A GLFES and DFT Technique for Feature Selection in High-Dimensional Imbalanced dataset

T.Deepa

Research scholar, Karpagam University ,
Coimbatore, Tamil Nadu,India
DeepaRaman12@gmail.com

Dr.M.Punithavalli

Director, Sri Ramakrishna Engineering College
Coimbatore, Tamil Nadu,India
mpunitha_srcw@yahoo.co.in

Abstract:
Feature selection has been an active research area in pattern recognition, statistics ,and data mining communities Feature selection, is a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. Feature Selection in High-Dimensional Imbalanced Dataset (where one class outnumbers the others) plays a significant task in the field of Data mining. Discarding data and adding data sometimes may affect the performance. This paper proposes a new approach GLFES (Granularity learning Fuzzy Evolutionary Sampling) and DFT (Defuzzification Technique) for Feature Selection. It is evaluated on micro array datasets.

Keywords*:* Imbalanced dataset, Feature Selection, Fuzzy Evolutionary Sampling, Defuzzification Technique.

1 Introduction

Feature selection plays in the field of data mining such as statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selections often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. A Dataset is said to be Imbalanced if one class (majority) outnumbers the others (minority).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 a new technique to balance the dataset and to select the appropriate features from the dataset.

2 Related work

Recently there has been a considerable amount of attention devoted to Feature selection in the field of data mining. It has several achievements in areas like bio-informatics, statistics etc. There are two issues in selecting features from high-dimensional Imbalanced dataset i) Balancing the dataset. ii) Selecting the appropriate features.

The EST technique and SVM Classification were used in the existing approach to balance the dataset and to select the appropriate features. Evolutionary Under Sampling (EUS) and Evolutionary Over Sampling (EOS) and E-SMOTE techniques were used to solve the Imbalance problem.

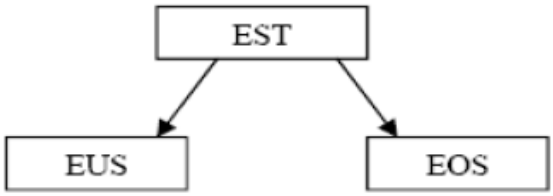

Figure 1: Classification of EST Technique

In this proposal approach GLFES and DFT techniques are proposed to balance the dataset and to select the appropriate features from the Imbalanced Dataset (IDS).

3 Methodology Used

Learning from Imbalanced Dataset is an important topic in the field of data mining. This problem is very representative since it appears in a variety of real-world applications including, medical applications, finance, telecommunications, biology and so on. Imbalanced Dataset occurs when the class distribution is not uniform. In this situation, the number of examples that represents one of the classes of the data-set (the concept of interest) is much lower than that of the other classes.
The previous work on this topic with Genetic algorithm shows a good behavior for the Evolutionary over sampling and Under sampling methods, especially in the case of the Evolutionary SMOTE methodology for balancing the dataset and SVM Classifiers were used to select the appropriate features but this method lack in accurate feature selection in high-dimensional IDS, to tackle with the above problem new techniques were developed in the proposed work.

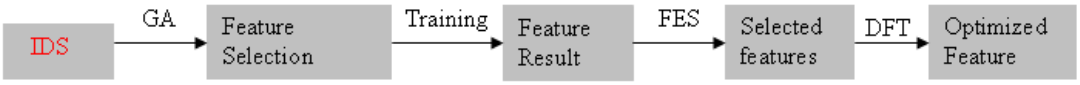

Figure 2: Steps in Proposed Work.

3.1. Granularity Learning Fuzzy Evolutionary Sampling (GLFES)

This contribution proposes a new method called GLFES (Granularity learning fuzzy Evolutionary Sampling).It is based on the traditional Genetic algorithm and on a Learning Methodology. It plays a vital role in both balancing and extracting the feature.

This is a standard generational algorithm for the IDS that allow us to select a set of variables (feature selection). Genetic algorithm works with a set of candidate solution called population. It retains the optimal solution after a series of iteration. In GA the data are transformed to 0's and 1's. The data which meets the requirement are selected and each individual's fitness is calculated with the help of the fitness function.

$$\text{Fitness} = \alpha + P / N \quad (1)$$

were α is the accuracy rate (ratio between true positive and true negative) and P is the total number of features i.e. 96 in the proposed work and N is the number of features selected.

The features that are selected by GA are learned individually and DB is built. This learning reduces the complexity in feature selection. In this, the feature that yields high granularity average is taken (i.e., which supports the work).

$$GL = TL - PL \quad (2)$$

Granularity Learning obtains a appropriate feature by finding the difference between the testing label (TL) and Predicted label (PL).

Fuzzy Rule based classification with the Genetic sampling shows a good result by avoiding the over fitting problem. It identifies the classes if the size of the majority class is higher than the minority class then the majority class is minimized to minority class without discarding the useful data. If the minority class is higher, this method adds up duplicate data and overcomes the over fitting problem. This process is handled by FES technique.FES balance the dataset and selects the appropriate feature. The steps in FES technique are

Step1: Create a random initial population.
Step2: Compute and save the fitness for each data using equation 1.
Step3: Define the selection probability.
Step4: Generate the selection probability.
Step5: Repeat step2 to 4 until the selection process is satisfied.
Step6: Train the selected population.
Step7: Find the predicted population.
Step8: Based on the rule imbalanced classis identified and FES technique is applied.

3.2. Defuzzification Technique (DFT)

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).

3.2.1. Centroid Defuzzification Technique (CDT)

This method is also known as centre of gravity or centre of area defuzzification. This technique was developed by Sugeno in 1985. This is the most commonly used technique and is very accurate.

The centroid method is one of the most popular algorithms for defuzzification because it provides smooth transitions from one output set to another, which is crucial to any control system. The centroid method is also known as the centre of mass or centre of moment method because it is based on calculating the area under the curve and then determining the centre of that area. Figure 3 shows the general form of two membership sets, μ_i, for i = 1, 2. ai is the centre of the i^{th} set, and Hi is the clipped height of the i^{th} set

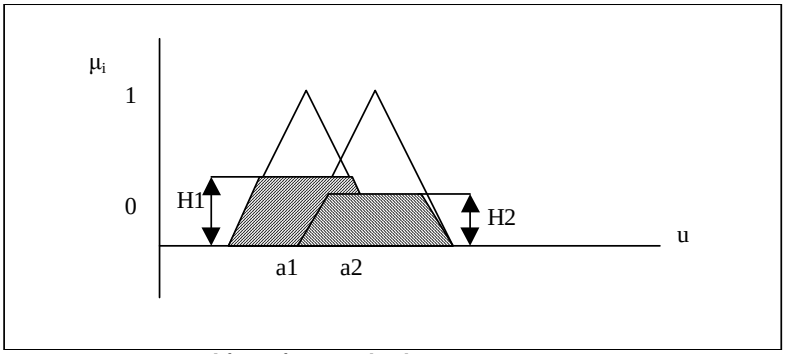

Figure 3: General form of two membership sets.

The calculation of the centroid is rather straightforward and can be calculated from equation (4). The centroid is calculated by multiplying the area of each set by its centre and then taking the sum of this product, this sum is then divided by the area under the entire curve.

$$Centroid = \frac{\sum_{i=1}^{N} a_i \int u_i du}{\sum_{i=1}^{N} \int u_i du} \quad (4)$$

Where ai is the center of the ith set, Hi is the clipped height of the ith set, N is the total number of set $\int u_i du$ is the area of the ith set. To calculate the area of the set below the curve ($\int u_i du$) is to find the area of the entire triangular set and then subtract from that the area of the upper triangle.

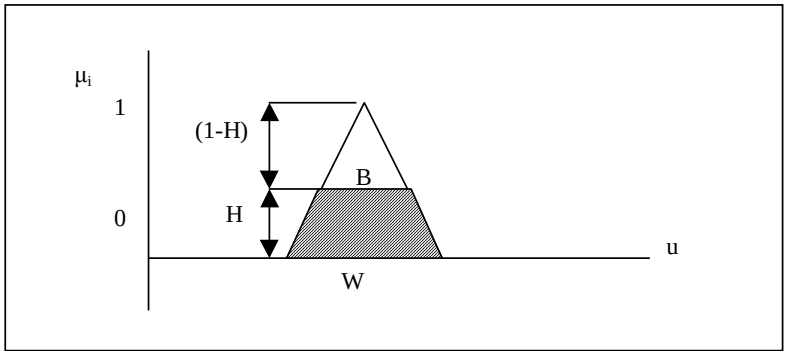

Figure 4: General form of area under the curve.

The area of the lower portion of the triangle can be computed by first determining the area of the whole triangle by Equation (5).

$$A_{total} = W \times 1 \quad (5)$$

where, A_{total} is the area of the full triangle, and W is the width of the base of the full triangle. Next, based on the law of similar triangles we can compute the width of the base of the upper triangle using Equations (6) and (7).

$$\frac{W}{1} = \frac{B}{1-H} \quad (6)$$

$$B = W \times (1-H) \quad (7)$$

where, B is the width of the base of the upper triangle, and H is the clipped height. Finally, we subtract the area of the upper triangle from the whole triangle in order to find the area, A, under the clipped height. This is expressed mathematically as:

$$A = \frac{1}{2}W - \frac{W \text{ x } (1-H)^2}{2} \tag{8}$$

We also need to consider the sets at either end of the membership range, which require a special case of the previous equations. In order to determine the area of these sets we break down the set into a triangle and a rectangle as shown in Figure 5.

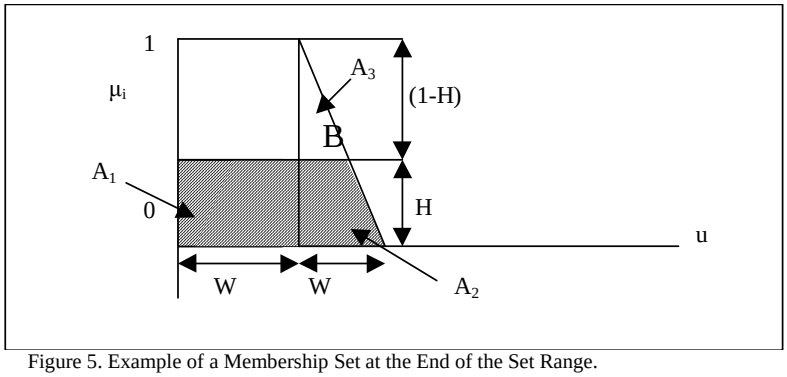

Figure 5. Example of a Membership Set at the End of the Set Range.

In this case, the area of interest A is A_1+A_2. We can calculate the areas of A_1 and A_2 as:

$$A_1 = WxH \tag{9}$$

$$A_2 = (A_2+A_3) - A_3 \tag{10}$$

We can determine A3 in the same manner as before, where, we use similar triangles to determine B. Using Equations (6) and (7) we can determine A_2.

$$A_2 = \frac{1}{2}W - \frac{W \times (1-H)^2}{2} \tag{11}$$

Finally, we calculate the total area by combining A_1 and A_2.

$$A = A_1 + A_2 = (W \times H) + (\frac{1}{2}W - \frac{W \times (1-H)^2}{2}) \tag{12}$$

Once the area of each set has been calculated then we use Equation (4) to find the centroid of all of the sets. This will yield the centroid of the curve, which is in fact, the location where the area under the curve to the right of the centroid is equal to the area under the curve to the left of the centroid. This will be the crisp value of the output set.

4 Experimental Result

The main objective of this work is to extract a subset of features from a high-dimensional imbalanced dataset and to balance the dataset. The GLFES is used to find the appropriate or suitable features based on the fitness value. The features are trained by Granularity method and it results in Predicted features. These features are then sampled or balanced using FES Technique. The resultant features are defuzzified to obtain the crisp features. The Centroid Defuzzification Technique is used to obtain the crisp features.

The experiment is carried out on a micro array dataset called Lymphoma, lung cancer and colon dataset. The original microarray dataset is shown in Table 1.This Initial Population is given as Input to GLFES It selects 96 features from the original space. Genetic algorithm achieves a solution through iterations, here it is iterated 5 times. From this features an appropriate feature is selected by using FES and DFT Technique. This produces crisp features.

Table 1. The Microarray Dataset.

Dataset name	Abbreviation	Attributes	Instances	Positive Ins
Lung Cancer	cancer	12534	181	31
Colon	Colon	2001	62	22
Lymphoma	Cancer	7130	77	19

Table 2. Experimental result showing the feature selected by FES and DFT

Dataset name	Features selected as Input	Features selected by FES (under)and DFT	Accuracy%	Features selected by FES (over)and DFT	Accuracy%
Lung Cancer	96	67	99	56	62
Colon	96	73	80	79	83
Lymphoma	96	85	97	66	82

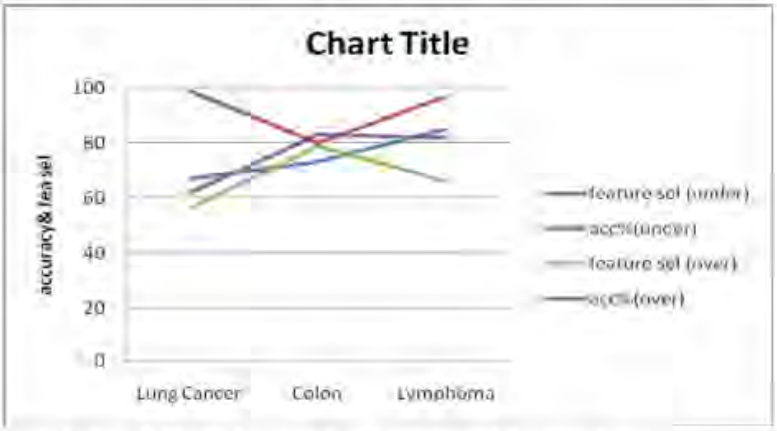

Figure 6.Chart showing the accuracy and Feature Selected

The above table and chart shows the features selected and their average accuracy percentage. The proposed work is compared with the existing approach in terms of accuracy and time.

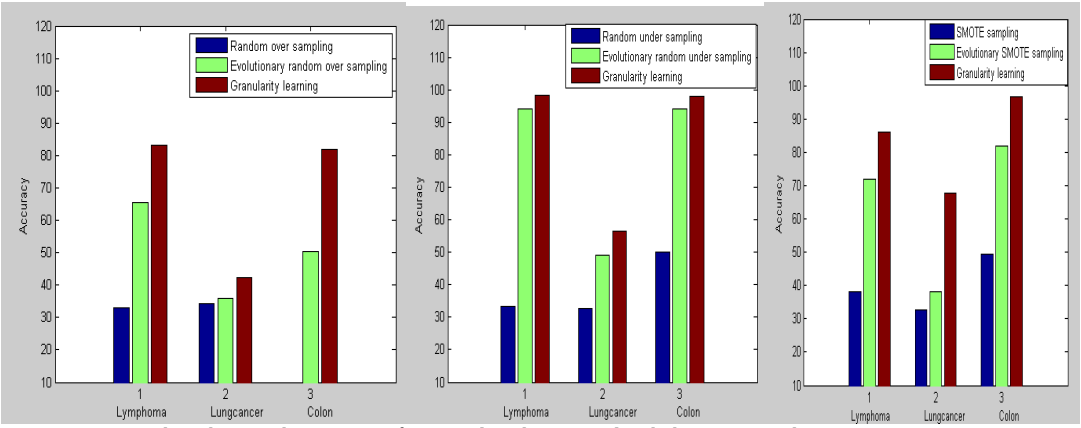

Figure 7: Chart showing the accuracy of proposed work compared with the existing techniques

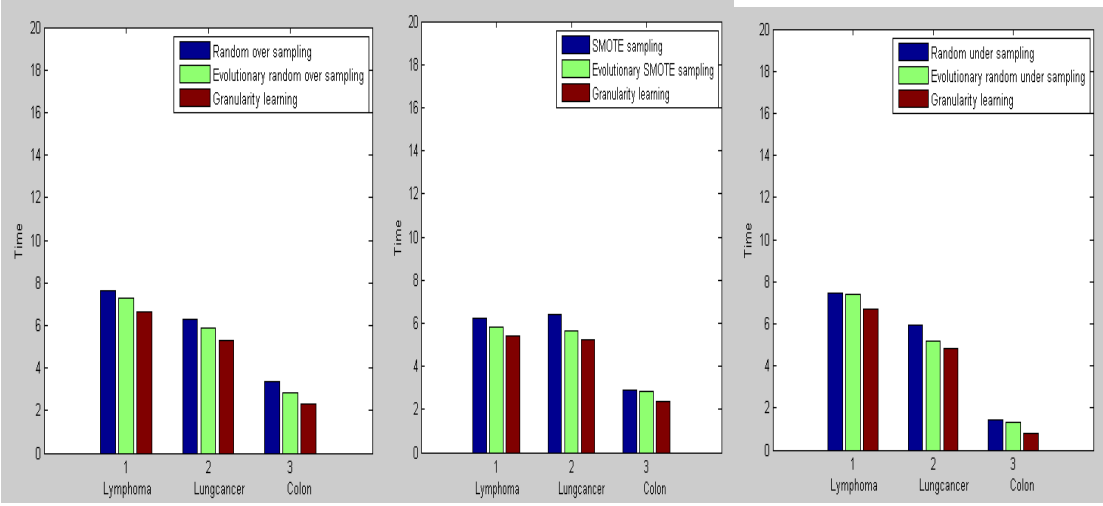

Figure 8: Chart showing the Time taken of proposed work compared with the existing technique

From the above figures it is concluded that the proposed yields better result than the existing work. In Future other sampling and optimization technique can be used to extract the features from the dataset.

5 Conclusion and Future work

This paper focused on extracting features from a high-dimensional Imbalanced dataset using GLFES and DFT technique with a Genetic algorithm, It solves the misclassification and over fitting problem. In this proposed work three types of micro array dataset was taken which is naturally imbalanced and the results were obtained with accurate features selected. In future more dataset can be considered, other sampling technique and optimization technique can be considered to extract the feature from the dataset.

REFERENCES

[1] E. A. P. A. Batista, R. C. Prati, and M. C.Monar (2004) "A study of the behavior of several methods for balancing machine learning training data."SIGKDD Explorations, 6(1):20- 29.

[2] In N. V. Chawla, N. Japkowicz, and A. Ko lcz,(2006) "Special Issue on learning from Imbalanced Dataset", Sigkdd Explorations Vol 6, issue 1, Pages1-6.

[3] In N. V. Chawla, N. Japkowicz, and A. Ko lcz, editors,(2003) proceedings of the ICML'2003

Workshop on Learning from Imbalanced Datasets.
[4] T. Jirapech-Umpai and S. Aitken.(2005) "Feature selection and classification for microarray data analysis: Eolutionary methods for identifying predictive genes. BMC bioinformatics, 6(1):148, 2005.
[5] Jason VanHulse, Amri NapoL itano, T.M, Khoshgoftaar (2009) "Feature Selection with High-dimensional Imbalanced data"IEEE Conference Data mining Tech.2009.
[6] H. Peng, F. Long, and C. Ding.(2005) "Feature selection based on mutual information: Criteria of Maxdependency, maxrelevance, and min- redundancy."IEEE Transactions on Pattern Analysis And Machine Intelligence, 27(8):1226–1238, 2005.
[7] T. S. Furey, N. Cristianini, N. Duffy, D. W.Bednarski, M.Schummer, and D. Haussler (2000) "Support vector machine classification and validation of cancer tissue samples using microarray expression data Bioinformatics, 16(10):906–914,
[8] I. Inza, P. Larraaga, R. Blanco, and A. J.Cerrolaza.(2004) "Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial Intelligence in Medicine, 31(2):91 – 103, Data Mining in Genomics and proteomics
[9] T.Deepa, Dr.M.Punithavalli, (2010) "Evaluating the performance of various filtering Techniques for Feature selection in High-dimensional Imbalanced Dataset", ICCIC IEEE Conf, Dec 2010.
[10] 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.
[11] T.Deepa, Dr.M.Punithavalli,(2011) "An E-SMOTE technique and SVM classification for feature Selection in High-dimensional Imbalanced Dataset", 2011 3rd International conf on Electronics computer technology, vol 2 pg 322-324.

Authors

T.Deepa, graduated with M.Sc in 2006 from Sri Ramakrishna College of arts & Science for Women, India and completed M.Phil from Bharathiar University, India during 2007-2008. Her areas of Interest include Software Engineering & Data Mining. She has about 4 years of teaching experience. Currently she is working as a Lecturer in Sri Ramakrishna college of Arts & Science for Women, India.

Dr. M.Punithavalli is presently working as Director & Head of the Dept of Computer Science, Dr.SNS College of arts and science for Women College, India. She has published more than twenty papers in National/&International journals. Her areas of interest includes E-Learning, Software Engineering, Data Mining, Networking and etc. She has about 16 years of teaching experience. She Is guiding many research scholars and has published many papers in National and International Conferences.