ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Mutual information for feature selection with missing data

Gauthier Doquire and Michel Verleysen *

Université catholique de Louvain - ICTEAM/Machine Learning Group
Place du Levant 3, 1348 Louvain-la-Neuve - Belgium

Abstract. Feature selection is an important task for many machine learning applications; moreover missing data are encoutered very often in practice. This paper proposes to adapt a nearest neighbors based mutual information estimator to handle missing data and to use it to achieve feature selection. Results on artificial and real world datasets show that the method is able to select important features without the need for any imputation algorithm. Moreover, experiments also indicate that selecting the features before imputing the data generally increases the precision of the prediction models.

1 INTRODUCTION

Missing data are likely to occur in machine learning and data mining. The origin of this missingness can be as diverse as the dysfunction of an equipment, the too high resolution of a sensor, people refusing to answer personal questions in a survey or the impossibility to conduct some measurements on certain patients [1]. Here, we assume the data to be missing completely at random (MCAR). This means that the probability of finding a missing value for the random variable X is not related to the values of X nor to any specific feature in the dataset.

Even with missing values, the task of feature selection remains of great importance for many tasks such as regression [2]. Nevertheless, to the best of our knwoledge, few work has been done on this topic until now. This paper proposes a way to achieve feature selection without the need to first impute the data. Indeed, a feature selection algorithm should be able to select relevant features independently of any imputation procedure because relying on imputation would bring a bias whose effect on the feature selection would be really hard to estimate. That is why this paper suggests estimating mutual information with the well known Kraskov k-nearest neighbors estimator [3] *directly* from the data with missing values. More precisely, a forward selection scheme is used with the mutual information criterion to sequentially build the set of selected features.

The remaining of the paper is organized as follows. The basic concepts of mutual information are introduced in Section 2, together with the Kraskov estimator used in the experiments. The methodology followed as well as the results are detailed in Section 3. Finally some conclusions and perspectives for future work are given in Section 4.

*G. Doquire is funded by a Belgian F.R.I.A grant.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

2 MUTUAL INFORMATION

Mutual information (MI) has been introduced by Shannon in 1948 [4]. It is a symmetric measurement of the dependance between two random variables X and Y. Its interest in feature selection is mainly due to its capacity to detect non-linear relationships between variables and to handle groups of vectors [5].

2.1 Basic definitions

More formally, the MI of two random variables X and Y is defined as follows:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{1}$$

where $H(X)$ is the entropy of X, defined for a continuous random variable as:

$$H(X) = -\int f_X(\zeta_X) \log f_X(\zeta_X)\, d\zeta_X \tag{2}$$

whith f_X the probability density function (pdf) of X.

MI can be thus be rewritten as:

$$I(X;Y) = \int\int f_{X,Y}(\zeta_X,\zeta_Y) \log \frac{f_{X,Y}(\zeta_X,\zeta_Y)}{f_X(\zeta_X) f_Y(\zeta_Y)}\, d\zeta_X\, d\zeta_Y \tag{3}$$

Since in practice neither f_X, f_Y nor $f_{X,Y}$ are known, the MI can not be directly computed but has to be estimated from the dataset.

In this paper, a nearest neighbors estimator introduced by Kraskov et al. [3] is used since it is expected to be more robust than other popular estimators such as histograms based estimators or kernel density estimators when working with high-dimensionnal data. Indeed it does not require the estimation of probability density function by dimension-sensitive methods such as Parzen windows [6]. Moreover, it has already proven to be succesful in the no missing value case [7].

2.2 The nearest neighbors estimator

The estimator is based on the Kozachenko-Leonenko estimator of entropy:

$$\hat{H}(X) = -\psi(K) + \psi(N) + \log(c_d) + \frac{d}{n}\sum_{n=1}^{N} \log(\epsilon_x(n)) \tag{4}$$

where ψ is the digamma function, K the number of nearest neighbors considered, N the number of samples in X, d the dimensionality of these samples, c_d the volume of a unitary ball and $\epsilon_x(n)$ twice the distance from the n^{th} observation in X, $x(n)$, to its K^{th} nearest neighbor. See [8] for more details.

Calling $\tau_x(n)$ the number of points whose distance from $x(n)$ is not greater than $0.5 \max(\epsilon_x(n), \epsilon_y(n))$, some mathematical developpements then lead to two slighlty different estimators [3]. Only one is considered here:

$$\hat{MI}(X,Y) = \psi(N) + \psi(K) - \frac{1}{K} - \frac{1}{N}\sum_{n=1}^{N}(\psi(\tau_x(n)) + \psi(\tau_y(n))) \tag{5}$$

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

2.3 MI estimation with missing values

As already mentionned, in this paper the mutual information is estimated from the data with missing values, thus before any possible imputation. Moreover, Kraskov estimator (5) is totaly determined by $\epsilon_x(n)$ and $\epsilon_y(n)$, and thus by the distance from each sample to its K^{th} nearest neighbor in each space. Usually, the Euclidean distance is used on both the X and Y space.

Here this distance is adapted to handle missing values. More precisely, the distance between two samples is computed by taking into account only the features with no missing values in both samples. The distance is then normalized with respect to the number of features used to compute it. This normalization is essential to ensure that distances between different pairs of samples remain comparable. Otherwise, the distances would obviously be larger between samples containing less missing values.

As an example, the distance between the vectors [3 • 8 7 2] and [1 9 • 3 4], where • denotes a missing value, is computed by considering the first, fourth and fifth elements of each vector; it is be equal to $(\frac{(3-1)^2+(7-3)^2+(2-4)^2}{3})^{\frac{1}{2}}$. Note that the term distance has been used abusively in this section since the similarity measure defined here is not guaranteed to obey the triangle inequality.

3 METHODOLOGY AND EXPERIMENTAL RESULTS

In this section we combine the MI criterion estimated as described above with a forward search procedure and show the effectiveness of this combination for feature selection with missing data. More precisely, the first feature selected is the one whose MI with the output is the highest. Then, at each step, we select the feature whose addition to the set of already selected features leads to the group having the maximal MI with the output.

3.1 Artificial datasets

In order to show the capacity of the proposed algorithm to select relevant features, three artificial regression problems are built. All have a training set containing 10 variables $X_1 \dots X_{10}$ that are uniformely distributed over [0 , 1].

The first problem is derived from Friedman [9]. The output Y_1 is defined as

$$Y_1 = 10 \sin(X_1\, X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon \tag{6}$$

where ϵ is a Gaussian noise with unit variance.
The second and third one have an output respectively computed as:

$$Y_2 = X_1\, X_2 + \sin(X_3) + X_4 + 0.2\, \epsilon, \tag{7}$$

$$Y_3 = \cos(2\, X_1)\ \cos(4\, X_2)\ \exp(X_2)\ \exp(2X_3) + 0.2\, \epsilon. \tag{8}$$

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

The sample size is 1000 for all three datasets and the K parameter of the estimator is set to 6 as suggested in [3]. Obviously, only the five first, four and three variables are relevant for the prediction in (6), (7) and (8) respectively.

For each equation, 10 datasets are generated which are then randomly filled with 1, 5, 10 and 20% of missing values before we apply our algorithm. For all of the 120 datasets generated this way, the proposed methodology always selects the relevant variables first even when 20% of the values are missing. This proves at least the interest of the approach and moreover indicates that an imputation procedure does not seem essential in order to select relevant features.

3.2 Real world datasets

In order to further assess the quality of the method, experiments are then carried out on 4 real world datasets. The goal here is again to show that relevant features can be selected and that it is preferable to carry out feature selection before imputation. To this end, the performances of a regression model for which the missing values are imputed before feature selection are compared to those obtained with features chosen by the approach described in this paper. The imputation phase is still necessary since little work has been done, especially in regression, to make prediction models able to handle missing values.

The prediction model used in the experiments is a radial basis function network (RBFN). The number of neurons and the width of the Gaussian kernel are determined using a 10-fold cross validation procedure. Two very popular imputation methods are used: the 10-nearest neighbors imputation and a regularized expectation-maximization (EM) algorithm proposed by Schneider [10].

The first real database used is the Delve census dataset[1] for which the 2048 first entries are kept. The first 1500 are then used as the training set and the remaining as the test set. The dimension of the data set is 104.

The second dataset is the nitrogen data set, containing 141 spectra discretized at 1050 different wavelengths[2] . As a preprocessing, each spectrum is represented using its coordinates in a B-splines base to reduce the amount of features to 105 [11]. There are 105 data points in the training set and 36 in the test set.

The third one is the Housing dataset, available from the UCI Machine Learning repository[3] and consisting of 13 features. 338 instances are kept for training the model and 168 for testing it.

The last one is the Mortgage dataset. It consists in 16 features describing the Economic data information of USA from 01/04/1980 to 02/04/2000[4] .

From each of these complete datasets, 10 new datasets are generated by randomly introducing 1, 5, 10 and 20% of missing values. Then the forward feature selection is conducted on the training part of the imputed datasets and of the datasets with missing values, until a maximum number m of features is reached. The maximum number of features selected is arbitrarily fixed to half

[1] http://www.idrc-chambersburg.org/index.html
[2] http://kerouac.pharm.uky.edu/asrg/cnirs/
[3] http://archive.ics.uci.edu/ml/index.html
[4] http://www.stls.frb.org/fred/data/zip.html

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

%	EM before	EM after	KNN before	KNN after
1	0,687 ± 0,029	0,679 ± 0,029	0,706 ± 0,083	0,705 ± 0,041
5	0,729 ± 0,007	0,724 ± 0,007	0,792 ± 0,021	0,721 ± 0,242
10	0,752 ± 0,053	0,725 ± 0,023	1,019 ± 0,125	0,758 ± 0,029
20	0,772 ± 0,041	0,768 ± 0,037	0,830 ± 0,092	0,821 ± 0,339

Table 1: Prediction performances for the Nitrogen dataset with a RBFN model.

%	EM before	EM after	KNN before	KNN after
1	0,1819 ± 0,069	0,151 ± 0,052	0,171 ± 0,091	0,218 ± 0,098
5	0,2680 ± 0,075	0,266 ± 0,064	0,356 ± 0,157	0,266 ± 0,134
10	0,3068 ± 0,075	0,190 ± 0,062	1,006 ± 0,463	0,255 ± 0,110
20	0,3964 ± 0,178	0,224 ± 0,072	1,621 ± 0,284	0,507 ± 0,271

Table 2: Prediction performances for the Mortgage dataset with a RBFN model.

%	EM before	EM after	KNN before	KNN after
1	7,730 ± 0,610	7,634 ± 0,792	7,718 ± 0,453	7,606 ± 1,224
5	8,153 ± 0,934	7,039 ± 0,247	7,016 ± 0,739	6,995 ± 0,456
10	7,684 ± 0,398	7,373 ± 0,273	7,892 ± 2,121	7,488 ± 0,552
20	8,364 ± 0,464	7,646± 0,415	9,357 ± 0,990	7,611 ± 0,398

Table 3: Prediction performances for the Housing dataset with a RBFN model.

%	EM before	EM after	KNN before	KNN after
1	1,937 ± 0,090	1,892 ± 0,054	1,597 ± 0,111	1,902 ± 0,110
5	2,148 ± 0,052	1,911 ± 0,073	2,700 ± 0,282	2,272 ± 0,098
10	2,416 ± 0,392	1,972 ± 0,113	3,451 ± 0,146	2,674 ± 0,252
20	2,600 ± 0,126	2,174 ± 0,126	4,504 ± 0,257	2,872 ± 0,143

Table 4: Prediction performances for the Delve dataset with a RBFN model.

the number of original features with a maximum of 20. This allows us to see which methods can quickly select the most relevant features.

3.2.1 Results

The results are presented in Tables 1 to 4. They correspond, for each percentage (%) of missing values, to the average best RMSE observed on each of the 10 datasets together with the standard deviation over these 10 results. The imputation strategy is referred to by 'EM' or 'KNN' and 'before' or 'after', depending on when imputation has been done.

Imputing after feature selection leads on average to better performances than the traditionnal method in all but 2 of the 32 cases. Moreover, it is in most of the cases less sensitive to the increase of the proportion missing values. As an example one can observe a reduction of the RMSE by more than 25% for KNN imputation and by more than 50% for EM imputation with both models when 20% of the values are missing on the Mortgage dataset. More generally, the RBFN model always performs better with the proposed approach as soon as 5% of values at least are missing.

The same experiments were carried out with a k-nearest neighbors predictor; their results confirm those obtained with the RBFN model. Indeed, selecting

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

features before imputing missing values led to better performances in 29 out of the 32 cases. Due to space limitations, the full results are not presented here.

4 CONCLUSIONS AND FUTURE WORK

In this paper, a new methodology to achieve feature selection for data with missing values is introduced. The approach consists in a simple forward search procedure, combined with the MI criteria. MI is estimated directly from the uncomplete data with a nearest neighbors estimator using truncated distances. The feature selection is thus independent of any imputation procedure.

Results on three artificial and four real-world databases show the efficiency of this new procedure to select relevant variables. Moreover, results also indicate that better performances can be achieved with regression models by conducting feature selection prior to a potential imputation.

As the similarity measure employed in this paper is no longer a metric, future work should focus on the effects of such a measure on the estimators of entropy and mutual information. Comparisons with inputation techniques based on self-organizing maps [12] could also be performed, since they use a similar proximity measure as the one considered here.

References

[1] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[2] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[3] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, 2004.

[4] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948.

[5] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5:537–550, 1994.

[6] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[7] F. Rossi, A. Lendasse, D. François, V. Wertz, and M. Verleysen. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems*, 80(2):215–226, 2006.

[8] L. F. Kozachenko and N. Leonenko. Sample estimate of the entropy of a random vector. *Problems Inform. Transmission*, 23:95–101, 1987.

[9] J. Friedman. Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19(1), 1991.

[10] T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, 14:2001, 2001.

[11] F. Rossi, N. Delannay, B. Conan-Guez, and M. Verleysen. Representation of functional data in neural networks. *Neurocomputing*, 64:183 – 210, 2005.

[12] A. Sorjamaa, A. Lendasse, and E. Séverin. Combination of SOMs for fast missing value imputation. *Proceedings of MASHS 2010, Modèles et Apprentissage en Sciences Humaines et Sociale*, 2010.