

Multiclass Classification and Support Vector Machine

Yashima Ahuja¹

¹ Lovely Professional University

Received: 12 December 2011 Accepted: 5 January 2012 Published: 15 January 2012

Abstract

In this paper we have studied the concept and need of Multiclass classification in scientific research. Various classification approaches are discussed in brief. Support Vector Machines (SVM) has well known record in Binary Classification. Our major emphasis in this paper is to study the fitness of Support Vector Machines in multiclass classification.

Index terms—

1 Introduction

Multiclass classification is a major requirement in field of science and engineering because multiclass discrimination of objects is a serious problem in science and engineering. Multiclass classification is always considered complex than binary classification. In binary classification, only the decision boundaries of 1 class are to be known and rest (complement of first class) is considered as second class where as in multiclass classification, several boundaries are essential for that reason. This may lead to increase the probability of error because of constructions of many decision boundaries. Various classification methods like decision tree, KNN (knearest) are widely used for multiclass classification.

In this paper, our goal is to investigate the problem of multiclass classification and to propose an efficient method for the purpose. Support vector machines are highly accurate and able to model complex non-linear decision boundaries. The Major limitations of SVM are time complexity is very high. It leads to low-speed solutions and Svm are well known for binary classification.

To address the problem of multiclass classification, researchers have used 3 methods-1. One vs one 2. One vs all 3. Directed acyclic graph After analyzing the literature, we have found that DAG-SVM is considered to be best for multiclass classification.

2 II.

3 Motivation

Multiclass problem can be expressed and solved as a single optimization problem, the following models proposed in the works by Suykens and Author : Research Scholar, Lovely Professional University, Jalandhar (Punjab) India. E-mail : er.yashima@yahoo.com Author : Assistant Professor, Lovely Professional University, Jalandhar (Punjab) India. Suykens and Vandewalle proposed a least square multi-support vector machine (LS-MSVM) for multi-category problems that can be interpreted as regularized least squares. It leads to a straightforward and rapid algorithm for generating linear or non-linear classifiers, obtained by solving a single linear system of equations. The approach also has alike interpretations as the reduced kernel multi-class machine. The drawback with LS-SVM is that sparseness is lost i.e. most of support vector becomes non-zero [13].

Szedmak et al. proposed a multiclass model for large sample sizes and number of features. In their formulation, the OAA framework is used, minimizing the L₁ norm of the normal vector w of the separating hyper planes. Their formulation solves k SVM optimization problems simultaneously and since there are no interactions between the variables of the k SVM problems, their method-which essentially is the OAA but considers all data at once produces the same solution as the separated k SVM problems with the L_1 norm in the objective function. Because their formulation considers all data at once, creating a large scale optimization problem, the size of the problem can be drawback if it gets too large [14].

In addition, Oladunni & Trafalis projected a regularized kernel multi-classification model by means of a reduced kernel approach. The resulting optimization problem solves for $(\frac{1}{2} - \frac{1}{2k})$ classifiers concurrently, and as there are no relations between the variables of the $(\frac{1}{2} - \frac{1}{2k})$

classification problems, the method can be considered a pair wise multiclass classification method that considers all data at one time. The formulation produces the similar solution as that of the independently solved $(\frac{1}{2} - \frac{1}{2k})$ binary classification problems. The method works well, however just like the SVM variant of Szedmak et al, the size of the problem can be a drawback. Consideration of all the data at once transforms the problem into a large scale optimization problem which can be very expensive to compute. Also, sparseness is lost [15].

The objective of this study is aimed towards the formulation of a piecewise kernel based multi-class model. The resulting optimization problem is linear programming problem with a linear cost function and linear constraints in dual space induced by a kernel function.

III.

5 Support vector machine

Machine Learning is ability to enable the computer to learn. It uses algorithm and techniques which perform different tasks and activities to provide efficient learning [23]. Our main problem is that how can we represent complex patterns and how to exclude bogus patterns. Support Vector Machine is a machine learning tool used for classification and regression. Support Vector Machine is based on supervised learning which classifies points to one of two disjoint half-spaces [2]. It uses nonlinear mapping to convert the original data into higher dimension. Its objective is to construct a function which will correctly predict the class to which the new point belongs and the old points belong. With an appropriate nonlinear mapping, two data sets can always be divided by hyperplane. Hyperplane separates the tuples of one class from another and defines decision boundary. There are many hyper planes that separate the data but only one will achieve maximum separation. The main reason behind maximum margin or separation because if we use a decision boundary to classify, it may end up nearer to one set of datasets compared to others [12]. This was the case when data is linear but mostly we find data is non-linear and data set is inseparable then we use kernels.

The core purpose of SVM is to separate the data with decision boundary and extends it to non-linear boundaries using kernel trick [12]. Major benefit of Svm is versatile means different Kernel functions can be specified for the decision function. General kernels are provided, but it is also possible to specify custom kernels. SVM becomes prominent when we used pixel maps as input; it gives accuracy equivalent to neural networks with elaborated features in a handwriting recognition task. Support vector machine is used for many applications such as text categorization, pattern recognition, face recognition, handwriting analysis but especially for classification and regression applications. Neural Networks are easier to apply than support vector machine but sometimes it provides insufficient results. Even the perceptron learning algorithms (e.g. gradient descent) are slower than SVM learning. Svm has been found to be unbeaten when used for pattern classification problems. One of the major challenge is that of choosing a suitable kernel for given application [12]. But there are many standard or default choices such as Gaussian or polynomial kernel but if these prove worthless then more elaborate kernels are needed.

Traditional Classification approaches perform weakly when working directly because of high dimensionality of data but support vector machine can avoid the pitfalls of very high dimensionality representations. Support vector machine is the most promising technique and approach as compared to others. Support vector machine scales fairly well to high dimensional data and the trade-off between classifier complexity and error can be controlled explicitly. Another benefit of SVMs and kernel methods is that one can design and use a kernel for a particular problem that could be applied directly to the data without the need for a feature extraction process. It is particularly important in problems where a lot of structure of the data is lost by the feature extraction process. Example is text processing. Limitations of Svm are speed, size both in training and testing [28]. Discrete data presents another problem. Most severe difficulty with SVMs is the high algorithmic complexity and extensive memory requirements. In short we can say that the development of SVM is an utterly different from standard algorithms used for learning and SVM provides a fresh insight into this learning.

IV.

6 Binary classification using svm

For binary classification problems, the idea behind SVM is to split the data in finest method. Binary classification is used when we need to classify the two data sets. There are numerous examples of Binary classification like try-outs (one either makes or fails to make the team), claim size (large claims are above some threshold and small claims below), and fingerprint identification (matched or unmatched). Support vector machines are primarily designed for 2-class classification problems [2]. Support Vector Machine consider 2 approaches-1. Case when the data are linearly separable 2. Case when the data are non-linearly separable

7 NON-LINEAR AND LINEAR SEPARATION

8 Non-Linearly

Linearly separable

Let us consider first case; there are many linear decision boundaries that divide the data. But only one of these achieves maximum division. The main purpose we need it is because if we use a decision boundary to classify, it may end up nearer to one set of datasets compared to others and we do not want this to happen and thus concept of maximum margin classifier or hyper

9 G

plane as an apparent solution. Support vectors are the data points that lie closest to the decision surface. Support Vectors can be described as those data points that the margins pushes up against. They are the most difficult to classify [75]. The major problem here is to find the only optimal margin of the separating hyperplane $0 = +b \times W^T$

, the one that provides maximum margin between the classes. This margin guarantees the lowest rate of misclassification. The further advantage of margin would be avoiding local minima and better classification [12].
-class 1 -class 2

10 DIAGRAM SMALL AND LARGE MARGIN SEPARATION

To allow some flexibility in separating decision boundaries, SVM models have a cost parameter C that controls the tradeoff between allowing training errors and forcing strict margins. It creates a soft margin that permits some misclassifications. Now, in second case data are not linearly separable i.e. in such cases no straight line can be found that would divide the classes. Linear svm's can be extended to generate nonlinear SVM'S for classification of linearly inseparable data. Such svm are capable of finding nonlinear decision boundaries.

V .

11 Multiclass classification

SVMs were mainly proposed to deal with binary classification but in today's life, we mostly have huge amount of data which we want to classify. Time series data represent quantities or trace the values taken by a variable over a period such as a month, year etc. Examples are stock market, price indexing etc. In this there will be more than two classes. So this creates the need of multiclass classification. Multiclass classification means classification with more than two classes.

Before introducing SVM, we have different kinds of multiclass techniques [79]. Firstly we will distinguish it on the basis of direct and indirect approach (via binary). Direct Approaches includes k-nearest neighbor, decision tree and bayes classification, linear classifications like perceptron. Multiclass classifications through binary include One-vs-one and One-vs-all, Directed acyclic graph svm, Error correcting output codes.

12 ??(??, ??)

$= ??(?? \text{ ? ? ? ? }) \text{ } 2 \text{ ? ? } ??=1$

Decision Tree is a flowchart like structure where each internal node denotes a test on an attribute, every branch represents as a result of the test and each leaf node holds a class label. The topmost node in a tree is the root node. In classification, attribute values of the tuple are tested against the decision tree. Decision Trees can be easily converted into classification rules. Decision trees are popular because it doesn't require any domain knowledge, parameter setting and can handle multidimensional data with fast speed and good accuracy [1].

Bayes classification predicts class membership probabilities such as probability that a given tuple belongs to a particular class. It is based on bayes theorem. Bayes theorem provides a way of calculating posterior probability $(X) \times H \text{ P } / \text{ of } H \text{ conditioned on } X. () () () () X \text{ P } H \text{ P } H \text{ X } P \text{ X } H \text{ P } / / =$

Bayesian classifiers have the minimum error rate in comparison to all other classifiers but in practice this is not always the case sometimes inaccuracies in assumptions such as lack of available probability data [1]. Now let us consider the case Multiclass classification using Binary. In SVM, The idea of using a hyperplane to separate the data into two groups sounds well when there are only two target categories, but how does SVM handle case where the target variable has more than two categories or values? Numerous approaches have been suggested, but there are two most popular approaches described below [11].

In general, the most frequent method has been to construct one-versus-rest classifiers (usually referred to as "one-versus-all" or OVA classification) where each category is split out and all of the other categories are merged and to choose the class which classifies the test data with greatest margin. In validation phase, a pattern is presented to each one of the binary classifiers and then classifier which provides a positive output indicates the output class. In numerous cases, the positive outcome is not unique and some tie-breaking techniques are compulsory. The most familiar approach uses the confidence of the classifiers to decide the last outcome, predicting the class from the classifier with the maximum confidence. Rather than having a score matrix, when

dealing with the outcomes of OVA classifiers (where i, r in $[0, 1]$ is the confidence for class i) a score vector is used: In validation phase, a pattern is presented to each one of the binary classifiers. The output of a classifier given by ij, r in $[0, 1]$ is the confidence of the binary classifier discriminating classes i and j in favor of the former class. Class with the largest confidence is the output class of a classifier. These outputs are represented by a score matrix $R: R = \begin{bmatrix} ? & ? & r_{12} & ? & r_{1m} & r_{21} & ? & ? & r_{2m} & ? & ? & r_{m1} & r_{m2} & ? & ? & ? \end{bmatrix}$

There is so much difference between these two approaches. One-vs-all strategy constructs by fitting one classifier per class where as one-vs-one strategy constructs one classifier per pair of classes. Benefit of one-vs-all approach is its interpretability. This is the most commonly used strategy and is a fair default option. One-vs-one classifier doesn't scale well with n samples.

Directed Acyclic Graph can be used to represent a set of programs where the input, output, or execution of one or more programs is dependent on one or more other programs. The programs are nodes in the graph, and the edges (arcs) identify the dependencies [25]. A directed acyclic graph is a directed graph that contains no cycles.

13 VI.

14 Conclusion

In today's era, SVM are majorly used in binary class classification. In real time data, it is commonly found that data is to be classified in more than two classes. KNN, Decision Tree, Bayesian Classification are widely used for multiclass classification. KNN are lazy learners and hence perform slowly. Decision Trees are good for the purpose but highly complex if the depth of tree is on higher side. Bayesian Classifiers need good sample of data with high probable states, which is not always possible. In SVM, The idea of using a hyperplane to separate the data into two groups sounds well when there are only two target categories. For Multiclass Classification, One-vs-all strategy constructs by fitting one classifier per class where as one-vs-one strategy constructs one classifier per pair of classes. Benefit of one-vs-all approach is its interpretability. This is the most commonly used strategy and is a fair default option. One-vs-one classifier doesn't scale well with n samples.

15 Global Journal of Computer Science and Technology

Volume XII Issue XI Version I



Figure 1:

183 [Svm] , Wiki Svm . <http://www.dataminingtools.net/wiki/svm.php>

184 [Hsu and Lin (2002)] ‘A Comparison of Methods for Multiclass Support Vector Machines’. Chih-Wei Hsu , Chih-
185 Jen Lin . *IEEE Transactions on Neural Networks* March 2002. 13 (2) .

186 [Oladunni and Trafalis (2006)] ‘A pair wise reduced kernel-based multi-classification Tikhonov regularization
187 machine’. T B Oladunni , Trafalis . *Proceedings of the International Joint Conference on Neural Networks*
188 *(IJCNN '06)*, (the International Joint Conference on Neural Networks (IJCNN '06)Vancouver, BC, Canada)
189 July 2006. IEEE Press. p. .

190 [By Jiawei Han and Micheline Kamber (ed.)] *Data Mining Concepts and Techniques book*, By Jiawei Han and
191 Micheline Kamber (ed.)

192 [Szedmak et al. ()] ‘Multiclass classification by L1 norm support vector machine’. S Szedmak , J Shawe-Taylor
193 , C J Saunders , D R Hardoon . *Pattern Recognition and Machine Learning in Computer Vision Workshop*,
194 (Grenoble, France) May 02-04, 2004.

195 [Suykens and Vandewalle ()] ‘Multiclass least squares support vector machine classifiers’. J Ak Suykens , J
196 Vandewalle . *Proceedings. of Joint Coif. On Neural Networks (IJCNN'99)*, (of Joint Coif. On Neural Networks
197 (IJCNN'99)Washington, D.C.) 1999c.

198 [Olutayo et al. ()] ‘Piecewise Multi-Classification Support Vector Machines’. O Olutayo , Gaurav Oladunni ,
199 Singhal . *Proceedings of International Joint Conference on Neural Networks*, (International Joint Conference
200 on Neural NetworksAtlanta, Georgia, USA) June 14-19, 2009.

201 [Auria et al.] *Support Vector Machines (SVM) as a Technique for Solvency Analysis*, Laura Auria , A Rouslan ,
202 Moro .