PREPROCESSING "RAW" DATA SETS AS AN IMPORTANT ASPECT OF INTELLIGENT INFORMATION PROCESSING

Sergii Konovalenko

***Abstract**: The paper highlighted the problems of selection methods for the preparation of the "raw" data for subsequent processing in the systems of intelligent information analysis. Each application domain contains a lot of different types of identification characteristics. Most algorithms are unable to work directly with all types and formats of data. Therefore, the preparation or the transformation of the input data set is an integral part of analysis system design. In the paper discusses methods of transformation of continuous and discrete types of data useful for analysis of the vector set. On the example of the domain "Information customs control" has been shown as configured training set for recognition of risks violation of customs legislation, based on a neural network the type multilayer perceptron. Were also considered methods of forming pseudographic patterns from the input sequence data initially did not a graphic of origin.*

***Keywords**: preprocessing, transformation data*

***ACM Classification Keywords**: D.2 SOFTWARE ENGINEERING – D.2.12 Interoperability – Data mapping*

Introduction

Building Systems intelligent information processing provides themselves stage of the preparing input data. Since in most cases, the information processing system receives the input data for analysis from various sources, there is a need to bring them to a suitable format for consideration. The primary source of data storage and may serve a database of commercial and government organizations, submitted documents, the Internet, i.e., as much as possible information that might be useful for decision making [Byuyul, 2005]. Given the fact that intelligent systems have the properties of learning, it is important to pay attention to the pre-treatment and preparation of input data sets. Failure to do so, we deteriorate the quality of information analysis system (pattern recognition, classification, etc.), and in some cases it will even make it impossible to adequately perceive the input vector data. For example, in order to train a multilayer perceptron (MLP) address a specific problem, the problem must be formulated in terms of a set of input vectors $x = [x_0, x_1, ..., x_i]^T$ and their associated reference output values $y = [y_1, y_2, ..., y_j]^T$ (standards) [Swingler, 1996]. Almost any set of domain identification characteristics of classes of information processed, is polytypic character which allows to allocate a for the consideration of actual problem as a choice of methods of preparation and transformation of its input data set for subsequent correct processing to the system analysis.

Problem definition

The purpose of this paper is consideration of the theoretical and practical application of methods of preparing input data for data mining systems, in connection with which there is a need in the following tasks:

1. To consider and to group the main methods and means of preparing the "raw" data;
2. Give an example in the domain of "information of customs control";
3. Identify possible methods for the preparation of data for training the neural network classifier.

Preprocessing of the "RAW" dataset

The functioning systems of intellectual information processing involves himself several important steps (Fig. 1):

1. Getting information from the external (internal) sources, e.g. reception;
2. Preparing Data (preprocessing);
3. Information processing;
4. Interpretation of results (postprocessing).

As mentioned above, preparation of input data, or the so-called training set (x, y) is one of the most important aspects in the creation of systems analysis. The quality of the training set has a strong influence on the model's ability to perform tasks (e.g. neural networks) [Swingler, 1996].

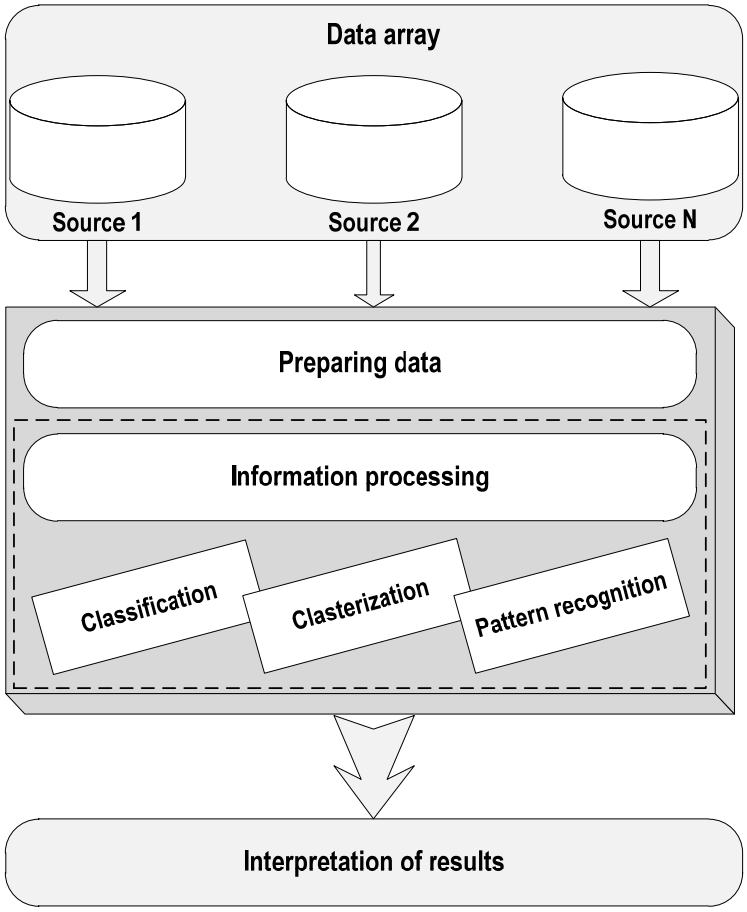

Fig. 1. Model of the information processing system

As seen in Figure 1 is the last stage of postprocessing of data, i.e. interpretation of the results. Different methods of analysis can provide information on various types of output data. In the simplest case – a set of digits (output vector), and more complex – models and rules. Therefore, it is also important to make the interpretation of this result is used in the context of the subject area. This stage performs the reverse conversion of data from the preprocessing task.

What represents the preprocessing of input data? Preprocessing is the process of preparing data for analysis, during which they are brought into conformity with the requirements determined by the specific problem to be solved (subject area) and use the model processing (analysis), the information received. As a rule, preprocessing of data includes two directions [BaseGroup]:

1. Cleaning and optimization;
2. Transformation and normalization.

Cleanup is performed to eliminate the factors that reduce the quality of data and hamper the work of analysis algorithms. It involves the processing of duplicates, inconsistencies, and fictitious values, restoration and filling gaps, smoothing and cleaning of data from the noise suppression and editing of anomalous values. In addition, the cleaning process restores violations of the structure, completeness and integrity of the data are converted incorrect format. Optimization of the data as an element of preprocessing includes reducing the dimension of input data, the identification and exclusion of irrelevant attributes. The main difference between optimization of the cleaning is that the factors that are fixed in the cleaning process, significantly reduce the accuracy of the solution of the problem or make it impossible to work analysis algorithms. Problems solved with optimization, adapting the data to a specific problem and increase the efficiency of their analysis.

Table 1. Methods of data preprocessing

Method	Problem	Solution
Cleaning the data	*Contradictory information*	1. Delete records; 2. Correct entries, selecting the most probable event.
	Gaps in data	1. Approximation (ordered sets of data); 2. Identification of the most verisimilar values (unordered information).
	Abnormal values	1. The value is removed; 2. The value is replaced by the nearest boundary value.
	Noise	1. Spectral analysis (cleaned frequently and marginal variations in some of the main signal); 2. Autoregressive methods (removal of noise from the signal describing function).
	Data input errors	1. Format-logic control
Optimization of data	*Reducing the dimension of input data (the identification and exclusion of irrelevant features)*	1. The method main components; 2. Multidimensional scaling; 3. Neural network techniques (Hopfield network, Kohonen network); 4. Using the analysis of the entropy; 5. Auto-associative networks.

As for the transformation and normalization of data, this step is necessary to bring the information to understand the terms used by the analytical model. This includes operations such as casting, quantization, coding, and so on. Each method of analysis requires that the original data were in any particular form. For example, neural networks only work with numeric data, and they should be normalized [Haykin, 1998].

Table 2. Methods of data preprocessing

Kind of information processing	Method	Solution
Data Transformation and Normalization	*Type conversion*	Convert a variable of one type to another type of value.
	Coding	It is used to encode qualitative data types or ranges of numerical types.
	Scaling	1. Decimal scaling; 2. Minimax normalization (2); 3. Normalizing the standard deviation (2).
	Quantization	1. The homogeneous (linear) quantization (3); 2. Quantization on the level

There are several types of data. Each of them is treated differently. Consider the basic data types (Fig. 2).

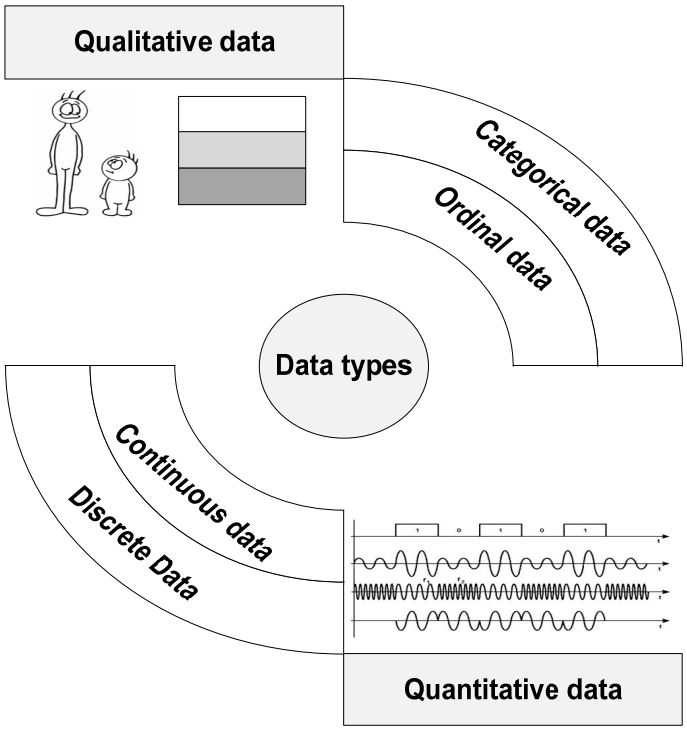

Fig. 2. Categories of the data types

The qualitative data. They represent some of the properties of objects that can't quantify (e.g., color, gender, person, etc.).

Quantitative data. They represent some of the properties of objects that take numeric values. Quantitative information in the analysis can be distributed on a scale, breaking it into equal intervals.

Consider using as a model for intelligent information processing neural networks, where there is a need to bring a training sample to the form that defines the activation function of computing neurons in the network [Bishop, 2006]. The choice of data for network training and processing is the most difficult stage of the solution. A set of training data must satisfy several criteria [BaseGroup]:

1. Representativeness – the data are intended to illustrate the true state of affairs in the subject area;
2. Consistency – is conflicting evidence in the training set will lead to poor quality of the learning network.

These criteria provide for himself the whole complex of actions preprocessing of different types of input data. As a rule, neural networks like multilayer perceptron using sigmoid activation function. That is, the input vector $x = [x_0, x_1, ..., x_i]^T$ must be brought to the range [0...1] or [-1...1] by (1) or (2).

$$x_i = \frac{x_i}{\max(X)}. \quad (1)$$

$$x_i = \frac{x_i - \mu}{S}, \quad (2)$$

where S – range $(\max(X) - \min(X)$ or standard deviation, μ – average value.

If you need to partition the continuous value into segments of equal length, the quantization can be performed as the initial value of the division by a constant value (quantization step) and the integral part of the quotient:

$$y_q = \frac{y - y_0}{h}, \quad (3)$$

where h – quantization step.

Qualitative data types shall be coded as can't be directly fed to the input of data processing systems. Typically, to each value is associated with a specific numeric value in the required range. For example:

Table 3. Example coding of the variable "Color"

Variable	Source value	Coding value	
		Decimal number	Binary number
Color	*Red*	1	00
	Blue	2	01
	Green	3	10
	White	4	11

Similarly, to encoded a vector output signal of the neural network.

In such a way, we examined the methods of preparation and transformation of the original "raw" data for systems analysts and data processing.

Preparing training sets "Information of customs control"

On the example of the domain "Information customs control" [WCO] will preprocessing training set for recognition of risks violation of customs legislation, based on of neural network the type multilayer perceptron [Moroz, 2011].

For example, the goods had been taken such a category as "Microcontrollers and Microcomputers" (8542 21 50 00 - Number of Classifier). An example of the process of forming the input vector is presented in the Table 4.

The components of the input vector are encoded by the following principle:

1. The variable X_0 is encoded binary numbers;
2. The variables X_1… X_5 are transformed by the formula (1) to the range [0 ... 1];
3. Variables X_6… X_7 take only three values, so we put them in compliance with three numeric values of {0, 0.5, 1}. These values match the level of risk {"Low," "Moderate," "High"} [ASYCUDA].

Table 4. Forming the input vector "Information of customs control"

№	Identification characteristics	Data types	Accepted values	Coding value
X_0	Country of Origin	string	Offshore	00 (bin)
			The EU countries	01 (bin)
			The EEA countries	10 (bin)
			other countries	11 (bin)
X_1	Product Code	integer	In accordance with the classifier	Range [0…1]
X_2	Customs cost	float	In accordance with the customs declaration	Range [0…1]
X_3	Quantity of goods	integer	Number of units or batches of delivery	Range [0…1]
X_4	Weight of goods	float	Weight unit of goods or the supply of the party	Range [0…1]
X_5	Invoice cost of a product	float	In accordance with the customs declaration	Range [0…1]
X_6	The difference gross and net product	float	no more 5%	0
			from 5% to 8%	0.5
			more than 8%	1
X_7	The history of the participant of foreign economic activity	string	black list	0
			gray list	0.5
			white list	1

In such a way input vector was transformed to a common format and range – [0 ... 1], which is suitable for the activation function. Now it can be input into the neural network used for training and classification.

Another way to prepare the input data is a representation of the input vector as a graphical image. Consider a vector X of 7 items and transform them into pseudographic dimension image of 7x7 (Fig. 3).

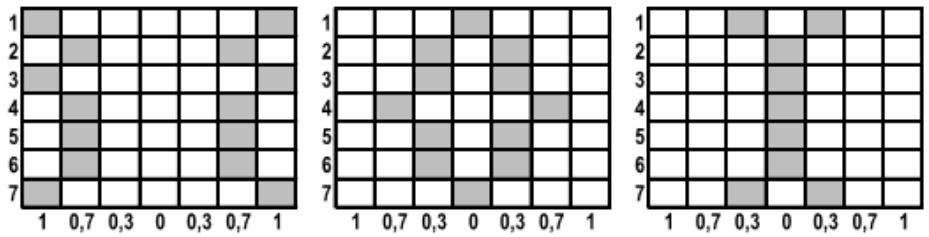

Fig. 3. Pseudographic patterns: {"High","Moderate","Low"}.

Graphic matrix is horizontally symmetrical range of risk levels (1; 0.7; 0.3; 0; 0.3; 0.7; 1), and vertically – the elements of our input vector (1...7). Each value is the input vector is estimated in accordance with the profile and level of risk (the anomaly) from 0 to 1. And then painted the cell at the intersection of the corresponding element

of the vector and its significance level of risk. The symmetry of the level of risk necessary for a better perception of the network image. In Fig. 3 shows examples of the formation of the input vector in pseudographic the first column corresponds to a “high” level of risk, the second and third – the “moderate” and “low”.

Then these symbols are the inputs to the neural network learning and recognition.

As a result of the creation of graphic images, improved quality and representativeness of the training set, both for neuroclassifier, and for the system designer analysis.

Conclusion

As a result of this work have been considered theoretically methods of preparation of input data.

The paper discusses methods for the conversion of continuous and discrete types of data suitable for analysis of the vector set. On the example of the domain "Information customs control" has been shown how a training set for recognition of risks violation of customs legislation, based on the type of neural network multilayer perceptron. Were also considered methods of forming pseudographic images from the input sequence data was not originally a graphic of origin.

Application of methods of preparation of the data allows you to turn "raw" data into high-quality training set, which is adequately and correctly displays the domain.

Bibliography

[Byuyul, 2005] A. Byuyul, P. Tsefel, SPSS: Art of information processing, (in Russian), DiaSoft, 2005.

[Swingler, 1996] Kevin Swingler. Applying Neural Networks: A Practical Guide. Morgan Kaufmann; Pap/Disk edition, 1996, p. 303.

[BaseGroup] BaseGroup Labs. Technology of data analysis [Online]. Available: http://www.basegroup.ru

[Haykin, 1998] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice Hall, 1998.

[Bishop, 2006] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), 1st ed. New York: Springer-Verlag, 2006.

[WCO] World Customs Organization. WCO course on risk assessment, profiling and targeting [Online]. Available: http://www.wcoomd.org

[Moroz, 2011] B. Moroz, S. Konovalenko, “Aplication of neural networks within of the concept E-customs,” (in Russian), in Information Science and Computing, book 22, Applicable Information Models, K. Markov, V. Velychko, Ed., 1st ed. Sofia, Bulgaria: ITHEA, 2011, pp. 104-110.

[ASYCUDA] Automated SYstem for CUstoms DAta (ASYCUDA). Risk Management [Online]. Available: http://www.asycuda.org

Authors' Information

Sergii Konovalenko *– Ph.D. student, inspector of customs service of 3rd rank, head of the laboratory of information systems and processes in customs affair of the Ukrainian Academy of Customs, 2/4 Dzerzhinsky str., Dnipropetrovs'k, 49044, UKRAINE; e-mail: customslab@rambler.ru*

Major Fields of Scientific Research: computation intelligence, machine learning