Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Belief-Rule-Base Inference Method Based on Gradient Descent with Momentum

FIRST A. AUTHOR[1], (Fellow, IEEE), SECOND B. AUTHOR[2], AND THIRD C. AUTHOR, JR.[3], (Member, IEEE)

[1]National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov)
[2]Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu)
[3]Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA

Corresponding author: First A. Author (e-mail: author@ boulder.nist.gov).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

ABSTRACT

The belief-rule-base (BRB) inference methodology using evidential reasonging (ER) approach is widely used in different fields, such as fault diagnosis, system identification and decision analysis. Since the calculation of the individual matching degree in the inference process needs to set the candidate value of the antecedent attribute parameter, it is easy to cause the rule zero activation problem, and the intelligent algorithm for parameter training is not as efficient and accurate as the gradient method, but the partial derivative of the gradient method is difficult to construct. In this paper, we propose a new belief rule structure and its gradient training method, aiming to solve zero activation during the inference process and improve inference accuray. We first used the Gaussian function to calculate the similarity of each attribute instead of the original method to avoid the zero activation problem caused by the non-adjacent candidate values of the antecedent attributes. Then, according to the new activation weight calculation method, the corresponding distance-sensitive parameter is set for each attribute, and the weight parameter of each rule is cancelled. This simplifies the calculation of rule weights in the inference process and enables the partial derivatives of the parameters of the inference system to be easily constructed. Finally, we use the momentum optimization gradient stochastic descent method to train the new brb system, which improves the training speed and accuracy compared to the ordinary gradient method. Experiments on several public classification datasets are conducted to validate the proposed approach compared with some recent existing works. The experimental results show that the proposed approach have a better performence in accuray and time consumption.

INDEX TERMS belief rule base, structure optimization, stochastic gradient descent, momentum optimization.

I. INTRODUCTION

It is well known that rule-based intelligent systems are one of the most common frameworks for expressing various types of knowledge. The rule-based system has certain expression and processing capabilities through the use of existing human knowledge, and at the same time has the flexibility to deal with ambiguity, incompleteness, uncertainty and to combine different types of input data formats. In various rule-based systems for solving complex problems, belief rule-based inference methodology using evidential reasonging approach (ER) proposed by Yang *et al*. [1] baesd on conventional IF-THEN rules [2], Dempster-Shafer theory of evidence [3], [4], decision theory [5] and fuzzy set theory [6] shows its powerful function of representing and processing uncertain information. By introducing a belief distribution structure in the rules, this methodology can effectively handle incomplete and uncertain information involved in the datasets and widely used in various problem in different fields such as oil pipeline leak detection [7], military capability estimation [8], consumer behavior prediction [9] and so on.

In the inference process of the belief rule base (BRB) system, the attribute weight, rule weight, belief distribution and other parameters directly affect the final accuray. Yang *et al*. [10] proposed optimization models for training BRB system using fmincon solver in Matlab, Chang *et al*. [11], [12] proposed an algorithm for training parameters in BRB system based on gradient and dichotomy methods, Wu *et al*. [13] used the accelerating of gradient algorithm to improve the convergence accuray and convergence speed. There are also a series of intelligent algorithms such as the particle swarm algorithm proposed by Su *et al*. [14] and the differential evolution algorithm proposed by Wang *et al*. [15] have excellent training effects on the BRB system. Liu *et al*. [16] introduces the belief distribution structure into the antecedent attributes and uses training data to build an extended belief rule base (EBRB) system, which simplifies the construction of the rule base and improves the inference speed.

At present, the parameter optimization model of the BRB system is mostly based on various intelligent algorithms. However, the process of those intelligent algorithms is complicated and there are many intermediate training parameters. When the conventional gradient method trains the parameters of the BRB system, the partial derivative of each parameter is difficult to construct, and the limit method is needed to solve the approximate value of the partial derivative [13]. The step size is limited by the corresponding parameter constraints and a search is required to find the best effective step size [11]–[13]. The EBRB system does not introduce a parameter training process, which makes the system have higher requirements for the representativeness of the training data used to construct the rule base. In the case of a large number of rules, it is necessary to perform rule reduction or use the data structure to optimize the storage and activation process of the rules. Because the conventional BRB system includes the rule attribute reference level setting, its potential zero activation problem may lead to the failure of the inference system.

In response to the above problems, a series of optimization modifications are proposed for the system structure and reasoning process, including:

1) We propose a new antecedent structure that does not need to set the attribute reference level, and propose a Gaussian function-based rule weight activation method for the new rule antecedent structure, which can effectively avoid the zero activation problem and has the feature of generating rules from the training data like EBRB.

2) We change the method of setting the weight of the global same antecedent attribute in the conventional BRB system, and set the corresponding rule attribute weight parameter for each rule, so that each rule has a finer activation granularity. On this basis, the rule weight and its related normalization process are cancelled, which simplifies the evidential reasoning process.

3) We further present a normalized exponential function to preprocess the restricted parameters to avoid the problem of parameter failure during the training process.

The remainder of this paper is organized as follows: Section II introduces the conventional BRB system and our further improvements for common problems in the system. In Section III, we give the preprocessing method of the training model and prove that the gradient descent method can be effectively applied to the newly proposed BRB system. In Section IV, we compare the effects of different gradient descent parameters on training speed and inference accuracy. Experiments on a series of public data sets prove that the newly proposed BRB model and its training method have the best performance. Section V concludes this paper.

II. BRB SYSTEM WITH NEW ATTRIBUTE STRUCTURE AND RULE ACTIVATION WEIGHT CALCULATION METHOD

The BRB system proposed by Yang *et al*. [1] mainly refers to the rule activation and evidence reasoning method on the belief rule base. This section will briefly introduce the related concepts of the BRB system and propose the solutions for the common defects of the conventional BRB system.

A. REPRESENTATION OF BELIEF RULE BASE

On the basis of the conventional production rules, Yang *et al*. [1] proposed the expression form of the belief rules by introducing the belief distribution structure, the rule antecedent attribute parameter and the rule weight parameter. The specific expression is as follows:

$$\begin{aligned} &R_k : if\{X_1 is A_1^k \wedge \cdots \wedge X_{T_k} is A_{T_k}^k\} \\ &then\{(D_1, \beta_1^k), \cdots, (D_N, \beta_N^k)\}, \sum_{i=1}^{N} \beta_i^k \leq 1 \end{aligned} \tag{1}$$

The equal sign is obtained when the rule information is complete. Each rule has its rule weight θ_k, antecedent attribute weight $\delta_1, \delta_2, \cdots, \delta_{T_k}$. A_i^k represents the candidate reference value selected by the rule on the i-th attribute and β_i^k represents the belief degree of the rule in the i-th result attribute. On this basis, the extended belief rule base system introduces a belief distribution structure to the antecedent attributes, and its rule form is expressed as follows:

$$\begin{aligned} &R_k : if\{[(A_{11}^k, \alpha_{11}^k), \cdots, (A_{1J_1}^k, \alpha_{1J_1}^k)] \wedge \\ &\cdots \wedge [(A_{T_k 1}^k, \alpha_{T_k 1}^k), \cdots, (A_{T_k J_{T_k}}^k, \alpha_{T_k J_{T_k}}^k)]\} \\ &then\{(D_1, \beta_1^k), \cdots, (D_N, \beta_N^k)\}, \sum_{i=1}^{N} \beta_i^k \leq 1 \end{aligned} \tag{2}$$

The extended belief rule base converts the original training data into antecedent attributes with a belief distribution form. For the input data $X^k = (x_1^k, \cdots, x_T^k)$, convert the i-th attribute parameter to construct the i-th antecedent attribute of the corresponding rule with a belief distribution form:

$$\begin{aligned} &\alpha_{ij}^k = \frac{\gamma_{i(j+1)} - x_i^k}{\gamma_{i(j+1)} - \gamma_{ij}}, \gamma_{ij} \leq x_i^k \leq \gamma_{i(j+1)} \\ &\alpha_{i(j+1)}^k = 1 - \alpha_{ij}^k, \gamma_{ij} \leq x_i^k \leq \gamma_{i(j+1)} \\ &\alpha_{it}^k = 0, t = 1, \cdots, (j-1), (j+2), \cdots, J_i \end{aligned} \tag{3}$$

According to the same conversion method, the values of original data on other attributes can be converted into the corresponding berlief distribution form. We can also obtain the belief distribution form of the rule result attribute according to this method.

B. EVIDENCE REASONING APPROACH BASED ON BELIEF RULE BASE

The calculation and synthesis of activation weights for each rule in the rule base is the core part of the inference system of the belief rule base. The whole process mainly includes two steps: calculate the activation weight, and synthesize the rules according to the activation weight. The calculation of the activation weight of each rule in the belief rule base can be regarded as calculating the belief distribution similarity on each attribute and combining their results. Euclidean distance is used to calculate the individual matching degree of the i-th attribute. After converting the input data to have the same belief distribution form as the corresponding attribute, the individual matching degree of the attribute is calculated as:

$$S_i^k = 1 - d_i^k = 1 - \sqrt{\frac{\sum_{j=1}^{J_i}(\alpha_{i,j} - \alpha_{i,j}^k)^2}{2}} \tag{4}$$

After the individual matching degree of each attribute is calculated, the individual matching degrees of all attributes are aggregated. The aggregation function in the form of conjunctive rules is:

$$\alpha_k = \prod_{i=1}^{T_k}(S_i^k)^{\overline{\delta_i}}, \overline{\delta_i} = \frac{\delta_i}{\max_{j=1,\cdots,T_k}\delta_j} \tag{5}$$

The activation weight of this rule is calculated by the following formula:

$$w_k = \frac{\theta_k \alpha_k}{\sum_{l=1}^{L}\theta_l \alpha_l} \tag{6}$$

Rule weight normalization operation makes all weights satisfy $0 \le w_k \le 1, \sum w_k = 1$.

After the rule weight calculation is completed, all the rules are synthesized and the inference result is obtained. First, the belief distribution of the rule is transformed into the corresponding probability quality information:

$$m_{j,k} = w_k \beta_j^k, j = 1, \cdots, N \tag{7}$$

$$m_{D,k} = 1 - \sum_{j=1}^{N} m_{j,k} = 1 - w_k \sum_{j=1}^{N}\beta_j^k \tag{8}$$

$$\overline{m}_{D,k} = 1 - w_k \tag{9}$$

$$\widetilde{m}_{D,k} = w_k(1 - \sum_{j=1}^{N}\beta_j^k) \tag{10}$$

$m_{j,k}$ represents the credibility of the k rule on the j result attribute, where $\overline{m}_{D,k}$ represents the credibility that the k-th rule is not assigned to any result attribute, and $\widetilde{m}_{D,k}$ represents the credibility of the missing reference attribute of the result of the k-th rule. The total uncertainty credibility is given by $m_{D,k} = \overline{m}_{D,k} + \widetilde{m}_{D,k}$. Synthesize the credibility information of all rules and obtain the final confidence result of each result attribute:

$$m_j = k[\prod_{i=1}^{L}(m_{j,i} + m_{D,i}) - \prod_{i=1}^{L} m_{D,i}], j = 1, \cdots, N \tag{11}$$

$$\overline{m}_D = n[\prod_{i=1}^{L}\overline{m}_{D,i}], \quad \widetilde{m}_D = k[\prod_{i=1}^{L} m_{D,i} - \prod_{i=1}^{L}\overline{m}_{H,i}] \tag{12}$$

$$k = [\sum_{j=1}^{N}\prod_{i=1}^{L}(m_{j,i} + m_{D,i}) - (N-1)\prod_{i=1}^{L} m_{D,i}]^{-1} \tag{13}$$

$$\beta_j = \frac{m_j}{1 - \overline{m}_D}, j = 1, \cdots, N \tag{14}$$

$$\beta_D = \frac{\widetilde{m}_D}{1 - \overline{m}_D} \tag{15}$$

C. NEW ATTRIBUTE STRUCTURE AND RULE ACTIVATION WEIGHT CALCULATION METHOD

The process of generating rules in the rule base requires artificial setting of candidate reference values for the antecedent attribute information, and when calculating the rule activation weight, if the input attribute information is not in the adjacent area of the rule attribute reference value, the rule cannot be activated. If all the rules in the library are not activated, the reasoning system will fail. In order to solve the above problems, we proposes an improved form of belief rules and corresponding activation weight calculation method as follows:

$$\begin{aligned} &R_k : if(x_1^k, \cdots, x_{T_k}^k) \\ &then\{(D_1, \beta_1^k), \cdots, (D_N, \beta_N^k)\}, \sum_{i=1}^{N}\beta_i^k \le 1 \end{aligned} \tag{16}$$

The simplified belief rule structure can directly use the training data to generate the rule antecedent attribute information without manually setting the candidate reference values of the antecedent attributes.

Using antecedent attribute belief distribution similarity as the activation weight calculation method is no longer applicable to the simplified confidence rule form. In order to perform effective weight activation, we uses Gaussian function to calculate the individual matching degree for activation weight calculation. The degree of individual matching of input data $X(x_1, \cdots, x_{T_k})$ and rule R_k : $if(x_1^k, \cdots, x_{T_k}^k) then\{(D_1, \beta_1^k), \cdots, (D_N, \beta_N^k)\}$ on i-th attribute is calculated using the Gaussian function as:

$$S_i^k = e^{-[a_i^k \times (x_i - x_i^k)]^2} \tag{17}$$

The parameter a_i^k represents the sensitivity of the i-th attribute to the distance at the position x_i^k. When the distance between the rule antecedent attribute and the input data remains unchanged, the value of parameter a inversely proportional to the matching degree. The activation weight

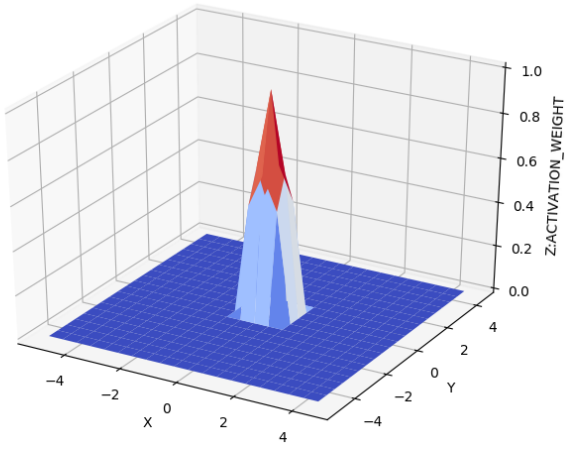

FIGURE 1. activation weight calculated by conventional methods

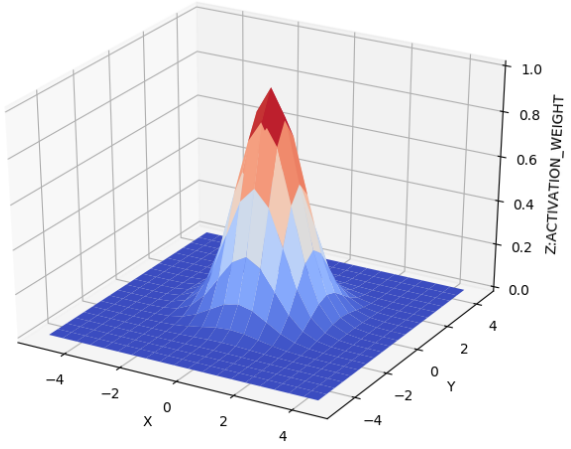

FIGURE 2. activation weight calculated by Gaussian function methods

of a single rule under conjunctive conditions is calculated by the following formula:

$$w_k = \prod_{i=1}^{T_k} S_i^k = e^{-\sum_{i=1}^{T_k}[a_i^k(x_i - x_i^k)]^2} \tag{18}$$

Assuming a rule with two attributes x and y located at the origin, the conventional method and Gaussian function method are used to calculate the activation weights. Set the reference candidate values on the x-attributes and y-attributes to be both $[-4, -3, -2, -1, 0, 1, 2, 3, 4]$, and set the distance-sensitive parameter a of each attribute at the origin is 0.5. For the convenience of calculation, we omitted the setting of rule weight. The two activation weight distributions shown in Figure 1 and Figure 2 can be obtained.

According to Figure 1, we can know that the activation weight calculated by the input data that is not in the adjacent area of the candidate value selected by the rule is zero. If the activation weight of all rules is zero, the inference cannot be performed. However, the input data in Figure 2 smoothly drops close to zero according to the distance from the rule and will not take a value of zero. This eliminates the impact of rule zero activation on system inference performance.

Another benefit brought by the new rule antecedent attribute structure and rule activation weight calculation method is that there is no need to adjust the activation weight of the rule by the attribute weight and rule weight. By adjusting the distance-sensitive parameters on each attribute of each rule, a good activation effect and activation granularity can be obtained. At the same time, due to the characteristics of the Gaussian function, the activation weight of each rule belongs to $(0, 1]$ without unnecessary normalization operations. This greatly simplifies the redundant weight adjustment and calculation in the inference process

III. MOMENTUM OPTIMIZED GRADIENT DESCENT TRAINING PARAMETER

When the conventional gradient method is applied to the parameter training process of the inference system of the belief rule base, it is difficult to construct the partial derivative formula of the rule attribute and the training step is restricted by the parameter constraints. The above-mentioned belief rule base reasoning system with improved structure and activation method avoids the difficulty of obtaining partial derivatives in conventional belief rule systems.

In this section we will:

1) Calculate the partial derivatives of the parameters of each part of the new brb system

2) Prove that the BRB inference system is differentiate

3) Introduce exponential normalization function for preprocessing to avoid specific constraints during parameter training

4) Introduce the stochastic gradient descent method using momentum optimization

A. PARTIAL DERIVATIVE OF THE PARAMETERS OF THE BRB SYSTEM

The reasoning process of the improved BRB system is shown in Figure 3. According to the data flow path of the inference system, we can use the compound function chain derivation rule to obtain the partial derivative of the output to different parameters of the inference system.

Since the model construction and experiment in this paper are carried out with complete data, the evidence reasoning process can be simplified. The belief distribution of the rule result attribute does not include uncertain information, that is, for any k-th rule:

$$\sum_{i=1}^{N} \beta_{ik} = 1(k = 1, \cdots, L) \tag{19}$$

$$m_{D,k} = \overline{m}_{D,k}, \widetilde{m}_{D,k} = 0 \tag{20}$$

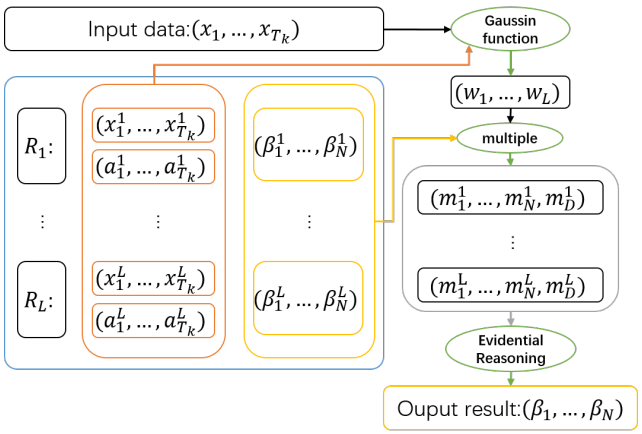

FIGURE 3. Inference system operation process

In the case of completeness, the j-th result attribute is expressed as:

$$\beta_j = \frac{\prod_{i=1}^{L}(m_{j,i} + m_{D,i}) - \prod_{i=1}^{L} m_{D,i}}{\sum_{t=1}^{N}\prod_{i=1}^{L}(m_{t,i} + m_{D,i}) - N \times \prod_{i=1}^{L} m_{D,i}} \tag{21}$$

Then the result of the j-th result attribute before normalization is expressed as:

$$\overline{\beta}_j = \prod_{i=1}^{L}(m_{j,i} + m_{D,i}) - \prod_{i=1}^{L} m_{D,i}, \beta_j = \frac{\overline{\beta}_j}{\sum_{k=1}^{N}\overline{\beta}_k} \tag{22}$$

Substituting $m_{j,i} = w_i\beta_{j,i}$ and $m_{D,i} = 1 - w_i$ into the j-th result attribute before normalization expression:

$$\overline{\beta}_j = \prod_{i=1}^{L}(w_i\beta_j^i + 1 - w_i) - \prod_{i=1}^{L}(1 - w_i) \tag{23}$$

We can get the partial derivative of the i-th result attribute β_i to the j-th unnormalized result attribute $\overline{\beta}_j$ as:

$$\frac{d\beta_i}{d\overline{\beta}_j} = \begin{cases} \dfrac{\sum_{k\neq j}^{N}\overline{\beta}_k}{(\sum_{k=1}^{N}\overline{\beta}_k)^2}, j = i \\ -\dfrac{\beta_i}{(\sum_{k=1}^{N}\overline{\beta}_k)^2}, j \neq i \end{cases} \tag{24}$$

Similarly, the partial derivative of the j-th unnormalized result attribute $\overline{\beta}_j$ to the activation weight of the k-th rule and the j-th result attribute of k-th rule can be obtained as:

$$\frac{d\overline{\beta}_j}{dw_k} = (\beta_j^k - 1)\prod_{i=1,i\neq k}^{L}(w_i\beta_j^i + 1 - w_i) + \prod_{i=1,i\neq k}^{L}(1 - w_i) \tag{25}$$

$$\frac{d\overline{\beta}_j}{d\beta_j^k} = w_k\prod_{i=1,i\neq k}^{L}(w_i\beta_j^i + 1 - w_i) \tag{26}$$

According to the activation weight expression of the k-th rule, we can obtain its partial derivatives of the rule antecedent attribute parameter and the corresponding attribute distance-sensitive parameter respectively:

$$\frac{dw_k}{dx_l^k} = 2(a_l^k)^2(x_l - x_l^k)e^{-\sum_{i=1}^{T_k}[a_i^k(x_i - x_i^k)]^2} \tag{27}$$

$$\frac{dw_k}{da_l^k} = 2a_l^k x_l^k(x_l - x_l^k)e^{-\sum_{i=1}^{T_k}[a_i^k(x_i - x_i^k)]^2} \tag{28}$$

Set the loss function expression of the final output result to $loss = Loss(\beta_1, \cdots, \beta_N)$, according to the compound function chain derivation rule, we can obtain the partial derivative of the final loss on each parameter.

$$\frac{dloss}{d\beta_j^k} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{dloss}{d\beta_i}\frac{d\beta_i}{d\overline{\beta}_j}\frac{d\overline{\beta}_j}{d\beta_j^k} \tag{29}$$

$$\frac{dloss}{dx_l^k} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{dloss}{d\beta_i}\frac{d\beta_i}{d\overline{\beta}_j}\frac{d\overline{\beta}_j}{dw_k}\frac{dw_k}{dx_l^k} \tag{30}$$

$$\frac{dloss}{da_l^k} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{dloss}{d\beta_i}\frac{d\beta_i}{d\overline{\beta}_j}\frac{d\overline{\beta}_j}{dw_k}\frac{dw_k}{da_l^k} \tag{31}$$

B. DIFFERENTIABLE PROOF OF BRB SYSTEM

The evidential reasoning process of the belief rule base is a multivariate compound function process. According to the differentiable condition of the multivariate compound function, each intermediate function must satisfy the differential condition, and the partial derivative of each variable must exist and be continuous. Since any of the partial derivatives above are only obtained by elementary functions through four arithmetic operations and compound, the partial derivatives of any parameter are continuous in its domain. The existence and continuous partial derivative of any parameter can prove that the whole inference system is differentiable.

When the appropriate loss function is selected, the final result loss is differentiable to all the parameters of the model, which provides conditions for using the gradient method to optimize the model parameters. The gradient of the loss result on the parameters of each part of the model can be obtained. The gradient of the loss function on all the rule antecedent attribute parameters is:

$$\nabla_x loss = \begin{bmatrix} \frac{dloss}{dx_1^1} & \cdots & \frac{dloss}{dx_{T_k}^1} \\ \vdots & \ddots & \vdots \\ \frac{dloss}{dx_1^L} & \cdots & \frac{dloss}{dx_{T_k}^L} \end{bmatrix} \tag{32}$$

The gradient of the loss function on the distance-sensitive parameters of all rules is:

$$\nabla_a loss = \begin{bmatrix} \frac{dloss}{da_1^1} & \cdots & \frac{dloss}{da_{T_k}^1} \\ \vdots & \ddots & \vdots \\ \frac{dloss}{da_1^L} & \cdots & \frac{dloss}{da_{T_k}^L} \end{bmatrix} \tag{33}$$

The gradient of the loss function on all rule result attribute parameters is:

$$\nabla_{\beta} loss = \begin{bmatrix} \frac{dloss}{d\beta_{1,1}} & \cdots & \frac{dloss}{d\beta_{N,1}} \\ \vdots & \ddots & \vdots \\ \frac{dloss}{d\beta_{1,L}} & \cdots & \frac{dloss}{d\beta_{N,L}} \end{bmatrix} \tag{34}$$

According to the belief distribution output by the inference system and the loss function of the result, the gradient of each part of the parameters can be optimized by updating the parameters along the negative gradient direction.

C. EXPONENTIAL NORMALIZATION FUNCTION PREPROCESSING

In the training process, in order to meet the restriction that the sum of the result attributes of each rule is one and each result attribute is non-negative, we use the exponential normalization function to preprocess the result attribute parameters during the training process:

$$\beta_j^k = \frac{e^{\overline{\beta_j^k}}}{\sum_{i=1}^{N} e^{\overline{\beta_i^k}}} (\sum_{i=1}^{N} \beta_j^k = 1, \beta_j^k > 0) \tag{35}$$

D. STOCHASTIC GRADIENT DESCENT WITH MOMENTUM OPTIMIZATION

The optimization process of using the gradient descent method to update the parameters of the inference model is given by the following equation:

$$M_{new}(x, a, \beta) = M_{old} - \mu \nabla_{M_{old}} loss \tag{36}$$

The gradient information is given according to the loss function of the final output result, and the learning rate μ is the updated step length information that needs to be set. Chang [11] and Wu [13] used dichotomy in the gradient training process to iteratively find the optimal step size in the constraint space and added perturbation parameters when the gradient is zero to avoid the training process stagnation. For the application of new rule structures, activation methods, and preprocessing steps, the gradient training update step size is no longer limited. We use momentum-optimized stochastic gradient descent method for faster training.

The stochastic gradient descent method is an iterative optimization method when the objective function is differentiable, It used a random subset of the training data to calculate the gradient value as the estimated value of the entire training data gradient, which reduces the computational burden in high-dimensional optimization problems.

The output of the loss function in the conventional gradient descent method is determined by all samples, and the model parameters are updated according to its gradient:

$$loss = \frac{1}{n} \sum_{i}^{n} Loss(\beta_1^i, ..., \beta_N^i) \tag{37}$$

We randomly selected a single sample as the estimated value of the average value of the loss function output on all samples to update the model parameters:

$$loss = Loss(\beta_1^r, ..., \beta_N^r) \tag{38}$$

The obvious disadvantage of the stochastic gradient descent method is that its update direction is completely dependent on the gradient of the current sample and is very unstable. The momentum method is used to solve this problem. The momentum method improves the stability and speed of the training method by retaining a certain degree of historical gradient information and combining it with the current sample gradient. It also enhances the ability to get rid of local optimal solutions.

The momentum method uses a weighted fusion method to synthesize the historical gradient information and the current sample gradient, and use it as an update parameter for this round of training:

$$v_t = \nu v_{t-1} + \mu \nabla_{loss} M, M = M - v_t \tag{39}$$

Initial v_0 is set to zero and set ν to represent the ratio of retaining historical gradient information. The same direction of the current gradient and the historical gradient will increase the speed of parameter training in this gradient direction, The different directions of the current gradient and the historical gradient will inhibit the current gradient from causing parameter training oscillations.

IV. EXPERIMENTAL RESULTS

In this section, we first conduct a comparison test of the training performance of the gradient method under different momentum parameters, and then conduct a comparison test of the performance of the BRB system and conventional machine learning algorithms on public data sets. Finally summarize all the experimental results.

A. EXPERIMENTAL ENVIROMENT

The experiment runs on a Ubuntu 20.04 system equipped with Intel© CoreTM i5 8500@3.0GHz CPU, 16GB RAM and GeForce GTX 1060 Graphics. Use TensorFlow 2.0 to build the evidential reasoning framework of the BRB system and use Scikit-learn machine learning library to collect and clean datasets.

B. PERFORMANCE OF GRADIENT DESCENT METHOD WITH DIFFERENT MOMENTUM PARAMETERS

Liu [16] proved that the confidence rule library can approximate any function. In this section, a nonlinear multi-extreme function is introduced to compare the training performance of the stochastic gradient descent method under different momentum parameters. The nonlinear multi-extreme function is as follows:

$$f(x) = e^{-(x-2)^2} + 0.5e^{-(x+2)^2}, x \in [-5, 5] \tag{40}$$

In the defined domain, 1000 pieces of data are uniformly selected as the fitting data set and the mean square error is

TABLE 1. Initial rule information

rule	antecedent x	sensitive a	result consequent
1	-5	1	$(-1, 1, -1, -1, -1)$
2	-2	1	$(-1, -1, 1, -1, -1)$
3	0	1	$(-1, 1, -1, -1, -1)$
4	2	1	$(-1, -1, -1, 1, -1)$
5	5	1	$(-1, 1, -1, -1, -1)$

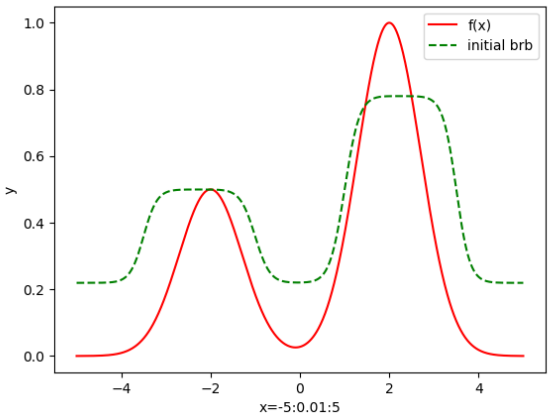

FIGURE 4. untrained BRB system output

used as the loss function. According to each extreme point on the function curve, the rule base result attribute evaluation level and corresponding utility value can be set:

$$\{D_1, D_2, D_3, D_4, D_5\} = \{-0.5, 0, 0.5, 1, 1.5\}$$

Select the five extreme points on the function curve to convert to the corresponding rule structure to build a belief rule library. Set the default distance-sensitive parameter of each rule to 1.0. The initial rule information is shown in Table 1. We can get the initial untrained brb system output Figure 4 Set the number of training samples in each batch to 128 and the learning rate μ is 0.001 for 1000 rounds of training. Set the momentum optimization parameters ν to 0.0 (non-momentum optimization), 0.5, 0.9 and 0.99 to compare their fitting performance.

Figure 5 shows the mean square error loss of each batch under different momentum parameters. Through the decreasing curve of the mean square error loss function, we can find that smaller momentum optimization parameters can not significantly improve the model performance The model with the momentum parameter value of 0.5 has never been below 0.001 during the training process. The value of the mean square error loss function of the model with the momentum parameter value of 0.9 and 0.99 dropped rapidly to 1×10^{-4}, and reached the level of 1×10^{-5} at the end of the training.

Figure 6 clearly shows that higher momentum parameter values can obtain better fitting performance, greatly reducing the distance between the fitted curve and the original curve.

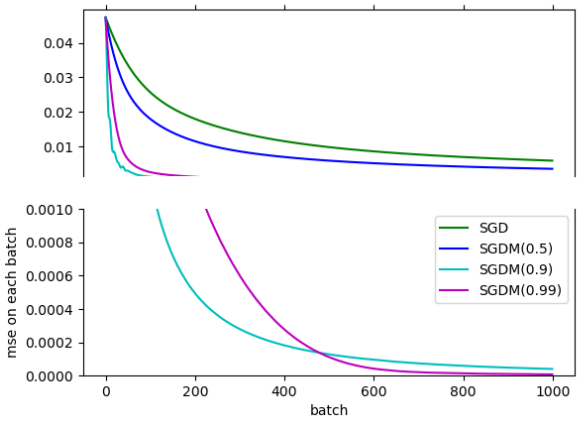

FIGURE 5. mean square error loss under different momentum parameters

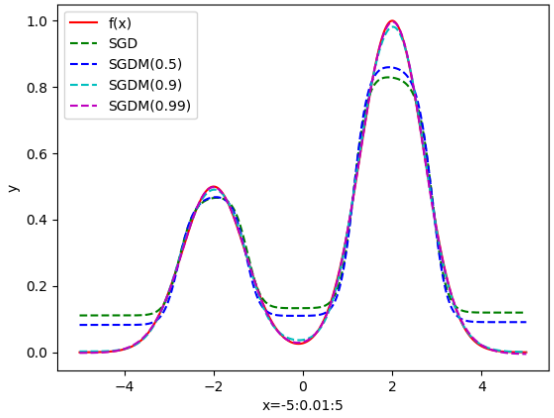

FIGURE 6. fitting performance of different momentum parameter models

Table 2 records the information of each rule in the model with the momentum parameter value of 0.99 after the training is completed. It can be found that all the distance-sensitive parameters have become smaller after training, that is, the activation area of each rule has increased, and the result distribution is significantly increased in their respective utility levels, while other levels have decreased significantly. The antecedent attribute information of each rule, namely the value of x, is only slightly adjusted before and after training.

We compare the performance of the final trained BRB

TABLE 2. Trained rule information

rule	antecedent x	sensitive a	result consequent
1	-4.5	0.20	$(0.10, 1.5, -1.8, -1.7, -1.7)$
2	-2.0	0.44	$(-1.3, -1.5, 0.56, -0.46, -0.47)$
3	0.080	0.50	$(-0.15, 1.4, -1.5, -1.5, -1.6)$
4	2.0	0.49	$(-1.8, -2.0, -1.9, 1.5, 0.59)$
5	4.7	0.21	$(0.12, 1.5, -1.8, -1.7, -1.7)$

TABLE 3. Accuracy and training time using different optimization algorithms

method	mean square error	running time
fmincon	5.16×10^{-5}	496.17
Chen-BRB	6.3228×10^{-5}	-
Wang-BRB	3.9284×10^{-5}	386.63
Li-BRB	3.3322×10^{-5}	357
SGD-BRB	5.90×10^{-3}	11.60
SGDM-BRB(0.5)	3.50×10^{-3}	10.82
SGDM-BRB(0.9)	4.01×10^{-5}	11.09
SGDM-BRB(0.99)	7.40×10^{-6}	13.42

TABLE 4. Details of the classification datasets

Dataset name	#Instances	#Features	#Classes
iris	150	5	3
wine	178	14	3
diabetes	768	9	2
ecoli	336	8	8
glass	214	10	6
seeds	210	8	3
yeast	1484	9	10

system with the previous conventional BRB system. In Table 3, we compare the final mean square error results and training time of these models. It is obvious that the gradient method after the momentum optimization has excellent performance in improving the inference accuracy of the BRB system, but the training model without the effective momentum parameter optimization cannot reach the loss level of the conventional method. And the running speed of gradient method is much lower than the conventional BRB system optimized by intelligent algorithm and other gradient descent algorithm.

C. EXPERIMENT ON PUBLIC CLASSIFICATION DATASETS

In this section, we select 7 UCI public classification datasets that are widely used to test the performance of various classifiers to evaluate the inference performance of our BRB system. Table 4 shows the detailed information of these classification datasets. We repeat independent 10-fold cross-validation experiments for ten times to obtain the final results. The final inference results obtained by the belief rule inference system trained using momentum optimization stochastic gradient method (SGDM-BRB) will be compared with the results obtained by other machine learning methods and belief rule base inference system. The machine methods for comparison include KNN, Naive Bayes (NB) and C4.5, support vector machine (SVM) and their results are cited from previous papers [22], [27]. The BRB systems for comparison include SRA-EBRB [20], MVP-EBRB [21] and BA-EBRB [22].

To enable a unified classification process, first perform a standardized operation on the original data:

$$x' = \frac{x - \overline{x}}{\delta}$$

where $\overline{x}$ is the mean of x and δ is the standard deviation of x. Then we select cross entropy as the loss function to train the BRB inference system to improve the classification accuracy. For a multi-classification task with N categories, the loss function on each sample y and the corresponding prediction result $\overline{y}$ is defined as follows:

$$Loss(y, \overline{y}) = -\sum_{i=1}^{N} y_i \log \overline{y}_i$$

For the experiment on each dataset, 32 training sample data are randomly selected as the initial rule and convert the classification result into the corresponding belief result distribution according to the above conversion method. The rule corresponding to the k-th training sample is:

$$X_k : (x_1, ..., x_{T_k}), Y_k : c, 1 \leq c \leq N \tag{41}$$

$$\begin{aligned} &R_k : if(x'_1, ..., x'_{T_k}) \\ &then\{(\overline{\beta_1^k}, -1), ..., (\overline{\beta_c^k}, 1), ..., (\overline{\beta_N^k}, -1)\} \end{aligned} \tag{42}$$

In the process of parameter optimization using momentum optimization stochastic gradient descent method, the learning rate is set to 0.001, the momentum optimization parameter is 0.99. There are 128 training samples in each batch, and 2000 trainings are performed on the dataset.

As shown in Table 5, by comparing with the accuracies of some machine learning approaches as well as the improved BRB systems, it is proved that the SGDM-BRB system can produce satisfactory accuracies. For the SGDM-BRB system, it is clear from Table 5 that the 96.50% accuracy of dataset iris is worse than the 96.67% accuracies obtained from KNN and SVM, the 85.43% accuracy of dataset ecoli is worse than the 85.71% accuracy obtained from KNN and the 85.61% accuracy obtained from MVP-BRB, and the 75.29% accuracy of dataset pima is worse than the 76.30% accuracy obtained from NB. For the remaining four datasets, the accuracy of the proposed BRB system outperforms all listed studies. Among conventional machine learning methods and other improved BRB systems, SGDM-BRB has the highest average ranking.

V. CONCLUSION

This paper proposes a new rule structure and its activation method and the corresponding momentum optimization gradient training method. Avoid the rule zero activation problem and improve the inference accuracy at the same time, and achieves performance beyond conventional machine learning algorithms and improved BRB system. The further conclusions of this paper are summarized as follow: 1): Using the method of combining attribute distancesensitive parameters and Gaussian function instead of the conventional belief distribution method to calculate the individual matching degree effectively avoids the defect that the conventional method may cause the individual matching degree to be 0. This makes the activation weight change more gentle and has good fitting performance. 2): The stochastic gradient method combined with larger momentum optimization parameters greatly improves the accuracy and convergence speed of the model.

TABLE 5. Accuracy of SGDM-BRB compare with conventional classification approaches

	KNN	NB	C4.5	SVM	SRA-EBRB	MVP-EBRB	BA-EBRB	SGDM-BRB
iris	96.67(**1**)	96.00(4)	96.00(4)	96.67(**1**)	94.80(8)	95.87(6)	95.26(7)	96.50(3)
wine	96.05(5)			96.40(4)	96.85(3)		97.02(2)	97.44(**1**)
diabetes	74.09(3)	76.30(**1**)	73.82(4)	65.10(8)	71.71(7)	72.59(5)	72.32(6)	75.29(2)
ecoli	85.71(**1**)	85.42(4)	84.23(5)	75.60(7)	84.85(5)	85.61(2)		85.43(3)
glass	66.63(7)	48.60(8)	66.82(6)	68.69(5)	73.08(2)	72.06(4)	72.32(3)	74.75(**1**)
seeds	92.38(2)	91.43(6)	91.90(8)	90.48(7)	91.24(7)	92.38(2)	93.95(4)	94.02(**1**)
yeast	58.22(3)	57.61(4)	55.39(7)	43.26(8)	56.85(6)	57.49(5)	58.63(2)	59.49(**1**)
average rank	3.14(2)	4.5(5)	5.66(7)	5.71(8)	5.42(6)	4(3)	4(3)	1.71(**1**)

Due to its good fitting performance, future research will focus on using integrated methods to further improve inference performance and reduce potential over-fitting risks.

REFERENCES

[1] YANG J B, LIU J, WANG J, et al. "Belief rule-base inference methodology using the evidential reasoning approach-rimer," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2006, 2(36):266-285.

[2] SUN R. "Robust reasoning: integrating rule-based and similarity-based reasoning," *Artificial Intelligence*, 1995, 2(75):241-295.

[3] DEMPSTER A P. "A generalization of bayesian inference," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1968, 2(30): 205-232.

[4] SHAFER G, SMITH A F M. "A mathematical theory of evidence[J]," *Biometrics*, 1976, 3(32):703.

[5] YOON K, HWANG C L. "Multiple attribute decision making," *Thousand Oaks, CA: Sage Publications*, 1995.

[6] ZADEH L. "Fuzzy sets," *Information and Control*, 1965, 3(8):338-353.

[7] ZHOU Z J, HU C H, YANG J B, et al. "Online updating belief rule based system for pipeline leak detection under expert intervention," *Expert Systems with Applications*, 2009, 4(36):7700-7709.

[8] JIANG J, LI X, JIE ZHOU Z, et al. "Weapon system capability assessment under uncertainty based on the evidential reasoning approach," *Expert Systems with Applications*, 2011.

[9] YANG Y, FU C, CHEN Y W, et al. "A belief rule based expert system for predicting consumer preference in new product development," *Knowledge-Based Systems*, 2016(94):105-113.

[10] YANG J B, LIU J, XU D L, et al. "Optimization models for training belief-rule-based systems," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2007, 4(37):569-585.

[11] CHANG R, ZHANG S. "An Algorithm for Training Param eters in Belief Rule-bases Based on Gradient Methods with Optimization Step Size," *Journal of North China Institute of Water Conservancy and Hydroelectric Power*, 2011, 1(32):154-157.

[12] CHANG R Y, WANG H, YANG J B. "An algorithm for training parameters in belief rule-bases based on the gradient and dichotomy methods," *Systems Engineering*, 2007.

[13] WU W K, YANG L H, FU Y G, et al. "Parameter Training Approach for Belief Rule Base Using the Accelerating of Gradient Algorithm," *Journal of Frontiers of Computer Science and Technology*, 2014, 8(8):989-1001.

[14] SU Q, YANG L H, FU Y G,et al. "Parameter training approach based on variable particle swarm optimization for belief rule base," *Journal of Computer Applications*, 2014, 34(8):2161-2165.

[15] WANG H J, YANG L H, FU Y G H, et al. "Differential Evolutionary Algorithm for Parameter Training of Belief Rule Base under Expert Intervention," *Computer Science*, 2015, 42(5):88-93.

[16] LIU J, MARTINEZ L, CALZADA A C, et al. "A novel beliefrule base representation, generation and its inference methodology," *Knowledge-Based Systems*, 2013(53):129-141.

[17] ROBBINS H, MONRO S. "A stochastic approximation method," *The Annals of Mathematical Statistics*, 1951, 3(22):400-407.

[18] KIWIEL K C. "Convergence and efficiency of subgradient methods for quasiconvex minimization," *Mathematical Programming*, 2001,1(90):1-25.

[19] RUMELHART D E, HINTON G E, WILLIAMS R J. "Learning representations by back-propagating errors," *Nature*, 1986, 6088(323):533-536.

[20] LIN Y Q, FU Y G. "A rule activation method for extended belief rule base based on improved similarity measures," *Journal of University of Science and Technology of China*, 2018, 48(1):21-27.

[21] LIN Y Q, FU Y G, SU Q, et al. "A rule activation method for extended belief rule base with vp-tree and mvp-tree," *Journal of Intelligent and Fuzzy Systems*, 2017, 33(6):3695-3705.

[22] FANG W, GONG X, LIU G, et al. "A balance adjusting approach of extended belief-rule-based system for imbalanced classification problem," *IEEE Access*, 2020, 8(19419049):41201-41212.

[23] SHANNON C E. "A mathematical theory of communication," *Bell System Technical Journal*, 1948, 4(27):623-656.

[24] CALZADA A, LIU J, WANG H, et al. "A new dynamic rule activation method for extended belief rule-based systems," *IEEE Transactions on Knowledge and Data Engineering*, 2015, 4(27):880-894.

[25] YANG L H. "New activation weight calculation and parameter optimization for extended belief rule-based system based on sensitivity analysis," *Knowledge and Information Systems*, 2018, 60(2):837-878.

[26] WANG Y. "Parameter learning for an intuitionistic fuzzy belief rulebased systems based on weight and reliability," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 2019, 23(2):219-228.

[27] ZHU H. "A minimum centre distance rule activation method for extended belief rule-based classification systems," *Applied Soft Computing*, 2020, 91:106214.

[28] JIA Q. "A novel fault detection model based on atanassov's intervalvalued intuitionistic fuzzy sets, belief rule base and evidential reasoning," *IEEE Access*, 2020, 8:4551-4567.

[29] G. Holmes, A. Donkin, I.H. Witten. "WEKA: A machine learning workbench, in: Conference on Intelligent Information Systems," 2002.

FIRST A. AUTHOR (M'76–SM'81–F'87) and all authors may include biographies. Biographies are often not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author's educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year the degree was earned. The author's major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author's last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (publisher name, year) similar to a reference. Current and previous research interests end the paragraph. The third paragraph begins with the author's title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, it should be of good quality, and professional-looking. Following are two examples of an author's biography.

SECOND B. AUTHOR was born in Greenwich Village, New York, NY, USA in 1977. He received the B.S. and M.S. degrees in aerospace engineering from the University of Virginia, Charlottesville, in 2001 and the Ph.D. degree in mechanical engineering from Drexel University, Philadelphia, PA, in 2008.

From 2001 to 2004, he was a Research Assistant with the Princeton Plasma Physics Laboratory. Since 2009, he has been an Assistant Professor with the Mechanical Engineering Department, Texas A&M University, College Station. He is the author of three books, more than 150 articles, and more than 70 inventions. His research interests include high-pressure and high-density nonthermal plasma discharge processes and applications, microscale plasma discharges, discharges in liquids, spectroscopic diagnostics, plasma propulsion, and innovation plasma applications. He is an Associate Editor of the journal *Earth, Moon, Planets*, and holds two patents.

Dr. Author was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, and the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011.

THIRD C. AUTHOR, JR. (M'87) received the B.S. degree in mechanical engineering from National Chung Cheng University, Chiayi, Taiwan, in 2004 and the M.S. degree in mechanical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2006. He is currently pursuing the Ph.D. degree in mechanical engineering at Texas A&M University, College Station, TX, USA.

From 2008 to 2009, he was a Research Assistant with the Institute of Physics, Academia Sinica, Tapei, Taiwan. His research interest includes the development of surface processing and biological/medical treatment techniques using nonthermal atmospheric pressure plasmas, fundamental study of plasma sources, and fabrication of micro- or nanostructured surfaces.

Mr. Author's awards and honors include the Frew Fellowship (Australian Academy of Science), the I. I. Rabi Prize (APS), the European Frequency and Time Forum Award, the Carl Zeiss Research Award, the William F. Meggers Award and the Adolph Lomb Medal (OSA).