ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Synthetic over-sampling in the empirical feature space

M. Pérez-Ortiz, P.A. Gutiérrez and C. Hervás-Martínez *

University of Córdoba, Dept. of Computer Science and Numerical Analysis
Rabanales Campus, Albert Einstein building, 14071 - Córdoba, Spain

Abstract. The imbalanced nature of some real-world data is one of the current challenges for machine learning, giving rise to different approaches to handling it. However, preprocessing methods operate in the original input space, presenting distortions when combined with the kernel classifiers, which make use of the feature space. This paper explores the notion of empirical feature space (a Euclidean space which is isomorphic to the feature space) to develop a kernel-based synthetic over-sampling technique, which maintains the main properties of the kernel mapping. The proposal achieves better results than the same oversampling method applied to the original input space.

1 Introduction

Imbalanced classification is currently receiving a lot of attention from the pattern recognition and machine learning communities [1, 2]. Often, the minority class happens to be more important than the majority one, but it may also be much more difficult to model and identify complex underlying behaviour patterns due to the low number of available samples. Since most traditional learning systems have been designed to work on balanced data, they will usually be focused on improving overall performance and be biased towards the majority class, consequently harming the minority one [3]. To cope with this issue, several algorithms have been designed over the years to over-sample minority samples and to under-sample the majority ones, the Synthethic Minority Over-sampling Technique [1] (SMOTE) being one of the most representative for the first group, among others.

At the same time, kernel methods [4] have been spreading rapidly and gaining more acceptance from machine learning researchers due to their good generalization ability and their determinism. These methods make use of the so-called kernel trick, which implicitly maps their inputs into a high-dimensional feature space via a function $\Phi(\cdot)$, in order to compute non-linear decision regions. When these methodologies are combined with other preprocessing techniques which operate in the input space (such as over-sampling techniques), some obvious distortions are found, given that they operate in different spaces. The ideal approach would be to preprocess the training patterns in the feature space, although this is not possible since the only information available is the dot products of their images. To deal with this issue, this paper makes use of the notion

*This work has been partially subsidized by the TIN2011-22794 project of the Spanish Ministerial Commission of Science and Technology (MICYT), FEDER funds and the P08-TIC-3745 project of the "Junta de Andalucía" (Spain).

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

of empirical feature space [5, 6], which has demonstrated to preserve the geometrical structure of the original feature space, given that distances and angles in the feature space are uniquely determined by dot products and that the dot products of the corresponding images are the original kernel values. This empirical feature space is Euclidean, so it provides a tractable framework to study the spatial distribution of $\Phi(\cdot)$ [7, 8], to measure class separability [6] and to optimize the kernel [6, 9]. Besides, the notion of empirical kernel feature space has been used for the kernelization of all kinds of linear machines [10, 11], with the advantage that the algorithm does not need to be formulated to deal with dot products between data points. This paper focuses on the idea of performing over-sampling in the empirical feature space, instead of in the input space. This Euclidean space is isomorphic to the feature space, hence we hypothesize that the synthetic patterns generated would be better adapted to the kernel machine classifier.

The idea of performing over-sampling in the feature space was researched in [12] (note that in our case, it is performed in the empirical feature space). In this previous work, the synthetic instances were generated by using the geometric interpretation of the dot products in the kernel matrix, and the pre-images of the synthetic instances were approximated based on a distance relation between the feature space and the input one, since inverse mapping $\Phi(\cdot)^{-1}$ from the feature space to input space is not available. Our proposal is free of the computational cost and assumptions of this inverse mapping approximation.

The paper is organized as follows: Section II shows a description of the methodology used; Section III describes the experimental study and analyses the results obtained; and finally, Section IV outlines some conclusions.

2 Methodology

The methodology proposed is based on applying the SMOTE over-sampling technique to the empirical feature space. Consequently, the notion of empirical feature space is described, along with a presentation of how to extend SMOTE to better handle imbalaced datasets when applied to kernel classifiers.

2.1 Empirical feature space

Assume that $\mathcal{X} \subseteq \mathbb{R}^d$ is a nonempty collection of objects, and $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is the set of their vector representations in some input space. Thus, $\mathbf{x} \in \mathcal{X}$, where this d-dimensional $\mathcal{X}$ set represents the training set. Assume that k is a symmetric real-valued kernel function, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\Phi : \mathcal{X} \to \mathcal{H}$ be a mapping of patterns from $\mathcal{X}$ to a high-dimensional or infinite-dimensional Hilbert Space $\mathcal{H}$. When applying the kernel trick, the only information available about the images of the input patters in $\mathcal{H}$ is their dot or inner product, which is represented by the kernel function computed using the original input patterns, $k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}$. The kernel trick turns a linear decision region in $\mathcal{H}$ into a nonlinear decision in $\mathcal{X}$. Therefore, instead of working directly with $\mathbf{x}$, a pattern is now represented by its similarity to all other points in the input

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

domain, organized in a Gram matrix containing the kernel function values for all the training data, $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. The requirement of positive definite or reproducing kernels is important since the use of these matrices is a key assumption in convex programming, ensuring in practice that kernel algorithms converge to a relevant solution. Hence, since any given Gram matrix $\mathbf{K}$ of rank r will be a symmetrical positive-semidefinite matrix, it can be diagonalised as:

$$\mathbf{K}_{(m\times m)} = \mathbf{P}_{(m\times r)} \cdot \mathbf{M}_{(r\times r)} \cdot \mathbf{P}^{\mathrm{T}}_{(r\times m)}, \tag{1}$$

where $(\cdot)^{\mathrm{T}}$ is the transpose operation, $\mathbf{M}$ is a diagonal matrix containing the r positive eigenvalues of $\mathbf{K}$ in decreasing order, and $\mathbf{P}$ consists of the eigenvectors associated to those r eigenvalues. The empirical feature space is a Euclidean space preserving the dot product information about $\mathcal{H}$ contained in $\mathbf{K}$. The mapping from the input space to a r-dimensional empirical feature space can be defined, $\Phi^e_r : \mathcal{X} \to \mathbb{R}^r$, where r is the rank of $\mathbf{K}$. This space is isomorphic to the embedded feature space $\mathcal{H}$, but presents all the advantages of being Euclidean:

$$\Phi^e_r : \mathbf{x}_i \to \mathbf{M}^{-1/2} \cdot \mathbf{P}^{\mathrm{T}} \cdot (k(\mathbf{x}_i, \mathbf{x}_1), \ldots, k(\mathbf{x}_i, \mathbf{x}_m))^{\mathrm{T}}. \tag{2}$$

It is easy to check that the kernel matrix of the training images obtained by this transformation is $\mathbf{K}$, when considering the standard dot product [5, 6]. Note that this transformation corresponds to the principal component analysis *whitening* step [13], although applied to the kernel matrix, instead of the covariance matrix. Although the whole set of all r positive eigenvalues has been considered in this paper, a smaller set (for example, a p-dimensional set) could also be considered by choosing the p dominant eigenvalues and their associated eigenvectors. This would limit the dimensionality of the empirical feature space.

2.2 Synthetic minority over-sampling in the empirical feature space

The proposal consists of using the empirical feature space to apply preprocessing algorithms, whose results would better suit the kernel machine classifier later considered. In this paper, the SMOTE algorithm was selected to decrease the problems caused by imbalanced datasets when applying a kernel classifier.

First of all, we compute the empirical feature space of the training set. $\mathbf{T}^e_{(m\times r)}$ is the matrix generated by applying the Φ^e_r transformation (equation (2)) to the training patterns. After this, the standard SMOTE algorithm [1] is run over the minority class images of this $\mathbf{T}^e$ matrix, resulting in the generation of n new synthetic images, arranged in the matrix $\mathbf{S}^e_{(n\times r)}$.

Synthetic samples will be used to complete the kernel matrix, by obtaining their dot product with respect to the rest of the training patterns, i.e. $\mathbf{KS}^e_{i,j} = \mathbf{T}^e_i \cdot \mathbf{S}^e_j, 1 \leq i \leq m, 1 \leq j \leq n$, and with respect to themselves $\mathbf{SS}^e_{i,j} = \mathbf{S}^e_i \cdot \mathbf{S}^e_j, 1 \leq i, j \leq n$, where $\mathbf{T}^e_i$ is the empirical space representation of the i-th training pattern, and $\mathbf{S}^e_i$ is the i-th synthetic sample previously generated. Using these matrices, the over-sampled training Gram matrix $\mathbf{K}^*$ will be composed as follows:

$$\mathbf{K}^*_{(m+n)\times(m+n)} = \begin{pmatrix} \mathbf{K}_{(m\times m)} & \mathbf{KS}^e_{(m\times n)} \\ \left(\mathbf{KS}^e_{(m\times n)}\right)^{\mathrm{T}} & \mathbf{SS}_{(n\times n)} \end{pmatrix}, \tag{3}$$

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

where $\mathbf{K}$ is the original kernel matrix. For the generalization phase, the same steps are considered to complete the test kernel matrix, taking into account that the empirical feature space images of the test patterns are derived using the same Φ_r^e transformation (considering only the training data).

3 Experimental results

The proposal has been evaluated considering the Support Vector Classifier (SVC) [14] and the SMOTE technique [1]. This proposal (Empirical feature space SMOTE, E-SMOTE) is compared to the original SMOTE applied in the input space, and to the results without over-sampling. 8 binary benchmark datasets from the UCI repository with different imbalance ratios (proportion of majority patterns with respect to minority ones) have been tested. Some multiclass datasets have also been considered by grouping some classes, e.g. ecoli1 represents the ecoli dataset when considering class 1 versus the rest, and glass0146vs2 is the glass dataset when grouping classes 0, 1, 4 and 5 versus class 2.

A stratified 10-fold technique was performed to divide the data, and the results are taken as mean and standard deviation of the selected measures over the 10 test sets. The Gaussian kernel was used. The kernel width and the cost parameter of SVC was selected within the values $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$, by means of a nested 5-fold method applied to the training set. The number of synthetic patterns generated was that needed to balance the distributions, i.e. after applying SMOTE, the number of majority and minority patterns were the same. $k = 5$ nearest neighbours were evaluted to generate synthetic samples.

The results have been reported in terms of three metrics: 1) the well-known Accuracy metric (Acc); 2) the Geometric Mean of the sensitivities ($GM = \sqrt{S_{\mathrm{p}} \cdot S_{\mathrm{n}}}$), where S_{p} is the sensitivity for the positive class (ratio of correctly classified patterns considering only the positive class) and S_{n} is the sensitivity for the negative one; and 3) the Minimum Sensitivity [15] ($MS = \min\{S_p, S_n\}$). GM and MS are specially aimed at measuring the performance of a classifier when handling imbalanced data. The measure considered during the hyperparameter selection was GM, given its robustness when considering imbalanced datasets. All the test results of these experiments can be seen in Table 1.

From the results obtained, several conclusions can be drawn. Firstly, the good performance of the proposal can be appreciated analysing GM and MS measures, where it can be seen that the application of the over-sampling technique in the empirical feature space outperforms the results achieved when applying it in the original input space. Standard deviations of GM and MS measures are high considering that the number of patterns of the minority class in the test set can be very low and that misclassifying a single positive pattern can result in drastic variations of S_p (and consequently in GM and MS). However, these standard deviations tend to be lower for the proposed method, which could indicate that it is more stable. Concerning Acc, the proposal achieves comparable results to those obtained by the other methods (especially for low IR values).

Finally, the non-parametric Friedman's test [16] (with $\alpha = 0.1$) has been

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Table 1: Results achieved by the three methodologies considered, where IR refers to the imbalance ratio of a dataset.

Dataset	IR	Method	*Acc*	*GM*	*MS*
colic	1.72	E-SMOTE+SVC	**82.60 ± 3.93**	**81.03 ± 5.19**	**74.36 ± 8.82**
		SMOTE+SVC	*81.52 ± 5.51*	*79.29 ± 6.04*	71.85 ± 9.71
		SVC	*81.52 ± 5.07*	79.15 ± 6.45	*72.19 ± 11.38*
breast	2.36	E-SMOTE+SVC	**69.61 ± 8.73**	**64.72 ± 9.48**	**55.17 ± 11.43**
		SMOTE+SVC	*67.19 ± 11.62*	*62.76 ± 9.36*	53.06 ± 12.41
		SVC	65.79 ± 8.96	44.29 ± 19.43	*28.89 ± 17.07*
haberman	3.15	E-SMOTE+SVC	*68.94 ± 9.39*	*60.59 ± 13.21*	*49.77 ± 18.31*
		SMOTE+SVC	**70.27 ± 8.24**	**61.68 ± 11.00**	**50.25 ± 16.94**
		SVC	68.30 ± 8.97	45.10 ± 9.66	26.81 ± 13.67
ecoli1	3.36	E-SMOTE+SVC	*87.19 ± 4.94*	**86.43 ± 6.90**	**80.09 ± 9.64**
		SMOTE+SVC	86.60 ± 6.05	*86.03 ± 7.17*	*79.32 ± 9.55*
		SVC	**90.13 ± 5.19**	81.51 ± 18.54	72.77 ± 26.58
spectfheart	3.84	E-SMOTE+SVC	*76.11 ± 7.45*	**77.63 ± 6.48**	**69.45 ± 8.14**
		SMOTE+SVC	75.01 ± 8.36	*75.16 ± 8.96*	*67.81 ± 9.13*
		SVC	**76.82 ± 8.59**	58.26 ± 24.33	45.67 ± 24.14
glass0146vs2	11.05	E-SMOTE+SVC	82.62 ± 10.70	**64.47 ± 36.61**	**56.52 ± 34.05**
		SMOTE+SVC	*86.45 ± 10.72*	*49.99 ± 44.59*	*44.21 ± 41.73*
		SVC	**88.38 ± 4.99**	16.05 ± 34.27	13.42 ± 29.42
cleveland0vs451	12.31	E-SMOTE+SVC	*93.56 ± 6.48*	**96.44 ± 3.59**	**93.13 ± 6.88**
		SMOTE+SVC	92.42 ± 6.25	*93.18 ± 8.61*	*87.50 ± 14.73*
		SVC	**94.87 ± 4.12**	72.96 ± 40.24	68.13 ± 40.77
yeast2vs8	23.10	E-SMOTE+SVC	90.65 ± 7.43	**67.12 ± 37.20**	**58.26 ± 35.53**
		SMOTE+SVC	*96.69 ± 2.60*	55.05 ± 39.82	45.00 ± 36.89
		SVC	**97.93 ± 1.37**	*65.25 ± 36.84*	*54.78 ± 36.60*

The best method is in **bold** face and the second one in *italics*

applied to the mean rankings for the three measures considered, rejecting the null-hypothesis that all algorithms perform similarly for GM and MS, and accepting it for Acc. The confidence interval was $C_0 = (0, F_{(\alpha=0.1)} = 2.73)$, and the corresponding F-value was $0.76 \in C_0$, $22.87 \notin C_0$ and $14.33 \notin C_0$ for Acc, GM and MS respectively. Furthermore, the Holm test (using the E-SMOTE+SVC as control method) has also been applied concluding that there are statistically significant differences for $\alpha = 0.1$ for GM and MS, when comparing the control method to all the others.

4 Conclusions

This paper proposes the idea of performing preprocessing techniques in the empirical feature space when applying kernel classifiers. We focus on the imbalanced binary classification context, and the proposal has been tested with the standard SVC and the SMOTE over-sampling method, achieving better Geometric Mean and Minimum Sensitivity results than when applying the same preprocessing in the original input space. Given that the over-sampling technique operates in r dimensions (kernel matrix rank), instead of d (dimensionality of the input space), what is noteworthy is its applicability to bioinformatics datasets where the number of features tend to be much higher than the number of samples ($r << d$), and where imbalanced datasets are commonly found. Additionally,

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

as an advantage of the method, there is no need to treat the data attributes differently (taking into account their nature) since all of them are real, unlike in the original SMOTE. This proposal can be extended by considering other different over-sampling methods or under-sampling ones and by extending the experiments with more datasets and methods.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[2] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 39, pp. 281–288, Feb. 2009.

[3] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[4] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[5] B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, pp. 1000–1017, 1999.

[6] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Transactions on Neural Networks*, vol. 16, no. 2, pp. 460–474, 2005.

[7] F. Yan, K. Mikolajczyk, J. Kittler, and M. A. Tahir, "Combining multiple kernels by augmenting the kernel matrix," in *Proc. of the 9th International Workshop on Multiple Classifier Systems (MCS)*, vol. 5997, pp. 175–184, Springer, 2010.

[8] X. Liang, "Feature space versus empirical kernel map and row kernel space in SVMs," *Neural Computing and Applications*, vol. 19, pp. 487–498, Apr. 2010.

[9] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Learning with the optimized data-dependent kernel," in *Proc. of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, vol. 6, pp. 95–, IEEE Computer Society, 2004.

[10] S. Abe and K. Onishi, "Sparse least squares support vector regressors trained in the reduced empirical feature space," in *Proc. of the 17th international conference on Artificial neural networks*, ICANN, pp. 527–536, Springer-Verlag, 2007.

[11] H. Xiong, "A unified framework for kernelization: The empirical kernel feature space," in *Chinese Conference on Pattern Recognition (CCPR)*, pp. 1 –5, nov. 2009.

[12] Z.-Q. Zeng and J. Gao, "Improving svm classification with imbalance data set," in *Proc. of the 16th International Conference on Neural Information Processing: Part I*, ICONIP '09, (Berlin, Heidelberg), pp. 389–398, Springer-Verlag, 2009.

[13] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 460–474, 1998.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] J. C. Fernández-Caballero, F. J. Martínez-Estudillo, C. Hervás-Martínez, and P. A. Gutiérrez, "Sensitivity versus accuracy in multiclass problems using memetic pareto evolutionary neural networks," *IEEE Transactions on Neural Networks*, vol. 21, no. 5, pp. 750–770, 2010.

[16] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.