ISSN: 1692-7257 - Volumen 1 – Número 21 - 2013

Revista Colombiana de Tecnologías de Avanzada

Recibido: 01 de agosto de 2012
Aceptado: 23 de octubre de 2012

# DIFFERENTIATIONS OF OBJECTS IN DIFFUSE DATABASES

# DIFERENCIACIONES DE OBJETOS EN BASES DE DATOS DIFUSAS

**PhD. Amaury Caballero*; PhD. Gabriel Velasco*; PhD. Aldo Pardo García****

* **Florida International University**, Miami, USA; ** **Universidad de Pamplona,**
Email: {caballer, gvelasco}@fiu.edu, apardo13@hotmail.com

**Abstract:** The paper presents a method for classification of objects in the case, where different parameters are used, and each parameter can vary for each object and parameter into certain range. Diffuse data bases are frequently present due to the fact that conditions of transmission or variation in the sensors can produce different received values for the same object and parameter. The method uses the basic concepts of information theory. Calculating the loose of information due to coincidences in the same parameter for different objects, it is possible to find out the smaller number of parameters for the discrimination under certain error restrictions, and from there, the different objects or classes can be discriminated.

**Keywords:** Diffuse Databases, Classification, Information Theory.

**Resumen:** El artículo presenta un método para la clasificación de los objetos, en donde los parámetros para cada objeto pueden variar, así como sus valores en cierto rango. Las Bases de datos difusas están presentes con frecuencia debido al hecho de que las condiciones de transmisión o variación en los sensores pueden producir diferentes valores recibidos para el mismo objeto y parámetro. El método utiliza los conceptos básicos de la teoría de la información. Cálculo de la pérdida de información debido a la coincidencia en el mismo parámetro para objetos diferentes, es posible encontrar el menor número de parámetros para la discriminación bajo ciertas restricciones de error, y a partir de allí, los objetos o clases diferentes pueden ser discriminados.

**Palabras clave:** Bases de datos difusas, Clasificación, teoría de la información.

## 1. INTRODUCTION

Extraction and discovery of information in a database has been growing due to the development, among other factors, of accurate and inexpensive sensors. Several works have been devoted to this question. Among them it is possible to cite (Ying Chieh, *et al.* 2006), (Battiti, 1991), (Granger, *et al.*, 2000) and (Caballero *et al.*, 2010). Frequently, in the differentiation among different objects or conditions, it is necessary to use more than one parameter. As the number of objects is increased and their properties are more similar, the number of differentiating parameters should be also increased.

In the case of transmitting numerical information, these numerical values can be changed into some interval for each object or parameter.

Let's accept the following definitions:

| | |
|---|---|
| $A_k$ | Attribute # *k.* |
| $U_i$, $U_j$ | Object (class) *i* or *j* respectively. |
| $l_i^k$, $l_j^k$ | Minimum values for classes *i* and *j*. |
| $u_i^k$, $u_j^k$ | Maximum values for classes *i* and *j*. |

The region of coincidence or misclassification rate of the two attributes may vary from 0 when there is not coincidence at all, to 1 when the two attributes coincide completely.

In general, the probability that objects in class $u_i$ be misclassified into class $u_j$ according to attribute $k$ has been slightly modified from (Leung, *et al.*, 2007) and given by:

$a_{ij}^{k}$ = a

$$if\ [l_i^k, u_i^k] \bigcap [l_j^k, u_j^k] = 0; \quad (1)$$

$$a_{ij}^{k} = min\{(u_i^k - l_j^k, u_j^k - l_i^k)/(u_i^k - l_i^k)\},1\},$$
$$if\ [l_i^k, u_i^k] \bigcap [l_j^k, u_j^k] \neq 0 \quad (2)$$

Where $a_{ij}^{k}$ is the probability that objects in class $U_i$ are misclassified into class $U_j$ as per the attribute $A_k$,.

Note that in general

$a_{ij}^{k} \neq a_{ji}^{k}$.

From the previous result, it was defined the maximum mutual classification error between classes $u_i$ and $u_j$ for attribute $k$ as

$$ß_{ij}^{k} = max\ \{ a_{ij}^{k}, a_{ji}^{k}\} \quad (3)$$

where $ß_{ij}^{k} = ß_{ji}^{k}$.

It results logical to think that when the two classes coincide for some parameter k, the obtained information from this parameter for discriminating between classes *i* and *j* is 0, and that it increases as the coincidence diminishes. This leads to the representation of this information, from Shannon and Hartley definition, using a logarithmic scale. Here the logarithm is used to provide the additive characteristic for independent uncertainty. For expressing it with logarithms base 10, it is given as

$$I_{ij}^{k} = -(\log ß_{ij}^{k}) \quad [Hartley] \quad (4)$$

Similarly, the minimum information required for the classification between two classes *i* and *j* for an attribute *k* is given by

$$I_a^k = L = -\log a \ [Hartley]. \quad (5)$$

Where a is the permissible misclassification error between classes for any attribute. This value shall be defined from the beginning of the classification, and should be bigger than zero. If $I_{ij}^{k} \geq I_a^k$, the two classes can be separated using the attribute k with the classification error given by a

## 2. PROPOSED METHOD

The proposed method can be explained from the following steps, also indicated in Figure 1:

**Step 1:** Select the value for *a* and calculate the value of *L* from $L = -\log_{10} a$ (Localization **L**)

**Step 2:** Calculate all the values of $I_{ij}^{k}$ from equation (4) and create Table 2 (**T2**).

**Step 3:** If $\sum_{k=1}^{m} I_{ij}^{k} < L$ for some calculated rows, mark those rows (**Vector ND)**

**Step 4:** Check each row containing only one *L* value and select the attribute corresponding to this value. If the same attribute is repeated in more than one row, select it only once. Then mark all the rows containing the selected attribute, and indicate the selected objects (**Vectors ATU** and **UD)**. If there are other rows with the value "*L"* go to Step 5. If not, move to Step 6.

**Step 5**: Among the non-selected attributes, select the attribute (column) with the value *L* shown in the largest number of rows. If this number is the same for more than one attribute, then select the attribute with the largest values in row S (shown in Table 2, in the example). Mark all the non-marked rows containing the selected attribute and objects (**Vectors ATU** and **UD)**. Repeat Step 5, until no row containing the value *L* is analyzed, then go to the following step. If all the rows are marked, the selection process is completed and the useful attributes are those selected.

**Step 6:** Select the value with maximum $I_{ij}^{k}$ on the first non-marked row. If the corresponding attribute has not been selected, select it and look for the next maximum $I_{ij}^{k}$ in the same row. Add them together and check whether the result satisfies S $I_{ij}^{k} \geq L$. If not, repeat the process with other attributes in the same row until the indicated condition is met. Mark this row. All the new selected attributes and objects will be included in the final selection (**Vectors ATD** and **NUD)**. Move to the next non-marked row until all the rows have been analyzed.

The method was proposed in (Leung, *et al.*, 2007).

The following parameters have been defined:

**L** ---- Localization for [$\log_{10}$ a ]

**T1** ---- Table for $a_{ij}^{k}$

**T2** ---- Table for calculated $I_{ij}^{k}$

**SA** ---- Location in T2, showing the addition of all the $I_{ij}^{k}$ for each row

**SO** ---- Location in T2, showing the addition of all the $I_{ij}^{k}$ for each column

**ST** ---- Temporary location of $I_{ij}^{k}$

**Vector ATU** ---- Vector showing the attributes that discriminate selected objects without any uncertainty for the accepted error a [number of rows =n, where n = max. k]

**Vector ATD** ---- Vector showing the attributes that discriminate selected objects with uncertainty for the accepted error a [number of rows =n, where n = max. k].

**Vector UD** ---- Vector showing the discriminated objects without any uncertainty for the accepted error a [number of rows =m, where m = max. number of objects].

**Vector NUD** ---- Vector showing the discriminated objects with some uncertainty for the accepted error a [number of rows =m, where m = max. number of objects].

**Vector ND** ---- Vector showing objects that can't be discriminated for the accepted error a [number of rows =m, where m = max. number of objects]

## 3. EXAMPLE OF APPLICATION

The interval-valued information system presented by Y. Leung et al. is partially presented in Table 1, showing the first five classes only. As per Step 1, using $a$ = 0.2 ($I_a$ = 0.7), the results are shown in Table 2.

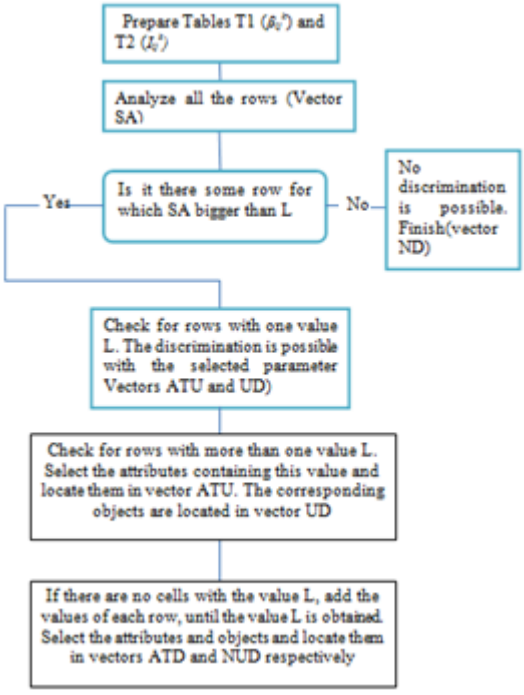



*Fig. 1: Generalized Block Diagram*

In this table, from Step 2, it is obtained that classes $U_2$ and $U_5$ cannot be discriminated. Step 4 gives on the second row the attribute $A_3$, which is selected, and rows $U_{13}$ and $U_{34}$ are marked.

*Table 1: Interval-Valued Information System*

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $U_1$ | 2.17; 2.86 | 2.45; 2.96 | 5.32; 7.23 | 3.21; 3.95 | 2.54; 3.12 |
| $U_2$ | 3.37; 4.75 | 3.43; 4.85 | 7.24; 10.47 | 4.00; 5.77 | 3.24; 4.70 |
| $U_3$ | 1.83; 2.70 | 1.78; 2.98 | 7.23; 10.27 | 2.96; 4.07 | 2.06; 2.79 |
| $U_4$ | 1.35; 2.12 | 1.42; 2.09 | 2.59; 3.93 | 1.87; 2.62 | 1.67; 2.32 |
| $U_5$ | 3.46; 5.35 | 3.37; 5.11 | 6.37; 10.28 | 3.76; 5.70 | 3.41; 5.28 |

*Table 2: Classification Error between Classes $U_i$ and $U_j$*

| $I_{ij}^k$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\sum_{k=1}^{m} I_{ij}^k$ |
|---|---|---|---|---|---|---|
| $U_{12}$ | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 3.5 |
| $U_{13}$ | 0.11 | 0 | 0.7 | 0 | 0.37 | 1.18 |
| $U_{14}$ | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 3.5 |
| $U_{15}$ | 0.7 | 0.7 | 0.35 | 0.58 | 0.7 | 3.03 |
| $U_{23}$ | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 3.5 |
| $U_{24}$ | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 3.5 |
| $U_{25}$ | 0.03 | 0 | 0.03 | 0.02 | 0 | **0.08** |
| $U_{34}$ | 0.42 | 0.38 | 0.7 | 0.7 | 0.4 | 2.6 |
| $U_{35}$ | 0.7 | 0.7 | 0 | 0.55 | 0.7 | 2.65 |
| $U_{45}$ | 0.7 | 0.7 | 0 | 0.7 | 0.7 | 2.8 |
| S | 0.56 | 0.38 | 0.38 | 1.15 | 0.77 | |

Applying Step 5, the first selected attribute is $A_5$. Rows $U_{12}$, $U_{14}$, $U_{15}$, $U_{23}$, $U_{24}$, $U_{35}$, and $U_{45}$ result marked. The selected attributes are $A_3$ and $A_5$. The following rules could be proposed:

*R*ule # 1: IF $A_3 \in$ [5.32; 7.23] and $A_5 \in$ [2.54; 3.12], THEN it is $U_1$.
*R*ule # 2: IF $A_3 \in$ [7.23; 10.27] and $A_5 \in$ [2.06; 2.79], THEN it is $U_3$.
*R*ule # 3: IF $A_3 \in$ [2.59; 3.93] and $A_5 \in$ [1.67; 2.32], THEN it is $U_4$.

The above results show that it is not possible to discriminate between classes 2 and 5, because adding the information given by all the attributes, the obtained information is smaller than the necessary for the accepted mutual classification error.

This information can be used as training values from where the rules are developed and can be applied for further received information under the same conditions.

## 4. OBTAINED RESULTS WITH THE PROGRAM

Based on the algorithm, a program has been developed. Several databases have been tested. The initial information for the iris database, taken as an example, is obtained from the original row database (Fisher, 1936), and is shown in Table 3, where the minimum and maximum values are obtained subtracting or adding (2s ) to the average value. The results for the runs with a = 0.1, and a = 0.31 are shown in Figure 2.

As can be noted, in the case of a = 0.1, only one attribute (PL), is necessary for identifying "setosa". The types "versicolor" and "virginica" cannot be differentiated using this small misclassification error. When the permissible error is bigger (a = 0.31), the number of necessary attributes for discriminating the three types is increased to two (PL and PW). In the last case, all the objects can be discriminated with the selected attributes, but the discrimination is subjected to a bigger error. The rules can be extracted from the program. For example, for a = 0.1, the following rule can be expressed:

*R*ule # 1: IF (*PL)* ∈ [1.11; 1.31]; THEN it is *Setosa*

The rest of the rules, for the different cases, can be extracted as well from the computer results as shown in Figure 2.

*Table 3: Iris Database and Attributes for Each Class*

| **Attribute** | **Setosa** | | | |
|---|---|---|---|---|
| | $x_{av}$ | s | Min | Max |
| SL | 5.0 | 0.35 | 4.3 | 5.7 |
| SW | 3.42 | 0.38 | 2.66 | 4.18 |
| PL | 1.45 | 0.11 | 1.24 | 1.67 |
| PW | 0.24 | 0.11 | 0.02 | 0.46 |
| **Attribute** | **Versicolor** | | | |
| | $x_{av}$ | s | Min | Max |
| SL | 5.94 | 0.52 | 4.9 | 6.98 |
| SW | 2.77 | 0.31 | 2.15 | 3.39 |
| PL | 4.26 | 0.47 | 3.52 | 5.2 |
| PW | 1.33 | 0.20 | 0.93 | 1.73 |
| **Attribute** | **Virginica** | | | |
| | $x_{av}$ | s | Min | Max |
| SL | 6.59 | 0.64 | 5.31 | 7.87 |
| SW | 2.91 | 0.32 | 2.27 | 3.55 |
| PL | 5.55 | 0.55 | 4.45 | 6.65 |
| PW | 2.03 | 0.27 | 1.49 | 2.57 |

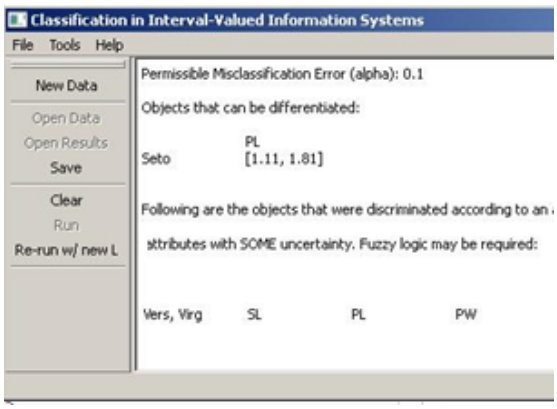
## 5. CONCLUSIONS

An algorithm and a program have been developed for attribute reduction in a classification process, using the concept of information (or loss of information) in diffuse databases. In this paper a logarithmic form for expressing the uncertainty in the differentiation between two classes has been used. One of the advantages of the method is that from it can be easily to be determined how far the solution is for each established misclassification error, as well as finding the minimum number of necessary parameters in a simple way. Another advantage is that it is easily seen from the table whether it is possible to discriminate between two classes or not.

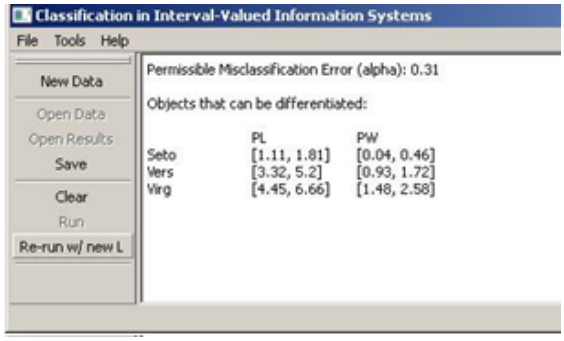
The obtained rules serve as a model, and the database can be used for training purposes for discriminating from the further received information from objects in the same process, under the same conditions.

In the presented case using the program, the discrimination is possible for all the objects if the misclassification error is big (a = 0.31), a more precise classification is possible making smaller the misclassification error (a = 0.1), but in this case not all the objects can be uniquely classified.

In this case as can be seen from the results obtained from the program, the discrimination is subjected to some uncertainty. The program gives the results indicating that it is not possible to discriminate the between the classes virginica and versicolor for small misclassification errors. Under such restrictions the simplest way is to apply fuzzy logic for finding the compatibility index for each object and attribute.




(a)




(b)

*Fig. 2. Obtained Results with the Program. a) a = 0.1, b) a = 0.31*

## REFERENCES

Ying Chieh, T., *et al.* (2006). Entropy-Based Fuzzy Rough Classification Approach for Extracting Classification Rules, *Expert systems with Applications*, No. 31, pp. 436-443.

Battiti, R. (1991). Using mutual information for Selecting Features in Supervised Neural Net Learning, *IEEE Transactions on Neural Networks*, No. 5, pp. 537-550.

Granger, E., *et al.* (2000). *Classification of Incomplete Data* Using *the Fuzzy ARTMAP Neural Network*, Technical Report. Boston U., Center for Adaptive Systems and Dept. of Adaptive Neural Systems, January.

Caballero A., *et al.* (2010). A practical Solution for the Classification in Interval-Valued Information systems, WSEAS Transactions *on Systems and Contro*l, Issue 9, Vol. 5, September, pp.735-744.

Leung, Y., *et al.* (2007). A Rough Set Approach for the Discovery of Classification Rules in Interval-valued Information Systems, *Int'l J. of Approximate Reasoning*, No. 47, pp. 233-246.

Fisher, R. (1936). *The use of Multiple Measurements in Taxonomic Problem*, Annals of Eugenics Cambridge University Press, V. 7, pp.179-138.