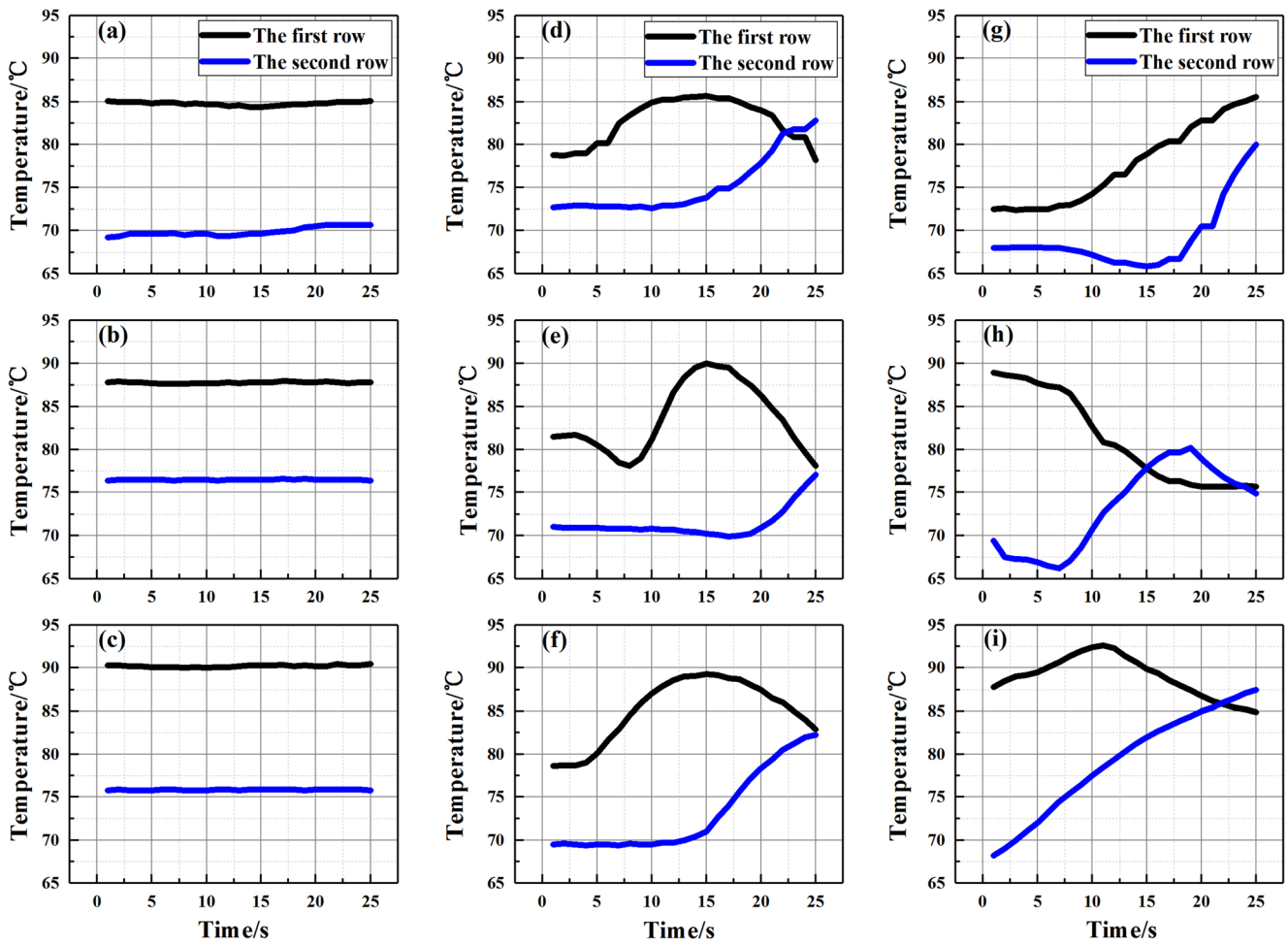Fig. 1 TCs arrangement in the mold and temperature variation of the breakout

Fig. 2 The temperature variation of different modes: **a**–**c** normal casting conditions, **d**–**f** breakout, and **g**–**i** false alarm

while that in the first row does not appear the rising trend. In Fig. 2(i), the temperature of TC in the first row has a typical change trend of rising first and then falling, but the temperature of TC in the second row rises simultaneously with that in the first row, regardless of the time lag caused by the propagation and expansion of breakout. It can be seen that the temperatures shown in Figs. 2(g)–(i) only have the local similarity compared with the breakout temperatures. However, there is still an essential difference between the temperature change trend shown in Figs. 2(g)–(i) and the change trend of the temperature during breakout, so the breakout prediction performance of the BPS still needs to be improved.

3 Determination of temperature typical feature vector

3.1 Feature extraction of temperature

The temperature features of the first and second rows' TCs which are close to the meniscus are extracted from the perspectives of temperature, change rate, temporal and spatial scale, respectively. The extraction process is shown in Fig. 3.

Figures 3(a) and (b) are feature extraction of temperature and change rate, respectively, in which the shadow grid denotes the Sliding Time Window (STW) of temperature and change rate. The features extracted in STW include:

1st_Rising_T_Amplitude(°C): amplitude of rising temperature of 1st row;

1st_Rising_T_Slope(°C/s): slope of rising temperature of 1st row;

1st_2nd_Timelag_T (s): temperature time-lag between 1st and 2nd rows;

1st_Rising_V_Ave(°C/s): average change rate of rising temperature of 1st row;

1st_Rising_V_Max (°C/s): maximum change rate of rising temperature of 1st row;

1st_Falling_V_Ave(°C/s): average change rate of falling temperature of 1st row;

2nd_Rising_V_Ave(°C/s): average change rate of rising temperature of 2nd row;

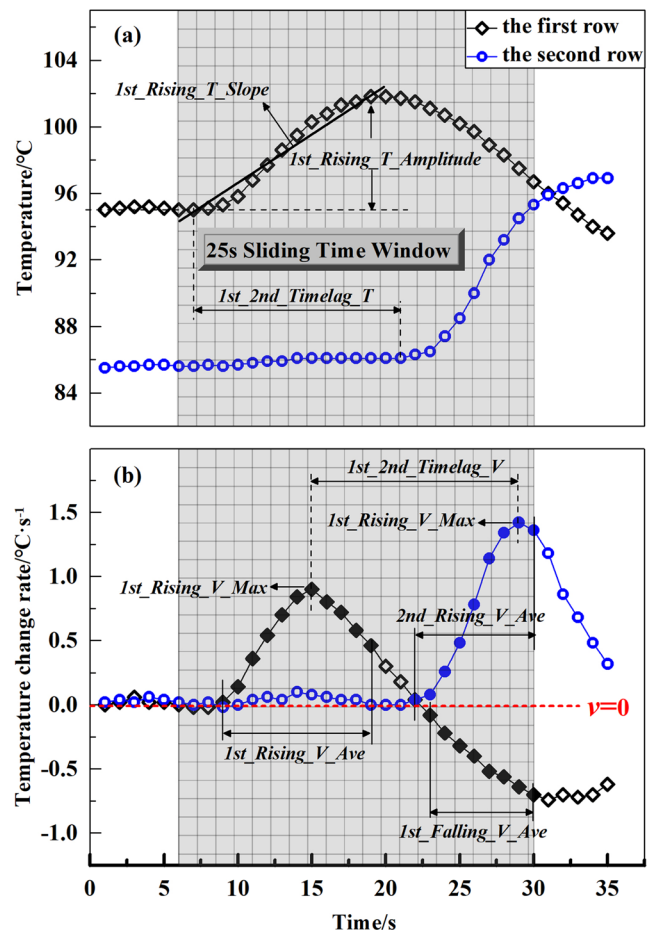2nd_Rising_V_Max (°C/s): maximum change rate of rising temperature of 2nd row;

Fig. 3 Feature extraction of temperature: **a** temperature, **b** temperature change rate

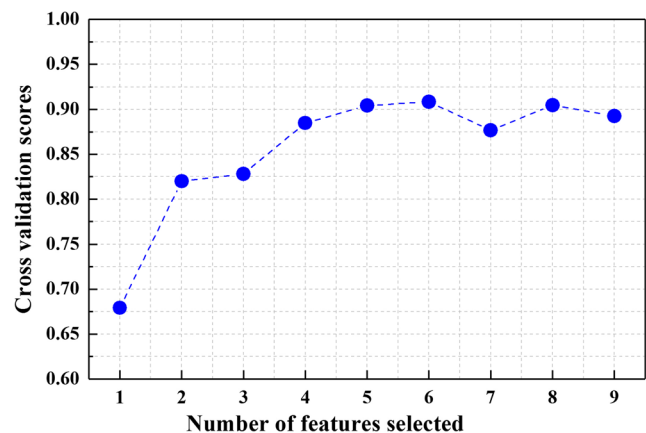1st_2nd_Timelag_V(s): time-lag of maximum change rate of rising temperature between 2nd and 1st rows.

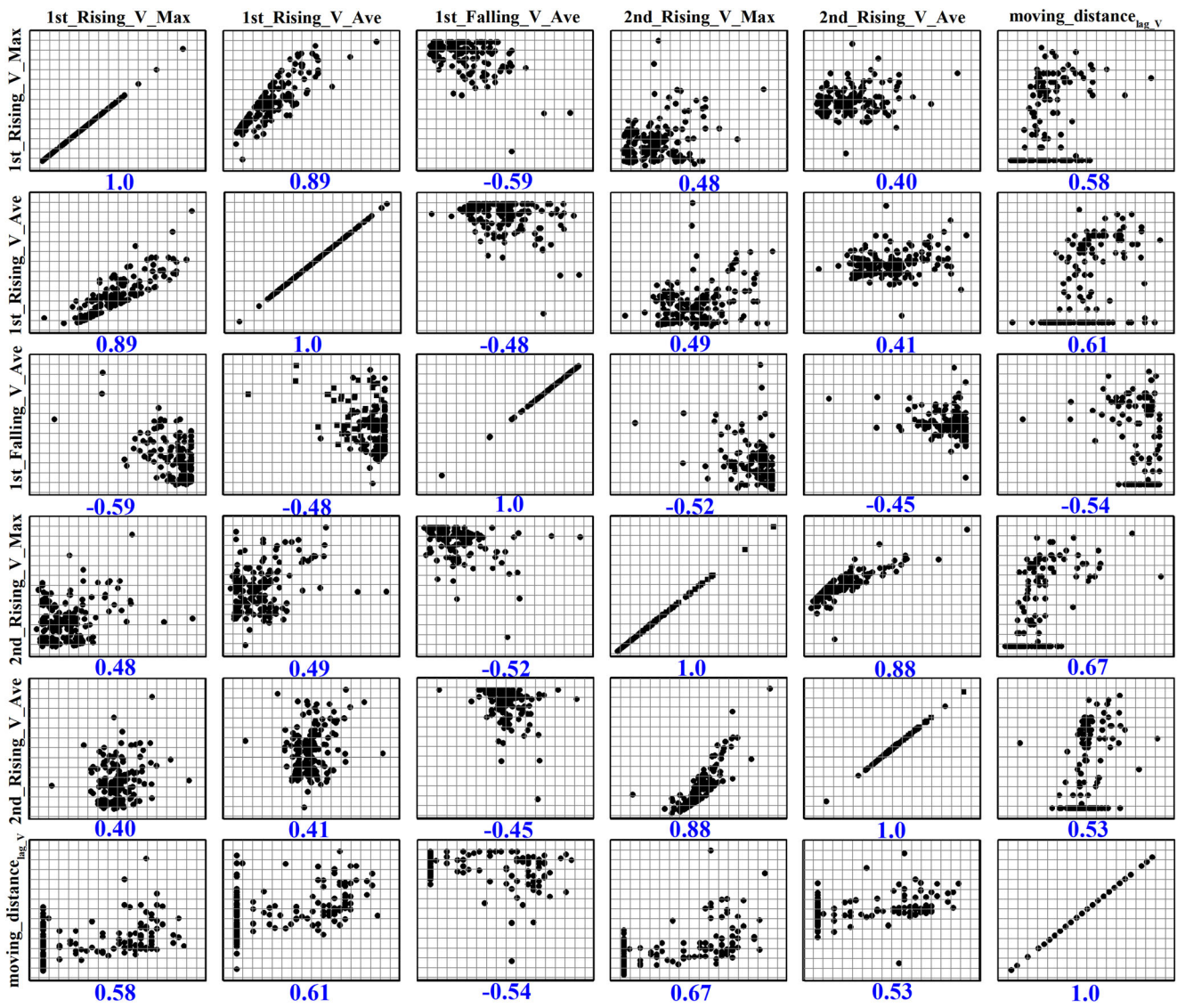Fig. 4 CVS results with different numbers of features selected

Fig. 5 The Pearson correlation coefficient matrix for each two typical features
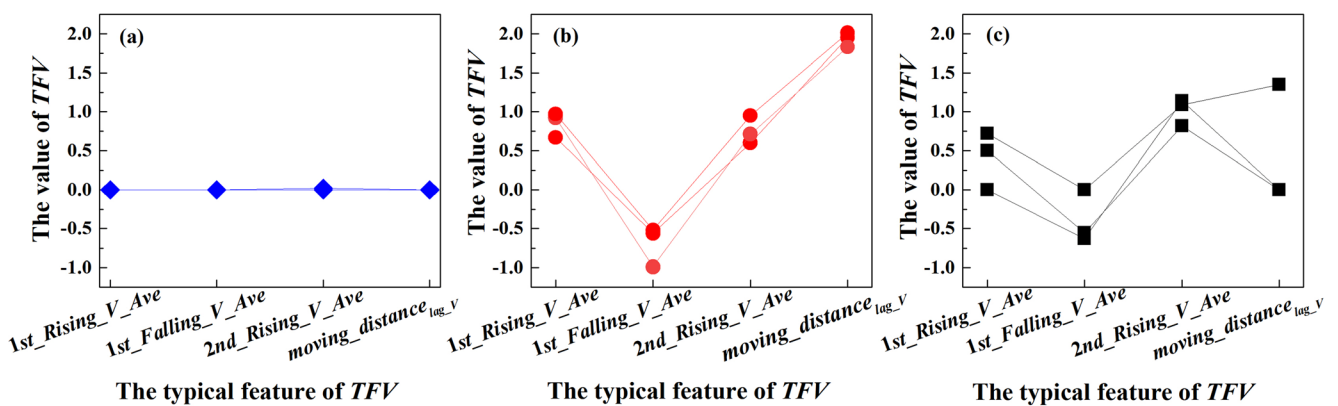

Fig. 6 The TFV constructed from the temperature in Fig. 2: **a** TFV of Fig. 2(a–c); **b** TFV of Fig. 2(a–c); **c** TFV of Fig. 2(a–c)

It can be seen from Fig. 3 that the value of 1st_2nd_Timelag_V is much larger than that of other features constructed from the same temperature, as well as 1st_2nd_Timelag_T. Therefore, in order to convert the value of each feature to the same scale, 1st_2nd_Timelag_V and 1st_2nd_Timelag_T are converted into the distances traveled by the slab, called $moving\_distance_{lag\_T}(dm)$ and $moving\_distance_{lag\_V}(dm)$. The conversion equations are as follows:

$$moving\_distance_{lag\_T} = \frac{10 \cdot V_c \cdot 1st\_2nd\_Timelag\_T}{60} \quad (1)$$

$$moving\_distance_{lag\_V} = \frac{10 \cdot V_c \cdot 1st\_2nd\_Timelag\_V}{60} \quad (2)$$

where, V_c denotes the casting speed.

3.2 Feature selection and dimension reduction

The features mentioned above comprehensively contain the important details of temperature and change rate that distinguish the temperature variation under breakout and normal casting conditions. However, some of them are strongly correlated and redundant. Therefore, it is necessary to use feature selection for reducing the negative impact of redundant features on the prediction effect. In addition, feature dimension can be used to increase the calculation speed for satisfying the real-time requirements of the prediction method.

3.2.1 Feature selection

Recursive Feature Elimination Cross-Validation (RFECV) is one of wrapper feature selection methods which performs RFE [11] method with cross-validation. RFECV calculates cross-validation scores (CVS) of all feature subsets by a learner model, and selects the subsets with the highest CVS. Figure 4 shows that CVS is up to 0.908 when the number of features selected is 6. Thus, six typical features are selected by RFECV automatically: 1st_Rising_V_Max, 1st_Rising_V_Ave, 1st_Falling_V_Ave, 2nd_Rising_V_Ave, 2nd_Rising_V_Max, and $moving\_distance_{lag\_V}$.

3.2.2 Dimension reduction

Then, this work calculates the Pearson correlation coefficient (PCC) between the six typical features selected above by Eq. (3) to construct the PCC matrix, as shown in Fig. 5:
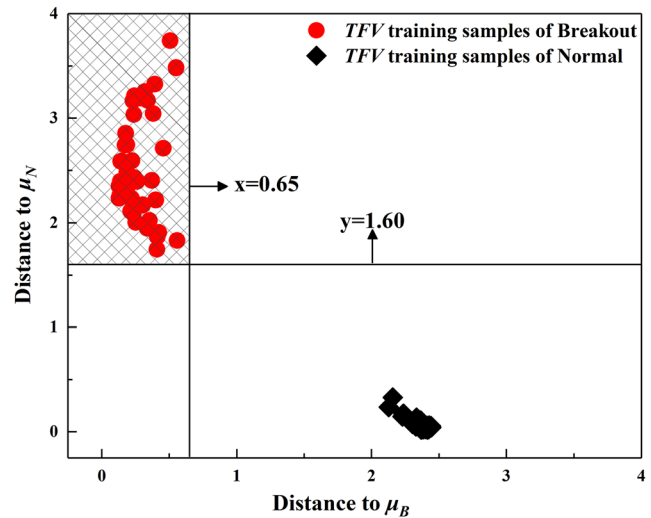

Fig. 7 The K-means clustering result of training TFV samples

$$R_{xy} = \frac{cov(x, y)}{\sigma_x \sigma_y} \quad (3)$$

where $cov(x, y)$ denotes the co-variance of x, y, and σ_x, σ_y denotes the variance of x, y.

As can be seen from Fig. 5, PCC of 1st_Rising_V_Max and 1st_Rising_V_Ave is 0.89, and PCC of 2nd_Rising_V_Max and 2nd_Rising_V_Ave is 0.88, which indicate that there is information redundancy in the characterization of the above features. Because 1st_Rising_V_Ave and 2nd_Rising_V_Ave can reflect the overall change trend of rising temperature, 1st_Rising_V_Max and 2nd_Rising_V_Max are eliminated.

Based on the above calculation process, the TFV consists of the remaining features, including: 1st_Rising_V_Ave, 1st_Falling_V_Ave, 2nd_Rising_V_Ave, and $moving\_distance_{lag\_V}$. For example, Fig. 6 shows the value of the TFV constructed from the temperature shown in Fig. 2, of which Figs. 6(a), (b), and (c) contain the TFV of temperature in Figs. 2 (a)–(c), (d)–(f), and (g)–(i), respectively.

Table 2   Distance between TFV samples in Fig. 6 and cluster centroids

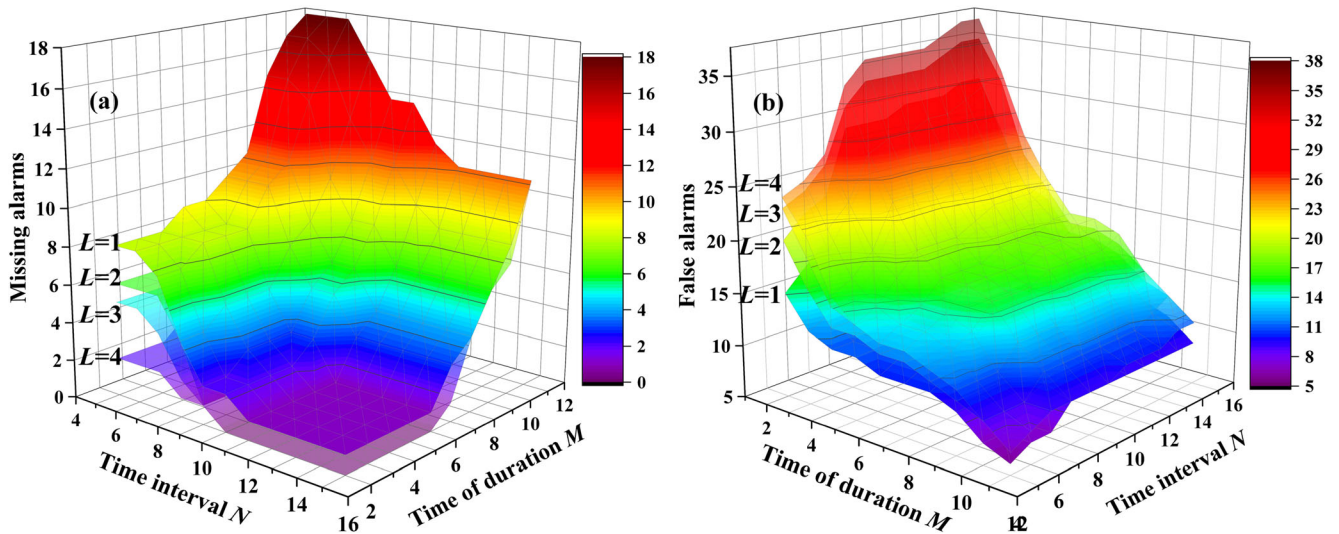|	TFV in Fig. 6(a)			TFV in Fig. 6(b)			TFV in Fig. 6(c)		
μ_B	2.38	2.39	2.39	0.45	0.58	0.15	0.77	2.11	1.96
μ_N	0.10	0.11	0.11	2.17	2.32	2.42	1.81	0.23	1.01

Fig. 8 The number of **a** missing and **b** false alarm with different *M*, *L*, and *N*

After all these, 90 TFV training samples, including 40 samples of breakout and 50 samples including normal casting conditions and false alarm, were extracted to construct TFV set *Q*.

4 K-means clustering of typical feature vector

4.1 K-means clustering

K-means [12] belongs to a rough clustering method. The training samples are divided into K non-overlapping clusters, in
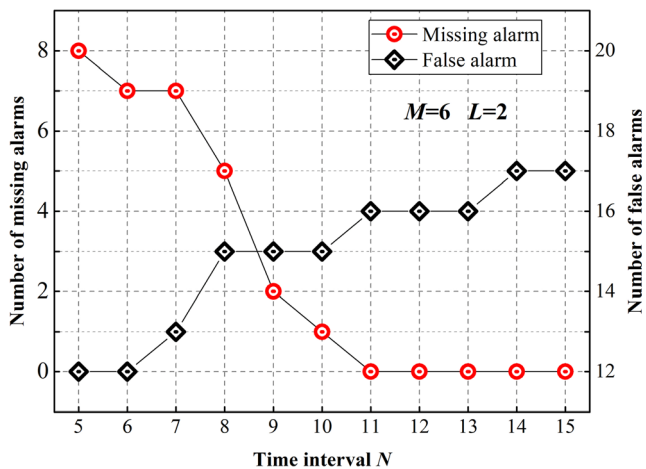

Fig. 9 The effect of *N* on the number of missing and false alarms (*M* = 6 and *L* = 2)
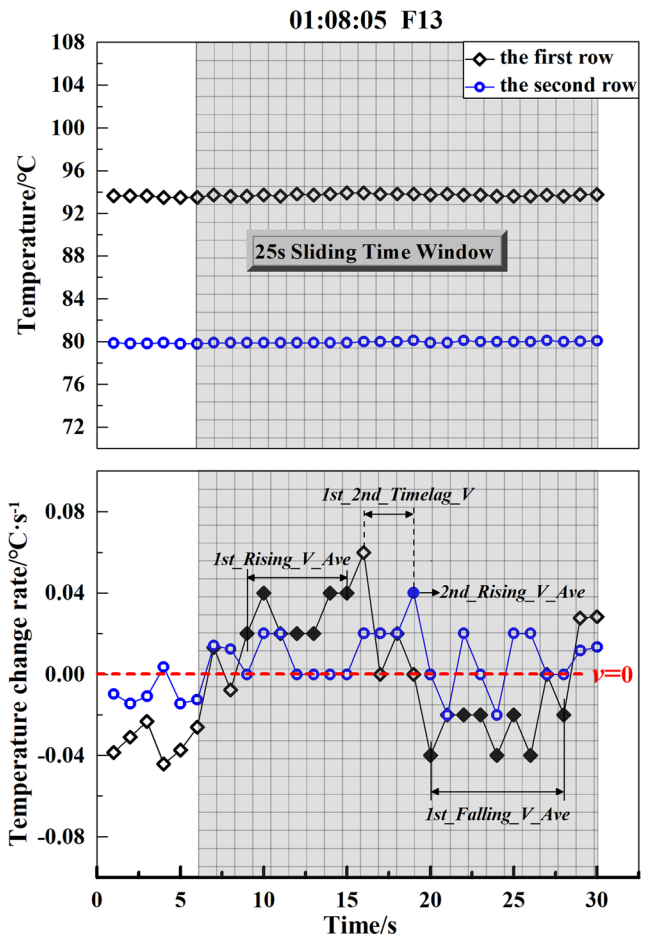

Fig. 10 TFV of normal conditions

which the similarity of samples is high in the same cluster and low in different clusters.

Because there are two temperature modes of normal and breakout, K is set to be 2 and K-means++ strategy [13, 14] is applied to select initial centroids. Then K-means clustering was performed on the set Q to obtain the breakout and normal clusters C_B, C_N and their centroids μ_B, μ_N, i.e.:

$$\mu_B = \{0.87, -0.50, 1.01, 1.92\}, \quad \mu_N$$
$$= \{0.08, -0.02, 0.07, 0.00\}$$

The clustering result is shown in Fig. 7, in which horizontal and vertical coordinates are distances of TFV to μ_B and μ_N.

As can be observed in Fig. 7, the TFV samples of normal and breakout separate from each other and locate in two regions. The former is represented by black diamonds, and the latter by red circles, which indicates that the samples falling into the area where the latter locates are most likely to be a breakout.

The above results demonstrate that the clustering model based on TFV can effectively gather the breakout sample and meanwhile separate normal samples. Therefore, a "region of breakout" (ROB) can be constructed by mode segmentation thresholds to identify and distinguish TFV samples. The thresholds are determined based on defining the smallest ROB under the premise of completely covering all the TFV breakout samples. In this respect, the thresholds are set, $x = 0.65$ and $y = 1.60$, and the ROB is shown in the shadow region of Fig. 7.

Table 2 demonstrates the distance between the TFV samples in Fig. 6 and cluster centroids μ_B, μ_N. It can be seen from Table 2 that the TFV samples of breakout in Fig. 6(b) are located in the ROB, while that under normal casting conditions and false alarm in Figs. 6(a) and (c) are all located out of the ROB.

4.2 Breakout prediction approach

1) TFV construction: obtain TFV from the temperature of all 40 columns TCs;
2) Distance measurement: calculate the distances between TFV and μ_B, μ_N to obtain $dist_B$ and $dist_N$;
3) Abnormal time marking: if $dist_B$ and $dist_N$ are located into the ROB, mark the time as $abnor_{time}$; otherwise, return to 1) and calculate the next second;
4) Abnormal TC column marking: if $dist_B$ and $dist_N$ satisfy the condition 3) for continuous M seconds, mark this TC column as TC_{col};

5) Adjacent TC column judgment: if at least one column of the left and right TC columns of TC_{col} satisfies the condition 4), meanwhile, the time interval of their $abnor_{time}$ is less than N seconds, mark the corresponding TC column as TC_{adj}, and the total number of columns of TC_{adj} is denoted as L;
6) Breakout identification: if corresponding TFV satisfy 4) and 5), issue a breakout alarm and then mark the abnormal TC columns; otherwise, return to 1) and continue checking and judging.

4.3 Determination of prediction parameters

The simultaneous judgments of M, L, and N greatly weaken the influence of the ROB on missing and false
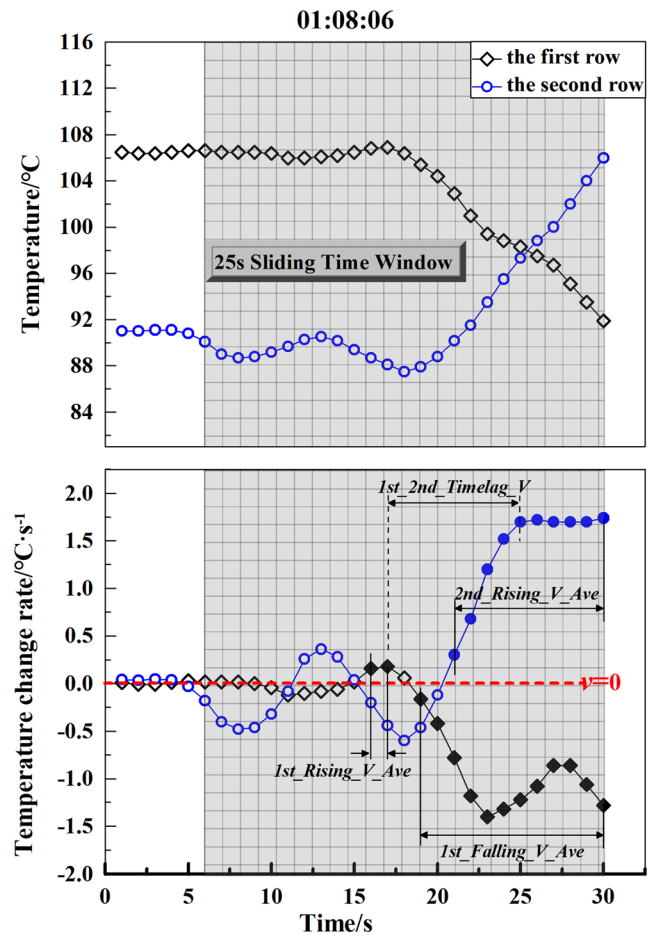

Fig. 11 TFV of false alarms

alarm. M characterizes the accumulation and evolution of abnormal time along the casting direction. L characterizes the propagation of sticking in the spatial horizontal direction. N characterizes the connection and transition of the sticking in temporal and spatial. To characterize the sticking propagation and expansion in temporal and spatial precisely, it is of great importance to regulate and match the prediction parameters reasonably, so as to avoid missing the true alarms and reduce the false alarms.

The prediction results of different combination parameters are illustrated in Fig. 8(a) and (b), which denote the number of missing and false alarms when $L =$ 1, 2, 3, and 4. The requirements of breakout prediction are that there is no missing alarms and the number of false alarms is as little as possible.

With the increase of M, the number of missing alarms increases and that of false alarms decreases, as shown in Fig. 8(a) and (b). In addition, the number of missing alarms is 0 when $M \leq 6$. Taking into account missing and false alarm comprehensively, M should be set to six. As observed in Fig. 8(b), the number of false alarms reduces significantly with the decrease of L. However, the missing alarm occurs when $L = 1$, as observed in Fig. 8(a). Hence, there is no missing alarm and the false alarms are relatively minimal when $L = 2$.
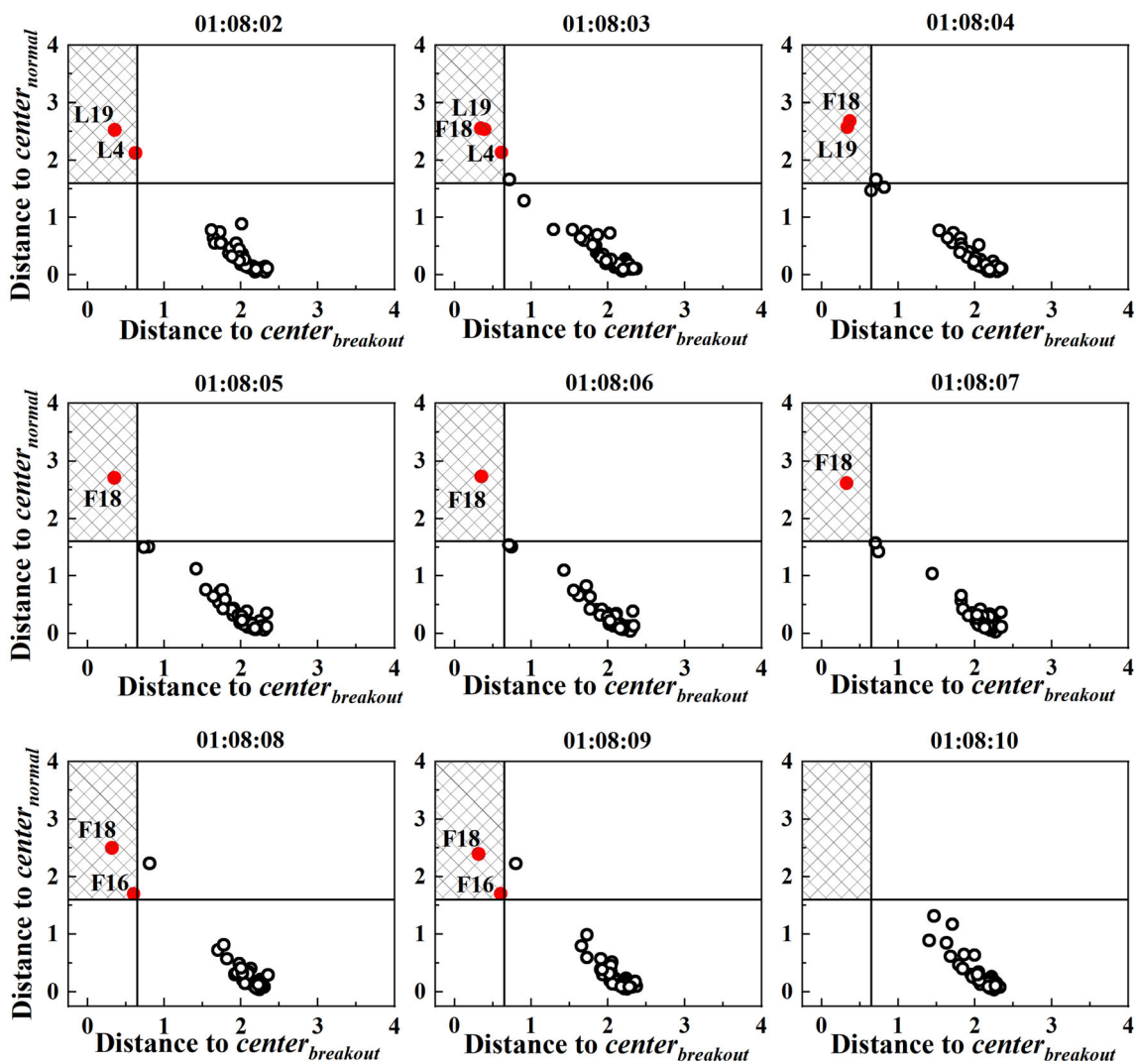

Fig. 12 The clustering results of TFV of normal conditions and false alarms

Figure 9(a) and (b) shows the number of missing and false alarms with the change of N when $M = 6$ and $L = 2$. As observed in Fig. 9(a) and (b), the number of missing alarms is zero and that of false alarms are the lowest when $N = 11$, 12, or 13. Based on the above analysis, this work takes $M = 6$, $L = 2$, and $N = 12$.

5 Result and discussion

The breakout prediction method built above was applied to test the historical temperature data from year 2016 to year 2018. In the online testing process, the TFVs are simultaneously constructed for the temperature corresponding to each moment of all the 40 columns of TC, and they are judged whether they meet the requirements in section 4.2.

Figures 10 and 11 show the temperature of TC in the first and second rows at 01:08:05 at the position of 13th and 18th columns of TC on the wide face, respectively. The former is the temperature under normal casting conditions, while the latter is the false alarm temperature judged by BPS. The temperature data including that under normal casting conditions and false alarm were tested online, and the test process is shown in Fig. 12.

As can be seen from Fig. 12, the TFV of 19th and 4th columns of TC on the loosed wide face both satisfy condition 3) at 01:08:02, which is marked as $abnor_{time}$, but condition 4) is not satisfied because the duration is less than 6 s. The 18th column of TCs on the fixed wide face is marked as TC_{col} due to the TFV is located in the ROB and lasts for 7 s from 01:08:03. Subsequently, the adjacent 16th column of TC on the fixed wide face is located in the ROB at 01:08:08, but only lasts for 2 s, which does not satisfy condition 5). Therefore, the above temperature variation is not considered as a breakout.

Figure 13 shows the temperature of TC in the first and second rows at 11:59:50 at the position of the 8th column of TC on the wide face. The temperature data under breakout were tested online, and the test process is shown in Fig. 14.

As observed in Fig. 14, all the TFV are far away from the ROB at 11:59:44. The 8th column on the fixed wide face locates in the ROB from 11:59:45 for 8 s, so it should be marked as TC_{col}. After the moment, the 9th column on the fixed wide face steps in ROB and maintains this status for continuous 6 s. The $abnor_{time}$ time interval between the 8th and 9th columns is only 2 s. Therefore, a breakout alarm is issued immediately, and the two abnormal TC columns are labeled. It is worth
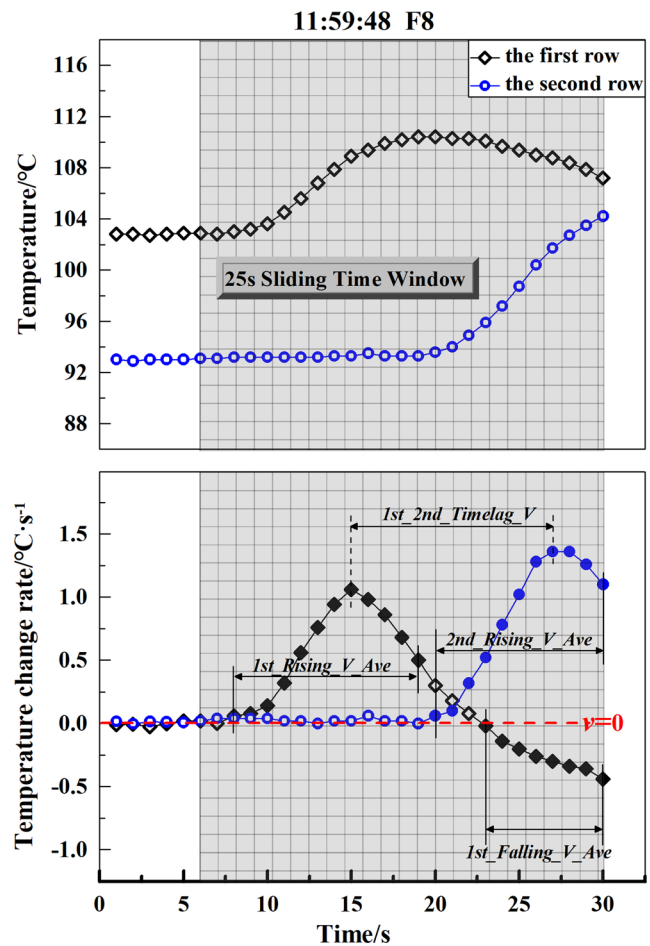

Fig. 13 TFV of breakout temperature

noting that the 7th column also stays for a few seconds after 11:59:50, which indicates that the breakout starts near the 8th column and extends to the adjacent 9th and 7th columns in a short time.

Table 3 shows the test results of the casting temperature data and its comparison with BPS. It can be seen from Table 3 that the proposed method can detect all 56 cases of true breakout, compared with BPS, the number of false alarms decreases from 555 to 217, which greatly reduces the false alarm rate.

It should be noted that no further modifications were made to M, N, and L after determining the prediction parameters during the test of a 3-year historical data, which is not possible to achieve by logical judgment and neural network methods.

6 Conclusion

In view of the shortcomings of existing breakout prediction methods in adaptability and migratability, the
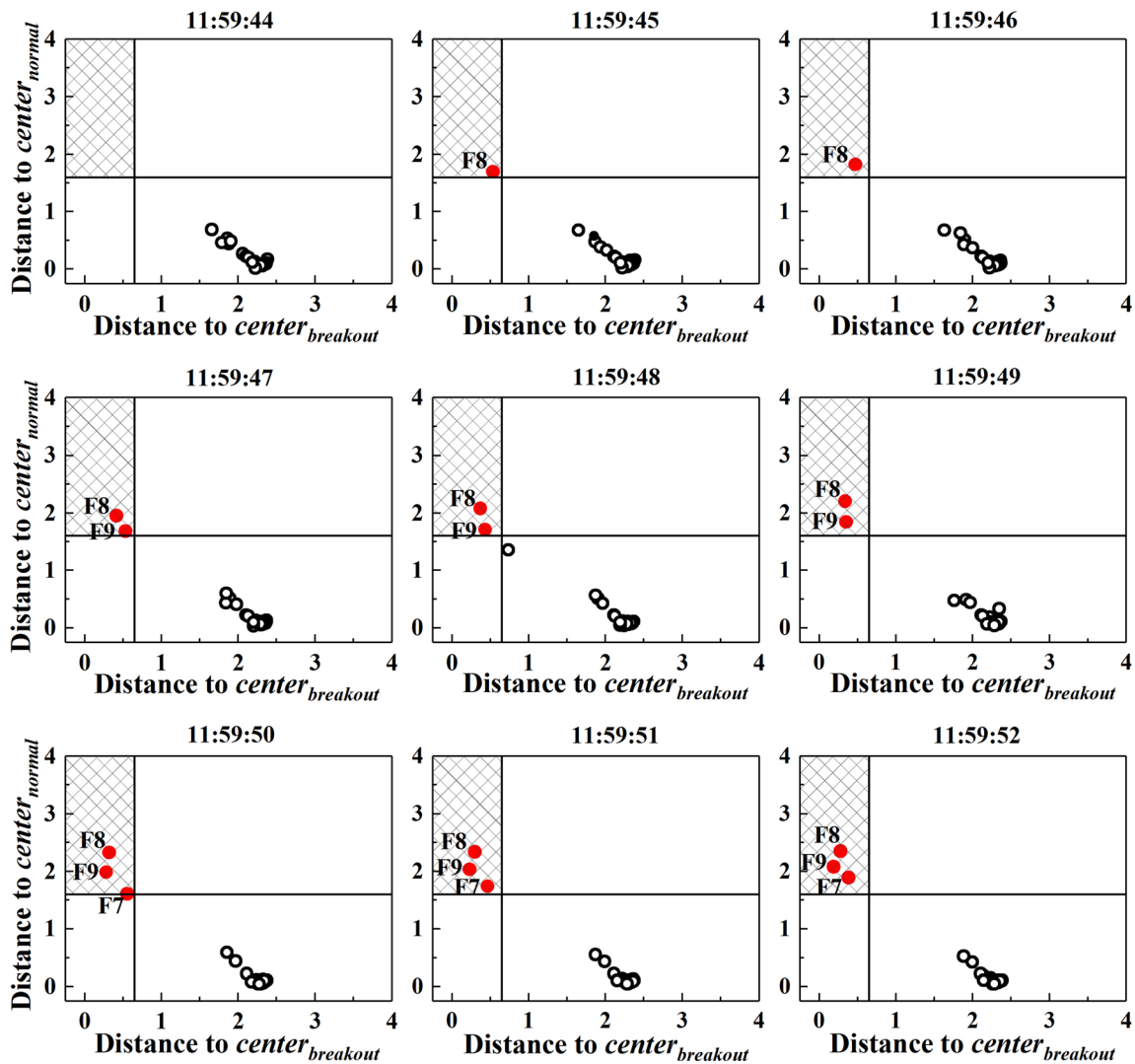
Fig. 14 The clustering results of TFV results of breakout

present work proposed a novel prediction method integrating the feature dimension reduction with the K-means clustering algorithm. In order to avoid the human intervention and meet the applicability requirements of

the online breakout detection, the TFV is extracted and constructed from the typical features of temperature and its change rate by using the feature selection and dimension reduction. The clustering model based on K-

Table 3 Comparison of the number of missing and false alarms

	BPS			The proposed method		
	True alarm	False alarm	Missing alarm	True alarm	False alarm	Missing alarm
2016	36	164	0	36	54	0
2017	6	279	0	6	103	0
2018	14	112	0	14	60	0
Total	56	555	0	56	217	0

means algorithm is built to obtain the TFVs' centroids and region partition, and the ROB is subsequently used to identify the breakout and normal casting conditions. The prediction result shows that the number of missing alarms is zero, and the false alarms are reduced from 555 to 217 compared with the on-line breakout prediction system (BPS). The proposed method gets rid of the limitations of traditional breakout prediction methods in parameter setting and training samples' preparation. At the same time, using TFV as the model input also greatly improves the detection speed, making the efficiency and accuracy of online detection and prediction of breakout improve significantly. The proposed approach provides a reference for application of machine learning in abnormality prediction during the continuous casting process.

References

1. Cheung N, Carcia A (2001) The use of a heuristic search technique for the optimization of quality of steel billets produced by continuous casting. Eng Appl Artif Intell 14(2):229–238. https://doi.org/10.1016/S0952-1976(00)00075-0
2. Kajitani T, Kato Y, Harada K, Saito K, Harashima K, Yamada W (2008) Mechanism of a hydrogen-induced sticker breakout in continuous casting of steel: influence of hydroxyl ions in Mould flux on heat transfer and lubrication in the continuous casting mould. ISIJ Int 48(9):1215–1224. https://doi.org/10.2355/isijinternational.48.1215
3. Roy PDS, Tiwari PK (2019) Knowledge discovery and predictive accuracy comparison of different classification algorithms for mould level fluctuation phenomenon in thin slab caster. J Intell Manuf 30(1):241–254. https://doi.org/10.1007/s10845-016-1242-x
4. Ma Y, Wang XD, Zang XY, Yao M, Zhang L, Ye SH (2010) Mould lubrication and friction behaviour with hydraulic oscillators in slab continuous casting. Ironmak Steelmak 37(2):112–118. https://doi.org/10.1179/030192309X12549935902347
5. Salah B, Zoheir M, Slimane Z, Jurgen B (2015) Inferential sensor-based adaptive principal components analysis of mould bath level for breakout defect detection and evaluation in continuous casting. Appl Soft Comput 34:120–128. https://doi.org/10.1016/j.asoc.2015.04.042
6. Zhang YX, Wang WL, Zhang HH (2016) Development of a mold cracking simulator: the study of breakout and crack formation in continuous casting. Metall Mater Trans B Process Metall Mater Process Sci 47(4):2244–2252. https://doi.org/10.1007/s11663-016-0705-y
7. Bhattacharya AK, Chithra K, Jatla SSVS, Srinivas PS (2004) Fuzzy diagnostics system for breakout prevention in continuous casting of steel. In: Proceeding of 5th World Congress on Intelligent Control and Automation, vol 4, pp 3141–3145
8. Liu Y, Wang XD, Du FM, Yao M, Gao YL, Wang FW, Wang JY (2017) Computer vision detection of mold breakout in slab continuous casting using an optimized neural network. Int J Adv Manuf Technol 88(1–4):557–564. https://doi.org/10.1007/s00170-016-8792-0
9. He F, Zhou L, Zheng ZH (2015) Novel mold breakout prediction and control technology in slab continuous casting. J Process Control 29:1–10. https://doi.org/10.1016/j.jprocont.2015.03.003
10. Duan HY, Wang XD, Bai Y, Yao M, Guo QT (2019) Application of k-means clustering for temperature timing characteristics in breakout prediction during continuous casting. Int J Adv Manuf Technol 106(11–12):4777–4787. https://doi.org/10.1007/s00170-019-04849-x
11. Chen GS, Zheng QZ (2018) Online chatter detection of the end milling based on wavelet packet transform and support vector machine recursive feature elimination. Int J Adv Manuf Technol 95(1–4):775–784. https://doi.org/10.1007/s00170-017-1242-9
12. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY (2002) An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 24(7):881–892. https://doi.org/10.1109/TPAMI.2002.1017616
13. AOzturk MM, Cavusoglu U, Zengin A (2015) A novel defect prediction method for web pages using k-means plus. Expert Syst Appl 42(19):6496–6506. https://doi.org/10.1016/j.eswa.2015.03.013
14. Xu YJ, Qu MY, Li ZY, MIn GY, Li KQ, Liu ZB (2014) Efficient k-means plus approximation with MapReduce. IEEE Trans Parallel Distrib Syst 25(12):3135–3144. https://doi.org/10.1109/TPDS.2014.2306193