Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Feature Selection Algorithms: Literature Review

P. Manikandan
Assistant Professor,
Department of Computer Application,
Thanthai Hans Roever College,
Perambalur, India

Dr. C. Jothi Venkateswaran
Dean & Head
Department of Computer Science,
Presidency College,
Chennai, India

Abstract: ***Feature selection is a term usually use in data mining to demonstrate the tools and techniques available for reducing inputs to a convenient size for processing and analysis. In this paper authors has reviewed the literature of feature selection algorithms such as well known attributes selection methods of FCBF, ReliefF, SVM-RFE, Random selection. This review of literature focuses on how feature selection techniques are used for different dataset.***

Keywords: Feature selection algorithms; literature review; subset evaluation;

I. INTRODUCTION

Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the choice of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis.

II. LITERATURE REVIEW

The importance of two separate kinds of knowledge, image dependent knowledge and image independent knowledge is discussed. Second, we develop an adaptive method for refining the feature set. This adaptive method analyzes the feature error to locate areas of the image that would lead to false matches. Then these areas are used to guide the search through feature space towards maximizing the class separation distance between the correct match and the false matches. Third, we introduce a self-diagnostic method for determining when apriori knowledge is necessary for finding the correct match. If the apriori knowledge is necessary then we use a surface reconstruction model to discriminate between match possibilities. Our algorithm is comprehensively tested against fixed feature set algorithms and against a traditional pyramid algorithm Michael S, ect..[9].

These approaches are simultaneous clustering and feature selection using aniching memetic algorithm. This approach makes feature selection an integral part of the global clustering search procedure and attempts to overcome the problem of identifying less promising locally optimal solutions in both clustering and feature selection, without making any a priori assumption about the number of clusters, Weiguo Sheng, etc[1]. Within the NMA_CFS procedure, a variable composite representation is devised to encode both feature selection and cluster centers with different numbers of clusters. Furthermore, local search operations are introduced to refine feature selection and cluster centers encoded in the chromosomes. Finally, a niching method is integrated to preserve the population diversity and prevent premature convergence. In an experimental evaluation, these demonstrate the effectiveness of the proposed approach by using both synthetic and real data.

Antonio Mucherino [2], the concept of β-consistent biclustering has been introduced for dealing with noisy data and experimental errors. However, the given definition for β-consistent biclustering is coherent only when sets containing non-negative data are considered. This paper extends the definition of β-consistent biclustering to negative data and shows, through computational experiments, that the employment of the new definition allows to perform better classifications on a well-known test problem.

SVM are attractive for the classification of remotely sensed data with some claims that the method is insensitive to the dimensionality of the data and so not requiring a dimensionality reduction analysis in pre-processing. This highlights a dependency of the accuracy of classification by a SVM on the dimensionality of the data and so the potential value of undertaking a feature selection analysis prior to classification. the accuracy derived from the use of a small number of features may be non-inferior to that derived from the use of a larger feature set providing potential advantages in relation to issues such as data storage and computational processing costs. Feature selection may, therefore, be a valuable analysis to include in pre-processing operations for classification by a SVM Mahesh Pal and Giles M. Foody [3].

Gustavo Camps-Valls [5] the Hilbert–Schmidt norm of the cross-covariance operator of mapped samples in the corresponding Hilbert spaces. The HSIC empirical estimator is easy to compute and has good theoretical and practical

© 2015, IJARCSSE All Rights Reserved

properties. Rather than using this estimate for maximizing the dependence between the selected features and the class labels, we propose the more sensitive criterion of minimizing the associated HSIC p-value. Results in multispectral, hyperspectral, and SAR data feature selection for classification show the good performance of the proposed approach.

The key idea is to decompose an arbitrarily complex nonlinear problem into a set of locally linear ones through local learning, and then learn feature relevance globally within the large margin framework. The proposed algorithm is based on well-established machine learning and numerical analysis techniques, without making any assumptions about the underlying data distribution. It is capable of processing many thousands of features within minutes on a personal computer, while maintaining a very high accuracy that is nearly insensitive to a growing number of irrelevant features. Theoretical analyses of the algorithm's sample complexity suggest that the algorithm has a logarithmical sample complexity with respect to the number of features. Experiments on eleven synthetic and real-world data sets demonstrate the viability of our formulation of the feature selection problem for supervised learning and the effectiveness of our algorithm Yijun Sun, ect..[4].

Bjorn Waske, ect..[11] MCSs based on SVM and random feature selection (RFS) are applied to explore the potential of a synergetic use of the two concepts. We investigated how the number of selected features and the size of the MCS influence classification accuracy using two hyper spectral data sets, from different environmental settings. In addition, experiments were conducted with a varying number of training samples. Accuracies are compared with regular SVM and random forests. Experimental results clearly demonstrate that the generation of an SVM-based classifier system with RFS significantly improves overall classification accuracy as well as producer's and user's accuracies. In addition, the ensemble strategy results in smoother, i.e., more realistic, classification maps than those from stand-alone SVM. Findings from the experiments were successfully transferred onto an additional hyper spectral data set.

Qi Cheng,ect..[6] A logistic regression (LR) model may be used to predict the probabilities of the classes on the basis of the input features, after ranking them according to their relative importance. In this letter, the LR model is applied for both the feature selection and the classification of remotely sensed images, where more informative soft classifications are produced naturally. The results indicate that, with fewer restrictive assumptions, the LR model is able to reduce the features substantially without any significant decrease in the classification accuracy of both the soft and hard classifications.

The proposed object recognition strategy in this paper has two main stages: single view and multi views processes Fatemeh Tabib Mahmoudi , etc..[7]. In the single view process, defining region's properties for each of the segmented regions, the object-based image analysis (OBIA) is performed independently on the individual views. In the second stage, the classified objects of all views are fused together through a decision-level fusion based on the scene contextual information in order to refine the classification results. Sensory information, analyzing visibility maps, height, and the structural characteristics of the multi views classified objects define the scene contextual information. Evaluation of the capabilities of the proposed context aware object recognition methodology is performed on two datasets: 1) multi angular Worldview-2 satellite images over Rio de Janeiro in Brazil and 2) multi views digital modular camera (DMC) aerial images over a complex urban area in Germany. The obtained results represent that using the contextual information together with a decision-level fusion of multi views, the object recognition difficulties and ambiguities are decreased and the overall accuracy and the kappa are gradually improved for both oftheWorldView-2 and the DMC datasets.

Pablo A. Estévez, ect..[8] A filter method of feature selection based on mutual information, called normalized mutual information feature selection (NMIFS), is presented. NMIFS is an enhancement over Battiti'sMIFS, MIFS-U, and mRMR methods. The average normalized mutual information is proposed as a measure of redundancy among features. NMIFS outperformed MIFS, MIFS-U, and Mrmr on several artificial and benchmark data sets without requiring a user-defined parameter. In addition, NMIFS is combined with a genetic algorithm to form a hybrid filter/wrapper method called GAMIFS. This includes an initialization procedure and a mutation operator based on NMIFS to speed up the convergence of the genetic algorithm. GAMIFS overcomes the limitations of incremental search algorithms that are unable to find dependencies between groups of features.

Lorenzo Bruzzone, ect..[10] This approach results in a more robust classification system with improved generalization properties with respect to standard feature-selection methods. The feature selection is accomplished by defining a multi objective criterion function made up of two terms: 1) a term that measures the class separability and 2) a term that evaluates the spatial invariance of the selected features. In order to assess the spatial invariance of the feature subset, we propose both a supervised method (which assumes that training samples acquired in two or more spatially disjoint areas are available) and a semi supervised method (which requires only a standard training set acquired in a single area of the scene and takes advantage of unlabeled samples selected in portions of the scene spatially disjoint from the training set). The choice for the supervised or semi supervised method depends on the available reference data. The multi objective problem is solved by an evolutionary algorithm that estimates the set of Pareto optimal solutions. Experiments carried out on a hyper spectral image acquired by the Hyperion sensor on a complex area confirmed the effectiveness of the proposed approach.

III. CONCLUSION

In this literature review are appearance of some well-known feature selection algorithms such as SVM, logistic regression, FCBF, ReliefF, SVM-RFE, Random selection, Best first search method, greedy stepwise search method and ranker search method. Many research papers are used develop, modifying of algorithms and evaluate subset selection.

© 2015, IJARCSSE All Rights Reserved

REFERENCE

[1] Weiguo Sheng, Xiaohui Liu, and Michael Fairhurst "A Niching Memetic Algorithm for Simultaneous Clustering and Feature Selection" IEEE transactions on knowledge and data engineering, vol. 20, no. 7, July 2008.

[2] Antonio Mucherino "Extending the definition of β-consistent biclustering for feature selection" Proceedings of the Federated Conference on Computer Science and Information Systems pp. 269–274, 2011 IEEE.

[3] Mahesh Pal and Giles M. Foody "Feature selection for classification of hyperspectral databy SVM" IEEE Transactions on Geoscience and Remote Sensing, 48, 2297-2307 (2010).

[4] Yijun Sun, Sinisa Todorovic, and Steve Goodison Local Learning Based Feature Selectionfor High Dimensional Data Analysis" IEEE transactions on pattern analysis and machine intelligence, vol. X, no. X, xx 20xx (2009).

[5] Gustavo Camps-Valls, J. Mooij and B. Schölkopf "Remote Sensing Feature Selection byKernel Dependence Measures" IEEE geoscience and remote sensing letters, vol. 7, no. 3, July 2010.

[6] Qi Cheng, Pramod K. Varshney, and Manoj K. Arora "Logistic Regression for Feature Selection and Soft Classification of Remote Sensing Data" IEEE geoscience and remote sensing letters, vol. 3, no. 4, October 2006.

[7] Fatemeh Tabib Mahmoudi, Farhad Samadzadegan, Peter Reinartz "Object Recognition Based on the Context Aware Decision-Level Fusion in Multiviews Imagery" IEEE journal of selected topics in applied earth observations and remote sensing 2014.

[8] Pablo A. Estévez, Michel Tesmer, Claudio A. Perez, Jacek M. Zurada "Normalized MutualInformation Feature Selection" IEEE transactions on neural networks, vol. 20, no.2, February 2009.

[9] Michael S. Lew. Thomas S. Huang, Kam Wong "Learning and Feature Selection in Stereo Matching" IEEEE transactions on pattern analysis and machine intelligence, vol. 16, no. 9, September 1994.

[10] Lorenzo Bruzzone, Claudio Persello, "A Novel Approach to the Selection of SpatiallyInvariant Features for the Classification of Hyperspectral Images With ImprovedGeneralization Capability" IEEE transactions on geoscience and remote sensing, vol. 47,no. 9, September 2009

[11] Bjorn Waske, Sebastian van der Linden,Jón Atli Benediktsson, , Andreas Rabe, and Patrick Hostert, "Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyper spectral Data" IEEE transactions on geoscience and remote sensing, vol. 48, no. 7, July 2010.

© 2015, IJARCSSE All Rights Reserved