Applied Soft Computing 36 (2015) 246–254

Contents lists available at ScienceDirect

Applied Soft Computing

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

Mixing numerical and categorical data in a Self-Organizing Map by means of frequency neurons

Carmelo del Coso, Diego Fustes, Carlos Dafonte*, Francisco J. Nóvoa, José M. Rodríguez-Pedreira, Bernardino Arcay

Fac. Informática, Universidade da Coruña, Campus deElviña, 15071 A Coruña, Spain

ARTICLE INFO

Article history:
Received 8 October 2014
Received in revised form 8 April 2015
Accepted 26 June 2015
Available online 29 July 2015

Keywords:
Self-Organizing Map
Categorical data
Mixed data
Big data

ABSTRACT

Even though Self-Organizing Maps (SOMs) constitute a powerful and essential tool for pattern recognition and data mining, the common SOM algorithm is not apt for processing categorical data, which is present in many real datasets. It is for this reason that the categorical values are commonly converted into a binary code, a solution that unfortunately distorts the network training and the posterior analysis. The present work proposes a SOM architecture that directly processes the categorical values, without the need of any previous transformation. This architecture is also capable of properly mixing numerical and categorical data, in such a manner that all the features adopt the same weight. The proposed implementation is scalable and the corresponding learning algorithm is described in detail. Finally, we demonstrate the effectiveness of the presented algorithm by applying it to several well-known datasets.

1. Introduction

Self-Organizing Maps (SOMs) are powerful unsupervised neural networks that provide simultaneous clustering and dimensionality reduction through the projection of a dataset into a two-dimensional lattice of neurons. SOMs are applied to a great many tasks, including data mining, pattern recognition, visualization, and data exploration. Not in vain has the original article from Kohonen [1] received more than 5000 citations worldwide. However, the original SOM algorithm was conceived for numerical datasets, since it was built on the Euclidean distance. This implies that many datasets with non-numerical data types, such as for example categorical values, cannot be analyzed with SOMs. There thus remains a big handicap for many real datasets that include categorical data, such as social, economic, and medical data.

Categorical data clustering could be considered as a special case of symbolic data clustering [2–4], in which categorical values are taken by the attributes of symbolic objects. However, most techniques used in the literature to cluster symbolic data are based on the hierarchical methodology, and are not efficient when clustering large datasets.

The problem of processing categorical data in machine learning and data mining was addressed by various authors, see [5,6]. Typically, the categorical variables are converted into binary codes, so that the original numerical algorithms can be applied to the transformed data. However, transformation into binary codes entails several issues, the main one being the increase of the dataset dimensionality, which complicates its storage and processing. Another approach consists in defining a new similarity metric that directly treats categorical data. On the one hand, this is an appealing solution in the sense that the data dimensionality is not increased, and that the results are easy to understand because the data are kept in the original format. On the other hand, it remains important to treat both numerical and categorical data. Several approaches exist in such direction, some of them dividing the dataset and dealing separately with the numerical and categorical parts, others using the combined data, but applying different algorithms for the numerical and the categorical parts. Some authors [7] recommend to build a classification of the observations restricted to the use of only the quantitative variables, applying a Kohonen classification followed by an Ascending Hierarchical [8] Algorithm, to define a new qualitative variable. Later, that new variable will be added to the other qualitative variables and it is possible to apply a Multiple Correspondence Analysis [9] or a KCMA [10] (Kohonen-based algorithm, which is analogous to the classical MCA) to all the qualitative variables.

Our work focuses on the algorithms that extend the numerical version of SOM to treat categorical data and integrate both, simultaneously, into the same procedure. In the field of clustering, and concretely that of SOMs, the issue of processing categorical data without the need of any binary coding was addressed in the works of Chen and Marques [11], Hsu [12], Hsu and Lin [13],

* Corresponding author.
E-mail address: dafonte@udc.es (C. Dafonte).

http://dx.doi.org/10.1016/j.asoc.2015.06.058
1568-4946/

and Lebbah and Benabdeslem [14]. They propose four different algorithms: NCSOM, CPrSOM, MixSOM, and GSOM, respectively. CPrSOM provides a probabilistic formalism for the SOM to treat categorical data. The authors claim that their method yields good results when applied to several real datasets. However, the proposed method is computationally hard and complex to implement. On the other hand, GSOM, MixSOM as well as NCSOM treat mixed numerical-categorical data. GSOM is based on the establishment of distance hierarchies, via domain experts, to unify the distances over numerical and categorical data. The GSOM authors set an online learning scheme, where the neuron prototypes are adapted by means of some heuristic rules. MixSOM also uses a hierarchical scheme: each attribute of training data, as well as its corresponding component of a MixSOM neuron, is associated with a distance hierarchy. The vector of a training instance and the prototype of a MixSOM are mapped to the hierarchies (aggregating distances between the corresponding mapping points). Finally, NCSOM is an extension of the SOM algorithm where the distance used for categorical variables is a direct matching procedure, i.e. 0 if the category is the same and 1 otherwise. The neuron adaptation is performed by computing the mode of each category within the objects assigned to the neuron and its neighbourhood. The NCSOM authors provide a clear and simple definition of the method as well as results with some well-known datasets from the UCI repository [15]. However, NCSOM gives more weight to categorical variables than to numerical variables and lacks deterministic behaviour, yielding poorer results.

This work presents a new algorithm, the Frequency neuron Mixed Self-Organizing Map (FMSOM), which is aimed at addressing the drawbacks present in the existing algorithms to train a SOM with mixed numerical-categorical data. FMSOM is based on the NCSOM algorithm, but incorporates the probability tables of the CPrSOM algorithm, and as such becomes able to train mixed data in an efficient and accurate fashion. FMSOM introduces a distance function that is more suitable when mixed data is used, since it gives the same strength to the categorical variables as to the numerical ones. In addition, FMSOM shows convergence after a number of iterations and its behaviour is deterministic after initialization, which is not the case with NCSOM. To demonstrate this, we present the results of the algorithm when applied to several known datasets.

The remainder of the article is organized as follows: Section 2 describes the presented algorithm, Section 3 exposes the algorithm implementation and its scalability through Map-Reduce, Section 4 presents the results obtained when the algorithm is applied to real datasets and, finally, Section 5 provides some discussion about the developed method.

2. The FMSOM algorithm

Classic SOM algorithms are conceived to be applied to numerical datasets and consequently do not provide good results when applied to categorical or mixed datasets. Based on the batch algorithm, the Frequency neuron Mixed Self-Organizing Map (FMSOM) creates a new model to deal with categorical features, which strongly improves the performance obtained by the SOM with categorical or mixed datasets. The main hypothesis is to preserve the original algorithm to treat the numerical part (after some preprocessing), while, for the categorical data, extending the neuron prototypes with a set of category frequency tables, one per categorical feature. The following subsections show the learning algorithm for this model, which is composed of three processes: competition, cooperation, and adaptation.

2.1. The competitive process

In the original SOM, for each input vector, the winner neuron is determined by the shortest geometric distance, using the Euclidean or Manhattan metric. Where for numerical datasets these measures are adequate, for non-ordinal features the results given by such metrics are incorrect. As these types of features lack ordering criteria, codification of values is completely arbitrary. It is therefore not possible to define a numerical distance between them, and the results obtained with such metrics are meaningless. For the aforementioned reasons, a new dissimilarity measure was created. The dissimilarity between numerical features is calculated as in the classic SOM algorithms, while the dissimilarity of categorical features is calculated on the basis of a probability measure.

P is the number of input vectors, I the number of map neurons, and F the number of features. We suppose that input vectors consist of n numerical features and $k = F - n$ categorical features, where $[\alpha_k^1, \ldots, \alpha_k^r]$ is the set of categories of the kth categorical feature. We mark $X_p = [x_{p1}, \ldots, x_{pF}]$, $p = [1, \ldots, P]$ as pth input vector and $W_i = [w_{i1}, \ldots, w_{in}, w_{in+1}, \ldots, w_{ik}]$ as the reference vector of the ith neuron, where $i = [1, \ldots, N]$ and $w_{in+1}, \ldots, w_{ik}$ are the probability vectors for each categorical feature. For the reasons stated above, the dissimilarity between an input vector X_p and a reference vector W_i is defined as the combination of its numerical and categorical dissimilarities, which are calculated independently following Eq. (1).

$$d(X_p, W_i) = Dn(X_p, W_i) + Dc(X_p, W_i) \tag{1}$$

The numerical dissimilarity is measured as in the classic SOM algorithms, using a numerical distance metric, such as the Euclidean distance (see Eq. (2)). To ensure that both numerical and categorical features have equal influence, numerical features are previously normalized in order to lie in the [0,1] range.

$$Dn(X_p, W_i) = \sqrt{\sum_{z=1}^{n} (X_{pz} - W_i)^2} \tag{2}$$

The categorical dissimilarity between an input vector and a reference vector is given by the sum of the partial dissimilarities obtained for each categorical feature. The dissimilarity for each categorical feature is measured as the probability of the reference vector not holding the category present on the input vector, given by the frequency table for the current feature and category.

$$Dc(X_p, W_i) = \sum_{z=n+1}^{k} (1 - W_{iz}[X_{pz}])^2 \tag{3}$$

2.2. The cooperative process

After computing the winner neuron $c(X_p)$, as described in the previous section, the cooperative process takes place. In this case the cooperation is similar to the common SOM batch algorithm. We use a Gaussian neighbourhood around the winner, defined as follows:

$$h(s, c(X_p), i) = \exp\left(\frac{-d^2}{2\sigma(s)^2}\right) \tag{4}$$

where $\sigma(s)$ is the neighbourhood radius for the actual epoch s, and d is defined as the number of connections that is necessary to continue in order to reach the neuron i from the winner neuron