by inverting the ratio of prior distributions between minority and majority class. Here instead, optimize the cost ratio locally.

3. Dealing with Class Imbalance with ABC-SVM

On dealing with imbalanced datasets, literature researchers focused on the data level and the classifier level. At the data level, the common task is the modification of the class distribution whereas at the classifier level many techniques were introduced such as manipulating classifiers internally, ensemble learning, one-class learning and CFL.

1.1 Cost-Free SVM

Support Vector Machines (SVM), which has strong mathematical foundations based on statistical learning theory that has been successfully adopted in various classification applications. SVM aims maximizing a margin in a hyper plane separating classes. However, it is overwhelmed by the majority class instances in the case of imbalanced datasets because the objective of regular SVM is to maximize the accuracy. In order to provide different costs associated with the two different kinds of errors, cost-sensitive SVM (CF-SVM) [15] is a good solution. CF-SVM is formulated as follows:

$$\min \frac{1}{2}\|w\|^2 + C_+ \sum_{i;j_t=+1} \xi_i + C_- \sum_{j;y_j=-1} \xi_i$$
$$\text{st } y_i[(w^T x_i) + b] \geq 1 - \xi_i \; \forall \; i = 1, L, .. n \; \xi_i \; \geq 0$$

where the C_+ is the higher misclassification cost of the positive class that is the primary interest, where $C\_$ is the lower misclassification cost of the negative class. Using the different error cost for the positive and negative classes, this SVM hyper plane could be pushed away from the positive instances. In this paper, we fix $C\_ = C$ and $C_+ = C \times C_{rf}$, where C and C_{rf} are respectively the regularization parameter and the ratio of misclassification cost. On the construction of cost sensitive SVM, this misclassification cost parameter plays an indispensable role. In general, the Radial Basis Function (RBF kernel) is a reasonable first choice for the classification of the nonlinear datasets, as it has fewer parameters (γ).

1.2 Optimized cost free SVM by measure of imbalanced data

SVM tries to minimize the regularized hinge loss; it is driven by an error based objective function. Yet, the obtained overall accuracy is not an appropriate evaluation measure for imbalanced data classification. Finally, there is an inevitable gap between the evaluation measure by which the classifier is to be evaluated and the objective function based on which the classifier is trained. The classifier for imbalanced data learning should be driven by more appropriate measures. We inject the appropri-ate measures into the objective function of the classifier in the training with ABC. The common evaluation for imbalanced data classification is G-mean and AUC. However for many classifiers, yet the learning process is still driven by error based objective functions. In this paper we explicitly treat the measure itself as the objective function when training the cost sensitive learning. Here, we designed a measure based training framework for dealing with imbalanced data classification issues. A wrapper paradigm proposed that discovers the amount of re-sampling for a dataset based on optimizing evaluation functions like the f-measure, and AUC. To date, there is no research about training the cost sensitive classifier with measure based objective functions. This is one important issue that hinders the performance of cost-sensitive learning.

Another important issue of applying the cost-sensitive learning algorithm to the imbalanced data is that the cost matrix is often unavailable for a problem domain. The misclassification cost, especially the ratio misclassification cost, plays a crucial role in the construction of a cost sensitive approach; the knowledge of misclassification costs is required for achieving expected classification result. However, the values of costs are commonly given by domain experts. They remain unknown in many domains where it is in fact difficult to specify the cost ratio information precisely. Also, it is not the correct method to set the cost ratio to the inverse of the imbalance ratio (the number of majority in-stances divided by the number of minority instances); especially it is not accurate for some classifier such as SVM. Heuristic approaches were used search the optimal cost matrix in some of the cost sensitive learning such as Genetic Algorithm or grid search to find the optimal cost setup. Feature subset selection and the intrinsic parameters of the classifier have a significant bearing on the performance, apart from the ratio misclassification cost information. These both factors are not only important for classification of imbalanced data, but also applicable for any kind of classification. The technique, feature selection is used for selecting a subset of discriminative features for building robust learning models by removing most irrelevant and redundant features from the data. Furthermore, this advanced optimal feature selection can concurrently achieve good accuracy and dimensionality reduction.

Unfortunately, the imbalanced data distributions are often accompanied by high dimensionality in real-world datasets such as bio-informatics, text classification and CAD (Computer Aided Detection). It is important to select features that can capture the high skew in the class distribution. Furthermore, proper intrinsic parameter setting of classifiers like regularization of cost parameter and the kernel functional parameter of SVM can improve the classification performance. It is necessary to use the grid search to optimize the kernel parameter and regulation parameters. Also, these three factors influence each other. Thus, we obtain the optimal ratio of cost of misclassification, feature subset selection and intrinsic parameters must occur simultaneously.

Based on the reasons above, our specific goal is to devise a strategy to automatically determine the optimal factors during training of the cost sensitive classifier oriented by the imbalanced evaluation criteria (G-mean and AUC). For binary class classification, there is only one parameter called cost parameter i.e., the relative cost information,

known as ratio misclassification cost factor *Crf*. Since the RBF kernel is selected for the cost sensitive SVM, γ and C are the parameters to be optimized. We need to combine the discrete and continuous values in the solution representation since the costs and parameters we intend to optimize are continuous while the feature subset is discrete as like each and every feature is represented by a 1 or 0 for whether it is selected or not. In our proposed method, the food source (solution) is randomly generated in the initial process. Then, each solution is evaluated through SVM classifier. The solving of SVM is generally a quadratic programming (QP) problem; sequential minimal optimization (SMO) will be applied in this study to optimize the computation time of training period of SVM.

SMO divides the large QP problem into series of smallest possible QP problem to avoid the intensive time and memories required. For this study, the radial basis function (RBF) is used as the kernel function of non-linear SVM classifier to learn and recognize pattern of input data from the training set. The equation of RBF function is defined as

$$K(x_i, x_j) = e^{-r(x_i, x_j)^2} \quad (1)$$

where r is a kernel parameter in RBF function. Then, a testing set is used to determine the classification accuracy for the input dataset. The fitness value is obtained by the classification accuracy of this testing dataset. For a small- and medium-sized data, the accurate value can be estimated by using the 10-fold cross-validation method. The method will randomly separate data into 10 subsets; one subset is used as a testing set while the remaining nine subsets are used as training sets. The process will be performed for 10 times in total. As a result, each subset will be used once as a validation or testing set. The accuracy of classification will be obtained from the average of all correctness values from 10 rounds. For the large-sized data, the holdout method is applied. This method divides the data into two parts for construction training and testing models. Each food source is modified based on the process of updating feasible solution by employed bees as expressed in the equation (2) where Φ is a random number in the range between [-1,1], negative one and one.

$$v_{ij} = x_{ij} + \Phi(x_{ij} - x_{kj}) \quad (2)$$

The equation (2) returns a numerical number of a new candidate solution vij from their current food source xij and their neighboring food source xkj. However, for dimension reduction these solutions must be binary numbers. Thus, this numeric solution must be converted into either 0 or 1 by using equation (3) and (4) as follows:

$$S(v_{ij}) = \frac{1}{1+e^{-v_{ij}}} \quad (3)$$

$$if\ (rand < S(v_{ij}))\ then\ v_{ij} = 1; else\ v_{ij} = 0 \quad (4)$$

The equation (3) is a sigmoid limiting function into the interval [0.0, 1.0]. Then, a random number between range [0.0, 1.0], rand, is used to determine the binary value of the solution in the equation (4). The new candidate solution must be evaluated with SVM classifier. If the new fitness value is better than the current one, the employed bees will replace its solution with this new candidate solution; otherwise, the new candidate solution will be ignored. Onlooker bees will select a food source according to the calculated probability of each food sources based on the equation (5), once after employed bees share information of their solutions.

$$P_i = \frac{fit_i}{\sum_{i=1}^{N} fit_i} \quad (5)$$

where P_i is the probability value of the solution i, N is the number of all solutions, and fit_i is the fitness value of the solution i. Thus, the solution with higher fitness value will have greater opportunity to be selected by the onlooker bees. After the onlooker bees select their desirable food source, the bees will perform the process of updating feasible solution similar to employed bees. The whole processes will be repeated until the termination criterion is reached. The dimension reduction process of ABC-SVM is summarized as follows:

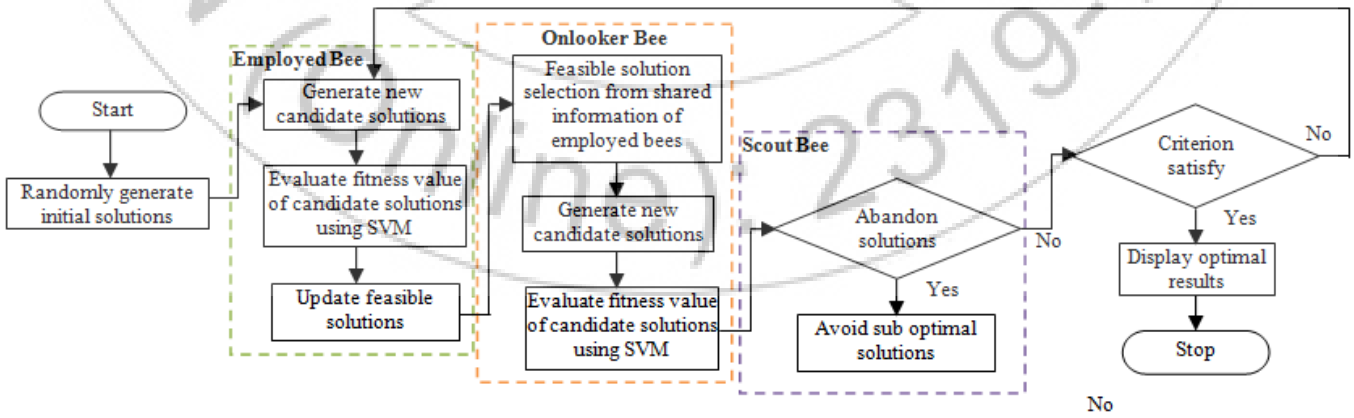

Figure 1: The flowchart of ABC-SVM method

4. Experimental Results and Discussion

1.3 Dataset Description

To evaluate the classification performance of our proposed method in different tasks of classification and to compare with other methods specifically devised for imba-lanced data, we tried several datasets from the UCI database. We used all available datasets from the combined sets used. This also ensures that we did not choose only the datasets on which our method performs better. The minority class label (+) is indicated in Table 1 and the chosen datasets contains diversity in the number of attributes and imbalance ratio. Also, these shown datasets have both continuous and categorical attributes of data. All the experiments are conducted by 10-fold cross-validation.

Table 1: The data sets used for experimentation the dataset name is appended with the label of the minority class (+)

Dataset	*Instances*	*Features*	*Class Imbalance*
Hepatitis (1)	155	19	1:4
Glass(7)	214	9	1:6
Segment(1)	2310	19	1:6
Anneal(5)	898	38	1:12
Soybean(12)	683	35	1:15

The comparison is conducted between our method and the other state-of-the-art imba-lanced data classifiers, such as the random under-sampling (RUS), SMOTE, SMOTEBoost, and SMOTE combined with asymmetric cost classifier. The re-sampling rate of under-sampling algorithm such as the SMOTE and SMOTEBoost is unknown. In order to compare equally, in our experiments, no matter whether it is under-sampling or over-sampling method, we used the evaluation measure as the optimization objective of the re-sampling method to search the optimal re-sampling level. The steps of both increment and decrement are set at 10%. This is a greedy search kind of approach which repeats the process greedily, upto no performance gains are observed.

Then the optimal rate of resampling is decided in an iterative fashion according to the evaluation metrics. Hence in the each fold, these data of training set is separated into training subset and validating subset for searching the appropriate rate parameters. The evaluation metrics are also used with the G-mean and AUC. For the CS-SVM with SMOTE, for each search of re-sampling, the optimal misclassification cost ratio is determined by searching under the evaluation measure guiding under the current over-sampling level of SMOTE.

1.4 Execution Time Comparison

This graph contains execution time of planned and existing system. The execution time of existing method is incredibly high compared with the planned system. The initial purpose of CF-ABC-SVM approach is there employing a screening methodology it'll be useful for mechanically effort class imbalance problem. This is shown in Fig.2. From the results it is observed that the proposed work takes less computational time to solve the class imbalance problem.

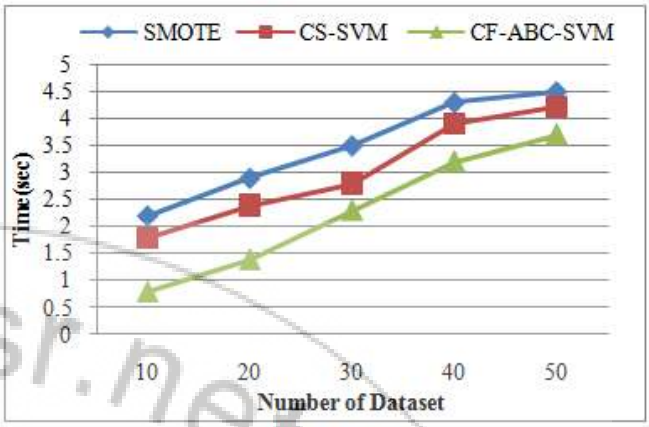

Figure 2: Execution Time Comparison

1.5 Accuracy

This graph contains execution time of planned and existing system. The accuracy of proposed system is incredibly high compared with the planned system. The initial purpose of CF-ABC-SVM approach is there employing a screening methodology it'll be useful for mechanically effort class imbalance problem. This is shown in Fig.3. From the results it is observed that the proposed work takes less computational time to solve the class imbalance problem.

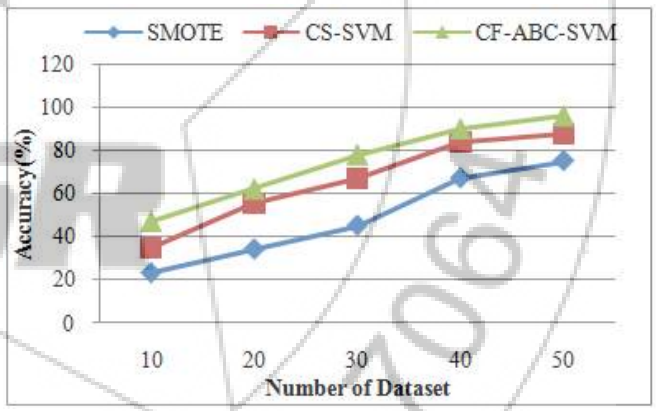

Figure 3: Accuracy Comparison

5. Conclusion

Learning with class imbalance is a challenging task. We propose a wrapper paradigm oriented by the evaluation measure of imbalanced dataset as objective function with respect to cost of misclassification, selection of feature subset and intrinsic parameters of SVM. Our measure oriented framework could wrap around an existing cost-sensitive classifier. The proposed method has been validated on some benchmark imbalanced data and real-world application. The obtained experimental results in this study have demonstrated that the proposed framework provides a very competitive solution to other existing state-of-the-arts methods, in terms of optimization of G-mean and AUC for conflicting imbalanced classification problems. These experimental results confirm the advantages of our approach that shows the promising

perspective and new understanding of cost sensitive learning. In the future research, we will extend the framework to the imbalanced multiclass data classification.

References

[1] Chawla, N.V., Japkowicz, N.& Kolcz, A. (2004): Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets 6 (1):1-6.
[2] Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. (2006): Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering:25-36.
[3] Lewis, D & Catlett, J.(1994). Heterogeneous uncertainty sampling for supervised learning. In Cohen, W. W. & Hirsh, H. (Eds.), Proceedings of ICML-94, 11th International Conference on Machine Learning, (pp. 148-156)., San Francisco. Morgan Kaufmann
[4] Murphy, P. & Aha, D.(1994). UCI repository of machine learning databases. University of California-Irvine, Department of Information and Computer Science.
[5] Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P.(2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence and Research 16, 321-357.
[6] Wu, G. & Chang, E. (2003). Class-Boundary Alignment for Imbalanced Dataset Learning. In ICML 2003 Workshop on Learning from Imbalanced Data Sets II, Washington, DC.
[7] Kubat, M., Holte, R., & Matwin, S.(1998). Machine learning for the detection of oil spills in satellite radar images. Machine Learning, 30, 195-215
[8] Weiss G., McCarthy K., Zabar B. (2007): Cost-sensitive learning vs. sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? IEEE ICDM, pp. 35–41.
[9] Yuan, B. & Liu, W.H. (2011): A Measure Oriented Training Scheme for Imbalanced Clas-sification Problems. Pacific-Asia Conference on Knowledge Discovery and Data Mining Workshop on Biologically Inspired Techniques for Data Mining. pp:293–303.
[10] Akbani, R., Kwek, S. & Japkowicz, N. (2004): Applying support vector machines to imba-lanced datasets. European conference on machine learning.
[11] Chawla, N.V., Cieslak, D.A., Hall, L. O. & Joshi, A. (2008): Automatically countering im-balance and its empirical relationship to cost. Utility-Based Data Mining: A Special issue of the International Journal Data Mining and Knowledge Discovery.
[12] Li, N., Tsang, I., Zhou, Z. (2012): Efficient Optimization of Performance Measures by Classifier Adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume: PP , Issue: 99, Page(s): 1.
[13] M.A. Maloof, "Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown," Proc. Workshop on Learning from Imbalanced Data Sets II. Int'l Conf. Machine Learning, 2003
[14] D.J. Hand, "Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve," Machine Learning, vol. 77, no. 1, pp. 103-123, Oct. 2009
[15] C.C. Friedel, U. R¨ uckert, and S. Kramer, "Cost Curves for Abstaining Classifiers," Proc. Workshop on ROC Analysis in Machine Learning. Int'l Conf. Machine Learning, 2006.
[16] B. Zadrozny and C. Elkan, "Learning and Making Decisions When Costs and Probabilities are Both Unknown," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 204-213, 2001.
[17] Y. Zhang and Z.-H. Zhou, "Cost-Sensitive Face Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 10, pp. 1758-1769, Oct. 2010.
[18] M. Wasikowski and X.-W. Chen, "Combating the Small Sample Class Imbalance Problem Using Feature Selection," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 10, pp. 1388-1400, Oct. 2010.
[19] S. Wang and X. Yao, "Relationships between Diversity of Classification Ensembles and Single-Class Performance Measures," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 1, pp. 206-219, Jan. 2013.
[20] T. Pietraszek, "Optimizing Abstaining Classifiers using ROC Analysis," Proc. Int'l Conf. Machine Learning, pp. 665-672, 2005.
[21] G. Fumera, F. Roli, and G. Giacinto, "Reject Option with Multiple Thresholds," Pattern Recognition, vol. 33, no. 12, pp. 2099-2101, 2000.
[22] M. Li and I.K. Sethi, "Confidence-Based Classifier Design," Pattern Recognition, vol. 39, no. 7, pp. 1230-1240, Jul. 2006.
[23] Xiaowan Zhang and Bao-Gang Hu, Senior Member, 'A New Strategy of Cost-Free Learning in the Class Imbalance Problem' IEEE 1041-4347 (c) 2013 IEEE