COMPARISON OF CLASSIFICATION METHODS: PERIL TO AVOID FOR BINARY AND MULTI-CLASS CLASSIFIER AND PROPOSE COMBINATION APPROACH.

[1]Shivani Raina,[2]Salima Velshi,[3]Neha Wattamwar,[4]Snehal Yeole
[1234]Computer Department
[1234]Vishwakarma Institute of Information Technology
[1]shivaniraina26@gmail.com[2]velshisalima@gmail.com[3]neha.wattamwar190@gmail.com[4]snehal.yeole207@gmail.com

ABSTRACT:

Classification plays an important role in various fields like Object recognition, text categorization etc. Studying classifiers for purpose of estimating probability for a certain object to belong to a specific class is crucial for classification .In this paper, we present a survey of four classifiers: Support vector machine, k Nearest Neighbour, Naive Bayes and Neural Network focusing on their merits and demerits.We will also shed light on combination of the above mentioned classifiers and present, how they overcome drawbacks of individual classifiers.

KEYWORDS: *Support vector machine classifier; k Nearest Neighbour classifier; Naive Bayes classifier; Neural Network; SNNB.*

1. INTRODUCTION

Classification is a method of finding category, also known as class, as output of a particular input which may be in the form of category, integer value etc. Classifier is an algorithm that performs classification. There are many classification algorithm used in practice, but this paper discusses only four.

2. RELATED WORK

2.1. K- nearest neighbor

The K-nearest neighbour(KNN) classification is classification method that identifies the category of input according to its known nearest neighbour's category. During training phase a set of instances, also known as features are given and during testing phase category of input is determined.KNN is based on a distance function that measures the difference or similarity between two instances.

KNN is simple algorithm but has been used to solve complex data mining problems. The classifier is effective and robust. Its disadvantage is its computing complexity for large training sets. This algorithm is an example of lazy learning that stores training data at training time and delays its learning until testing time. Thisis based on a distance function that measures the difference or similarity between two instances, [5] thus defining the distance function is crucial which decides the efficiency of the system.

2.2. Naïve bayes

The Bayesian classifier that uses the Naïve Bayes assumption and finds class is called Naïve Bayes classifier. [1]It is one of the most practical learning methods. The idea behind a Bayesian classifier is that, if an agent knows the class, it can predict the values of the other features from the given class and if it does not know the class, Bayes' rule can be used to predict the class from some of the given feature values.

This classifier is easy to implement as it predicts the class on the basis of some given features. It uses Bayes theorem for classification. So training and classification are fast. It considers only those features that are important for particular class.Handles real and discrete data.[2]It assumes class conditional independence resulting in loss of accuracy. Practically, dependencies exist among variables and these cannot be modelled by Naïve Bayesian classifier.

2.3.Support vector machine

SVM recognizes various kinds of the patterns .SVM construct hyper plane or a set of the hyper planes for the classification in high-dimensional space. The OSH(Optimal separating hyper plane) algorithm includes finding a pair of parallel hyper planes that separates the data having the largest perpendicular distance between them. SVMs calculates OSH directly from the training data using the geometric properties. [4]SVM through non-linear mapping maps the input space (ie the 32x32 pixel image) into a high dimensional feature space and then constructs the OSH in the feature space. Hence linear decision surfaces can be constructed in a feature space which corresponds to non-linear.

SVM appear to perform superior performance for the two object classification.SVM's can handle only binary classification problem. More SV's are requires as the noise level increases.SVM performs better than BP(Back propagation) when the noise is added in the testing phase.SVM is helpful in text and hypertext categorization, in recognition of the hand written characters and in medical science for the classification of proteins.

2.4. Neural network (backpropagation)

In artificial neural network, various algorithms are present which are used to train the system. Thesetraining algorithms in neural network work on the idea that the weights of each unit should be adjusted in such a way that the error between the desired output and the actual output is reduced. This process requires that the neural network computes the error derivative of the weight (EW) i.e. how the error changes as each weight is increased or decreased slightly. The Back propagation algorithm is the most widely used method for determining the EW.

The Back propagation algorithm is used for layered feed-forward ANN. This means that the artificial neurons are organized in layers and the signals are send in the forward direction i.e. from input to output and then the errors are propagated backwards.The training begins with random weights and the goal is to adjust them so that the error will be minimal.

The Back propagationalgorithm is used in various fields like intrusion detection in computer networks.

3. COMPARISON

After examining varied articles, we find useful in giving a comparative study of different classifiers along with their advantages and disadvantages.

Table1. Comparison of various classifiers.

KNN	Idea	classifies instances based on their similarity to instances in the training data
	Advantage	Simple, effective, robust
	Disadvantage	Large computation time, lazy learner
NAÏVE BAYES	Idea	based on the assumption that information about

		classes in the form of prior probabilities and distributions of patterns in the class are known
	Advantage	Easy to implement, provides fast training and classification.
	Disadvantage	Assumes conditional independence
SVM	Idea	Decompose M class problem into series of two class problem and construct several binary classifiers
	Advantage	superior performance for the two-object classification, efficient, reliable, less time complexity
	Disadvantage	Decrement in accuracy with increment in noise variance, can handle only binary classification.
NEURAL NETWORK (BACK-PROPAGATION)	Idea	The main idea of theBack propagation algorithm is to reduce the errors, until artificial neural network learns the training data.
	Advantages	able to form arbitrary complex nonlinear mapping, Widely used for training feedback networks and some recurrent networks.
	Disadvantage	There is an inability to know how to precisely generate any arbitrary mapping procedure, It is hard to know how many neurons and layers are necessary.

4. COMBINATION OF CLASSIFIERS

4.1. Selective neighborhood naïve bayes (SNNB)

KNN classifier mainly depends on value of k. An algorithm that gives best value of k for KNN is a must to be designed. To achieve this goal, the most direct method is try various k values and choose the best one. In SNNB, given a test instance,[6] various k values are tested. For each k value, a local naive Bayes is learned from the k nearest neighbors and is evaluated. After this, the mostaccurate naive Bayesian classifier is used to classify the test instance. Although SNNB demonstrates good classification performance, it has very high time complexity. The process searchingfor the best k value is very time-consuming. Thus naïve bayes is combined with KNN to determine the size of neighbourhood for finding k nearest neighbors for each test instance.

4.2. Combination of svm classifiers

A fuzzy logic system (FLS) is constructed for combining multiple SVM classifiers showing the performance of each individual classifier. Genetic algorithms (GAs) tune the fuzzy logic system for generating the optimal fuzzy logic system. In phase I, training data are trained on different SVMs.

Validation data are classified to obtain individual SVM AUCs and distances of validation data examples to SVM hyperplanes. In phase II, a GFS is constructed and fuzzy MFs are tuned by GAs in cross validation manner. Finally, in phase III, testing data are fed into the optimal fuzzy fusion system to make the final decision. In the fusion system that includes the combination of the three SVM classifiers, there are three

AUC inputs showing three SVM classifier performances, three distance inputs representing the classification results of a data example from three individual SVM classifiers, and one output indicating the final decision from the fusion system for the example. Individual SVMs are combined in a genetic fuzzy system and then genetic algorithms are applied for tuning the fuzzy MFs based on AUC measure. The experimental results show that the proposed genetic fuzzy system is more stable and more robust than individual SVMs. The combined SVM classifier from the genetic fuzzy fusion model is more accurate and it can also be used in medical science.

6. CONCLUSION

Thus we make survey on varied classifiers, highlighting the pros and cons of each. The combinations of classifiers discussed aims to reduce the training set to gain on speed and space efficiencies. We conclude by not proving one classifier as best, but as discussed in the paper, each one of them is effective in specific area and in special circumstances but combination of classifiers gives better result as compared to individual.

7. REFERENCES

[1]Raykar, V.C, Yu, S.Zhao,L.H.Valadez,G.H,Florin,C.Bogoni,L.Moy,Learning from Crowds, Journal of Machine Learning Research 2010.

[2]G.Luger, Artificial Intelligence:structures and strategies of complex problem solving, 6THEd.Addison Wesley 2009.

[3]B. Smith, R. Gosine ,Support Vector Machines for Object Recognition,

[4]Tomasz Malisiewicz ,Abhinav Gupta, Alexei A. Efros, Ensemble of Exemplar-SVMs for Object Detection and Beyond, Carnegie Mellon University.

[5]MILOUD-AOUIDATE Amal, BABA-ALI Ahmed Riadh, Survey of Nearest Neighbor Condensing Techniques, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 11, 2011.

[6]Liangxiao Jiang, ZhihuaCai, Dianhong Wang, Siwei Jiang, Survey of Improving K-Nearest-Neighbor for Classification, China University of Geosciences.