An iterative algorithm based on Fuzzy Support Vector Machine Classifier

Saeed Khodayi, Mehdi Fatan
Faculty of Computer & Electrical Engineering, Qazvin Islamic Azad University

Abstract
Regarding progresses in data collection and storage capabilities in recent decades, high- dimensional dataset is quickly increasing in different disciplines. Most of these datasets have many features and relatively less patterns. Most of these features are mostly unrelated and redundant which leads to reduction of classification algorithms performance. Therefore, feature selection is proposed for reducing the dimensions of problem and increasing the efficiency of classification algorithms. In this paper, a new method is presented for improving the function of data classification. In proposed method, Fuzzy Support Vector Machine classifier algorithms is used for removing unrelated features in data. The performance of proposed method is compared with the newest and most known methods based on support vector machine classifier. The results of tests showed that the proposed method has good performance due to classification accuracy.

Keywords:
data analysis, feature selection, classification, support vector machine

1. Introduction

Data analysis copes with huge amount of data, their processing and analysis as artificial intelligence connector, machine learning, statistics and database. Data analysis aims at knowledge extraction from dataset and turning it to a comprehensible structure for later use. In other words, data analysis, analyzes and explores huge datasets in knowledge discovery and extraction. An important point on data analyses applications such as patter recognition is high-dimensional dataset in which the number of features is much more than the number of patterns. For example, in classifying bioinformatic data which includes high-dimensional datasets, the classification parameters are increasing too. Therefore, classification performance dramatically decreases.

Pattern recognition is a branch of machine learning. It can be said that pattern recognition is raw data gain and making decision based on data classification. Most of researches on pattern recognition is related to supervised learning or non-supervised learning. Pattern recognition methods separate desired patterns of a dataset using previous knowledge on patterns or statistics of data. The patterns which are classified with this method are groups of measurements or observations which form certain areas in a multidimensional space. This feature is the main difference of pattern recognition and pattern matching in which pattern are recognized based on a specified pattern using completely precise and certain cases. Pattern recognition and pattern matching are main parts of image processing especially in machine vision.

High-dimensional dataset reduces the classification performance in two sides. In one hand, the calculation's volume is increased upon increasing data dimensions and on the other hand, the model which is made with high-dimensional data has low Interoperability and overfitting is increasing. Therefore, reducing the dimensions of problem leads to reducing the computational complexity and improves the performance of classification algorithms [3-1].

Two solutions are presented for dimensionality reduction: feature extraction and feature selection. In feature extraction, the primary spaces of features are mapped to a smaller space. In fact, in this solution, less features are created upon combining available features, such that these features have all (or most part of) available information in primary features. On the other hand, in feature selection, a subset of primary features is selected. Feature selection is considered as an important and useful technique in data preprocessing which increases the speed of machine learning algorithms and improves classifier performance. Feature selection has been recognized as an important and active research area in pattern recognition, machine learning and data analysis since 1970's and has been used widely in many areas such as text classification [6-4], face recognition [9-7], image recovery [11,10[, medical diagnosis [13,12], and in financial works [15,14].

In continuing, second part reviews different classifiers while part 3 deals with total concepts of SVM classifier and describing proposed method. In part 4, the proposed method is evaluated and finally part 5 is summing up and conclusion of paper.

2. Review on different classifiers

Pattern recognition and classification is one of the important applications of statistical methods in different disciplines. One of the main goals of modeling and classification in statistics is predicting based on realities

Manuscript received July 5, 2017
Manuscript revised July 20, 2017

and variables and available information about special subject. This is mainly done by methods such as regression, discriminate analysis, temporal series, classification, regression tree in statistics. Four cases of classifications which have most usages are as follow:

- Support vector machine
- Naive Bayes
- ID3 decision tree
- Nearest neighbor

Support vector machine is a supervised learning method which is used for classification and regression. This method is relatively new method indicated good efficiency in recent years compared to previous methods in classifying such as Perceptron Neural Network. Primary SVM algorithm was invented by Vladimir Vapnik in 1963 which was generalized in 1995 by Vapnik and Corinna Cortes for nonlinear mode.

Bayes classifier computes the entry probability per class $P(X|C_i)$ and obtains the posterior probability $P(C_i|X)$ using prior probability $P(C_i)$ and Bayes' rule.

The advantage of this method is that it works good with numerical data and with text data that means it is widely used in text classification. The basic problems of this method is improper estimation, high volume and complicated calculations for estimating required probabilities. Naïve Bayes method solves this problem upon assuming independent events while in real world, it is not true in most of the cases and it will not work if the dependency of events is high while it has acceptable efficiency in text classification.

Decision tree is a tree in which the samples are classified such that they grow downwards from the root and finally they reach on leaf nodes. The features of this tree are as follow:

- Any internal or non-leaf node is specified with a feature. This feature raised a question about input problem.
- In any internal node, there are branches for the possible answers to this question, each of which is specified with that answer.
- The leaves of this tree are specified with a class or group of answers.

The reason of their nominating with decision tree is that this tree shows the decision process for determining the class of an input example.

One of the other useful classifications of algorithm is nearest neighbor. The idea is that the points which are in neighborhood of a point are probably in one class. The advantage of this method is easy designing and implementing process of classification no need to learning process, less parameters and their high efficiency. That means it determines the input class well in most of the cases though this method has many computations and therefore it is very slow.

One of the methods of increasing the function of classification is reducing the dimensions of dataset. From among the most well-known methods of feature selection, we can mention information gain (IG), [16], gain rate (GR) [17], gain index (GI), Fisher score (FS) and minimal-redudancy-maximal-relevance (mRMR) [18].

3. Proposed Method based on Fuzzy Support Vector Machine

Support vector machine is one of the supervised learning methods which is used for classification and regression. This method is relatively new method indicated good efficiency in recent years compared to previous methods in classifying such as Perceptron Neural Network. SVM classification is based on linear classification of data and in linear division of data, it is tried to select a line which has more safety margin.

Finding optimal line for data is solved via QP methods which are known methods in solving restricted problems. Before linear division, we carry data to high-dimensional space via phi function so that the machine could classify data with high complexity. Lagrange duality theorems is used for solving high dimensional problems such that desired minimization problem is turned to its duality form where a simpler function named as core function which is vector of phi function instead of phi complex function which carries us to high dimensional space. We can use different core functions such as exponential, polynomial and sigmoid cores.

We have D educational dataset including n members which is defined as equation 1:

$$D = \{(X_i, Y_i) | X_i \in R^p, Y_i \in \{-1,1\}\} \qquad (1)$$

Where Y is 1 or -1 and each Xi is a P-dimensional real vector. The goal is finding separating hyperplane with most distance from margin points which separates points with Yi=1 from points with Yi=-1. Each hyperplane can be written as a collection of X points which meets equation 2:

$$W.x - b = 0 \qquad (2)$$

In equation 2, "." Is multiply, W is normal vector which is perpendicular to hyperplane. We want to select W and b such that it creates the most distance between parallel hyperplanes which separates data from each other. These hyperplanes are described via equation 3.

$$W.x - b = 1 \tag{3}$$

$$W.x - b = -1$$

If educational data are linear detachable, we could consider two hyperplanes in points margin such that they have no common point and then try to maximize their distance. The distance between these two planes is $\frac{2}{||W||}$ using geometry. So, we should minimize $||W||$. To prevent the entry of points to margin, we add the following conditions: for each i

$$W.x_i - b >= 1 \qquad \text{For } x_i \text{ of the first class} \tag{4}$$

$$W.x_i - b <= -1 \qquad \text{For } x_i \text{ of the second class}$$

It can be written as equation 5:

$$Y_i(W.x_i - b) >= -1 \qquad \text{for all } 1<= i <=n \tag{5}$$

Having put them next to each other an optimization problem is obtained as equation 6.

Minimize |W|| Subject to for any i=1,….,n $Y_i(W.x_i) >= 1$ (6)

support vector machine estimates super-linear based on graphs such that samples can be classified via these super-linear. The super-linear which is estimated by support vector machine is Kernel function. Kernel function can be linear and polynomial. When we use support vector machine, three problems should be considered:

- Selection of a proper Kernel function
- Selection of an optimal subset of features
- Adjusting proper parameters for Kernel function

SVM is mainly binary separation. In previous part, the theoretic basis of support vector machine were described for classifying two classes. A multi-class pattern recognition can be achieved via combining two-class support vector machines. There are usually two views for this end. One of them is the strategy of “one for all” for classifying each pair of class and remaining classes. The other strategy is “one for one” for classification of each pair. For multi-class problems, the general solution is reducing multi-class problem to several binary problems. Each of problems are solved with a binary separator. Then the output of SVM binary separator are combined with each other and therefore multi-class problem is solved. Due to raised idea for SVM classifier, later Fuzzy support vector machine (FSVM) classification algorithm was presented for improving this method. The initial plan of FSVM was only presented for two-class classification problems while later on this initial plan was developed and it was used for classification of multi-class problems. The results of studies showed that generally FSVM method is superior that SVM method.

In proposed method, for feature selection via FSVM classifier, it is done as follow: in each repetition, subset of selected feature is studied and according to error rate, it is tried to change the feature subset such that classification error become minimized. In continuing, pseudo code of the proposed method is shown for selection of optimal subset. In fact, in each repetition of algorithm via equation 7, it is tried to change the selected feature subset such that the classification error reduces.

$$\mathrm{a}^{(k+1)} = \mathrm{a}rgmin \frac{1}{2(\lambda_2+\rho)} \mathrm{a}^T YXX^T Y\mathrm{a} + (\frac{YX(\rho u^{a(k)} - \mu^{(k)}}{\lambda_2+\rho} - 1)^T \mathrm{a} \tag{7}$$

4. Evaluating proposed method

In this part, the performance of proposed method is evaluated in feature selection problem. In this regard, the proposed method is compared with one of the newest methods of feature selection based on SVM named as ADMM-DrSVM [19]. In this paper, several dataset with different features were used for evaluating proposed method and comparing its performance with other methods of feature selection. These datasets include WDBC, Sonar, Arrththmia and Colon. The general details of these datasets are shown in table 1.

Table 1: details of used dataset

dataset	features	classes	Patterns
WDBC	30	2	569
Sonar	60	2	208
Arrththmia	279	16	452
Colon	2000	2	62

For studying the efficiency of proposed methods on different datasets, some tests were done on them. In the tests on proposed method, the datasets were randomly divided into educational data, validation data and tentative data. For this end, 50% of datasets were considered as educational data, 25% as validation data and remaining 25% as tentative data. Meanwhile, in all tests, having specified the educational, tentative and validation sets, each feature selection method is repeated ten times and mean of ten different performance was used for comparing different methods.

In tests, the comparing methods were compared with each other based on criteria for classification accuracy. Table 2 shows mean of classification accuracy via SVM classifier for different methods of feature selection. As it can be seen in table, in most of datasets, the proposed methods have best performance. For example, in Sonar and DrSVM dataset, the classification accuracy for proposed method is 84.96 and 81.36 respectively. Meanwhile, classification accuracy for Sonar dataset is 76.05 when all features are selected. In this dataset, the proposed method has better performance and when all features are selected, we have lowest classification accuracy. Mean of classification accuracy in all datasets is 88.11, 85.25% for proposed method and DrSVM respectively. Meanwhile, mean of classification accuracy for all datasets is 81.57 when all features are selected.

Table 2: comparing the performance of different methods due to classification accuracy

Dataset	FSVM	DrSVM	SVM
Wine	94.45	91.76	93.27
WDBC	94.67	93.72	92.22
Sonar	84.96	81.36	76.05
Colon	78.39	74.19	64.75
Average	88.11	85.25	81.57

5. Conclusion

Regarding progresses in data collection technology and increasing data storage capabilities in recent decades, high-dimensional dataset is quickly increasing. Support vector machine is one of the supervised learning methods which is used for classification and regression. This method is relatively new method indicated good efficiency in recent years compared to previous methods in classifying such as Perceptron Neural Network. In this paper, Fuzzy support vector machine is used for reducing the dimensions of dataset. Therefore, the classification accuracy is increased via this method. Evaluation of proposed method and comparing its performance with other previous methods showed that the proposed method has proper performance and in most of datasets, it has best performance among different methods.

Refrences

[1] Cadenas, J.M., M.C. Garrido, and R. Martínez, Feature subset selection Filter–Wrapper based on low quality data. Expert Systems with Applications, 2013. 40(16): p. 6241-6252.

[2] Liu, Y. and Y.F. Zheng, FS_SFS: A novel feature selection method for support vector machines. Pattern Recognition, 2006. 39(7): p. 1333-1345.

[3] Xin Sun, et al., Using cooperative game theory to optimize the feature selection problem. Neurocomputing, 2012. 97: p. 86-93.

[4] Aghdam, M.H., N. Ghasem-Aghaee, and M.E. Basiri, Text feature selection using ant colony optimization. Expert Systems with Applications, 2009. 36(3): p. 6843-6853.

[5] Jung-Yi Jiang, Ren-Jia Liou, and S.-J. Lee, A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification. IEEE Transactions Knowledge and Data Engineering, 2011. 23(3): p. 335 - 349

[6] Uğuz, H., A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 2011. 24(7): p. 1024-1032.

[7] Kanan, H.R. and K. Faez, An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system. Applied Mathematics and Computation, 2008. 205(2): p. 716-725.

[8] Vignolo, L.D., D.H. Milone, and J. Scharcanski, Feature selection for face recognition based on multi-objective evolutionary wrappers. Expert Systems with Applications, 2013. 40(13): p. 5077-5084.

[9] Zini, L., et al., Structured multi-class feature selection with an application to face recognition. Pattern Recognition Letters, 2014.

[10] da Silva, S.F., et al., Improving the ranking quality of medical image retrieval using a genetic feature selection method. Decision Support Systems, 2011. 51(4): p. 810-820.

[11] Rashedi, E., H. Nezamabadi-pour, and S. Saryazdi, A simultaneous feature adaptation and feature selection method for content-based image retrieval systems. Knowledge-Based Systems, 2013. 39: p. 85-94.

[12] Inbarani, H.H., A.T. Azar, and G. Jothi, Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Comput Methods Programs Biomed, 2014. 113(1): p. 175-85.

[13] Jaganathan, P. and R. Kuppuchamy, A threshold fuzzy entropy based feature selection for medical database classification. Comput Biol Med, 2013. 43(12): p. 2222-9.

[14] Cheng-Lung Huang and C.-Y. Tsai, A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. Expert Systems with Applications, 2009. 36(2): p. Pages 1529-1539.

[15] Ronen Meiri and J. Zahavi, Using simulated annealing to optimize the feature selection problem in marketing applications. European Journal of Operational Research, 2006. 171(3): p. 842–858.

[16] Raileanu, L.E. and K. Stoffel, Theoretical comparison between the Gini index and information gain criteria. Ann. Math. Artif. Intell. 41, 2004: p. 77-93.

[17] Mitchell, T.M., Machine Learning. McGraw-Hill,NewYork., 1997.

[18] Hanchuan Peng, Fuhui Long, and C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions Pattern Analysis and Machine Intelligence, 2005. 27(8): p. 1226 - 1238

[19] Liu, D., et al., An iterative SVM approach to feature selection and classification in high-dimensional datasets. Pattern Recognition, 2013. 46(9): p. 2531-2537