Classification with Mixed Numeric and Categorical Data Using Improved Extension Theory

Cheng-Hsiang Liu

***Abstract*—Classification is a chief issue in decision science and knowledge discovery, hence the recent development of several classification methods. Compared to other methods, extension theory (ET) does not require a particular learning process, and its calculation is both fast and simple. This paper proposes improved extension theory (IET) to supersede the shortcoming of ET, such as failure to implement data classification when attributes are categorical. This study assesses IET performance according to six real-world datasets. Comparisons with other classifiers (i.e., ET and decision trees) illustrate the effectiveness of the proposed IET.**

***Index Terms*—Extension theory, Classical domain, Classification, Mixed numeric and categorical data**

I. Introduction

Extension theory (ET), first proposed by Cai [1], is a pattern classification algorithm. Given a set of n labeled examples $D_n = \{(\overrightarrow{X_1}, Y_1), (\overrightarrow{X_2}, Y_2), \ldots, (\overrightarrow{X_n}, Y_n)\}$ with input vectors $\overrightarrow{X_i} \in \Re^d$ and class labels $Y_i \in \{\omega_1, \omega_2, \ldots, \omega_m\}$, ET classifies an unseen pattern $\overrightarrow{X_k} = (x_{k1}, x_{k2}, x_{k3}, \ldots, x_{kd})$ to the class Y_k with the highest membership grade. Identifying a class with the highest membership grade requires defining a set of matter-element models and an extended correlation function. An ordered ternary $R = (N, c, v)$ is necessary to describe a matter for transformation called a matter-element. Three fundamental elements are included in a matter-element (*R*): the name of the matter (*N*), the characteristic of the matter (c), and the value of the matter characteristic (*v*). A matter may have numerous characteristics. Assuming that *C* is a characteristic vector, then $C = [c_1, c_2, ..., c_d]$, and assuming that *V* is the same as *C*, thus also a vector, then $V = [v_1, v_2, ..., v_d]$, and a multidimensional matter-element is defined as follows:

$$R = (N, C, V) = \begin{bmatrix} N & c_1 & v_1 \\ & c_2 & v_2 \\ & \vdots & \vdots \\ & c_d & v_d \end{bmatrix} \quad (1)$$

If the value of a characteristic has a classical domain or range, the matter-element can be defined for the classical domain as follows [2]:

$$R = (N, C, V) = \begin{bmatrix} N & c_1 & v_1 \\ & c_2 & v_2 \\ & \vdots & \vdots \\ & c_d & v_d \end{bmatrix} = \begin{bmatrix} N & c_1 & (v_1^L, v_1^U) \\ & c_2 & (v_2^L, v_2^U) \\ & \vdots & \vdots \\ & c_d & (v_d^L, v_d^U) \end{bmatrix} \quad (2)$$

The first step of the extension classification method is to formulate matter-element models of m categories, performed as follows:

$$R_j = (N_j, C, V_j) = \begin{bmatrix} N_j & c_1 & v_{1j} \\ & c_2 & v_{2j} \\ & \vdots & \vdots \\ & c_d & v_{dj} \end{bmatrix} = \begin{bmatrix} N_j & c_1 & (v_{1j}^L, v_{1j}^U) \\ & c_2 & (v_{2j}^L, v_{2j}^U) \\ & \vdots & \vdots \\ & c_d & (v_{dj}^L, v_{dj}^U) \end{bmatrix}, \quad j = 1, 2, \ldots, m \quad (3)$$

Similarly, a matter-element R_p of *P* can be described as follows:

$$R_p = (N_p, C, V_p) = \begin{bmatrix} N_p & c_1 & v_{1p} \\ & c_2 & v_{2p} \\ & \vdots & \vdots \\ & c_d & v_{dp} \end{bmatrix} = \begin{bmatrix} N_p & c_1 & (v_{1p}^L, v_{1p}^U) \\ & c_2 & (v_{2p}^L, v_{2p}^U) \\ & \vdots & \vdots \\ & c_d & (v_{dp}^L, v_{dp}^U) \end{bmatrix} \quad (4)$$

where V_p are the ranges of *C*, called the neighborhood domains. The value range (v_{tj}^L, v_{tj}^U) of the classical domain for each characteristic can be obtained from previous experience, or determined according to the lower and upper bounds of the field-test records. The value range (v_{tp}^L, v_{tp}^U) of the neighborhood domain can be obtained from previous experience, or determined from the maximum and minimum values of each characteristic in the statistical records [2-5].

After the formulation of the element-matter models of classification categories, ET defines an extended correlation function by *K(x)* to quantify the relationship between an element and a set. The correlation functions have numerous forms dependent on application. A common extended correlation function can be defined as Equation (5) [4]. The extended correlation function is shown in Figure 1. Figure 1

Manuscript received January 06, 2013; revised January 23, 2013. This work was supported by National Science Council of Taiwan under grant no. NSC 101-2221-E-020-009.

Cheng-Hsiang Liu is with the Department of Industrial Management, National Pingtung University of Science & Technology, Pingtung 91201, Taiwan. (phone: +886-8-7703202, ext. 7717; fax: +886-8-7740321; e-mail: chliu.uw@gmail.com).

shows that: (1) $x_{kt} \in v_{tj}, K_{tj}(x_{kt}) \geq 0$, that is, the classic set, indicating the degree to which x_{kt} belongs to v_{tj}; (2) $x_{kt} \in v_{tp} - v_{tj}, -1 \leq K_{tj}(x_{kt}) < 0$, that is, the extension set, meaning that element x_{kt} is apparently outside v_{tj}, but still has a probability of becoming a part of the set if conditions change; (3) $x_{kt} \notin v_{tj}, K_{tj}(x_{kt}) < -1$, that is, the negative set, implying that element x_{kt} cannot be in v_{tj}.

$$K_{tj}(x_{kt}) = \begin{cases} \dfrac{-2\rho(x_{kt}, v_{tj})}{\left|v_{tj}^U - v_{tj}^L\right|}, & if \ x_{kt} \in v_{tj} \\ \dfrac{\rho(x_{kt}, v_{tj})}{\rho(x_{kt}, v_{tp}) - \rho(x_{kt}, v_{tj})}, & if \ x_{kt} \notin v_{tj} \end{cases}, \ j = 1, 2, 3, \ldots, m; \ t = 1, 2, ,3, \ldots, d \tag{5}$$

where

$$\rho(x_{kt}, v_{tj}) = \left| x_{kt} - \frac{v_{tj}^U + v_{tj}^L}{2} \right| - \frac{v_{tj}^U - v_{tj}^L}{2} \tag{6}$$

The membership grade of the unseen pattern $\overrightarrow{X_k}$ with the class ω_j is calculated as Equation (7), where W_{tj} denotes the significance of every class feature in the classification process. The maximum value of the membership grade determines the class label of the unseen pattern $\overrightarrow{X_k}$.

$$\lambda_{kj} = \sum_{t=1}^{d} W_{tj} \times K_{tj}(x_{kt}), \quad j = 1, 2, \ldots, m \tag{7}$$

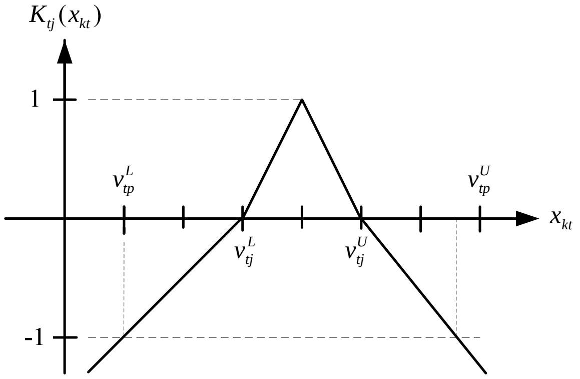

Figure 1. Extended correlation function [3]

Extension theory has shown promising results in fault diagnoses [2-4][6] and in tracking the maximum power point of photovoltaic (PV) arrays [5]. Traditional extension classification methods rely on experiences to set rules for the classical domain [2], which is a tedious and complicated step in the classification process. Liu [7] proposed a modified extension theory (MET) to overcome the above shortcoming. Experimental results indicate that the MET consistently achieved better or comparable results than the traditional ET. For more details on MET, refer to Liu [7]. However, MET is limited to numeric data. Therefore, this study presents a improved extension theory (IET) based on MET to supersede the shortcoming of limiting to numeric data. The experiments, using six real-world datasets, demonstrated that IET surpasses ET in terms of classification accuracy. Additionally, the IET is superior or comparable to other classification method presently in use, such as DT.

The remainder of this paper is structured as follows: Section 2 introduces the proposed modified extension theory, Section 3 presents an evaluation of the proposed method, and finally, the last section states several conclusions and suggestions for future considerations.

II. Improved Extension Theory

This study modifies the MET of Liu [7] in order to overcome the shortcoming of limiting to numeric data. This work incorporates the use of the concept of frequency-based center (FBC) into MET to deal with categorical data. The proposed method is hereafter called the improved extension theory (IET). This section will present the mathematical descriptions of IET. The proposed IET is described as follows:

Step 1: Choose values of the classical domains (v_{tj}^L, v_{tj}^U) for $j = 1, 2, 3, \ldots, m$ and $t = 1, 2, ,3, \ldots, d$. To determine the range of the classical domain for the characteristic c_t of class w_j, a largest sphere centered on class-center $\overrightarrow{X^j}$ was constructed, which excludes all training examples from other classes. Attaining this sphere involves setting the sphere radius to

$$r_j = \min_{i:Y_i \neq w_j} d(\overrightarrow{X^j}, \overrightarrow{X_i}) - \varepsilon \tag{8}$$

$$d(\overrightarrow{X^j}, \overrightarrow{X_i}) = \left(\sum_{t=1}^{d} d_{ij}^t \right)^{1/2} \tag{9}$$

where $\varepsilon > 0$ is an arbitrary small number and $\overrightarrow{X^j} = (\overline{x_1^j}, \overline{x_2^j}, \overline{x_3^j}, \ldots, \overline{x_d^j})$. For a numeric characteristic, $\overline{x_t^j}$ is represented by the mean of all values for that characteristic ct in the corresponding class w_j. For a categorical characteristic, $\overline{x_t^j}$ is represented by a proportional distribution of all its values for that characteristic c_t in the corresponding class w_j, which is called a frequency-based center (FBC), proposed by Ordonez [8]. The example in Figure 2 displays 3 categorical characteristics. The number in each rectangle in Figure 2(a) refers to occurrences of each characteristic value (c_{tl}: the l-th value of a characteristic c_t) counted from the examples of a class. Figure 2(b) is an example of FBC resulting from Figure 2(a). The computation of d_{ij}^t in Equation (9) is regarded as Equation (10).

$$d_{ij}^t = \begin{cases} \left|\overline{x_t^j} - x_{it}\right|^2 & ,if \ characteristic \ c_t \ is \ numeric \\ (1 - f_{tj}(x_{it}))^2 & ,if \ characteristic \ c_t \ is \ categorical \end{cases} \tag{10}$$

where $f_{tj}(x_{it})$ refers to the frequency of the characteristic value x_{it}, counted from examples of a class w_j. The example $\overrightarrow{X_i}$ is included in a set L_j if $Y_i = w_j$ and $d(\overrightarrow{X^j}, \overrightarrow{X_i}) \leq r_j$. For a numeric characteristic, the range of the classical domain (v_{tj}^L, v_{tj}^U) can be determined based on the values of the characteristic c_t of the examples in L_j as equations (11) and (12). For a categorical characteristic, the classical domain of the characteristic (v_{tj}^L, v_{tj}^U) becomes a set containing all values of the characteristic ct of training examples that are in L_j.

$$v_{tj}^L = \min_{i:i\in L_j}\{x_{it}\}, \quad t = 1, 2, \ldots, d \ \textit{and} \ j = 1, 2, \ldots, m \tag{11}$$

$$v_{tj}^U = \max_{i:i\in L_j}\{x_{it}\}, \quad t = 1, 2, \ldots, d \ \textit{and} \ j = 1, 2, \ldots, m \tag{12}$$

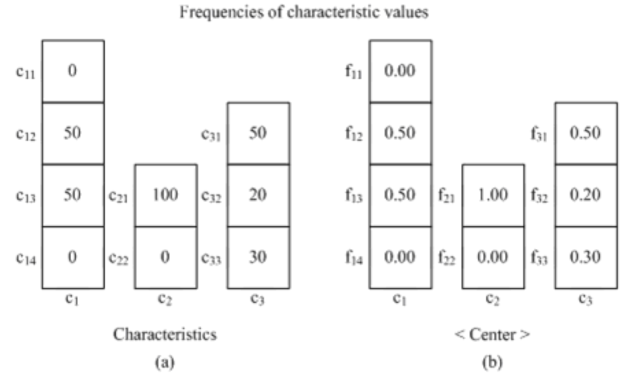

Figure 2. Frequency-based center (FBC)

Step 2: Read the testing pattern $\overrightarrow{X_k} = (x_{k1}, x_{k2}, x_{k3}, \ldots, x_{kd})$.

Step 3: Calculate the correlation degrees of the testing pattern $\overrightarrow{X_k}$ for each characteristic of each matter-element model (class), using equations (13)-(15).

$$\rho(x_{kt}, v_{tj}) = \left|x_{kt} - \overline{v_{tj}}\right| - \frac{v_{tj}^U - v_{tj}^L}{2} \tag{13}$$

$$\overline{v_{tj}} = \frac{\sum_{i\in L_j} x_{it}}{\left|L_j\right|} \tag{14}$$

$$\begin{array}{l} \textit{If characteristic } c_t \textit{ is numeric Then} \\ \quad K_{tj}(x_{kt}) = \begin{cases} \dfrac{-2\rho(x_{kt}, v_{tj})}{\left|v_{tj}^U - v_{tj}^L\right|}, & \textit{if } x_{kt} \in v_{tj} \\ \dfrac{\rho(x_{kt}, v_{tj})}{\rho(x_{kt}, v_{tp}) - \rho(x_{kt}, v_{tj})}, & \textit{if } x_{kt} \notin v_{tj} \end{cases} \\ \quad j = 1, 2, 3, \ldots, m; \ t = 1, 2, ,3, \ldots, d \\ \textit{Else} \\ \quad K_{tj}(x_{kt}) = \begin{cases} 1, & \textit{if } x_{kt} \in v_{tj} \\ -1, & \textit{if } x_{kt} \notin v_{tj} \end{cases}, \quad j = 1, 2, 3, \ldots, m; \ t = 1, 2, ,3, \ldots, d \\ \textit{End} \end{array} \tag{15}$$

Since $\frac{v_{tj}^U + v_{tj}^L}{2}$ in Equation (6) has been replaced by $\overline{v_{tj}}$ in Equation (13) to provide useful summaries of asymmetrical data, an additional feasibility checking routine is performed for $K_{tj}(x_{kt})$ after each calculation, as shown in Equation (16).

```
if $x_{kt} \notin v_{tj}$ then
    $b = \begin{cases} v_{tj}^L, & \textit{if } x_{kt} < v_{tj}^L \\ v_{tj}^U, & \textit{f } x_{kt} > v_{tj}^U \end{cases}$
    $K_1 = \dfrac{\rho(b, v_{tj})}{\rho(b, v_{pj}) - \rho(b, v_{tj})}$                    (16)
    $K_2 = \dfrac{-2\rho(b, v_{tj})}{\left|v_{tj}^U - v_{tj}^L\right|}$
    if $K_1 > K_2$ then
        $K_{tj}(x_{kt}) = K_{tj}(x_{kt}) - (K_1 - K_2)$
    end if
end if
```

Step 4: The membership grade of the unseen pattern $\overrightarrow{X_k}$ with the class ω_j is calculated as Equation (17).

$$\lambda_{kj} = \sum_{t=1}^{d} K_{tj}(x_{kt}), \quad j = 1, 2, \ldots, m \tag{17}$$

Step 5: Rank the membership grades and find the maximum value of the membership grade to determine the class label of the testing pattern. The classification rule is shown as Equation (18).

$$\textit{if } (\lambda_{kj} = \underset{j}{Max}(\lambda_{kj})) \textit{ then } (\textit{the class label of } \overrightarrow{X_k} \textit{ is } w_j) \tag{18}$$

Step 6: Return to Step 2 for the next testing pattern once the classification of the first pattern has been completed, until they have all been finished.

III. Experimental Results

This section reveals that IET is superior in effectiveness and performance in terms of classification accuracy compared to previous classifiers. A comparison with decision trees (DT), and extension theory (ET) verifies this claim. The performed experiments are on six publicly available datasets from the UCI repository (http://archive.ics.uci.edu/ml/). Table 1 describes several characteristics of the used domains. Comparative results show classification on two different types of datasets: pure numeric (Blood) and mixed numeric and categorical datasets (Australian, Cars, Cleveland, Crx, and Hepatitis). Each experiment involved Ten-fold cross-validation. Each dataset was randomly divided into 10 partitions, and each classifier was given a training set comprising 9 partitions from which the classifier returned a classification model to classify the remaining partition. Ten such trials were run for each dataset with each classifier, using a different partition out of the 10, as the test set for each trial. Since ET is limited to numeric data, the classical domain for a categorical characteristic is represented by a set for all values present in the field-test records. Thereafter, $K_{tj}(x_{kt}) = 1$ if the value of the characteristic c_t of the testing example k falls within the classical domain of the characteristic c_t of the class w_j, else $K_{tj}(x_{kt}) = -1$.

Table 1. Basic information of the six datasets from UCI

data set	number of attributes	number of classes	data size
Australian	14	2	460
Blood	4	2	748
Cars	8	3	261
Cleveland	13	2	296
Crx	15	2	653
Hepatitis	19	2	80

Table 2 shows the average error rate over the 10 trials for each domain. Clearly, ET is seemingly inappropriate for classifying these datasets. For most datasets, DT outperforms ET, except for the Hepatitis dataset. Table 2 shows that, on average, the accuracy of IET is higher than that of ET. These results are encouraging because the proposed IET is comparable to that of DT. Statistical significance tests can further ascertain these findings.

Table 2. The error rate on classifiers

Data set	IET	ET	DT
Australian	16.30%	58.26%	18.26%
Blood	25.53%	63.90%	24.73%
Cars	0.38%	0.77%	0.38%
Cleveland	19.26%	30.74%	20.95%
Crx	13.48%	64.78%	13.78%
Hepatitis	11.25%	17.50%	17.50%

The paired *t*-test was applied to detect statistically significant differences in the performance of every pair of classifiers. Table 3 shows the results of the paired *t*-tests. The classifiers are listed in descending order of performance, grouped into homogeneous subsets labeled with a different letter, if the difference between the means of performance measurements of the two classifiers in the subset is not significantly beyond the prescribed α=0.05 level. Based on the compared measurements, the classifier with "A" is significantly superior to the classifier with "B", and the classifier with "B" is significantly superior to the classifier with "C". Testing results revealed that for the four datasets (Australian, Blood, Cleveland, and Crx), IET constructed a classifier not significantly different from DT, but statistically significantly more accurate than its antecedent component, ET. For the Hepatitis dataset, the accuracy of the classifier produced by MET is identical to the best of the other classifiers. For the Cars data-set, no statistically significant differences emerged between IET and the other classifiers. As the experimental results clearly show, IET can function effectively for pure numeric, as well as mixed numeric and categorical datasets.

Table 3. Results of paired *t*-test for IET, ET, and DT

Data set	Classifier	Result	Data set	Classifier	Result
Australian	IET	A	Cleveland	IET	A
	DT	A		DT	A
	ET	B		ET	B
Blood	DT	A	Crx	IET	A
	IET	A		DT	A
	ET	B		ET	B
Cars	IET	A	Hepatitis	IET	A
	DT	A		DT	B
	ET	A		ET	B

IV. Conclusions

This study proposed improved extension theory (IET), a classification method for mixed numeric and categorical datasets, and presented an evaluation of its performance. The characteristic of IET is to incorporate the concept of frequency-based center into MET of Liu [7] for dealing with categorical characteristics. The results obtained with IET over a number of real-world datasets are encouraging. The tests on several real-world datasets demonstrated that the performance of IET is superior to that of ET, and comparable in performance to decision trees (DT). The simplicity of IET and its impressive performance makes it an appealing tool for pattern classification.

References

[1] W. Cai, "The extension set and incompatibility problem," *Journal of Scientific Exploration*, vol. 1, pp. 610-614, 1983.

[2] M.-H. Wang, Y.-F. Tseng, H.-C. Chen and K.-H. Chao, "A novel clustering algorithm based on the extension theory and genetic algorithm," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8269-8276, 2009.

[3] J. Ye, "Application of extension theory in misfire fault diagnosis of gasoline engines," *Expert Systems with Application*, vol. 36, no. 2, pp. 1217-1221, 2009.

[4] M.-H. Wang, "Application of extension theory to vibration fault diagnosis of generator sets," *IEE Proceedings Generation Transmission and Distribution*, vol. 151, no. 4, pp. 503-508, 2004.

[5] K.-H. Chao and C.-J. Li, "An intelligent maximum power point tracking method based on extension theory for PV systems," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1050-1055, 2010.

[6] M.-H. Wang and C.-P. Hung, "Extension neural network and its applications," *Neural Networks*, vol. 16, no. 5-6, pp. 779-784, 2003.

[7] C.-H. Liu, "Extending extension theory for classifying data with numerical values," *Neural Computing and Applications*, DOI: 10.1007/s00521-011-0795-z, 2012.

[8] C. Ordonez, "Clustering binary data streams with K-means," *Proc. ACM SIGMOD workshop on Data Mining and Knowledge Discovery*, pp. 12-19, 2003.