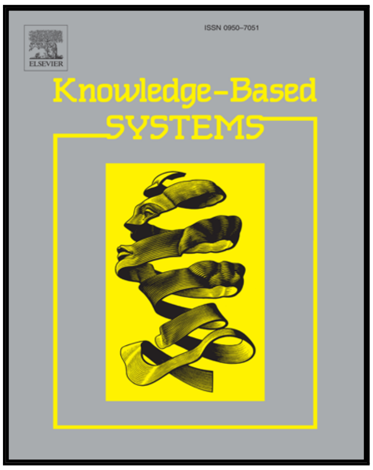
Data Classification using Evidence Reasoning Rule

DOI:
10.1016/j.knosys.2016.11.001

Citation for published version (APA):
Xu, X., Zheng, J., Yang, J-B., Xu, D., & Chen, Y-W. (2017). Data Classification using Evidence Reasoning Rule. *Knowledge-Based Systems*, *116*, 144-151. https://doi.org/10.1016/j.knosys.2016.11.001

Published in:
Knowledge-Based Systems



Data Classification using Evidence Reasoning Rule

Xiaobin Xu , Jin Zheng , Jian-bo Yang , Dong-ling Xu ,
Yu-wang Chen

PII: S0950-7051(16)30428-2
DOI: 10.1016/j.knosys.2016.11.001
Reference: KNOSYS 3725

To appear in: *Knowledge-Based Systems*

Received date: 16 June 2016
Revised date: 21 August 2016
Accepted date: 1 November 2016

Data Classification using Evidence Reasoning Rule

Xiaobin Xu[1], Jin Zheng [1], Jian-bo Yang [2], Dong-ling Xu[2], Yu-wang Chen[2]

[1] Institute of System Science and Control Engineering, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China; [2]Decision and Cognitive Sciences Research Centre, The University of Manchester, Manchester M15 6PB, UK.

Abstract- In Dempster-Shafer evidence theory (DST) based classifier design, Dempster's combination (DC) rule is commonly used as a multi-attribute classifier to combine evidence collected from different attributes. The main aim of this paper is to present a classification method using a novel combination rule i.e., the evidence reasoning (ER) rule. As an improvement of the DC rule, the newly proposed ER rule defines the reliability and weight of evidence. The former indicates the ability of attribute or its evidence to provide correct assessment for classification problem, and the latter reflects the relative important of evidence in comparison with other evidence when they need to be combined. The ER rule-based classification procedure is expatiated from evidence acquisition and estimation of evidence reliability and weight to combination of evidence. It is a purely data-driven approach without making any assumptions about the relationships between attributes and class memberships, and the specific statistic distributions of attribute data. Experiential results on five popular benchmark databases taken from University of California Irvine (UCI) machine learning database show high classification accuracy that is competitive with other classical and mainstream classifiers.

Keyword- Date classification, Dempster-Shafer evidence theory (DST), Evidential reasoning (ER) rule, Reliability and weight of evidence, Sequential linear programming (SLP)

1. Introduction

Classification problem is one of the most important issues in data mining and knowledge discovery [1]. Its purpose is to fall a sample with unknown class into a basket with a label of specific class where an appropriate classifier should be used to analyze the attributes of this sample. Classification are fundamental to many theoretical and practical applications, including pattern recognition [2-4], fault diagnosis [5-7], and image processing[8-10], etc. Many well-known methods have been proposed to solve classification problems, including k nearest neighbors (k-NN) [11], support vector machine [12], naive Bayes [13], Bayes net [14], decision tree learner [15], random forest [16], and other latest techniques, such as gravitational inspired classifier [17], feature vector graph-based classifier [18], and Learning Automata(LA)-based classifier[19] , and so on.

From the perspective of uncertain information processing, the imprecision or even incorrectness of classification results is likely to be caused by the fact that the values of attributes of a sample cannot categorically point to a certain class, that is to say, the boundaries of attributes among different classes are commonly imprecise, or even overlapping[20-22]. As a result, Dempster-Shafer evidence theory (DST) can provide an available mechanism to deal with the classification imprecision. In detail, a frame of discernment (FoD) needs to be firstly determined which includes all preassigned class memberships. The next step is to obtain a basic belief assignment (BBA), i.e., a belief distribution (BD) function, in which the belief degrees are used to measure the extents to which the attributes of a sample supports each class and the subsets of the classes. Such a BBA or a BD can be also named as a piece of evidence. There are different ways for generating BBAs from different types of attribute information. The typical ways include core sample [20], neural network [21], k-NN [22], expert system [23] and so on. The final step is to use Dempster's combination (DC) rule to fuse these BBAs and then make a classification decision according to the fused results. The aim of combination is to reduce the classification imprecision by fusing multi-source attribute information.

Recently, the evidential reasoning (ER) rule has been established to advance the seminal Dempster-Shafer evidence theory [24-28] and the original ER algorithm [29-32]. Compared with the DC rule, the main advance of the ER rule is to propose a novel concept of weighted evidence (WE) and extend to WE with Reliability (WER) in order to characterize evidence in complement of BBA or BD introduced in the DST. As a result, the implementation of the orthogonal sum operation on WEs and WERs leads to the establishment of the new ER rule [33]. The most important property of the ER rule is that it constitutes a generic conjunctive probabilistic reasoning process, or a generalized Bayesian inference process which can be implemented on the power set of FoD. It has been proved that (1) the DC rule is a special case of the ER rule when each piece of evidence is fully reliable, and (2) the original ER algorithm is also a special case when the weights of all pieces of evidence being normalized are equal to their respective reliabilities [33]. The evidence reasoning procedure consists of the belief structure for modelling various types of uncertainty [34-35], the rule and utility based information transformation techniques [30], and the ER algorithm for information aggregation [31], etc. In the past twenty years, the ER algorithm has been widely applied to many system and decision analysis problems as surveyed by Xu [35]. Furthermore, the ER algorithm has been introduced to extend traditional If-Then rule based systems to belief rule based (BRB) systems [28]. The BRB methodology employs the informative belief structure to represent various types of information and knowledge with uncertainties and shows the capability of approximating any linear and nonlinear relationships across a wide variety of application areas. Recently, the BRB also have been applied for solving classification problem in [36-38]. However, a problem of BRB is the high multiplicative complexity on the number of referential values of attributes in the belief rule base [28].

Given that the ER rule has explicitly generalised the DST and the original ER algorithm, it becomes perfectly logical and also extremely important to revisit and further extend those techniques which were previously developed from the latter two methods. This paper presents a novel classification method using the ER rule. In detail, the likelihoods of class membership for the referential values of each attribute can be calculated by statistical analysis on training samples with known classes, and then the evidence for each attribute can be acquired by the normalization of likelihoods; the reliability of evidence can be estimated by analyzing the classifying ability of each attribute; an optimization model using sequential linear programming (SLP) is proposed to obtain the optimal weight and referential values of each attribute; finally, the ER rule is used to combine the pieces of evidence provided by all attributes of a sample and then make a classification decision according to the fused results.

The rest of the paper is organised as follows: Section 2 briefly introduces the concepts and properties of the ER rule. Section 3 details the ER rule-based classification method. In Section 4, an experiment on the well-known Iris database shows the specific procedure of the proposed method, and then it is compared with other six classical classifiers to demonstrate its superiority by using five popular benchmark databases taken from University of California Irvine (UCI) machine learning database. Some concluding remarks are presented in Section 5.

2. Outline of the ER rule

In this section, the ER rule [33,39] is briefly introduced. Suppose Θ={h_1,h_2,…,h_N} is a set of mutually exclusive and collectively exhaustive hypotheses. Θ is referred to as a frame of discernment. The power set of Θ consists of all its subsets, denoted by P(Θ) or 2^{Θ}. A piece of evidence is profiled by a belief distribution as follows

$$e_j = \{(\theta, p_{\theta,j}) \mid \forall \theta \subseteq \Theta, \textstyle\sum_{\theta\subseteq\Theta} p_{\theta,j} = 1\} \quad (1)$$

where (θ,$p_{\theta,j}$) is an element of evidence e_j, representing that the evidence points to proposition θ with the degree of $p_{\theta,j}$ referred to as probability or degree of belief in general. θ can be any subset of Θ or any element of P(Θ) except for the empty set. (θ,$p_{\theta,j}$) is referred to as a focal element of e_j if $p_{\theta,j}>0$.

In the ER rule, reliability r_j and weight w_j of evidence e_j are defined. Reliability r_j represents the ability of the information source, where e_j is generated, to provide correct assessment or solution for a given problem. The reliability of a piece of evidence is the inherent property of the evidence. The weight w_j of evidence can be used to reflect its relative importance in comparison with other evidence and determined according to who uses the evidence. This means that weight w_j can be subjective and different from reliability r_j in situations where different pieces of evidence are generated from different sources and measured in different ways. A so-called weighted belief distribution with reliability can be defined as follows

$$m_j = \{(\theta, \tilde{m}_{\theta,j}) \mid \forall \theta \subseteq \Theta; (P(\Theta), \tilde{m}_{P(\Theta),j})\} \quad (2)$$

where $\tilde{m}_{\theta,j}$ measures the degree of support for θ from e_j with both the weight and reliability of e_j taken into account, defined as follow

$$\tilde{m}_{\theta,j} = \begin{cases} 0 & \theta = \varnothing \\ c_{rw,j} m_{\theta,j} & \theta \subseteq \Theta, \theta \neq \varnothing \\ c_{rw,j}(1 - r_j) & \theta = P(\Theta) \end{cases} \quad (3)$$

where $m_{\theta,j} = w_j p_{\theta,j}$, $c_{rw,j} = 1/(1 + w_j - r_j)$ is a normalization factor, which is uniquely determined to satisfy $\sum_{\theta\subseteq\Theta} \tilde{m}_{\theta,j} + \tilde{m}_{P(\Theta),j} = 1$ given that $\sum_{\theta\subseteq\Theta} p_{\theta,j} = 1$. Compared with Shafer's discounting method, the critical difference is that in the ER rule, the degree of residual support (1-r_j) defined as the unreliability of evidence is earmarked to the power set for redistribution instead of assigning it specifically to the frame of discernment. That is because $p_{\Theta,j}$ is the inner characteristics of e_j, so it should be equally discounted by $c_{rw,j}$ as the other propositions θ. e_j and m_j will keep the same probability characteristics according to this operation. Based on the above definition, the degree of residual support of a piece of evidence reflects the unreliability of the evidence.

If two pieces of evidences e_1 and e_2 are independent, the combined degree of belief to which e_1 and e_2 jointly support proposition θ, denoted by $p_{\theta,e(2)}$, can be generated by the ER fusion rule as follows

$$p_{\theta,e(2)} = \begin{cases} 0 & \theta = \varnothing \\ \dfrac{\hat{m}_{\theta,e(2)}}{\sum_{D\subseteq\Theta}\hat{m}_{D,e(2)}} & \theta\subseteq\Theta, \theta\neq\varnothing \end{cases} \tag{4}$$

$$\hat{m}_{\theta,e(2)} = [(1-r_2)m_{\theta,1} + (1-r_1)m_{\theta,2}] + \sum_{B\cap C=\theta} m_{B,1}m_{C,2} \quad \forall\theta\subseteq\Theta$$

The recursive formulae of the ER rule to combine multiple pieces of evidence in any order are also given in [33], where it is proven that Dempster's rule is a special case of the above ER rule when each piece of evidence e_j in question is assumed to be fully reliable, or r_j=1 for all j.

3. The data-driven ER methodology for Data Classification

This section aims to develop an ER-based classifier for such an N-class classification problem, where a sample data set including K samples can be identified by M attributes $\boldsymbol{x} = \{x_1, x_2, ..., x_M\}$. This sample set has to be classified in $\Theta=\{y_1,y_2,..., y_n,\ldots, y_N\}$, y_n is the n^{th} class membership. The inputs of the ER-based classifier are the M attributes of a sample, and the output is the estimated class membership which this sample belongs to. The detailed modelling and estimating procedure is described as follows.

3.1 Acquiring evidence from the attribute data

Our previous work [39] explored the relationship between Bayes' rule in statistical inference and the ER rule for conjunctive combination of independent evidence, and found that the normalisation of likelihoods in Bayesian paradigm results in the equivalent evidence in the ER paradigm with the evidential meaning of data kept intact in the process. Here, we use the normalisation of likelihoods to acquire the evidence from the training sample set

$$S=\{[\boldsymbol{x}^k,y_k] \mid \boldsymbol{x}^k = (x_1^k,...,x_i^k,...,x_M^k), x_i^k \in S_i, y_k\in\Theta, k=1, 2,\ldots,K_s\}$$

here, $K_s \le K$, S_i is the value domain of x_i.

Firstly, the relationship between attribute x_i and class y needs to be approximatively transformed into the relationships between the referential values $\boldsymbol{A}_i=\{A_j^i \mid j=1,\ldots,J_i\}$ of x_i and class y. Here, as adjustable parameters, the initial values of A_j^i can be given according to expert knowledge or random rule without any prior knowledge, and subsequently trained using sample data under a certain optimization target; secondly, for a specific value x_i^k, its similarity distribution $S_I(x_i^k)$ about the referential values $\boldsymbol{A}_i$ can be generated by the following information transformation technique [30]:

$$S_I(x_i^k) = \{(A_j^i, \alpha_{i,j}) \mid j=1,...,J_i; i=1,...,M\} \tag{5a}$$

where

$$\alpha_{i,j} = \frac{A_{j+1}^i - x_i^k}{A_{j+1}^i - A_j^i},\ \alpha_{i,j+1} = 1 - \alpha_{i,j} \qquad A_j^i \le x_i^k \le A_{j+1}^i \tag{5b}$$

$$\alpha_{i,j'} = 0 \qquad j' = 1,..., J_i,\ j' \ne j, j+1 \tag{5c}$$

$\alpha_{i,j}$ represents the similarity degree to which x_i^k matches the referential value A_j^i.

Hence, a sample pair (x_i^k,y_k) in the set S can be transformed and uniquely represented as a similarity distribution $(\alpha_{i,j}, \alpha_{i,j+1})$ for the class y_k. Table 1 shows the statistical result of casting all sample pairs in S in the form of the similarity degree. Here, $a_{n,j}$ is the sum of the similarity degrees of the sample pairs whose attribute values, e.g. x_i^k, match the referential value A_j^i and also belong to the class y_n. $\delta_n = \sum_{j=1}^{J_i} a_{n,j}$ is the sum of the similarity degrees of the sample pairs that belong to class y_n. $\eta_j = \sum_{n=1}^{N} a_{n,j}$ is the sum of the similarity degrees of the sample pairs whose attribute values, e.g. x_i^k, match A_j^i. Obviously, $\sum_{n=1}^{N} \delta_n = \sum_{j=1}^{J_i} \eta_j = K_s$.

Table 1 The casting result of sample pairs (x_i, y) on attribute x_i

y_k \ x_i	A_1^i	…	A_j^i	…	$A_{J_i}^i$	Total
y_1	$a_{1,1}$	…	$a_{1,j}$	…	a_{1,J_i}	δ_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
y_n	$a_{n,1}$	…	$a_{n,j}$	…	a_{n,J_i}	δ_n
⋮	⋮	⋮	⋮	⋮	⋮	⋮
y_N	$a_{N,1}$	…	$a_{N,j}$	…	a_{N,J_i}	δ_N
Total cast	η_1	…	η_j	…	η_{J_i}	K_s

According to Table1, we can construct the likelihood (denoted as $c_{n,j}$) to which x_i is identified as A_j^i given the known class y_n as follows:

$$c_{n,j} = p(A_j^i \mid y_n) = \frac{a_{n,j}}{\delta_n} \tag{6}$$

Thereupon, a piece of evidence e_j^i with the weight w_j^i corresponding to A_j^i can be defined as shown in Table 2.

Table 2 The belief matrix of the input attribute x_i

x_i / y_k	e_1^i	$\ldots$	e_j^i	$\ldots$	$e_{J_i}^i$
	A_1^i	$\ldots$	A_j^i	$\ldots$	$A_{J_i}^i$
y_1	$\beta_{1,1}^i$	$\ldots$	$\beta_{1,j}^i$	$\ldots$	β_{i,J_i}^i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$\beta_{n,1}^i$	$\ldots$	$\beta_{n,j}^i$	$\ldots$	β_{n,J_i}^i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_N	$\beta_{N,1}^i$	$\ldots$	$\beta_{N,j}^i$	$\ldots$	β_{N,J_i}^i

In Table 2, the degree of belief $\beta_{n,j}^i$ about e_j^i can be calculated by the normalization of likelihoods $c_{1,j}$, $c_{2,j}$,…, $c_{N,j}$

$$\beta_{n,j}^i = \frac{c_{n,j}}{\sum_{k=1}^{N} c_{k,j}} \tag{7}$$

More precisely, $\beta_{n,j}^i$ is the probability that a sample is believed to belong to the class y_n given that the input attribute x_i takes the referential value A_j^i. Thus, Table 2 can be regarded as a belief matrix for characterizing the relationship between the attribute x_i and the class y_n .

3.2 Evaluating the reliability of evidence

The reliability of evidence represents the classifying ability of its corresponding attributes. Apparently, the more reliable the attribute x_i is, the more samples can be classified by it individually. That is to say, there are not or few overlaps among the intervals of the attribute values for the different class memberships. Hence, the reliability of information source x_i can be defined as follows

$$r_i = \frac{Q_i}{\max\limits_{l, l\in\{1,2,..,M\}}(Q_l)} \tag{8}$$

where Q_i is the number of such samples that can be directly identified as a specific class by the attribute $x_{i.}$ Equation (8) implies that if $\max\limits_{l, l\in\{1,2,...,M\}}(Q_l) = Q_i$ then x_i is the most reliable attribute (r_i=1). The reliabilities of the other attributes are measured by comparing with the most reliable one.

The calculation of the reliability can be demonstrated by a typical three-class case as shown in Fig.1. Here, the maximum and minimum values of the attribute x_i for the classes C_s (s=1, 2, 3) need to be determined as $\max^s$ and $\min^s$ respectively by sample statistics, and then these boundary points divide the value domain of x_i into three single-class intervals ([$\min^1$, $\min^2$] for C_1, [$\max^1$, $\min^3$] for C_2, [$\max^2$, $\max^3$] for C_3) and two double-class overlapping intervals ([$\min^2$, $\max^1$] for C_1&C_2, [$\min^3$,$\max^2$] for C_2&C_3). If the number of samples in the five intervals are q_1, q_2, q_3 and o_1, o_2

respectively, then Q=q_1+q_2+q_3 for the attribute x_i. It means that the Q samples can be clearly identified by x_i, but O=o_1+o_2 samples are confused and hardly classified by x_i only.

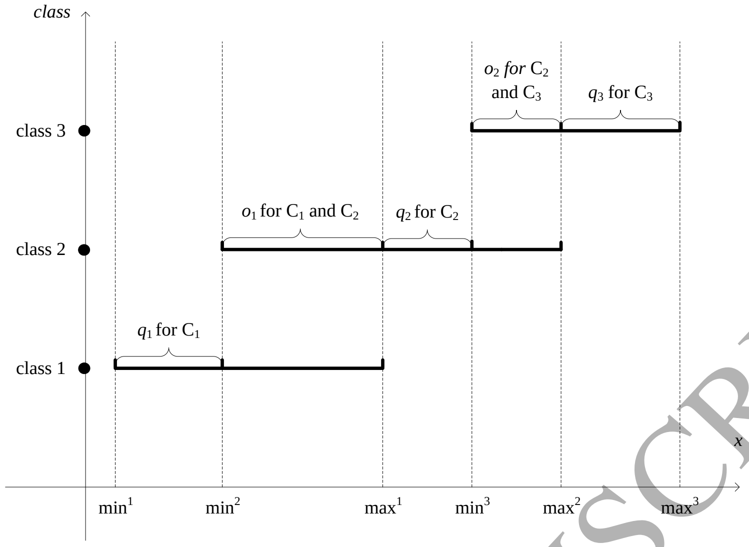


Fig. 1. The distribution of samples about a typical three-class case for the attribute x

3.3 Combination of activated evidence

For a certain sample with M attribute values $\boldsymbol{x}^k = (x_1^k, ..., x_i^k, ..., x_M^k)$, x_i^k of $\boldsymbol{x}^k$ will activate the adjacent two pieces of evidence e_j^i with the weight w_j^i and e_{j+1}^i with the weight w_{j+1}^i if x_i^k takes value in interval $[A_j^i, A_{j+1}^i]$. Thus, the piece of evidence e_i of x_i^k with the belief degree $p_{n,i}$ can be calculated as the weighted sum of e_j^i, e_{j+1}^i

$$e_i=\{(y_n,p_{n,i}),n=1,...,N\} \tag{9}$$

$$p_{n,i} = \alpha_{i,j}\beta_{n,j}^i + \alpha_{i,j+1}\beta_{n,j+1}^i \tag{10}$$

here, $p_{n,i}$ denotes the belief degree to which the class is believed to be y_n given that x_i^k activates e_j^i and e_{j+1}^i. Accordingly, in the same way, the weight w_i of e_i can be similarly calculated as the weighted sum of w_j^i, w_{j+1}^i

$$w_i = \alpha_{i,j}w_j^i + \alpha_{i,j+1}w_{j+1}^i \tag{11}$$

After obtaining all M pieces of evidence e_1,e_2,...,e_M about M attribute values by Equations (9) and (10), we can use the ER rule in Equation (4) to combine them with weights and reliabilities to yield the following fused result

$$O(\boldsymbol{x}^k)=\{(y_n, p_{n,e(M)}), n=1,\ldots,N\} \tag{12}$$

The initial weight w_j^i of e_j^i can be set as its reliability r_i, because it is believed that the evidence with high reliability should be of relatively high importance in comparison to other evidence. Of course, it needs to be trained through data-driven optimization method which will be discussed in the following section. According to the fused result $O(\boldsymbol{x}^k)$, we can estimate that the sample $\boldsymbol{x}^k$ belongs to such class that has the maximum degree of belief.

As a result, the initial ER rule-based classifier is constructed by the above three subsections.

3.4 Training of ER-based classifier parameters using SLP

3.4.1 Optimization model for the ER-based classifier

So far, we have constructed the ER rule-based classifier with the initial parameters including the attribute referential values $\boldsymbol{A}_i$={ A_j^i |j=1,…,J_i} and the weights W={ w_j^i |i=1,..,M; j=1,…,J_i}. However, the initial ER-based classifier may not accurately model the complex causal relationship between the attribute x_i (i=1,..,M) and the actual classes due to the assumed initial values of the above parameters. Therefore, it is extremely important to train these parameters using sample dataset S so that the performance of the classifier can be improved. Here, an optimization model based on mean squared error (MSE) is presented as

$$\min\nolimits_{\boldsymbol{P}} \xi(\boldsymbol{P}) \tag{13a}$$

Here, the objective function

$$\xi(\boldsymbol{P})=\sum\nolimits_{k=1}^{T_s} d_E(O(\boldsymbol{x}^k), V^k) \tag{13b}$$

$\boldsymbol{P}$={ $A_{j'}^i$, w_j^i | i=1,...,M; j' =2,...,J_i-1; j=1,...,J_i} stands for the parameters to be optimized, the remaining parameters A_1^i and $A_{J_i}^i$ are given as $\min\limits_{k,k\in S_{x_i}}(x_i^k)$, $\max\limits_{k,k\in S_{x_i}}(x_i^k)$ respectively, because they are all fixed boundary values. d_E stands for the Euclidean distance between the fused result $O(\boldsymbol{x}^k)=(p_{1,e(M)}, p_{2,e(M)},\ldots, p_{N,e(M)})$ (the belief vector form of Equation (12)) and the reference belief vector V^k with the categorical belief degree assigned to the class y_k that $\boldsymbol{x}^k$ actually belongs to. For example, for a typical three-class case, if $\boldsymbol{x}^k$ actually points to y_2, then the corresponding reference vectors is V^k=(0,1,0). Equations (14a) and (14b) represent the bound constraints which the adjustable parameters need to satisfy

$$0 \le w_j^i \le 1\,, i=1,\ldots,M,\ j=1,\ldots,J_i \tag{14a}$$

$$A_{j'-1}^i < A_{j'}^i < A_{j'+1}^i\ ,\ j'=2,\ldots,J_i-1 \tag{14b}$$

In the following section, the sequential linear programming (SLP) method will be introduced to solve this optimization problem. With the optimization of the parameters ***P***, the belief matrix in Table 2 will also reach to optimal values.

3.4.2 SLP for training parameters of the classifier

SLP is initially known as a method of approximation programming [40]. It is also one of the easiest strategies for solving nonlinear optimization problems. The guiding principle of this strategy is to approximate a nonlinear program by a series of linear programs using first-order Taylor series expansions.

The kernel of the SLP is the linear programming solver, which is easily available with the development of Simplex [41] and Interior-point methods [42]. The linearly approximated model is easily constructed since the required first-order derivatives can be easily obtained by using analytical methods or finite difference methods [43]. It avoids the complexity associated with deriving expressions for high-order derivatives. the advantage and disadvantage of the SLP are analyzed in detail in reference [44]. The main steps of the optimization process using SLP are outlined as follows:

Step 1: Calculation of first-order gradients of optimization objective function.

According to the optimization model of the ER classifier in Subsection 3.4.1, the first-order derivations of the objective function $\xi(\boldsymbol{P})$ to the parameters $\boldsymbol{A}_i$ and w_i need to be calculated respectively so that it can be linearized as follows

$$\xi(\boldsymbol{P}) \approx \xi(\boldsymbol{P}_0) + \xi'(\boldsymbol{P}_0)(\boldsymbol{P} - \boldsymbol{P}_0) \tag{15}$$

where $\boldsymbol{P}_0$ stands for a given initial point. Therefore, the nonlinear optimization problem $\min_P \xi(\boldsymbol{P})$ is converted into such a linear programming problem $\min_P \xi'(\boldsymbol{P}_0)\boldsymbol{P}$.

Step 2: Determination of move limits.

The proper determination of move limits is critical for the successful implementation of the technique. The performance of SLP is very sensitive to the definition of proper move limits for all variables. Large move limits may cause inaccuracies and prevent convergence; on the other hand, small move limits may lead to a large number of iterations and excessive computational efforts. A variety of techniques have been proposed to define the move limits [45-49]. In the context of classification problems, a common way is adopted to determine the move limits [44]. Firstly, the upper bounds of adjustable parameters *UB*(***P***) can be acquired from Equations (14a)-(14b) as follows:

$$UB(w_i) = 1,\, i = 1, ..., M \tag{16a}$$

$$UB(A_j^i) = A_{J_i}^i,\; j = 1, ..., J_i \tag{16b}$$

Then, the initial move limits are set to be 10% of the above upper bounds.

Step 3: Acquisition of the optimal solution using linear programming.

After the implementation of **Step 1** and **Step 2**, the nonlinear objective function $\xi(\boldsymbol{P})$ can be linearized at a given initial point $\boldsymbol{P}_0$, around the point, and a search space is established using the initial move limits of all variables. Therefore, linear programming technology (such as Interior-point methods) can be adopted for this search process. If the intersection of the established search space and the linearized feasible space is empty, then the move limits need to be increased for expanding the search space. If the intersection is not empty, the optimal solution of the linearized programming problem will be searched. The obtained optimal solution is subsequently used as a new basic point to re-linearize the original nonlinear programming problem. The process is repeated recursively until some stopping criterion is satisfied.

Step 4: Stopping criteria.

The SLP iteration process will be stopped if a) the move limits of all variables have been reduced to be significantly small, or b) the values of both the variables and the objective function are not significantly different in two successive iterations.

4. Experiments

This section includes some experimental results to demonstrate the proposed method for pattern classification tasks. One representative example is firstly analyzed to show the specific implementing procedure of this classifier, followed by a comparison study with other well-known classifier using five popular benchmark databases.

4.1 The ER-based classifier for Iris data classification

In this section, we conducted an experiment on the Iris dataset [50], which is a well-known benchmark dataset in pattern classification. The Iris data involve classification of three classes of the Iris flowers, namely, Iris Setosa (y_1), Iris Versicolour (y_2) and Iris Virginica (y_3). Each class of Iris flowers has 50 samples with four attributes: sepal length (x_1) in centimeter (cm), sepal width (x_2) in cm, petal length (x_3) in cm, and petal width (x_4) in cm.

Firstly, 30 of 50 samples in each class are randomly selected to construct the training sample set $S=\{[x_i^k, y_k] | x_i^k \in S_i, y_k \in \Theta, k=1, 2, \ldots, K_s, K_s=90, i=1,...,4\}$, in which, $S_1=[4.3, 7.9]$, $S_2=[2, 4.4]$, $S_3=[1, 6.9]$, $S_4=[0.1, 2.5]$; the remaining 60 samples are used to test. According to experiences, we initialize referential points $\boldsymbol{A}_1=\{4, 5, 5.5, 6.5, 7, 8\}$ for x_1, $\boldsymbol{A}_2=\{2, 2.5, 3, 3.5, 4, 4.5\}$ for x_2, $\boldsymbol{A}_3=\{1, 3, 4, 5, 6, 7\}$ for x_3, and $\boldsymbol{A}_4=\{0, 0.5, 1, 1.5, 2, 2.5\}$ for x_4. For these training samples, the initial casting results and the belief matrixes can be acquired successively according to the information transformation technique and normalization of likelihoods respectively as introduced in Subsection 3.1.

Secondly, the numbers of samples that can be directly identified by single attribute are calculated as $P_1=28$, $P_2=7$, $P_3=113$, $P_4=112$, and then the reliability evaluation method in Subsection 3.2 is used to calculate the reliability of the evidence provided by each attribute as $r_1=0.2478$, $r_2=0.0619$, $r_3=1$, $r_4=0.9912$, respectively. As for the weights w_j^i ($i=1,..,4$; $j=1,...,J_i$) of the evidence provided by x_i, its initial values are given as $w_j^i=r_i$ as it is believed that the evidence with high reliability should be of relatively high importance in comparison to other evidence before the training of parameters.

Thirdly, for given attribute values of a sample from the training set *S*, its initial estimated class can be acquired by combining the activated evidence using the ER rule as described in Subsection 3.3. According to the MSE-based optimization method in Subsection 3.4, the optimal parameters ***P*** can be obtained by using the training samples. Furthermore, corresponding to these training samples, Tables 3-10 list the trained casting results and belief matrixes respectively for the four attributes. Table 11 lists the trained evidence weights of the four attributes.

Table 3 The trained casting result of the training sample pairs (x_1^k ,y_k)

x_1 y	A_1^1	A_2^1	A_3^1	A_4^1	A_5^1	A_6^1	Total
	4	4.6401	5.6747	6.4684	7.1035	8	
1	1.4696	17.2586	11.05	0.2218	0	0	30
2	0	2.1239	17.3805	9.037	1.4586	0	30
3	0	0.821	7.7636	13.9708	5.3447	2.1	30
Total	1.4696	20.2035	36.194	23.2295	6.8033	2.1	90

Table 4 The trained casting result of the training sample pairs (x_2^k ,y_k)

x_2 y	A_1^2	A_2^2	A_3^2	A_4^2	A_5^2	A_6^2	Total
	2	2.2992	2.8258	3.0674	3.4268	4.5	
1	0	0	1.8091	7.1467	16.0577	4.9866	30
2	0.6632	6.2927	12.4621	8.8282	1.7539	0	30
3	0	3.9635	12.9898	7.0892	5.1007	0.8569	30
Total	0.6632	10.2561	27.261	23.064	22.9123	5.8434	90

Table 5 The trained casting result of the training sample pairs (x_3^k ,y_k)

x_3 y	A_1^3	A_2^3	A_3^3	A_4^3	A_5^3	A_6^3	Total
	1	2.6994	4.1406	5.0473	6.0123	7	
1	21.644	8.356	0	0	0	0	30
2	0	3.2522	19.1192	7.574	0.0546	0	30
3	0	0	1.5781	16.5446	10.611	1.2663	30
Total	21.644	11.6081	20.6973	24.1186	10.6656	1.2663	90

Table 6 The trained casting result of the training sample pairs (x_4^k ,y_k)

x_4 y	A_1^4	A_2^4	A_3^4	A_4^4	A_5^4	A_6^4	Total
	0	0.3982	1.0733	1.5223	2.0049	2.5	
1	11.4421	18.0971	0.4607	0	0	0	30
2	0	0.5428	13.74	14.4515	1.2657	0	30
3	0	0	0.322	6.3992	16.7324	6.5464	30
Total	11.4421	18.6399	14.5227	20.8507	17.9982	6.5464	90

Table 7 The trained belief matrix of the attribute x_1

x_1 y	e_1^1	e_2^1	e_3^1	e_4^1	e_5^1	e_6^1
	A_1^1	A_2^1	A_3^1	A_4^1	A_5^1	A_6^1
	4	4.6401	5.6747	6.4684	7.1035	8
1	1	0.8542	0.3053	0.0095	0	0
2	0	0.1051	0.4802	0.389	0.2144	0
3	0	0.0406	0.2145	0.6014	0.7856	1

Table 8 The trained belief matrix of the attribute x_2

x_2 y	e_1^2	e_2^2	e_3^2	e_4^2	e_5^2	e_6^2
	A_1^2	A_2^2	A_3^2	A_4^2	A_5^2	A_6^2
	2	2.2992	2.8258	3.0674	3.4268	4.5
1	0	0	0.0664	0.3099	0.7008	0.8534
2	1	0.6136	0.4571	0.3828	0.0765	0
3	0	0.3864	0.4765	0.3074	0.2226	0.1466

Table 9 The trained belief matrix of the attribute x_3

x_3 y	e_1^3	e_2^3	e_3^3	e_4^3	e_5^3	e_6^3
	A_1^3	A_2^3	A_3^3	A_4^3	A_5^3	A_6^3
	1	2.6994	4.1406	5.0473	6.0123	7
1	1	0.7198	0	0	0	0
2	0	0.2802	0.9238	0.3140	0.0051	0
3	0	0	0.0762	0.686	0.9949	1

Table 10 The trained belief matrix of the attribute x_4

x_4 y	e_1^4	e_2^4	e_3^4	e_4^4	e_5^4	e_6^4
	A_1^4	A_2^4	A_3^4	A_4^4	A_5^4	A_6^4
	0	0.3982	1.0733	1.5223	2.0049	2.5
1	1	0.9709	0.0317	0	0	0
2	0	0.0291	0.9461	0.6931	0.0703	0
3	0	0	0.0222	0.3069	0.9297	1

Table 11 The trained evidence weights w_j^i of the attribute x_i

x_1	w_1^1	w_2^1	w_3^1	w_4^1	w_5^1	w_6^1
	0.2473	0.2473	0.1983	0.1983	0.1983	0.1983
x_2	w_1^2	w_2^2	w_3^2	w_4^2	w_5^2	w_6^2
	0.0124	0.0124	0.0124	0.0614	0.0614	0.0124
x_3	w_1^3	w_2^3	w_3^3	w_4^3	w_5^3	w_6^3
	0.9995	1	0.9505	0.9505	0.9995	0.9995
x_4	w_1^4	w_2^4	w_3^4	w_4^4	w_5^4	w_6^4
	1	1	0.9417	0.9417	0.9907	0.9907

As a result, for a sample in the testing set, its class label can be estimated by the trained ER-based classifier. For example, for a sample [x_1=6.5, x_2=2.8,x_3=4.6, x_4=1.5, y_2] in the testing set, its attribute value x_1 activates e_4^1 and e_5^1 with the similarity degrees $\alpha_{1,4}$=0.9502, $\alpha_{1,5}$=0.0498 respectively; x_2 activates e_2^2 and e_3^2 with the similarity degrees $\alpha_{2,2}$=0.049, $\alpha_{2,3}$=0.951 respectively; x_3 activates e_3^3 and e_4^3 with the similarity degrees $\alpha_{3,3}$=0.4933, $\alpha_{3,4}$=0.5067 respectively; and x_4 activates e_3^4 and e_4^4 with the similarity degrees $\alpha_{4,3}$=0.0497, $\alpha_{4,4}$=0.9503 respectively. Using Equations (10)-(11), we have e_1={(y_1,0.009), (y_2,0.3803), (y_3,0.6106)}, e_2={(y_1,0.0631), (y_2,0.4648), (y_3,0.4721)},e_3={(y_1,0), (y_2,0.6148), (y_3,0.3852)}, e_4={(y_1,0.0016), (y_2,0.7057), (y_3,0.2928)}, w_1=0.1983, w_2=0.0124, w_3=0.9505, w_4=0.9417, and then by using ER rule, the combined result can be obtained as $O(\boldsymbol{x}^k)$={(y_1, 0), (y_2,0.7817), (y_3, 0.2183)}, where y_2 has the maximum degree of belief. So we predict this testing sample belongs to Iris Versicolour, which coincides with its actual class. Table 12 shows the confusion matrix of the testing samples given by the trained ER-based classifier.

Table 12 The confusion matrix of the testing samples given by the ER-based classifier

		Predicted class			Total
		y_1	y_2	y_3	
Actual class	y_1	20	0	0	20
	y_2	0	19	1	20
	y_3	0	0	20	20

In order to get more reliable result to reflect the performance of the ER-based classifier, we repeat the above experiment 100 times, and then calculate the following three performance values: the average classification accuracy (ACA) of the initial ER-based classifier for the training set is 96.11%, the ACA of the trained ER-based (T-ER) classifier for the training set is 96.89%, and the ACA of the T-ER classifier for the testing set is 96.33%.

4.2 Comparisons with the existing classifiers on five datasets

To further verify the validity of the proposed ER-based classifier, the five well-known classifiers are compared with it, which include naive Bayes [13], Bayes net [14], decision tree learner (REP Tree) [15], random forest [16], one nearest neighbor (1-NN) [51]. The recent DC rule-based classifier (DC-core sample) [52] is also compared.

Besides the Iris dataset, the other four datasets from the UC Irvine Machine Learning Repository [53],as shown in Table 13, are also used for the comparison study. The *Heart* dataset (Statlog collection) is concerned with predicting the presence or absence of heart disease, which is based on some general information about a patient and some test results. The *Wine* dataset is the result of chemical analysis of three types of wines grown in the same region in Italy but derived from three different cultivars. The *Haberman* dataset (Haberman’s Survival Dataset) contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago’s Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The *Iris* data set, as mentioned above, is perhaps the best known database in pattern recognition literature. The *Seeds* dataset comprises the measurements of seven geometric parameters of kernels belonging to three different varieties of wheat. Table 13 shows the general information about these datasets.

Similarly, in the above datasets, 60% of the sample data are randomly selected from each class dataset to build the training dataset, while the remaining is served as test data. Table 14 presents the ACA indices of the T-ER classifier together with the others for the six datasets. It can be seen that there is no universal and perfect classifier for all datasets, but as a whole, the ACA of the T-ER classifier is a bit higher than the others.

Table 13 General information about the five datasets

Dataset	Sample	Class	Attribute
Heart	270	3	13
Wine	178	3	13
Haberman	306	2	3
Iris	150	3	4
Seed	210	3	7

Table 14 The ACAs of the different classifiers

	Naive Bayes	Bayes net	REP Tree	Random forest	1-NN	DC-Core samples	T-ER
Heart	0.8056	0.7593	0.7778	0.8056	0.7500	0.7778	**0.8420**
Wine	0.9718	0.9859	0.8592	0.9718	0.9437	0.9069	**0.9783**
Haberman	0.7623	0.7787	0.7377	0.6639	0.6537	0.8000	**0.7424**
Iris	0.9333	0.9167	0.9167	0.9333	0.9000	0.9667	**0.9633**
Seeds	0.8810	0.9048	0.8810	0.8926	0.8690	0.9048	**0.8956**
Average	0.8708	0.8691	0.8345	0.8543	0.8233	0.8712	**0.8843**

5. Conclusion

In this paper, an ER rule-based classifier is proposed to solve the classification problem. An initial ER rule-based classifier is firstly constructed by acquiring evidence and evaluating the reliability of evidence from the training data, and then the SLP technique is used as the optimization algorithm to update the parameters of the classifier. The advantages of the proposed method are summarized as follows.

1) It completely depends on the available data or is purely data-driven. In particular, the generation process of evidence based on sample casting and normalization of likelihoods can transfer the information in data into the corresponding evidence without any information loss or distortion.
2) The reliability of evidence can objectively reflect the identification ability of attribute and its evidence. On the other hand, when combination is done, the weight of evidence embodies its relative importance compared with other evidence. Obviously, these two factors are very essential and vital for multi-source information fusion.

The experimental results based on five datasets validate the efficiency of the proposed ER rule-based classifier, and confirm that this method can be easily used in many practical applications.

Acknowledgements

This work was supported by the NSFC (No. 61433001, 61374123, 61573076, 61573275, U1509203), and the Zhejiang Province Research Program Project of Commonweal Technology Application (No. 2016C31071), and the University Students' Scientific and Technological Innovation Activity Plan of Zhejiang province (Xin-Miao Talent plan).

Reference

[1] F. Azuaje, I.H. Witten, E. Frank, data mining: practical machine learning tools and techniques, Biomedical Engineering Online 5(1) (2006) 1-2.

[2] P. Melin, O. Castillo, A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition, Applied Soft Computing 21(5) (2014) 568-577.

[3] L. Liu, Y. Nie, L. Lin, W. Li, Z. Huang, S. Xie, et al, Pattern recognition of multiple excitation autofluorescence spectra for colon tissue classification, Photodiagnosis & Photodynamic Therapy 10(2) (2013)111-119.

[4] R. Casini, P.G. Judge, T.A. Schad, Removal of spectro-polarimetric fringes by two-dimensional pattern recognition, Astrophysical Journal 756(2) (2012) 828-842.

[5] X.B. Xu, S.B Li, X.J. Song, C.L. Wen, D.L. Xu, The optimal design of industrial alarm systems based on evidence theory, Control Engineering Practice 46(2016) 142-156.

[6] X.B. Xu, Z. Zhou, C.L. Wen, Data fusion algorithm of fault diagnosis considering sensor measurement uncertainty, International Journal on Smart Sensing and Intelligent Systems 6(1) (2013) 172-190.

[7] X.B. Xu, Z. Zhang, D.L. Xu, Y.W. Chen, Interval-valued evidence updating with reliability and sensitivity analysis for fault diagnosis. International Journal of Computational Intelligence Systems (2016) 396-415.

[8] Z. W. Lu, L. W. Wang, Learning descriptive visual representation for image classification and annotation, Pattern Recognition 48 (2015) 498-508.

[9] G. Camps-Valls, D. Tuia, L. Bruzzone, J. A. Benediktsson, Advances in hyperspectral image classification, IEEE Signal Processing Magazine 31(1) (2014) 45 - 54.

[10]O. Boiman, E. Shechtman, M. Irani, In defense of Nearest-Neighbor based image classification, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 69, IEEE, 2008, pp.1-8.

[11]M. Aci, M. Avci, K nearest neighbor reinforced expectation maximization method, Expert Systems with Applications 38 (2011) 12585–12591.

[12]J. A. Nasiri, N. M. Charkari, S. Jalili. Least squares twin multi-class classification support vector machine, Pattern Recognition 48 (2015) 984-992.

[13]I. Rish, An empirical study of the naive Bayes classifier, in: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, 2001, pp. 41-46.

[14]N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2–3) (1997) 131–163.

[15]Y. Freund, L. Mason, The alternating decision tree learning algorithm, in: ICML,vol. 99, 1999, pp. 124–133.

[16]L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5-32.

[17]M. A. Sanchez, O. Castillo, J. R. Castro, P. Melin, Fuzzy granular gravitational clustering algorithm for multivariate data, Information Sciences 279 (2014) 498-511.

[18]G. D. Zhao, Y. Wu, F. Q. Chen, J. M. Zhang, J. Bai. Effective feature selection using feature vector graph for classification, Neurocomputing 151 (2015)376-389.

[19]S. Afshar, M. Mosleh, M. Kheyrandish, Presenting a new multiclass classifier based on learning automata, Neurocomputing 104 (2013) 97-104.

[20]C. Zhang, Y. Hu, F. T. S. Chan, R. Sadiq, Y. Deng, A new method to determine basic probability assignment using core samples, Knowledge-Based Systems 69(1) (2014) 140-149.

[21]T. Denoeux, A neural network classifier based on Dempster-Shafer theory, IEEE Trans. Syst., Man Cybernet., Part A: Syst. Hum. 30 (2) (2000) 131-150.

[22]T. Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE Trans. Syst., Man Cybernet. 25 (5) (1995) 804-813.

[23]L. Dymova, P. Sevastianov, P. Bartosiewicz, A new approach to the rule-base evidential reasoning: stock trading expert system application, Expert Systems with Applications 37(8) (2010) 5564-5576.

[24]A.P. Dempster, Upper and lower probabilities induced by a multi-valued mapping, Annals of Mathematical Statistics 38(2) (1967) 325-339.

[25]A.P. Dempster, A generalization of Bayesian inference, Journal of the Royal Statistical Society. Series B 30 (1968) 205-247.

[26]G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, New Jersey,1976.

[27]Shafer, G., & Pearl, J., Readings in uncertain reasoning, Morgan Kaufmann Publishers Inc., 1990.

[28]Y. Chen, Y. W. Chen, X. B. Xu, C. C. Pan, J. B. Yang, G. K. Yang, A data-driven approximate causal inference model using the evidential reasoning rule, Knowledge-Based Systems 88(2015) 264-272.

[29]D.L. Xu, J.B. Yang, Y.M. Wang, The evidential reasoning approach for multi-attribute decision analysis under interval uncertainty, European Journal of Operational Research 174(3) (2006) 1914-1943.

[30]J.B. Yang, Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties, European Journal of Operational Research 131(1) (2001) 31-61.

[31]J.B. Yang, D.L. Xu, On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty, Systems Man & Cybernetics Part A Systems & Humans IEEE Transactions on 32(3) (2002) 289-304.

[32]J.B. Yang, Y.M. Wang, D.L. Xu, K.S. Chin, The evidential reasoning approach for MADA under both probabilistic and fuzzy uncertainties, European Journal of Operational Research 171(1) (2006) 309-343.

[33]J.B. Yang, D.L. Xu, Evidential reasoning rule for evidence combination, Artificial Intelligence 205(205) (2013) 1-29.

[34]J.B. Yang, M.G. Singh, An evidential reasoning approach for multiple-attribute decision making with uncertainty, IEEE Transactions on Systems Man & Cybernetics 24(1) (1994) 1-18.

[35]D.L. Xu, An introduction and survey of the evidential reasoning approach for multiple criteria decision analysis, Annals of Operations Research 195(195) (2012) 163-187.

[36]L. Jiao, Q. Pan, T. Denœux, Y. Liang, X. Feng, Belief rule-based classification system: Extension of FRBCS in belief functions framework, Information Sciences, 309 (2015) 26-49.

[37]L. Chang, Z.J. Zhou, Y. You, L. Yang, Z. Zhou, Belief rule based expert system for classification problems with new rule activation and weight calculation procedures. Information Sciences, 336 (2015) 75-91.

[38]G. Kong, D.L. Xu, J.B. Yang, X. Yin, T. Wang, B. Jiang, Y. Hu, Belief rule-based inference for predicting trauma outcome. Knowledge-Based Systems, 95 (2015) 35-44.

[39]J.B. Yang, D.L. Xu, A study on generalising Bayesian inference to evidential reasoning, in Belief Functions: Theory and Applications, Springer International Publishing, New York, 2014, pp.180-189.

[40]R.E. Griffith, R.A. Stewart. A nonlinear programming technique for the optimization of continuous processing systems, Management Science 7(1961) 379-392.

[41]G.B. Dantzig, Making progress during a stall in the simplex algorithm. Linear Algebra & Its Applications, s 114-115(1989) 251-259.

[42]N. Karmarkar, An Interior-Point Approach to NP-Complete Problems, Proceedings of the 1st Integer Programming and Combinatorial Optimization Conference, University of Waterloo Press, 1990, pp.351-366.

[43]B.R.D. Richtmyer, K.W. Morton, Difference Methods for Initial Value Problems, Wiley, New York, 2010.

[44]Y.W. Chen, D.L. Xu, J.B. Yang, Effective learning of belief rule based systems with sequential linear programming, International Conference on Automation & Computing, 2010.

[45]R.T. Haftka, Z. Gürdal. Elements of Structural Optimization (3rd ed.). Kluwer, Boston, MA. 1993.

[46]L. Lamberti, C. Pappalettere, Move limits definition in structural optimization with sequential linear programming, Part I: Optimization algorithm and Part II: Numerical examples, Computers and Structures, 81(2003) 197-238.

[47]T.Y. Chen. Calculation of the move limits for the sequential linear programming method. International Journal for Numerical Methods in Engineering, 36(1993) 2661-2679.

[48]T.Y. Chen. A comprehensive solution for enhancing the efficiency and the robustness of the SLP algorithm. Computers & Structures, 66(4)(1998) 373-384.

[49]B.A. Wujek, J.E. Renaud. New adaptive move-limit management strategy for approximate optimization, Parts 1 and 2. AIAA Journal, 36(1998) 1911-1934.

[50]R. Fisher, The use of multiple measurements in taxonomic problems, Annals of Human Genetics 7 (2) (1936) 179-188.

[51]T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21-27.

[52]C. Zhang, Y. Hu, F. T. S. Chan, R. Sadiq, Y. Deng, A new method to determine basic probability assignment using core samples, Knowledge-Based Systems 69(1)(2014) 140-149.

[53] A. Asuncion, D. Newman, UCI machine learning repository, 2007.