International Journal of Research in Advent Technology, Vol.3, No.12, December 2015
E-ISSN: 2321-9637
Available online at www.ijrat.org

A study on the Class Imbalance classification using Fuzzy Total margin based Support vector Machine

S.Lavanya[1], Dr.S.Palaniswami[2], R.Premalatha[3]
Assistant Professor, Department of CSE, Anna University Regional Campus, Coimbatore, India[1]
Principal, Government college of Engineering, Bodinayakanur, Tamilnadu, India[2]
PG Scholar, Department of CSE, Anna University Regional Campus, Coimbatore, India[3]

Abstract-The classification of imbalanced data is a difficult challenge for machine learning with help parameter selection and classification techniques. Employment of the traditional classifiers like SVM will leads to the overfitting to the class. Fuzzy total margin based support vector machine (FTM-SVM) way to handle the class imbalance learning (CIL) is analysed to identify the outliers in the feature vector, incorporates total margin algorithm, different cost functions and the proper method of fuzzification of the penalty into FTM-SVM and defines them in nonlinear case. Comparison with state-of-the-art data stream classification techniques creates the effectiveness of the proposed approach. And we conclude the classification of imbalanced data is a analysed in detail with machine learning principles, feature selection and classification techniques. Employment of the traditional classifiers like SVM has leads to the overfitting to the class. Fuzzy total margin based support vector machine (FTM-SVM) method to handle the class imbalance learning (CIL) is analysed to identify the outliers in the feature vector, incorporates total margin algorithm, different cost functions and the proper approach of fuzzification of the penalty into FTM-SVM and formulates them in nonlinear case. Comparison with state-of-the-art data stream classification techniques establishes the effectiveness of the proposed approach.

Index Terms- outlier detection, Classification, Class Imbalance Data, Data Classification

1 INTRODUCTION

Class imbalance learning (CIL) is an emerging topic that is attracting growing attention. It aims to tackle the combined concern of online learning [1] and class imbalance learning [2]. Different from incremental learning that processes data in clutch, online learning here means study from data examples "one-by-one" without storing and reprocessing observed examples. Class imbalance learning handles a type of classification problems where some classes of data are densely under represented in relation to other classes. With these above two problems, connected class imbalance learning deals with data streams where data appear progressively and the class distribution is imbalanced. Although online learning and class imbalance learning have been well learned in the literature individually, the combined problem has not been discussed much. When both problems of online learning and class imbalance exist, new challenges and interesting research questions arise, with regards to the assumed accuracy on the minority class and adaptively to dynamic environments. The difficulty of learning from imbalanced data is caused due to or absolutely underrepresented class that cannot draw equal consideration to the learning algorithm related to the majority class. It often leads to very specific classification rules or lost rules for the minority class without much observation ability for future prediction. Total margin-based adaptive fuzzy support vector machines (TAF-SVM) [4] and fuzzy support vector machines for class imbalance learning (FSVM-CIL) [3] are emerging. TAF-SVM not only mitigate the overfitting problem influenced by the outliers and noise with the approach of fuzzification of the damages, but also corrects the change of the optimal separating hyperplane(OSH) owing to the imbalanced data sets by using different cost functions. The total margin algorithm, fuzzy membership functions and different cost functions are attached in the traditional SVM so that the TAF-SVM is reformulated in both linear and nonlinear cases. In the FSVM-CIL, fuzzy participation values for training examples are authorized to reduce the effect of both the problems of CIL and the disputes of outliers and noise under the principle of cost sensitive learning. FSVM-CIL can be used to handle CIL problem in the existence of outliers and noise.. The proposed method incorporates total margin algorithm, different cost functions and the fitting approach of fuzzification of the penalty. FTM-SVM introduces variety of cost functions and the applicable fuzzy membership functions. Therefore, FTM-SVM can be used to mitigate the problem of CIL better than some existing CIL learning methods. FTM-SVM introduces fuzzy membership functions with strong methods, this is not sensitive to outliers

and noise and can compromise with this overfitting problem when the data sets contain some outliers and noise examples. FTM-SVM has good universality ability because it introduces the total margin algorithm to replace traditional soft margin algorithm.So we considered six forms of fuzzy-membership values and got six FTM-SVM settings. We calculated the FTM-SVM method on two artificial data sets and imbalanced data sets and compared its performance with existing CIL methods. Experimental results show that the proposed FTM-SVM method has higher F Measure and precision and recall values than some existing CIL methods. The Rest of the paper is organized as follows In Section 2, some related work is reviewed, In section 3, Outline of the Work is analysed. Section 4 Provides review of literatures. In Section 5, paper is concluded.

2. RELATED WORKS

2.1. Diversity in Classification Ensembles

Diversity of ensembles has been a hot topic during the past few years. It is frequently agreed that the success of ensemble is applied to diversity, the degree of disagreement within an ensemble. In the regression text, it has before been quantified and measured explicitly in terms of the correlation between individual learners. In the classification context, it is loosely described as “making errors on different examples”.

2. 2. Ensemble Methods

Ensemble learning methods have become a major category of solutions for class imbalance learning, due to their flexibility and ability to improve generalization. First, an ensemble method is applicable to most classification algorithms. Second, it’s easy to fix with resampling techniques. Third, combining multiple classifiers is able to reduce the error bias/variance [5]. These attractive features lead to a variety of ensemble methods proposed to handle imbalanced data sets from the data and algorithm levels.

2.3. Synthetic Minority Oversampling Technique (SMOTE)

The algorithm which occupies the minority class feature space vitally placing synthetic examples on the line segment connecting two minority instances. SMOTE has been shown to improve the classification accuracy on the minority class over other standard approaches.

OUTLINE OF THE SURVAY

2.4 Data labelling and Preprocessing

Slack variables are popularized. Slack variables can be delibrated to be treated as counts of outliers and noise in the data sets because they can grant the margin constraints to be destroyed and can cause each training points to have smaller or even negative margin. Thus, the method that takes account of the slack variables is applied to handle the data that have outliers and noise. The noise data pre-processed using the stop word removal and stemming process, in order to reduce data size prior to clustering and Classification..

2.5 Feature Selection

The target of feature selection, in general, is to select a subset of features that allows a classifier to reach optimal performance, where j is a user-specified parameter. The curse of dimensionality tells us that if many of the characteristics are noisy, and the cost of using a classifier can be very high, and the performance may be severely hindered, so feature selection algorithm traced from mutual information measure and the entropy computation is characterized using the mutual information measure for identification of a suitable feature subset with less redundancy.

2.6.Data Classification using SupportVector Machine and Fuzzy Total Margin

We analyze sophisticated SVM; the maximal margin is an important concept. In order to optimize the maximal margin, by maximizing the minimum distance between support vectors and the separating hyper plane, Optimal Separating Hyper plane with OSH is seen with maximal mar-gin of separation. However, only few support vectors could cause the loss of information because most information is contained in the majority of the pre-processed training set. Therefore, it can improve the generalization error bound. The surplus variables measure the distance in between the correctly classified data points and the hyper plane respectively. In addition to minimizing the sum of slack variables during maximizing the margin of separation proposed by soft margin algorithm, Fuzzy membership values is used as training examples are assigned to reduce the effect of both the problems of CIL and the problems of outliers and noise under the principle based on cost sensitive learning

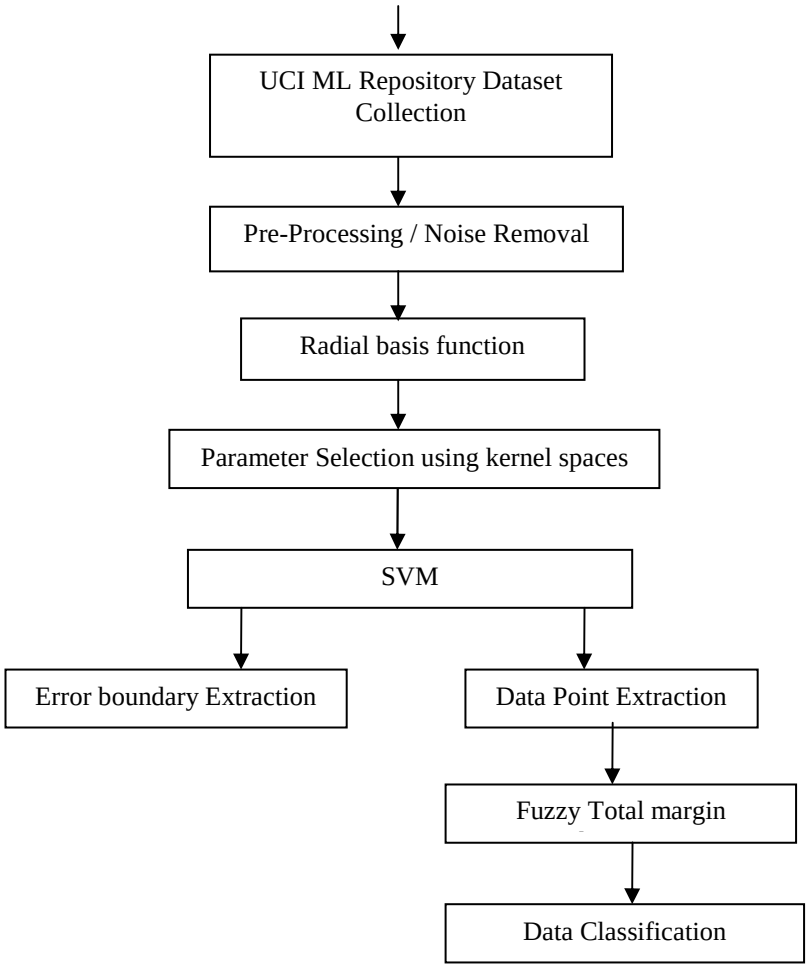
3.Equations

The FSVM method assigns different fuzzy membership values or weights, for different examples to reflect different importance in their own classes.In the proposed FTM-SVM method,we defined these membership functions as follows:

$$\mu_i^+ = f(x_i)r^+,\ \forall x_i \in S_f^+, \quad \mu_i^- = f(x_i)r^-,\ \forall x_i \in S_f^-$$

where f(xi) generates a value between 0 and 1, and f(xi) reflectsthe importance of xiin its own class. We assigned r+= 1 and r−= r,where r is the minority-to-majority class ratio. Therefore, according to this assignment of values, a positive-class example can takea membership value in the [0, 1] interval, negative-classexample can take a membership value in the [0, r], where r < 1.

4.1 Flow Diagram:



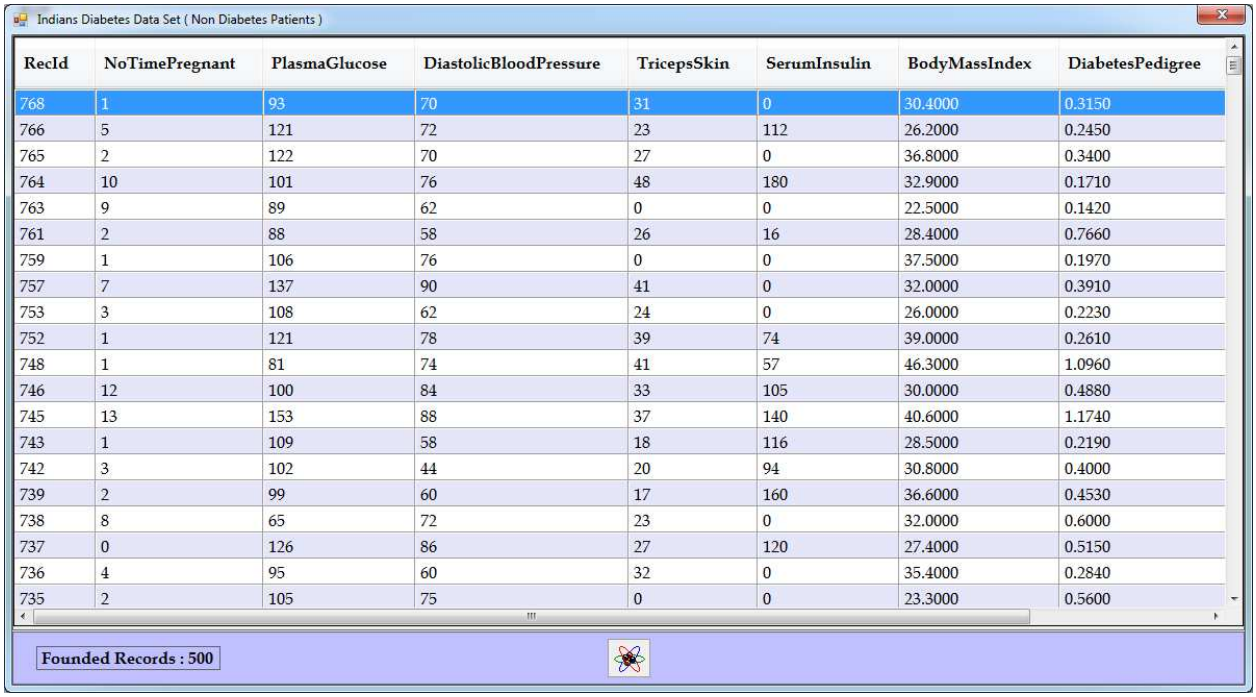
4.2 TABLE

Indians Diabetes Data Set (Diabetes Patients)

RecId	NoTimePregnant	PlasmaGlucose	DiastolicBloodPressure	TricepsSkin	SerumInsulin	BodyMassIndex	DiabetesPedigree
767	1	126	60	0	0	30.1000	0.3490
762	9	170	74	31	0	44.0000	0.4030
760	6	190	92	0	0	35.5000	0.2780
758	0	123	72	0	0	36.3000	0.2580
756	1	128	88	39	110	36.5000	1.0570
755	8	154	78	32	0	32.4000	0.4430
754	0	181	88	44	510	43.3000	0.2220
751	4	136	70	0	0	31.2000	1.1820
750	6	162	62	0	0	24.3000	0.1780
749	3	187	70	22	200	36.4000	0.4080
747	1	147	94	41	0	49.3000	0.3580
744	9	140	94	0	0	32.7000	0.7340
741	11	120	80	37	150	42.3000	0.7850
740	1	102	74	0	0	39.5000	0.2930
733	2	174	88	37	120	44.5000	0.6460
732	8	120	86	0	0	28.4000	0.2590
731	3	130	78	23	79	28.4000	0.3230
723	1	149	68	29	127	29.3000	0.3490
720	5	97	76	27	0	35.6000	0.3780
717	3	173	78	39	185	33.8000	0.9700

Founded Records : 268

4.3 TABLE



RecId	NoTimePregnant	PlasmaGlucose	DiastolicBloodPressure	TricepsSkin	SerumInsulin	BodyMassIndex	DiabetesPedigree
768	1	93	70	31	0	30.4000	0.3150
766	5	121	72	23	112	26.2000	0.2450
765	2	122	70	27	0	36.8000	0.3400
764	10	101	76	48	180	32.9000	0.1710
763	9	89	62	0	0	22.5000	0.1420
761	2	88	58	26	16	28.4000	0.7660
759	1	106	76	0	0	37.5000	0.1970
757	7	137	90	41	0	32.0000	0.3910
753	3	108	62	24	0	26.0000	0.2230
752	1	121	78	39	74	39.0000	0.2610
748	1	81	74	41	57	46.3000	1.0960
746	12	100	84	33	105	30.0000	0.4880
745	13	153	88	37	140	40.6000	1.1740
743	1	109	58	18	116	28.5000	0.2190
742	3	102	44	20	94	30.8000	0.4000
739	2	99	60	17	160	36.6000	0.4530
738	8	65	72	23	0	32.0000	0.6000
737	0	126	86	27	120	27.4000	0.5150
736	4	95	60	32	0	35.4000	0.2840
735	2	105	75	0	0	23.3000	0.5600

REFERENCE

[1] L. L. Minku, "Online ensemble learning in the presence of concept drifts," Ph.D. dissertation, School Comput. Sci., Univ. Birmingham, Birmingham, U.K., 2010.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[3] R. Batuwita, V. Palade, FSVM-CIL: fuzzy support vector machines for classimbalance learning, IEEE Trans. Fuzzy Syst. 18 (June (3)) (2010) 558–571.

[4] Y. Liu, Y. Chen, Face recognition using total margin-based adaptive fuzzysupport vector machines, IEEE Trans. Neural Netw. 18 (January (1)) (2007)178–192.

[5]. Y. Sun, M.S. Kamel, A.K. Wong, and Y. Wang, "Cost-Sensitive Boosting for Classification of Imbalanced Data," Pattern Recognition, vol. 40, no. 12, pp. 3358-3378, 2007. [5] H. Xue, S. Chen, Q. Yang, Structural regularized support vector machine: aframework for structural large margin classifier, IEEE Trans. Neural Netw. 22(April (4)) (2011) 573–587.

[6] R. Batuwita, V. Palade, FSVM-CIL: fuzzy support vector machines for classimbalance learning, IEEE Trans. Fuzzy Syst. 18 (June (3)) (2010) 558–571.

[7]X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalancelearning, IEEE Trans Syst. Man Cybern. B: Cybern. 39 (2) (April 2009) 539–550.

[8] R. Batuwita, V. Palade, Efficient resampling methods for training support vectormachines with imbalanced datasets, in: Proc. of IJCNN 2010, 2010 IEEE World Congress on Comp. Intelligence-WCCI 2010, Barcelona-Spain, July 2010.