The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:

The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.89 e-ISSN: 2251-7545

A Feature Selection Method for Imbalance Data sets

Ilnaz Jamali
Department of Electrical and Computer Engineering
Shiraz University
Shiraz, Iran
ilnazjamali@yahoo.com

Sattar Hashemi
Department of Electrical and Computer Engineering
Shiraz University
Shiraz, Iran
s_hashemi@shirazu.ac.ir

***Abstract*—This paper introduces a game theoretic framework for feature selection in imbalance data sets. In this method which is called FSSH (Feature Selection based on Shapley value), first some coalitions will be constructed and the marginal importance of each feature in its coalition will be computed. Then, the weighted mean of each feature's value considered as the Shapley value. Finally features will be ranked according to their Shapley value and high ranked features will be selected in the realm of feature selection. Experimental results and comparison with several existing feature selection methods show the advantages of presented approach across the data sets adopted in this study.**

Keywords—feature selection, imbalance data sets, game theory, Shapley value.

I. INTRODUCTION

One of the greatest challenges in machine learning and data mining researches is class imbalance problem. Imbalance problems can appear in two different types of data sets: binary problems, where one of the two classes comprises considerably more samples than the other and multiclass problems, where each class contains a tiny fraction of the samples. However, any data set that exhibits an unequal distribution between its classes can be considered as imbalanced [4]. Imbalanced data sets introduce a significant reduction in performance of standard classifiers when they are invoked to learn data underlying concepts. The problem becomes even more sever when imbalanced data sets are involved with high dimensions.

In what follows we focus on optimizing performance measure which is the area under the ROC curve [7].

In this paper we use the Shapley value for ranking and finally selecting variables in an attempt to maximize the performance of a classifier on unseen data. Shapley value provides a set of contributions [5], which stands for the unique fair division of the game's worth among the different features. Hence, the true importance of each feature can be measured from a contribution of this feature to a function, that is, the part it plays in the successful performance of that function [6, 12, 13].

The rest of this paper is organized as follows: Section 2 introduces the necessary background from game theory in detail. In this Section the payoff function which is suitable for imbalance problems will be presented. Our proposed method is also explained in this Section. Experimental results and comparison with some well known feature selection methods on imbalance problems are discussed in Section 3. Section 4 concludes the conclusion and finally in Section 5 we will introduce our future work.

II. PROPOSED METHOD

A. Background

In game theory, a cooperative game is a game where coalitions of players (S) compete with each other in order to achieve high payoff. In other words, a coalition of players cooperates, and obtains a certain overall gain from that cooperation. A coalitional game is a pair of (N, v) which N is a finite number of players, and $v(S)$ is a real value assigned to each coalition S. In each coalition, contribution value of each player is calculated by a value function, which assigns a real value to each player. In other words, the value can be considered as a power of player. The value function is based on Shapley value. This can be considered as a measure of the utility of players in a game. Shapley value is defined as follows,

$$\Delta_i(S) = v(S \cup \{i\}) - v(S) \qquad (1)$$

$$\Phi_i(v) = \frac{1}{n!}\sum_{\pi\in\Pi} \Delta_i(S_i(\pi)) \qquad (2)$$

Where Π is the set of coalitions over N and $S_i(\pi)$ is the set of players appearing before the i^{th} player in

The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.89 e-ISSN: 2251-7545

coalition π. The Shapley value of a player is a weighted average of its contribution over all possible coalition of players. In addition, The Shapley value is one way to distribute the total gains to the players. In the realm of feature selection, *N* refers to the number of features [5]. The Shapley value also has an axiomatic foundation which will be considered below:

Axiom 1 (normalization):
For any game (N, v) it holds that $\sum_{i\in N} \Phi_i(v) = v(N)$
In the context of feature selection, this axiom implies that the performance on the dataset is divided fully between the different features [5].

Axiom 2 (Permutation invariance or symmetry):
For any (N, v) and permutation π on N it holds that $\Phi_i(v) = \Phi_{\pi(i)}(\pi v)$
This axiom implies that the value is not altered by arbitrarily renaming or reordering the features [5].

Axiom 3 (Preservation of carrier or dummy property):
For any game (N, v) such that $v(S \cup \{i\}) = v(S)$ for every S$\subseteq N$ it holds that $\Phi_i(v) = 0$
This axiom implies that a dummy feature that does not influence the classifier's performance indeed receives a contribution value 0[5].

B. Performance Measure

In this method we attempt to identify the most characterizing features which can minimize the classification error. For this purpose, first contribution of the paper has focused on the feature selection based on Shapley value. The second contribution of the paper shows how to use the AUC as a statistical measure to evaluate classifier's performance, and also as a value for a coalition of features. So, we use Area under the ROC Curve (AUC) to evaluate each subset of features. Whereas the datasets in this work have supposed to have only two classes, we can label each instance by negative or positive. For each positive sample (or negative ones and so on), we consider k nearest neighbors around it (k=1 reports the best for imbalance problems [9]) and constitute a criterion (3). To calculate the numerator, we set it initially to zero and for every positive sample of V that the positive samples between k nearest neighbors around it (SF (z)) are more than the negative ones, it is incremented by one and if the positives and negatives are equal, it is incremented by 0.5. The denominator is the multiplication of the positive samples of V by the negatives [1], [7].

$$V(F) = \frac{|\{(z,z\prime)\in v^2, y<y', S_F(X)<S_F(X')\}|}{\{(z,z\prime)\in v^2, y<y'\}} \quad (3)$$

C. Feature Selection based on Shapley Value (FSSH)

The process of main algorithm is described as follow. First of all, the dataset has been split into the unseen data and training sets. Next, the 10-fold cross validation has been used. Then in each fold, the Shapley value procedure is called. In our feature selection approach, we use the Shapley value to estimate the contribution value of a feature in the task of feature selection. First some coalitions will be constructed. It has been shown that we only need to construct coalitions with three features [5]. Then, the marginal importance of each feature in its coalition is computed using (1). At the end of this procedure, the weighted mean of each feature's value is considered as the Shapley value (2), finally features will be ranked according to their contribution value which is their Shapley value. Then, we can select the high ranked features in the realm of feature selection.

Algorithm 1. Pseudo code of the proposed method

Step 1: *Partition the dataset into training and testing sets. After that, training set is divided to training and validation sets.*
Step 2: *Start training phase.*
Step 3: *Construct coalitions.*
- *In this step, we only need to construct coalitions with three members.*

Step 4: *Calculate AUC for the coalitions.*
Step 5: *Computing the marginal importance for each feature according to the AUC value of the coalition.*
Step 6: *Take average on marginal importance of each feature as the Shapley value.*
Step7: *Rank all features from maximum to minimum value according to their Shapley value.*
Step8*: select 2% of high ranked features.*
Step 9: *Validation.*
Step 10: *Go to step 2 until 10 times.*
Step 11: *Testing.*

III. EXPERIMENTS

A. Benchmarks

The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.89 e-ISSN: 2251-7545

Table 1 indicates the characteristics of eight benchmark datasets that are used to validate this method.

TABLE I. CHARACTERISTICS OF DATA SETS USED IN EXPERIMENTS.

Name of Dataset	Number of features	Number of samples	Reference
SONAR	60	208	Sonar Dataset from UCI Machine Learning Repository[18]
IONOSPHERE	34	351	Ionosphere Dataset from UCI Machine Learning Repository[18]
SPAMBASE	57	4601	spambase Dataset from UCI Machine Learning Repository[20]
LUNG	12533	180	Lung Cancer Data Set [18].
PROSTATE	15154	89	Prostate Cancer Data Set [18].
OVARIAN_1	15154	116	Ovarian Cancer Data Set [18].
OVARIAN_2	15154	116	Ovarian Cancer Data Set [18].
LEUKAEMIA	7129	73	Leukemia Molecular Data Set [18].

B. Evaluation Metrics

As pointed out by many authors, the performance of a classification process over imbalanced data sets should not be expressed in terms of the plain accuracy or error rates [14, 15, 16]. The use of these simple measures might produce misleading conclusions since they do not take into account misclassification costs, are strongly biased to favor the majority class, and are sensitive to class skews [18].

Some measures have been proposed to evaluate classifiers in imbalanced scenarios. Two well known examples are Receiver Operating Characteristic (ROC) curve, the area under the ROC curve (AUC) [17], and the f -measure [15]. All these measures are combinations of error/accuracy rates measured separately on each class. [18]. in this paper we use F1 measure which equally weights precision and recall and AUC measure to evaluate our method while using linear SVM as a classifier. For more information about the AUC measures, interested reader can refer to [8].

We calculate the F1 measure according to the formulate bellow. The notations tp, fp and fn are respectively stands for true positive, false positive and false negative.

$$recall = \frac{tp}{tp + fn} \quad (1)$$

$$pecision = \frac{tp}{tp + fp} \quad (1)$$

$$F1 = \frac{2 * percision * recall}{percision + recall} \quad (1)$$

C. Analysis and Observation

Tables below compare different methods according to their AUC and F1 level on imbalance dataset.

We compare proposed method (FSSH) with other feature ranking approaches while using 2% of high ranked features. This number of features appears to be the point where the best performing feature selection metrics peak across each evaluation statistic. With more than 2% features selected, these metrics see a significant decline in performance. The goal of data mining is to make the best predictions possible; on high dimensional imbalanced data sets, it appears that we only need to select 2% of features to attain this peak performance [9, 10,11].

Table 2 presents the performance of LSVM classifier across different data sets in term of AUC evaluation statistics.

In all datasets except IONOSPHERE and LEUKAEMIA proposed method can improve the classification performance of LSVM in the term of AUC more than other mentioned methods. In IONOSPHERE and LEUKAEMIA datasets FAST[9] method performs better than the proposed approach while selecting more features. It has been shown that on IONOSPHERE data set, FAST method select 14 features to reach its maximum performance(0.868) while proposed method selects 7 features and reaches to the performance of 0.865 in the term of AUC evaluation statistics.

On LEUKAEMIA data set, FAST method needs to select 52 features in order to reach to its maximum performance of 0.814, while proposed method select 28 features to peak its maximum performance which is 0.81 in term of AUC measure.

Table 2 : The performance of LSVM classifier across different data set in term of AUC evaluation statistics. The Number of selected features is given in brackets.

The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.89 e-ISSN: 2251-7545

DATA SET	FSSH	FAST	IG	PCC	S2N
SONAR	**0.782 (14)**	0.750 (16)	0.710 (4)	0.780 (19)	0.760 (34)
IONOSPHERE	0.865 (7)	**0.868 (14)**	0.850 (4)	0.860 (10)	0.850 (34)
SPAMBASE	**0.940 (12)**	0.880 (12)	0.830 (16)	0.845 (10)	0.831 (22)
LUNG	**0.832 (19)**	0.831 (26)	0.810 (51)	0.790 (52)	0.782 (10)
PROSTATE	**0.820 (20)**	0.797 (32)	0.780 (34)	0.798 (67)	0.781 (9)
OVARIAN_1	**0.843 (25)**	0.818 (28)	0.790 (43)	0.832 (48)	0.772 (14)
OVARIAN_2	**0.832 (21)**	0.820 (24)	0.770 (40)	0.796 (39)	0.760 (15)
LEUKAEMIA	0.810 (28)	**0.814 (52)**	0.760 (46)	0.812 (45)	0.780 (9)

Table below compare the performance of LSVM classifier across different data sets. It has been shown that in all data sets except IONOSPHERE proposed method performs better than other methods. In IONOSPHERE data set, FAST method performs better than others in the term of F1 evaluation statistics while selecting more features than proposed approach.

Table 3: The performance of LSVM classifier across different data set in term of F1 evaluation statistics. The Number of selected features is given in brackets.

DATA SET	FSSH	FAST	IG	PCC	S2N
SONAR	**0.800 (15)**	0.770 (35)	0.761 (21)	0.760 (39)	0.757 (60)
IONOSPHERE	0.910 (13)	**0.922 (15)**	0.910 (30)	0.920 (9)	0.910 (34)
SPAMBASE	**0.920 (9)**	0.860 (17)	0.910 (14)	0.880 (19)	0.890 (22)
LUNG	**0.920 (19)**	0.900 (25)	0.890 (29)	0.900 (26)	0.890 (52)
PROSTATE	0.920 (24)	**0.930 (32)**	0.880 (42)	0.890 (43)	0.880 (67)
OVARIAN_1	**0.910 (17)**	0.890 (29)	0.880 (47)	0.900 (40)	0.880 (48)
OVARIAN_2	**0.915 (21)**	0.900 (34)	0.870 (45)	0.910 (39)	0.890 (48)
LEUKAEMIA	**0.920 (23)**	0.910 (37)	0.890 (38)	0.880 (40)	0.900 (52)

It should be mentioned that reducing the training and testing time is very important in data mining filed. It has been shown that in comparison with other methods, the proposed method achieves to its maximum performance level while selecting very small number of features which is really a great advantage for this method.

IV. CONCLUSION

In this paper, we propose a feature selection approach for imbalance datasets. The algorithm FSSH (feature Selection based on Shapely value) is proposed and compared with some state of the art approaches which proposed for imbalance data sets. Evaluation on different data sets shows that proposed method is a great candidate for feature selection in most applications, especially when selecting very small numbers of features. Feature Assessment by Sliding Thresholds (FAST) are consistently close to the best average performance across each of the evaluation statistics. Both these metrics would be a good choice for use on an arbitrary imbalanced data set but FSSH is the most efficient method.

V. FUTURE WORK

The performance of the classifier usually decreases while selecting number of features. This problem refers to some irrelevant and noisy features. In our future work we try to use the game theoretic concepts to present a method for detecting irrelevant and noisy features and selecting the most important ones.

The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.89 e-ISSN: 2251-7545

REFRENCES

[1] M. Lindenbaum, S. Markovitch, D. Rusakov, "Selective sampling for nearest neighbor classifiers", Machine learning, vol. 54, pp. 125– 152, 2004.

[2] A.I. Schein, L.H. Ungar, "Active learning for logistic regression: an evaluation", Machine Learning, vol. 68, pp. 235–265, 2007.

[3] S. Cohen, G. Dror, and E. Ruppin, "Feature selection via coalitional game", Theory Neural Computation, vol. 19, pp. 1939–1961, 2007.

[4] N. Chawla, N. Japkowicz and A. Kolcz, Eds, Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets, 2003.

[5] L. S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, Contributions to the Theory of Games, volume II of Annals of Mathematics Studies 28, pages 307–317. Princeton University Press, Princeton, 1953.

[6] A. Keinan, B. Sandbank, C. C. Hilgetag, I. Meilijson, and E. Ruppin, "Fair attribution of functional contribution in artificial and biological networks," *Neural Computuation*, vol. 16, no. 9, pp. 1887–1915, 2004.

[7] S. M. Hazrati, A. Hamzeh, S.Hashemi, "A Game Theoretic Framework for Feature Selection," Fuzzy System and Knowledge Discovery, Chongqing, FSKD 2012, in press.

[8] X. Chen and M. Wasikowski, "FAST: A ROC-Based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems," Proc. ACM SIGKDD '08, pp. 124-133, 2008.

[9] M. Wasikowski and X. Chen, "Combating the small sample class imbalance problem using feature selection", IEEE Transactions on knowledge and data engineering, 2009.

[10] G. Forman, "An extensive empirical study of feature selection metrics for text classification," Journal of Machine Learning Research, vol. 3, pp. 1289–1305, 2003.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol. 286, pp. 531–537, 1999.

[12] T.M. Cover, "The Best Two Independent Measurements Are Not the Two Best," IEEE Trans. Systems, Man, and Cybernetics, vol. 4, pp. 116-117, 1974.

[13] T. Cover and J. Thomas, Elements of Information Theory. New York: Wiley, 1991.

[14] Chawla, N. V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Exploration Newsletters, vol. 6, pp.1–6, 2006.

[15] Daskalaki, S., Kopanas, I., Avouris, N.: Evaluation of classifiers for an uneven class distribu-tion problem. Applied Artificial Intelligence ,vol.20 pp.381–417, 2006.

[16] Elazmeh, W. and Japkowicz, N., Matwin, S.: Evaluating misclassifications in imbalanced data. In: Proc. 17th European Conference on Machine Learning pp.126–137, 2006.

[17] Bradley, P. W.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, vol 30 ,pp.1145–1159, 1997.

[18] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. BloomÞeld,E.S. Lander. "Molecular ClassiÞcation of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". SCIENCE vol. 286, pp. 531-537, 1999