Classification methods and models for automatic determination of goods code by foreign economic activity goods nomenclature

J.U. Sevinov and G.R. Khamroev

Tashkent State Technical University, Tashkent, Uzbekistan

Abstract. Classification methods and models for automatic determination of goods code according to the commodity nomenclature of foreign economic activity (CN FEA) are considered. In the classification procedure, the object is reduced to one or more classes based on the results of the comparison to evaluate the proximity and make a conclusion. The relationship between vectors calculated by scalar multiplication was used as a measure of proximity. In order to determine the most informative symbols for classes, linguostatistical methods based on the information about the a priori probability of occurrence of terms and symbols were used. In doing so, macro-average and micro-average methods, which are considered effective in assessing the quality of classification for several classes, were used. Also, on the basis of syntactic and linguostatistical analysis, a generalized scheme of the process of automatic classification of goods in column 31_1 of the goods declaration graph is proposed. This recommended method of automatic identification of the CN FEA code allows participants of foreign economic activity to ease the processes of filling out declarations for goods.

1 Introduction

The main part of the customs declaration is the reliable classification of goods. Operations on the classification of goods include not only making a decision on classification in the field of determining its customs-tariff position, but also determining whether the goods comply with the restrictions and prohibitions set for notarized goods, making a final decision on the goods, as well as qualifying the deed of the entity transporting the goods across the customs border. is also important for [1,2].

In the modern market, there is a sufficient amount of software designed for customs clearance of goods. When filling out the 33 columns of the customs declaration for goods that we use, their interface, in the best case, suggests the use of the relevant reference CN FEA and its explanations, which are explained to it in the applicable section [3].

Customs practice has shown that the analysis and systematization of classification errors, one of the necessary conditions for their occurrence, is the weak methodological support of the software that provides the identification of goods [2,4].

* Corresponding author: sevinovjasur@gmail.com
This article discusses the methods and classification models used in the automatic identification of the commodity code according to CN FEA, a hierarchical classifier of commodity codes. The product code is a 10-digit number. The first two digits of the number indicate the commodity group of the CN FEA, the next two the commodity position, the next two the subheading, and then the subheading. In the work, a set of goods with the same CN FEA code is considered as a class.

2 Objects and methods

The mathematical model of the classification problem can be expressed as follows [5,6]. A set of \( L \) objects, also \( \{ \ldots \} \), there is a collection \( N \) of classes, here \( 1, \ldots, iN \). Each object corresponds to one or more classes. Each \( i \) class is somewhat formalized \( S \) is represented by the description.

The classification procedure consists in making changes to them, based on this \( l \) one or more \( i \) \( S \) fits the description, \( l \) comes to the conclusion that it belongs to the class. Then the classification process can be represented by an algebraic system of the following form:

\[
< L, K, S, R, f >,
\]

where:
- \( L \) – a set of objects to be classified;
- \( K \) – a collection of classes;
- \( S \) – a set of class descriptions;
- \( R \) – a set of relationships has the following characteristics: any class corresponds to one description, and the reverse requirement is not necessary;
- \( f \) – the classification operation in a view.

There are no restrictions on the description of, so there are cases where some objects can be assigned to more than one class at the same time.

In addition to the problem of structured classification, there is the problem of training classification, which is based on some a priori information. 

As a measure of closeness, for example, the cosine of the angle between vectors calculated by scalar multiplication can be used:

\[
\text{Cosine} = \frac{\text{Vector}_1 \cdot \text{Vector}_2}{||\text{Vector}_1|| \cdot ||\text{Vector}_2||}
\]
Based on the above, it is necessary to create a description of objects to solve the problem of automatic classification. Typically, classification algorithms deal with vectors in the feature space. The quality of the classification largely depends on how the objects are converted into a vector image.

Approaches to the formation of the feature vector of the object can be significantly different. In the simplest case, each character corresponds to the character of one of the objects in the considered set. The value of the corresponding coordinate of the vector can also be determined in different ways: for example, if the given attribute belongs to the object, the value can be equal to one, and in the opposite case it is equal to zero; can be calculated using more complex formulas. The choice of feature weights significantly affects the quality of object classification.

Most classification algorithms work with a "character-object" matrix consisting of feature vectors of objects as input. This matrix is used to construct a proximity matrix, and the proximity matrix is used to determine the set of objects most similar to the data [7,20].

There are several reasons for trying to reduce the size of the character field, such as:

- in the classification of objects in the field of a given subject, signs unrelated to this field can hide similarities between objects;
- the high size of the character space reduces the performance of the classification algorithm.

The following methods can be used to reduce the size of the character field, in particular:

- remove the marks related to all considered classification objects;
- use of linguistic methods: grouping word forms from dictionaries and theories according to normal forms and combining normal forms into synonymous groups. An extended version of the same method can be based on the use of semantic network [8] and group terms based on more complex types of relationships. Some classification algorithms use not individual terms as elements of the document vector, but separated nominal or verbal groups, related names, set phrases [9]. Various methods of resolving homonymy and polysemy can be used as assistants to clarify the values of the components of the document vector [8,10,11];
- use of linguostatistical methods based on information about the a priori probability of occurrence of terms and signs to determine the most informative signs for classes, which can be calculated according to the following formula:

\[
MI(b_k | K_i) = \sum_{b_k \in \{\square\}} \sum_{K_i \in \{\square\}} P(b_k | K_i) \frac{P(b_k | K_i)}{P(b_k | P(K_i))}, \quad (3)
\]

where:

- \(f_i\) – number of objects containing characters;
- \(f_i\) – the number of objects included in the class;
- \(f_i\) – the number of all objects;
- \(P(b_k | K_i)\) – signs and \(P(b_k | P(K_i))\) joint distribution probability of the class.

\[
r_{ij} = \frac{(S_i | S_j)}{\sqrt{|S_i| \times |S_j|}}, \quad (2)
\]
So, if \( k \) belongs to \( K \), and if the distributions of the class are statistically independent, in that case \( M(I | b_k K) = 0 \).

If there is a functional relationship between characters and the occurrence of a class, in that case \( M(I | b K) \) is the maximum.

Many variance estimates (or significance coefficients) are based on a measure called the inverse frequency of sign, which is given by the following expression [11]:

\[
\omega_i = \frac{n_i}{f_i}, \quad 0 \leq \omega_i \leq \frac{n_i}{N}, \quad (4)
\]

Here \( n_i \) – total number of classified objects; \( f_i \) – frequency of character \( i \).

Such a scale weights the entire character field. \( 0 < \omega_i < 1 \) has been divided into clusters. At the same time, \( \omega_i = 1 \) if \( f_i = \frac{n}{N} \) divide into clusters. At the same time, \( \omega_i = 1 \) if \( f_i = \frac{n}{N} \) characters with fall into the last cluster – the most "thematically specific" symbol cluster, and the first cluster contains the most common symbols for the set of objects.

There are several indicators for evaluating the effectiveness of automatic classification methods [2,13]. The most commonly used are precision \( (p) \) and recall \( (r) \), which are also used to evaluate the quality of information retrieval. The following values are used to determine the completeness and accuracy of the classifier:

- \( a \) – the number of correctly classified objects;
- \( b \) – the number of incorrectly classified objects;
- \( c \) – the number of incorrectly rejected objects.

The method of classification as correct and incorrect classification refers to cases where the analyzed object belongs to a certain class, which is accepted by some expert as a correct or incorrect decision, respectively. In false object rejection, the classifier does not include the object in a class that the expert believes is incorrect.

It follows that completeness is the ratio of correctly classified objects to the total number of objects belonging to a class, and precision is the ratio of correctly classified objects to the total number of objects belonging to a class. Values are usually measured as percentages, so for an ideal algorithm the recall and precision would be 100%.

To evaluate the classification quality for several classes, two averaging methods are used: macro-averaging and micro-averaging [14,19]:

\[
\begin{align*}
\text{reconmar} & = \frac{\sum_{i=1}^{n} a_i - c_i}{n} \quad \text{precmar} = \frac{\sum_{i=1}^{n} a_i} {n} \\
\text{reconmic} & = \frac{\sum_{i=1}^{n} a_i - c_i}{\sum_{i=1}^{n} a_i + b_i} \quad \text{precmic} = \frac{\sum_{i=1}^{n} a_i} {\sum_{i=1}^{n} a_i + b_i} 
\end{align*}
\]

(5)

The macro average reflects the efficiency of the method on average classes, so it is often used.

There are also integral estimates of classification methods constructed using recall and precision. The most famous of them are:

\[
F = \frac{2rp}{r + p} \quad (7)
\]
more generally defined as:

\[ F_\beta = \frac{(\beta + 1) pr}{\beta p + r} \]

One of the problems that arise when evaluating the quality of an object classification method is the selection of a set of tests. As a rule, this is a set of classified objects on which an automatic classification algorithm can be trained. At the same time, it should be taken into account that it is not possible to obtain reliable quality estimates by testing the method on the same set of objects used for training.

There are different ways of forming the training and test sets, and for successful training of an algorithm, the classes should be represented in approximately the same proportion in the training set. However, if there is insufficient data or the partitioning process fails to form the training set, one of the classes may be dominant. This can lead to "mistakes" in the learning process, and the dominant class is considered the most [6,15]. The cross-validation method allows you to avoid this.

The proposed method uses a vector representation of an object and a class [7,16]. The description of object \( I \) is presented in the following form:

\[
I = \begin{pmatrix}
I_{1k} \\
\vdots \\
I_{nk}
\end{pmatrix}
\]

If character belongs to -object; 1 otherwise; 0.

Let's consider the vector matrix line as a description of the classes:

\[
S = \begin{pmatrix}
S_{11} & \cdots & S_{1j} & \cdots & S_{1D} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
S_{T1} & \cdots & S_{Tj} & \cdots & S_{TD}
\end{pmatrix}
\]

It is proposed to use one of the similarity indicators calculated on the basis of statistical indicators as a similarity coefficient. It can be the correlation coefficient, Sorala, Djakkarda, Rodgers-Tanimoto, Daysa and other well-known variants of vector proximity measures.
\[ S \times I = I \]  

(10)

\[ \bar{b}_i = \sum_{j=1}^{D} s_{ij} b_j \]

(11)

\[ r = \max_i b_r \]

The automatic classification process is represented in Figure 1 by three steps. In the first step, a classification is initially created, which is a set of descriptions for each class. Based on the processing of a set of declarations for goods, column 33 is determined using column 31_1, which contains the CN FEA code of the declared goods and its detailed description.

At the next stage, based on the obtained classifier, a "character-class" proximity matrix containing the values of statistical proximity coefficients is created, which are calculated using the indicators of the joint and separate occurrence of signs and classes.

The third stage is the actual classification of goods. For declared goods, a description is drawn up on the basis of column 31_1 of the customs declaration, which is represented by a set of symbols [3,17,21]. Next, a quantitative assessment of the closeness of the description of the goods to the CN FEA classes is carried out. For this purpose, the general affinity coefficient of the symbols included in the description of the declared goods is calculated using the "character-class" proximity matrix built in the previous step. Choosing the maximum value of the common affinity determines whether it belongs to one of the classes.

As mentioned above, the effectiveness of automatic classification largely depends on the correct and complete filling of column 31 of the customs declaration describing the goods. An uneducated user cannot perform such a task. Therefore, this method includes, as a result, determining several most suitable CN FEA codes and offering the user to indicate the class of goods to which the user belongs, or to specify identifying marks that allow the algorithm to determine the only suitable commodity code [3,18].

The direct interaction of users with information resources, without intermediaries, is the main important factor that determines the directions of development of modern information systems, including customs [22]. This forces the creators of information systems to pay more and more attention to automation tools in the performance of various functions. The proposed method of automatic identification of the CN FEA code is intended for exactly such purposes - to facilitate the work of participants of foreign economic activity in filling out declarations for goods.
3 Conclusion

The methods and classification models used in the automatic identification of the commodity code by CN FEA, a hierarchical classifier of commodity codes, and the set of commodities with the same CN FEA code as a class were considered. The presentation of an effective mathematical model of customs inspection and classification of goods at customs inspection was shown. A description of the three-stage automatic classification process of classification methods and models for automatic determination of the goods code according to the foreign economic activity commodity nomenclature was created.

References
