The Imbalanced Problem in Morphological Galaxy Classification

Jorge de la Calleja[1], Gladis Huerta[1], Olac Fuentes[2], Antonio Benitez[1], Eduardo López Domínguez[1], and Ma. Auxilio Medina[1]

[1] Ingeniería en Informática, Universidad Politécnica de Puebla, Puebla, 72640, México
{jdelacalleja,ghuerta,abenitez,elopez,mmedina}@uppuebla.edu.mx
[2] Computer Science Department, University of Texas at El Paso, Texas, 79968, U.S.A.
ofuentes@utep.edu

Abstract. In this paper we present an experimental study of the performance of six machine learning algorithms applied to morphological galaxy classification. We also address the learning approach from imbalanced data sets, inherent to many real-world applications, such as astronomical data analysis problems. We used two over-sampling techniques: SMOTE and Resampling, and we vary the amount of generated instances for classification. Our experimental results show that the learning method Random Forest with Resampling obtain the best results for three, five and seven galaxy types, with a F-measure about .99 for all cases.

Keywords: machine learning, imbalanced data sets, galaxies.

1 Introduction

Imbalanced class problems are often encountered in many real world applications. The problem occurs when the number of instances in one class heavily outnumbers the instances in the other class. With imbalanced data sets we will have biased classifiers that obtain high predictive accuracy over the majority class, but poor predictive accuracy over the minority class which is generally the class of interest. Some examples of domains that present an imbalanced class are: text classification, detection of fraudulent telephone calls, disease detection, astronomical object classification, and many others.

A short time ago, there has been a great deal of interest from astronomers in applying machine learning techniques in order to solve astronomical problems such as classification of galaxies, classification of stars, classification of binary stars, galaxy/star discrimination, astronomical object classification in spectral images, among others. Although they have used a wide variety of learning algorithms, these approaches have not addressed the class imbalance inherent to this kind of problems.

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 533–540, 2010.

We present an experimental study of the performance of six machine learning algorithms applied to morphological galaxy classification considering the imbalanced data set problem. Classification of galaxy images is one of the most important challenges for astronomers because it provides significant clues about the origin and evolution of the Universe. The paper is organized as follows: Section 2 describes related work to deal with imbalanced data sets and a brief introduction of galaxy classification. In Section 3 we describe the methods used for doing the experiments. In Section 4 we show experimental results and finally in Section 5 conclusions and future work are presented.

2 Related Work

2.1 Imbalanced Data Sets

The class imbalance problem has received much attention from the machine learning community. This problem has been addressed in two main approaches: internal and external approaches. The first approach consists of modifying or creating new learning methods, while in the second approach, sampling techniques are used in order to build a more balanced data set. We now present some works proposed to deal with imbalanced data sets.

Kubat and Matwin [8] presented a heuristic under-sampling method to balance the data set in order to eliminate noisy, borderline, and redundant training examples of the majority class, keeping the original population of the minority class. Chawla et al. [3], devised a method called Synthetic Minority Oversampling Technique (SMOTE). This technique creates new synthetic examples from the minority class. SMOTEBoost is an approach introduced by Chawla et al. [4] that combines SMOTE with the boosting ensemble. Han et al. presented two new minority over-sampling methods: borderline-SMOTE1 and borderline-SMOTE2, in which only the minority examples near the borderline are oversampled. Hongyu and Herna [7] introduced a method that combines boosting and data generation (DataBoost-IM), that achieved comparable and slightly better predictions, when using G-mean and F-measures metrics. Liu et al. [9] proposed an ensemble of SVMs with an integrated sampling technique, which combines both over-sampling and under-sampling. Wang and Japkowicz [15] proposed the boosting-SVMs with Asymmetric Cost algorithm, and they obtained very good results for the majority class as well as the minority class.

2.2 Automated Galaxy Classification

Increasing astronomical data is becoming available in quantities vastly too large to analyze by traditional methods. Therefore, automated and robust tools are required for any kind of analysis, such as morphological classification of galaxies.

Recently, there has been a great deal of interest from astronomers in applying machine learning techniques to solve astronomical problems, such as morphological galaxy classification. The morphology of galaxies is generally an important issue in the large scale study of the Universe. Galaxy classification is the first step

towards a greater understanding of the origin an formation process of galaxies, and the evolution process of the Universe. The easiest way to classify galaxies is by their shape, and Edwin Hubble devised a basic scheme for classify them into three main types: Spirals, Ellipticals and Irregulars.

Automated classification of galaxies has been tackled using several machine learning techniques [1,5,11,12,13,16,17] such as neural networks, decision trees, ensembles of classifiers, instance-based methods, self organized maps, random forest, just to name a few. Nevertheless, they have not have not addressed the class imbalance inherent to this problem.

3 The Methods

In this section we briefly describe the learning methods we used for the experiments. For a deeper introduction of the algorithms we recommend the reader review the references.

3.1 Naive Bayes Classifier

The Naive Bayes classifier [10] is a probabilistic algorithm based on the assumption that the attribute values are conditionally independent given the target values. The Naive Bayes classifier applies to learning tasks where each instance x can be described as a tuple of attribute values $a_1, a_2, \ldots a_n$ and the target function $f(x)$ can take on any value from a finite set V. When a new instance x is presented, the Naive Bayes classifier assigns to it the most probable target value by applying the rule:

$$f(x) = argmax_{v_j \epsilon V} P(v_j) \Pi_i P(a_i \mid v_j) \tag{1}$$

The learning task of the Naive Bayes is to build a hypothesis by estimating the different $P(v_i)$ and $P(a_i \mid v_j)$ terms based on their frequencies over the training data.

3.2 C4.5

C4.5 [10] operates by recursively splitting a training set based on feature values to produce a tree such that each example can end up in only one leaf. An initial feature is chosen as the root of the tree, and the examples are split among branches based on the feature value for each example. If the values are continuous, then each branch takes a certain range of values. Then a new feature is chosen, and the process is repeated for the remaining examples. Then the tree is converted to an equivalent rule set, which is pruned.

3.3 Radial Basis Function Networks

A radial basis function network (RBFNet) [10] is a variant of artificial neural networks. RBFNets are embedded in a two layer neural network, where each

hidden unit implements a radial activated function. The output units implement a weighted sum of hidden unit outputs. Due to their nonlinear approximation properties, RBFNets are able to model complex tasks.

3.4 Random Forest

Random forest [2] is an ensemble of unpruned classification trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Prediction is made by aggregating the predictions of the ensemble. Random forest generally yields better performance than single tree classifiers such as C4.5.

3.5 Support Vector Machines

Support Vector Machines (SVMs) [14] are based on the Structural Risk Minimization principle from computational learning theory. This principle provides a formal mechanism to select a hypothesis from a hypothesis space for learning from finite training data sets. The aim of SVMs is to compute the hyperplane that best separates a set of training examples. Two cases are analyzed: the linear separable case and the non-linear separable case. In the first case we are looking for the optimal hyperplane in the set of hyper-planes separating the given training examples. The optimal hyperplane maximizes the sum of the distances to the closest positive and negative training examples (considering only two classes). The second case is solved by mapping training examples to a high-dimensional feature space using kernel functions. In this space the decision boundary is linear and we can apply the first case. There are several kernels such as polynomial, radial basis functions, neural networks, Fourier series, and splines, among others; that are chosen depending on the application.

3.6 SMOTE

The Synthetic Minority Over-sampling TEchnique [3] is an over-sampling method to deal with imbalanced data sets. This technique operates in the feature space rather than the data space. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

4 Experimental Results

The experiments were done using a data set of 310 galaxy images to classify three types of galaxies (E, S, Irr), and 293 for five (E, S0, Sa+Sb, Sc+Sd, Irr) and seven classes (E, S0, Sa, Sb, Sc, Sd, Irr). The minority class was represented by the Irregular galaxies, about 3.5% for three types and 3.7% for five and seven types.

We used 13 features to perform the classification task, which were obtained in an automated manner using principal component analysis (details can be found in [5]).

We used the Naive Bayes classifier, a RBF Network (a normalized Gaussian radial basis function network), SMO (Support Vector Machines), J48 (a particular C4.5 implementation), J48graft (J48 with pruning) and Random Forest that are implemented in Weka[1]. For SVMs we use a linear kernel, with 1.0 for the complexity constant and 0.0010 for rescale kernel. In the case of RBF we use logistic regression applied to k-means clusters as basis functions, with 2 clusters, a minimum standard deviation of 0.1, until convergence is reached. We also used the SMOTE and Resampling methods for over-sampling, testing different amounts for generating new instances: 100%, 200% and 500%.

In learning imbalanced data sets, accuracy is often not a good measure of performance, because a classifier that labels everything with the majority class can still achieve very high accuracy. We use metrics such as precision, recall, and F-measure to evaluate the performance of the learning algorithms. These metrics have been widely used for comparison and can be defined as:

$$Recall = TP/(TP + FN) \tag{2}$$

$$Precision = TP/(TP + FP) \tag{3}$$

$$F - measure = 2 \times Recall \times Precision/(Recall + Precision) \tag{4}$$

where TP and TN denote the number of positive and negative examples that are classified correctly, while FN and FP denote the number of misclassified positive and negative examples, respectively.

Tables 1, 2 and 3 show the results for three, five and seven classes, respectively. This results were obtained by averaging ten runs of 10-fold cross-validation for each learning method. Analyzing the tree-class case, we can observe that Random Forest obtained the best results for SMOTE-100%, SMOTE-500%, Resampling-100%, Resampling-200% and Resampling-500%. We can also note that J48, J48graft and RBFNet obtained very good results for SMOTE as well as for Resampling.

For the five-class case, when we use SMOTE, RBFNet obtained the best results for 100% and 200%, while Random Forest obtained the best precision, recall and F-measure using 500%. For the case of Resampling, Random Forest again obtained the best results with over .77 for 100%, over .91 for 200% and over .99 for 500%.

Finally, for the seven-class case, RBFNet obtained the best results in eight of the nine results for SMOTE. However, Random Forest again obtained the best results for Resampling, with about .75, .91 and .99, for 100%, 200% and 500%, respectively.

[1] Weka is a collection of machine learning algorithms for data mining tasks. `http://www.cs.waikato.ac.nz/ml/weka/`

Table 1. Results for 3 types of galaxies

SMOTE									
	100%			200%			500%		
	Prec	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m
Naive Bayes	0.8678	0.8189	0.8333	0.8493	0.7995	0.8154	0.8001	0.7616	0.7719
J48	0.8310	0.8358	0.8323	0.8137	0.8166	0.8145	0.7890	0.7937	0.7908
J48graft	0.8384	0.8519	0.8421	0.8218	0.8364	0.8272	0.7890	0.7937	0.7908
RBFNet	0.8632	0.8834	0.8632	0.8464	0.8645	0.8485	0.8076	0.8191	0.8090
Random Forest	0.8825	0.8919	0.8615	0.8462	0.8622	0.8303	0.8684	0.8706	0.8559
SMO	0.7660	0.8750	0.8170	0.7160	0.8460	0.7760	0.5930	0.7700	0.6700
Resampling									
	100%			200%			500%		
	Prec	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m
Naive Bayes	0.8879	0.8777	0.8790	0.8927	0.8589	0.8691	0.9885	0.9882	0.9881
J48	0.9336	0.9345	0.9330	0.9724	0.9715	0.9716	0.9945	0.9944	0.9944
J48graft	0.9468	0.9468	0.9440	0.9820	0.9820	0.9818	0.9953	0.9952	0.9952
RBFNet	0.9272	0.9280	0.9257	0.9320	0.9296	0.9275	0.9186	0.9271	0.9189
Random Forest	0.9689	0.9682	0.9659	0.9888	0.9888	0.9885	0.9997	0.9997	0.9997
SMO	0.7980	0.8940	0.8430	0.7840	0.8850	0.8320	0.8180	0.9050	0.8590

Table 2. Results for 5 types of galaxies

SMOTE									
	100%			200%			500%		
	Prec	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m
Naive Bayes	0.4481	0.4588	0.4420	0.4248	0.4334	0.4171	0.4565	0.4570	0.4449
J48	0.4034	0.4207	0.4045	0.3947	0.4144	0.3996	0.4188	0.4354	0.4233
J48graft	0.4080	0.4280	0.4107	0.3954	0.4159	0.4008	0.4197	0.4382	0.4244
RBFNet	0.4667	0.4914	0.4686	0.4586	0.4813	0.4599	0.4808	0.4949	0.4788
Random Forest	0.4719	0.4684	0.4590	0.4537	0.4507	0.4439	0.5236	0.5208	0.5181
SMO	0.2772	0.4574	0.2997	0.2731	0.4414	0.2824	0.3597	0.4178	0.2892
Resampling									
	100%			200%			500%		
	Prec	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m
Naive Bayes	0.4659	0.4453	0.4466	0.5115	0.4875	0.4892	0.4804	0.4875	0.4656
J48	0.7163	0.7149	0.7143	0.8699	0.8692	0.8694	0.9729	0.9726	0.9724
J48graft	0.7146	0.7133	0.7123	0.8666	0.8662	0.8660	0.9729	0.9727	0.9725
RBFNet	0.6005	0.6005	0.5954	0.5991	0.6020	0.5902	0.5828	0.5924	0.5652
Random Forest	0.7786	0.7744	0.7744	0.9132	0.9123	0.9122	0.9941	0.9939	0.9939
SMO	0.3913	0.4990	0.4352	0.4142	0.5196	0.4263	0.3853	0.4705	0.3380

Table 3. Results for 7 types of galaxies

SMOTE									
	100%			200%			500%		
	Prec	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m
Naive Bayes	0.3407	0.3760	0.3446	0.3525	0.3829	0.3544	0.3503	0.3831	0.3550
J48	0.3369	0.3455	0.3393	0.3424	0.3478	0.3429	0.3598	0.3642	0.3611
J48graft	0.3395	0.3505	0.3422	0.3413	0.3508	0.3434	0.3623	0.3693	0.3652
RBFNet	0.3739	0.4300	0.3846	0.3769	0.4254	0.3880	0.4231	0.4632	0.4335
Random Forest	0.3496	0.3943	0.3641	0.3800	0.4093	0.3860	0.4198	0.4548	0.4326
SMO	0.1950	0.4420	0.2710	0.1850	0.4300	0.2590	0.1635	0.3966	0.2287
Resampling									
	100%			200%			500%		
	Prec	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m	*Prec*	*Rec*	*F*-m
Naive Bayes	0.4141	0.3977	0.3875	0.5571	0.5592	0.5386	0.4804	0.4875	0.4656
J48	0.6715	0.6708	0.6686	0.8272	0.8247	0.8248	0.9729	0.9726	0.9724
J48graft	0.6657	0.6634	0.6617	0.8285	0.8263	0.8262	0.9729	0.9727	0.9725
RBFNet	0.5564	0.5569	0.5474	0.5590	0.5700	0.5433	0.5828	0.5924	0.5652
Random Forest	0.7582	0.7540	0.7488	0.9166	0.9148	0.9145	0.9941	0.9939	0.9939
SMO	0.2631	0.3936	0.2522	0.1950	0.4415	0.2707	0.3853	0.4705	0.3380

5 Conclusions

We have presented an experimental study of the performance of six machine learning algorithms applied to galaxy classification, addressing the imbalanced class inherent to this astronomical problem. From the results we can say that Random Forest was the best classifier for most of the cases, nevertheless, RBFNets and J48 obtained very good results. In addition, we can mention that the Resampling technique obtained better results than SMOTE in almost all cases for all the classifiers.

Future work includes addressing the class imbalanced problem in other domains such as text classification, astronomical classification in wide-field images, biological structures, where the imbalanced problem is very common.

Acknowledgements

First author wants to thank PROMEP for supporting this research work under grant UPPUE-PTC-023.

References

1. Bazell, D., Aha, D.: Ensembles of classifiers for morphological galaxy classificacion. The Astrophysical Journal 548, 219–233 (2001)
2. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
3. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.P.: SMOTE: synthetic minority oversampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)

4. Chawla, N., Lazarevic, A., Hall, L., Bowyer, K.: SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
5. De la Calleja, J., Fuentes, O.: Machine learning and image analysis for morphological galaxy classification. Montly Notices of the Royal Astronomical Society 349, 87–93 (2004)
6. Han, H., Wang, W., Mao, B.: Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005)
7. Hongyu, G., Herna, L.V.: Learning from imbalanced data sets with boosting and data generation: The databoost-IM approach. SIGKDD Explor. Newsl. 6(1), 30–39 (2004)
8. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186 (1997)
9. Liu, Y., An, A., Huang, X.: Boosting predicion accuracy on imbalanced datasets with svm ensembles. In: Ng, W.-K., Kitsuregawa, M., Li, J., Chang, K. (eds.) PAKDD 2006. LNCS (LNAI), vol. 3918, pp. 107–118. Springer, Heidelberg (2006)
10. Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
11. Mohamed, M.A., Atta, M.M.: Classification of galaxies using transformed domain features. Internartional Journal of Computer Science and Network Security 10(2), 86–91 (2010)
12. Naim, A., Lahav, O., Sodre Jr., L., Storrie-Lombardi, M.: Automated morphological classification of apm galaxies by supervised artificial neural networks. Monthly Notices of the Royal Astronomical Society 275, 567 (1995)
13. Philip, N., Wadadekar, Y., Kembhavi, A., Joseph, K.: A difference boosting neural network for automated star-galaxy classification. Astronomy and Astrophysics 385, 1119–1126 (2002)
14. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
15. Wang, B., Japkowicz, N.: Boosting support vector machines for imbalanced data sets. In: An, A., Matwin, S., Raś, Z.W., Ślęzak, D. (eds.) Foundations of Intelligent Systems. LNCS (LNAI), vol. 4994, pp. 38–47. Springer, Heidelberg (2008)
16. Yagi, M., Nakamura, Y., Doi, M., Shimasaku, K., Okamura, S.: Morphological classification of nearby galaxies based on asymmetry and luminosity concentration. Monthly Notices of the Royal Astronomical Society 368(1), 211–220 (2006)
17. Zhang, Y., Zhao, Y.: Automated clustering algorithms for classification of astronomical objects. The Astrophysical Journal 422, 1113–1121 (2004)