An Innovative Optimization Algorithm for Feature Selection –A Comparative study

T.DEEPA
Research Scholar, Department of Computer Science
Karpagam University
Coimbatore, Tamil Nadu.

Dr.M.PUNITHAVALLI
Director& Head, MCA Department
Sri Ramakrishna Engineering College
Coimbatore, Tamil Nadu.

***Abstract*— The most crucial task in data mining is that extracting a subset of feature (Feature selection) from the unequal class distribution (Imbalanced dataset). This paper proposes a study of various methods for feature selection and balancing the dataset. It concludes that the PSO optimization with SVM classification works better for selecting a subset of feature from the sparse dataset.**

Keywords-Feature Selection, Imbalanced dataset, Sampling, SVM Classification, fuzzy Evolutionary Sampling,Defuzzification, PSO Optimization.

I. INTRODUCTION

There are two issues in mining data from a high-dimensional data space. The former is the unequal class distribution in which one class outnumbers the other called *Imbalanced dataset.* The later is selecting the subset of features from the original dataset known as *Feature selection.* Extracting a literal feature from the High-Dimensional Imbalanced dataset without Duplication and loss of data is a crucial task. This paper focus on the above said problem and proposes by comparing various techniques to balance the dataset and to select the appropriate features from the dataset.

II. RELATED WORK

Feature Selection is a spirited research area in many fields of data mining such as bio-informatics, statistics, etc. To extract features from the original data space there are three common techniques such as Filter techniques, Wrapper method and embedded methods. Filtering is the simplest technique to select the features, it is independent of the classifier and wrapper method involves classifier to extract the features. The disadvantage of the existing technique is that filtering suits for small training sets and in case of speed it fails and wrapper method is computationally very expensive. The embedded method results in poor classifier performance.

The proposed work accounts the problems in the existing technique and suggests the appropriate solution by comparing various techniques that overrides the problem encountered in selecting feature from the Imbalanced dataset.

III. METHODOLOGY USED

The optimal Feature selection is the spotted area in this proposed work. To achieve that, the problem is cut into two critical branches i) Balancing the Dataset ii) selecting the optimal feature. The methodologies are also classified based on the above two aspects. The proposed work compares various techniques and concludes the best technique used for feature selection in high-dimensional Imbalanced dataset.

A. Dataset Balancing Methods

1) *Random Sampling:* It is divided into three types Random under Sampling, Random over Sampling and Smote Sampling. Random under-sampling is a non-heuristic method that aims to balance class distribution through the random elimination of majority class examples and over-sampling balance class distribution through the random replication of minority class examples and SMOTE sampling is the combinational approach of both random under and over sampling. The drawback of this approach is loss of useful data and over fitting problem.
2) *Evolutionary Sampling Techniques:* It is implemented with the help of Genetic algorithm. The data are converted to 0's and 1's.The fitness value for each chromosomes, based on the result the features are selected either by under sampling or over sampling or smote sampling. The over fitting and loss of data are minimized.
3) *Granularity Learning and Fuzzy Evolutionary Sampling Technique:* It is a semi-supervised learning technique with the combination of fuzzy to minimize the loss of data. The initial feature is the one that satisfies the training and testing instances. Before classification it is sampled with the evolutionary sampling technique.

B. Optimal Feature Selection Methods

1) SVM(SUPPORT VECTOR MACHINE) CLASSIFICATION:

Support Vector Machine is a most powerful tool among the machine learning methods. In this proposed work it is used to overcome the disadvantages of classification

such as over fitting and misclassification while extracting Features from high-dimensional space. It is a supervised learning method to classify the dataset. The SVM method is well suited for high-dimensional data space. It lays a line between majority and minority class and classifies the data.

2) *Defuzzification Technique:* Defuzzification is the process of producing a quantifiable result in fuzzy rule. It is a useful tool for making specific decision. The set of rules is applied to the fuzzified input. The output of each rule is fuzzy. These fuzzy outputs need to be converted into a scalar output quantity the process of converting the fuzzy output is called defuzzification. Before an output is defuzzified all the fuzzy outputs of the system are aggregated with an union operator. The union is the *max* of the set of given membership functions and can be expressed as

$$\mu_A = \bigcup_i (\mu_i(x)) \quad (3)$$

A common and useful DFT technique is used in the proposed work called CDT (Centroid Defuzzification Technique).

3) *Particle Swarm Optimization:* PSO was proposed by Kennedy in 1995. It provides high-performance in large space. In our proposed work PSO is an added feature which selects the optimal feature from the original data set.

Table1.Algorithm for PSO

Step 1: Initialize the population - locations and velocities
Step 2: Evaluate the fitness of the individual particle (pf)
Step 3: Keep track of the individual's highest fitness (gb)
Step 4: Modify velocities based on (pf) and (gb) position
Step5: Update the particles position
Step 6: Terminate if the condition is met
Step 7: Go to Step 2
From the above steps we can select optimal features

IV. RESULTS AND DISCUSSION

The Experiment is carried out on micro array dataset called Lymphoma, Lung cancer and colon data set. These dataset are normally imbalanced.

The factors such as Accuracy, number of feature selected, Error and time of each sampling and feature selection methods are considered.

ACCURACY COMPARISION:

Dataset	EROS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	73.2668	82.7982	90.9969	93.7942	94.3176
LungCaner	75.9285	80.6851	87.0524	94.2362	95.7515
Colon	72.0023	88.4348	89.2373	92.4475	95.2797

Table 2. Accuracy comparison of over sampling

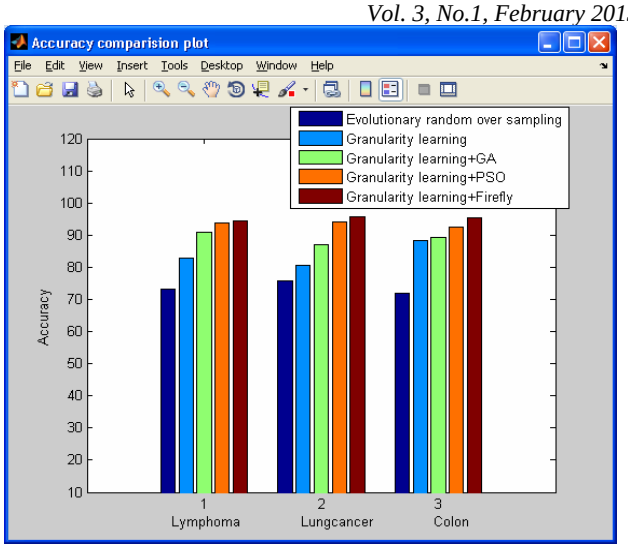

Figure 1: Accuracy comparison of over sampling

Dataset	ERUS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	66.8511	75.8169	82.3055	84.6409	91.3498
LungCaner	70.5786	80.3422	81.3878	88.6932	89.0990
Colon	70.9338	81.0806	82.5627	87.7236	95.5473

Table 2: Accuracy comparison of under sampling

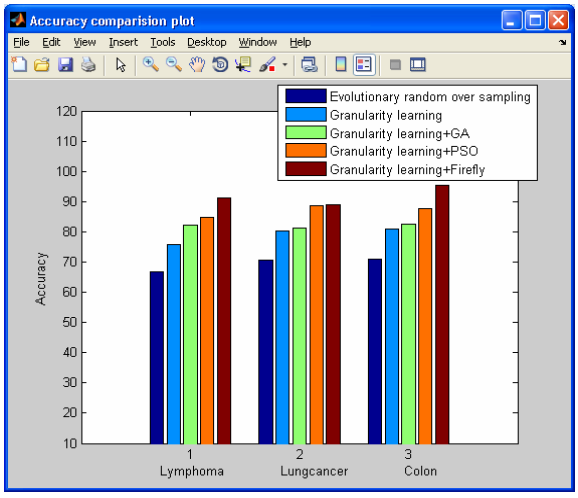

Figure 2 : Accuracy comparison of under sampling

Dataset	ESS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	71.8428	75.9962	81.2542	85.5539	85.9542
LungCaner	71.6612	75.4905	84.3022	86.6444	92.9210
Colon	75.5035	77.4436	79.9300	85.8869	89.1958

Table 4 Accuracy comparison of SMOTE sampling

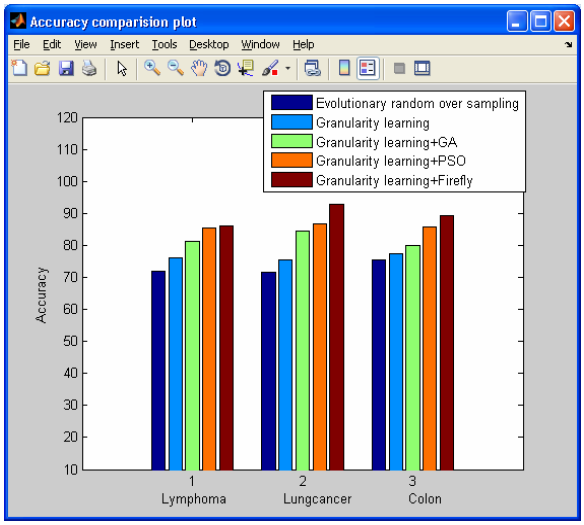

Figure 3: Accuracy comparison of SMOTE sampling

FEATURE SELECTION COMPARISION

Dataset	EROS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	51	50	45	41	39
LungCaner	51	50	45	42	39
Colon	52	49	46	41	39

Table 5: Feature selection comparison of over sampling

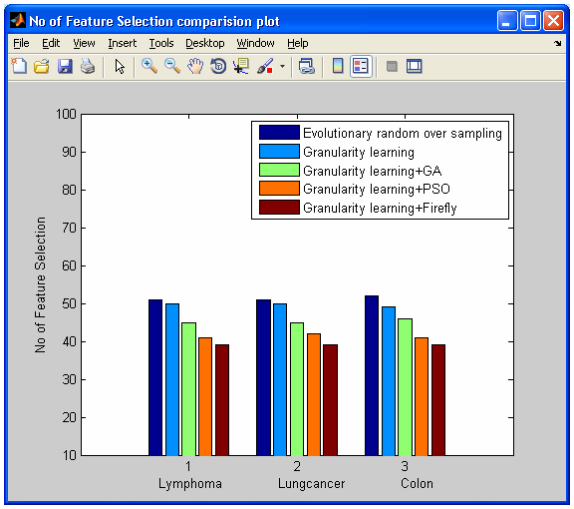

Figure 4: Feature selection comparison of over sampling

Dataset	EUS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	51	50	44	41	41
LungCaner	50	48	45	43	40
Colon	50	48	44	42	40

Table 6 Feature selection comparison of under sampling

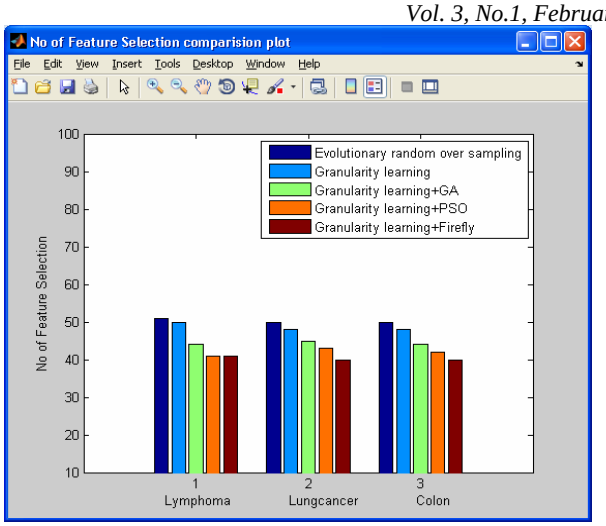

Figure 5: Feature selection comparison of under sampling

Dataset	ESS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	50	49	46	41	39
LungCaner	50	49	46	41	39
Colon	49	48	45	43	39

Table 7: Feature selection comparison of SMOTE sampling

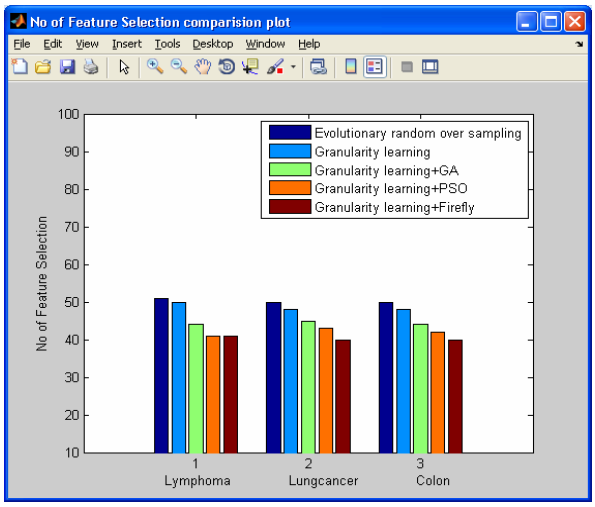

Figure 6: Feature selection comparison of SMOTE sampling

ERROR COMPARISION

Dataset	EROS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	0.9290	0.9121	0.8874	0.8205	0.7840
LungCaner	0.9757	0.9452	0.9449	0.8701	0.7019
Colon	0.9963	0.8986	0.8745	0.8282	0.7542

Table 8 Error comparison of over sampling

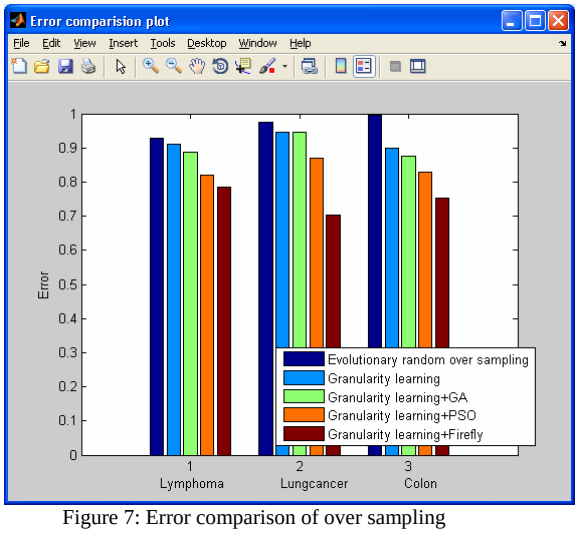

Figure 7: Error comparison of over sampling

Dataset	EROS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	0.8634	0.8542	0.8205	0.8037	0.6731
LungCaner	0.9104	0.8940	0.8813	0.8294	0.6876
Colon	0.8633	0.8346	0.8208	0.8043	0.7318

Table 9 Error comparison of under sampling

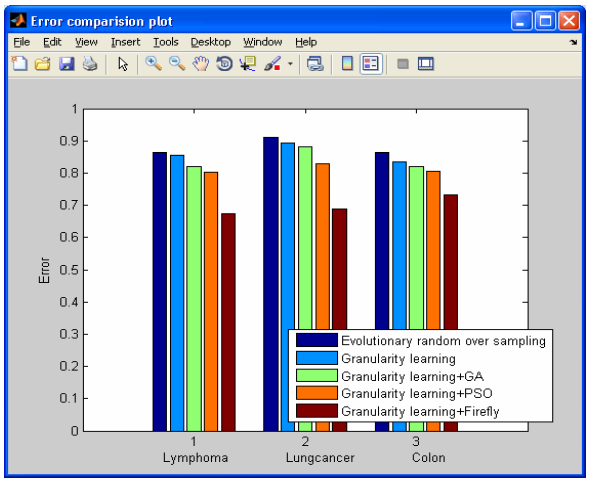

Figure 8: Error comparison of under sampling

Dataset	ESS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	0.9152	0.8275	0.8035	0.7370	0.7047
LungCaner	0.8957	0.8302	0.8204	0.7497	0.7147
Colon	0.8563	0.8417	0.8356	0.7325	0.7160

Table 10 Error comparison of SMOTE sampling

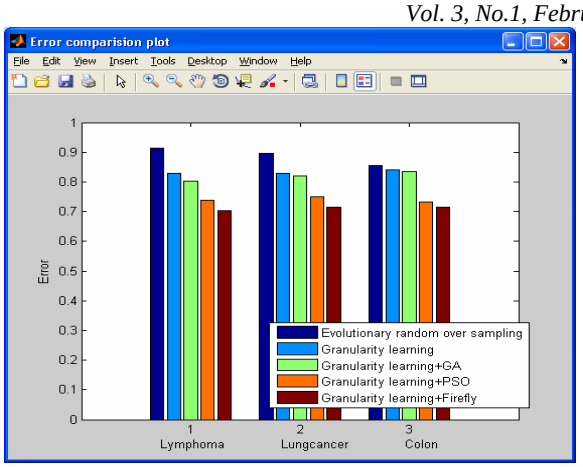

Figure 9: Error comparison of SMOTE sampling

TIME COMPARISION

Dataset	EROS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	2.4672	1.9579	1.9049	1.8685	1.2776
LungCaner	2.2139	2.1619	2.0817	1.9600	1.4000
Colon	2.4651	2.3690	2.1413	1.7570	1.0214

Table 11 Time comparison of over sampling

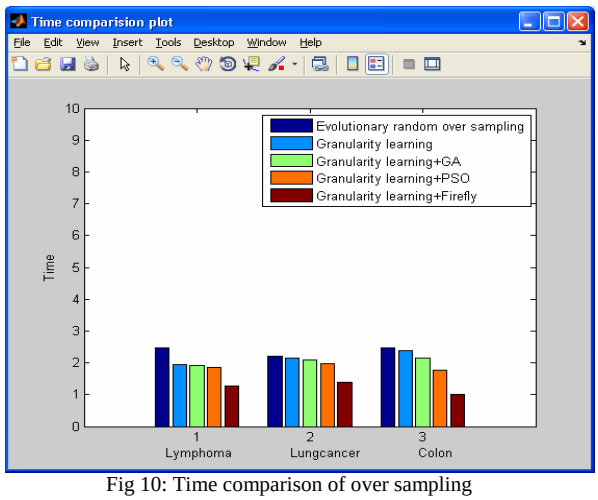

Fig 10: Time comparison of over sampling

Dataset	ERUS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	2.3827	2.3080	2.1360	1.8723	1.1090
LungCaner	2.6183	2.0596	2.0532	1.9494	1.9047
Colon	2.2798	1.9674	1.9292	1.3761	1.3516

Table 12 Time comparison of under sampling

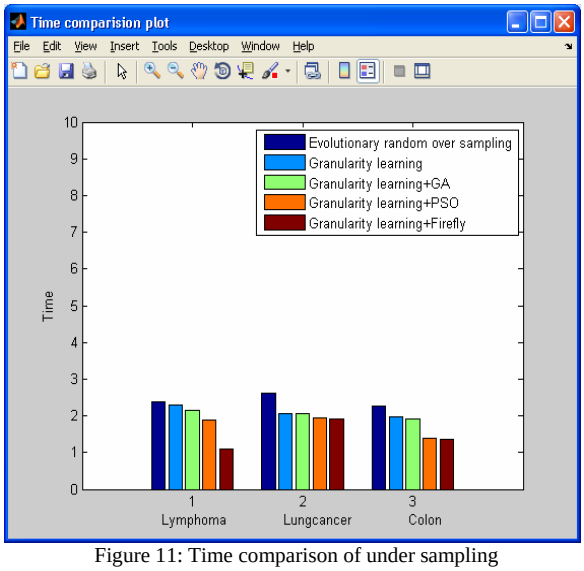

Figure 11: Time comparison of under sampling

Dataset	ESS	GL	GL+GA	GL+PSO	GL+FA
Lymphoma	0.9152	0.8275	0.8035	0.7370	0.7047
LungCaner	0.8957	0.8302	0.8204	0.7497	0.7147
Colon	0.8563	0.8417	0.8356	0.7325	0.7160

Table 13 Time comparison of SMOTE sampling

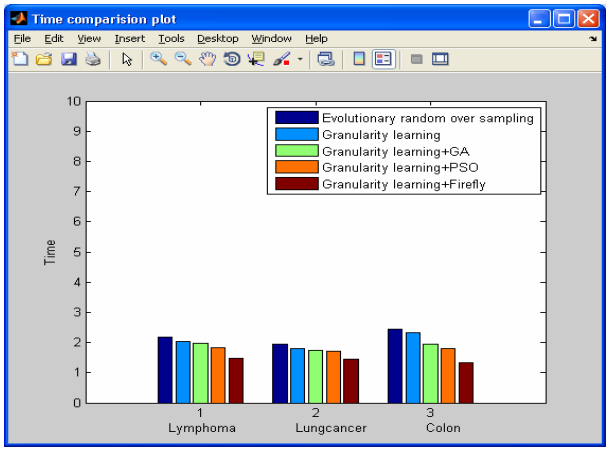

Figure 12 : Time comparison of SMOTE

V. CONCLUSION AND FUTURE WORK.

The above result shows that the SMOTE sampling and PSO optimization works better. Since PSO faces convergence problem to overcome that new optimization algorithm can be included.

The position that are calculated in the PSO algorithm sometimes may or may not be optimal so other criteria can be included to derive the best position of the particle.

REFERENCES

[1] O.Cardon,F.Herrera,P.Villar, “A genetic learning process for the scaling factors,granularity and contexts of fuzzy rule-based system database” Elsevier Information Sciences 2001 85-107.

[2] T. Jirapech-Umpai and S. Aitkin. “Feature selection and classification for microarray data analysis: Evolutionary methods for Identifying predictive genes.BMC bioinformatics, 6(1):148, 2005.

[3] L N. V. Chawla, L. O. Hall, K. W. Bowyer, and W.P. Kegelmeyer “SMOTE:Synthetic Minority Oversampling Technique”. *Journal of Artificial Intelligence Research*, 16:321- 357, 2002.

[4] T.Deepa, Dr.M.Punithavalli,(2011) “A New Sampling technique and SVM classification for feature selection in High-dimensional Imbalanced Dataset”,3rd International conf on Electronics computer technology, vol 5 pg 401-404.

[5] T.Deepa, Dr.M.Punithavalli, (2010) “Evaluating the performance of various filtering Techniques for feature selection in High-dimensional Imbalanced Dataset”, ITFR journal, Aug 2011. Vol 1, No1. Pgs 1-4.

[6] T.Deepa, Dr.M.Punithavalli, (2010) “An E-SMOTE Sampling and SVM classification for feature selection in High-dimensional Imbalanced Dataset”, 3rd International conf on Electronics computer technology, vol 2 pg 322-324.

[7] T.Deepa, Dr.M.Punithavalli, (2010) “A GLFES and DFT Technique for feature selection in High-dimensional Imbalanced Dataset”, IJCSE,Vol 3,no:2, April-May 2012.Pgs 336-343.

[8] Fernandez, A., Garc´ıa, S., Del Jesus, M.J., Herrera, F.: A study of the behavior of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. Fuzzy Sets and systems159(18), 2378–2398 (2008).

[9] Pedro Villar, Alberto Fern´andez, and Francisco Herrera “A Genetic Algorithm for Feature Selection and Granularity Learning in Fuzzy Rule-Based Classification Systems for Highly ImbalancedData-Sets” Springer 2010,pgs741-750.

[10] Cui1, J. Zeng, and G. Sun, “A Fast Particle Swarm Optimization,”Int. J. of Innovative Computing, Information and Control, vol. 4, no6, pp. 1365–1380,2006.

[11] E. Elbeltagia, T. Hegazyb, and D. Grierson, “Comparison among five evolutionary-based optimization algorithms,” Advanced Engineering Informatics, vol. 19, pp. 43–53, 2005.

[12] Cano, J. R., Herrera, F., and Lozano, M ”Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study”. IEEE Transactions on Evolutionary Computation,7(6):561–575.(2003).