Bio-Medical Materials and Engineering 26 (2015) S1541–S1547
DOI 10.3233/BME-151453
IOS Press

Restricted Boltzmann machines based oversampling and semi-supervised learning for false positive reduction in breast CAD

Peng Cao[a,b*], Xiaoli Liu[a,b], Hang Bao[a,b], Jinzhu Yang[b] and Dazhe Zhao[a,b]
[a]*Medical Image Computing Laboratory of Ministry of Education, Northeastern University, 110819, Shenyang, China*
[b]*College of Information Science and Engineering, Northeastern University, 110819, Shenyang, China*

Abstract. The false-positive reduction (FPR) is a crucial step in the computer aided detection system for the breast. The issues of imbalanced data distribution and the limitation of labeled samples complicate the classification procedure. To overcome these challenges, we propose oversampling and semi-supervised learning methods based on the restricted Boltzmann machines (RBMs) to solve the classification of imbalanced data with a few labeled samples. To evaluate the proposed method, we conducted a comprehensive performance study and compared its results with the commonly used techniques. Experiments on benchmark dataset of DDSM demonstrate the effectiveness of the RBMs based oversampling and semi-supervised learning method in terms of geometric mean (G-mean) for false positive reduction in Breast CAD.

Keywords: Breast computer aided detection, false-positive reduction, imbalanced data learning, semi-supervised learning, restricted Boltzmann machines

1. Introduction

Breast cancer has become one of the major public health issues of women in developed countries and an early detection of masses is important for early-stage breast cancer diagnosis [1]. Classification of masses from mammograms is a critical step in computer aided detection and diagnosis for breast cancer.

Computer aided detection (CAD) can provide the initial mass detection which may help an expert radiologists in their decision-making. A CAD scheme for mass detection in mammography can be broadly divided into two major stages: (1) suspicious candidate masses are detected and then (2) the false positive masses (FPs) are reduced while retaining the true positives (TPs). Furthermore, there are two key challenges to be solved in the false-positive reduction step.

1) Imbalanced data distribution between TPs and FPs

For finding the suspicious mass, the initial detection of the CAD requires high sensitivity, so it produces a number of false positives. The false-positive reduction (FPR) step is a critical part in the

* Address for Corresponding: Peng Cao, College of Information Science and Engineering, Northeastern University, 110819, Shenyang, China. Tel.: (86 24) 8366 5418; Fax: (86 24) 8366 3446; E-mail: caopeng@ise.neu.edu.cn.

breast mass detection system [2-4]. The major issues in the FPR are that the two classes are skewed and have unequal misclassification costs. Typically, the positive class carries a higher cost of misclassification, making the common classification methods inappropriate. This is a typical "class imbalance problem" [5]. Class imbalanced data has detrimental effects on the performance of conventional classifiers. However, in the potential mass classification, this problem has attracted less attention.

2) Limitation of labeled candidate instances

In order to make the CAD systems perform well, a large number of instances labeled are required for constructing classifier model. However, the size of the labelled instances are usually insufficient since that the task of making label for each instances is time-consuming [6]. An effective strategy is required to develop a model from limited knowledge on a large amount of unlabeled instances to enhance the performance of the learned model.

To address the above-mentioned challenges, we designed two methods based on the RBMs to reduce the number of false positives in Breast CAD, named OSRBM and SSRBM respectively. The OSRBM algorithm models the actual data distributions and generates new instances based on the learned probability for imbalanced data learning. After balancing the distribution, the SSRBM algorithm uses the unlabeled instances to improve the performance of the hypothesis trained from the labeled instances.

2. RBMs based probabilistic over-sampling algorithm, OSRBM

The imbalanced data issue usually occurs in computer-aided detection systems since the healthy class (negative class) is more predominant in comparison to the diseased class (positive class), including other CAD, such as lung [7] and colon [8]. This imbalanced data has detrimental effects on the performance of conventional classifiers. Typically, the classifiers attempt to reduce global error without taking the data distribution into consideration. As a consequence, all instances are misclassified as negative for high classification accuracy.

The over-sampling method is the most straightforward method for solving the issue of imbalanced data learning. The most popular over-sampling method is called SMOTE, which creates synthetic samples between each minority instance and one of its neighbors [9]. SMOTE generates synthetic instances throughout the line segments by joining all of the k minority class nearest neighbors of each minority class instance. SMOTE can also address the between-class imbalance issue. Furthermore, SMOTE can increase the amount and significance of the minority class in the critical region for decision and predication. In order to generate more appropriate instances, several SMOTE based methods are combined with the ensemble framework are proposed, such as SMOTEBoost [10] and RAMOBoost algorithm [11]. However, the SMOTE based algorithm manipulates the instances blindly without considering the data distribution on the whole feature space, resulting in creation of wrong instances under the complex distribution. Moreover, to generate more accurate instances and generalize the decision region for the mass class, probabilistic oversampling can be used by estimated probability distributions that model the actual data distributions from the training data. By this approach, we can avoid the possibility of the synthetically generated training samples actually belonging to any other class in case of class overlapping [12]. However, the prior distribution is unknown for the mass class. Therefore, here we propose a RBMs based generative probabilistic oversampling method to generate samples without requiring the prior probability function.

The Restricted Boltzmann Machines (RBMs) have attracted considerable interest in machine

learning as they have a strong representation ability in learning useful features from input data in recent years [13]. The RBMs are two-layer generative probabilistic models that can atomically learn and model a probability distribution from the training data. RBMs consist of a set of hidden units h in the second layer, a set of observation units' v in the input layer, and symmetric connections (weight matrix W) between the two. The parameterized RBMs can also model the joint probability distributions over visible and hidden vectors through an energy function. Here, we propose a novel oversampling based RBMs model, called OSRBM. The procedure of the OSRBM is performed according to the distribution probability without jeopardizing structure of data. OSRBM works as follows:

Step 1: Learning the probability distribution

The generative RBMs can be constructed on the training samples by adjusting the RBMs parameters. Therefore the model fits a finite number of training data and estimates the distribution of the underlying probability distribution from an unknown target distribution.

Step 2: Generating the initial samples by SMOTE

We use SMOTE to increase the amount of the minority class at first, this enhances the significance of the minority class in the decision region. The new data generated by the oversampling technique is used as the initial sample of Gibbs sampling. This helps in achieving a faster convergence of the generated samples with the minority class distribution.

Step 3: Generating new instances with Gibbs sampling

After successful learning, Gibbs sampling is used for generating data from the joint probability of multiple random variables. During the Gibbs sampling, a Markov chain is constructed by updating each variable based on its conditional probability distribution. This Gibbs sampling produces multiple Markov chains, each starting with a different minority class sample, instead of one very long chain as done in conventional Gibbs sampling. The univariate conditional distribution of each dimension is represented by $P(X_i|x_1^{(t+1)},\ldots, x_{i-1}^{(t+1)},\ldots, x_{i-1}^{(t+1)}, x_n^{(t+1)})$ and is used to choose attribute values that form a synthetically generated sample, and is further repeated for all the elements.

The procedure of Gibbs sampling depends on Burn-in and Lag. Burn-in is the number of sample generation iterations that are required for the samples to reach a stationary distribution. Lag is the number of consecutive samples that are discarded from the Markov chain following each accepted sample to avoid autocorrelation between the consecutive samples. This algorithm is described in Algorithm 1, and the procedure described above can be visualized in Figure 1.

3. RBMs based semi-supervised classification algorithm, SSRBM

Algorithm 1 OSRBM algorithm

Input: S: minority class Training set; T: Number of iterations; N:minority dimensions; bi: burn-in period; D: size of minority ; lag

Output: $new\_samples$

Construct RBMs model learnt from S

for $d=1$ *to* D **do**

  $X^{(0)}=SMOTE(X_d)$

  for $t=1$ *to* T **do**

    for $i=1$ *to* N **do**

      $x_i^{(t+1)}\sim P(X_i|\, x_1^{(t+1)},\ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)},\ldots, x_n^{(t)},)$

  if $t>bi$ *and* t *mod*$(lag)=0$

  $new\_samples= new\_samples+X^{(t)}$

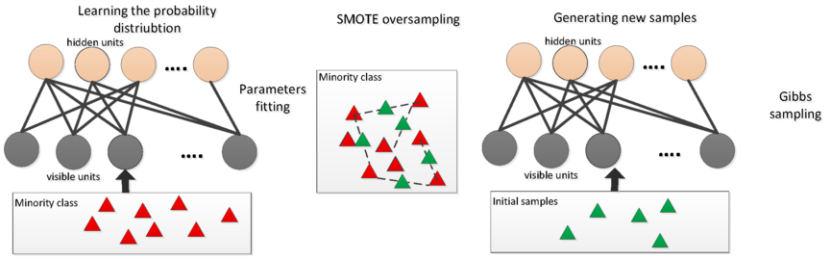

Fig. 1. The procedure of OSRBM algorithm.

In the medical domain, it is difficult to obtain a large amount of instances labeled by radiologist, but there are abundant unlabeled instances. In supervised learning, sophisticated classifiers require sufficient training of labeled data. When the labeled data is not enough, the unlabeled data is exploited and combined with the labeled data by semi-supervised learning methods. This improves the performance of classifier model.

We propose a hybrid of supervised and unsupervised training method based on RBMs. Both the labeled instances and unlabeled instances are used for the training. The RBMs framework is originally a good generative model of the unlabeled data. In order to perform semi-supervised learning, we can combine a generative objective function ($L_{unlabel}$) with a discriminative objective function (L_{label}).

$$L_{overall} = L_{label} + L_{unlabel} = -\sum_{i=1}^{|T|} \log p(y_i \mid x_i) + \sum_{i=1}^{|U|} \log p(x_i) \tag{1}$$

In the objective function, L_{label} is used to optimize the posterior probability $p(y|x)$ directly, and $L_{unlabel}$ is used to optimize $p(x)$ on the unlabeled data. To optimize this overall objective function, a contrastive divergence approximation is used:

$$\frac{\partial \log p(x_i)}{\partial \theta} = -E_{y,h|x_i}\left[\frac{\partial E(y_i, x_i, h)}{\partial \theta}\right] + E_{y,x,h}\left[\frac{\partial E(y, x, h)}{\partial \theta}\right] \tag{2}$$

Where θ are parameters of RBMs. These two terms denote the expectations under the distribution specified by $P(y,h|x)$ and $P(y,x,h)$, respectively.

4. Experiment

4.1. Potential mass candidates detection

The suspicious regions in the mammograms used for evaluating the proposed FPR algorithms are from the Digital Database for Screening Mammography (DDSM). As the masses of cancers grow from the center to surround they have a high concentration value of the gradient vector. Therefore, we use iris filter to detect the potential mass. Iris filter is a filter which can adaptively obtain the concentration of gradient vectors [14]. The threshold value in our experiment was calculated adaptively for each mammogram. This threshold value is estimated as the gray level value at the 95% of the cumulative distributive function. Potential masses can be detected and segmented from the

background using a threshold value in the enhanced image processed by iris filter. Using these parameters, 1792 ROIs (Region of Interest) were recognized and extracted from the DDSM database where each true mass region is indicated and delineated by physicians. In each mammogram image, false positive class besides the true one labels the other ROIs. In total, 158 were depicted as a true mass, while the rest 742 were normal (false positive candidates). The imbalanced data distribution occurs as the iris filter has a limited capacity of differentiating masses from normal breast tissues. Each ROIs were scaled to 80 pixels by 80 pixels size.

4.2. Evaluating the SSRBM for few labeled mass candidates

The experiment was designed to evaluate the effectiveness of SSRBM algorithm. We empirically assessed SSRBM learning against the state-of-the art methods for semi-supervised learning, such as self-training, co-training and co-forest. In order to simulate different amount of unlabeled data, four different unlabeled rates (20%, 40%, 60% and 80%) are studied in our experiments. Moreover, we also employ hidden unit activations in order to minimize the dimensionality of output.

For each data set with each unlabeled rate, the results were averaged by 10-fold cross validation. All the semi-supervised learning methods are combined with the OSRBM algorithm for equal comparison. The over-sampling ratio is set to 200%. For the other three comparative methods, five types of features (a total of 20 features) are extracted for each ROI: iris filter descriptors, gray level descriptors, texture descriptors, contour-related descriptors, and morphological descriptors [15].

The metric for representing the performance of each classifier is chosen by the geometric mean (G-mean). This was done as the overall accuracy is not an appropriate metric for evaluating imbalanced data. Moreover, raising the sensitivity inevitably decreases the specificity; and therefore we need to consider the overall performance. G-mean is the geometric mean of specificity and sensitivity, which is commonly utilized for evaluating or optimizing for imbalanced data [2]. G-mean is defined as $G\text{-}mean=(Sensitivity*Specificity)^{1/2}$.

These experimental results show that each semi-supervised learning method can improve the performance of the hypothesis learned on a small amount of labeled ROI instances by exploiting the large amount of unlabeled instances. Besides the G-mean, the ROC curves of different semi-supervised learning methods are shown in Figure 2. This helps to visualize the performance over all instances with 20% unleaded rate. From the study of Table 1 and Figure 2, it is apparent that SSRBM achieved a higher G-mean and AUC value compared to the other contender methods. As we know, the representation of ROI is critical for discrimination between true masses and false masses. However, it is difficult to obtain good feature representations by human efforts. These results also demonstrate that the feature learned by RBMs is good feature representation of mass ROI.

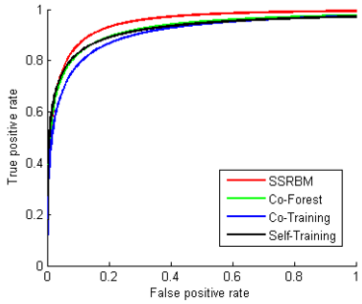

Fig. 2. The true positive rate/false positive rate (ROC) curves on the dataset.

Table 1

Experimental results of semi-supervised learning methods for candidate mass classification under different unlabeled rates

Unlabeled rate	Metric	Supervised Learning	Semi-supervised Learning			
		RBM	**SSRBM**	**Self-Training**	**Co-Training**	**Co-Forest**
80%	AUC	0.773	**0.824**	0.817	0.811	0.817
	G-mean	0.722	**0.781**	0.752	0.738	0.772
	(Sen, Spec)	(0.708, 0.736)	(0.734,0.831)	(0.781,0.724)	(0.763,0.714)	(0.752,0.793)
60%	AUC	0.825	**0.846**	0.835	0.829	0.840
	G-mean	0.739	0.804	0.778	0.756	**0.809**
	(Sen, Spec)	(0.745, 0.733)	(0.767,0.843)	(0.802,0.755)	(0.774,0.738)	(0.762,0.858)
40%	AUC	0.824	**0.877**	0.856	0.844	**0.877**
	G-mean	0.748	**0.823**	0.789	0.762	0.817
	(Sen, Spec)	(0.749, 0.748)	(0.799,0.848)	(0.814,0.764)	(0.807,0.719)	(0.793,0.841)
20%	AUC	0.840	**0.889**	0.872	0.866	0.879
	G-mean	0.762	**0.851**	0.813	0.782	0.821
	(Sen, Spec)	(0.779,0.745)	(0.822,0.881)	(0.845,0.782)	(0.833,0.734)	(0.829,0.813)

4.3. Evaluating the OSRBM for imbalanced mass candidate data

As is previously known, the optimal over-sampling ratio may be unknown but the parameter of over-sampling ratio plays a vital role in the performance of imbalanced data learning. Many over-sampling methods, over-sample the minority class into a completely balanced training set in the literature. However it is not an appropriate method of comparison. It is desirable for an over-sampling method to be robust with respect to the different over-sampling ratio. We empirically evaluate the OSRBM algorithm against three other over-sampling algorithms including SMOTE, SMOTEBoost and RAMOBoost.

The 10-fold cross validation method is employed for comparison of the different sampling methods. In each fold, only the L is used to estimate the distribution and generate new samples. The unlabeled rate is 40%. We show that OSRBM performs empirically better than SMOTE, SMOTEBoost and RAMOBoost sampling with respect to G-mean based on the SSRSM learning. From the Figure 3, we see that OSRBM performs well compared to other common over-sampling methods with different over-sampling ratio. OSRBM can be extended to more potential regions rather than being limited to the line between the positive instance and it is selected nearest neighbors. Moreover, this result demonstrates that OSRBM is more robust to over-sampling ratio compared to the performance of other resampling methods.

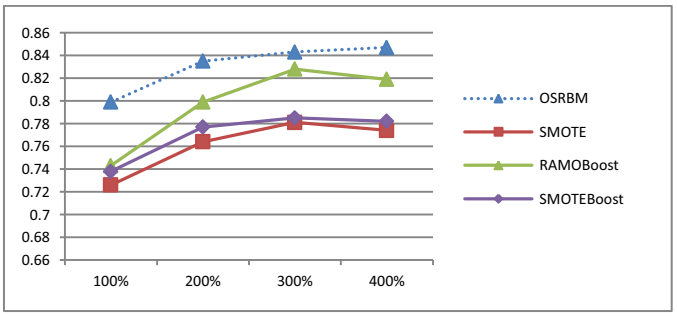

Fig. 3. The performance of four comparative methods while tuning over-sampling ratio in terms of G-mean.

5. Conclusion

This paper proposes two new RBMs based approaches to deal with the issue of constructing a classification model from imbalanced data with only few labeled positive examples. This probabilistic oversampling strategy based on the RBMs can generate more accurate instances than the state-of-the art sampling methods. Moreover, the hybrid of supervised and unsupervised training method based on RBMs outperforms the other reported methods. The methods proposed in this paper can be applied and evaluated on other potential lesion detection or diagnosis, such as nodule and polyp. In the future, we will extend our methods to the classification of the multi-class semi-supervised imbalanced data.

Acknowledgment

This research was supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under grant 2014BAI17B01, the Fundamental Research Funds for the Central Universities under Grant N140403004, N140402003, as well as N140407001, and the Postdoctoral Science Foundation of China 2015M570254.

References

[1] R. Siegel, J. Ma, Z. Zou and A. Jemal, Cancer statistics, CA: A Cancer Journal for Clinicians **64** (2014), 9-29.
[2] P. Cao, J.Z. Yang, W. Li, et al., Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD, Computerized Medical Imaging and Graphics **38** (2014), 137-150.
[3] B.C. Patel and G.R. Sinha, Abnormality detection and classification in computer-aided diagnosis (CAD) of breast cancer images, Journal of Medical Imaging and Health Informatics **4** (2014), 881-885.
[4] F.B. Garma and M.A. Hassan, Classification of breast tissue as normal or abnormal based on texture analysis of digital Mammogram, Journal of Medical Imaging and Health Informatics **4** (2014), 647-653.
[5] H. He and E.A. Garcia, Learning from imbalanced data, IEEE Transaction on Knowledge and Data Engineering **21** (2009), 1263-1284.
[6] M. Li and Z.H. Zhou, Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples, IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans **37** (2007), 1088-1098.
[7] P. Cao, D. Zhao and Z. Osmar, Cost sensitive adaptive random subspace ensemble for computer-aided nodule detection, 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal, 2013, pp. 20-22.
[8] X. Yang, Y. Zheng, M. Siddique, et al., Learning from imbalanced data: A comparative study for colon CAD, Medical Imaging, International Society for Optics and Photonics, 2008, Orlando, Florida, USA, 69150R.
[9] N.V. Chawla, K.W. Bowyer, L.O. Hall, et al., SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research **16** (2002), 321–357.
[10] N.V. Chawla, A. Lazarevic, L.O. Hall, et al., SMOTEBoost: Improving prediction of the minority class in boosting, Proceedings of Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases 2003, Cavtat-Dubrovnik, Croatia, pp. 107-119.
[11] S. Chen, H. He, E.A. Garcia, Ramoboost: Ranked minority oversampling in boosting, IEEE Transactions on Neural Networks **21** (2010), 1624-1642.
[12] A. Liu, J. Ghosh and C.E. Martin, Generative oversampling for mining imbalanced datasets, Proceedings of the 7th IEEE International Conference on Data Mining, 2007, Omaha, Nebraska, USA, 66-72.
[13] H. Larochelle, M. Mandel, R. Pascanu, et al., Learning algorithms for the classification restricted boltzmann machine, The Journal of Machine Learning Research **13** (2012), 643-669.
[14] T. Terada, Y. Fukumizu, H. Yamauchi, et al., Detecting mass and its region in mammograms using mean shift segmentation and Iris Filter, International Symposium on Communications and Information Technologies, Gold Coast, Australia, 2010, pp. 1176-1179.
[15] C. Varela, P.G. Tahoces, A.J. Mndez, M. Souto and J. J.Vidal, Computerized detection of breast masses in digitized mammograms, Computers in Biology and Medicine **37** (2007), 214–26.