Expert Systems with Applications 41 (2014) 4752–4761

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Addressing remitting behavior using an ordinal classification approach

Pilar Campoy-Muñoz [a,*], Pedro Antonio Gutiérrez [b], César Hervás-Martínez [b]

[a] Department of Economics, Universidad Loyola Andalucía, Escritor Castilla Aguayo 4, 14004 Córdoba, Spain
[b] Department of Computer Science and Numerical Analysis, University of Córdoba, Campus de Rabanales, Edificio Albert Einstein, Planta 3, 14071 Córdoba, Spain

ARTICLE INFO

Keywords:
Nominal classification
Ordinal classification
Support Vector Machine
Remittances

ABSTRACT

The remittance market represents a great business opportunity for financial institutions given the increasing volume of these capital flows throughout the world. However, the corresponding business strategy could be costly and time consuming because immigrants do not respond to general media campaigns. In this paper, the remitting behavior of immigrants have been addressed by a classification approach that predicts the remittance levels sent by immigrants according to their individual characteristics, thereby identifying the most profitable customers within this group. To do so, five nominal and two ordinal classifiers were applied to an immigrant sample and their resulting performances were compared. The ordinal classifiers achieved the best results; the Support Vector Machine with Ordered Partitions (SVMOP) yielded the best model, providing information needed to draw remitting profiles that are useful for financial institutions. The Support Vector Machine with Explicit Constraints (SVOREX), however, achieved the second best results, and these results are presented graphically to study misclassified patterns in a natural and simple way. Thus, financial institutions can use this ordinal SVM-based approach as a tool to generate valuable information to develop their remittance business strategy.

1. Introduction

Over the last decades, remittance flows have increased dramatically and in parallel to the surge in international migration. In the 1990–2013 period the estimated number of immigrants in the world grew from 154.2 million to 231.5 million (United Nations Population Division, 2013), a 1.5 fold increase; while the amount of remittances grew from US$64 billion to US$549 billion, a 8.6 fold increase (World Bank, 2013). This trend is expected to continue with an average annual growth rate of over 8%, with remittance flows reaching over $700 billion worldwide by 2016 (Ratha, Eigen-Zucchi, Plaza, Wyss, & Yi, 2013). However, these official figures are likely to understate total remittances due to the flows sent through informal channels, which could account for an increase of at least 50% more than recorded (Ratha, 2009).

Remittance flows result from the addition of thousands of individual transfers made by immigrants who send money back home to their relatives and friends (Adams Jr., 2009). Estimates from IFAD (2007) show that remittances to developing countries, which account for more than 70% of worldwide flows, are made through more than 1.5 billion separate financial transactions. The size of the flows and the fees for handling these transactions, which cost approximately 9% on average globally (Ratha et al., 2013), constitute a potential revenue stream for operators in remittance markets. This partially explains the increasing interest of banks to enter into the remittance business; a market largely dominated by a small number of Money Transmitters Operators (MTOs) who enjoy more than a 90% share (Alberola & Salvado, 2006). The other motivation is rooted in the potential to cross sell additional banking products to immigrants and the recipients of their remittances (Fransen & Andersson, 2012; Paulson, Singer, Newberger, & Smith, 2006; Suki, 2007).

To market their remittance services, banks must cater to the specific preferences of this new customer base with facilities such as local branches near immigrants' residences that have opening hours that are compatible with immigrants' work hours, tellers with language skills, and new products and approaches tailored to immigrants' needs. Financial institutions must also carry out advertising and street promotion at the local level because the immigrant collective does not respond to general media campaigns (Orozco, 2004). The implementation of this marketing strategy demands a considerable expense that can be recouped quicker by properly selecting the targeted group of immigrants, especially keeping in mind that not all immigrants are remitters and not all remitters send the same amount of money. In the latter respect, Orozco (2012) notes the relevance of immigrant profiles in determining the money sent but, to the best of our knowledge, only a few studies have attempted to determine the influence of

* Corresponding author. Tel.: +34 957 222 100.
E-mail addresses: mpcampoy@uloyola.es (P. Campoy-Muñoz), pagutierrez@uco.es (P.A. Gutiérrez), chervas@uco.es (C. Hervás-Martínez).

http://dx.doi.org/10.1016/j.eswa.2014.01.036
0957-4174/

immigrant's characteristics on their remittance levels (Hoddinott, 1992; Nziramasanga & Yoder, 2013). These prior studies are based on traditional statistical techniques, mainly ordered response models, which take into account the ordering nature of the dependent variable. However, based on their strict assumptions such as linearity, normality of variable distribution, and independent predictors, these traditional techniques are limited in their application to real-world problems.

On the other hand, classification models have been employed in market segmentation (Bloom, 2004; Suh, Noh, & Suh, 1999; Tsiotsou, 2006) as well as other retail banking services (Dinh & Kleimeier, 2007; Huang, Chen, & Wang, 2007; Malhotra & Malhotra, 2003) but not in the analysis of remitters. These models assign a set of observations or patterns into predefined groups or classes based on various features related to the patterns, encompassing both linear and non-linear approaches. Comparing Artificial Neural Networks (ANNs) and statistical models, most studies have highlighted that ANNs performed better. In credit risk classification, (Malhotra & Malhotra, 2003) proved that ANNs are more accurate compared to multiple discriminant analysis. Along the same lines, (Blanco, Pino-Mejías, Lara, & Rayo, 2013) also found out than ANNs yield better results than linear discriminant analysis and logistic regression. In spite of satisfactory results, ANNs have some disadvantages: difficulty in determining the optimal network topology, its "black box" nature, and the risk of overfitting. Conversely, the SVM approach overcomes the hurdles of overfitting and offers a unique globally optimal solution. In fact, Li, Shiue, and Huang (2006) proved that SVM outperformed ANNs and was more computationally efficient at evaluating consumer loans. In addition, Huang et al. (2007) found out that the SVM-based classifiers achieved an identical classificatory accuracy with relatively few input features.

Classification algorithms are developed to accommodate the nature of real world decision-making problems. While most of them have focused on predicting data labels from classes defined in a nominal way, new algorithms from the machine learning field are being formulated to address problems where there is an inherent ordering between classes (Chu & Keerthi, 2007; Fernández-Navarro, Campoy-Muñoz, Paz-Marín, Hervás-Martínez, & Yao, 2013; Hsieh & Hung, 2010). The application of standard nominal algorithms implies that all errors are treated equally when it is clear that the penalization of errors should increase as the distance between the predicted category and the real one increases (Gutiérrez, Pérez-Ortiz, Fernández-Navarro, Sánchez-Monedero, & Hervás-Martínez, 2012). This penalization should be taken into account both for classifier construction and evaluation. For example, if a classifier confuses an immigrant who does not remit money with one who sends the highest level of remittances, this type of error should be penalized more than confusing a non-remitter with an immigrant who remits a small amount of money. Ordinal regression algorithms tend to obtain classifiers that predict categories as closely as possible to the real ones. Another advantage of ordinal regression algorithms is that some of them allow patterns to be projected into a real line. In our case, this helps to identify those immigrants who are in the boundary between two levels of remittance to rank them in different classes.

Therefore, the primary objectives of the research are: (1) to compare the predictive ability of nominal and ordinal algorithms for classifying immigrants by level of remittances, using available data from national surveys; (2) to draw up the remitting profile of the different segments of remitters based on the models obtained by the best classifiers; and (3) to determine which of those immigrants are incorrectly classified and the reason for those errors, by analyzing the corresponding projections. Sample evidence from recent Ecuadorian immigrants in Spain, whose transfers represent roughly 30% of the total annual remittance inflow received by Ecuador (Central Bank of Ecuador, 2013), were used in research. The selected nominal and ordinal classifiers are some of the most widely used approaches in machine learning: MLogistic, SLogistc, MultiLayer Perceptron (MLP) and Radial Basis Function (RBF) neural networks, and the Support Vector Machine (SVM) with "1 *versus* 1" configuration as representative of the nominal SVM-based approach. Several extensions have recently been developed to apply SVM to ordinal outcomes. From these, two classifiers were selected: the SVM with the Ordered Partitions model, based on binary decomposition of the target variable, and the SVM for Ordinal Regression with Explicit Constraints, belonging to the threshold model category.

The remainder of the paper is structured as follows. The next section presents the definition of the problem and the performance measures for nominal and ordinal classifiers. Section 3 briefly describes the classifiers employed. In Section 4, the dataset, the experimental design, and the results are explained. Finally, Section 5 closes the paper with some concluding remarks.

2. Problem definition

The study of the remitting behavior of immigrants will be addressed as a classification problem. Let $\mathbf{x}$ be the input vector of immigrant K characteristics, where $\mathbf{x} \in X \subseteq \Re^K$, and y the label class vector of remittance levels, where $y \in Y = \{C_1, C_2, \ldots, C_Q\}$. The objective is to find a classification rule or function $f{:}X \to Y$ by using a training set $T = \{(\mathbf{x_i}, y_i),\ 1 \leqslant i \leqslant N_1\}$ of N_1 randomly selected patterns, so that for a given value of $\mathbf{x_i}$, the function f estimates the corresponding value of y_i, i.e., $f(\mathbf{x_i}) = y_i$, in such a manner that the function f minimizes an error measure in an independent generalization set $G = \{(\mathbf{x_i}, y_i),\ 1 \leqslant i \leqslant N_2\}$ within the remaining N_2 patterns, to ensure good performance of the function.

Taking into account the economic nature of the remittances, we can define different class labels based on the amount of money sent in such way that an ordered relationship exists between the labels ($C_1 \prec C_2 \prec \cdots \prec C_Q$). This ordered rank implies an additional restriction for the classification problem. As a result, the nominal multi-class classification problem turns into an ordinal classification one, which can be tackled by means of regression techniques. Under the regression approach, the differently labeled classes $\{C_1, C_2, \ldots, C_Q\}$ are cast into real values $\{r_1, r_2, \ldots, r_Q\}$, where $r_i \in \Re$, and then standard regression techniques can be applied. However, mapping this ordinal scale by assigning numerical values hampers the performance of the regression models because the distance between classes is unknown. Ordinal classifiers are specifically built to exploit the existing order among classes of the target variable. To do this, ordinal classifiers take into account the rank of the label, that is, the position of the label on an ordinal scale, which is usually expressed by the following function $O(C_q) = q$, $q \in \{1, 2, \ldots, Q\}$.

When the ordinal nature of the target variable is not obvious or has been defined *a posteriori*, just as in the case here where remittance levels have been established among immigrants, nominal classifiers can also be applied to ordinal problems and yield a better performance. However, the nominal ones completely ignore the ordering of the labels by considering them as independent classes; hence, they usually require larger training sets with respect to ordinal approaches (Kramer, Widmer, Pfahringer, & De Groeve, 2001).

The performance of both ordinal and nominal classifiers (f) can be evaluated by means of several metrics based on the contingency or confusion matrix calculated for the generalization set G. The confusion matrix $M(f)$ for a classification problem with Q classes and N_1 patterns is defined follows:

$$M(f) = \left\{ n_{ij}, \sum_{i,j}^{Q} n_{ij} = N_2 \right\} \tag{1}$$