International Journal of Pure and Applied Mathematics
Volume 117 No. 20 2017, 873-879
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

PARALLEL SELECTIVE SAMPLING USING RELEVANCE VECTOR MACHINE FOR IMBALANCE DATA

M. Athitya Kumaraguru[1], Viji Vinod[2], N. Rajkumar[3]
[1] Research Scholar, Department of Computer Applications, Dr. M.G.R. Educational and Research Institute University Maduravoyal, Chennai, Tamil Nadu 600095
[2] Professor, Department of Computer Applications, Dr. M.G.R. Educational and Research Institute University Maduravoyal, Chennai, Tamil Nadu 600095
[3] Research Executive, Accendere Knowledge Management Services Pvt. Ltd New Delhi, Delhi 110044
athityaguru@yahoo.co.in[1], vijivino@gmail.com[2], rajkumar.mnm@gmail.com[3]

Abstract

Many areas where large data sets are used there comes the problem to determine the outcome from that dataset. The primary cause of this problem is the size of the data i.e., very large size data the second cause is the imbalanced class. There are many numbers of classification algorithm and many applications to overcome this big data problem but there are extreme limitations, in the data set. Many algorithm and application have been used to resolve this problem in the literature. A major preprocess technique parallel selective sampling method along with the classification algorithm Relevance Vector Machine. This PSS will choose data from majority class and reduce the imbalance in the big data. A major technique is used in the literature is PSS-SVM proposed by D'Addabbo et al., 2015. But the existing algorithm is failed to produce result with many probabilities. To over this problem in this paper we propose a method which combines preprocessing technique and a classification algorithm Parallel Selective Sampling with Relevance Vector Machine (PSS-RVM). This PSS-RVM will produce a finest result then the existing algorithm in the literature.

Keywords PSS, SVM, RVM, RUSBOOST, UNDERSAMPLING, OVERSAMPLING

Introduction

Many classifications real-time applications have to find the rare event occur in the large data's. The area were very huge data set have to be handled, the number of training examples will be very huge and the classes are strongly imbalanced and minority class. There are two problematic issues first one computational complexity depends on the size of the data, the second one is high rate of correction detection in the minority class.

There are many classification algorithms with performance degradation and several limitations due to class imbalance. In the existing classification technique, it is hard to deal with many domains and the computational cost is too high due to very huge data set. Other than tis the classification process is very difficult due the strong imbalance and the complexity of the training is purely dependent on the size of the data. However, some of the classification algorithms are sensitive on the class imbalance due to that drop in the performance in the minority class. Due to this imbalance between majority and minority class there is an alter in the class boundary. There are some other classification schemes where this limitation is common ex: Multi-layer perception (MLP) and Logistic Regression (LR).

To overcome this problem there is procedure which is well known in the literature is called "Undersampling" method. This method will perform sampling a small number of samples from the majority class and reduce both the number of data and imbalance. This method generally improves the performance of the classification and computational complexity will be reduced (Mohd. Faheem Khan et al., 2011). This is the potential disadvantage of the altering the distribution of the majority class. It will reduce the performance of the classification if the sampled pattern of the majority class does not represent the original distribution. This drawback comes true when the class pattern is very small in the majority class. So, other technique for the large data set because: (1) by modifying cost for misclassified patterns belonging to the minority class, without changing the number of original data (L. Bruzzone et al., 1997), (2) by increasing the total number of examples by copying patterns from the minority class to balance the ratio of classes (" oversampling" method) (N. Chawla et al.,2002), (3) by combining oversampling and undersampling techniques (Q. Wang et al., 2014).

There are several techniques in the literature to select the samples: the example-selection is merge with the learning algorithm or the examples are filtered before they passing the scheme of classification. These methods generally work by keeping the original ratio between the classes. To overcome this problem the pre-processing of the data before the classification step. In this framework, a very interesting method has been developed by Evgeniou and Pontil in (T. Evgeniou et al., 2002). Based on the Euclidean distance method the preprocessing algorithm

computes cluster of points in each class and center point of the cluster is substitute in each cluster. These methods did not focus on huge and imbalance data. Recently a new method for big and imbalanced data classification has been proposed. It is called as cost sensitive support vector machine using randomized dual coordinate descent method which belongs to class embedded method. In this method, both learning algorithm and example are embedded and the classifier is depended. This method is examined using data which is huge and strongly class imbalance. The rate of minority class is computed by CSVM-RDCD.

A valuable alternative method is proposed in the literature which is classifier independent and it will adjust only the distribution of the original training set. The method is called as Parallel Selective Sampling (PSS) which is a filtering method which can be combine with Relevance Vector Machine (Tipping. M. E 2000). This Relevance Vector Machine will select only relevant data from majority class and used in classification process by RVM (Bishop. C.M et al., 2000). This PSS is designed for parallel and Distributed computing due to computational complexity. This PSS-RVM will predict accurate statistical result.

Approaches

Parallel Selective Sampling

The PSS method can be used to preprocess very large training data with significant skew between classes. The examples have to be reduced only from the majority class a method called undersampling is used. The Tomek link characterize as a pair of nearest neighbors of opposite classes (GowrishankarKasilingam et al., 2015). It is based on the computation of Tomek links (I. Tomek et at., 1976), defined as a pair of nearest neighbors of opposite classes. Given {E1, . . ., En} ∈ Rk, a pair (Ei, Ej) is called a Tomek link if Ei and Ej have different labels, and there is not an El such that d (Ei, El) < d (Ei, Ej) or d (Ej, El) < d (Ei, Ej), where d (·, ·) is the Euclidean distance. Here Tomek links are used to remove samples of majority class staying in areas of input space dense of data belonging to the same class (FariborzParandin et al., 2015).

Let S = {(x1, y1), . . ., (x, y)} be the training set, where xi ∈ Rk and yi ∈ {0, 1}, ∀i = 1, . . ., We define S0 the set of 0 training data be-longing to class y = 0 and S1 the set of 1 training data belonging to class y = 1, with 0 1. PSS achieves a reduced training set whose percentage M% of the minority class on the total number of examples is chosen by the user (D'Addabbo et al., 2015).

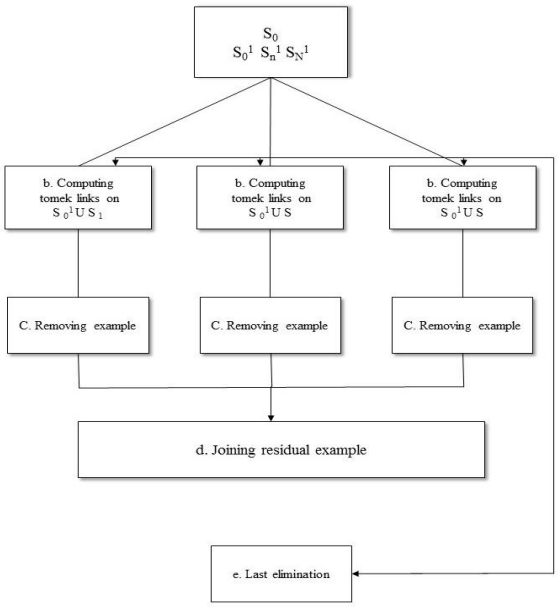

Fig.1: Block diagram of PSS

Data Partitioning

TheS_0set is divided into N subsetS_0^nwithn = 1, 2, . . ., N, with N set by the user. In this way, N different undersampling procedures are performed in parallel computation (see Fig. 1).

For each S_0^n, with n = 1, 2, . . ., N, the following steps are per-formed:

Computing Tomek links

Let us define the set L^n of all examples in the majority class $S0^n$ that are first neighbors of one sample in S1, that is L^n = {x ∈ S0n| (x, z) is Tomek link on S1 ∪ $S0^n$, z ∈ S1}.

Removing examples

Let us randomly select$\bar{x} \in D^n = S_0^n \backslash T^n$; thefollowing steps are performed

- The Tomek link ($\bar{x}$, $\bar{z}$) is computed over the data set $\bar{x} \cup S_1$, with $\bar{z} \in S_1$;
- The Euclidean distances d ($\bar{x}$, x) are computed for each x ∈ S_0^n;
- Let us define the subset L ={x ∈ S_0^n|d ($\bar{x}$, x)< d ($\bar{x}$, $\bar{z}$)}, (see the red circumference in The Tomek link (x^*, $\bar{z}$) in $\bar{z} \cup$ L is computed, i.e. x^* is defined as the first neighbor in L of $\bar{z}$;
- Let us define the set R ={x ∈ L|d ($\bar{x}$, x)<(d ($\bar{x}$, $\bar{z}$)– d (x^*, $\bar{z}$))} Let us delete all the points in R that are not Tomek links, i.e. each x ∈ R with R ={x ∈ R|x ∈/ T^n}. The remaining data points of the majority class are contained in S_0^n= $S_0^n \backslash R$;

- If the classes are balanced, the algorithm goes to the following step d; otherwise it randomly selects $x^{-} \in D^{n} = S_0^{n} \backslash T^{n}$ and repeats the previous issues (step c).

Joining Residual Examples

The majority class data, selected by each parallel computation, are then joined.

last elimination

The procedure previously described (step c) is repeated achieving a final reduced training set whose M% belongs to the minority class.

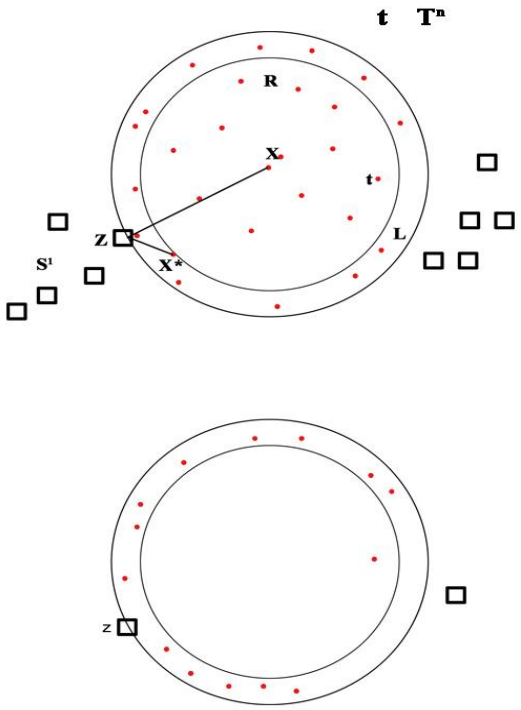

Fig. 2. Removing examples step (c) of PSS

Relevance Vector Machine

The new classification technique which belong to Bayesian model of generalized linear model named as Relevance vector machine(RVM) (Tipping. M 2001). This RVM overcome many disadvantages of existing classification algorithm like SVM, RUSBoost. Some disadvantages of SVM is prediction are not probabilistic, make liberal use of kernel functions, kernel function must satisfy Mercer's condition, cross-validation procedure (Rafi. M et al., 2013). The key function of RVM are non-linear probabilistic model, distribution of weight that gives solution sparse, can produce fewer decision than SVM with accurate result, don't need any parameter during the training session, don't need any kernel fulfill the mercer's condition.

Data set	Training set size	% of minority data	No. of attributes
A1	13580	7.1%	6
A2	14550	1%	6
A3	20855	0.5%	6

Table 1: Summary of training data set

Data Description

In this study, we use three real data set have been fetched from activity recognition system data set belongs to UCI repository having 42240 instances and 6 classes. We extract two classes from this data set and data is divided as training data and test data.

- A1: Bending vs Cycling (6305 vs 7275 training set; 3153 vs 3638 test sets)
- A2: Cycling vs Walking (7275 vs 7275 training set; 3638 vs 3638 test sets)
- A3: Bending vs all (6305 vs 14550 training set; 3153 vs 7275 test sets)

Evaluation of experimental

The Result of an objective function is lacking for classification tasks with hard data imbalancing classes. For example, let us consider a classification problem in the classes, minority class as 5% patterns and majority class as 95% patterns. If a naive approach of classifying is done all patterns will be classified into the majority class, it would achieve 99% of accuracy.

Consider dice D, precision P, recall R and relative overlap R.O

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$D = \frac{2 * P * R}{P + R} = \frac{2 * TP}{(TP + FP) + (TP + FN)}$$

$$R.O. = \frac{FP}{TP + FP + FN}$$

- TP: True Positive (actual positive data which are correctly classified)
- FP: False Positive (negative data classified as positive)
- TN: True Negative (the actual negative data which are correctly classified)
- FN: False Negative (positives incorrectly classified as negatives)

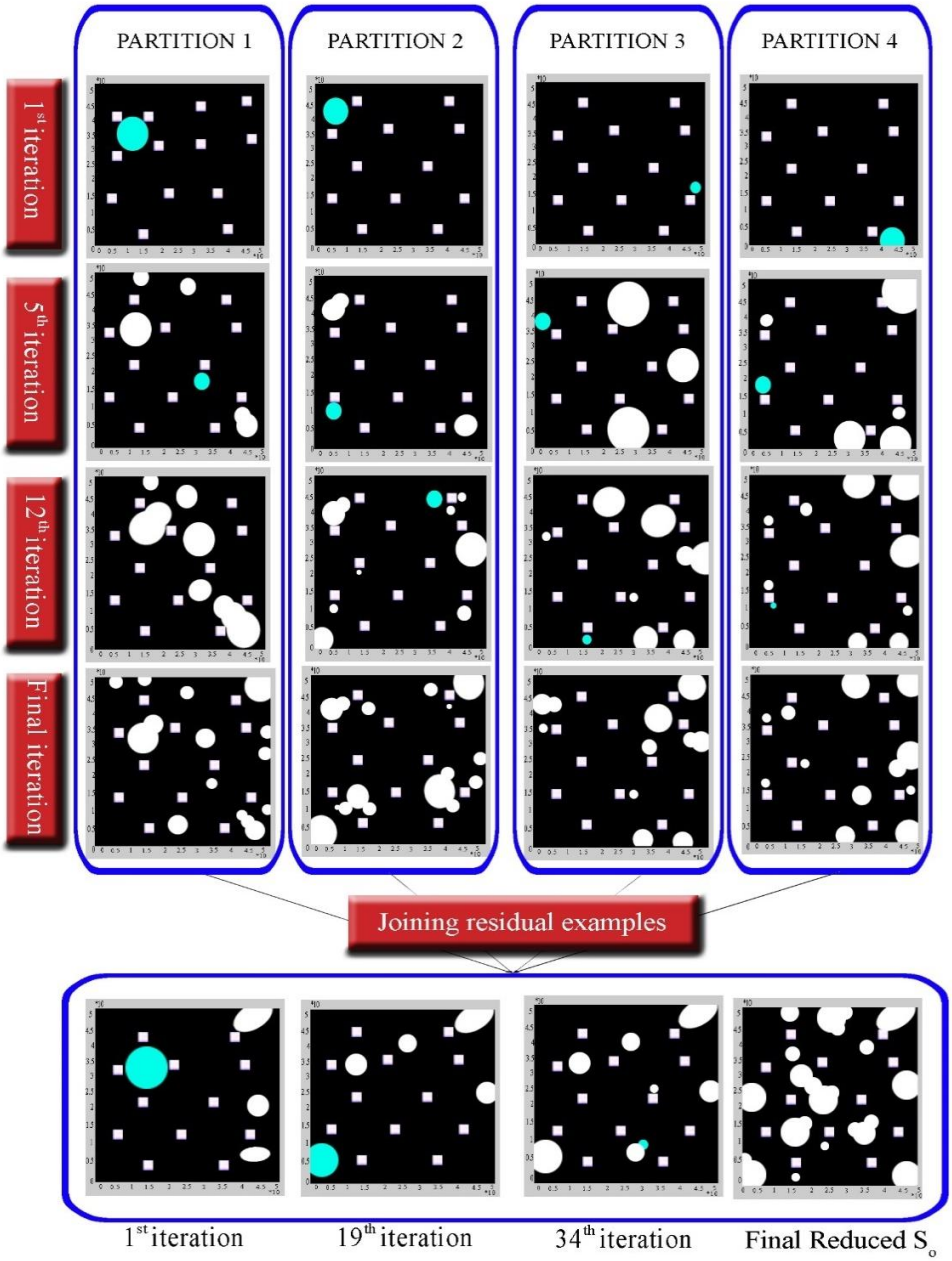

Fig.2: Undersampling of the Real data set A1 by PSS. After the partition of S1 in four subsets, the 1st, 5th, 12th and final iterations are shown. The residual examples of the four subsets are joined in the first picture (1st iteration) of last row. Then, the 19th and 34th iterations of PSS to the joined residual examples and the final reduced majority class (S0) are shown.

The D value is used to merge precision and recall into a single metric for convenience. The R.O. accounts for the fraction of TP on the total number of true and predicted positive examples.

Result

In this paper, we compared the performance of PSS-RVM with RUSBoost. This RUSBoost is a Boosting-based sampling algorithm that handles class imbalance randomly removing examples from the majority class until the balance is achieved.

The data set A1 contained 13580 examples and 7.1% is majority class. This data set is small in size and less imbalance. So, the performance of the both the algorithm is good.

The data set A2 contained 1% majority class and 14550 training set and PSS-RVM had a better and short computation.

The data set A3 contained 0.5% majority class and 20855. The processing time is 152s by PSS-RVM and showed greater performance than the RUSBoost.

In this analysis of 3 data set with huge amount of data's and strongly class imbalance. The minority class is reduced to improve the performance of evaluation. In this case RUSBoost does not show any better performance on this 3-huge data set. The advantage of implementing this PSS-RVM is its distinct performance in large & class imbalanced data's.

A1	RUSBoost n.trees=1000 n.leaf=5	PSS-RVM M = 15%, RBF-σ= 0.8 C = 10
Time (s)	962	45
Dice	90.7	90.8
Precision	89.9	90.7
Recall	91.5	90.8
Relative overlap	82.9	83.1
A2	RUSBoost n.trees=100 n.leaf=5	PSS-RVM M = 10%, RBF-σ= 2 C = 10
Time (s)	38	81
Dice	99.4	99.7
Precision	99.2	99.8
Recall	99.6	99.7
Relative overlap	98.8	99.7
A3	RUSBoost n.trees=1500 n.leaf=5	PSS-RVM M = 15%, RBF-σ= 1 C = 10
Time (s)	1984	152
Dice	87.8	87.5
Precision	83.4	84.4
Recall	92.9	91.0
Relative overlap	78.3	77.9

Table 2: Summary of the experimental results on the synthetic data sets: a) evaluation metrics computed on test set and computational time of PSS-SVM required for both preprocessing and training, b) mean values (standard deviation) of evaluation metrics computed on test set and computational time on 10 iterations of random Undersampling SVM. Optimal parameters (SVM kernel, regularization parameter C and desired percentage M) are shown.

Conclusion

In this paper, the preprocessing technique PSS is combined with RVM. The comparison between PSS-RVM and RUSBoost is carried on the three datasets. This PSS-RVM showed excellent performance than RUSBoost. Our analysis suggested that this PSS-RVM will be valued alternative for RUSBoost and showed a greater performance on imbalanced data. PSS presented a great advantage to perform in parallel computing and reduce the computational time.

References

1. L. Bruzzone, S. Serpico, Classification of imbalanced remote-sensing data by neural networks, Pattern Recognit. Lett. 18 (11) (1997) 1323–1328.
2. Q. Wang, A hybrid sampling SVM approach to imbalanced data classification, Ab-str. Appl. Anal. 2014 (2014), doi:10.1155/2014/972786. Article ID 972786, 7 pages.
3. N. Chawla, L. Hall, W. Kegelmeyer, Smote: synthetic minority oversampling techniques, J. Artif. Intell. Res. 16 (2002) 321–357.
4. T. Evgeniou, M. Pontil, Support vector machines with clustering for training with very large datasets, in: I. Vlahavas, C. Spyropoulos (Eds.), Methods and Applications of Artificial Intelligence, Springer-Verlag, Berlin, 2002, pp. 346–354.
5. I. Tomek, Two modifications of cnn, IEEE Trans. Syst. Man Cybernet. 6 (11) (1976) 769–772.
6. D'Addabbo. A, Maglietta. R, "Parallel selective sampling method for imbalanced and large data classification", Pattern Recognition Letters, 2015, (62), pp.61–67.
7. Rafi. M, Shaikh. M.S, "A comparison of SVM and RVM for Document Classification", Procedia Computer Science, 2013, (00), pp.3–8.
8. Tipping. M, "Sparse Bayesian Learning and the Relevance Vector Mach", Journal of Machine Learning Research, 2001, (1), pp.211–244.
9. Bishop. C.M, Tipping. M.E, "Variational relevance vector machines", Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, 2000, (1), pp.46–53.
10. Tipping. M. E, "The Relevance Vector Machine", Advances in Neural Information Processing Systems, (2000), (12), pp 652–658.

11. GowrishankarKasilingam, JagadeeshPasupuleti, “Coordination of PSS and PID Controller for Power System Stability Enhancement – Overview”, Indian Journal of Science and Technology, 2015 Jan, 8(2), Doi no:10.17485/ijst/2015/v8i2/58441.
12. FariborzParandin, Ali Mohammadi, HosainSariri, “Adaptive Multi Machine PSS Design for Low Frequency Oscillations Damping”,Indian Journal of Science and Technology,2012 Dec, 5(12), Doi no:10.17485/ijst/2012/v5i12/30609.
13. Mohd. Faheem Khan, Gaurav Chauhan, A. K. Jaitly, “An Approach to overcome Imbalance Datasets of Eukaryotic Genomes during the Analysis by Machine Learning Technique (SVM)”, Indian Journal of Science and Technology,2011 May, 4(5), Doi no:10.17485/ijst/2011/v4i5/30053.