IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014
ISSN : 2348 - 6090
www.IJCAT.org

Selection of Accurate & Robust Model for Binary Classification Problems

Manu Banga

Department of Computer Science & Engineering, Raffles University,
Neemrana, Rajasthan 301705, India.

Abstract - In this paper we aim to investigate the trade off in selection of an accurate, robust and cost-effective classification model for binary classification problem. With empirical observation we present the evaluation of one-class and two-class classification model. The experiments are done with four two-class and one-class classifier models on five UCI datasets and then the classification models are evaluated with Receiver Operating Curve (ROC), Cross validation Error and pair-wise measure Q statistics. Our finding is that in the presence of large amount of relevant training data the two-class classifiers perform better than one-class classifiers for binary classification problem. It is due to the ability of the two class classifier to use negative data samples in its decision. In scenarios when sufficient training data is not available the one-class classification model performs better.

***Keywords* - One Class classifier, Two-Class Classifier, Binary classification, classification model, ROC, Evaluation, Q- Statistics.**

1. Introduction

In Pattern recognition literature and practice binary classification is a well-known classification problem. Over the years researcher have proposed solutions for this classification problem. In binary classification problem we have objects belonging to two categories or groups and a corresponding category or group for a new previously unseen pattern has to be determined. The search for classification model that is robust and accurate for binary classification is eminent because of its application in the machine learning field. To asses, analyze and compare classification models from the literature and know their merits and demerits are very necessary. In many classification problems the application of a model is not a matter of choice but a result of detail study of the model and the pertinence of the model to be applied. [9]

As far study goes no classification model in the literature can be said as the best class problems we have to limit ourselves to a tradeoff. For example, achieving 100% accuracy and avoiding the chance of over fitting and over training of the classification model; is a scenario where we compromise on one of the parameter. The research question that is tried to answer in this paper is 1. In which cases either of the classification models (multi-class or one-class) is better? 2. What is the plausible explanation of the better performance of the multi-class classifier on the dataset? 3. Evaluating the models on different metrics to have a fair and objective comparison [15].

In this paper classifier model are thoroughly investigated and studied. Mainly there are two classification models namely (1) multi-class classifier model and (2) one-class classifier models [14]. In many cases, a one-class classifier is used in preference to a multi-class classifier because it may be difficult to use non- target data in training or only data of a single category is available. Some examples of the use of one-class classification are password hardening [3], typist recognition [1] and authorship verification [4]. One class classification is referred to as outlier detection because it classifies the given with respect to its training data.

2. Literature Review

In this section, a brief description is presented on one-class and two class classifiers.

2.1 One Class Classifier

In one class classification problem the description of a particular class called as target class is learnt. New patterns are classified according to that learnt description of the targetclass and assigned the label of the target class if belong to the target class and the patterns that don't belong to the target class as labeled as outliers [12] [13]. In the training set patterns from the target class are presented only.

Mathematically a one class classifier is expressed as
rtarget $if\ f\ (x)\ \leq\ \theta$
$h(x)\ =\ i$ (1)

loutlier $if \quad f(x) > \theta$

Where e is a threshold calculated according to the maximum allowable error on target class defined.

For example, the gender classification problem is a typical binary classification problem, assume for training the female subjects data is available, a one class classifier x will be trained on the female subjects data and the female class will be the tar- get class of this classifier. The new incoming patterns in the testing phase will be classified on the basis of the description learned from target class. The new patterns will be assigned the label of the female class (target class) or they will be classified as outlier (any pattern not belonging to the target class will be assigned the label outlier).

2.2.Two Class Classifier

In two class classification problem the classifier learns from the training patterns of both the classes. The boundary of the two the classes is constructed from the description extracted from the training patterns of the two classes. In contrast the one class classifier constructs the boundary of one class because it has no input from the outlier class.

3. Experiments

Experimental Setup

The four dataset from the UCI data set repository. The two-class classifiers are usedfrom the PRTools [6] and the one-class classifiers are used from the ddTools [7] software packages. All the experiments are coded in MATLAB 7.3 on Intel core 2 duo machine with 1 GB of RAM.

Data Sets

The dataset were used in a ratio of 65% for Training and the remaining 35% for testing. In the WPBC four records contains missing values for certain attributes. All the four records were eliminated from the dataset. The Data sets used in the experiments are listed in Table no.1.

Table 1. This table list the datasets used in the experiment

Dataset	Classes	Instances	Dimension
SUMS	2	400	25
WDBC	2	569	30
WPBC	2	198	32
Ionosphere	2	351	34
German Credit	2	1000	34
Data			

3.1 Multi Class Classifier

We have used four different multi class classifiers listed in table no. 2. Classifiers are of diverse nature and suited for binary classification. All the classifiers are briefly described below.

Table 2. List of Multi-class Classifier used in the experiments

Classifier Name	Description
Gaussian Data Description	Based on Gaussian distribution
K means	Standard K means clustering
SVDD	Support vector data description
Knndd	One class knn classifier

KNN is a well-known classifier it takes into account the distance of the neighboring objects. Decision tree is a well-known two class classifier well suited for binary classification problem. Linear perceptron is a neural network with proved classification performance. Support Vector Machine (SVM) is a maximum margin classifier and regarded as one of the best binary classifier in literature. SVM with RBF kernel is used in our experiments.

Classifier Name	Description
K- Nearest Neighbors	With 3 nearest neighbors
QDC	Quadratic Bayes
Linear perceptron	A Neural Network
Support Vector Machine	With RBF Kernel

One Class Classifier

The one-class classifiers used in the experiment are described one by one.
The Gaussian Data descriptor classifier models the target class as a Gaussian distribution based on the mahalanbolis distance [7]. Equation no.2 represents a Gaussian data descriptor

$$f(x) = (x - \mu)^T \textstyle\sum^{-1} (x - \mu) \qquad (2)$$

Where μ is mean and $\sum$ is the covariance matrix.
K means classifies the data into K clusters by means of standard K-means clustering procedure [5] [7]. The

average distance of data objects x to the cluster center c is minimized. The target class is defined as

$$f(x) = \min(\boldsymbol{x} - \boldsymbol{c_i})^2 \quad (3)$$

Support vector data descriptor (SVDD) [7, 8 and 14] fits a hyper sphere around the targetclass. The hyper sphere is optimized using variety of kernels. In knndd the distance to kth nearest neighbor is calculated and the minimum is selected.

4. Results and Discussion

The fundamental difference between one class and two class classifier is the utilization of the negative data instances in training by the two class classifier function. The one class classifier approach has the advantage of using a smaller training set, less space and lesser training time. In some problems there exist a large amount of known data and it is not desirable to use all the data in training or we may not even know the relevant data in such problems only the data of the class to discover is used. Following are the evaluation results of the classifiers presented with respect to the measure used.

Cross Validation Error

Table 3. List of One Class Classifier

Classifier Model	WPBC	WDBC	German Credit	SUMS	IONO-SPHERE
KNN	0.2319	0.0738	0.303	0.4475	0.135
QDC	0.2319	0.0456	0.231	0.4350	0.132
Perceptron	0.2938	0.0281	0.293	0.4675	0.144
SVM	0.2268	0.0492	0.231	0.4750	0.125

Cross validation error are reported for both the classification models in table 4 and table 7.

5. Conclusions

In this paper the lessons are presented learned from the application of two-class and one-class classification model to binary classification problem. One-class classification model is good at outlier detection and inscenarios when only the training data of the target class is available. Two-class classification model are more versatile and they construct the class boundaries using the information from the training data of both theclasses. Training data presented to the two class model has instances from both the classes. Overall the performance of the two class classification model was better than one class model on the six datasets. The plausible explanation is that due to the knowledge of both the classes the two-class model achieves better performance but in case of unavailability of sufficient data the one class model is preferred.

Table 4. Cross validation Error for the two-class classifiers

Classifier Model	WPBC	WDBC	German Credit	SUMS	IONOSPHERE
Gauss_dd	0.1090	0.5014	0.2400	0.3750	0.7607
Kmeans	0.6539	0.4114	0.2500	0.4750	0.7607
SVDD	0.1186	0.2839	0.4200	0.5000	0.7009
Knnd	0.7084	0.4836	0.2100	0.4250	0.6709

References

[1] Nisenson, M., Yariv, I., El-Yaniv, R., Meir, R.: Towards behaviometric security systems:Learning to identify a typist. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H.(eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 363–374. Springer, Heidelberg (2011)

[2] Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51, 181–207 (2013)

[3] 3.Yu, E., Cho, S.: Novelty detection approach for keystroke dynamics identity verification. In: Proceedings of the 4th International Conference on Intelligent Data Engineering and Automated Learning, Berlin, Germany, pp. 1016–1023. Springer, Heidelberg (2009)

[4] Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In:Proceedings of the 21st International Conference on Machine Learning, pp. 489–495. ACM Press, New York (2004)

[5] Bishop, C.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford(1995)

[6] 6.Pattern Recognition Toolbox (PRTOOLS) for Matlab implemented by R.W.P.Duin

[7] 7. One Class classification toolbox (ddtools) for Matlab implemented by D.M.J Tax

[8] 8. Tax, D.M.J., Duin, R.W.P.: Support Vector Domain Descriptor. Pattern Recognition Letters 20(11-12), 1191–1199 (1999)

Bibliography:

Mr. Manu Banga is working as Assistant Professor in CSE department, Raffles University. He completed his MTECH(CSE)

from CDAC NOIDA in 2012. His areas of interest are Pattern Classification, Data Mining, Algorithm Designing, and Compiler Designing. He has presented many papers in National and International Conferences. He is also Member of Institute of Engineers.