On the Classification of Imbalanced Datasets

Arun Kumar M.N
Research Scholar
Dept. of Electronics and Communication Engg.
P.E.S. College of Engineering
Mandya, India

H.S. Sheshadri
Professor
Dept. of Electronics and Communication Engg.
P.E.S. College of Enginering
Mandya, India

ABSTRACT

In recent research the classifications of imbalanced data sets have received considerable attention. It is natural that due to the class imbalance the classifier tends to favour majority class. In this paper we investigate the performance of different methods for handling data imbalance in the microcalcification classification which is a classical example for data imbalance problem. Micro calcifications are very tiny deposits of calcium that appear as small bright spots in the mammogram. Classification of microcalcification clusters from mammograms plays an important role in computer-aided diagnosis for early detection of breast cancer. In this paper, we review in brief the state of the art techniques in the framework of imbalanced data sets and investigate the performance of different methods for microcalcification classification.

Keywords

Classification, Microcalcification, Imbalanced data sets, Mammography.

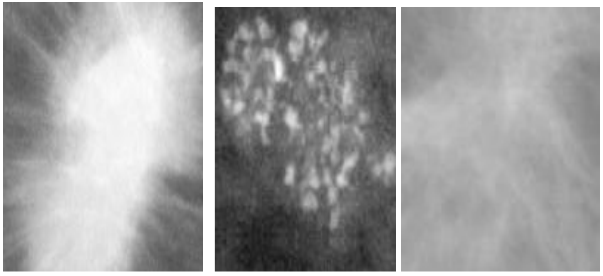
1.INTRODUCTION

Imbalanced data classification often arises in many practical applications. Many classification approaches are developed by assuming the underlying training set is evenly distributed. However, those approaches are faced with a severe bias problem when the training set is a highly imbalanced distribution. There are many real-world problems those are faced with severe problem of learning for imbalanced class.

The imbalanced data [1, 2, 3, 4, 5, 6] cause classifiers to perform poorly on the minority class. When the data are highly imbalanced many existing methods tend to misclassify the minority class. The imbalanced problem has attracted much attention in machine learning community. There are many real-world applications that are faced with class imbalance problem. Some examples for these applications include fault diagnosis, anomaly detection, medical diagnosis, e-mail foldering, face recognition etc.

Breast cancer [54, 55, 56, 57, 58, 59, 60, 61] is the most frequently occurring cancer and one of the leading causes of death among women. Early diagnosis and subsequent treatment can significantly improve the chance of survival for patients with breast cancer. Mammography is the most effective method for the detection of early breast cancer. Mammograms are among the most difficult radiological images to interpret by radiologists. Its images are often very poor in contrast. It can show different features and patterns depending on breast anatomy and tissues density. Radiologists do not detect all breast cancers that are retrospectively detected on the mammograms.

Mammogram facilitates a manner of automated screening which acts as a "second reader" to the physician. There are two major processes involved in the development of techniques that perform automatic analysis and interpretation of breast mammogram: Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx). Detection is the ability to identify potential abnormalities, such as microcalcification, masses, and architectural distortions. Diagnosis is the ability to characterize or classify a detected abnormal entity (architectural distortion, masses, microcalcification clusters) as being either benign or malignant. Figure 1 shows the three important abnormalities which are to be detected by any CAD system on digital mammograms. There are a series of tasks done before CADe algorithms can identify suspicious regions in a mammogram. These include: pre-processing, segmentation, feature extraction, and classification.



a) Spiculations b) Calcifications c) Circumscribed lesions

Fig. 1 Important Abnormalities

Classification of real (potential) microcalcification in a mammogram is a classical example for class imbalance problem. In this paper we review novel classification methods for imbalance data problems and investigate the ability of different methods for handling data imbalance in the microcalcification classification. Paper is organized as follows. Related works that deals with class imbalance problem and microcalcification classification are discussed in section 2. Investigation of the performance of different methods in the microcalcification classification along with pre-processing method is described in section 3. Conclusion is drawn in section 4.

2. LITERATURE SURVEY

Many methods for imbalanced data for binary class and multi class problems are developed. Methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19] are summarized in table 1. Many other methods developed are briefed out in the following paragraphs.

Vladimir Nikulin et.al. proposed [20] an ensemble based method on a large number of balanced random sets. The proposed method is general and may be implemented in conjunction with different base-learners. They have concluded further improvement that can be achieved using more advanced pre-processing technique. Albert Orriols-Puig et.al. [21] investigated the capabilities of evolutionary on-line rule-based systems which extracts knowledge from imbalanced data. Proposed methods are flexible that can be adapted to detect class imbalance cases and find suitable models. A combination of random forest based techniques and sampling methods to identify the potential buyers are proposed [22] by Jie Gu. Proposed method showed its favourable performance improvement in terms of the area under the ROC. Bhavani Raskutti [23] reported methods on balancing training data which has two classes of interest in heavily unbalanced proportions. They focused on the case of supervised learning with support vector machines. Gang Wu et.al. proposed the class-boundary alignment algorithm [24] to augment SVMs to deal with imbalanced training-data problems posed by applications like image retrieval, video surveillance, and gene profiling. Mikel Galar et.al. proposed [25] taxonomy for ensemble-based methods to address the class imbalance where each proposal can be categorized depending on the inner ensemble methodology in which it is based. They developed a thorough empirical comparison by other approaches. In the paper [26] authors reviewed the existing methods for solving the class imbalance problem. They have discussed the various metrics used to evaluate the performance of classifiers. By combining both unsupervised and supervised learning they introduced a new approach to solve the class imbalance problem.

In the paper [27] authors discussed the problem of class or cost imbalance. By outlining some of the trends Nitesh V. Chawla et.al. attempted to chart out the progress in related areas of learning from imbalanced data sets. They discussed many issues that stimulate new directions and solutions. These directions and solutions lead to both theoretical insight and practical applications. Suzan Koknar-Tezel et.al. presented an innovative approach [28] that augments the minority class by adding synthetic points in distance spaces. The proposed approach used Support Vector Machines for classification. Haibo He et.al. discussed [29] new challenges and critical problem in the knowledge discovery and the imbalanced learning problem. Their discussions of the fundamental nature of the imbalanced learning problem, the state-of-the-art solutions used to address critical problem, and the several major assessment techniques used to evaluate the problem served as a comprehensive resource for existing and future knowledge discovery and data engineering researchers and practitioners. Seyda Ertekin et.al. developed an efficient method [30] for active selection of informative instances from a randomly picked small pool of samples. In the proposed method they have maintained the same or achieving even higher g-means values by using less number of training instances in the SVM model. Xing-Ming Zhao in their article [31] presents a new technique for protein classification with imbalanced data. They proposed a new algorithm to overcome the imbalanced problem in protein classification. J. Burez, D et.al. used more appropriate [32] evaluation metrics. These metrics are AUC and lift. They investigated the increase in performance of random and advanced under-sampling and two specific modelling techniques. Baek Hwan Cho et.al. accurately predicted the onset of diabetic nephropathy. They effectively applied various machine learning techniques [33] to irregular and unbalanced diabetes dataset. Those methods include support vector machine (SVM) classification and feature selection methods. For evaluating imbalanced dataset authors introduced a novel measure [34] as a better alternative. They provide a theoretical background for the new evaluation technique that is designed to cope with cost biases. Sireesha Rodda et.al. in their paper [35] presented a method in learning from multi-class imbalanced dataset. The proposed method does not break dataset into a series of binary class datasets. For estimating the quality of classification rule they introduced a new measure called normalized strength score.

Many methods [36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53] are developed for the microcalcification classification. Many methods proposed in the literature results in less classification accuracy in terms of Az (Area under ROC curve). The reason is the difficulty in dealing with imbalanced data set. Many machine learning algorithm fail in dealing with imbalanced data set. We will investigate the performance of different methods for the classification of microcalcification.

3. EXPERIMENTAL ANALYSIS

Data collection of imbalanced datasets related to microcalcification for experimental analysis is extracted from MIAS database and this involves some steps. These steps are discussed in the following 2 subsections. Last subsection summarizes the performance of different methods for microcalcification classification.

3.1 Preprocessing

We have applied some preprocessing steps to remove the artifacts, and noises in the mammograms. Also the pectoral muscle is segmented to reduce the processing area. Median filter of size 3X3 is used to remove the noise. It is the nonlinear filter used to remove the impulse noise from an image. To remove the artifacts the mammogram is opened using a suitable structuring element, and then it is reconstructed. It is followed by thresholding the difference image with value 102. Morphological operators are applied to smooth irregularities and Sobel edge detector is used to detect the edge. Detected breast contour is superimposed on original image. Figure 2 shows the results of these steps for a MIAS image.

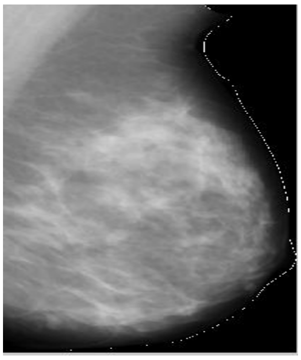


Fig 2. Breast contour superimposed on the original image

Segmentation of pectoral muscle uses wavelet decomposition (Daubechies Wavelet) of fourth level and edge detection using Canny filter. Figure 3 shows the result of these steps in a MIAS image.

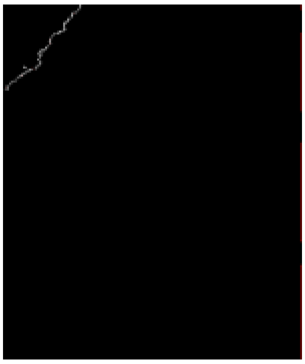


Fig 3. Pectoral muscle identified in the above image

Identified pectoral muscle edge is superimposed on original image. For more details of our work refer [62]. Then binary images are created from each segmented image that is intended for helping automatic cropping procedure so the image will contain only the ROI.

3.2 Detection of Signals (Potential Micro calcification)

This data collection phase uses only 22 MIAS images preprocessed in the previous phase since these images contain the microcalcification inside the boundaries of breast. To detect the signals, which are called as the potential microcalcification, construct a DoG filters and apply sequentially to the images obtained from previous phase by varying the binarization threshold. The points obtained after the application of each filter is added to the points obtained previously. The remaining DoG filter is used in the same way to the points obtained. These points are selected as the signals (potential microcalcification) according to the region area, gray level, and gray gradient. The resultant signals are preclassified into microcalcification and not-microcalcification according to the available truth information. Only 0.4% of signals are microcalcification and 99.6% are not. We employed approach as in [63] for these steps. Because of this imbalanced distribution of classes data sampling is necessary. 35 features are extracted from each potential microcalcification related to their contrast and shape. Features are passed through a feature selection process and 5 relevant features are extracted.

3.3 Investigation of different methods

The main objective of the study is to evaluate the performance of the different classifiers. For data sampling we used SMOTE, Locally linear embedding, and four versions of SMOTE + Complementary Neural Network (CMTNN). In order to demonstrate that the proposed sampling technique can assist classification of imbalanced data, several classification algorithms are used. They are ANN, SVM, Bayesian Networks, and k-NN. The results of classification are evaluated and compared in terms of performance. Table 2 shows the comparison results.

4. CONCLUSION

There are many real-world applications, where uneven distribution of data patterns is very common. In these cases number of training samples of a minority class is much smaller compared to other majority classes. Microcalcification classification is one classical example for imbalanced data problem. In this paper, we review in brief the state of the art techniques in the framework of imbalanced data sets, and investigate the performance of different methods for handling data imbalance in the microcalcification classification. A range of alternative classifiers are selected and datasets are prepared or sampled in order to assess the performance. Future research work can be focused on medical images with intractable geometric complexity in data classification.

Table 1. Methods on class imbalanced data problems

Method used	Dataset used	Other methods taken for analysis	Remarks	Future work
[1] **Bi-class problem** Based on SVM and backward pruning technique	1484 samples collected from SWISS-PROT and Hyperspectral/Polarimetric target detection	SVM-based weighting method	Proposed method outperforms SVM-based weighting method.	Employ other pruning method such as branch and bound.
[2] **Bi-class problem** Linear embedding algorithm incorporated on conventional SMOTE algorithm	Three datasets are collected from several chest x-ray image databases. Sample size of three data sets is 1180, 1530, and 1164 with minority classes of .038%, 0.029%, and 0.042% respectively.	SMOTE algorithm	Proposed method outperforms conventional SMOTE algorithm Naïve Bayesian, K-NN and support vector Machine classifiers are used for classification	Application of proposed method on other medical images with intractable geometric complexity in data classification.
[3] **Bi-class problem** Adopted two methods. 1. [Cost-sensitive] Random Forest 2. Sampling technique + Ensemble idea	Oil, Mammography, Satimage, Hypothyroid, Euthyroid, KDD thrombin of size 937, 11183, 6435, 2520, 2640, and 2543 respectively. Percentage of Minority class is 4.4, 2.3, 9.7, 4.8, 9.1, and 7.6 respectively.	SHRINK, SMOTE, and SMOTEBoost	Proposed two methods outperform SHRINK, SMOTE, AND SMOTEBoost.	Performance of the two proposed methods under label noise.
[4] **Bi-class problem** Kernel-Based Two-Class Classifier using orthogonal forward selection	Synthetic data, Pima Indian diabetes data, Haberman data, ADI data, Satimage data	1-NN, 3-NN, SVM, SUPANOVA, RIPPER, SMOTE, SMOTE-Boost, BRF, WRF	Proposed method outperforms all other method.	Underlying assumption of the procedure may no longer be suitable for imbalanced data sets. So research on the suitability of the procedure over imbalanced data set is necessary
[5] **Bi-class problem** SVM-WEIGHT GSVM-RU SVM-SMOTE SVM-RANDU	Oil, Mammography, Satimage, Abalone (19 vs. other), Abalone (9 vs. 18), Yeast (ME2 vs.other), Yeast (CYT vs. POX)	WRF, AdaCost, KBA, SMOTE-Boost	GSVM-RU outperforms all other classifier	Future research can be focused on comparison study on other benchmarks with highly imbalanced data sets
[6] **Bi-class problem** Two different versions of infinitely imbalanced logistic regression with Gaussian k=1 and k=10.	Landmine, and Watermine dataset	LR	IILR outperforms LR	Make IILR approach suitable for datasets in which the data points of the rare class do not cluster tightly. Other future work can be focused on other applications with severe class imbalance.
[7] **Bi-class problem** Hierarchical fuzzy rule based classification system	22 data sets with low imbalance and 22 data sets with high imbalance. For details refer [7].	Chi-3, Chi-5, Ishibuchi05, E-Algorithm, c4.5	Proposed method outperforms all other methods.	Inclusion of a multi-objective GA for rule selection
[8] Sampling based on clustering (SBC), SBC	Synthetic dataset of 10,000 samples.	AT, RT, Nearmiss-2	For real dataset SBC outperforms all other method	Future research can be focused on comparison

with Nearmiss-1, SBC with Nearmiss-2, SBC with Nearmiss3, SBC with most distance, and SBC with most far	Census-Income datasets (real dataset) of 30,162 samples. 80% used for training and 20% used for testing. Overdue detection database (real dataset) of 62,309 for training and 63,532 for testing.		For synthetic dataset SBC and RT is most suited.	study on other benchmarks with highly imbalanced data sets.
[9] **Bi-class problem** Classifier established by Mahalanobis-Taguchi System (MTS)	Wisconsin, Letter, Heart disease Shuttle, Covtype	SDA, C4.5, BPN, SVMs	Proposed method outperforms all other method	Future research can be focused on the extension of MTS and the research to multiclass tasks
[10] **Bi-class problem** Information granulation based data mining	Two balanced datasets and unbalanced datasets from UCI machine learning repository	NN (BP)	Proposed method outperforms NN.	To reduce the data dimension more sophisticated tool can be used.
[11] **Bi-class problem** SVM based active learning (AL) selection strategy	8 datasets of Reuters, 5 datasets from Citeseer, 3 datasets from UCI, USPS, MNIST-8	Batch, US, SMOTE, DC	AL outperforms in 12 out of 18 cases. No risk of losing information	Future research work can be focused on medical images with intractable geometric complexity in data classification
[12] **Bi-class problem** Combination of supervised learning and unsupervised clustering to handle imbalanced data set.	Liver, Hepatitis, Pima diabetes, and Wisconsin	RPROP	Modified RPROP (Mod-RPROP) has higher values of G-mean and F-measure than RPROP. But similar classification rate for both.	Future research can be focused on highly imbalanced data sets of very less imbalanced ratio < 0.1
[13] **Bi-class problem** Four combination of SMOTE and complementary neural network (CMTNN) to handle problem of classifying imbalanced data. Three classifiers used are ANN, k-NN, and SVM	Pima Indian diabetes data, German credit data, Haberman's survival data, SPECT Heart data.	ENN, Tomek links, and SMOTE to handle problem of classifying imbalanced data.	For ANN classifier, four combination of SMOTE and CMTNN outperforms others. For SVM classifier Tomeklinks and one combination of (SMOTE+CMTNN) outperform other. For k-NN classifier, two combination of SMOTE + CMTNN	The number of instances removed by the proposed method is a major issue in the classification accuracy of the imbalanced data and this issue has a future scope.
[14] **Multi-class problem** Cost sensitive boosting algorithm (AdaC2.M1)	Car data, New thyroid data, and Nursery data from UCI machine learning database	AdaBoost.M1	AdaC2.M1 outperforms Adaboost.M1	Further research can be focused on the modification of AdaC2.M1 where the learning speed is a crucial issue.
[15] **Bi-class problem** Clustering based classification (CLU-SUMO)	Datasets from MediaMill Challenge Problem	SUMO, NB, NN, 3-NN, Ada, DTree, MP	CLU-SUMO outperforms all other methods	To avoid the over fitting problem experiments pertain to the influence of the cluster size K of the proposed framework should be conducted
[16] **Bi-class problem** Active Learning Heuristic (SALH) + Wilcoxon-Mann-Whitney (WMW)	6 unbalanced datasets from UCI repository Abalone, Thyroid, Opt.Digits, Solar Flare, Adult, Liver	C4.5, Naïve Bayes, Canonical GP, GP under SALH	SALH+WWW outperforms all other method	Future work can be focused on the significance of fitness function in GP classifier
[17] **Bi-class problem** k-Nearest Neighbor Approach (1-NN, 3-NN, 5-NN, 7-NN, 9-NN)	Dataset of 300 MEDLINE abstracts	C5.0	5-NN outperforms all other methods	Future research work can be focused on Investigation of the effect of over-sampling on the k-NN approach.
[18**] Bi-Class problem** Ensemble-based learning algorithm (DataBoost.IM)	16 datasets from UCI data repository	C4.5, AdaBoost MI, DataBoost, AdaCost, CSB2, SMOTEBoost	DataBoost.IM outperforms all other methods	Extend DataBoost.IM for noisy data, and multi-class problem.
[19] **Multi-class problem** 3 classifier (min. distance , k-NN, and fuzzy k- nearest) + 22 data preprocessing methods	6 different types of weld flaws	----	1-NN + AHC-KM preprocessing method is best suited among all.	Future work can be focused on development of evaluation criteria for multi-class, application of feature selection together with data preprocessing method, and application of other classifiers on the same data sets

Table 2. Performance of classifiers

Sampling methods	Classifier	Sensitivity	Specificity
SMOTE	Bayesian	71.91	82.
	k-NN	75.06	83.2
	SVM	81.3	80.76
LLE+SMOTE	Bayesian	72.56	83.31
	k-NN	76.45	84.23
	SVM	84.34	83.31
CMTNN (Tech.1)+ SMOTE	k-NN	72.0	84.21
	ANN	77.31	85.45
	SVM	79.0	86.41
CMTNN (Tech.2)+ SMOTE	k-NN	71.4	83.01
	ANN	83.2	79.32
	SVM	84.1	79.1
SMOTE+ CMTNN (Tech.1)	k-NN	72.3	80.9
	ANN	85.3	84.2
	SVM	86.01	**87.5**
SMOTE+ CMTNN (Tech.2)	k-NN	82.3	85.91
	ANN	79.4	86.01
	SVM	**86.4**	81.5

4. REFERENCES

[1] Xue-wen Chen, Byron Gerlach, and David Casasent, “ Pruning support vectors for imbalanced data classification”, Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005.

[2] Juanjuan Wang, Mantao Xu, Hui Wang, Jiwu Zhang, “Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding”, ICSP2006 Proceedings.

[3] Chao Chen, Andy Liaw , and Leo Breiman, “Using random forest to learn imbalanced data”, www.stat.berkeley.edu/tech-reports/666.pdf

[4] Xia Hong, “A kernel-based two-class classifier for imbalanced data sets”, IEEE Transactions on Neural Networks, Vol. 18, NO. 1, January2007

[5] Yuchun Tang, Nitesh V. Chawla,, and Sven Krasser, “SVMs modeling for highly Imbalanced classification “, IEEE Transactions On Systems, MAN, and Cybernetics, Vol. 39, No. 1, FEBRUARY 2009

[6] David P. Williams, Vincent Myers, and Miranda Schatten Silvious, “Mine classification with imbalanced data “, IEEE Geoscience and Remote Sensing Letters, Vol. 6, No. 3, July 2009.

[7] Alberto Fernández, María José del Jesus, Francisco Herrera, “Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets “, International Journal of Approximate Reasoning 50 (2009) 561–577.

[8] Show-Jane Yen, Yue-Shi Lee, “Cluster-based under-sampling approaches for imbalanced data distributions “, Elsevier, Expert Systems with Applications 36 (2009) 5718–5727.

[9] Chao-Ton Su and Yu-Hsiang Hsiao, “An evaluation of the robustness of MTS for imbalanced data “, IEEE Transactions on knowledge and data engineering, Vol. 19, No. 10, October 2007.

[10] Mu-Chen Chen a, Long-Sheng Chen, Chun-Chin Hsu, Wei-Rong Zeng , “An information granulation based data mining approach for classifying imbalanced data “, Elsevier, Information Sciences 178 (2008) 3214–3227.

[11] Jian Huang, L´eon Bottou, C. Lee Giles, “Learning on the border: Active learning in imbalanced data classification “, CIKM’07, November 6–8, 2007, Lisboa, Portugal. ACM 978-1-59593-803-9/07/0011

[12] Son Lam Phung, Abdesselam Bouzerdoum, Giang Hoang Nguyen, “Learning pattern classification tasks with imbalanced data sets “, http://ro.uow.edu.au

[13] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung, “Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm”, http://researchrepository.murdoch.edu.au

[14] Yanmin Sun, Mohamed S Kamel, and Yang Wang, “Boosting for learning multiple classes with imbalanced class distribution “, Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 0-7695-2701-9/06 © 2006 IEEE.

[15] Chao Chen and Mei-Ling Shyu, “Clustering-based binary-class classification for imbalanced data sets“, rvc.eng.miami.edu/paper/2011/IRI11_clustering.pdf

[16] John Doucette and Malcolm I. Heywood, “GP Classification under imbalanced data sets: Active sub-sampling and AUC approximation“,www.cs.dal.ca/~mheywood/Xfiles/publications/JohnEuroGP08.pdf

[17] Zhang and Indrajeet Mani, “kNN approach to unbalanced data distribution: A case study involving information Extraction”, Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003

[18] Hongyu Guo, Herna L Viktor, “Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach”, Sigkdd Explorations. Volume 6, Issue 1 - Page 30-39

[19] T. Warren Liao, "Classification of weld flaws with imbalanced class data ", Elsevier, Expert Systems with Applications 35 (2008) 1041–1052

[20] Vladimir Nikulin, Geoffrey J. McLachlan, and Shu Kay Ng, "Ensemble approach for the classification of imbalanced data ", www.maths.uq.edu.au/~gjm/nmn_lncs09.pdf

[21] Albert Orriols-Puig · Ester Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets ", Soft Computing (2009) 13:213–225, Springer-Verlag 2008

[22] Jie Gu, "Random Forest Based Imbalanced Data Cleaning and Classification", http://lamda.nju.edu.cn/conf/pakdd07/dmc07/reports/P251.pdf

[23] Bhavani Raskutti, Kovalczyk, "Extreme re-balancing for SVMs: A case Study", Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003.

[24] Gang Wu, Edward Y. Chang, "Class-boundary alignment for imbalanced dataset learning", Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003

[25] Mikel Galar, Alberto Fern´andez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera, "A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and Hybrid-Based Approaches ", IEEE Transactions on systems, MAN, and Cybernetics— 2011.

[26] Giang Hoang NguyenAbdesselam, "Learning pattern classification tasks with imbalanced data sets",

[27] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz, "Editorial: Special issue on learning from imbalanced data sets ", Sigkdd Explorations. Volume 6, Issue 1 - Page 1-6, 2004

[28] Suzan K¨oknar-Tezel, Longin Jan Latecki, "Improving SVM classification on imbalanced data sets in distance spaces ", ICDM 2009.

[29] Haibo He, and Edwardo A. Garcia, "Learning from Imbalanced Data ", IEEE transactions on knowledge and data engineering, Vol. 21, No. 9, September 2009

[30] Seyda Ertekin, Jian Huang, C. Lee Giles, "Active learning for class imbalance problem ", www.personal.psu.edu/juh177/pubs/SIGIR2007_short.pdf

[31] Xing-Ming Zhao, Xin Li, Luonan Chen, and Kazuyuki Aihara1," Protein classification with imbalanced data", 2008; 70:1125–1132. VVC 2007

[32] J. Burez, D. Van den Poel, "Handling class imbalance in customer churn prediction ", Expert Systems with Applications 36 (2009) 4626–4636

[33] Baek Hwan Cho, Hwanjo Yu , Kwang-Won Kim , Tae Hyun Kim, In Young Kim, Sun I. Kim, "Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods predict diabetic nephropathy using visualization and feature selection methods ", Elsevier, Artificial Intelligence in Medicine (2008) 42, 37—53

[34] Cheng G. Weng Josiah Poon, "A new evaluation measure for imbalanced datasets ", crpit.com/confpapers/CRPITV87Weng.pdf

[35] Sireesha Rodda, Prof. Shashi Mogalla, "A normalized measure for estimating classification rules for multi-class imbalanced data sets ", International Journal of Engineering Science and Technology (IJEST). 2011

[36] L. Bocchi, G. Coppini, J. Nori, G. Valli, "Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks", Med. Eng. Phys. 26 (4) (2004) 303–312.

[37] Cancer Research UK: Key Facts on Breast Cancer, 2009. http://info.cancerresearchuk.org/cancerstats/types/breast/.

[38] M. De Santo, M. Molinara, F. Tortorella, M. Vento, "Automatic classification of clustered microcalcifications by a multiple expert system", Pattern Recogn. 36(7) (2003) 1467–1477.

[39] I. El-Naqa, Y. Yang, N.P. Galatsanos, et al., "A similarity learning approach to content-based image retrieval: application to digital mammography", IEEE Trans. Med. Imaging 23 (10) (2004) 1233–1244.

[40] I. El-Naqa, Y. Yang, M.N. Wernick, et al., "A support vector machine approach for detection of microcalcifications", IEEE Trans. Med. Imaging 21 (12) (2002) 1552–1563.

[41] J. Ge, B. Sahiner, L.M. Hadjiiski, H.-P. Chan, J. Wei, M.A. Helvie, C. Zhou, "Computer aided detection of clusters of microcalcifications on full field digital mammograms", Med. Phys. 33 (8) (2006) 2975–2988.

[42] L. Hadjiiski, B. Sahiner, H.-P. Chan, et al., "Classification of malignant and benign masses based on hybrid ART2LDA approach", IEEE Trans. Med. Imaging 18 (12) (1999) 1178–1187.

[43] M. Kallergi, "Computer-aided diagnosis of mammographic microcalcification clusters", Med. Phys. 31 (2) (2004) 314–326.

[44] R. Nakayama, Y. Uchiyama, K. Yamamoto, et al., "Computer- aided diagnosis scheme using a filter bank for detection of microcalcification clusters in mammograms", IEEE Trans. Biomed. Eng. 53 (2) (2006) 273–283.

[45] A. Papadopoulosab, D.I. Fotiadisb, A. Likasb, "Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines", Artif. Intell. Med. 34 (2) (2005) 141–150.

[46] P. Sajda, C. Spence, J. Pearson, "Learning contextual relationships in mammograms using a hierarchical pyramid neural network", IEEE Trans. Med. Imaging 21 (3) (2002) 239–250.

[47] H. Soltanian-Zadeha, F. Rafiee-Radc, S. Pourabdollah-Nejad, "Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms", Pattern Recognition. 37 (10) (2004) 1973–1986.

[48] J. Ren, D. Wang, J. Jiang, "Effective recognition of MCCs in mammograms using an improved neural classifier", Eng. Appl. Artif. Intell. 24 (4) (2011) 638–645.

[49] B. Verma, J. Zakos, "A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques", IEEE Trans. Inform. Technol. Biomed. 5 (1) (2001) 46–54.

[50] L. Wei, Y. Wei, Y. Yang, R.M. Nishikawab, "Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis", Pattern Recognition. 42 (6) (2009) 1126–1132.

[51] L. Wei, Y. Yang, R.M. Nishikawa, et al., "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications", IEEE Trans. Med. Imaging 24 (3) (2005) 371–380.

[52] D. Soria, J.M. Garibaldi, F. Ambrogi, et al., "A non-parametric version of the Naive Bayes classifier", Knowledge Based System. 24 (6) (2011) 775–784.

[53] M. Salamo, M. Lopez-Sanchez, "Adaptive case-based reasoning using retention and forgetting strategies", Knowledge Based System. 24 (2) (2011) 230–247

[54] Sheshadri H.S, and Kandaswamy A, "Detection of breast cancer tumor based on morphological watershed algorithm", GVIP, 2005, pp. 17-21.

[55] Sheshadri H.S, and Kandaswamy A, "Experimental investigation on mammogram segmentation for early detection of breast cancer", Journal of Computerized Medical Imaging and Graphics, Elsevier science Vol. 31, 2005, 46-48

[56] Sheshadri H.S. and Kandaswamy A, "Mammogram image analysis using recursive watershed algorithm", National Journal of Technology, Vol. 1, No. 1, 2004, pp. 73-77.

[57] Sheshadri H.S, and Kandaswamy A, "Computer aided decision system for early detection of breast cancer", Indian Journal of Medical research, Vol. 124, 2006, pp. 149-154

[58] S. Oporto-D´ıaz, R. R. Hernandez-Cisneros and H. Terashima-Mar´ın, "Detection of microcalcification clusters in mammograms using a difference of optimized Gaussian filters", in Proceedings of the Second International Conference on Image Analysis and Recognition, ICIAR 2005, Toronto, ON, Canada, pp. 998–1005, 2005.

[59] Karssemeijer, N and Hendrikis, L. (1997). Computer assisted reading of mammograms Eur. Radiol. (7), 743-748

[60]Kim, J, K and Park H. W. (1999). Statistical textural features for detection of microcalcifications in digitized mammograms. IEEE Transactions on Medical Imaging (18), 231-238

[61] Mushlin, R and Shapiro, K, D.(1998). Estimating the Accuracy of screening mammography: A meta analysis. Journal of Preventive Medicine vol.14 (2)143-153

[62] Arun kumar M.N and H.S. Sheshadri, "Breast contour extraction and pectoral muscle segmentation in digital mammograms", International Journal of Computer Science and Information Security, Vol 9, No.2, February 2011.

[63] Rolando R. Hern´andez-Cisneros and Hugo Terashima, "Evolutionary Neural Networks Applied To The Classification Of Microcalcification Clusters In Digital Mammograms", 2006 IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006