International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue-17
December 2017

Neighborhood Component Feature Selection for High-Dimensional Data

N.Chandramouli & K.Gouri Charanya,

[1]*Assistant Professor, M.Tech, Department Of Cse,*

Vaageswari College Of Engineering, Karimnagar, T.S, India.

ABSTRACT: *Feature selection is of considerable importance in data mining and machine learning, especially for high dimensional data. In this paper, we propose a novel nearest neighbor-based feature weighting algorithm, which learns a feature weighting vector by maximizing the expected leave-one-out classification accuracy with a regularization term. The algorithm makes no parametric assumptions about the distribution of the data and scales naturally to multiclass problems. Experiments conducted on artificial and real data sets demonstrate that the proposed algorithm is largely insensitive to the increase in the number of irrelevant features and performs better than the state-of the-art methods in most cases..*

Keywords: *feature selection, feature weighting, nearest neighbor.*

I. INTRODUCTION

With the emergence of a great quantity of high dimensional data in various applications, including information retrieval, automated text categorization, combinatorial chemistry and bioinformatics, feature selection has become more and more important in data mining and machine learning. Feature selection is the technique of selecting a small subset from a given set of features by eliminating irrelevant and redundant features. Proper feature selection not only reduces the dimensions of features and hence amount of data used in learning, but also alleviates the effect of the curse of dimensionality to improve algorithms' generalization performance. Furthermore, it also increases the execution speed and the models' interpretability.

Generally speaking, feature selection algorithms now usually fall into one of the three categories: filter, wrapper and embedded methods. In the filter model, feature selection is done by evaluating feature subset with the criterion functions characterizing the intrinsic properties of the training data, such as interclass distance (e.g., Fisher score), statistical measures (e.g., Chi-squared) and information theoretic measures, not involving the optimization of performance of any specific classifier directly. On the contrary, the last two methods are closely related with specified classification algorithms and perform better than filter methods in most cases. The wrapper model requires one predetermined classifier in feature selection and uses its performance to evaluate the goodness of selected feature subsets. Since the classifier need always be trained for each feature subsets considered, wrapper methods are computationally intensive and thus often intractable for large-scale feature selection problems. In the embedded model, feature selection is built into the classifier construction and gradient descent method is usually used to optimize the feature weights, which indicate the relevance between the corresponding features and the target concept. The advantages of the embedded methods are that they are not only less prone to over fitting but also computationally much more efficient than wrapper methods. In particular, many SVM-based embedded methods have been proposed. More comprehensive reviews on feature selection methodologies can be referred.

Nearest neighbor is a simple and efficient nonlinear decision rule and often yields competitive results compared with the state-of-the-art classification methods, such as support vector machines and neural network. Recently, several nearest neighbor-based feature weighting methods, including RELIEF [7], Simba [8], RGS [9], IRELIEF [10], LMFW [11], Lmba [12] and FSSun [13], have been successfully developed and shown the better performance on high-dimensional data analysis. Inspired by the previous work [14], we propose a novel nearest neighbor-based feature selection method called neighborhood component feature selection (NCFS) algorithm. The proposed algorithm uses gradient ascent technique to maximize the expected leave-one-out classification accuracy with a regularization term. Experiments conducted on artificial and real data sets show that NCFS is almost insensitive to the increase in the number of irrelevant features and performs better than Simba, LMFW and FSSun in most cases. The rest of the paper is organized as follows. Section 2 proposes a novel feature selection algorithm based on neighborhood component. Section 3 summarizes some related feature selection approaches. Experiments conducted on toy data and real microarray datasets to evaluate the effectiveness of the proposed algorithm are presented in Section 4. Finally, conclusions are given in Section 5.

II. RELATED WORK

In authors introduced a novel concept, predominant correlation, and propose a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. The efficiency and effectiveness of our method is demonstrated through extensive comparisons with other methods using real-world data of high dimensionality. In [9] authors present a novel concept predominant correlation and propose a new algorithm that can effectively select good features based on correlation analysis with less than quadratic time complexity. A correlation based measure used in this approach. Two approaches classical linear correlation and Information theory are used. The algorithm used is FCBF, Fast correlation based filter. In [10] authors introduced the importance of removing redundant genes in sample classification and pointed out the necessity of studying feature redundancy. And proposed a redundancy based filter method with two desirable properties. It does not require the selection of any threshold in determining feature relevance or redundancy and it combines sequential forward selection with elimination, which substantially reduces the number of feature pairs to be evaluated in redundancy analysis. In [7] authors proposed a new framework of efficient feature selection via relevance and redundancy analysis, and a correlation-based method which uses C-correlation for relevance analysis and both C- and F-correlations for redundancy analysis. A new feature selection algorithm is implemented and evaluated through extensive experiments comparing with three representative feature selection algorithms. The feature selection results are further verified by two different learning algorithms. In [5] authors present an integrated approach to intelligent feature selection. They introduce a unifying platform which serves an intermediate step toward building an integrated system for intelligent feature selection and illustrate the idea through a preliminary system based on research. The unifying platform is one necessary step toward building an integrated system for intelligent feature selection. The ultimate goal for intelligent feature selection is to create an integrated system that will automatically recommend the most suitable algorithm to the user while hiding all technical details irrelevant to an application. In [13] authors present an optimization tool for attribute selection. This paper formulates and validates a method for selecting optimal attribute subset based on correlation using Genetic algorithm, where genetic algorithm used as optimal search tool for selecting subset of attributes. Correlation between the attributes will decide the fitness of individual to take part in mating. Fitness function for GA is a simple function, which assigns a rank to individual attribute on the basis of correlation coefficients. Since strongly correlated attributes cannot be the part of DW together, only those attributes shall be fit to take part in the crossover operations that are having lower correlation coefficients. In [11] authors generalised the ensemble approach for feature selection. So that it can be used in conjunction with many subset evaluation techniques, and search algorithms. A recently developed heuristic algorithm harmony search is employed to demonstrate the approaches. The key advantage of FSE is that the performance of the feature selection procedure is no longer depended upon one selected subset, making this technique potentially more flexible and robust in dealing with high dimensional and large datasets. In [14] authors identify the problems associated with clustering of gene expression data, using traditional clustering methods, mainly due to the high dimensionality of the data involved. For this reason, subspace clustering techniques can be used to uncover the complex relationships found in data since they evaluate features only on a subset of the data. Differentiating between the nearest and the farthest neighbors becomes extremely difficult in high dimensional data spaces. Hence a thoughtful choice of the proximity measure has to be made to ensure the effectiveness of a clustering technique. In [12] authors proposed a framework for feature selection which avoids implicitly handling feature redundancy and turns to efficient elimination of redundant features via explicitly handling feature redundancy. This framework composed of two steps: analysis of relevance determines the subset of relevant features by removing irrelevant ones, and analysis of redundancy determines and eliminates redundant features from relevant ones and thus produces the final subset. A novel clustering based feature subset selection algorithm for high dimensional data.

III. PROPOSED SYSTEM

The research on feature selection is still in process since past two decades. Mutual information (MI) is oftenly used to select a relevant features. Forward selection and backward elimination are the two methods used in the statistical variable selection problem. Forward selection method is utilized in many of the successful FS algorithms in high dimensional data. Backward elimination method is not used in practical application because of huge number of features. A problem with the forward selection method is, change in a decision of the initial feature, which creates a different features subset and varies in the stability. It is known as stability problem in FS.

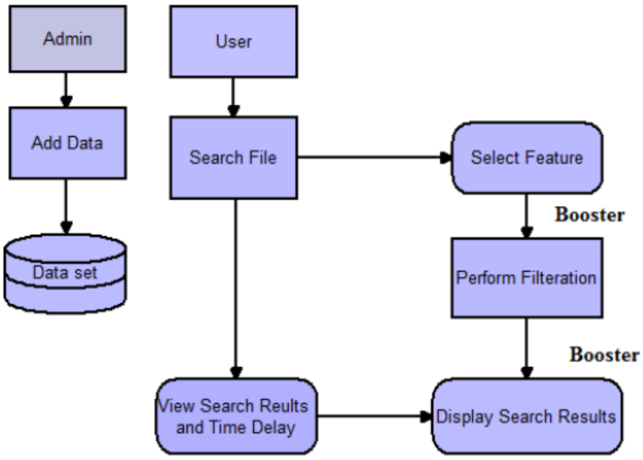

Figure 2: System Architecture

The application is all about a search engine. Here the user can search for the file which is required. Before searching it is necessary that the file should be present in the database. So for that reason admin will add the file into the database which further creates the data set. Once the file is stored in database a user can search for the required file. When the user search for a file multiple operations are performed. It first selects the feature, perform the filtration on the selected feature and then display the search result. Here Booster helps to obtain the results as soon as possible, it works from the time of search till the results are displayed. And a Q-statistic measure shows that how many number of files are present related with the search keyword which we have inserted. Booster also shows the time delay to extract the file. After obtaining the searched file, a user can download the document or images, he can also view the time delay and file count.

IV. DATA SOURCE

For the purposes of evaluating the performance and effectiveness of our proposed FAST algorithm, verifying whether or not the method is potentially useful in practice, and allowing other researchers to confirm our results, 35 publicly available data sets were used. The numbers of features of the 35 data sets vary from 37 to 49152 with a mean of 7874. The dimensionality of the 54.3% data sets exceed 5000, of which 28.6% data sets have more than 10000 features. The 35 data sets cover a range of application domains such as text, image and bio microarray data classification.

Feature selection in loss-based classification

Feature selection in loss-based classification As mentioned above, variable selection-capable penalty functions such as the L1 and SCAD can be applied to the regularization framework to achieve variable selection when dealing with data with many predictor variables. Examples include the L1 SVM (Zhu, Rosset, Hastie, and Tibshirani, 2003), SCAD SVM (Zhang, Ahn, Lin, and Park, 2006), SCAD logistic regression (Fan and Peng, 2004). These methods work fine for the case with a fair number of predictor variables. However the remarkable recent development of computing power and other technology has allowed scientists to collect data of unprecedented size and complexity. Examples include data from microarrays, proteomics, functional MRI, SNPs and others. When dealing with such high or ultra-high dimensional data, the usefulness of these methods becomes limited. In order to handle linear regression with ultra-high dimensional data, Fan and Lv (2008) proposed the sure independence screening (SIS) to reduce the dimensionality from ultra-high p to a fairly high d. It works by ranking predictor variables according to the absolute value of the marginal correlation between the response variable and each individual predictor variable and selecting the top ranked d predictor variables. This screening step is followed by applying a refined method such as the SCAD to these d predictor variables that have been selected. In a fairly general asymptotic framework, this simple but effective correlation learning is shown to have the sure screening property even for the case of exponentially growing dimensionality, that is, the screening retains the true important predictor variables with probability tending to one exponentially fast. The SIS methodology may break down if a predictor variable is marginally unrelated, but jointly related with the response, or if a predictor variable is jointly uncorrelated with the response but has higher marginal correlation with the response than some important predictors. In the former case, the important feature has already been screened out at the first stage, whereas in the latter case, the unimportant feature is ranked too high by the independent screening technique. Iterative SIS (ISIS) was proposed to overcome these difficulties by using more fully the joint covariate information while retaining computational expedience and stability as in SIS. Basically, ISIS works by iteratively applying SIS to recruit a small number of predictors, computing residuals based on the model fitted using these recruited variables, and then using the working residuals as the response variable to continue recruiting new predictors. Numerical examples in Fan and Lv (2008) have demonstrated the improvement of ISIS. The crucial step is to compute the working residuals, which is easy for the least-squares regression problem but not obvious for other problems. By sidestepping the computation of working residuals, Fan et al. (2008) has extended (I)SIS to a general pseudo-likelihood framework, which includes generalized linear models as a special case. Roughly they use the additional contribution of each predictor variable given the variables that have been recruited to rank and recruit new predictors.

V. EXPERIMENTAL SETUP

To evaluate the performance of our proposed FAST algorithm and compare it with other feature selection. algorithms in a fair and reasonable way, we set up our experimental study as follows. 1) The proposed algorithm is compared with five different types of representative feature selection algorithms. Relief searches for nearest neighbors of instances of different classes and weights features according to how well they differentiate instances of different classes. The other three feature selection algorithms are based on subset evaluation. CFS exploits best-first search based on the evaluation of a subset that contains features highly correlated with the tar-get concept, yet uncorrelated with each other. The Consist method searches for the minimal subset that separates classes as consistently as the full set can under best-first search strategy. FOCUS-SF is a variation of FOCUS [2]. FOCUS has the same evaluation strategy as Consist, but it examines all subsets of features. Considering the time efficiency, FOUCS-SF replaces exhaustive search in FOCUS with sequential forward selection.

Four different types of classification algorithms are employed to classify data sets before and after feature selection. They are (i) the probability-based Naive Bayes (NB), (ii) the tree-based C4.5, (iii) the instance-based lazy learning algorithm IB1, and (iv) the rule-based RIPPER, respectively. Naive Bayes utilizes a probabilistic method for classification by multiplying the individual probabilities of every feature-value pair. This algorithm assumes independence among the features and even then provides excellent classification results. Decision tree learning algorithm C4.5 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. The tree comprises of nodes (features) that are selected by information entropy. Instance-based learner IB1 is a single-nearest-neighbor algorithm, and it classifies entities taking the class of the closest associated vectors in the training set via distance metrics. It is the simplest among the algorithms used in our study. Inductive rule learner RIPPER (Repeated Incremental Pruning to Produce Error Reduction) is a propositional rule learner that defines a rule based detection model and seeks to improve it iteratively by using different heuristic techniques. The constructed rule set is then used to classify new instances.

When evaluating the performance of the feature subset selection algorithms, four metrics, (i) the proportion of selected features (ii) the time to obtain the feature subset, (iii) the classification accuracy, and (iv) the Win/Draw/Loss record, are used. The proportion of selected features is the ratio of the number of features selected by a feature selection algorithm to the original number of features of a data set. The Win/Draw/Loss record presents three values on a given measure, i.e. the numbers of data sets for which our proposed algorithm FAST obtains better, equal, and worse performance than other five feature selection algorithms, respectively. The measure can be the proportion of selected features, the runtime to obtain a feature subset, and the classification accuracy, respectively

VI. CONCLUSION

In this paper, we present a novel feature weighting method in the context of NN. The proposed method, which is called NCFS, uses the gradient ascent technique to maximize the expected leave-one-out classification accuracy with a regularization term. The effectiveness of this algorithm has been evaluated through a number of experiments involving a toy data and eight microarray datasets. Meanwhile, the impact of two parameters, the kernel width σ and the regularization parameter λ, has been studied empirically. Overall, the proposed method is insensitive to a specific choice of the two parameters.

REFERENCES

[1] H. Liu, E. Dougherty, J. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons et al., “Evolving feature selection,” IEEE Intelligent Systems, vol. 20, no. 6, pp. 46–76, 2005.

[2] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Feature Extraction: Foundations and Applications. SpringerVerlag, 2006.

[3] S. Maldonado, R. Weber, and J. Basak, “Simultaneous feature selection and classification using kernel-penalized support vector machines,” Information Sciences, vol. 18, pp. 115–128, 2011.

[4] J. Miranda, R. Montoya, and R. Weber, “Linear penalization support vector machines for feature selection,” Pattern Recognition and Machine Intelligence, pp. 188–192, 2005.

[5] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” The Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[6] Y. Saeys, I. Inza, and P. Larra˜naga, “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.

[7] K. Kira and L. Rendell, “A practical approach to feature selection,” in Proceedings of the ninth international workshop on Machine learning, 1992, pp. 249–256.

[8] R. Gilad-Bachrach, A. Navot, and N. Tishby, “Margin based feature selection-theory and algorithms,” in Proceedings of the twenty-first international conference on Machine learning, 2004, pp. 43–50.

[9] A. Navot, L. Shpigelman, N. Tishby, and E. Vaadia, “Nearest neighbor based feature selection for regression and its application to neural activity,” in Advances in Neural

Information Processing Systems 18. MIT Press, 2006, pp. 995–1002.

[10] Y. Sun, “Iterative relief for feature weighting: Algorithms, theories, and applications,” IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 1035– 1051, 2007.

[11] B. Chen, H. Liu, J. Chai, and Z. Bao, “Large margin feature weighting method via linear programming,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 10, pp. 1475–1488, 2009.

[12] Y. Li and B. Lu, “Feature selection based on loss-margin of nearest neighbor classification,” Pattern Recognition, vol. 42, no. 9, pp. 1914–1921, 2009.

[13] Y. Sun, S. Todorovic, and S. Goodison, “Local learning based feature selection for high dimensional data analysis,” IEEE transactions on pattern analysis and machine intelligence, vol. 99, 2009.

[14] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, “Neighbourhood components analysis,” in Advances in Neural Information Processing Systems 17. MIT Press, 2005, pp. 513–520.

[15] K. Weinberger, J. Blitzer, and L. Saul, “Distance metric learning for large margin nearest neighbor classification,” in Advances in Neural Information Processing Systems 18. Cambridge, MA: MIT Press, 2006, pp. 1473–1480.