The Classification of Imbalanced Spatial Data

Alina Lazar
Department of Computer Science and Information Systems
Youngstown State University
Youngstown, OH 44555
alazar@ysu.edu

Bradley A. Shellito
Department of Geography
Youngstown State University
Youngstown, OH 44555
bashellito@ysu.edu

Abstract

This paper describes a method of improving the prediction of urbanization. The four datasets used in this study were extracted using Geographical Information Systems (GIS). Each dataset contains seven independent variables related to urban development and a class label which denotes the urban areas versus the rural areas. Two classification methods Support Vector Machines (SVM) and Neural Networks (NN) were used in previous studies to perform the two-class classification task. Previous results achieved high accuracies but low sensitivity, because of the imbalanced feature of the datasets. There are several ways to deal with imbalanced data, but two sampling methods are compared in this study.

Introduction

The aim of this paper is to show that class imbalance has a powerful impact on the performance of binary classification algorithms. Most machine learning algorithms provide models with better performances when trained using balanced training datasets. However, most of the real-world datasets from various domains like medical diagnosis, document classification, fraud and intrusion detection are highly imbalanced towards the positive or the minority class.

In general, classification algorithms are designed to optimize the overall accuracy performance. However, for imbalanced data, good accuracy does not mean that most examples from the minority class were correctly classified. Therefore, additional performance measures like recall, f-measure, g-means, AUC should be included when we study imbalanced problems.

One common approach to solve the imbalance problem is to sample the data to build an equally distributed training dataset. Several sampling techniques were proposed and analyzed in the literature (Van Hulse, Khoshgoftaar, and Napolitano 2007) including random under-sampling, random over-sampling and more intelligent sampling techniques. A second class of methods uses meta-costs and assigns different penalties for the misclassified instances, depending on their true class. The problem with this type of methods is that it is hard to come up with a good penalty cost. The last type of methods is the algorithmic-based approach. They tweak the classifier to accommodate imbalanced datasets. The algorithm-based methods use meta-learning (Liu, An, and Huang 2006, Zhu 2007) or on-line active learning (Ertekin et al. 2007) to build better classifiers. Different combinations of these methods were also reported.

Real-world imbalanced datasets come from diverse application areas like medical diagnosis, fraud detection, intrusion detection, gene profiling, and object detection from satellite images (Kubat, Holte, and Matwin 1998). Our study investigates the effect of two sampling techniques when applied on four large GIS datasets with an imbalance ratio between 2.4 and 12.5. The four datasets contain over a million instances each, therefore there is no need to use over-sampling. Besides that, over-sampling is known to introduce excessive noise and ambiguity. Instead, the sampling methods considered were random sampling, under-sampling and the Wilson's editing algorithm in combination.

SVM and NN were used before in various studies to predict urbanization and land cover with almost similar results, but different prediction patterns (Lazar and Shellito 2005, Shellito and Lazar 2005). Even if SVM itself does not provide a mechanism to deal with imbalanced data, it can be easily modified. SVM builds the decision boundary on a limited number of instances that are close to the boundary, being unaffected by instances far away from the boundary. This observation can be used as an active learning selection strategy that provides a balanced training set for the early training stages of the SVM algorithm (Ertekin et al. 2007).

In the Background section we summarize related studies that deal with the problem of imbalanced datasets. The section Support Vector Machines and Multi-Layer Perceptrons presents the methods used, while the section describing our experiments presets a comparison between

random sampling, under-sampling and Wilson's editing. The last section presents the conclusions.

Background

Previous research (Lazar and Shellito 2005, Pijanowski et al. 2005, Pijanowski et al. 2002, Pijanowski et al. 2001, Shellito and Lazar 2005, Shellito and Pijanowski 2003) has shown that classification methods such as Support Vector Machines (SVM) and Neural Networks (NN) can be successfully used to predict patterns of urbanization in large datasets. SVM and NN can then be used as predictive tools to determine if grid cells can be accurately predicted as urban or non-urban cells. The effectiveness of the predictive capability of the SVM and NN can be measured through standard accuracy and other measures.

The dataset generated for Mahoning County had over 1,000,000 instances and the imbalanced ratio was approximately 5:1. Even if the accuracy for both SVM and NN were over 90%, the recall was quite low 55%.

Lately, several studies dealt with imbalanced datasets and their effect on classification performance; however none of the studies included datasets with over a million instances.

Extensive experimental results using several sampling techniques combined with several classification methods applied on several datasets were reported by (Van Hulse, Khoshgoftaar, and Napolitano 2007). The sampling techniques considered were: random minority oversampling, random majority oversampling, one-side selection, Wilson's editing, SMOTE (Akbani, Kwek, and Japkowicz 2004), borderline SMOTE and cluster-based oversampling. They concluded that some of the more complicated sampling techniques especially one-side selection and cluster-based oversampling exhibit inferior performance in comparison with some of the simple sampling techniques.

Support Vector Machines

The machine learning algorithms named support vector machines proposed by (Vapnik 1999) consist of two important steps. Firstly, the dot product of the data points in the feature space, called the kernel, is computed. Secondly, a hyperplane learning algorithm is applied to the kernel.

Let (x_i, y_i), $i = 1,\ldots,l$, be the training set of examples. The decision $y_i \in \{-1, 1\}$ is associated with each input instance $x_i \in R^N$ for a binary classification task. In order to find a linear separating hyperplane with good generalization abilities, for the input data points, the set of hyperplanes $\langle w,\mathbf{x}\rangle+b=0$ is considered. The optimal hyperplane can be determined by maximizing the distance between the hyperplane and the closest input data points. The hyperplane is the solution of the following problem:

$$\min_{w\in R^l\times R^l, b\in R} \tau(w) = \frac{1}{2}\|w\|^2 \quad (1)$$

where $y_i(\langle w, x_i\rangle + b) \geq 1$ for all $i = 1,\ldots, l$.

One challenge is that in practice an ideal separating hyperplane may not exist due to a large overlap between input data points from the two classes. In order to make the algorithm flexible a noise variable $\varepsilon_i \geq 0$ for all $i = 1,\ldots,l$, is introduced in the objective function as follows:

$$\min_{w\in R^l\times R^l, b\in R} \tau(w, \varepsilon_i) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\varepsilon_i \quad (2)$$

when $y_i(\langle w, x_i\rangle + b) \geq 1-\varepsilon_i$ for all $i = 1,\ldots, l$.

By using Lagrange multipliers the previous problem can be formulated as the following convex maximization problem (Liu, An, and Huang 2006):

$$W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \quad (3)$$

when the following conditions are true, $0 \leq \alpha_i \leq C$ for all $i = 1,\ldots, l$, and $\sum_{i=1}^{l}\alpha_i y_i = 0$. Here the positive constant C controls the trade-off between the maximization of (3) and the training error minimization, $\sum\varepsilon_i$.

From the optimal hyperplane equation the decision function for classification can be generated. For any unknown instance x the decision will be made based on:

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b) \quad (4)$$

which geometrically corresponds to the distance of the unknown instance to the hyperplane.

The method described until now works well on linear problems. Function K, the kernel from (4) enables good results for nonlinear decision problems. The dot product of the initial input space is called the new higher-dimensional feature space.

$$K : R^l \times R^l \rightarrow R, K(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle \quad (5)$$

A polynomial kernel, the radial basis and the sigmoid function are suitable kernels with similar behavior in terms of the resulting accuracy and they can be tuned by changing the values of the parameters. There is no good method to choose the best kernel function. The results reported in this paper were obtained by using the following radial basis function (Schölkopf and Smola 2002) as kernel.

$$K(x_i, x_j) = \exp(-\frac{\left\| x_i - x_j \right\|^2}{2\gamma^2}) \quad (6)$$

Multi-layer Perceptron Neural Networks

The multi-layer perceptron (MLP) (Witten and Frank 2000) is a popular technique because of its well-known ability to perform arbitrary mappings, not only classifications. Usually built out of three or four layers of neurons, the input layer, the hidden layers and the output layer, this network of neurons can be trained to identify almost any input-output function. The back-propagation algorithm used for the training process adjusts the synaptic weighs of the neurons according with the error at the output. During the first step of the algorithm the predicted outputs are calculated using the input values and the network weights. Afterwards, in the backward pass the partial derivatives of the cost function are propagated back through the network and the weights are adjusted accordingly.

The problem with the MLP methods is that they are susceptible to converge towards local minimums. MLP methods are considered as "black box", because it is impossible to obtain snap-shots of the process.

Sampling Methods

Since the datasets considered have over a million instances we decided to investigate under-sampling (US). This sampling technique discards random instances from the majority class until the two classes are equally represented. The other sampling method used in this study is called Wilson's editing (Barandela el al. 2004) (WE). A k-means nearest neighbor classification procedure is used with k=3 to classify each instance in the training set using all the remaining instances. Afterwards, all the instances from the majority class that were misclassified are removed.

Performance Metrics

Especially in the case of imbalanced datasets, classification accuracy alone is not the best metric to evaluate a classifier. Several other performance metrics can be used in order to get a more comprehensive picture of the classifier's capabilities.

Recall or sensitivity is the metric that measures the accuracy on the positive instances, It can be defined as *TruePositive* / (*TruePositive* + *FalseNegative*). Specificity measures the accuracy on the negative instances and can be defined as *TrueNegative* / (*TrueNegative* + *FalsePositive*). Both sensitivity and specificity are incorporated in the g-means measure (Ertekin et al. 2007), which is defined as square root from sensitivity * specificity.

Datasets

Seven broad predictor variables, which aid in describing the distribution of urbanization within the counties, were constructed using ESRI's ArcGIS 9.2 software package. ArcGIS allows for modeling of a vast array of geospatial techniques, including the cell-by-cell raster models. These variables were chosen as they reflect large-scale factors that influence the patterns of urbanization and general urban trends for the region, as well as being similar to GIS variables for urban modeling within the Midwest (Pijanowski et al. 2005, Pijanowski et al. 2002, Pijanowski 2001, Shellito and Pijanowski 2003). The variables constructed were:

a. Distance to City Centers
b. Distance to Highways
c. Distance to Interstates
d. Distance to Railroads
e. Distance to Lakes
f. Distance to Rivers
g. Density of Agriculture

For the county, a series of base layers was compiled to build the variables. The NLCD (National Land Cover Database) 2001 data was used for location of urban areas and as a source of agricultural data. Base layers for highways, interstates, and railways were drawn from US Census 2000 TIGER files. Lakes and rivers data was derived from Ohio Department of Transportation (ODOT) data. All base layers were projected into the UTM (Universal Transverse Mercator) projection and used to develop the predictor variables in raster format at 30m resolution. Distance variables were created by calculating the Euclidian distance of each cell from the closest feature in the base layers. The density variable was constructed by using a 3x3 moving window neighborhood operation and summing up the number of base layer grid cells in the neighborhood. Urban land was identified by selecting all grid cells with the "developed" classification in the NLCD dataset.

Predictor variables for each county were constructed by incorporating data from their bordering Ohio counties, to simulate the influence of nearby spatial factors outside the county borders (for instance, the proximity of a nearby city center in a bordering county could potentially effect the urban development within the target county). The resultant predictor variables created at this multi-county level were then clipped down to the boundaries of the chosen county and used in the analysis.

This type of data was extracted for four counties from the state of Ohio: Delaware, Holmes, Mahoning and Medina. All four resulting datasets contain more than a million instances each. Table 1 shows for each county dataset how many instances belong to the positive class, how many instance belong to the negative class and the ratio between the positive and the negative instances. All datasets are

mildly imbalanced from a 2.4:1 ratio for Mahoning County to a 12.5:1 ratio for Holmes County.

Table 1. Number of Training Instances and Ratios

	# Positive	*# Negative*	*Ratio*
Delaware	209,765	1,106,749	5.2761:1
Holmes	90,164	1,129,403	12.5260:1
Mahoning	353,411	868,423	2.4576:1
Medina	228,819	987,405	4.3152:1

For the first set of experiments we used two classifiers, the SVM and the Multi-Layer Perceptron (MLP). We used the libSVM (Chang and Lin 2001) software to run the parameter search, the training and the testing for SVM and Weka for the MLP. The experiments were similar with the experiments reported in (Lazar and Shellito 2005, Shellito and Lazar 2005) for Mahoning County.

Random stratified sampling, which maintains the ratio of positive versus negative instances in the datasets, was used to generate datasets of 10,000 instances for the parameter search and datasets of 50,000 for training sets.

A grid parameter search was performed for the SVM classifier and the values for the two parameters C and gamma are listed below in table 2.

Table 2. Parameters C and gamma for the LibSVM

	C	gamma
Delaware	8192	0.125
Holmes	2048	0.5
Mahoning	2	32
Medina	128	0.125

Next, both classifiers SVM and MLP were trained on the 50,000 instances datasets and the models obtained were tested using the entire datasets. The results obtained are reported in Table 3. For each dataset and for each classifier (SVM and MLP) three performance metrics are listed: accuracy, recall and g-means.

Table 3. Classification Performances for NN and SVM

		Del	Hol	Mah	Med
Accuracy	SVM	91.11	92.86	87.87	87.67
	MLP	80.36	93.8	85.72	87.53
Recall	SVM	57.35	4.19	70.44	47.13
	MLP	18.86	21.68	70.85	49.36
G-means	SVM	74.78	20.47	81.79	67.64
	MLP	40.02	46.46	80.63	68.97

The results show that even if SVM has higher accuracy for three of the datasets MLP has higher recall, so a better classification of the positive instances for three of the datasets. Recall has the largest values for the Mahoning County dataset, which also has the lowest imbalanced ratio. Looking at the low recall values for the other three datasets, we need to investigate ways to better classify the instances from the positive class. Experiments using different sampling techniques are reported in the next section.

Experiments

We run experiments using RapidMiner (Mierswa et al. 2006) on the four datasets Delaware, Holmes, Mahoning and Medina as follows. For each dataset we performed 5 runs of a five-fold cross validation with the libSVM software. The rbf kernel was used. The two parameters C and gamma where changed to values previously found by running a grid parameter search.

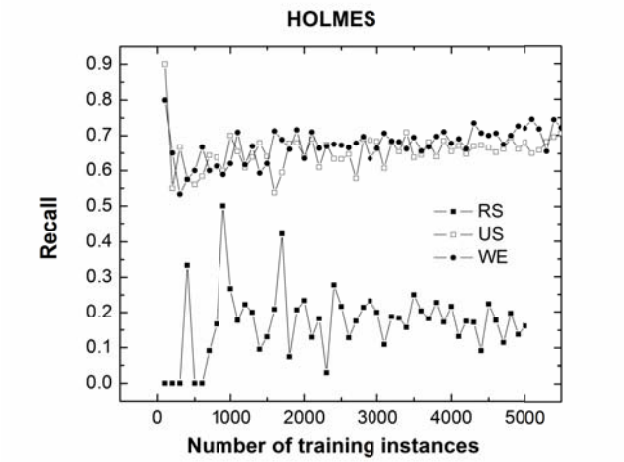

Figure 1. Recall for Holmes County Dataset

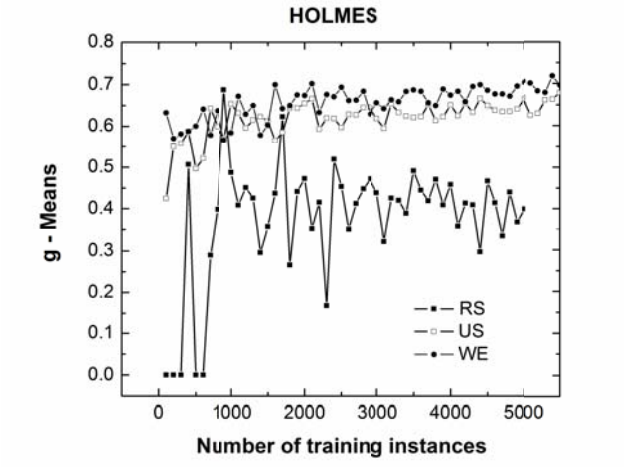

Figure 2. G-means for Holmes County Dataset

Three sampling techniques were used: random stratified sampling (RS), equal under-sampling (US) and Wilson's

editing sampling (WE). Each experiment was iterated through subsample datasets with sizes between 100 and 5000, with a step of 100.

The results are shown on two counties Holmes and Medina, due to space limitation. The Holmes County has the highest imbalanced ratio of approximately 12.5 and Medina has a 4.3 imbalanced ratio.

All four figures show that both under-sampling and Wilson's editing sampling have a great influence on the classification performance of the SVM learner. As accuracy is not relevant in the case of imbalanced datasets we looked at recall and g-means. The Wilson'e editing worked only slightly better than the equal under-sampling, but required extensive preprocessing. The biggest difference in performance can be seen in Figure 1 with the recall for the Holmes County.

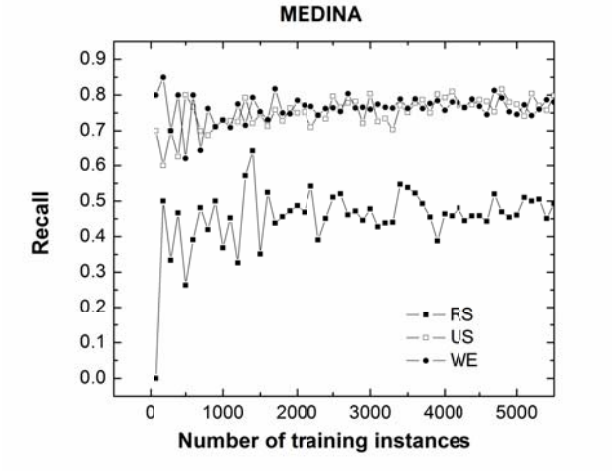

Figure 3. Recall for Medina County Dataset

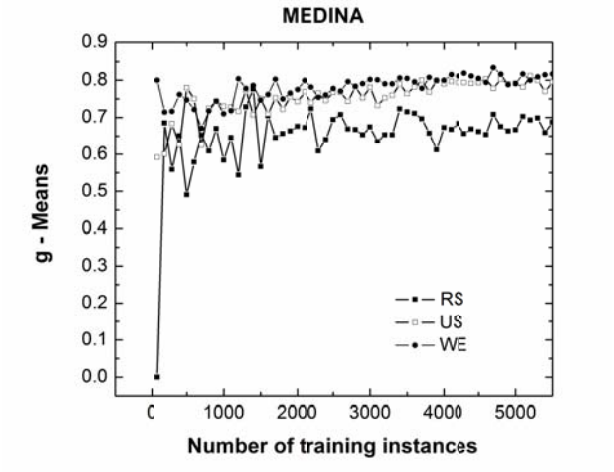

Figure 4. G-means for Medina County Dataset

Conclusions

We have presented an experimental analysis performed on large imbalanced GIS extracted datasets. The goal was to find what sampling techniques improve the classification performance especially for the minority class. It is important in the case of imbalanced datasets that additional performance measures like recall and g-means are compared in addition to the usual accuracy. We concluded that both equal under-sampling and Wilson's editing work better that just simple random stratified sampling, but there is no significant difference between the two.

Further research may investigate how other learners like Neural Networks or Decision Trees perform with under-sampling and Wilson's editing sampling. Over-sampling, cost-sensitive learning, and meta-learners are other alternatives that can be used to improve the performance for our datasets.

References

Akbani, R.; Kwek, S.; and Japkowicz N. 2004. Applying support vector machines to imbalanced datasets. Proceedings of European Conference on Machine Learning. 39-50. Pisa, Italy, Springer-Verlag, Germany.

Barandela, R.; Valdovinos, R. M.; Sanchez J. S.; and Ferri, F. J. 2004. The Imbalanced Training Sample Problem: Under or Over Sampling? In Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition (SSPR/SPR'04), *Lecture Notes in Computer Science* 3138: 806-814.

Chang, C.; and Lin, C-J. 2001
LIBSVM : a library for support vector machines, 2001. Software at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Last accessed 01/15/2011.

Cristianini, N; and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge, England.

Ertekin, S.; Huang, J.; Bottou, L.; and Lee Giles, C. 2007. Learning on the border: active learning in imbalanced data classification. In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (Lisbon, Portugal, November 06 - 10, 2007). CIKM '07. 127-136. ACM, New York, NY.

Koggalage, R.; and Halgamuge, S. 2004. "Reducing the Number of Training Samples for Fast Support Vector Machine Classification." *Neural Information Processing – Letters and Reviews* 2 (3): 57-65.

Kubat, M.; Holte, R. C.; and Matwin. S. 1998. Machine Learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3): 195-215.

Lazar, A.; and Shellito, B. A. 2005. Comparing Machine Learning Classification Schemes – a GIS Approach. In Proceedings of

ICMLA'2005: The 2005 International Conference on Machine Learning and Applications, IEEE.

Liu, Y.; An A.; and Huang, X. 2006. Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. *Lecture Notes in Artificial Intelligence*, vol. 3918: 107-118.

Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; and Euler, T. 2006. YALE: Rapid Prototyping for Complex Data Mining Task. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06).

Pijanowski, B.; Pithadia, S.; Shellito, B. A.; and Alexandridis, K. 2005. Calibrating a neural network based urban change model for two metropolitan areas of the upper Midwest of the United States. *International Journal of Geographical Information Science* 19 (2): 197-216.

Pijanowski, B.; Brown, D.; Shellito, B. A.; and Manik, G. 2002. Use of Neural Networks and GIS to Predict Land Use Change. *Computers, Environment, and Urban Systems* 26(6): 553-575.

Pijanowski, B.; Shellito, B. A.; Bauer, M. and Sawaya, K. 2001. "Using GIS, Artificial Neural Networks and Remote Sensing to Model Urban Change in the Minneapolis-St. Paul and Detroit Metropolitan Areas." In Proceedings of the ASPRS Annual Conference, St. Louis, MO.

Schölkopf, B.; and Smola, A. 2002. *Learning with Kernels*. MIT Press, Cambridge Massachusetts.

Shellito, B. A.; and Lazar, A. 2005. Applying Support Vector Machines and GIS to Urban Pattern Recognition. In Papers of the Applied Geography Conferences, volume 28.

Shellito, B. A.; and Pijanowski, B. 2003. "Using Neural Nets to Model the Spatial Distribution of Seasonal Homes." *Cartography and Geographic Information Science* 30 (3): 281-290.

Van Hulse, J.; Khoshgoftaar, T. M.; and Napolitano. A. 2007. Experimental perspectives on learning from imbalanced data. In Proceedings of the 24th International Conference on Machine Learning (Corvalis, Oregon, June 20 - 24, 2007). Z. Ghahramani, Ed. ICML '07, vol. 227. ACM, New York, NY, 935-942.

Vapnik, V. N. 1999. *The Nature of Statistical Learning Theory*, 2nd edition, Springer-Verlag, New York, NY.

Witten, I. H.; and Frank, E. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*, Morgan Kaufmann Publishers, Burlington, MA.

Zhu, X. 2007. Lazy Bagging for Classifying Imbalanced Data. In Seventh IEEE International Conference on Data Mining. 763-768. Omaha, NE.